TIME AND CAUSALITY IN GENOMICS DATA

by
Rossin James Erbe

A dissertation submitted to Johns Hopkins University in conformity with the requirements

for the degree of Doctor of Philosophy

Baltimore, Maryland
February 2023

Abstract

The ability to sequence the genomic information that describes individual cell states has provided enormous insight into biological systems. However, to sequence the genomic information within a cell, the cell must be killed, preventing measurements from the future states that cell would have occupied had it been allowed to survive. Thus, sequencing measurements only provide a single snapshot in time of cellular genomic states. Often the ultimate goal of an analysis is to derive mechanistic insight into the biology of a system or process from the data. However, such mechanistic, causal inference is almost impossible without temporal information because causality in standard formulations is based on the concept of connected causes and effects through time.

This thesis has interacted with time in genomics data in several ways. The first contribution of this thesis is a neural network-based model that attempts to predict future single-cell transcriptomic states from single-cell transcriptomics data sets. This work demonstrates that using metabolic labeling data sets, future RNA states are estimable within the same cell in the short term, providing a proof of principle that can be expanded as genomics data sets with a temporal dimension become more common.

The second contribution of this thesis is a simulation of molecular cell states over time, which is able to demonstrate how single time points from cells do not allow for robust mechanistic inference. Further, the simulation conforms to observations that mRNA expression and expression of the corresponding protein are often poorly correlated and provides mechanistic explanations for how this occurs.

The final contribution relates to time in a different sense, analyzing the impact of human age on biomarkers used for cancer immunotherapy. We found that older individuals possessed a number of favorable biomarkers at higher levels than their younger counterparts, possibly explaining clinical observations that older individuals do no worse than younger individuals on immune checkpoint therapies despite the usual anticorrelation between patient age and effective immune responses.

Primary Reader and Advisor: Elana J. Fertig

Secondary Reader: Joel S. Bader

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The ideal outcome of most genomic studies is to uncover the mechanisms that underly the biological system of interest. As the first step in my efforts to understand how to robustly infer mechanism from genomics data I performed an in depth literature review of methods that use graphical networks to describe the casual regulatory relationships that control biological systems. This review was published in Volume 82, Issue 2 of *Molecular Cell* and is reprinted here with permission as an introduction to this thesis.

A vast web of interdependent molecular interactions governs biological systems and allows organisms to function. This network of interactions is highly complex, involving reactions at many molecular scales (e.g. from the level of genes to the level of cells) (Schaffer and Ideker, 2021). To effectively model such complex systems, it is worthwhile to examine the many molecular levels from which one might approach this challenge (Figure 1.1). At the molecular scale of the gene, researchers often attempt to understand the tens of thousands of different genes that drive the biological operations of complex multicellular life. Even when studied in isolation, understanding the function of each of these genes is a monumental task, and many human genes have not yet been extensively characterized (Su and Hogenesch, 2007), (Stoeger et al., 2018). Moreover, genes do not act in isolation; their function is inextricably tied to the rest of the biological system. For example, transcription factors (TFs) concurrently regulate the expression of multiple genes, possibly even the gene coding for the regulating TF itself. The expression of each gene as an RNA product is thought to be primarily controlled by its epigenetic state and the activity of regulatory proteins and functional RNAs (Harmston and Lenhard, 2013), (Bhan and Mandal, 2014). However, the exact nature of this relationship is still not well characterized for most genes. The protein products of these genes likewise often cannot be well understood in isolation, but must be placed in a network of other interacting proteins to accomplish a cellular task such as signal transduction, catalysis, or molecular transport. Post-transcriptional modifications and functional non-coding RNAs further impact cellular function and introduce another plethora of interactors that may be involved in a given cellular process (Cech and Steitz, 2014), (Yao et al., 2019), (Kuijjer et al., 2020). Each cell then interacts with other cells in the wider context of a microenvironment, a tissue, and the organism as a whole.

The regulatory complexity underlying biological processes and disease demonstrates the challenges of accurately modeling these systems. Bulk and single-cell profiling technologies are now commonly used to provide insight about the variety of molecular and cellular actors in biological processes. These technologies generate high-dimensional data sets that require specialized computational methodologies to interpret (Davis-Marcisak et al., 2021). Thus, the growth in molecular profiling technologies has been mirrored by the advance of a wide variety of machine learning methods for high-throughput data analysis. This review describes machine learning methods for high-throughput data analysis that are designed to model the interactions between biological effectors such as genes, proteins, metabolites, and cells. We focus on methods that are predominantly based on graphical networks (Figure 1.2, Table 1.1), which explicitly model the interactions or regulatory relationships (called edges) between nodes (molecular effectors such as genes, proteins, metabolites, or cells).

1.1 Gene network inference aims to capture the mechanistic regulatory relationships underlying gene expression

A wide variety of computational methodologies have been developed for gene regulatory network inference, a graphical network modeling approach to elucidating gene function and regulation. The ultimate goal of gene network inference is to uncover the regulatory biology of a particular system, often as it relates to a pathological phenotype. Graphical network methods have been designed to predict interactions algorithmically based on high-throughput molecular data, prior experimental knowledge, or a combination of the two. The resulting networks can be analyzed to yield humanly interpretable insights about the biological system under study from a convoluted web of molecular interactions (Figure 1.3). Network metrics called centrality measures (Table 1.1), which have been used widely for analysis of webpage and social networks, can also be applied to biological network inference. For these biological applications, network metrics can be calculated to identify key nodes (e.g. genes or proteins) in a system that may act as regulatory hubs controlling the biological process being studied, though their usefulness for this purpose in biological networks still requires thorough experimental validation. Another strategy for identifying critical parts of the network are optimization algorithms such as PCSF (Akhmedov et al., 2017) and SAMNet (Gosline et al., 2012), which have been applied in biological networks to find a smaller subsection of the network containing the nodes and edges with the largest regulatory influence in the data. Networks can also be used to generate specific mechanistic hypotheses by examining the causal predictions made by network structure. For example, if a network predicts that a specific gene regulates a set of genes that are all thought to contribute to a disease phenotype, that gene could be predicted as a molecular target to treat said phenotype. Thus, the structure of the network implies that node to be a potentially useful target, due to its regulatory relationship with several other implicated factors. In this way, the goal of these graphical network methods is to distill relatively simple insights from the immense complexity of biological systems.

1.2 Accurately modeling biological systems using gene network inference requires thorough consideration of experimental design

Gene network inference methods have been developed to predict regulatory interactions based upon the dependencies between genes in both bulk and single-cell expression data (Nguyen et al., 2021), (Mercatelli et al., 2020). The regulatory networks that can be inferred depend on the biological context and study design for the genomics data that are input to the network inference methods. The biological context is critical to consider because it is impossible to infer regulatory information about systems that are not active in the samples used to produce the data. For example, many of the regulatory processes of cell division will not be possible to infer from data derived from quiescent cells. Additionally, highly stable systems will be difficult to glean much regulatory information from - if a gene's expression undergoes close to zero variance in a data set, the genes that exert regulatory control of it will not be able to be determined.

The choice of bulk or single-cell data also impacts regulatory network inference. The main drawback of bulk data is that if it is drawn from a heterogenous mixture of cells, the expression signal from different cell types may be difficult to distinguish. The relevant regulatory interactions in different cell types may be different due to differing epigenetic landscapes, thus confounding the regulatory signal from the data. Single-cell sequencing allows individual cell types to be modeled separately, but technical dropouts (specifically genes that were expressed in the cell but zero counts are returned from sequencing due to measurement error) introduce additional challenges for predicting accurate regulatory relationships between genes because one cannot be sure if a zero occurs because of regulatory control or measurement error. In cases with known biological networks, these structures can be embedded in single-cell analysis algorithms to enhance data analysis (Elyanow et al., 2020).

1.3 Computational methods infer gene interactions through undirected networks and causal regulatory mechanisms via directed networks

After the experiment has been performed, computational methods are needed to infer regulatory networks from the resulting high-throughput datasets. The approaches for gene network inference can be generally classified into those that produce undirected networks - the interactions predicted between genes do not specify which is the regulator and which is the target - and directed networks, which attempt to make that distinction computationally (Figure 1.2). Additionally, a wide array of visualization tools have been developed that further support the network-based interpretation and inference of high-throughput datasets, notably the Cytoscape platform (Shannon et al., 2003), (Otasek et al., 2019).

Among undirected network inference methods, the foundational approach uses Pearson correlation statistics between the expression values of pairs of genes to predict regulatory relationships between genes (Stuart et al., 2003). While this may appear to be a simplistic approach, correlation based methods have been found to recover known regulatory interactions better than more complex methods on several data sets (Stone et al., 2021). However, they come with the caveat that genes with correlated expression are not necessarily functionally related. To overcome this limitation, another method utilizes the concept of Mutual Information, which measures how much one can know about the expression of gene X, given that you know the expression of gene Y. This method is popularly employed by the ARACNE algorithm (Margolin et al., 2006). Partial information decomposition (PIDC) has also been applied to refine results to functional interactions between genes (Chan et al., 2017). Partial information decomposition is used to measure statistical dependencies between three variables. This is applied to gene network inference by calculating the unique information between genes X and Y, divided by the information provided by every other gene Z in the data set. The results of which are used to determine the confidence the algorithm places in a regulatory link between X and Y.

The approaches described above for gene network inference all produce undirected networks: they estimate whether pairs of genes have a regulatory interaction between them but do not predict which gene is the target and which is the regulator. Therefore, these approaches require prior knowledge of gene regulation (e.g. which genes are known TFs) to distinguish the directionality of regulatory relationships. To handle cases in which prior information is unavailable or incomplete, another class of regulatory inference algorithms has been developed to infer directed networks without this reliance on prior biological knowledge. A prominent method that has performed well relative to other methods at recapitulating experimentally determined regulatory interactions, GENIE3, uses ensembles of decision trees to predict the likelihood of a regulatory link between genes based on how useful the expression of gene X is in predicting the expression of gene Y (Huynh-Thu et al., 2010), (Aibar et al., 2017). Decision tree ensembles can be thought of as a model that learns many general “rules of thumb” (e.g. when gene A is above expression level X, gene B is almost always above expression level Y) about the system they are employed to predict. From those many rules, a single consensus prediction is made by a vote among all the trees (do they predict that gene A can in general be used to predict gene B?). The degree to which a gene can predict another is returned as a score of how confident the method is in the regulatory link between two genes.

The measurement noise and molecular noise in transcription introduce technical variation in gene expression datasets (Tunnacliffe and Chubb, 2020), often propagating to the inferred network. Therefore, other approaches aim to concurrently infer a directed network while reducing the noise from the input expression data. The scTenifoldNet method first produces a baseline directed network using principal components (PC) regression (Osorio et al., 2020). PC regression performs principal components analysis (PCA), which decomposes the expression data into new variables (PCs) that describe the data’s uncorrelated sources of variance. These PCs can then be used to predict the expression of each gene in turn. Based on the value of each PC for predicting a target gene’s expression, an inference can be made about the effect of each other gene on the target gene’s expression. The resulting gene interaction network does not yet correct for technical variation in gene expression data. Therefore, this process is repeated for subsamples of the total expression data. Several networks are thus produced, the agreement between which can be used to determine which parts arise from technical variation and which correspond to regulatory biology (Osorio et al., 2020).

The methods introduced thus far are generally intended to analyze gene expression data collected from a single time point. However, datasets with measurements of gene expression over time can enhance the inference of directed networks. Expression changes in one gene that precede or follow another can better implicate a causal relationship than estimates made from a single point in time. Therefore, other approaches have been developed to model gene interactions as a system of equations with respect to time. The set of putative regulators of a gene can be determined and used to produce equations that predict how a gene’s expression values will change over time. These equations can then be solved and related to time course data through mathematical approaches such as differential equations. While using such methods with bulk RNA-seq data requires explicit time course data, transitions in cellular state that occur over time can be estimated computationally from single-cell datasets using trajectory inference methods, providing a pseudo-temporal framework in which to use these methods (Trapnell et al., 2014), (Saelens et al., 2019). Single-cell regulatory inference algorithms such as SCODE have been developed to perform temporal modeling using differential equations based on trajectory estimates of cell-state transitions from single-cell RNA-seq (Matsumoto et al., 2017). However, differential equations require models of the biological mechanisms through which genes interact, which may be unknown a priori and lack sufficient data to parameterize. Therefore, several other network inference methods instead perform statistical tests of whether the time series for one gene forecasts another (Granger Causality), again based upon trajectory estimates from single-cell data, such as SINCERITIES (Papili Gao et al., 2018) and SINGE (Deshpande et al., 2019). The Scribe method is capable of using any time-ordered set of single-cell data as input and uses an estimation of causality from information theory called directed information to identify direct regulatory links between genes (Qiu et al., 2020). These methods thus yield a network that is intended to account for changes in cell state over time in its regulatory predictions.

1.4 Benchmarking the accuracy of gene regulatory networks enables selection of inference methodologies and priorities for new algorithm development

With this wide array of network inference methods, standards for judging their relative merits are fundamental. Benchmarking computational algorithms requires applying them to datasets with a known ground truth state in order to assess performance. The two main approaches generally used for benchmarking gene network inference algorithms are based on either simulated datasets with known network structure or regulatory databases that contain experimentally determined interactions.

Simulated benchmarks use a pre-defined network structure to simulate what expression profiles might look like given a known set of regulatory interactions. In some cases, gene expression datasets are simulated based upon randomized network structures. In these cases, algorithm performance is typically benchmarked in multiple simulations to test the variance of performance for a given network structure and sensitivity across a range of network parameters. However, the simulated networks may not reflect the structure of true biological networks. In other cases, the networks used in these simulated datasets are based on prior biological knowledge of gene interactions. For example, GeneNetWeaver uses a known network of regulatory interactions (such as one that has been fully experimentally determined in S.cerevisiae or E.coli) to estimate how expression of gene products would change over time according to a system of equations that allows for both additive and multiplicative regulatory interactions (Schaffter et al., 2011). Such simulations provide a very clean way of benchmarking network inference methods because all regulatory relationships are already known and the data only contains as much noise as is introduced purposefully by the researchers to maintain biological realism. Methods can be scored against how many of the known regulatory interactions each correctly predicts without concerns about whether this reference of interactions might be incomplete or incorrect.

However, benchmarking methods in the context they must ultimately be used in (experimental expression data) is desirable to robustly demonstrate a model's effectiveness, especially since the assumptions necessary to simulate data may bias them in a way that does not reflect real biological systems. More complex networks of interactors with more and less predictable sources of noise will usually provide a more accurate representation of the context in which these methods will be applied. Furthermore, the performance of a method has been shown to sometimes differ substantially between simulated and experimental tests (Pratapa et al., 2020). Predictions from network inference methods are most commonly validated against ChIP-seq, ChIP-chip, and gene perturbation experiments. Often non-specific databases of gene interactions are used for these evaluations, and thus the context (cell type, epigenetic state, metabolic state) in which the interaction was determined may not be the same as in the data set that the gene network inference method is applied to. Generally, this limitation can be minimized by examining only highly variably expressed genes in a data set. Then, if a gene is not undergoing regulation or is epigenetically repressed, the method will not try to predict its regulators, since that information does not exist in the data. However, there may be cases in which genes are variably expressed, but are capable of acting in other, currently inactive, processes, which may lead to the appearance of the network method failing to identify a regulatory link that could have been inferred from the data. In this way, experimental benchmarking has more potential to incorrectly label a network as having generated false negative results, but provides a more realistic context than simulation based benchmarks. These benchmarks are also expected to be incomplete descriptions of gene regulatory networks, which may additionally lead to incorrect identification of regulatory interactions as false positives. Ideally, algorithms should be tested on both types of benchmarks as each can reveal distinct properties of algorithm performance.

One of the main reasons such a wide variety of network inference approaches have been developed is because different approaches perform better at reconstructing experimentally determined and simulated regulatory interactions in different data sets and contexts. Furthermore, no single method is currently capable of achieving a universally high prediction accuracy across simulated or experimental benchmark data sets, based on several independent assessments (Greenfield et al., 2010), (Chen and Mar, 2018), (Pratapa et al., 2020), (Stone et al., 2021). Across these evaluations, the PIDC and GENIE3 (or methods based on GENIE3) methods have been pointed out as performing particularly well at capturing experimentally determined interactions in real expression data, though even these generally well performing methods occasionally yield poor performances (Greenfield et al., 2010), (Chen and Mar, 2018), (Pratapa et al., 2020), (Stone et al., 2021). The high error rates observed on some data sets could be plausibly attributed to any or all of the following factors: the need to include multiple omics data sets to improve predictions, the need for more robust algorithms to distinguish direct interactions from indirect interactions between genes and their products, molecular noise in transcription levels, measurement noise from sequencing, and methodological problems with the benchmarks used. Community-wide data science challenges, including notably the Dialogue on Reverse Engineering Assessment and Methods (DREAM) challenge, have been developed to facilitate widespread validation of network inference methods from simulated and experimental datasets (Marbach et al., 2012), (Hill et al., 2016). Standardizing datasets for benchmarking enables robust comparison of methods against a common ground truth and facilitates the independence of simulated datasets from the assumptions used in developing an algorithm (Camacho et al., 2018).

1.5 Inference of multi-scale intracellular networks requires multi-omics analysis methods

While building a regulatory graphical network from a high-throughput transcriptional data set is a highly complex endeavor, it is still a considerable simplification of cellular processes. Within each cell, the DNA sequence, chromatin conformation, epigenetic modifications, gene expression, protein expression, protein modifications, and metabolites form a complex web of causal factors that produce cellular phenotypes (Figure 1.1). These multi-scale processes are more accurately modeled from multi-omics datasets that characterize these molecular scales (Table 1.3). In particular, elucidating the entire chain of causality by which cellular processes generate a phenotype of interest requires following events across different molecular levels (Schaffer and Ideker, 2021) (Figure 1.4). Additionally, inferring networks from only a single data modality can lead to identifying interactions that appear to be only conditionally valid, due to the differing epigenetic context of the cell. For example, in gene network inference, using data from one cellular context, a method may correctly identify a TF-gene regulatory link, but with data from a different context may fail to identify the same relationship. This could occur because the gene's promoter in the second case was in a heterochromatic conformation and not accessible for the TF to bind or because a genetic variant altered TF binding affinity. Incorporating variant calling and epigenetic data could help resolve such problems in gene network inference, particularly as technologies to profile transcriptomics and chromatin state from the same single cell become more widely available.

Integrating prior knowledge of transcription factor targets, either from databases or binding assays, can be used to refine inference of TF-gene regulation from expression data. An approach called BETA integrates ChIP-seq of TFs and expression data to infer TF-gene regulation (Wang et al., 2013). BETA predicts both whether a TF is activating or repressing gene expression and which genes are the TF's direct targets, based on the statistical relationship between TF binding and differential gene expression. Similarly, the post-hoc statistics can be applied to matrix factorization to incorporate existing databases of TF regulation and patterns in gene expression to score the context-specific TF regulation of genes, which can be used to identify genes that are co-regulated or that are regulated by multiple TFs (Fertig et al., 2013). Both of these methods attempt to discover the regulatory structure of biological systems using multiple data types through transcriptional regulatory networks, and serve as an important foundation as similar methods are developed for emerging single-cell data sets.

In the case of genetic variants, an approach has been developed to determine the impact of individual genetic variation on gene expression networks using EGRET (Weighill et al., 2021). The authors reason that given the substantial proportion of functional genetic variants that appear to mediate their effect via differences in gene expression (Zhu et al., 2016), gene regulatory networks may differ between individuals in important ways. EGRET builds a general gene network based on prior knowledge, experimental TF cooperativity, and gene expression data, which it can then update based on genetic variant data to produce a different gene network for each individual. The mechanistic regulatory impact of genetic variants can be thus inferred, which the authors validate using cell lines with known genetic differences (Weighill et al., 2021).

While multi-omics analysis can provide a more complete description of cellular processes, it also introduces several new challenges for analysis (Lê Cao et al., 2021). In the context of regulatory networks, the most immediate challenge is in combining the information across multiple data modalities into a single network. Alternatively, networks could be defined separately from each data modality, but then it is necessary to address a similar challenge: how to model the interactions between those separate networks. One approach that has been developed to address this type of problem is the field of multilayer networks (Kivela et al., 2014), which formulate networks with distinct layers that each contain nodes of a specific type. This framework is applied by (Liu et al., 2020) using large-scale databases to produce a multilayer network containing one layer each for genes, proteins, and metabolites. The multilayer network thus produced was shown to be robust at recovering the importance of genes that are required for cellular function or had been annotated as critical cancer genes.

Predicting causal relationships between molecular effectors is also more complex when multiple levels of molecular effectors are involved, due to the need to account for possible interactions both within and between modalities. Determining the order of cause and effect is also an even larger challenge in this context, especially when processes such as gene regulation are often cyclic, making many of the best-developed causal inference frameworks, such as directed acyclic graphs (Pearl, 1995), unusable. One possible approach is to model only acyclic processes, but ignoring feedback loops in biological systems will often omit substantial information. Technical variation arising from different sources of noise, variance, or batch effects across the different data modalities also must be accounted for to avoid biasing results.

The COSMOS method (Dugourd et al., 2021) attempts to navigate the many obstacles of multi-omics network modeling with an approach based on prior knowledge and their previously developed method for network analysis within a single data modality (Liu et al., 2019). COSMOS finds prior knowledge networks that provide relationships between transcriptomics, phosphoproteomics, and metabolomics data using the OmniPath protein-protein and gene regulatory interaction database (Türei et al., 2016) and the Recon3D metabolomics database (Brunk et al., 2018). These prior knowledge networks are then refined by removing interactions that create incorrect predictions when applied to the transcriptomics, proteomics, and metabolomics data sets provided as input. It then removes interactions that lead to incoherent predictions (e.g. two molecules that should be correlated end up being anti-correlated). The network is further filtered based on the expression differences observed in the biological context of interest, which yield a set of genes, proteins, and metabolites that are differentially regulated. These molecular effectors are used to produce the final network, which only includes nodes (genes, proteins, or metabolites) a set number of regulatory steps away from the differentially regulated starting nodes. COSMOS is additionally incapable of forming loops, which is a possible limitation of the method, but also allows causal analysis to be applied with much less difficulty. From this network, the regulatory effects of the perturbed molecules can be causally inferred. A statistical test for gene set analysis can then be applied to determine whether the genes identified in the network are annotated to pathways with known biological relevance (Dugourd et al., 2021). COSMOS is limited to producing a subnetwork connected to differentially regulated molecules, which appears to be a strength in that it focuses the method on relevant biological differences. Yet it also creates the inability to reach relevant molecular effectors that are either more distantly regulated or are not included in the prior knowledge network. COSMOS appears to be a significant step in network modeling across multi-omics data sets. However, more work still needs to be done to robustly model the wide variety of regulatory interactions that control biological systems at a multiscale molecular level.

1.6 Intercellular networks model signaling between cells, altering intracellular dynamics and producing large-scale phenotypes

In isolation, even a highly robust model of the internal operations of a single cell would often be insufficient to characterize many phenotypes, due to the importance of intercellular signaling. For example, intercellular signaling has been shown to be critical for cellular differentiation (Kirouac et al., 2010), (Basson, 2012), organ homeostasis

(Arneson et al., 2018), (Wang et al., 2020), the cellular response to aging (Ximerakis et al., 2019), and the cellular response to disease (Fernandez et al., 2019), particularly cancer (Vaske et al., 2010), (Kumar et al., 2018), (Baghban et al., 2020). In general, the collective processes and interactions of many cells produce the tissue-scale and organism-scale phenotypes that are the primary focus of biomedical research. Thus, characterizing these interactions as a graphical network model provides a valuable framework to understand many phenotypes of interest in terms of the interactions of the cells that produce them.

Several methods have been developed to model cell-cell interactions, generally in the form of ligand-receptor interactions at the cell surface. Many models also include predictions of the downstream effects these interactions will have within the cells involved. Here we will again focus on those methods that model regulatory interactions explicitly as graphical networks. Generally, these methods produce a score of cell and receptor interactions, and then model the effects these interactions will have on the expression of genes regulated downstream of the receptors (Wang et al., 2019), (Browaeys et al., 2020), (Cherry et al., 2021). This feature allows these methods to describe the impact of intercellular interactions on intracellular processes, which seems likely to be a necessary feature to fully understand many cellular phenotypes. However, none of the methods thus developed are able to model interactions between the downstream signaling effects of multiple different receptors, which may be a significant limitation in some circumstances.

NicheNet uses prior knowledge of ligand-receptor interactions and gene regulatory networks along with bulk or single-cell transcriptomics data to predict activated receptors (Browaeys et al., 2020). These predictions are made using a personalized PageRank metric, an adaptation of the method developed by Google to score and rank web pages in their search engine (Page et al., 1998). Here, it is instead used to produce a score for ligand-receptor interactions. Another method, SoptSC, approaches the problem with a greater emphasis on cell clusters, taking single-cell expression data and a set of known receptors and their cognate ligands as input to calculate the similarity between each cell's expression profile. This information is compiled into a matrix, from which the method creates a cell-cell interaction network and clusters the cells, ultimately allowing inferences of signaling pathways activated between cell clusters (Wang et al., 2019). DOMINO similarly emphasizes cell clusters, but focuses more on TF activity as well as receptor-ligand activation (Cherry et al., 2021). DOMINO uses the results of SCENIC (Aibar et al., 2017), a method that builds on GENIE3 to score TF activity, combined with prior knowledge networks of ligand-receptor pairs to determine interactions between cell types and the activated ligands and TFs within each cell type. NATMI takes a slightly different approach (Hou et al., 2020), focusing on learning interactions between cells using bulk or single-cell expression data, not addressing the specifics of how these regulatory interactions impact downstream gene expression. NATMI uses large-scale ligand-receptor databases to create prior knowledge networks. It then calculates a weight for each interaction between genes, based on three expression based metrics from the data set of interest, which are used to determine cell type interactions (Hou et al., 2020).

Due to the wide array of cell-cell interactions that play roles in cancer (Kumar et al., 2018), (Baghban et al., 2020), CCCExplorer was developed specifically for use with tumor data (Choi et al., 2015). CCCExplorer identifies differentially expressed ligands in cells in the tumor microenvironment as well as expressed receptors on tumor cells. It then uses expression data from tumor cells to find expressed TFs, combining prior knowledge of each TF’s regulated genes to determine the probability that the corresponding pathway is active. These data are combined to identify active signaling branches, which are further combined to generate a crosstalk network. This network is used to identify regulations between the tumor microenvironment and the tumor cellular phenotype (Choi et al., 2015).

While the methods discussed thus far all produce predictions of cell-cell interactions, an important consideration is often the question of where cells are interacting in a particular tissue, which may be highly relevant to phenotype. To account for this, SpaOTsc maps single-cell transcriptomics to spatial data sets (such as in situ hybridization) and uses the spatial element to inform the prediction of cell-cell interactions and how these impact gene regulation (Cang and Nie, 2020). This is accomplished using partial information decomposition, which calculates the statistical dependencies between three variables (e.g. is gene A important to the relationship between genes B and C?), and ensembles of decision trees, which in effect combine many “rules of thumb” that are computationally learned from the data to produce consensus predictions (Table 1.4).

1.7 Overview of Validation of Network Models for Biological Insight

In order to ensure that computational models are capable of generating robust biological insight for users, they must be thoroughly tested for accuracy and biological relevance. This is particularly essential given the complexity of network analysis for high-throughput profiling data. Several different strategies have been employed for validation of biological network methods, each with strengths and weaknesses.

Simulated data is generally the first test a network method is subjected to and is produced by assuming a particular network structure and generating data using a mathematical model (e.g. if we know Gene 1 upregulates Gene 2 which downregulates Gene 3, what might expression data from this system look like based on what we know about the dynamics of gene regulation?). The strengths of simulated data tests are that the correct network is known as a certainty and it is quick and inexpensive to do large numbers of tests across different contexts. However, these simulations must rely on machine-coded assumptions to generate data sets. When the assumptions of the model do not adequately conform to the biological processes being simulated, they can produce output that lacks some characteristics of genuine data sets.

Another strategy for benchmarking uses databases of interactions that are known to occur in an organism, then scores the model against the number that it identifies correctly when tested on real biological data. While this approach has the advantage of working with the sort of data the method is intended to be used on in practice, performance assessments will be biased by the incompleteness of existing databases. Furthermore, in cases that the database was generated from data that came from a different context than the data input to the network model, the network may, correctly, not identify some context dependent interactions and be penalized incorrectly.

Finally, mechanistic experiments can be performed in the same context as the data fed into the network model, providing the most reliable feedback on the usefulness of a model at identifying biologically relevant regulatory interactions. While providing a gold standard, the large number of perturbations required for high-throughput validation can make such efforts both cost and time prohibitive at a genome-wide scale. However, a limited set of experiments can greatly increase the confidence given to other predicted interactions made by the network model, if those tested are validated.

1.8 Applications of network methods enable computational prediction of perturbations at scale and prioritization of targets for validation

Biologically, the value of gene regulatory network inference is that it can be used to discover interactions between genes. Producing this comprehensive understanding of the regulatory mechanisms of a biological system allows for the application of additional computational techniques to predict the impact of interventions on phenotypes (Sonawane et al., 2019), (Belyaeva et al., 2020). These methods use the network to go beyond associating variables to predicting the experimental results of an intervention to the biological system (e.g. a perturbation). Understanding the mechanistic contribution of a single gene to a particular biological process or phenotype is often the work of years or even decades using traditional experimental tools. Network methods may be able to aid investigations about the role of genes and their products in biological systems by generating *in silico* hypotheses regarding the mechanistic impact of altering gene expression levels (Figure 1.5). This information can guide candidate prioritization and selection for more highly time-intensive experiments to accelerate mechanistic biological discovery.

While some of this information can be provided by high-throughput knock-out screening methods such as Perturb-seq (Dixit et al., 2016), the reasons why a particular knockout has the impact that it does may still be opaque after such experiments. The advantage of gene network analyses is that they can provide both a prediction of the end result of a perturbation and a mechanistic account of why that result was produced, which may be critical for fully understanding biological processes and rational drug design.

The scTenifoldKnk method aims to computationally predict gene knockout (KO) experiments using what the authors term virtual KO screens (Osorio et al., 2021). The method produces a directed gene regulatory network using single-cell transcriptomics data from unperturbed cells by applying their scTenifold network inference method, as described in (Osorio et al., 2020). The virtual knockout is then performed using the adjacency matrix (Table 1.1). A gene is “knocked out” by setting the entries for the target gene to zero. This creates a version of the network in which the gene is no longer acting, simulating the results of a KO. The two networks are then compared, which can be used to evaluate which genes will be differentially expressed as a result of the gene KO. Within these putative differentially expressed genes, scTenifoldKnk searches for enrichment of known gene sets. The authors show that the gene sets found to be enriched in these virtual KO differentially expressed genes are often related to the known biology of the system being studied. For example, genes predicted to be perturbed by a *CFTR* gene KO are enriched for ABC transporter disorder and abnormal surfactant secretion pathways, which would be expected given the known functions of the *CFTR* gene. This capacity to predict differential expressed genes enriched in pathways that would be expected based on the known function of a gene is shown across several different cellular contexts (Osorio et al., 2021). The authors additionally perform more direct experimental validation of the predicted differentially expressed genes. When they perform an experimental KO of *Malat1* in mouse pancreatic cells, they predict 167 perturbations in other genes. However, only four of those predictions overlap with the 1695 experimental differentially expressed genes they found between the WT and KO cells (Osorio et al., 2021). This result indicates that while the general biological significance of a KO may be recovered by the method (e.g. it predicts there will be shifts in pathways that are known to be associated with the biology of the system), the precise transcriptomic effects are not. This result suggests further development of such methods will be required to achieve the ideal of establishing a robust causal model of gene network interactions that can make accurate predictions of the transcriptomic effects of a gene KO.

CellBox is another method designed to predict experimental results computationally. However, instead of predicting the results of a gene KO, CellBox is designed to predict the results of drug perturbations on phenotypes of interest. It uses bulk proteomics data from drug perturbation experiments, in which a phenotype of interest was measured, to fit a system of ordinary differential equations. These equations can then be used to predict the phenotypic effect of unseen drug treatments and combinations of drug treatments (Yuan et al., 2021). CellBox provides the functionality of varying drug concentration as well as treatment type, which can allow many more permutations to be predicted than would normally be experimentally feasible. CellBox thus allows a small amount of drug screening data to be generalized to predict the outcome of arbitrary drug dosages and combinations at a network and phenotypic level. The authors highlight the potentially great value of CellBox for evaluating combination therapies for cancer. Oncology may be a specifically useful application of such methods due to both the potential for drug synergy as well as the logic that the more therapeutics a tumor has to evolve resistance to in order to survive, the less likely resistance is to develop (Bayat Mokhtari et al., 2017). If each drug requires a separate genetic or epigenetic event for a cancer cell to acquire resistance to it, it will be much less likely to undergo sufficient evolutions to evade being killed by the effects of at least one of the treatments.

Given the complex regulatory relationships that exist in tumor cells and their cellular microenvironment, oncology is a field in which graphical network models may be particularly valuable. A recent study by (Zhou et al., 2021) leverages both gene regulatory and cell-cell interactions models to analyze single-cell RNA-seq data from triple-negative breast cancer patients, providing an informative example of how network methods can be applied at multiple biological levels to glean insights into complex systems. The authors used the CellPhoneDB method (Efremova et al., 2020) to identify ligand-receptor pairs from their data, from which they were able to determine the dominant regulatory role of macrophages in the tumor microenvironment of the patients studied, particularly noting EGFR-amphiregulin interactions in patients with basal-like tumors. The study further constructed TF-target based gene regulatory networks using GENIE3 (Huynh-Thu et al., 2010), which they analyzed via centrality metrics, measures of node (in this case gene) importance that are generally in some way based upon how many regulatory interactions a gene is involved in. Some centrality measures only account for direct interactions, while others include information about how many interactions the interactors of a node have as well. These centrality measures were used to predict critical genes, capturing known important genes such as *MYC* and identifying *ETV6* as an activated critical gene across all subtypes. This use of centrality metrics is a simple and highly useful approach to identifying key nodes that may warrant further examination and experimental testing of their importance in the biological system of study.

Experimental validation is critical to ensuring the reliability of computational methods and is particularly important when dealing with highly complex models such as network methods. An advantage of the network methods and subsequent predictions is that these analyses can prioritize candidate targets for validation experiments. Such validation can yield much greater confidence in the ability of a method to capture the underlying biology of a system or phenotype. The study describing CCCExplorer provides an excellent example of this type of validation (Choi et al., 2015). CCCExplorer predicted that the high *IL6* expression in tumor-associated macrophages in their data activated the IL6 receptor on the tumor cells, activating the STAT3 pathway. They established an *in vitro* system of macrophages and the same type of tumor cells in which tumor conditioned media upregulated *IL6* in wild type macrophages, which in turn increased phosphorylated-STAT3 levels in the tumor cells more than ten-fold. Additionally, macrophages with *IL6* knocked out did not upregulate phosphorylated-STAT3 in the tumor cells. This kind of validation experiment is able to demonstrate the ability of a computational method to not only capture already known interactions, but to identify novel relationships that have important effects on the biological system of study. This sort of validation is critical to establish sufficient confidence in these methods beyond computational benchmarking so that they can begin to help guide experimental planning and therapeutic development.

1.9 Future Directions for Network Biology Research

Graphical network methods provide a model to understand the complexity and sheer number of interacting molecular effectors that contribute to cellular and organism level phenotypes. Progress is ongoing and many improvements have been made in the ability of these methods to model the relationships between molecular effectors and translate these regulatory models into meaningful insights into biological systems and the phenotypes they produce. These methods allow researchers to identify regulatory controls active within a cell, which can be used to generate hypotheses about how to manipulate a biological process to treat disease. Given the complexity of biological systems, such insights may in some cases be extremely difficult to achieve without a model capable of containing many of the molecular effectors at play.

Network methods currently are used to yield insights into regulatory biology, protein and metabolic interactions, intercellular interactions, and how this molecular web translates into phenotypes. However, there are still several areas in which significant further research is warranted. One of the highest priority areas is the fact that gene network inference methods often still do not perform reliably in benchmark experiments on either experimental gold standards or simulated data sets (Chen and Mar, 2018), (Pratapa et al., 2020), (Stone et al., 2021), as discussed in the *Benchmarking the accuracy of gene regulatory networks enables selection of inference methodologies and priorities for new algorithm development* subsection. Predicting whether genes are causally interacting or merely correlated, dealing with transcriptional and measurement noise, and cellular heterogeneity still pose major challenges for the field. Identifying strategies for handling these issues is a crucial area of ongoing research. Another important problem in the field is that few existing network methods integrate across omics data sets. Many approaches do not include multi-omics data for reasons of complexity, computational capacity, or data availability. While a challenging problem, incorporating information across molecular scales is necessary to accurately model the regulatory biology of many cellular processes and diseases. Finally, many studies also do not provide experimental validation of the novel predictions their methods make. Though such validation requires substantial investments of researchers' time and resources, if a method is intended to generate hypotheses worthy of further investigation, such validation seems critical to providing users the confidence to plan experiments based on a computational method's predictions.

While this review primarily focuses on more recently developed algorithms for emerging single-cell technologies, several foundational methods developed for older microarray and bulk profiling technologies have continued relevance for analyses of these emerging datasets. The solid foundation of mathematical insight into how to model biological interactions has allowed these models to continue to be useful even as network methods are updated and refined. We note that (Camacho et al., 2018), (Sonawane et al., 2019) provide additional reviews of a range of computational methodologies for biological network methods, providing greater detail on the methods, while we focus more on their specific biological applications in this review. Some of the major recent developments in network modeling have been based around accounting for technical features of biological data sets, such as sources of noise and heterogeneity e.g. (Osorio et al., 2020), as well as providing tools to more easily ascertain the biological significance of network models e.g. (Aibar et al., 2017).

As algorithms and validation develop to accurately model disease and biological systems with network methods, they have the potential to become more powerful tools for therapeutic development. Much of the time required to develop new treatments or discover the main drivers of some biological process is spent finding a relatively high confidence target and understanding the mechanism of action. Thus, prioritizing functional candidates through network methods could significantly improve the speed of preclinical studies for therapeutic development and studies exploring pathways and complex interacting mechanisms in biological systems.

Figures and Tables

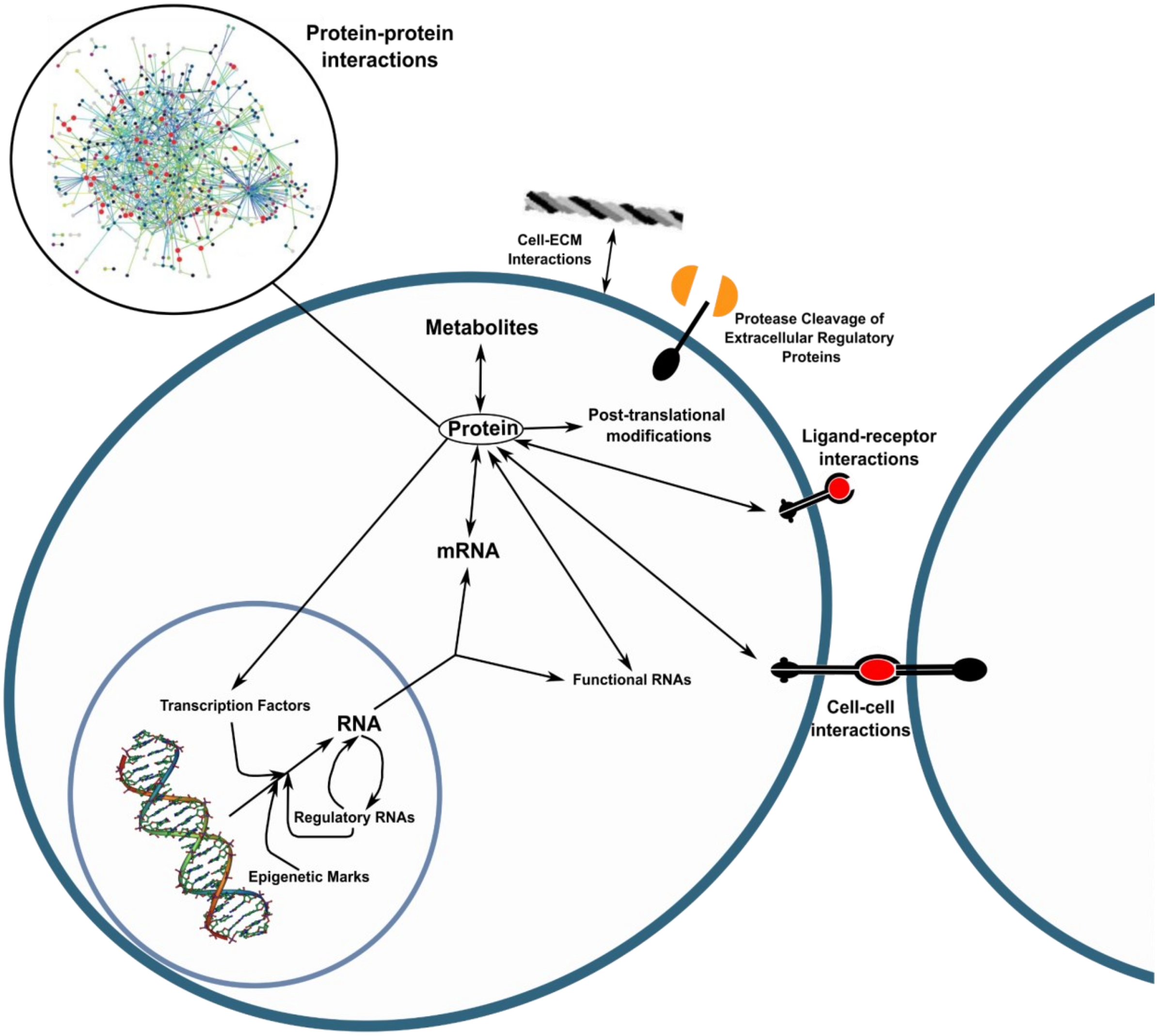
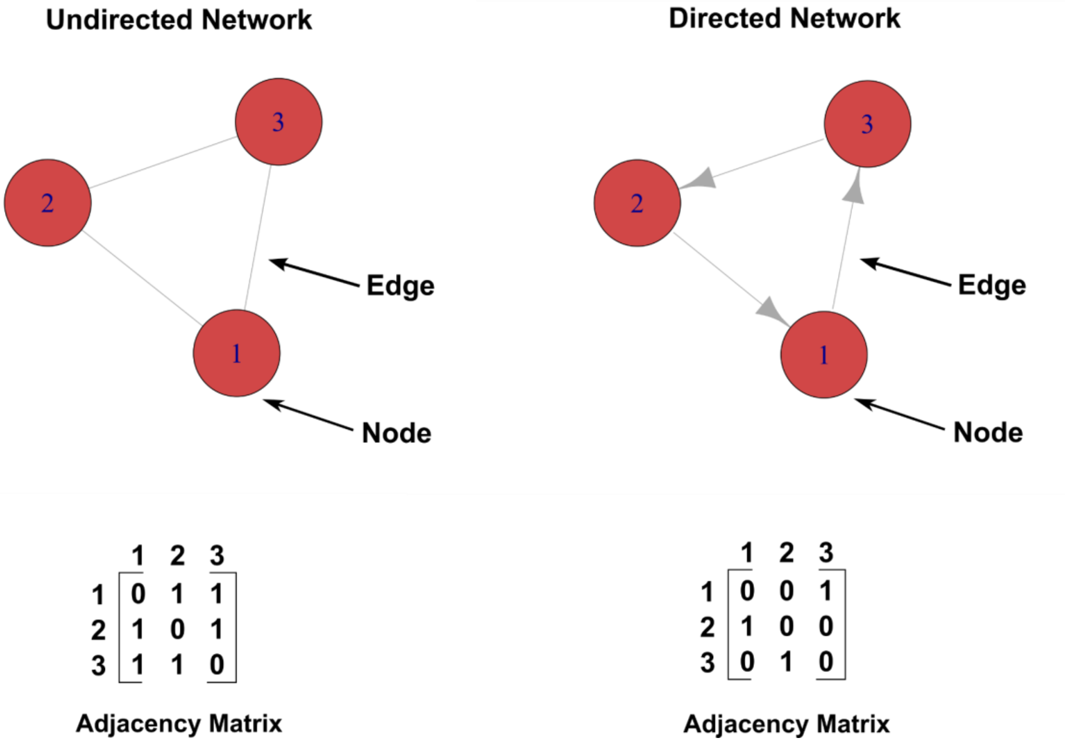

Protein-protein interactions
Cell-ECM Interactions
Protease Cleavage of Extracellular Regulatory Proteins
Metabolites
Protein
Post-translational modifications
Ligand-receptor interactions
mRNA
Functional RNAs
Cell-cell interactions
Transcription Factors
RNA
Regulatory RNAs
Epigenetic Marks

Figure 1.1 Molecular Interactors in Biological Systems - Diagram of the interactions across molecular scales that are involved in the biological processes between and within cells, including insoluble regulatory proteins and interactions with the extracellular matrix (ECM). A protein-protein interaction network is shown in the top left, demonstrating the interactive complexity that can exist within a single molecular scale. Includes as components DNA Overview 2 by Michael Ströck, licensed under Creative Commons CC

BY-SA 3.0, The protein interaction network of Treponema pallidum by Titz et al., licensed under Creative Commons CC BY 1.0, a cropped version of Collagen biosynthesis by GKFK, licensed under Creative Commons CC BY-SA 3.0, and a cropped version of Ligand-receptor interaction by Rit Rajarshi, licensed under Creative Commons CC BY-SA 4.0.



$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 1 \\ 3 & 1 & 1 & 0 \end{array}$$

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 0 & 0 & 1 \\ 2 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \end{array}$$

Figure 1.2 Directed and Undirected Graphs and their Adjacency Matrices - Diagram of the basic structure of an undirected and directed network graph. Each is composed of nodes (which in biological systems generally represent molecules such as genes or proteins) connected by edges (which in biological systems generally represent regulatory or direct functional relationships). Undirected networks only assert that a

relationship exists between nodes, and this relationship is presented as symmetric. This feature is reflected in the symmetric adjacency matrix, a matrix representation of the network. In row 1, the given values are 0, 1, and 1, indicating that node 1 is not connected to itself, but is connected to nodes 2 and 3. The columns can be read the same way for undirected networks, hence the symmetry of the matrix. Directed networks, by contrast, assert the directionality of the relationship between nodes. In biological networks, this is often intended to indicate that one node is the regulator and the other node is the target. The corresponding adjacency matrix is read slightly differently, where each row indicates the edges going out from that node, while each column represents the edges coming in. Thus, the values 0, 0, 1 in row 1 indicate that node 1 has an edge going into node 3, but not the other two nodes.

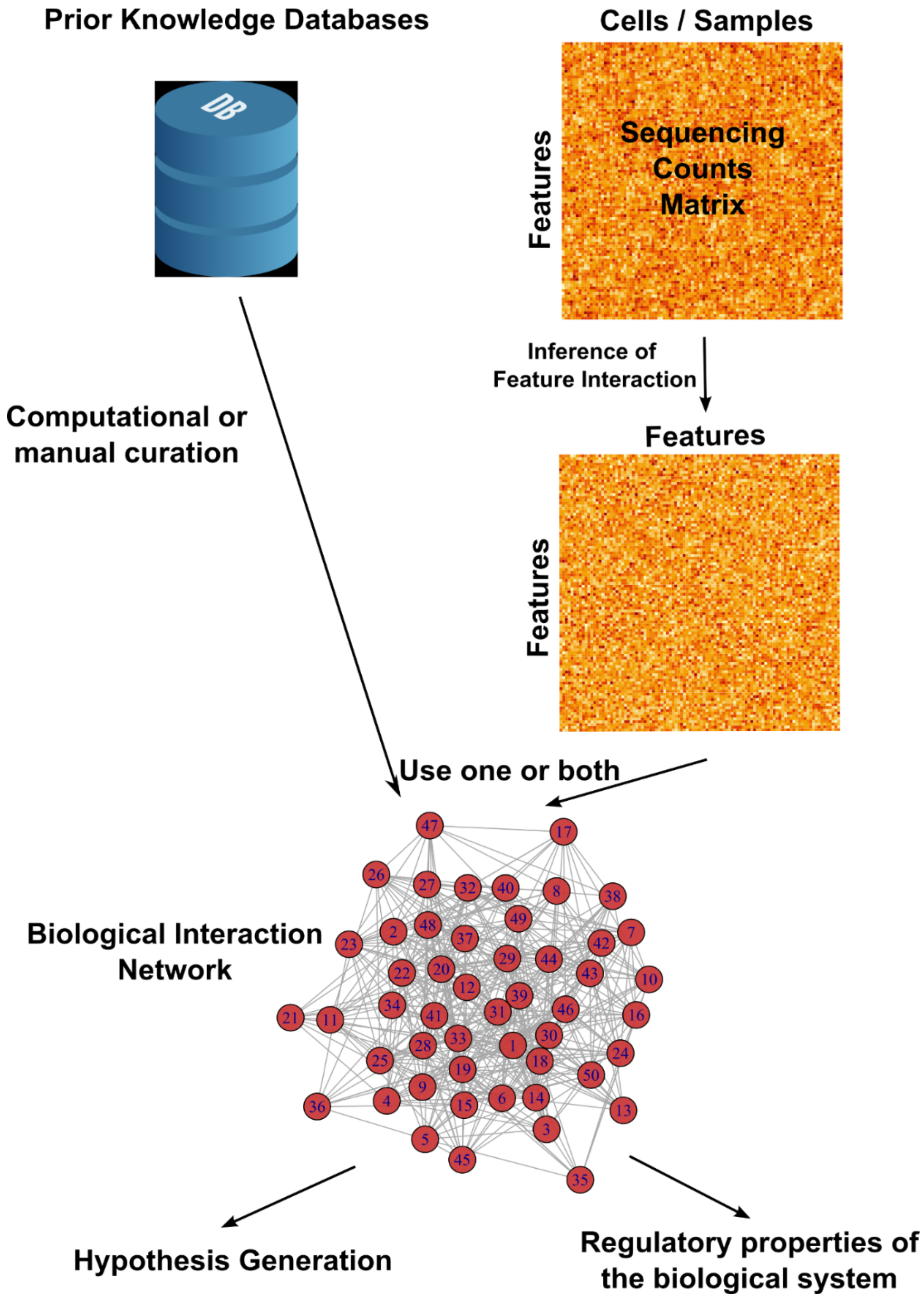


Figure 1.3 Building a Biological Network Graph - Graphical network models are generally created using prior knowledge databases, high-throughput molecular data, or some combination of both. Molecular data is usually summarized as a sequencing counts or

abundance matrix describing features such as genes, proteins, or sequencing peaks present in each cell or sample. An algorithm is then applied to this data to determine the likelihood that these features regulate one another. We diagram these feature correspondence predictions as a feature by feature matrix with each element of the matrix giving the confidence of the algorithm in an interaction between two molecular features. These predictions can then either be used in isolation or combined with prior knowledge of feature interactions (which are generally computationally or manually curated to suit the particular application) to produce a graphical network of interactions underlying a biological system.

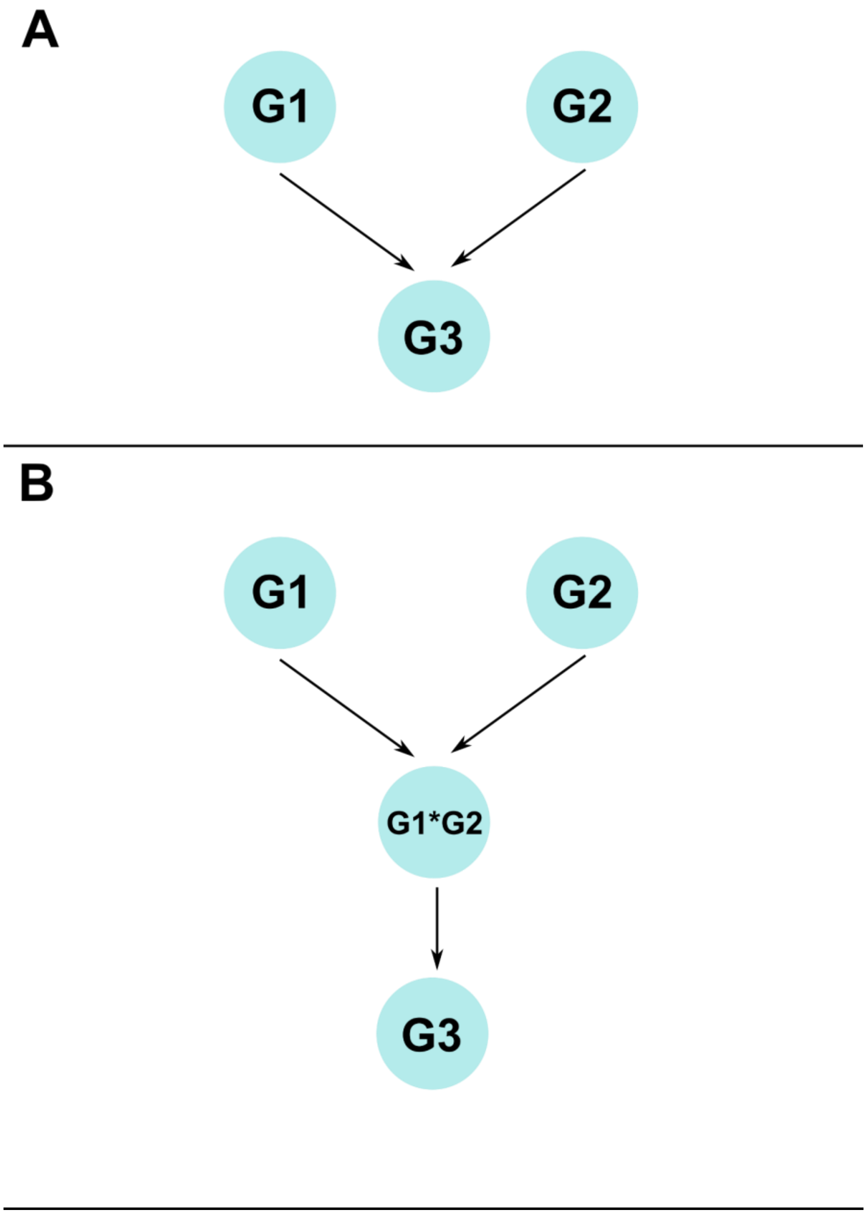
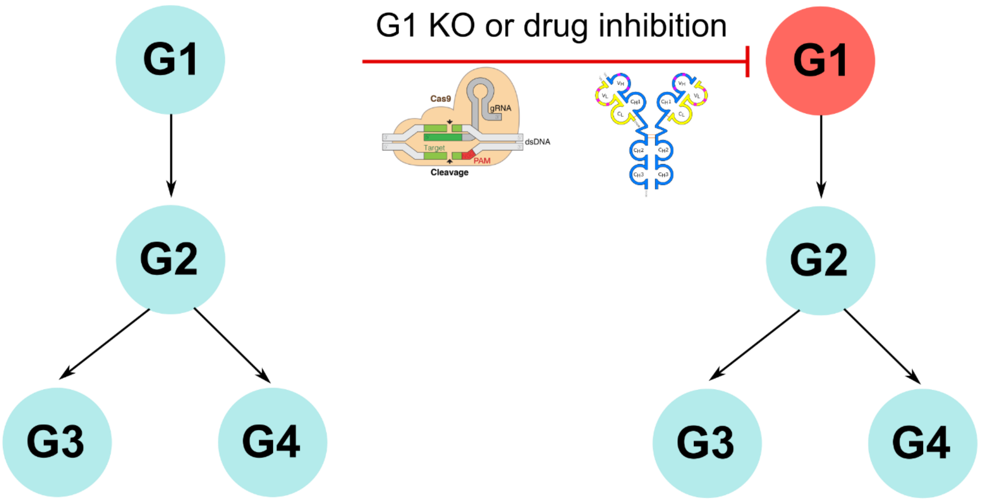


Figure 1.4 Multi-scale models are necessary to capture some biological interactions - Panel A shows a possible example of a gene regulatory structure in which two genes, G1 and G2, both regulate a third gene, G3. In general this situation poses no particular

problem for gene network inference. However, if the regulation of G3 requires both G1 and G2 to be expressed for either regulatory effect to occur, panel A does not adequately describe the regulatory relationships between these three genes. If G1 is expressed and G2 is not, the regulatory link from G1 to G3 is then spurious, as is the link from G2 to G3 in the opposite situation. However, when both genes are expressed, both links appear valid. Panel B demonstrates a network that can capture this possible regulatory structure, in which the products of G1 and G2 form a complex (G1*G2), which is the direct regulator of G3. Including combinations of gene products as nodes creates a multi-scale network, which will exponentially increase the number of possible interactions to consider, the necessary cost of dealing with the type of regulatory behavior given in this example.



Trancriptional effect propagates through the network:

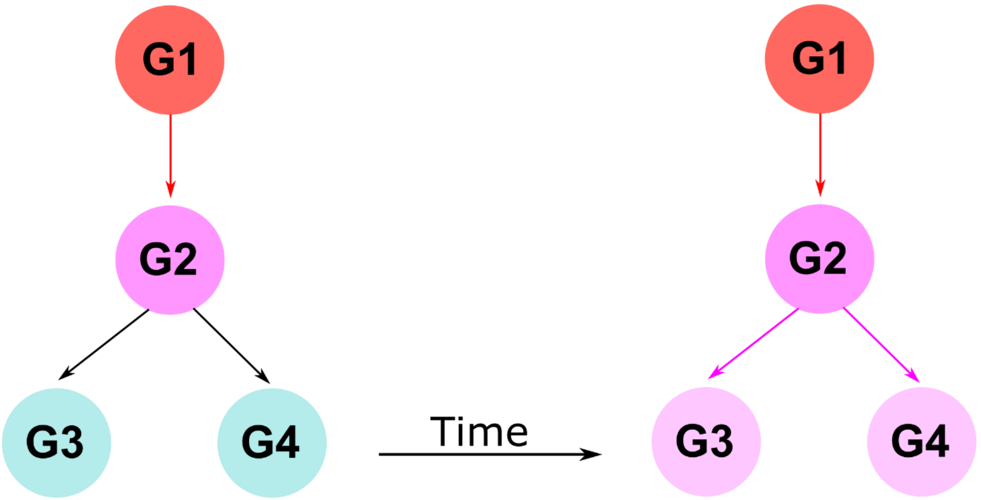


Figure 1.5 A network model of a transcriptomic perturbation - Illustration of how a genetic perturbation can be modeled using a graphical network. After a KO or inhibition by a drug, the network describes which genes will be transcriptionally affected by this perturbation and in which order. By quantifying these relationships the transcriptional impact can be predicted, along with the mechanistic steps that would produce it. This diagram can be generalized to interactions between proteins or other molecular effectors. Includes as components GRNA-Cas9 by Marius Walter, licensed under Creative Commons CC BY-SA 4.0 and Antibody_structureA by Michael Jeltsch, licensed under Creative Commons CC BY-NC-SA 4.0.

Term	**Definition**
Node	An entity in the network that is capable of interacting with other entities
Edge	The interactions or relationships between nodes
Degree	The number of edges a node is connected to
Directed network	A network in which edges only can go in one direction (e.g. A->B is different from B->A)
Undirected network	A network in which edges are not directed (e.g. A-B implies that A and B are equal interactors)
Centrality	A measure of node importance, which can be determined using several different metrics; generally in some way describes the number of paths in the network that pass through a node or how many other nodes it is connected to.
Adjacency matrix	A matrix representation of a graphical network in which the values of the entries represent the interactions or relationships between nodes. The size of the matrix is n by n, where n is the total number of nodes in the network.

Table 1.1 - Basic Graphical Network Terminology

Table defining general terminology used to describe graphical network methods.

Method	Algorithm type	Bulk or single-cell	Directed or Undirected	Citation	Code Source
Pearson Correlation	Correlation	Both	undirected	(Stuart et al., 2003)	Various
PIDC	Partial information decomposition	Both	undirected	(Chan et al., 2017)	https://github.com/Tchanders/NetworkInference.jl
ARACNE	Mutual information	Both	undirected	(Margolin et al., 2006)	https://github.com/califano-lab/GPU-ARACNE
GENIE3	Decision tree ensembles	Both	directed	(Huynh-Thu et al., 2010)	https://arboreto.readthedocs.io/en/latest/index.html
SCODE	Ordinary differential equations	Single-cell	directed	(Matsumoto et al., 2017)	https://github.com/hmatsu1226/SCODE
SINCERITIES	Granger causality	Single-cell	directed	(Papili Gao et al., 2018)	https://github.com/CABSEL/SINCERITIES
SINGE	Granger causality	Single-cell	directed	(Deshpande et al., 2019)	https://github.com/gitter-lab/SINGE
Scribe	Directed Information	Single-cell	directed	(Qiu et al., 2020)	https://github.com/cole-trapnell-lab/Scribe

scTenifoldKnk	Principal components regression and tensor decomposition	Single-cell	directed	(Osorio et al., 2021)	https://github.com/cailab-tamu/scTenifoldKnk
CellBox	Ordinary differential equations	Bulk proteomics	directed	(Yuan et al., 2021)	https://github.com/sanderlab/CellBox

Table 1.2 - Gene Network Inference Methods

Table containing methods described in this review for or involving gene/protein network inference. These methods use transcriptomics data as input unless otherwise indicated.

Experiment	**Data Type**	**Output**	**Application for network modeling**
RNA-seq/scRNA-seq	Transcriptomics	Sequences of expressed transcripts	Inferring regulatory relationships between gene expression levels
ATAC-seq/scATAC-seq	Chromatin conformation	Sequences of DNA that are in an open conformation	Identifying DNA sequences that are undergoing epigenetic regulation and which regions can express transcripts
Methyl-seq/scMethyl-seq	DNA methylation	Methylated regions of DNA	Identifying DNA sequences that are methylated and are thus unlikely to be able to express transcripts

ChIP-seq/scChIP-seq	Protein binding to DNA	Sequences of DNA with a particular protein/proteins bound	Determining where particular regulatory proteins are binding in the genome
Protein Mass Spectrometry	Proteomics	Abundance of molecules with specific mass/charge ratio	Estimate protein abundance and protein interaction networks
Protein microarrays	Proteomics	Abundance of a set of proteins	Estimate protein abundance and protein interaction networks for a particular set of proteins
CyTOF	Proteomics	Abundance and location of a set of proteins	Estimate protein abundance and protein interaction networks for a particular set of proteins, including a spatial element
CITE-seq	Transcriptomics and proteomics	Single-cell transcriptomics and abundance of cell surface proteins	Infer relationships between gene expression and to cell surface protein abundance
Metabolite Mass Spectrometry	Metabolomics	Abundance of molecules with specific mass/charge ratio	Estimate the relationships between metabolite levels, to data from other experiments
NMR Spectroscopy	Metabolomics	Abundances of organic and some inorganic molecules	Estimate the relationships between metabolite levels, to data from other experiments

Table 1.3 - High Throughput Technologies for Network Modeling

A list of high-throughput experiments, their outputs, and how these can be potentially applied for biological network modeling.

Method	Algorithm type	Data Type(s)	Citation	Code Source
NicheNet	PageRank	Bulk or single-cell transcriptomics	(Browaeys et al., 2020)	https://github.com/saeyslab/nichenetr
SoptSC	Similarity matrix	Single-cell transcriptomics	(Wang et al., 2019)	https://github.com/WangShuxiong/SoptSC
DOMINO	Decision trees and correlation	Bulk or single-cell transcriptomics	(Cherry et al., 2021)	https://github.com/chris-cherry/domino
SpaOTsc	Partial information decomposition and decision tree ensembles	Single-cell transcriptomics + Spatially resolved data	(Cang and Nie, 2020)	https://github.com/zcang/SpaOTsc
CCCExplorer	Pathway activation probability	Bulk or single-cell transcriptomics	(Choi et al., 2015)	https://github.com/methodistsmab/CCCExplorer
NATMI	Prior knowledge weighting	Bulk or single-cell transcriptomics	(Hou et al., 2020)	https://github.com/asrhou/NATMI

Table 1.4 - Intercellular Graphical Network Methods

Table containing methods for intercellular network modeling described in this review.

Chapter 2

Transcriptomic Forecasting with Neural ODEs

Introduction

Cells are dynamic and constantly changing. Predicting their future molecular states enables greater understanding of how biological systems will change naturally and in response to perturbation. A limitation of single cell RNA sequencing (scRNA-seq) technologies is that they destroy the cell to measure its molecular state. Therefore, scRNA-seq cannot track the specific molecular trajectory of an individual cell over time. Rather, scRNA-seq yields statistical samples from populations of cells. Performing additional time course experiments can increase the information available about cellular dynamics and cell state changes over time in a biological process. Time course designs can provide substantial information about dynamics of a biological system of interest, but are costly and limited to the time period over which they are measured. While many single-cell technologies do not dynamically profile the molecular state of an individual cell, new metabolic labeling technologies and live cell imaging methods are emerging that are starting to unlock the potential for longitudinal sampling of the molecular states of cells. As these technologies develop, new computational algorithms are needed to determine the distinct transcriptomic states each cell occupied in the past and estimate how cell states will evolve.

Predicting cellular dynamics requires models of both cellular phenotypes and their underlying molecular states. Trajectory inference methods have been widely applied to scRNA-seq data to estimate transitions between cell states (Saelens et al., 2019). Building on the foundation of pseudotime, trajectory inference methods infer the ordering of cells based upon the relative distance of their expression profiles, often incorporating information from low dimensional embeddings or distance between subgroups to define a trajectory of cellular dynamics (Trapnell et al., 2014), (Reid and Wernisch, 2016), (Saelens et al., 2019). These algorithms have been applied to scRNA-seq data collected at different time points (Trapnell et al., 2014), (Reid and Wernisch, 2016), (Schiebinger et al., 2019), along a developmental trajectory (Chen et al., 2019), and through disease states (Campbell and Yau, 2018). Another form of trajectory inference uses optimal transport methods to order cells along a time course by calculating the shortest path in expression space between cell states in a Waddington landscape (Schiebinger et al., 2019), optimal transport with neural ordinary differential equations (Tong et al., 2020), or optimal transport modeled using Jordan-Kinderlehrer-Otto flow learned by an input convex neural network (Bunne et al., 2022). Notably, trajectory inference methods are focused solely on ordering cells and require further extensions to model the gene expression values of these cells forward or backward in time or to account for tracing of individual cells.

RNA velocity, rather than focusing on ordering cells, investigates dynamic cellular processes by estimating the change in expression occurring in each cell based upon the ratio of spliced to unspliced transcripts (La Manno et al., 2018), (Bergen et al., 2020). These methods are most commonly applied to overlay predicted steady state of gene expression onto embeddings, and thus model changes in cellular phenotypes. Velocity methods have been extended to additionally estimate the cellular direction at the level of translation, by comparing spliced counts with protein data, called protein acceleration (Gorin et al., 2020). However, these RNA velocity and protein acceleration methods do not make predictions about future or past cell states beyond the immediate changes in expression. To predict expression state changes further into the future, vector fields have been applied to the concept of RNA velocity to allow for prediction of future states (Qiu et al., 2022), (Chen et al., 2022). One of these methods, called Dynamo, additionally suggests the use of metabolic labeling scRNAseq variants (Battich et al., 2020), (Qiu et al., 2020), (Hendriks et al., 2019), (Erhard et al., 2019), (Cao et al., 2020), in which cells are treated with a modified uridine for a set period of time before they are harvested for sequencing. This modified uridine is incorporated into the RNAs produced in that labeling period, which allows more recently produced transcripts to be distinguished from older ones. While RNA velocity vector fields methods predict the future cellular expression states of the cells in the data for multiple time steps into the future, they requires that all predicted future states fall within the UMAP or gene-dimensional expression space observed in the data set (Qiu et al., 2022), meaning unseen cell expression states cannot be identified.

To estimate the future transcriptomic states of single cells with dynamic measurements, we have developed a neural ordinary differential equation (neural ODE) (Chen et al., 2018) based method, RNAForecaster. RNAForecaster uses count data as input from two time points in the same cell. This sort of data is not available when using standard scRNAseq protocols, but can be provided using labeled and unlabeled counts from metabolic labeling transcriptomic profiling techniques such as scEU-seq (Battich et al., 2020). The counts from the earlier point in time are provided to the input layer of the neural network, which attempts to predict the expression of each gene at the later time point. This prediction is compared to the actual expression at the later time point to train the network. In metabolic labeling data, where the length of the labeling period is known, this allows for the network to forecast expression in real time. The key distinguishing feature of RNAForecaster from trajectory and RNA velocity methods is that it does not depend on a particular lower dimensional embedding of the data but takes input in the gene dimensional space. Therefore, RNAForecaster does not limit its predictions of future transcriptomic states to the expression space of the input data and attempts to generalize beyond the expression values observed in training. Specifically, training the method on expression values for each gene instead of relying on an embedding of the data provides the potential for this method to predict previously unseen transcriptomic states over a limited time period.

We demonstrate the predictive accuracy of RNAForecaster in simulated data, where we can establish a ground truth regarding future cellular expression states. We then apply RNAForecaster to scEU-seq from constitutively dividing cells and demonstrate that the model can predict the transcriptomic direction of cell cycle progression each hour for three days after the initial expression state is provided to RNAForecaster. Altogether, these analyses demonstrate the utility of neural ODEs for short term forecasting of future expression states from temporally resolved single-cell data.

2.1 RNAForecaster is a neural ODE based method for predicting future transcriptomic states

We designed RNAForecaster as a neural ODE (Chen et al., 2018) method that can leverage single cell transcriptomics profiling methods with temporal resolution to learn to predict future time points over limited time periods. To enable this analysis, the input to RNAForecaster is two single cell RNA count matrices. The model does not depend on the source of these matrices, only that they measure the same genes and cells in two adjacent time points (Figure 2.1A). RNAForecaster requires that these two matrices measure the same cell in each column, rather than similar cells harvested separately for sequencing. This is required because RNAForecaster attempts to estimate the future transcriptomic states of each individual cell. We denote the first gene expression matrix as time point t = 0 and the second as time point t = 1. These matrices are used to train a neural ODE.

The training process begins with each cell from the matrix of data from time t = 0 forming an input vector, where the log counts for each gene fill one node in the input layer (Figure 2.1B). The weights connecting the nodes of the hidden layer(s) and the output layer create an activation function. The output of the activation function represents a prediction of each gene’s expression at time point t = 1. These predictions are then compared with the actual expression level of each gene at time t = 1 based on that input matrix. The mean squared error (MSE) between these values is the loss function of the network. As opposed to standard neural network implementations, weights are updated differently in a neural ODE. Specifically, backpropagation is performed using an ordinary differential equations solver, allowing the network to have a continuous depth and constant memory cost. Thus, the network can yield performant predictions without using many hidden layers and maintains a constant memory requirement, making it computationally cheaper to train than most deep learning alternatives (Chen et al., 2018). After the network is trained, the predicted expression values from the output layer can be fed back into the input layer (Figure 2.1B), allowing the network to predict the cellular transcriptional state at future time steps. These predictions can be repeated recursively until an arbitrary time t = n, although the propagation of error with each step will cause the prediction error to generally increase over time.

The use of the ODE solver for backpropagation explicitly models dynamical systems, such as the evolution of gene expression values over time, making neural ODEs particularly well suited to predicting future transcriptional states. Additionally, as the neural network does not require a large number of layers to be performant (Chen et al., 2018) it is able to solve this prediction task in a computationally tractable manner. This is a critical feature because using thousands of variable genes as input creates a very large number of network parameters, which would produce a very computationally demanding network to train using other deep neural network architectures. Further, neural ODEs have been found to be particularly accurate at time series predictions relative to other neural network variations (Chen et al., 2018).

To illustrate the prediction task RNAForecaster performs, we provide an example of a sample cell with ten genes (Figure 2.1C). RNAForecaster is trained on the first two time points from this cell. By training RNAForecaster on many similar simulated cells, each with two time points from the same cell, it can learn the relationships between genes and generalize to make future predictions beyond the gene expression space it was trained on. This challenge of generalizing to a diverse array of transcriptomic states and determination of the temporal limits of predictability will be the focus of the applications we discuss.

2.2 RNAForecaster makes accurate predictions in future expression data outside its training set in simulated single cell transcriptomic data

We generate simulated data to benchmark the feasibility of estimating future transcript counts with RNAForecaster. Simulated temporally resolved single cell expression data was generated using BoolODE (Pratapa et al., 2020) as described in the methods. Briefly, this algorithm simulates gene expression from a system of ordinary differential equations from a known gene regulatory network and incorporates a model that allows for transcriptional busting, and thus contains stochastic elements. To recapitulate the way in which expression data provides a single time snapshot of gene expression, BoolODE simulates a cell's expression at hundreds of time points and then samples one for inclusion in the output counts matrix. Here, we leverage these additional future expression states as the ground truth for comparison with the predictions of RNAForecaster. To create each simulated data set, we generated a random ten gene network of regulatory relationships between genes. We generated over one hundred randomly generated networks, each of which was used to simulate a single-cell data set with 2000 cells and ten genes at 801 simulated time points.

We begin by training RNAForecaster on two simulated time points, each from the same cell. Standard scRNA-seq data sets are not able to produce this type of training data, but these simulated data allow for a validation of the general principle underlying RNAForecaster. A set of cells is randomly selected as the training set (80%) and the remaining cells form a validation set. We additionally trained a five hidden layer multilayer perceptron (MLP) for comparison on the same data. This MLP model is a feed-forward neural network that can provide a comparison for prediction accuracy using a simple network architecture (see methods for details). The MLP is used as a comparison to benchmark the performance of the neural ODE against the most standard neural network architecture. We first compare predictions of expression at $t = 1$ in the held out validation set using the expression at $t = 0$ as input. The RNAForecaster neural ODE significantly outperforms the MLP model on the validation data ($p < 1e-16$) (Figure 2.2A), though both methods accurately predict the first time point, with the average mean squared error (MSE) across simulations below 0.015 for both networks.

To determine the temporal range over which predictions can be made, we then tested the ability of the models to predict simulated expression for the next fifty time points for each cell. While the error in both methods increases over time, the neural ODE outperforms the MLP model significantly at all time points ($p < 1e-11$) (Figure 2.2B). Error propagates more quickly in the MLP and we additionally observe the presence of extreme outliers in the MLP predictions as early as $t = 10$, that fall within the same range of the worst predictions of the neural ODE at $t = 50$. The presence of these inaccurate outlier predictions suggests that the MLP is more likely to make poor predictions when faced with expression states outside the distribution encountered in its training data, a phenomenon termed “catastrophic forgetting” (French, 1999). To illustrate these MSE values in terms of the simulated expression levels, we take a closer look at a particular cell in Figure 2.2C. We selected this example as an approximately average performance by both the neural ODE and MLP. Here the MLP maintains a MSE under 0.02 until time t = 8, after which the predictions are inaccurate (median MSE of 0.42 over time points). The neural ODE, in contrast, is a better, though far from perfect fit to the simulated data (median MSE of 0.054). In some cases, the neural ODE predictions demonstrate a closer fit to the data, with median MSE values as low as 0.017 across fifty time points. However, we observe some poor fits with the neural ODE as well, producing median MSE values as high as 1.58.

In order to understand why some neural ODE solutions perform substantially better than other solutions, we examined the impact of different gradient descent initializations. Due to the recursive application of the neural ODE, the random seed used to initialize stochastic gradient descent influences predictions substantially. Even with the exact same training data, neural ODEs with different initializations can yield highly divergent predictions after fifty time points (Figure 2.2D). We observe that the predictions at time t = 1 are very similar from differently initialized networks, as we would expect, given that the networks are trained on the exact same data. However, because stochastic gradient descent can find many local minima, the weights are somewhat different. When making recursive predictions with the network, these differences in the weights compound, which often leads to very different predictions at later time points

(Figure 2.2D). We observe that some differently initialized neural ODEs perform better than others on a given example, but not uniformly better across all examples. This observation suggests that each differently initialized neural ODE may learn slightly different information about how to predict future expression states.

Ensemble based predictions leveraging information across multiple simulations from varied parameters have been shown to improve predictions of complex dynamical systems (Fertig et al., 2007), and are readily adaptable from weather prediction to forecasting biological systems (Kostelich et al., 2011). Therefore, we take an ensemble approach to improve RNAForecaster's ability to leverage the slightly different information learned by each network and handle variation in prediction accuracy. Using a different random seed to initialize gradient descent for each network, we train multiple neural ODEs and then evaluate the predictions of each, taking the median prediction as the final expression level estimate. This approach substantially outperforms a single network across simulated data sets (Figure 2.2E). As expected, at time t = 1 there is no significant difference in prediction accuracy, but there is a significant difference by t = 10 ($p < 1e-6$) and the magnitude of the difference in MSE loss increases with t. Most notably, the ensembles are much less vulnerable to catastrophic forgetting. If one network has extreme outliers in its prediction of gene expression profiles, it will usually be overruled by the others. We find that ten networks are sufficient to achieve most of the accuracy gains we can achieve through ensembling. Twenty-five networks yields no significant improvement over ten networks at any of the fifty time points, though the average MSE across simulations is slightly lower.

To determine the limits over which RNAForecaster can generalize robustly outside its training data, we simulated a single-cell data set using the bifurcating cell lineage gene regulatory network proposed by the authors of BoolODE (Pratapa et al., 2020). In the simulation, the cells progress in a single lineage for about 375 simulated time points, after which they bifurcate into two distinct lineages (Figure 2.3A). We trained RNAForecaster on time points 365 and 366 across 2000 cells to determine whether training immediately before the bifurcation provided sufficient information for RNAForecaster to predict cell lineage after the bifurcation. We find that RNAForecaster can differentiate the two lineages over 100 predicted time points (Figure 2.3B) despite being trained before the bifurcation. We compared the MSE on these 100 time points against the median MSE across the random simulations described above. The MSE was lower than the median level on 42 of 100 time points, indicating that RNAForecaster performed comparably in this case to the random network simulations. Immediately before the bifurcation, the expression differences in the cells appear sufficient to indicate which lineage a cell will become. This result demonstrates the ability of RNAForecaster to make predictions outside the space of its training data. However, if trained well before the bifurcation, we hypothesized that it should not be possible to reliably predict cell lineage fate on a per cell basis. To test this hypothesis, we trained RNAForecaster at time points 250 and 251, well in advance of the bifurcation, and predicted through the next 200 time points. We find that predictions break down at the bifurcation point, estimating transcriptional states that fall into a new cluster of cells in the UMAP that did not exist in the simulation (Figure 2.3C-D). The MSE predictions are likewise poor, worse than the random network median on all 200 time points. This result still leaves uncertainty as to whether the predictions of the second model are worse because of the stochastic effects and recursive error that are unavoidable when predicting from further away in time or due to the model being unable to learn key predictive relationships at the earlier time point. To distinguish these possibilities, we applied the second model (trained at the earlier time points) to predict one hundred time points forward from the cells the first model was trained upon. In this simulation, we find that the second model does predict the bifurcation (Figure 2.3E-F), though slightly less accurately than the first model. This result indicates that the major predictive relationships can be learned throughout the lineage of the cells, but that predictions only remain accurate over a limited time period due to stochastic effects and error propagation.

2.3 RNAForecaster can predict gene expression states beyond the space of the cell states used as input

One distinguishing ability of RNAForecaster from trajectory inference single-cell algorithms is that it can predict gene expression values outside of the space in a two-dimensional embedding occupied by the input data. To demonstrate this feature, we modified the BoolODE simulation to introduce a knock out (KO) of a single gene after one hundred simulated time points, after which the simulation continues with that gene's value set to zero. The simulated knock out of a single gene further introduces changes in the expression values of all other genes over time, leading to a divergent cluster of cells in UMAP space relative to the simulated cells from earlier time points (Figure 2.4A).

In order to determine whether RNAForecaster could predict into this space without being trained on it, we trained an ensemble of ten neural ODEs based on only the two time points before the KO simulation began. We then interrogated RNAForecaster's predictions of future gene expression profiles at time points after the

KO occurred. The predicted expression profiles cluster distinctly with the KO simulated data, despite being trained on none of these cells (Figure 2.4A-B), illustrating the ability of RNAForecaster to make accurate predictions outside the input space it was trained on. Across each of the simulated data sets used to evaluate the ensemble network performance, we simulated a data set from the same regulatory network with a gene KO. We find that RNAForecaster, using an ensemble of ten networks, is capable of producing comparably accurate predictions over fifty time points to those it made on the simulations that did not introduce a KO. We observe a small but statistically significant decrease in prediction accuracy at all time points (the mean difference is less than 0.05 MSE for time points 1 to 30 and less than 0.1 MSE for all time points) (Figure 2.4C). A loss in predictive accuracy from the distributional shift a KO causes is expected, but the small size of the difference demonstrates that RNAForecaster is able to produce accurate predictions for most simulated KOs through 50 time points.

2.4 RNAForecaster predicts the direction of cell cycle related transcriptomic changes over 72 hours from metabolic labeled scEU-seq data

In order to perform the recursive predictions that allow RNAForecaster to make predictions into the future, we need input data that can approximate the t = 0 and t = 1 matrices we used with simulated data. Critically, these count matrices must contain the two time points from the same cell. Metabolic labeling single-cell RNA-seq is the method we use to accomplish this. With metabolic labeling protocols (such as the scEU-seq protocol (Battich et al., 2020) we will use here) cells are labeled with 4sU modified uridine for a specified time period (Figure 2.5A). The cells are then harvested for single-cell RNA-seq. The 4sU labeled transcripts can be identified as those that were produced within the labeling period. The other sequenced transcripts were produced before the labeling began. This provides the temporal information we need to train RNAForecaster. The input matrix for the input layer consists of the unlabeled counts matrix plus estimated degraded transcripts. Metabolic labeling data allows for an estimate of the degradation rate in real time, as described by (Qiu et al., 2022). This count matrix is able to represent the total transcripts in the cell at time t = 0. The total counts at t = 1 are then provided by the unlabeled + labeled counts together (Figure 2.5A).

Now that we have a framework that allows us to train RNAForecaster on a biological data sets, we need a method for assessing its performance. We cannot get a series of expression levels in the same cell, preventing a direct assessment of per gene error over time. However, we can train RNAForecaster in a context where we know the general future expression path the cells should take, such as the cell cycle. To validate the method, we can test if RNAForecaster is able to predict the transcriptomic changes that are required for cell cycle progression. For this validation, we employed a scEU-seq data set from immortalized human retinal pigment epithelium (RPE) cells, published by (Battich et al., 2020). RNAForecaster was trained on the 405 cells in the data with a one hour labeling period, using a ten network ensemble. Once trained, RNAForecaster was used to predict the future expression levels in each cell for 72 hours. To score each cell's position in the cell cycle, we used tricycle, an R package that projects gene expression data onto an embedding of well characterized cycling cells to create a continuous score for cell cycle position (Zheng et al., 2021). These scores range from 0 to 2π, allowing cell cycle scores to be visualized on a circle (Figure 2.5B). The expression level predictions of RNAForecaster are then likewise scored by tricycle. The degree to which they are ordered from 0 to 2π over time is then assessed to estimate the degree to which RNAForecaster accurately predicts cell cycle progression.

The tricycle scores are highly ordered with respect to time through 72 hours of predictions in most cells (Figure 2.6A). For all 72 hours, the predictions are significantly more ordered than randomly generated scores ($p < 1e-16$). The ordering of these scores relative to random was further checked by generating 10 million random sets of tricycle scores, none of which achieved an ordering score equal to or greater than the median order score from the RNAForecaster predictions. Within the RNAForecaster predictions, the order of the scores decreased significantly each day ($p < 0.001$), indicating the fidelity to the cell cycle and general quality of predictions decreased the further predictions were into the future, as expected due to error propagation with recursive prediction.

A challenge we dealt with in this data was a tendency of the neural ODE to eventually start predicting extremely high transcript counts (Figure 2.6B). This likely results as an example of catastrophe when the network encounters input that is sufficiently dissimilar to what it was trained on. Then, in an example of positive feedback in the predictions, the predicted gene expression levels begin to go towards infinity. In order to control these extreme values in the prediction model, we set a realistic prior on the upper bound of expression values given the expression distribution observed in the training data. Enforcing these priors yields predictions that have similar median total counts/cell as in the scEU-seq data, even after 72 recursive predictions (Figure 2.6B).

These maximum counts priors also improve performance on cell cycle ordering scores by a small but statistically significant amount ($p < 0.05$) (Figure 2.6C).

Tricycle is used to score where RNAForecaster's predictions are in the cell cycle at each hour over the three day period, allowing us to determine the rate at which the predictions are moving through the cell cycle. While the order of scores strongly reflects the cell cycle, the rate of progression predicted by RNAForecaster is much slower than expected. Immortalized RPE cells usually replicate about once every 24 hours, and the RNAForecaster predictions proceed much more slowly (Figure 2.6D-F). Cells generally are predicted to progress steadily through the cell cycle, tracking the order of the cell cycle, but falling well short of the expected three completed cycles. This observation indicates that RNAForecaster learns the standard movement of cell cycle related genes, but is unable to recognize rarer regulatory events that lead to large changes in expression over shorter time periods. A relatively small number of cells were available for training RNAForecaster (405) which may contribute to this issue. Particularly for less common events, a larger data set could allow RNAForecaster to more adequately learn to model these gene expression dynamics. Alternatively, predicting shifts that are less frequently observed in the training data may be a weakness of the RNAForecaster's neural network architecture. Despite this limitation, accurately predicting the order of changes in cell cycle related genes across the different cell cycle stages using only a relatively small training set demonstrates the ability of RNAForecaster to estimate future expression states in single cells.

2.5 Discussion

RNAForecaster is a tool for generalizing temporal relationships in single cell transcriptomic data. Through a neural ODE, we attempt to learn activation functions that predict the expression level of a gene in terms of the previous expression levels of other genes. We demonstrate that it is possible to forecast future expression states in single-cell data that have a temporal dimension. The accuracy of the predicted gene expression states depends on the time period over which RNAForecaster is applied, with a reasonable degree of accuracy over short to intermediate time periods. RNAForecaster can thus provide valuable insight into the dynamics of transcription and transcriptional regulation over time. Through simulated data, we demonstrate that KOs and bifurcations can be predicted if the method can predict starting at time points shortly before the event occurs. In order to better capture the relationships between genes that allow prediction of their future expression states, it would be ideal to train the RNAForecaster network on more time points within the same single cell. While this is currently unavailable across at scale, recent techniques allow imaging of a small number of genes in cells over time (Cawte et al., 2020), (Wang et al., 2022). As these methods improve, RNAForecaster could be trained with longer time series, likely improving its accuracy and the span of time over which it can make accurate predictions. Altogether, future research should evaluate the limits of predictability using this model over diverse timescales and biological conditions.

The reliance on ODEs in our framework results in estimates of smooth temporal trajectories of gene expression. However, it is important to consider the predictions RNAForecaster produces in the context of the biology of transcription. RNA transcription occurs in bursts, and thus appears stochastic (Tunnacliffe and Chubb, 2020). Therefore, RNAForecaster's predictions should not be interpreted as an estimate of the exact counts in a cell, since this is not precisely predictable. Rather, RNAForecaster should be thought of as estimating the expected value of the distribution of a gene's counts in a cell.

In simulated data, we observed that significant gains in prediction accuracy could be attained by using an ensemble of networks to forecast future expression states instead of a single network. Given the recursive application of the network to make future predictions, the impact of small differences in the network weights can lead to large differences in predictions when two different initializations of stochastic gradient descent are used. The slightly different local minima found by these different gradient descent initializations often had strengths and weaknesses in their ability to generalize to expression levels outside their training data. We attempted to combine these strengths through a simple ensemble approach where we use the median predictions of the networks. However, even with the ensembling approach, prediction accuracy decreased substantially over time in the simulated data as error propagated and predictions trended further outside the domain the neural network was trained in. At the same time, catastrophically poor predictions occurred at a much lower rate, which does increase the time scale in which RNAForecaster is applicable. The downside of this approach is the increased computational resources required. However, training ten networks using a GPU is faster on single cell transcriptomics sized data sets than training one on a CPU, which makes GPU training highly preferred even without ensembling.

Using metabolic labeling scRNA-seq protocols, RNAForecaster can make recursive predictions about future transcriptomic states. These protocols are currently largely confined to *in vitro* studies, which limits the application of RNAForecaster in many *in vivo* contexts. *In vitro*, RNAForecaster was capable of accurately predicting the general direction of cell cycle related expression changes over 72 hours. However, the model failed to recapitulate the speed of cell cycle progress. This may reflect an inability of the model to predict cell cycle checkpoints and the ensuing transitions between stages, or some other rare regulatory event in the cell cycle. A larger training set might allow RNAForecaster to better capture these less common events, given that the training data contained only 405 cells. However, it may be the case that the current structure of RNAForecaster lacks the capability to handle these sorts of exceptions to the transcriptomic changes it sees in most cells. Future work integrating attention based architecture (Vaswani et al., 2017) into RNAForecaster could potentially allow the model to differentiate, for example, the expression changes within a cell cycle stage from the changes at the end of the stage after checkpoints are passed.

We observed that the neural ODE tended to make extremely high predictions of expression values after the previous predictions had departed sufficiently from its training data. To handle this we enforced maximum expression level predictions for each gene based on the observed data, which constrained the model to obtain realistic expression levels on both an individual gene and overall cell level. These maximum expression levels can be justified from a Bayesian perspective, where we assign very low probability to seeing expression levels of a gene that are higher than a certain point. There is a difficult balance, however, between preventing unrealistic expression levels and removing valuable signal from the predictions. Preventing these kinds of extreme, unrealistic values is a major challenge termed as “catastrophe” in the field of machine learning, (French, 1999), and it may be further exacerbated for RNAForecaster due to the high dimensionality of the data and the recursive predictions required. The use of network ensembles helps alleviate this tendency to some extent. However, future work may be needed to teach the neural network about the prior probability for gene expression levels outside the range normally observed in cells, rather than having to enforce this prior after the fact.

Whereas many predictive methods, including notably Dynamo (Qiu et al., 2022), can estimate cellular states from RNA velocity vector fields estimated through splicing or metabolic labeling, RNAForecaster relies on temporal single-cell transcriptomics data tracing an individual cell currently enabled uniquely with metabolic labeling data (Qiu et al., 2022). Whereas these current methods aim to predict cellular states captured in the training data, RNAForecaster instead attempts to generalize its predictions to the full space of possible expression states. This formulation uniquely allows RNAForecaster to estimate the impact of perturbations that shift the expression state into part of the space not observed in the input, as well as future developmental or evolutionary states not captured in the input data. Another important difference between Dynamo and RNAForecaster is that Dynamo requires its input data to be smoothed using k-nearest neighbors averaging in order to compute RNA velocity (Qiu et al., 2022). This procedure essentially averages the cells that are close together in expression space, which may introduce some distortions or remove important variation (Gorin et al., 2022).

Several other methods have been proposed with the goal of predicting single-cell responses to perturbations. PerturbNet trains a generative neural network using perturbation single-cell data sets such as Perturb-seq (Dixit et al., 2016) to predict responses to genetic knockouts or knockdowns (Yu and Welch, 2022). This approach differs from RNAForecaster by focusing on learning from specific perturbations rather than learning general temporal relationships between genes. The scGen method also attempts to predict single-cell perturbation responses, using a combination of a variational autoencoder and a deep generative network to project what it has learned in its training data into unseen cell states (Lotfollahi et al., 2019). While RNAForecaster attempts to learn the predictive relationships between genes over time, scGen attempts to learn how cell states shift under perturbation using similar perturbations in similar cells. These approaches can yield valuable insight into future expression states after a perturbation and the correct choice of method for a particular use case will often depend on the particulars of a problem and the type of data that is most readily available.

The gold standard that many computational methods aspire to is inference of mechanistic interactions from high-throughput biological data sets. One advantage of neural ODEs is that they can yield greater interpretability than other neural network formulations (Chen et al., 2018). This feature may allow the relationships between genes that RNAForecaster learns to be interrogated, which could potentially yield mechanistic insight. The accuracy of gene network inference methods suggests that the high degree of correlation between genes makes prediction much easier and more robust than causal inference (Pratapa et al., 2020). However, extending RNAForecaster and other methods from the prediction of future gene expression states to mechanistic, molecular networks remains an important area of future research.

RNAForecaster demonstrates that future states in transcriptomic data with a temporal dimension can be estimable, even outside the expression space of the input data. As single-cell and machine learning technology improves, it may be possible to extend this capability to accurately predict counterfactuals regarding future cell states based on a variety of cellular factors. This could be used to predict the response of diseased cells to perturbations, potentially informing treatment options on a general or personalized level. Extending these techniques to personalized predictions of the effects of perturbations and therapies may enable predictive biology and medicine approaches (Fertig et al. 2021), (Stein-O’Brien, Ainsile, and Fertig 2021) and require new methods to quantify the limits of predictability of therapeutic outcomes across disease systems.

2.6 Methods and Software

Required Input Data and Preprocessing

RNAForecaster primarily requires two normalized single-cell RNA counts matrices as input. These counts should be from adjacent time points, such that the labels time t=0 and time t=1 can be reasonably applied and the cells in each matrix are identical. The main preprocessing steps needed are sparsity filtering and log normalization. Including genes that have high proportions of zeroes (default greater than 98%) can cause problems with gradient descent, and thus these genes must be removed. The only normalization applied is a log1p transform. In addition, filtering to highly variable genes is strongly recommended, and was performed for all biological data sets used.

When using metabolic labeling data, there is an additional preprocessing step to account for transcripts that degraded during the labeling period. The degradation rate is calculated using the slope between the labeled and total counts as described by (Qiu et al., 2022). A linear regression is fit between the two count matrices and the degradation rate is calculated as *-log(1-slope)* which estimates the number of transcripts degraded per labeling period. We then estimate the total counts at the beginning of the labeling period by adding each gene's degradation rate to the unlabeled count matrix. The resulting matrix becomes the time t=0 input matrix.

Neural ODE Training

By default, the input data matrices are divided into training and validation sets (default 80-20 split). The default number of nodes in the hidden layer is twice the number of nodes in the input layer to allow for interactions between genes. A neural ODE (Chen et al., 2018) is then trained using Flux.jl and DiffEqFlux.jl, using the Tsit5 ODE solver and a default error tolerance of 1e-3. Training occurs for a default of 10 epochs using a default learning rate of 0.005. The loss function is calculated as the mean squared error between the output nodes and each gene expression level in the time t=1 matrix.

We provide the option to check network stability on recursive predictions at this stage. Recursive predictions are made for a user-defined number of steps, checking on each step whether any expression levels are higher than any plausible level. If this is observed, the training process is restarted. These stability checks can prevent the frequency of catastrophe in network predictions outside the training distribution. With ensembling, these stability checks are largely unnecessary, but they provide an alternative for data sets where the network is less prone to catastrophe.

Creating an ensemble of networks simply repeats this process using a different random seed to initialize stochastic gradient descent for each network. The default number of networks trained is ten.

Recursive Predictions of Future Expression Levels

The process of estimating future expression requires the input of a trained neural ODE (or ensemble of them) and a set of initial expression states to predict from. These expression states are fed into the input nodes of the neural ODE and then the output is recorded and fed back into the input nodes, allowing for recursive predictions forward in time. Some prior knowledge and assumptions are enforced on the predictions by default. All predictions must be non-negative, as this is a constant characteristic of gene expression data. Additionally, expression level predictions that are higher than an allowed maximum are set to the maximum value (by default two times the maximum observed in the training data, in log space).

When estimating expression levels using an ensemble of networks, the above process is performed for each network and the median prediction is used.

Variational Autoencoder Input Option

We tested using a lower dimensional representation of the single cell transcriptomics data as input with a VAE. The VAE was trained on the first two time points of the data, the same portion that would be used to train the neural ODE. The encoder was then applied to the t=0 data and the lower dimensional matrix was used as input to the neural ODE, which was trained on the lower dimensional encoding of the t=1 matrix. The output lower dimensional representation was then decoded with the VAE back into individual gene expression levels.

Simulating single cell expression data with BoolODE

BoolODE was designed to simulate single-cell expression data sets on the basis of a network of gene-gene interactions (Pratapa et al., 2020). We generated 117 simulations because we wanted at least one hundred and we were concerned that some random networks might generate strange regulatory behaviors. To generate these 117 different simulations, we first needed 117 different gene-gene networks. These were created as random ten gene networks, where genes could have positive, negative, or no direct relationship with other genes. Each network was input to BoolODE, which simulates 801 time points of expression for 2000 cells in each simulation. Minor changes were made to BoolODE code (see https://github.com/FertigLab/RNAForecasterPaperCode), to generate output for the task of predicting future expression states.

The bifurcation simulated example was produced using the bifurcation gene network and initial conditions set created for and provided by BooODE (Pratapa et al., 2020). We simulated 2000 cells for 801 time points to yield the data used in this work. UMAP visualizations of the data were produced using Seurat version 4.0.1.

To simulate data sets with gene KOs, the simulations were allowed to run for 101 time points, at which point a gene's expression value was set to zero and was again set to zero on each future iteration, mimicking a KO gene.

Applying RNAForecaster to simulated data

We used the t=101 simulated count matrix as our t=0 for input to RNAForecaster, in order to give the simulation time to initialize and stabilize. RNAForecaster was thus trained on t=101 and t=102 for each simulation. Predictions were made for up to 200 time points later, up to t=300. The neural ODE was trained for 10 epochs, with 100 hidden layer nodes, and a learning rate of 0.005.

Ensembles were created using groups of 10 and 25 networks. Networks used 100 hidden layer nodes and stability checks were performed. Simulations where stability checks were not passed on fifteen iterations were excluded, leaving 111 simulations in the final set.

Comparison predictions using a feed-forward MLP model

For comparison with RNAForecaster's predictions in simulated data, we employed a simple five hidden layer, fully connected feed-forward neural network architecture. This MLP model was trained for 10 epochs and a learning rate of 0.005. The network node structure from the input to output layer is as follows: Dense(10,32), Dense(32,64), Dense(64, 100), Dense(100,100), Dense(100,64), Dense(64, 32), Dense(32,10).

Data download and processing

The scEU-seq data set from (Battich et al., 2020) was downloaded using the Dynamo python package to acquire the rpeLabeling.h5ad AnnData file produced by (Qiu et al., 2022).

Genes with more than 98% zero counts in either labeled or unlabeled count matrices were filtered from both matrices. The matrices were additionally filtered to genes by variance to genes in the top quartile. The degradation rate was then calculated using the slope between the labeled and total counts. A linear regression was fit between the two count matrices and the degradation rate was calculated as *-log(1-slope)* which estimates the number of transcripts degraded per labeling period. We then estimate the total counts at the beginning of the labeling period by adding each gene's degradation rate to the unlabeled count matrix. We then subset to those cells treated with 4sU uridine for 60 minutes, so that the labeling time is equal for all input cells. The resulting matrix becomes the time t=0 input matrix, to be compared to the total counts as the t=1 matrix in RNAForecaster.

Training RNAForecaster

RNAForecaster was trained as an ensemble of ten networks, training for 20 epochs on all 405 cells with a 60 minute labeling period. These networks were trained on a Nvidia Titan V GPU using a mini-batch size of 100 and a learning rate of 0.001. All other parameters use the default values.

Predicting future expression states and estimating their position in the cell cycle

RNAForecaster predicts the future expression levels in each cell from the total counts matrix into the future each hour for 72 hours. The maximum prediction for each gene is set to 1.2 log fold increase over the maximum value observed in the training data. Predictions were performed on a Nvidia Titan V GPU.

The resulting predictions were scored for their position in the cell cycle using the tricycle R package (Zheng et al., 2021). Tricycle creates a quantitative embedding of the cell cycle from scRNAseq data of cells with known cell cycle positions. This embedding ranges from 0 to 2π to represent the circular nature of the cell cycle. In this range, 0.5π to π is the approximate bounds of S phase, π to 1.75π G2M phase, and 1.75π to 0.25π G1 or G0 phase. This embedding is then projected into a target single cell RNA data set to approximate the cell cycle position of each cell. We applied tricycle to each initial cell state from the scEU-seq data and each expression state predicted by RNAForecaster.

To determine the degree to which the tricycle scores in the RNAForecaster predictions matched the order of the cell cycle, we developed an ordering metric. The predictions made in a cell receive a point in this metric if the score increases (or goes back around from 2π to 0), but does not increase by more than 0.75 in a one hour period. This metric additionally differentiates small decreases in cell cycle score from large ones by giving 0.2 points to decreases of less than 0.25. This is used to differentiate slight variation from substantial incorrect shifts in gene expression prediction. We additionally plotted the scores, both for individual cells as a line plot, and all together on a circle plot.

Figures

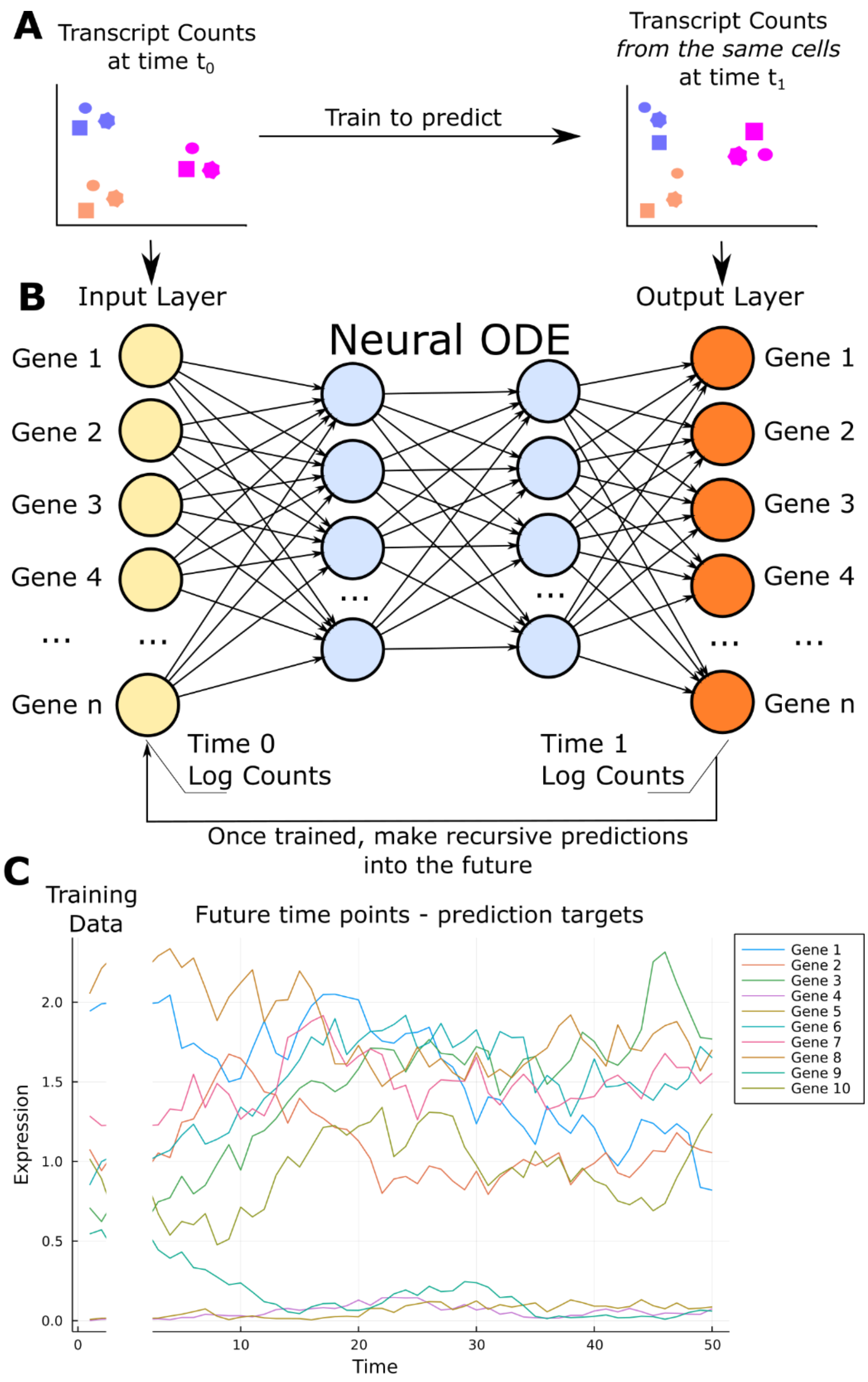


Figure 2.1 - Diagram of RNAForecaster

A Two count matrices are input to RNAForecaster, each containing the same genes and cells. The counts matrices are from adjacent time points from the same cells, labeled here as t=0 and t=1. **B** The t=0 counts for each cell are input to the input layer of a neural network. The output layer of the neural network has the same number of nodes as the input layer and is compared to the results from the same cell at t=1. The mean squared error between the two forms the loss function which is trained on using an ODE solver to produce a neural ODE. Once the network is trained the output can be fed into the input layer, allowing for prediction of the expression levels at the next time point, which can be repeated recursively to predict for t time steps. **C** A simulation of the expression levels in a cell, showing ten genes over fifty time points. RNAForecaster is trained on the first two time points, using multiple cells in order to learn some generalization of the temporal dynamics between genes. RNAForecaster then attempts to estimate expression of each gene at the later time points.

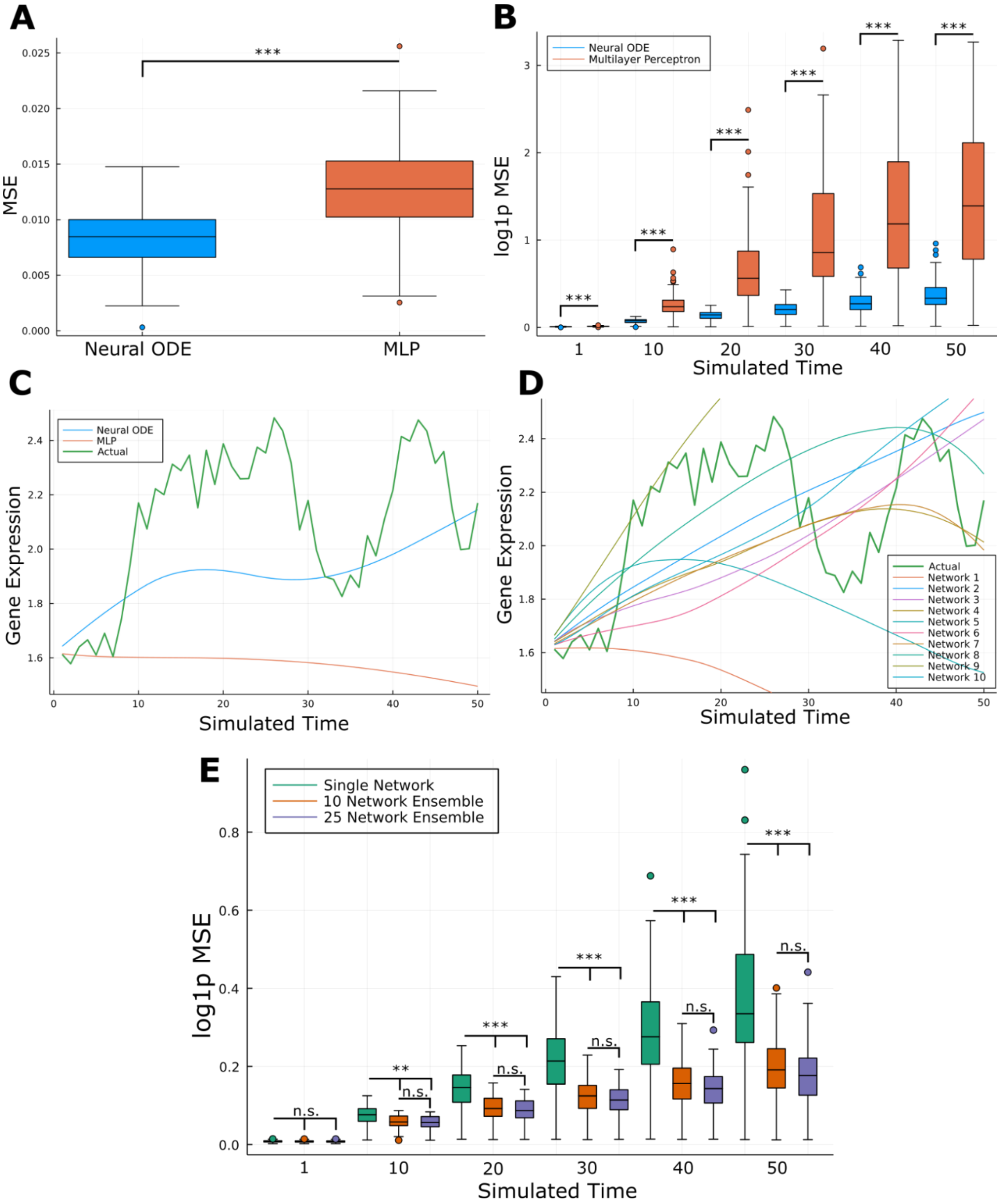


Figure 2.2 - RNAForecaster prediction accuracy in simulated single cell expression data

A Comparison of MSE loss on the 20% held out validation set of predictions from t=0 to t=1 between a neural ODE and a 5 hidden layer MLP, over all simulations. **B** Comparison of log MSE loss on the next 50 simulated time points between a neural ODE and a 5 hidden layer MLP. **C** A median example of expression prediction of a single gene in a single cell. The predictions of the neuralODE and MLP are shown. **D** The predictions of ten different neural ODEs, each trained using a different initialization of stochastic gradient descent, for the same gene and cell as C. **E** Log MSE loss comparison between a single network neural ODE vs the median predictions from a 10 or 25 network ensemble of neural ODEs. ** p < 1e-6 *** p<1e-10

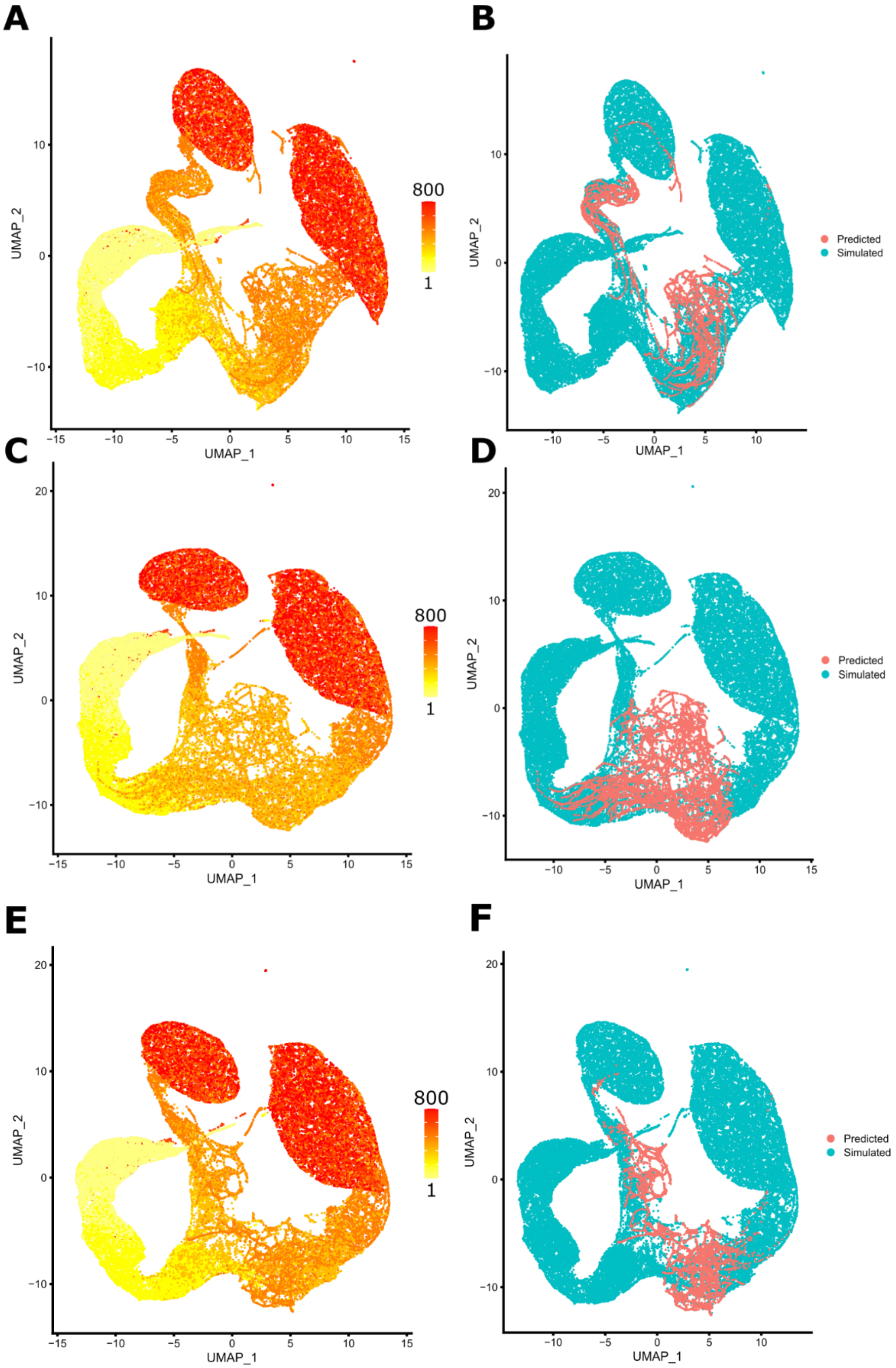
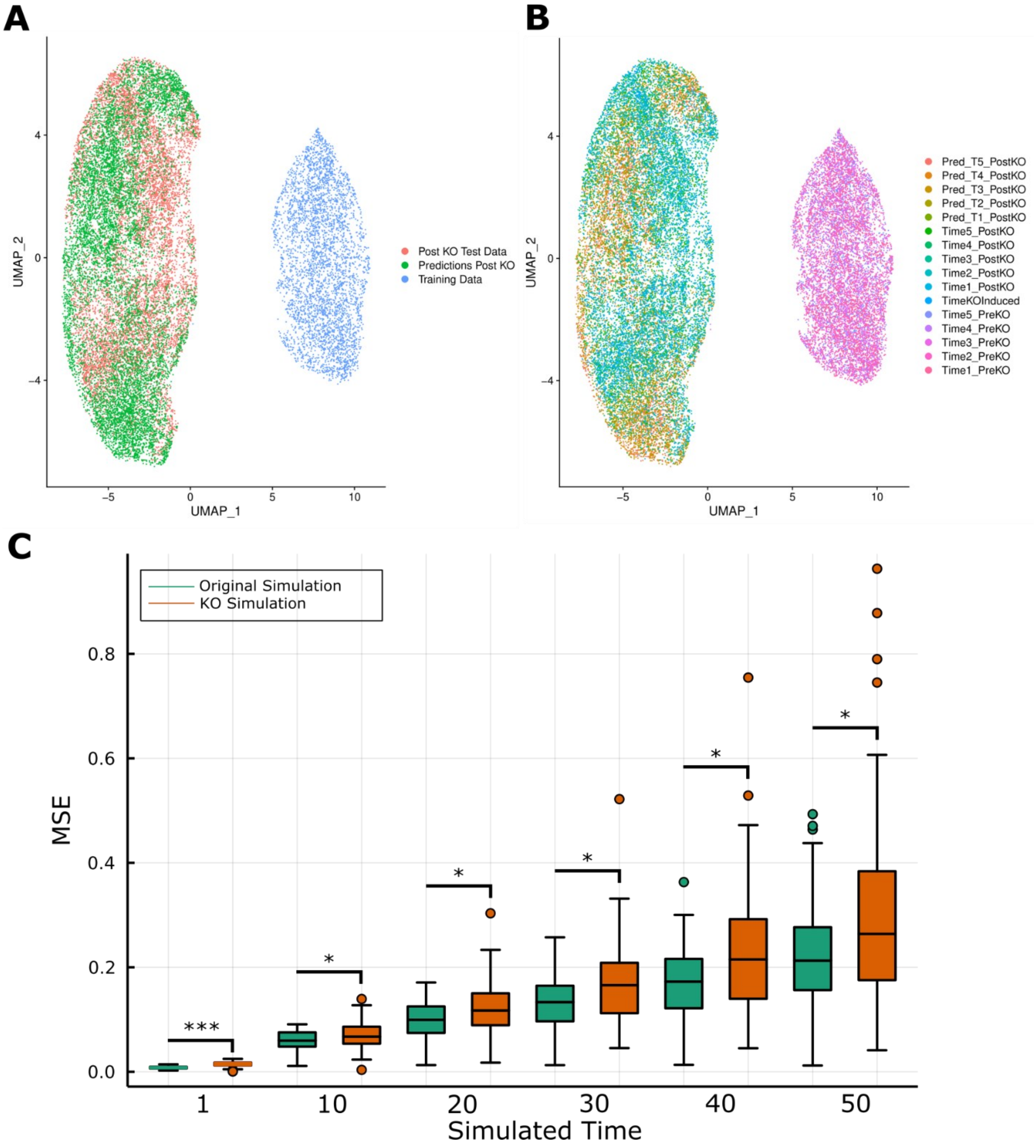

A
B
C
D
E
F
UMAP_1
UMAP_2
800
1
Predicted
Simulated
−15
−10
−5
0
5
10
15
20

Figure 2.3 – RNAForecaster predicts bifurcation of cells when trained on cells immediately prior to bifurcation

A UMAP of bifurcating cell simulation across 50 cells from simulated time point 1 to time point 800, plus RNAForecaster's predictions from time point 366 (just before the bifurcation) through next 100 time points. Colored by time point. **B** UMAP of same cells as in A, but colored by whether cells were from the ground truth simulation or RNAForecaster's predictions. **C** UMAP of bifurcating cell simulation across 50 cells from simulated time point 1 to time point 800, plus RNAForecaster's predictions from time point 251 through next 200 time points. Colored by time point. **D** UMAP of same cells as in C, but colored by whether cells were from the ground truth simulation or RNAForecaster's predictions. **E** UMAP of bifurcationg cell simulation across 50 cells from simulated time point 1 to time point 800, plus RNAForecaster's predictions using the model from C and D, predicted from time point 366 (just before the bifurcation) through the next 100 time points. Colored by time point. **F** UMAP of same cells as in E, but colored by whether cells were from the ground truth simulation or RNAForecaster's predictions.



<u>Figure 2.4 - RNAForecaster can predict the impact of a gene KO that moves cell expression outside the input space</u>

A A UMAP embedding of the training data from one simulation provided to RNAForecaster alongside the simulated data after a gene KO and RNAForecaster's estimations of expression states after KO. **B** UMAP from same simulation as A, labeled by time point and whether a cell was from the pre-KO simulations, post-KO simulation, or post-KO RNAForecaster prediction. **C** Boxplot comparing the MSE loss from the ten network ensembles shown previously in Figure 2.2 and the MSE loss from ten network ensembles onto simulated KO data, where the same gene networks were used to generate the simulations in both cases. * p < 0.01 ** p < 1e-6 *** p<1e-10

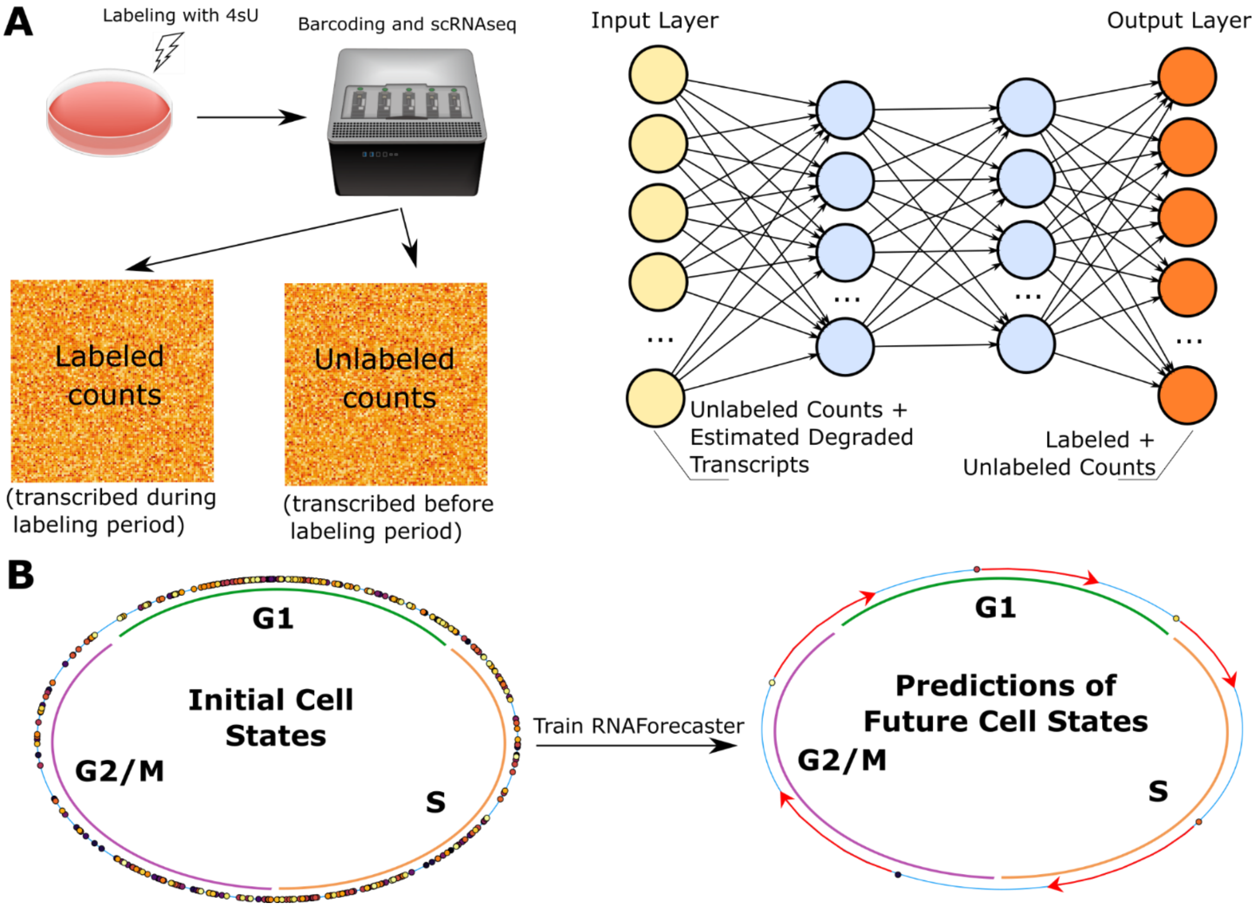

A
Labeling with 4sU
Barcoding and scRNAseq
Labeled counts
(transcribed during labeling period)
Unlabeled counts
(transcribed before labeling period)
Input Layer
Output Layer
...
Unlabeled Counts + Estimated Degraded Transcripts
Labeled + Unlabeled Counts
B
G1
Initial Cell States
G2/M
S
Train RNAForecaster
G1
Predictions of Future Cell States
G2/M
S

Figure 2.5 - Application of RNAForecaster to metabolic labeling single cell expression data

A Left is a diagram of the basic concept behind metabolic labeling protocols such as scEU-seq. On the right is a diagram illustrating how the output from metabolic labeling protocols is input to the RNAForecaster neural network. **B** Diagram showing the tricycle cell cycle scores of each one hour labeled cell from the Battich, et al (2020) scEU-seq retinal epithelium cell cycle data set. After these cells are used to train RNAForecaster, the future expression states of each cell can be predicted. These expression states can likewise be scored for cell cycle prediction and we can validate the predictions on whether they generally follow the expected trajectory of the cell cycle.

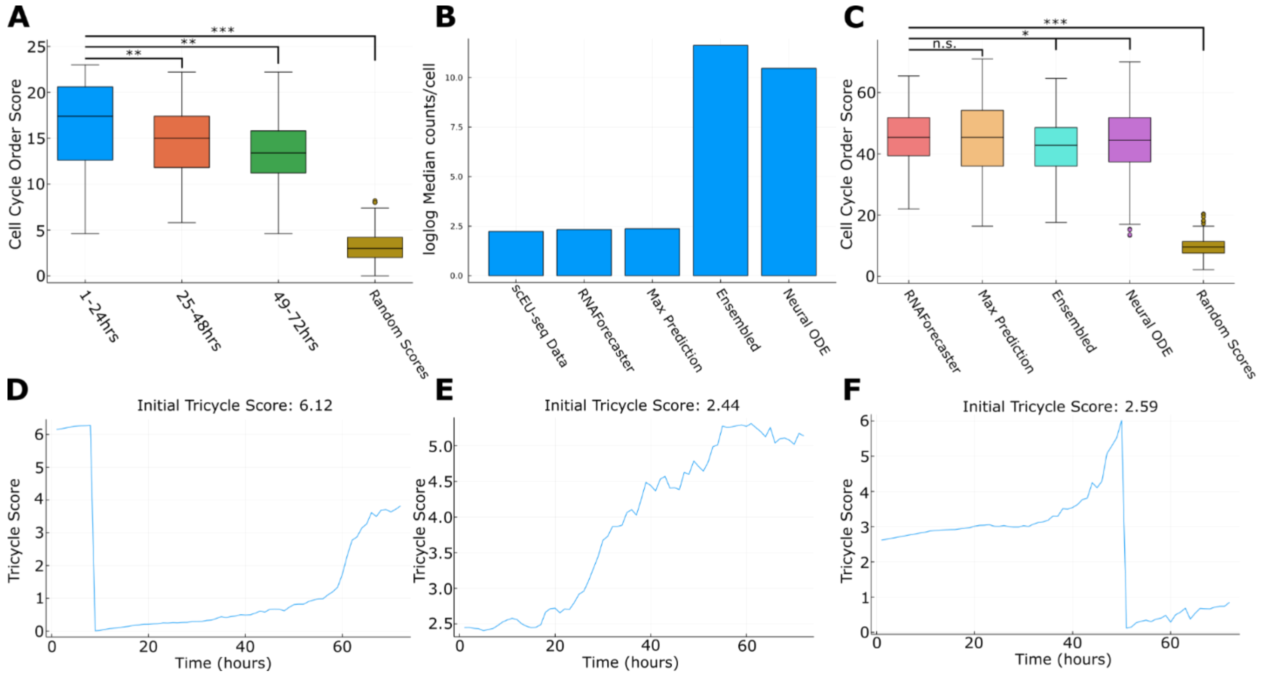


Figure 2.6 - Performance of RNAForecaster at forecasting the cell cycle

A Boxplot of a metric describing the order of the tricycle scores made using RNAForecaster's predictions. A higher score indicates the scores were more aligned with the order of the cell cycle. Compared to the metric when applied to randomly generated tricycle scores. **B** Barplot of the log-log median total counts per cell in the scEU-seq data set vs the output of different neural ODE implementations at the 72 hour prediction. **C** Boxplot of the tricycle score order metric for the neural ODE implementations shown in B. **D-F** Examples of tricycle scores on the RNAForecaster predictions in three cells. * $p < 0.05$ ** $p < 0.001$ *** $p < 1e-16$

Chapter 3

Mechanistic simulation of molecular cell states over time

Introduction

To understand the interplay between temporal changes from molecular interactions and the limitations of experimental data, mathematical simulations of molecular cell states can be useful to provide a known ground truth (Hill et al. 2016). In addition to providing benchmarks for the performance of network inference methods, using simulated genomics data generated from mathematical models can elucidate the kinds of temporal and regulatory dynamics that are likely to exist biologically. Moreover, simulations enable a broader range of evaluation of conditions and the requirements of datasets for inference as they can generate data under conditions that are infeasible or prohibitively expensive to generate experimentally. For example, simulations can readily provide temporally resolved multi-omics data from the same cells and low noise data. Thus, simulations can allow us to understand the impact of not capturing these dimensions in profiling methods on the resulting data and may illustrate cellular dynamics that are difficult to resolve from real cells.

A variety of simulation methods that attempt to generate data approximating cellular genomics have been proposed (Pratapa et al. 2020), (Das and Mitra 2021), (Herbach et al. 2017), (Hill et al. 2016), (Gorin et al. 2022), (Thornburg et al. 2022). However, most methods do not use a direct mechanistic model with biologically interpretable parameters to simulate molecular states over time, instead parameterizing the simulation with a mathematical model such as a set of ordinary or stochastic differential equations. Additionally, most simulations do not cover all of DNA accessibility, RNA expression, and protein expression. To our knowledge, both of these features have not been implemented in the same method in previous work. Therefore, we have generated software for Multi-Omic Mechanistic Simulations (MOMS) that simulates mechanistic molecular interactions using the general model displayed in Figure 3.1, is fully tunable regarding molecular parameters, and simulates based on the interplay between DNA accessibility, RNA expression, and protein expression states.

3.1 A mechanistic simulation of cellular states

The MOMS simulation outputs DNA accessibility, spliced and unspliced RNA counts, and protein counts for each gene at each time point (intended to approximate a one minute time difference from the previous data output) according to the mechanistic model illustrated in Figure 3.1. This model uses genes as its basic units, and the cells are assumed to have diploid genomes. Any set of genes can be used to simulate cell molecular states, given a list of genes where each gene has the required parameters specified: Baseline Transcription Probability, Transcription Factors, Transcriptional Repressors, Epigenetic Upregulators, Epigenetic Downregulators, Splicing Rate, RNA Half Life, Protein Half Life, Translation Initiation Probability, Translational Inhibitors, and

Protein Degradation Factors (Figure 3.1). This set of genes can be generated randomly or based on some known biological network input by the user.

Each copy of a gene can be either accessible or not accessible. For the simulation, this means the gene is either transcribed at a normal rate based on transcription factor binding (if accessible) or is much less likely to be transcribed because transcription factors cannot usually bind to the DNA (if not accessible). This is specified in the simulation in a binary manner, where each gene copy is accessible or not and there is a small (tunable) probability of transcription even when not accessible. If a gene is accessible, there is a baseline probability of the standard transcription factors binding per unit time of the simulation, which is intended to represent approximately one minute of real time. Each gene can be assigned a set of specific transcriptional activators and repressors that promote or inhibit RNA transcription, as well as epigenetic activators and repressors that can change the accessibility state. The impact of these regulatory genes is based on the concentration of the corresponding proteins. Thus, the volume of the nucleus and cytosol are specified in the simulation and the binding probability is calculated according to the equation 1 - e^-[protein]. The binding affinity is assumed to be equal for all proteins. Whether binding occurs at each unit time is evaluated with a random number generator using the specified binding probability. The splicing rate is specified for each gene and the spliced and unspliced RNA transcript counts are tracked separately. RNA degradation is controlled by the RNA half life parameter, which is specified for each gene.

Genes in the simulation are assumed to produce mRNAs. RNA is translated to protein based on a specified probability for how likely a spliced transcript is to bind a ribosome per unit time, as well as the concentration of translational inhibitors.

Translational inhibitors are protein products of specified genes. Protein degradation is determined by the protein half-life parameter specified for each gene, as well as the concentration of any protein specific degradation factors (i.e. other proteins that bind it and cause it to be degraded).

This simulation structure is probabilistic and thus the results are to some extent stochastic. As in our previous work, we select a probabilistic model as the basis of cell behavior because the precise moment molecules collide and bind is not usually predictable in advance without measuring cell state to a degree that is far beyond current technologies (Fertig et al. 2011).

Often we may want to randomly generate a gene set and its parameters rather than specifying each parameter manually. Thus, we provide functionality to generate the parameters for random gene sets, with parameter values that fall within normal biological ranges based on reference to experimental data. All parameter ranges are user tunable. For the simulations presented below, the baseline transcriptional probability was set between 0.01 and 0.001, the maximum number of transcriptional regulators was set to 3 proteins, and the maximum number of epigenetic regulators was set to 4 proteins. The splicing rate was randomly set for each gene between 5 and 10 time points (approximately 5 to 10 minutes). The RNA half life range was between 60 to 900 time points and the protein half-life range was between 720 and 3600 time points, intended to reflect the fact that most proteins degrade much slower than RNA (Schwanhäusser et al. 2011), (Cambridge et al. 2011). Each gene could have at most one translational inhibitor and at most two protein degradation factors. The probability of translation initiation for an unbound mRNA at each time point was set between 0.1 and 0.75.

The Julia software to run MOMS simulations is made freely available on GitHub at https://github.com/FertigLab/MOMSCellSimulations.

3.2 Impact of parameters on molecular cell states

The outcome of a simulation is determined by the parameters provided for each gene, the initial conditions of the simulated cell, plus a random number generator that determines the outcome of stochastic elements. A key to the variation in the simulations is gaining some insight into how much impact changing the random number generator seed and the initial conditions changes the observed cell states over time.

To do this, we randomly generated the parameters for ten genes and simulated a cell for 1000 time points (Figure 3.2A). The simulation was then repeated, changing only the random seed used for the random number generator used to evaluate whether protein binding occurs during a particular time interval, to determine how deterministic the expression levels of RNA and protein are in the simulation (Figure 3.2B). We observe general similarity, but with notable differences. The lowest expressed gene (at both the RNA and protein level) has a small increase in RNA expression at the RNA level, which leads to a larger increase at the protein level for most of the simulation period. For the other genes, little to no change in expression is observed at both the RNA and protein level. This result demonstrates that while the simulation is stochastic, the distribution of values of gene expression are predictable.

Each simulation is provided with a set of initial conditions to parameterize the interactions between genes. Each gene’s expression at the RNA and protein level begins at a random value between 0 and 2 molecules. In order to test how much impact the initial conditions of the simulation had on the molecular state in the long run, we generated a different set of initial conditions randomly for the same gene set used in Figure 3.2A. We then ran the simulation with the new initial conditions for 1000 time points (Figure 3.2C). Here, the expression of several genes changes substantially. Notably, the pink labeled protein, which is highest expressed in Figure 3.2A is three orders of magnitude lower expressed with different initial conditions. The lowest expressed green labeled gene from 3.2A is now expressed much more highly, at about the same level as the pink labeled gene. The turquoise labeled gene's expression also goes to zero at the RNA and protein level by the halfway mark in the simulation and does not make a resurgence. The other genes' expression remains similar to the previous initial conditions. These results indicate that the initial conditions can have a large impact on future cell states, even if the levels of all genes' expression are low initially.

With a mechanistic model of cell states that contains parameters corresponding to molecular mechanisms, we can perturb those parameters to attempt to understand the impact those parameters have on cellular states. Using the same baseline set of genes used in Figure 3.2A, we perturbed the splicing rate of each gene and reran the simulation (Figure 3.3A). Changing the splicing rate had some impact on the early states of the simulation, particularly at the RNA level, but in the long run very little difference in expression at the RNA or protein level is observed with the different splicing rates, which may indicate that splicing rate does not have a large impact on RNA and protein expression levels long term, at least when the splicing rate is only a small fraction of the half life of the molecules (as is usually the case in real cells) (Alpert, Herzel, and Neugebauer 2017).

Another important parameter is RNA and protein half-life. We observe that perturbing protein half life mostly yields a large change in protein expression of one gene, the pink gene, which is even higher expressed (Figure 3.3B). Only minor changes to RNA expression are observed, as might be expected. Perturbing RNA half-life leads to three genes having very low RNA expression levels. Protein expression of these genes swiftly follows suit, leaving those three genes at or near zero expression for the remainder of the simulation (Figure 3.3C).

3.3 Effect of gene perturbations on cellular states

Limitations to temporal profiling technologies currently make it difficult to track the impact of perturbations within the same single cell over time. We can evaluate how perturbations may impact cell states over time using simulated data. To evaluate the way a gene knockout impacts future expression states, we modified the simulation from Figure 3.2A. We allow the simulation to proceed for fifty time points and then allow no more of the RNA transcripts of the blue gene to be produced, mimicking the result of a fully deleterious mutation that leads to immediate RNA degradation (Figure 3.4A). The remaining RNA degrades rapidly, while the remaining protein degrades over the next ~150 time points. This knockout leads to substantial changes in gene expression, particularly at the protein level. The protein expression of the pink gene reaches its maximum three orders of magnitude lower than in the original simulation, while the tan gene ends three orders of magnitude higher. Two gene’s protein expression falls to zero, three stay relatively constant, and two others increase in expression about tenfold. Interestingly, only two of the observed changes in protein expression appear to be driven by changes in RNA expression level: the knockout gene and the orange gene. The other differences observed appear to be mediated via the direct regulation of proteins by proteins (both through degradation by inhibiting translation).

We compared this knockout simulation to perturbing the same gene at an epigenetic level. After time t=50, the blue gene was set to a non-accessible state (Figure 3.4B). The RNA expression slowly decreases over time as the remaining counts degrade, while a very low level of expression is allowed to occur despite general gene inaccessibility. Protein expression slowly falls as well, not reaching zero until 800 time points in. The slower change means the change in expression levels plays out slightly differently. The pink gene's protein expression falls to a similar level as in the knockout, however, the tan gene does not reach as high of an expression level and has similar expression as the turquoise gene, which is again increased in expression from the original simulation to the same level. The RNA expression of the orange gene does not start to fall until the last two hundred time points, leaving both its RNA and protein expression at close to the original level by the end of the simulation.

We can also investigate a wider variety of perturbations than can be performed in real cells using current technologies. For example, we can set the RNA expression level of a gene to zero instantaneously and then trace how each cell responds over time at an epigenetic, transcriptional, and proteomic level. In order to test what this sort of perturbation would look like, we created a simulation using the same gene set as Figure 3.2A and at time t = 100 we set the spliced RNA counts to zero. We repeated this perturbation every 100 time points through the end of the simulation. We observe that the expression of the gene rapidly returns to its original value after the first perturbation. This result suggests that the molecular state of the cell is able to reestablish the RNA expression of an upregulated gene, even in the absence of any RNA from that gene (Figure 3.4C). Despite rebounding in expression level five times, the expression profiles of the other genes in the simulation quickly reach approximately the same levels they occupy in the knockout simulation, which may indicate that even temporary loss of the blue gene's expression immediately begins moving the cells towards the state observed in the knockout simulation, while temporary reexpression of the blue gene does not restore the previous state.

3.4 Correlation and causality in a mechanistic simulation

The naïve expectation is that an mRNA and the corresponding protein will in most cases have strongly correlated expression values. However, due to the time lag between transcription and translation as well as due to the influence of other regulatory factors, this expectation will not necessarily hold. Profiling studies suggest that in many cases, an mRNA and its corresponding protein are almost entirely uncorrelated (Gry et al. 2009), (de Sousa Abreu et al. 2009), (Vogel and Marcotte 2012). In order to investigate this phenomenon, 100 simulations with different gene sets were generated and the correlation between the RNA and protein levels was assessed. The mean correlation between RNA and protein expression was 0.63, with a median of 0.80 (Figure 3.5A). The correlation coefficient for a substantial minority (10.8%) of genes was less than zero. Often, we see this occur because of the influence of another gene. For example, if gene 1 is strongly regulated at the protein level by gene 2 then gene 1 RNA and gene 1 protein may not be correlated because the primary determinant of gene 1 protein levels may be gene 2 protein levels. We also observe that some regulatory networks result in genes and proteins with higher correlations than others. Median correlation varies substantially between simulations randomly assigned different gene-level parameters as described previously (Figure 3.5B), with some simulations above 0.9 and others below zero. The simulations with median correlation below zero often have one or two genes that strongly regulate the protein levels of many other genes. These one or two genes are usually very highly expressed and thus largely make the expression level of other genes irrelevant to cell state due to the overwhelming influence of one to two genes. It is important to note that this low correlation is observed even with “perfect measurement” of the RNA and protein levels; without the measurement noise that would exist with any profiling technique used to measure the expression levels in real cells.

The goal of many genomics analyses is to extract mechanistic understanding from the data (Erbe et al. 2022). To accomplish this, many gene network inference methods have been developed, which attempt to infer causal regulatory relationships between genes using gene expression data (Margolin et al. 2006), (Chan, Stumpf, and Babtie 2017), (Huynh-Thu et al. 2010), (Osorio et al. 2020), (Matsumoto et al. 2017), (Papili Gao et al. 2018), (Deshpande et al. 2019), (Qiu et al. 2020). Despite the myriad of approaches, independent assessments indicate that these methods often cannot robustly predict known causal interactions between genes (Chen and Mar 2018), (Pratapa et al. 2020), (Stone et al. 2021). In order to better understand the challenge facing these network inference methods, we simulated 2000 cells with the same regulatory parameters and randomly selected an expression profile for each cell over 500 simulated time points. The resulting matrix mimics the output of single cell RNA-seq data, without any measurement noise. We performed this simulation experiment with both ten gene cells and 100 gene cells. We then used the resulting count matrix to find the mutual information each gene pair provided and selected the gene pairs with a causal regulatory interaction between them. The mutual information for causally interacting gene pairs was not significantly higher than non-causal genes for the ten or 100 gene data set (Figure 3.5C-D). This result indicates that the statistical relationships between genes that are not directly causal prevent differentiation of causal and non-causally related genes on the basis of predictive power, even in a noiseless simulation. While the rank distribution in the ten gene set is slightly shifted towards lower ranks, this shift is not nearly sufficient to reliably distinguish these gene pairs from non-causal ones. This result may help explain the difficulties the field of gene network inference has encountered.

In order to determine whether this result was based on the specifics of our simulation model or would also occur in other biologically-driven simulation frameworks models, we assessed whether the information between causally related genes was higher when using the BoolODE simulation (Pratapa et al. 2020). BoolODE takes a causal regulatory network as input and uses the regulatory relationships to parameterize a set of ordinary or stochastic differential equations to simulate single-cell gene expression values. We simulated 2000 cells from two different causal regulatory networks provided by BoolODE (Pratapa et al. 2020): one derived from studies of Gonadal Sex Determination (GSD) and the other designed to produce a trifurcation in cell lineage. In both of these simulations, the mutual information between a pair of genes was again insufficient to reliably distinguish causal gene pairs (Figure 3.6). While the GSD simulation does show a slightly higher proportion of causal gene pairs with high mutual information ranks, many of the causal gene pairs are still very low ranked, preventing robust causal inference for these gene pairs. The trifurcation simulation shows very little bias in rank distribution for the causal genes pairs, indicating there is nearly no signal distinguishing these gene pairs from random gene pairs. Taken together, the results from the MOMS simulation and BoolODE simulation suggest causal inference of direct mechanism from single-cell RNA-seq datasets may not be possible for a large subset of genes because the statistical relationship between expression values does not distinguish causally related and correlated gene pairs even in the clean simulated cases presented here.

3.5 Discussion

We present a mechanistic simulation of cell states across epigenetic, transcriptomic, and proteomic cell states, MOMS. MOMS includes parameters for splicing rate and degradation of genes, allowing the impact of these to be assessed. The simulations we have presented indicate that degradation rate parameters are substantially more impactful on the long term RNA and protein expression rate than splicing rate, though perturbations of either parameter have observable impacts on expression.

We examined the relationships between the expression values of RNAs and their corresponding proteins in order to determine if the proposed simulation matched reports from cells of frequent low correlations between these pairs. We did find a wide array of correlations between corresponding RNA and protein pairs, including many near or even below zero. This result suggests that our simulation captures important regulatory features that give rise to those dynamics *in vivo.* Additionally, the simulation provides explanations for how this phenomena can occur. Frequently, we find it is observed due to the overwhelming influence of another gene that obfuscates the relationship because protein or RNA expression is most strongly influenced by the other protein.

One limitation of MOMS is that it supplies output at discrete time intervals rather than allowing for continuous assessment of cell states. However, this choice allows for a highly computationally efficient simulation relative to a continuous model. For the purpose of allowing many simulations across many cells to be performed under many different perturbation conditions, we believe this tradeoff will often be worthwhile.

We further investigated the ability to distinguish correlation from causation using simulated RNA data. By randomly sampling cells from simulations over time (to approximate how scRNA-seq captures single time points of RNA expression) we find that causal gene pairs often have lower mutual information between them than other non-causal gene pairs. We further validated this result using a different type of simulation, BoolODE (Pratapa et al. 2020). Taken together, the inability to distinguish causal from non-causal gene pairs even in simulated data sets with a ground truth suggests robust causal gene network inference using only single cell RNA-seq count matrices as input may be subject to pervasive inaccuracies.

Figures

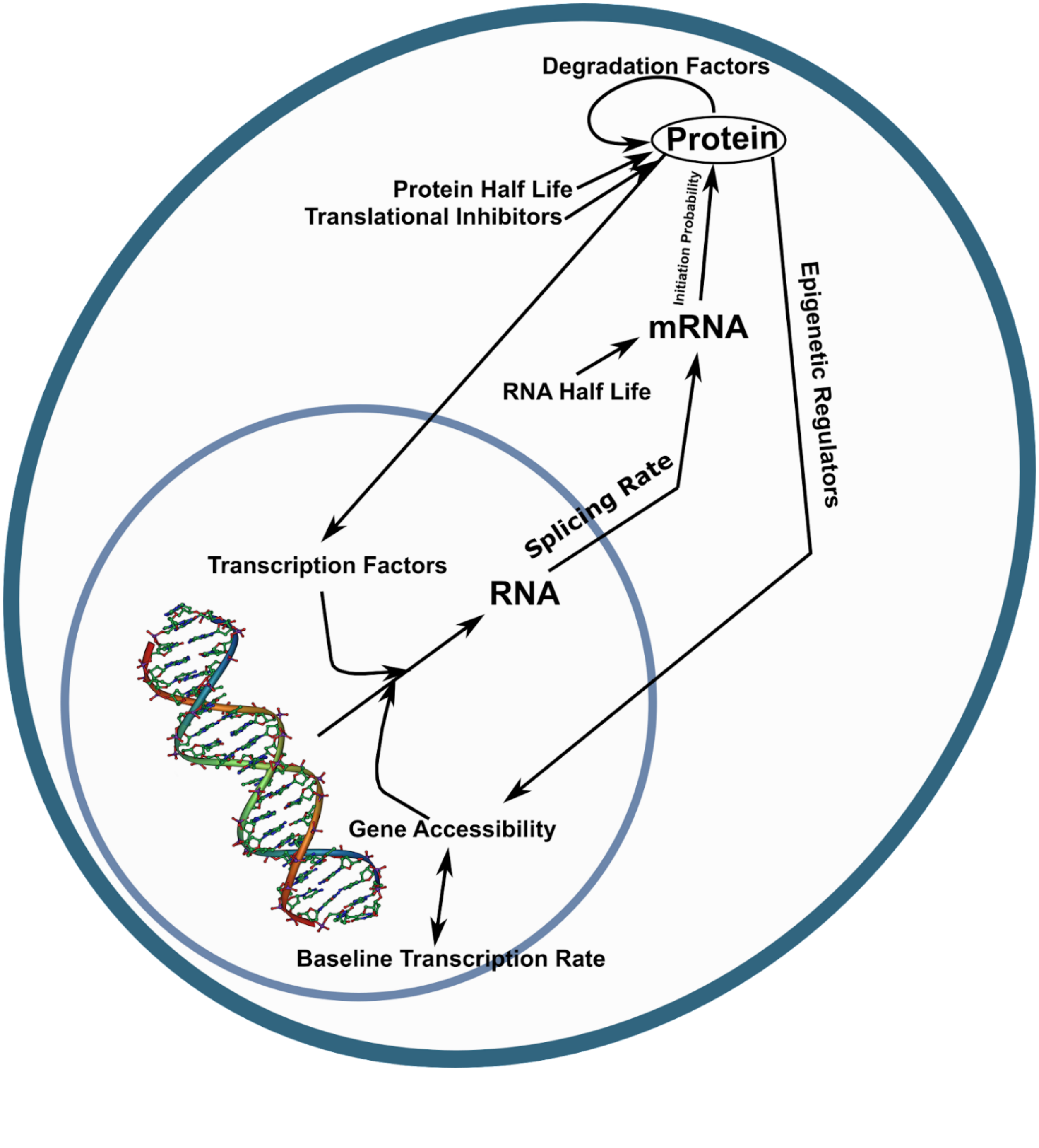

Degradation Factors
Protein
Protein Half Life
Translational Inhibitors
Initiation Probability
mRNA
Epigenetic Regulators
RNA Half Life
Splicing Rate
Transcription Factors
RNA
Gene Accessibility
Baseline Transcription Rate

Figure 3.1 – Overview of the Mechanistic Links and Parameters Incorporated in the Simulation

The simulation software proposed simulates genes with two copies that are accessible or not. The simulation accounts for baseline transcription rate and concentration of transcription factors when assessing whether transcription of a gene occurs. The RNA in the nucleus is then spliced before maturing to cytosolic mRNA which can be translated into protein, which can feedback onto the DNA as transcription factors or epigenetic regulators. Protein can additionally feedback on itself as a translational inhibitor or degradation factor. Otherwise, RNA and protein degradation occurs according to its half-life.

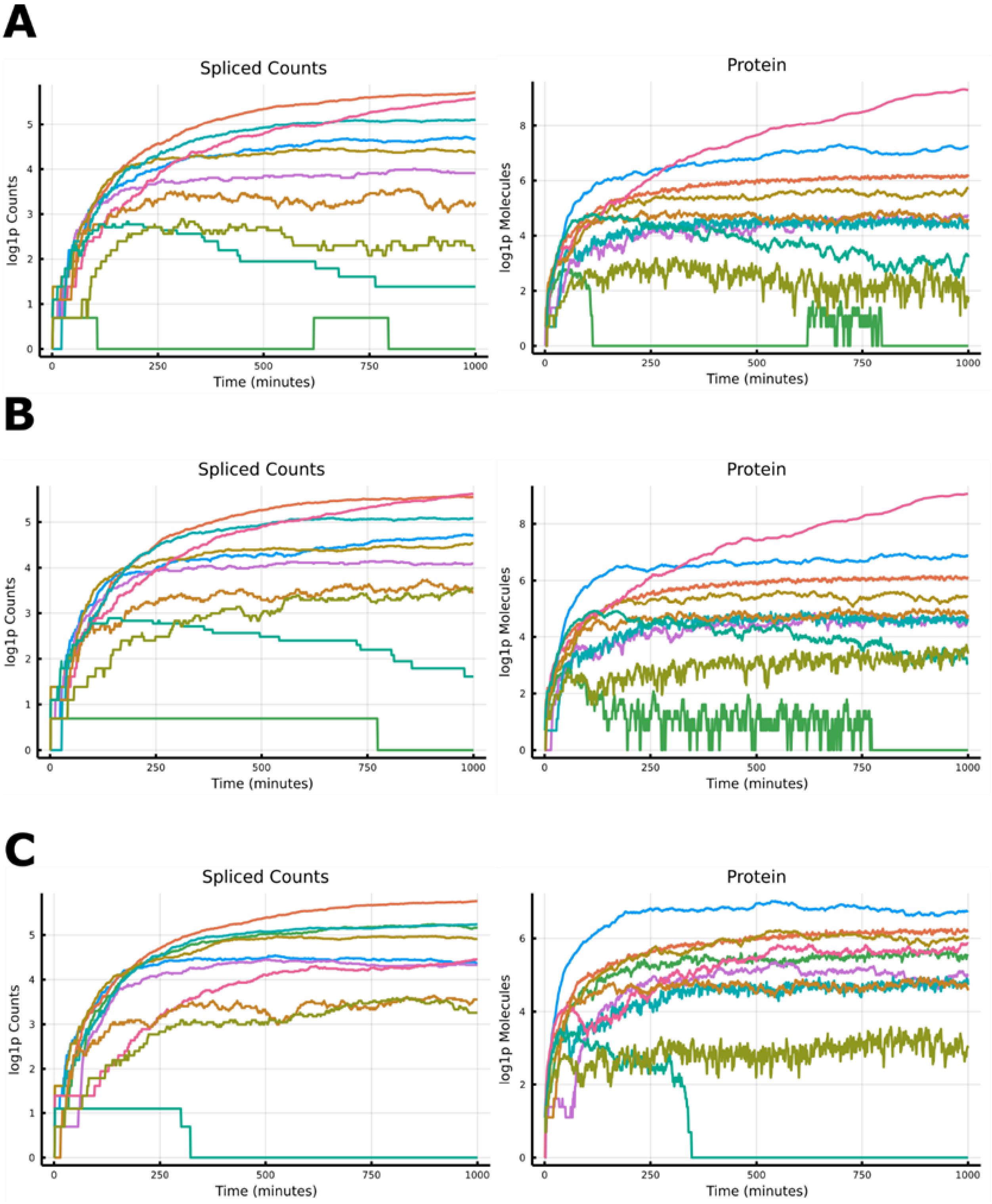

A
Spliced Counts
log1p Counts
0
1
2
3
4
5
0
250
500
750
1000
Time (minutes)
Protein
log1p Molecules
0
2
4
6
8
0
250
500
750
1000
Time (minutes)
B
Spliced Counts
log1p Counts
0
1
2
3
4
5
0
250
500
750
1000
Time (minutes)
Protein
log1p Molecules
0
2
4
6
8
0
250
500
750
1000
Time (minutes)
C
Spliced Counts
log1p Counts
0
1
2
3
4
5
0
250
500
750
1000
Time (minutes)
Protein
log1p Molecules
0
2
4
6
0
250
500
750
1000
Time (minutes)

Figure 3.2 Impact of Random Seed and Initial Conditions on Simulation of RNA and Protein Levels

A Spliced RNA counts and protein expression of a baseline ten gene simulation over 1000 time points. **B** Spliced RNA counts and protein expression with a different random seed than A. **C** Spliced RNA counts and protein expression with a different set of initial conditions from A.

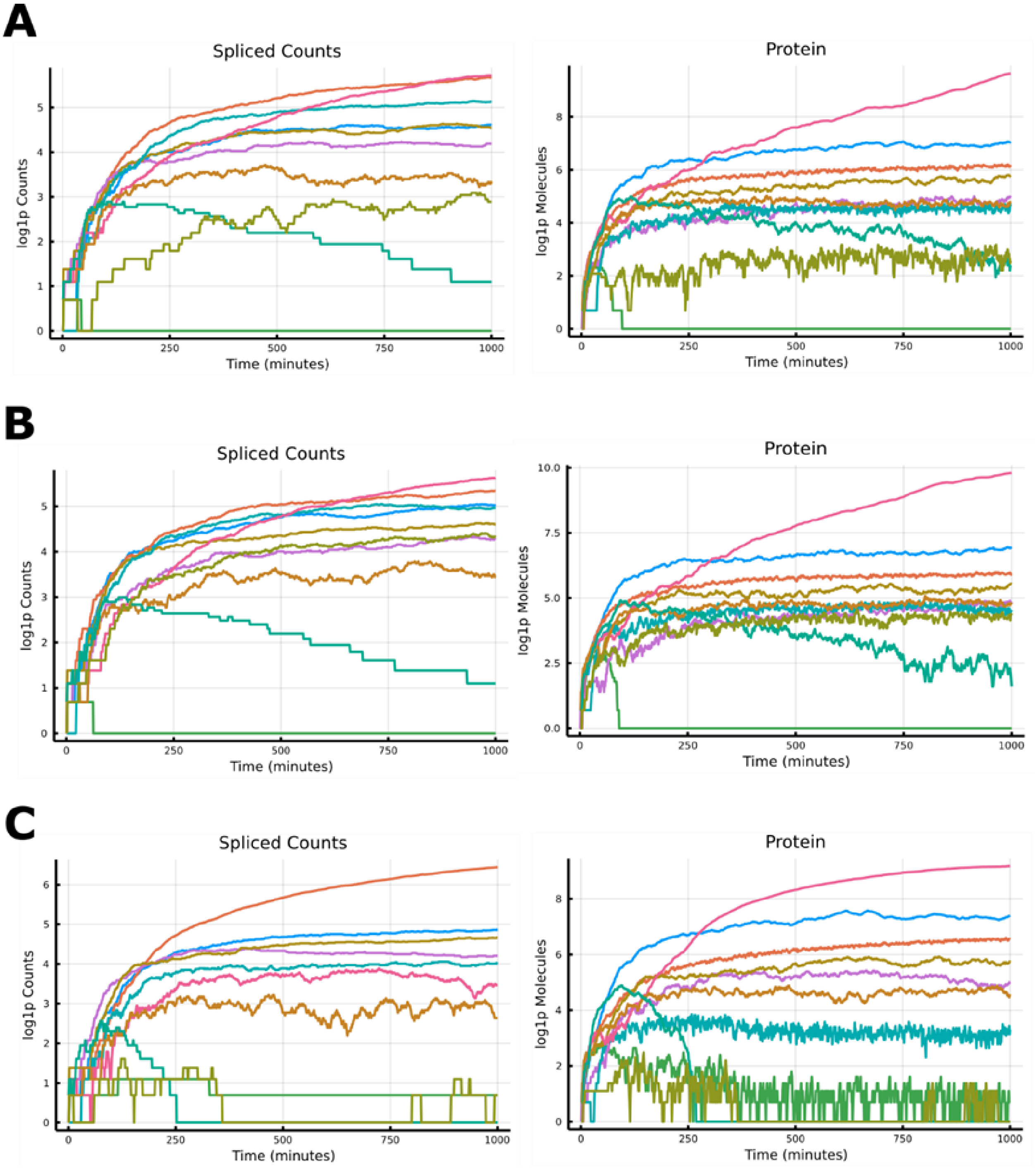


Figure 3.3 Impact of Changing Splicing Rate and Half Life Parameters

A Spliced RNA counts and protein expression after changing splicing rate of each gene. **B** Spliced RNA counts and protein expression after changing protein degradation rate of each gene. **C** Spliced RNA counts and protein expression after changing RNA degradation rate of each gene.

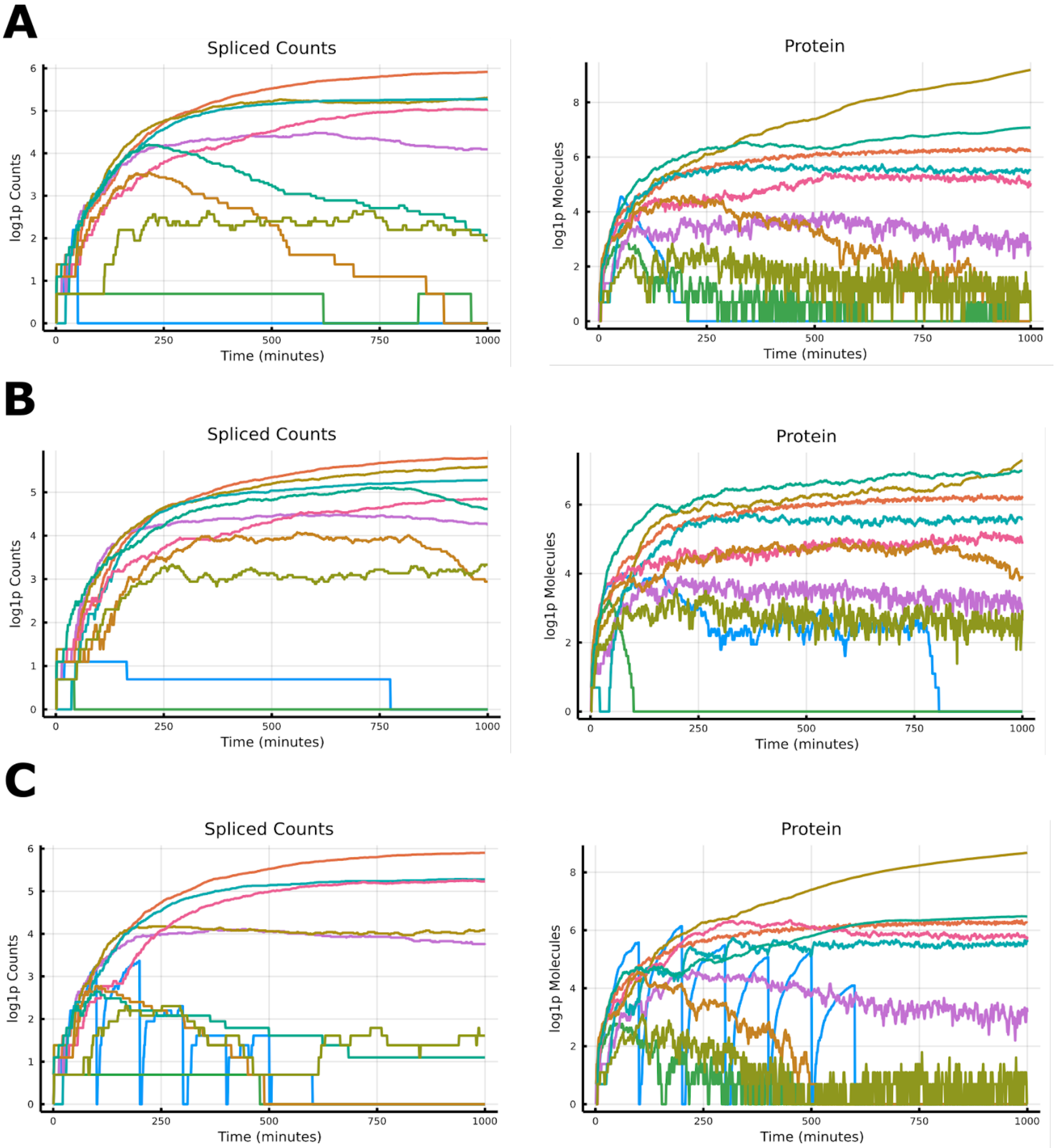
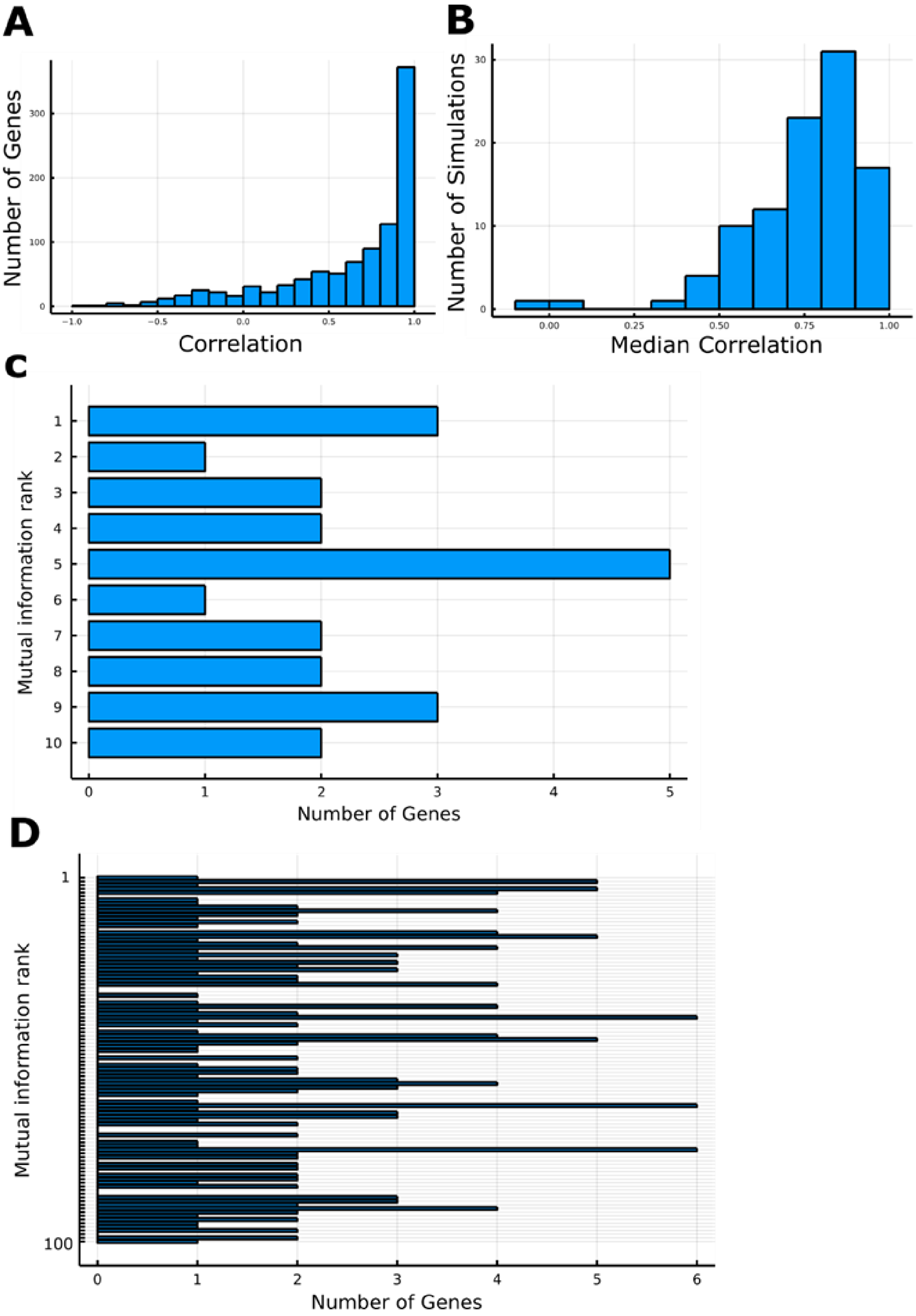


Figure 3.4 Impact of Perturbations of RNA and Protein Expression Levels over Time

A Spliced RNA counts and protein expression if gene 1 (blue) is set to an epigenetically repressed state after t=50. **B** Spliced RNA counts and protein expression if gene 1 (blue) is prevented from transcribing after t=50. **C** Spliced RNA counts and protein expression if gene 1 (blue) is set to zero every 100 time points.



A
Number of Genes
0
100
200
300
−1.0
−0.5
0.0
0.5
1.0
Correlation
B
Number of Simulations
0
10
20
30
0.00
0.25
0.50
0.75
1.00
Median Correlation
C
Mutual information rank
1
2
3
4
5
6
7
8
9
10
0
1
2
3
4
5
Number of Genes
D
Mutual information rank
1
100
0
1
2
3
4
5
6
Number of Genes

Figure 3.5 Correlation and Information between RNA and Proteins

A Histogram of the correlation of each corresponding RNA-protein pair across 100 ten gene simulations. **B** Histogram of the median correlation of each simulation between each corresponding RNA-protein pair across 100 ten gene simulations. **C** Histogram of the mutual information rank among all genes provided by each causal gene pair among 100 ten gene simulations. **D** Histogram of the mutual information rank among all genes provided by each causal gene pair among 100 hundred gene simulations.

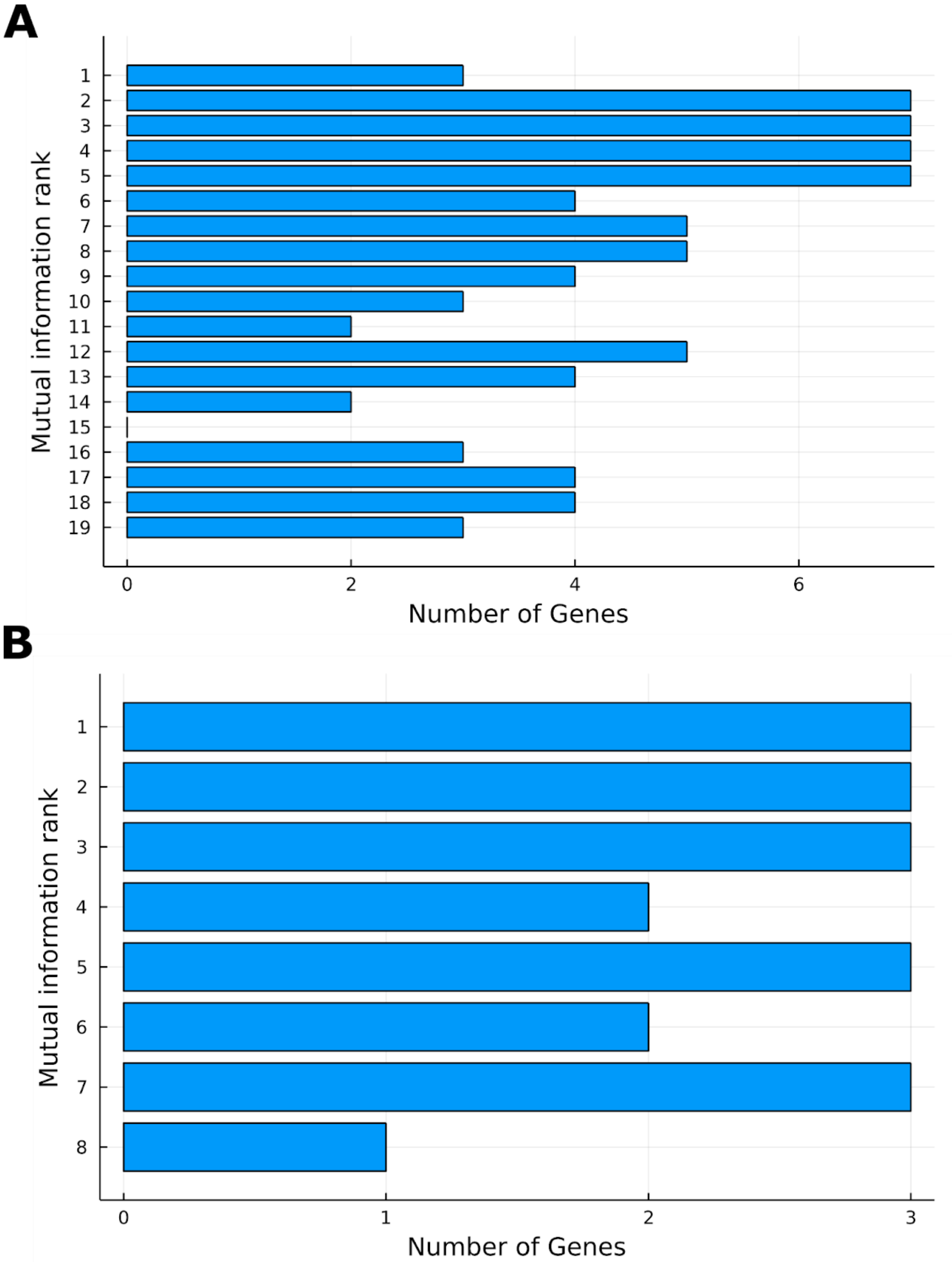


Figure 3.6 Mutual Information Between Causal Gene Pairs in BoolODE

A Histogram of the mutual information rank of causally related gene pairs compared to all gene possible gene pairs for each gene in the BoolODE GSD simulation across 2000 cells. **B** Histogram of the mutual information rank of causally related gene pairs compared to all gene possible gene pairs for each gene in the BoolODE trifurcation simulation across 2000 cells.

Chapter 4

Evaluating the impact of age on immune checkpoint therapy biomarkers

Introduction

The association of cancer incidence with age is well established and the phenomenon of age-related immune decline has been recognized for even longer (Gardner, 1980). Mutations and DNA methylation has been shown to accumulate with age and drive carcinogenesis (Tomasetti et al., 2017), (Horvath, 2013), (Klutstein et al., 2017), (Xie et al., 2018). Recent research has highlighted the specific changes that contribute to the general decline of the immune system that occurs as individuals age (Aw et al., 2007). Understanding the effect such alterations have on the anti-tumor immune response is critical for the informed development and application of immunotherapies to elderly patients.

Outside the context of cancer, older individuals are generally observed to have less effective immune responses to disease (Gardner, 1980). This observation is commonly associated with systemic immune aging. In particular, loss of T cell receptor (TCR) diversity (Britanova et al., 2014), decreased capacity of cytotoxic cells (Solana and Mariani, 2000), and increased inflammatory signaling (Franceschi et al., 2000) have been identified as age-related immune changes. These studies note the potential significance of these forms of immune aging on cancer, and indeed systemic immune aging has received considerable attention in the context of its effect on cancer development and progression (Fulop et al., 2017). Still, the potential translation of these findings to cancer therapeutics and patient care requires further comprehensive evaluation of the interplay between systemic immunity and the tumor immune microenvironment resulting from aging, particularly in the context of immunotherapy. In spite of the general immune decline associated with aging, the majority of clinical trial analyses suggest that elderly patients experience no reduced benefit or even increased benefit as compared to younger patients on ICB therapies (Kugel et al., 2018), (Elias et al., 2018), (Jain et al., 2019). However, there is still some contention on this point (Daste et al., 2017) and elderly patients are less likely to be treated with ICB therapies than their younger counterparts (Hurez et al., 2018), (Jain et al., 2019).

High-throughput molecular data from atlas studies provide new opportunities to comprehensively characterize the immune landscape of tumors (Thorsson et al., 2018), and are now sufficiently powered to evaluate aging-related changes (Wu et al., 2019), (Shah et al., 2020), (Chatsirisupachai et al., 2021). This study leverages genomics and clinical data from 9,523 patients across 31 cancer types from the Cancer Genome Atlas (TCGA), 37,961 patients across 8 cancer types from the Genomics Evidence Neoplasia Information Exchange (GENIE), 15,557 patients with breast, colon, or head and neck cancers from Caris Life Sciences (CLS), and 1,818 patients with breast cancer from Molecular Taxonomy of Breast Cancer (METABRIC), as well as genomics data from a pan-tissue reference of 948 non-cancer individuals from the Genotype-Tissue

Expression project (GTEx) (see Supplemental Data for a summary of patient characteristics in each cohort) to evaluate such age related changes in the tumor immune landscape. Because the immune microenvironment mediates ICB response, we focus our analysis of these these large-scale data to evaluate the impact of aging on the molecular and cellular biomarkers of ICB response, such as PDL1 expression (Patel and Kurzrock, 2015), tumor mutational burden (TMB) (Yarchoan et al., 2017), (Goodman et al., 2017), cell type composition of the ITME (Frankel et al., 2017), TCR diversity (Han et al., 2020), expression of other immune checkpoint genes (Taube, 2014), and expression of inflammation-related pathways such as interferon gamma (Cristescu et al., 2018), (Higgs et al., 2018) and TGFβ signaling (Tauriello et al., 2018). We further compile these analyses into a web application Cancer Associations with Molecular Aging (CAMA) to allow for further customized analyses of the cellular and molecular pathways altered with age pan-cancer. Our analyses from CAMA in the context of ICB biomarkers suggests that the aged ITME upregulates major pathways associated with immune response, although additional indicators of immune decline warrant future prospective clinical studies to provide databases of combined genomics and clinical data in order to directly evaluate the impact of age on the ITME in the context of ICB response.

4.1 Tumor mutational burden increases with age in most cancers, while T cell receptor diversity decreases

The large number of public domain genomics datasets from primary tumors and normal tissue in the literature provides the opportunity to characterize the impact of age on the ITME and ICB biomarkers. Due to the widespread use of tumor mutational burden

(TMB) as a primary clinical biomarker of ICB therapy (Yarchoan et al., 2017), (Goodman et al., 2017), we first examine the relationship of TMB with patient age. As has been previously reported among TCGA samples (Chalmers et al., 2017), (Qing et al., 2020), we find TMB significantly increases with patient age at diagnosis (1.02% increase per year of age, $p < 1 \times 10^{-16}$) (Figure 4.1A) pan-cancer in TCGA when modeling cancer type as a covariate. This association is further observed within most cancer types (Figure 4.1B-C), although both lung adenocarcinomas (-1.31% per year, q = 0.0072) and uterine carcinomas (-2.02% per year, q = 0.0022) demonstrate decreased tumor mutational burden with age. To validate these findings, we also investigate the relationship of age and TMB within the larger mutational data set provided by GENIE for eight ICB-approved cancer types. This analysis identifies a significant increase in TMB with age in all eight cohorts (Figure 4.1D-E). In contrast to TCGA, non-small cell lung cancer samples in the GENIE cohort show a small increase in TMB with age (0.2% increase per year, $q = 6.13 \times 10^{-4}$). We further identify significant increases in TMB among elderly CLS patients with colorectal and breast cancer ($q = 4.92 \times 10^{-15}$ and $q = 6.39 \times 10^{-11}$ respectively). In contrast to TCGA and GENIE, the null hypothesis is not rejected in HPV-negative head and neck cancers (q-value = 0.244). These results provide a robust indication of TMB increases with age across most if not all ICB approved cancer types.

The canonical interpretation of TMB as a biomarker for ICB therapy is that more mutations generally implies more immunogenic mutations, which in turn makes it more likely for an antigen to be displayed via MHC I that T cells are able to recognize, allowing them to target the corresponding tumor cells. Another factor in the likelihood of this recognition event is the number of antigens T cells infiltrating the tumor are able to recognize, defined by the TCR sequence carried by each T cell. The overall decline in the total number of unique TCR clones as part of the normal aging process (Yager et al., 2008), (Britanova et al., 2014), (Egorov et al., 2018) is well established in the literature. The process of thymic involution (the loss of thymus tissue with age) eventually ends the production of naive T cells and is the major driver of normal age-related decreases in T cell clonality (Aspinall and Andrew, 2000). However, the impact of carcinogenesis on age-related T cell clonality has not been fully characterized. To quantify aging-related changes in TCR clonality specific to the ITME, we leveraged estimates of TCR sequences previously generated with the miTCR algorithm (Bolotin et al., 2013) by (Thorsson et al., 2018) from RNA-seq data in TCGA to determine the association between TCR clonality and age. We define our metric of clonal diversity as the Shannon entropy multiplied by the number of unique clones divided by the total number of TCR sequencing reads to correct for variation in total number of T cells in each tumor sample. We determine that this TCR clonality measure significantly decreases with age for pan-cancer TCGA samples, including cancer type as a covariate (-0.0051 normalized Shannon entropy per year; $p = 1.48 \times 10^{-8}$) (Figure 4.1F), corresponding to a 0.26% predicted decrease per year in tumor TCR clonality relative to the mean normalized Shannon entropy of 1.95 observed pan-cancer. Among individual cancer types, we observe a significant increase in TCR clonality with age in patients with lung adenocarcinoma (0.013 per year; $q = 3.99 \times 10^{-3}$) and significant decreases in patients with breast (-0.01 per year; $q = 2.63 \times 10^{-3}$), uterine (-0.012 per year; $q = 0.011$), melanoma (-0.015 per year; $q = 3.99 \times 10^{-3}$), and gastric cancer (-0.01 per year; $q = 0.040$). These results indicate a general decrease in TCR clonality with age, though not necessarily a uniform one across cancer types.

4.2 Age correlates with ICB related gene expression among both patients with cancer and normal individuals

In addition to TMB and TCR clonality as biomarkers of ICB therapies, immune checkpoint gene expression can also be used as a biomarker for specific inhibitors. PDL1 expression is an established clinical biomarker to predict patient response to anti-PD1/PDL1 treatment (Patel and Kurzrock, 2015). More broadly, the efficacy of ICB immunotherapy is linked to the expression of target genes and their complementary receptors such as *PD1*, *PDL1*, *CTLA4*, *CD80*, and *CD86* (Taube, 2014), as well as to associated genes such as *PDL2, JAK2, LAG3, HAVCR2, TGFB1,* and *CXCL9* (Conway et al., 2018). While the expression of these genes is important for the efficacy of ICB therapy, their expression as a function of ageing has not been studied. In order to understand the relationship of the expression of these genes and age, we performed differential expression analysis in both TCGA and CLS tumor samples, as well as normal GTEx tissue samples.

In TCGA, we identify that of these listed genes, *PDL1, CD80, HAVCR2*, *LAG3*, *PDL2*, and *CXCL9* expression significantly increases with age (Figure 4.2A, p-values and effect sizes provided in Table 4.2), including cancer type as a covariate. We compare these findings to reference non-cancer samples from GTEx to assess whether there is any age-associated expression change in these genes in normal tissues (Figure 4.2A). As in the cancer tissue samples in TCGA, *PDL1*, *HAVCR2*, *LAG3*, *PDL2*, and *TGFB1* expression significantly increases with age among GTEx normal samples pan-tissue, while *JAK2* significantly decreases and no significant change is identified in *CD86* expression. *CTLA4* and *CD80* are very lowly detected across samples in GTEx and therefore do not enable comparison (see Methods). These results indicate that the gene expression differences observed in tumor samples are likely largely the result of the systemic effects of aging, possibly involving the higher levels of inflammation that have been reported in older individuals (Fulop et al., 2017), (Kovtonyuk et al., 2016).

We further investigate age-related changes in expression of these genes within each cancer type in CLS, TCGA, and METABRIC. Analysis of the CLS cohorts of colorectal, head and neck, and breast cancers identifies a significant increase in PDL1 expression via immunohistochemistry ($q = 1.03 \times 10^{-9}$), as well as increases in *HAVCR2* ($q = 0.0077$), *LAG3* ($q = 7 \times 10^{-4}$), and *PDL2* ($q = 0.0357$) RNA expression in colorectal cancer in elderly patients (Figure 4.2B) and a significant increase in *LAG3* expression ($q = 0.0112$) in patients with HPV-negative head and neck cancer (Figure 4.2C), while no significant changes in immune checkpoint gene expression were identified in the breast cancer cohort (Figure 4.2D). We identify significantly increased expression of *PDL2* and *CXCL9* in lung adenocarcinoma with age in TCGA. Head and neck, colorectal adenocarcinomas, and gastric cancer tumors in TCGA also demonstrate increased age-related expression trends in *PDL1* although they do not reach statistical significance, while melanoma, breast, bladder, and kidney cancers do not show any age-related association. We note that some TCGA studies have relatively low numbers of patient RNA-seq samples, limiting the statistical power of subtype-specific analyses, particularly when evaluating two highly heterogeneous variables (age and cancer type). Finally, we evaluate differential expression with patient age among METABRIC breast cancer samples and identify a significant decrease in *CD80* expression (q = 0.044), and no significant differences in the other immune checkpoint related genes assayed with age.

Gene set enrichment indicates age-related signaling changes in pathways associated with ICB response

To further evaluate the role of transcriptional regulation on ICB biomarkers, we performed additional analysis of several molecular pathways that have been shown to predict patient response to ICB therapies, including high Interferon Gamma signaling (Higgs et al., 2018), low TGFβ signaling (Tauriello et al., 2018), (Mariathasan et al., 2018), and low WNT pathway signaling (Xiao et al., 2018). These pathways are indicative of an immunostimulatory and immune-inhibitory tumor microenvironment, respectively. To determine if the expression of any of these pathways is altered with patient age, we perform differential expression and GO term enrichment on both TCGA tumor samples and GTEx normal samples. We observe increased enrichment of the GO_RESPONSE_TO_INTERFERON_GAMMA term in both TCGA tumors (normalized effect size (NES) = 2.05; q = 1.19 x 10^{-3}) and GTEx normal (NES = 2.37; q = 2.84 x 10^{-3}) samples with increasing age, decreased GO_RESPONSE_TO_TRANSFORMING_GROWTH_FACTOR_BETA in TCGA tumors (NES = -2.11; q = 1.03 x 10^{-3}), decreased signaling through the GO_CANONICAL_WNT_SIGNALING_PATHWAY in TCGA tumors (NES = -2.00; q = 1.03 x 10^{-3}), and decreased GO_POSITIVE_REGULATION_OF_CANONICAL_WNT_SIGNALING_PATHWAY in both TCGA tumors (NES = -1.57; q = 0.021) and GTEx normals (NES = -1.69; q = 5.64 x 10^{-3}) (Figure 4.3A).

We further sought to evaluate the impact of aging on these pathways within cancer types. Of particular note, we observe increased enrichment of Interferon Gamma signaling terms with age in most ICB-approved cancers: including colon, esophageal, head and neck, kidney, lung, and gastric cancer cohorts (Figure 4.3B). However, in melanoma and breast cancer cohorts interferon gamma signaling significantly decreases with age, and bladder cancers demonstrate no significant difference (Figure 4.3B). We identify decreased TGFβ signaling in breast, kidney, and gastric cancers, increased TGFβ signaling in lung and bladder cancers, and no significant change in the other aforementioned cohorts (Supplemental Figure 4.4A). We additionally observe decreased WNT signaling terms in breast, esophageal, kidney, melanoma, and gastric cohorts, increased signaling in lung and bladder cancers, and no significant change in colon and head and neck cohorts (Supplemental Figure 4.4B). While these results display heterogeneity in the relationship of age and the expression of tumor immune pathways, they suggest a general shift towards a more immunostimulatory signaling environment in older patients in most ICB-approved cancer types, which would be expected to improve response to ICB therapies. The similar association identified in the normal tissues corresponding to these tumor types from GTEx data further indicates that this shift may relate to the general increase in inflammation that has been repeatedly linked to biological aging (Kovtonyuk et al., 2016), (Fulop et al., 2017), (Franceschi et al., 2000).

4.3 Age related changes in promoter methylation align with most of the observed shifts in gene and pathway expression

Due to previous work suggesting that DNA methylation regulates tumor expression of *PDL1* (Asgarova et al., 2018), (Micevic et al., 2019), we hypothesize that, to the extent the observed expression increases in immune checkpoint genes occur within individual cancer types, they are driven by changes in DNA methylation. We leverage merged 450k and 27k methylation array data from TCGA (Thorsson et al., 2018) and use Illumina methylation array mappings to annotate CpGs to the promoters of specific genes. We find that of two probes annotated to the *PDL1* promoter region, methylation of one of the probes significantly decreases with age pan-cancer ($q = 3.27 \times 10^{-10}$; -0.3% of mean probe intensity per year of age), while the other does not demonstrate any significant change ($q = 0.232$). Methylation of CpGs annotated to the promoters of *LAG3*, *CTLA4*, *CD86*, *CD80*, and *HAVCR2* also decreases with age pan-cancer (Table 4.2). One CpG annotated to the TGFβ promoter is hypermethylated with age, while another has no significant change (Table 4.2). No CpGs within this data were annotated to *CXCL9*. We further investigate CpG methylation within individual cancer types. Similar to the gene expression patterns, we observe considerable heterogeneity in this data across cancer types. While most cancers approved for ICB therapy have decreasing promoter methylation trends with age among the majority of these CpGs, many do not reach statistical significance. However, both gastric and esophageal cancers demonstrate significant decreases in promoter methylation of *PDL1* and *CD86* with increasing age.

We additionally investigate whether age-related promoter methylation appears in concordance with observed changes in pathway expression pan-cancer in TCGA samples. While there is no significant change in methylation of gene promoters annotated to the GO_RESPONSE_TO_INTERFERON_GAMMA term (q = 0.663), GO_RESPONSE_TO_TRANSFORMING_GROWTH_FACTOR_BETA promoter methylation increases with age (NES = 1.97, q = 4.36 x 10^{-4}), as does GO_CANONICAL_WNT_SIGNALING_PATHWAY (NES = 2.05, q = 4.36 x 10^{-4}). These promoter methylation increases are concordant with the observed expression decreases of these pathways with increasing age. Taken together, these results suggest that age related methylation changes, as have been reported to occur in normal aging and oncogenesis (Easwaran and Baylin, 2019), (Easwaran et al., 2012), (Horvath, 2013), may drive some of the observed age-related expression-related changes in ICB therapy biomarkers.

4.4 Deconvolution of immune cell type abundance in tumor samples reveals an age-related decrease in T cell abundance and increase in macrophage abundance

Ultimately, ICB response relies on the balance between cellular subtypes contributing to immune attack and immunosuppression in the ITME. As a result, the immune cell infiltrate of the tumor microenvironment has been shown to be associated with response to ICB therapies, particularly the relative infiltration of T cells and NK cells with macrophages and MDSCs (Frankel et al., 2017). The large number of primary tumor transcriptional profiles across disease subtypes available from TCGA provides a unique cohort to estimate the impact of age on tumor immune cell composition. We apply the MIXTURE immune cell type deconvolution algorithm (Fernández et al., 2020) to infer the absolute proportions of immune cell types from RNA-sequencing data derived from pan-cancer TCGA samples. The algorithm provides an absolute proportion which describes the portion of total immune content that a particular immune cell type makes up in a sample, but is normalized to be comparable across all samples in the data set by multiplying the inferred relative proportion by a scaling factor that measures the total immune content in the sample. We then fit a linear model with age, including cancer type and patient sex as covariates, for each immune cell type to assess changes in immune cell infiltration as patients age. We find that overall T cell abundance slightly but significantly decreases with age in the ITME (-6.03×10^{-4} per year; mean proportion 0.198; q-value = 0.00175) while macrophages slightly but significantly increase in abundance (1.08×10^{-3} per year; mean proportion 0.662; $q = 4.45 \times 10^{-4}$). Detectable changes in the infiltration of NK cells, Dendritic cells, B cells, and other myeloid populations do not occur with age pan-cancer.

To compare the effect of aging in the ITME to that on the immune cell compositions of normal tissues, we applied MIXTURE to GTEx consortium RNA-sequencing data of post-mortem samples from individuals without cancer (GTEx Consortium et al., 2017) to infer cell type abundance across tissues. These results provide a non-cancer baseline for immune changes that occur across many individuals of varying ages to compare with our observations from tumor data. Similar to our TCGA and METABRIC analyses, we fit a linear model to each cell type in order to determine associations between cell type abundance and age both across and within normal tissues. In contrast to our findings in the pan-cancer ITME, in pan-tissue analyses we observe a significant increase in overall T cell absolute proportion with age (8.97×10^{-4} per year; mean proportion 0.106; q = 0.001). We further fail to find significant changes in macrophage levels (q = 0.870) with age. Additionally, we observe increases in NK cell proportion (0.0019 per year; 0.062 mean proportion; q = 5.13×10^{-14}) and decreases in other myeloid cell (monocytes, mast cells, eosinophils, neutrophils) (-0.0018 per year; 0.470 mean proportion; q = 0.0235) proportion that were not found among TCGA tumor samples. Recall that each effect size must be evaluated relative to the average proportion of immune infiltrate that cell type makes up (e.g. NK cells are expected to increase in abundance 153% over 50 years of life on average, while myeloid cells are expected to decrease only 19% over that same period despite essentially the same absolute proportion change per year). These results indicate differences between systemic immune aging and the effects of age on immune tumor infiltrate. Most notably, a very large systemic increase in NK cell abundance does not appear to be reflected in the tumors of older patients.

To determine the variance in age-related effects that occur within different cancer types, we then evaluate the association between age and immune composition for each cancer type with at least 100 samples that could be successfully deconvoluted by the MIXTURE algorithm. Non-significant deconvolution is generally due to a low content of the immune cells the algorithm searches for, and with this filtering only 8 tumor types in TCGA have over 100 samples after filtering. While several cancer types demonstrate age-related trends in T cell and macrophage abundance, these are only found to be statistically significant in breast cancers (Figure 4.4A). To determine if these results are robust across cohorts and whether lack of statistical significance in some cancer types is related to a lack of statistical power, we further examine large breast, head and neck, and colon cancer cohorts produced by CLS. As methodological validation, a different immune cell type deconvolution program, quanTIseq (Finotello et al., 2019), was used to estimate cell type abundance from RNA-seq data. Among 6,462 patients with breast cancer, a significant increase in M2 macrophage infiltration was identified with increasing age, while no significant difference was observed among infiltrating T cell abundance (Figure 4.4B). In contrast, within 7924 patients with colorectal cancer and 527 patients with HPV-negative head and neck cancer, no significant differences in macrophage or T cell immune cell fraction are observed. We further investigate this association among 1,818 METABRIC patients with breast cancer, again using MIXTURE for immune cell type deconvolution. We identify a similar decrease in T cell abundance with age (-6.57 x 10^{-4} per year; mean proportion 0.268; q = 0.00188) and increase in Macrophage abundance with age (1.38 x 10^{-3} per year; mean proportion 0.45; q = 8.21 x 10^{-9}) (Figure 4.4C), as well as a significant decrease in B cell abundance (4.22 x 10^{-4} per year; mean proportion 0.0554; q = 8.51 x 10^{-4}) that we did not observe in TCGA breast cancer data. This analysis thus identifies age-related macrophage proportion increases with patient age across three different breast cancer cohorts (TCGA, METABRIC, and CLS) using two different computational microdissection methods.

4.5 Patient age associates with little to no detectable difference in survival outcomes after ICB treatment

While the genomics datasets we have examined can help uncover the molecular and cellular pathways of mechanistic biomarkers for ICB that are altered by age, they cannot directly evaluate therapeutic response. Previous analyses of the impact of age on ICB therapeutic efficacy in clinical trials (Kugel et al., 2018), (Elias et al., 2018), (Daste et al., 2017), (Jain et al., 2019) have remarked on the limited numbers of older patients treated with ICB available for their analyses, and the need for further investigation of this subject. To provide additional insight into this question, we investigate the relationship between age and outcome. A recently published cohort of anti-PD-1 treated patients with renal cell carcinoma (Braun et al., 2020) had age available for 985 patients along with progression free survival (PFS) and overall survival (OS). We identify no statistically significant difference in progression free survival (PFS) or overall survival (OS) with age both based on a log rank test (p = 0.25 and p = 0.29, respectively) and multivariate cox proportional hazards analysis (HR = 0.994 [0.987-1.001], p = 0.09 and HR = 1.001 [0.994-1.009], p = 0.72, respectively), including sex, number of prior therapies, and metastatic origin as covariates. We also investigate survival differences in 11,888 ICB treated patients with melanoma, lung, kidney, head and neck, or urothelial cancers collected by the United States Department of Veterans Affairs (USVA) (La et al., 2020). A multivariate Cox proportional hazards model fit for overall patient survival, including cancer type and sex as covariates, identifies a statistically significant reduction in overall survival, of small effect size, for patients with increasing age (HR = 1.005 [1.001-1.009], p = 0.01).

4.6 High-throughput molecular databases inform an atlas of immune aging in cancer and healthy tissues

Understanding the impact of patient age on likelihood of response to immunotherapies is a subject of clear clinical relevance and investigating relevant biomarkers of said response forms the central focus of this work. Still, the comprehensive analysis of these data was based on general characterization of aging-related molecular shifts in tumors and the tumor microenvironment. To that end, we provide a web application containing these results to enable custom analyses of the relationship of age to molecular changes genome-wide: http://www.lab-apps.onc.jhmi.edu/CAMAAtlas. The Cancer Associations with Molecular Aging (CAMA) atlas is informed from analysis of 9,523 patients across 31 cancer types from TCGA, 37,961 patients across 8 cancer types from GENIE, 1,818 patients with breast cancer from METABRIC, and a pan-tissue reference of 948 non-cancer individuals from GTEx.

Briefly, the web-based application includes distinct panels for each of the analyses of distinct molecular modalities and datasets, based upon the data that are available from each cohort. The application allows for exploration of associations of TMB with age by cancer subtype in both TCGA and GENIE. The application further allows for customized evaluation of tumor-subtype changes relative to tissue-specific changes in gene expression through differential expression analyses in TCGA and GTEx, respectively. While the analyses presented in this study are limited to gene expression changes in ICB biomarkers, the CAMA web-application allows users to search for genes of interest across the entire genome allowing for evaluation of further age-related changes in the immune context and beyond. The application allows for further evaluation of the regulatory changes associated with these transcriptional alterations through GO enrichment analysis (TCGA and GTEx) and DNA methylation changes with patient age

(TCGA). The CAMA atlas is thus intended to act as an initial resource for further studies of the relationship between molecular features of cancers and aging. The relationship of a particular molecular feature (gene expression, gene promoter methylation, pathway enrichment, cell type abundance) with age can be queried by individual cancer type or across cancers. This atlas is meant to provide a resource that broadly characterizes cancer genomic associations with patient age and can be used to perform customized analyses. These relationships are often available in multiple cohorts, allowing for computational validation of identified associations.

4.7 Discussion of age related shifts in tumor molecular enivironment

This study presents an atlas of age-related shifts in the genomic, transcriptomic, and immune tumor environment. The effect of patient age on tumor characteristics has not been thoroughly explored in most cancer types. Here we analyze genomics and clinical databases from a total of 77,732 cancer patients with 31 different cancer types to generally characterize relevant associations between age and these molecular markers, which we provide the broad results of as the CAMA atlas http://www.lab-apps.onc.jhmi.edu/CAMAAtlas.

We hypothesize that the relationship between age and cancer makes understanding the impact of aging on cellular and molecular pathways an important consideration for precision medicine. Indeed, the general link between increased age and reduced immune effectiveness has naturally inspired caution and concern about the treatment of elderly patients with ICB therapies. Therefore, in this study, we leverage multiple large-scale cancer genomic cohorts to characterize the impact of age on established ICB biomarkers and contextualize previous clinical findings that older patients counterintuitively experience either no reduced benefit or increased benefit from ICB immunotherapies as compared to younger patients (Kugel et al., 2018), (Elias et al., 2018), (Jain et al., 2019). Our analysis identifies several possible explanations for these data based on currently established and developing predictors of ICB response. Patient age at diagnosis is associated with increases among several biomarkers associated with effective ICB response, including notably increased TMB, increased expression and decreased promoter methylation of immune checkpoint genes, increased Interferon Gamma signaling, decreased TGFβ signaling, and decreased canonical WNT signaling. The induction of these immunostimulatory biomarkers may be related to normal mutational accumulation with age, the increased inflammation that has been observed in normal systemic aging(Kovtonyuk et al., 2016), (Fulop et al., 2017), (Franceschi et al., 2000), and previously identified age-related methylation changes (Easwaran and Baylin, 2019), (Easwaran et al., 2012), (Horvath, 2013). Expected to act in opposition to these immune effects, we observe concurrent features of immunosuppression with age such as decreased TCR diversity and T cell infiltration as well as increased macrophage abundance, in some cancer types. However, it is critical to note that the effect size of TCR decreases with age in pan-cancer is quite small (on average a -0.26% change per year of age). Further, the decrease in T cell abundance and increase in macrophage abundance is not only small (on average, -0.3% and 0.16% change per year, respectively), it is also only statistically significant in one individual cancer type - breast. Altogether, these results support an adapting immune landscape with age that nonetheless retains characteristics associated with effective ICB response. Nonetheless, we note that all results of this work are correlative and thus a large-scale prospective study collecting genomics for immunotherapy treated elderly patients is warranted to generate a causal understanding of the effects of age on the immune response to cancer.

We complement our molecular studies with corresponding analysis of patient outcomes from large scale clinical databases for two large cohorts of ICB treated patients containing patients across an array of ages. Among the renal cell carcinoma cohort published by (Braun et al., 2020) we identify no significant difference in progression free survival or overall survival with age, supporting the results of previous clinical studies. However, among a large group of patients collected by the USVA, we find a small decrease in overall survival with age. It is notable that this slight overall survival difference observed could be related to general age-related frailty rather than differences in immunological efficacy. This point is supported by previous work published on this USVA cohort, which showed that a frailty status assessment considerably better differentiated therapeutic response in each cancer type than did patient age (La et al., 2020). The immunological biomarkers assessed in this study further support the interpretation that most of the small worsening in survival outcomes sometimes observed for older patients is the result of increased systemic frailty rather than decreased efficacy of the therapy itself. Future large-scale cohort studies of aged populations with combined outcomes, frailty measures, and genomics data are critical to fully delineate the relative impact of frailty and functional mechanisms of ICB response on its efficacy in the elderly population.

This study additionally includes normal tissues in our analyses in order to understand whether the associations noted appear to be a normal consequence of age or an interaction between aging and tumor biology. The associations established between age and ICB biomarkers largely recapitulate in GTEx normal samples (ICB gene expression, immune pathway enrichment), or, when not assessed here, have already been thoroughly established in the literature (TCR diversity (Aspinall and Andrew, 2000), (Yager et al., 2008), (Britanova et al., 2014), (Egorov et al., 2018), mutational accumulation (Morley, 1998)). The major exception identified to this concordance between tumor and normal aging is the large increase in NK cells with age in normal tissues, which has been previously identified in the literature (Solana and Mariani, 2000), (Gounder et al., 2018). This NK cell increase is not observed pan-cancer and is only observed in one cancer type cohort studied, Caris breast, where there was only a ~5% increase on average between the youngest and the oldest patients, compared to a ~150% increase in GTEx samples. This result suggests that while NK cell proportion increases with age, they either are not able to proportionately respond to immune stimuli and infiltrate into the aged tumor tissues or that aging biology interacts with tumor biology to inhibit the infiltration of NK cells. NK cells have been shown to play a significant role in ICB efficacy and general tumor immunity (Shimasaki et al., 2020), (Freeman et al., 2019), (Lo et al., 2020), (Jhunjhunwala et al., 2021) and thus this observation may be therapeutically relevant, particularly if these accumulated NK cells can be stimulated to infiltrate the tumors of elderly patients.

Beyond their relevance to ICB alone, the molecular and cellular changes inferred from the CAMA atlas may support selection of precision medicine strategies based on molecular and cellular changes in elderly patients. For example, we identify macrophage increases with age in 3 different breast cancer cohorts (TCGA-BRCA, METABRIC, CLS-Breast) with two different computational microdissection methods (MIXTURE and quanTIseq). Combination therapeutics to target immunosuppressive cells are emerging as a common therapeutic approach to sensitize tumors to immunotherapies. For example, there are several strategies currently in development to target tumor-associated macrophages (Chanmee et al., 2014),(Poh and Ernst, 2018),(Lee et al., 2019). These results suggest that elderly patients with breast cancer may be particularly promising candidates for these therapies. Thus, characterizing age-related changes in these distinct cellular populations in the tumor microenvironment can further illuminate combination therapeutic strategies specific for elderly patients.

To ensure that our data was sufficiently powered to analyze aging-related effects of tumors and their microenvironments, we leverage large scale databases that contain predominantly bulk profiling technologies. It is important to note the limitations of bulk expression data for some of the analyses in this work. Notably, our aging-related analyses of cell types relies on computational microdissection to provide estimates of proportional representation on each cell type in each sample studied. However, these techniques are only effective for samples with substantial immune infiltration, limiting the number of tumors that could be included in this analysis. Moreover, these bulk data do not enable discovery of cell-type specific molecular pathways that are altered by aging. Some computational methods have been developed to attempt to regress out effects of individual cell types on bulk expression data to perform such cell-type specific differential expression analysis. However, these techniques will be confounded in cases in which immune genes also serve as cell type markers, limiting the applicability of these techniques for the analyses in our atlas. A further limitation of computational microdissection methods used is that they estimate cell type abundance, but not cell state. Single cell data is essential to further evaluate immune cell functionality and quality in the ITME. While large-scale single cell studies of aging have been generated in healthy tissue for mouse models (Tabula Muris Consortium, 2020), to date these studies are for small cohorts in tumors that are not sufficiently powered to identify immune cell state transitions associated with aging. Therefore, future single-cell pan-cancer characterization from projects such as the Human Tumor Atlas Network (Rozenblatt-Rosen et al., 2020) will be critical to validate these results and further expand our atlas to delineate the role that aging-related changes to immune cell function play in cancer.

4.8 Methodological Details

Method Details

RNA-Sequencing Data

TCGA RNA-sequencing data processed and normalized according to https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/ was downloaded from the GDC Data Portal on August 8th, 2019, filtering for all TCGA samples with patients above 30 years of age. Patients under 30 were excluded to focus on ITME changes in adult populations, which are more likely to generalize to the majority of cancer patients.

GTEx RNA-sequencing counts version 8 were downloaded from the GTEx Portal on November 12th, 2019. Only individuals over 30 were included in the final analysis, to be comparable with filtering of TCGA.

METABRIC RNA-seq counts were downloaded from cBioPortal on October 20th, 2020 as provided by (Pereira et al., 2016).

TMB Data

To find the association of patient age and number of tumor mutations we downloaded the mutation counts provided for each sample pan-cancer in TCGA from the GDC data portal on August 8th, 2019. GENIE TMB counts were downloaded from the GDC data portal for all patients with cancers on January 13th, 2021.

TCR Clonality Data

TCR clonality was estimated from TCGA RNA-seq data using the miTCR algorithm (Bolotin et al., 2013), as previously published by (Thorsson et al., 2018). These data were published publicly on the GDC data portal and downloaded from the link provided in the Key Resources Table for use in this study.

DNA Methylation Data

Merged 450k and 27k DNA methylation array data preprocessed by (Thorsson et al., 2018) were published on the GDC data portal and downloaded from the link provided in the Key Resources Table for use in this study.

All statistical analyses were performed using R version 4.0.2. Statistical significance is evaluated as $p < 0.05$ after the Bonferroni-Hochberg procedure was applied in cases of multiple hypothesis testing.

Modeling the Age Associations of Number of Tumor Mutations and Normalized TCR Clonality

We transformed the TMB counts data with a natural log, which we used to fit a log linear model and Cox proportional hazards model, using cancer type as a covariate in the log linear model and cancer type and age at diagnosis as covariates for the Cox model. We additionally fit log linear models between TMB and patient diagnosis age within each TCGA cancer type study.

TCR clonality is assessed using miTCR (Bolotin et al., 2013) results previously published by Thorsson et al., 2018 (Thorsson et al., 2018). Our immune cell type deconvolution results demonstrate there may be decreased infiltration of T cells with increasing age, so to avoid biasing our results, the Shannon Entropy is multiplied by the number of unique TCR clones divided by the total number of TCR reads. We then fit a linear model for the association of age with this TCR clonality measure, including patient sex and cancer type as covariates. We again use a Cox Proportional hazards model to assess if normalized Shannon entropy is a relevant survival prognostic, using the same survival function and covariates as described above. We additionally fit linear models between normalized Shannon entropy and patient diagnosis age within each TCGA cancer type study.

Differential Expression Analysis with Age

Differential expression analyses from both TCGA and GTEx data were performed on all samples from individuals of at least 30 years of age. The R edgeR package version 3.30.3 was used for normalization and identification of differentially expressed genes with age. Age at diagnosis was modeled as a continuous variable, including cancer type as a covariate for the TCGA analysis and tissue type as a covariate for the GTEx analysis. Immune cell type proportions were included as covariates in each analysis to account for age-related differences in abundance. Genes were considered differentially expressed below an FDR adjusted p-value of 0.05. Differential expression analysis for diagnosis age was analogously performed on each cancer type separately that had at least 100 samples, though cancer type was naturally no longer included as a covariate.

Gene Set Enrichment Analysis

The fgsea R package version 1.14.0 (Sergushichev, 2016) was used to perform gene set enrichment analysis from differential expression results with age from TCGA and GTEx, produced as described above. GO terms were downloaded from MsigDB (Liberzon et al., 2011) using the msigdbr R package Version 7.2.1. GO enrichment was determined for all terms both pan-cancer and within each TCGA cancer type study and terms related to Interferon Gamma, TGFβ, and WNT were visualized.

Differential Methylation Analysis with Age

Merged 450k and 27k DNA methylation array data was used to examine the relationship between age and DNA mathylation. A linear model for diagnosis age was fit using R version 4.0.2 to data from each CpG, including cancer type as a covariate. CpG methylation was considered significantly different with age if the FDR adjusted p-value for the diagnosis age term was less than 0.05. Annotations of CpG sites to gene promoters were retrieved from the IlluminaHumanMethylation27k.db R package Version 1.4.8. The same process was repeated among each TCGA cancer type study, using a linear model between each CpG and patient diagnosis age. Gene set enrichment analysis was performed by using the differentially methylated CpGs that are annotated to gene promoters. This analysis was performed as described above using the R fgsea package version 1.14.0 (Sergushichev, 2016).

Immune Cell Type Deconvolution from Bulk RNA-Sequencing Data

The MIXTURE algorithm (Fernández et al., 2020) builds on the nu-Support Vector Regression framework used by CIBERSORT (Newman et al., 2015) for particular use with noisy tumor samples. MIXTURE applies Recursive Feature Selection to make the cell type deconvolution more robust to noise and collinearity, and was thus designed to improve performance on tumor data.

We run MIXTURE using a population-based null distribution and the nu-SVM Robust RFE method on the preprocessed RNA-sequencing data from both TCGA and GTEx. A signature expression matrix (LM22 from Newman et al) (Newman et al., 2015) is used to determine the proportion of 22 immune cell types in each sample. MIXTURE returns both relative and absolute proportions of immune cells. Absolute proportions were used for all analyses of TCGA and GTEx datasets. MIXTURE provides a p-value for the cell type deconvolution performed. Only samples with a deconvolution p-value less than 0.05 were used in the final analyses, leaving 3576 patient samples remaining in TCGA and 1689 in GTEx. A further 29 TCGA patients had received treatment prior to sample collection, and were removed to avoid biasing of results.

Modeling the Association of Immune Cell Type with Age

Linear models are fit to investigate the association between the absolute proportion of each immune cell type and the initial diagnosis age in TCGA. The models are fit separately for each cancer type as well as jointly with cancer type and patient sex as covariates. Significance is assessed using Benjamini-Hochberg FDR correction for multiple testing across all cell types tested.

Higher order cell types are defined by adding together individual substituent cell type values and dividing by the sum of all cell types, the result of which is used as the predictor variable in the linear model.

GTEx data was similarly analyzed using linear models, including sex and tissue type as covariates.

Survival Modeling for Patient Age among Braun et al. 2020 and USVA cohorts

We fit multivariate Cox proportional hazards models to survival data from 985 anti-PD1 treated patients with renal cell carcinoma collected by Braun et al. 2020 (using progression free survival and overall survival data provide in Supplementary Table 1 of Braun et al. 2020) and to 11,888 ICB treated patients with melanoma, lung, kidney, head and neck, or urothelial cancers, collected by the United States Department of Veteran Affairs. This model was fit using the R survival package version 3.1-12. We additionally produce Kaplan-Meier survival curves based on the Braun et al., 2020 data set, separating the curves into 65 and under and 66 and older age groups for each data set. These curves were fit using the R survival package version 3.1-12 and the R survminer package version 0.4.8.

Caris Life Sciences Data and Analyses

15,557 Caris samples were analyzed using next-generation sequencing (NextSeq, 592 Genes and WES, NovaSEQ), IHC and WTS (NovaSeq) (Caris Life Sciences, Phoenix, AZ). PD-L1 expression was tested by IHC using 28-8 and 22c3 (Agilent) and SP-142 (Spring Biosciences) (positive cut-off >1% for CRC and HNC, >5% for BC). TMB was measured by totaling somatic mutations per tumor. Immune checkpoint gene expression was normalized to the median expression in the lowest age quartile. Immune cell fraction was calculated by quanTIseq (Finotello et al., 2019). Immunotherapy biomarkers, immune checkpoint gene expression and immune cell fraction was compared across four age quartiles. Median transcripts per million (TPM) were normalized to the median TPM value in quartile 1. Statistical significance was determined using chi-square and Wilcoxon rank sum test and adjusted for multiple comparisons using the Benjamini-Hochberg procedure.

Figures

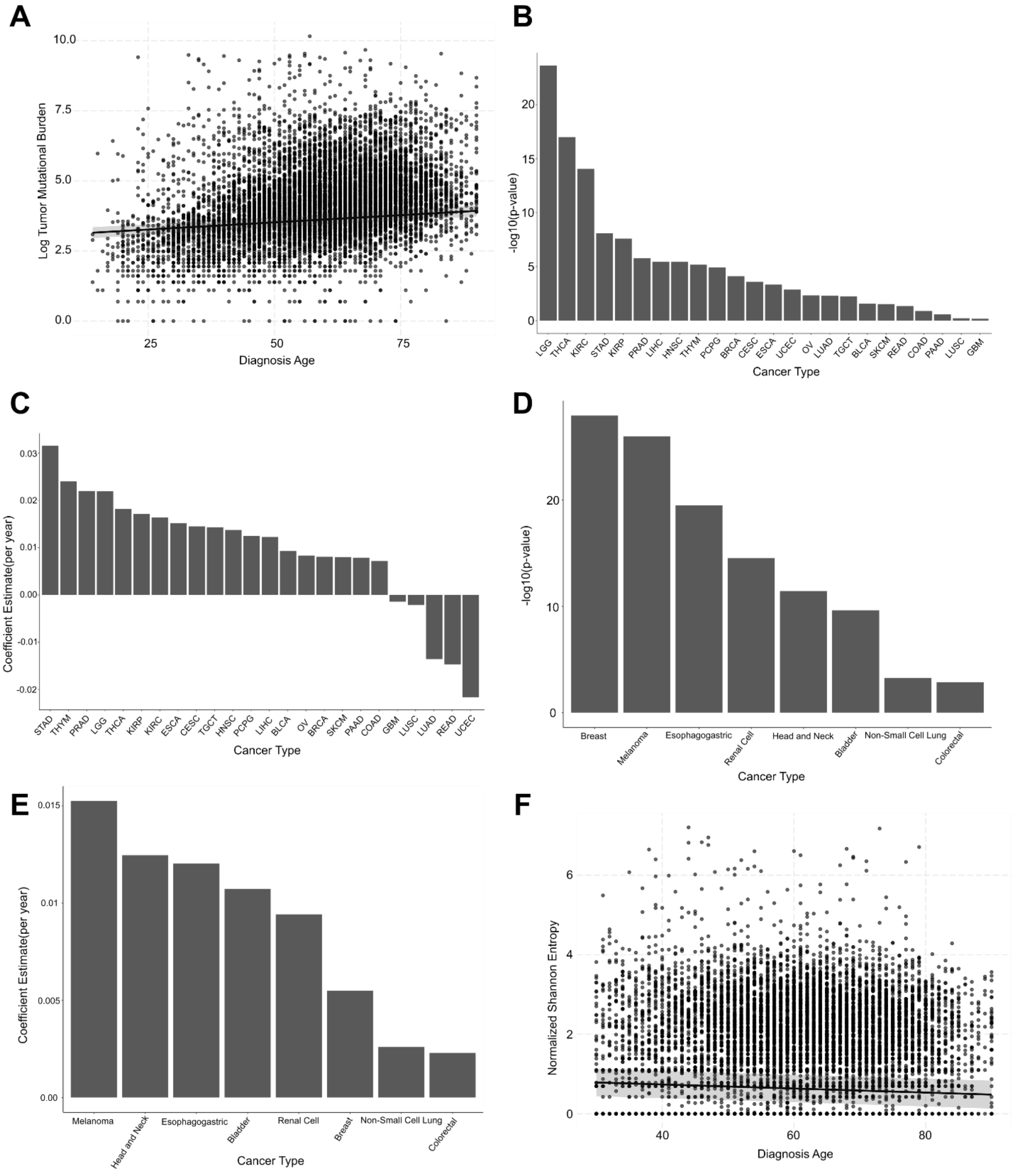
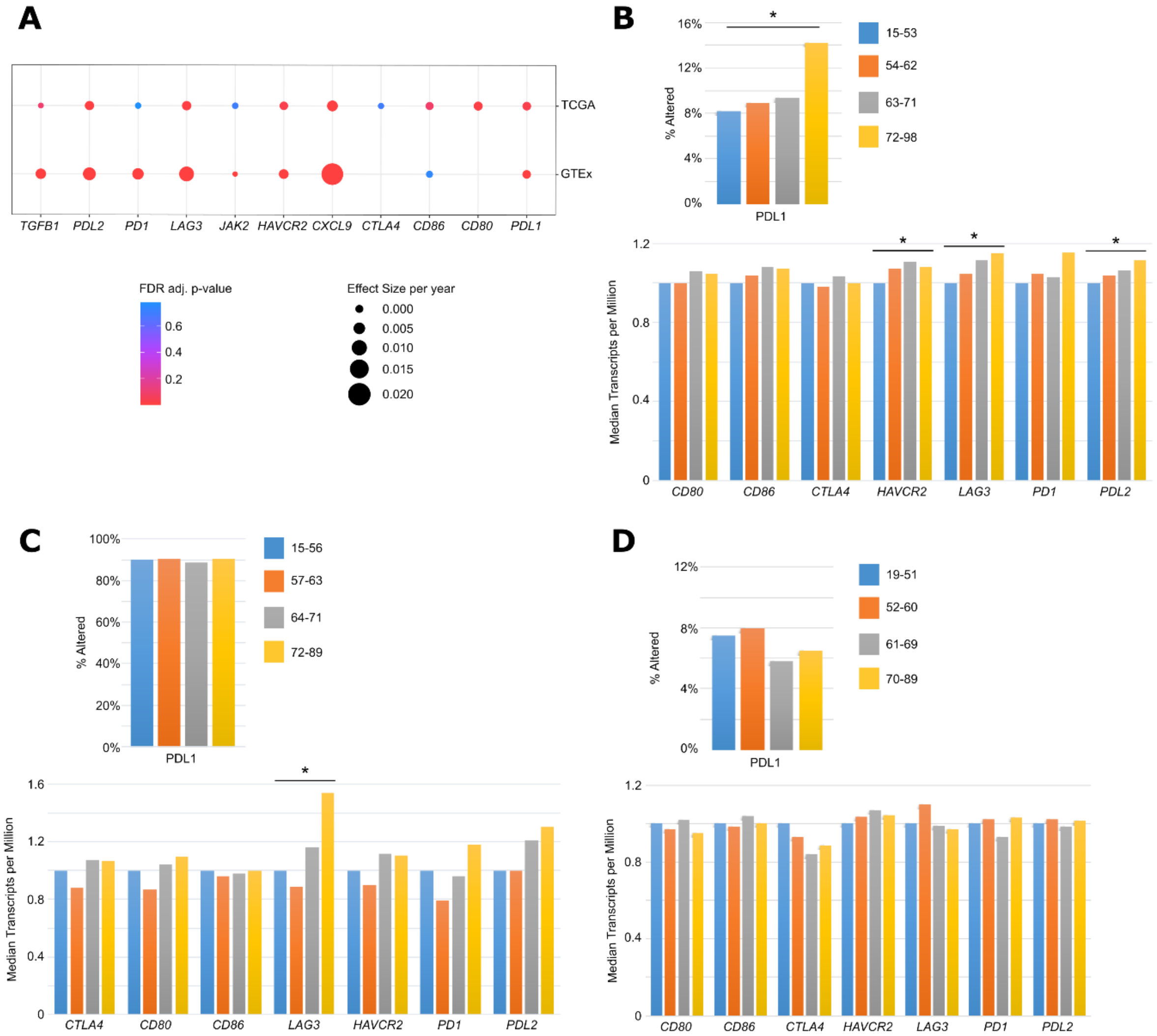

A
10.0
7.5
5.0
2.5
0.0
Log Tumor Mutational Burden
25
50
75
Diagnosis Age
B
20
15
10
5
0
-log10(p-value)
LGG
THCA
KIRC
STAD
KIRP
PRAD
LIHC
HNSC
THYM
PCPG
BRCA
CESC
ESCA
UCEC
OV
LUAD
TGCT
BLCA
SKCM
READ
COAD
PAAD
LUSC
GBM
Cancer Type
C
0.03
0.02
0.01
0.00
-0.01
-0.02
Coefficient Estimate(per year)
STAD
THYM
PRAD
LGG
THCA
KIRP
KIRC
ESCA
CESC
TGCT
HNSC
PCPG
LIHC
BLCA
OV
BRCA
SKCM
PAAD
COAD
GBM
LUSC
LUAD
READ
UCEC
Cancer Type
D
20
10
0
-log10(p-value)
Breast
Melanoma
Esophagogastric
Renal Cell
Head and Neck
Bladder
Non-Small Cell Lung
Colorectal
Cancer Type
E
0.015
0.01
0.005
0.00
Coefficient Estimate(per year)
Melanoma
Head and Neck
Esophagogastric
Bladder
Renal Cell
Breast
Non-Small Cell Lung
Colorectal
Cancer Type
F
6
4
2
0
Normalized Shannon Entropy
40
60
80
Diagnosis Age

Figure 4.1 - TMB generally increases and TCR diversity decreases with patient age at diagnosis

A Scatterplot of log2 tumor mutational burden by patient diagnosis age pan-cancer in TCGA data. The linear trend predicted by a multivariate linear model that includes cancer type as a covariate is shown. **B** Barplot of the negative log10 p-values for the age term of linear models fit for TMB in each TCGA cancer type study. **C** Barplot of the coefficient estimates per year for the age term of linear models fit for each TCGA cancer type study. Positive coefficients indicate increased mutational burden with increasing age. **D** Barplot of the negative log10 p-values for the age term of linear models fit for TMB within eight cancer types commonly treated with ICB therapies, data from GENIE. **E** Barplot of the coefficient estimates per year for the age term of linear models fit for TMB within eight cancer types commonly treated with ICB therapies in data from GENIE. **F** Scatterplot of normalized Shannon Entropy of TCR sequences by patient diagnosis age pan-cancer in TCGA data. The linear trend predicted by a multivariate linear model that includes cancer type as a covariate is shown.



<u>Figure 4.2 - Patient age at diagnosis correlates with increased expression of immune checkpoint genes in some cancer types</u>

A Dotplot of differential expression statistics for immune checkpoint therapy related genes with age. Compares results from pan-cancer TCGA samples and pan-tissue GTEx samples. **B** Caris Life Sciences colorectal cancer cohort PDL1 immunohistochemistry (top) and immune checkpoint gene expression data in median transcripts per million (bottom). * indicates a FDR adjusted p-value less than 0.05. **C**

Caris Life Sciences HPV-negative head and neck cancer cohort PDL1 immunohistochemistry (top) and immune checkpoint gene expression data in median transcripts per million (bottom). * indicates a FDR adjusted p-value less than 0.05. **D** Caris Life Sciences breast cancer cohort PDL1 immunohistochemistry (top) and immune checkpoint gene expression data in median transcripts per million (bottom).

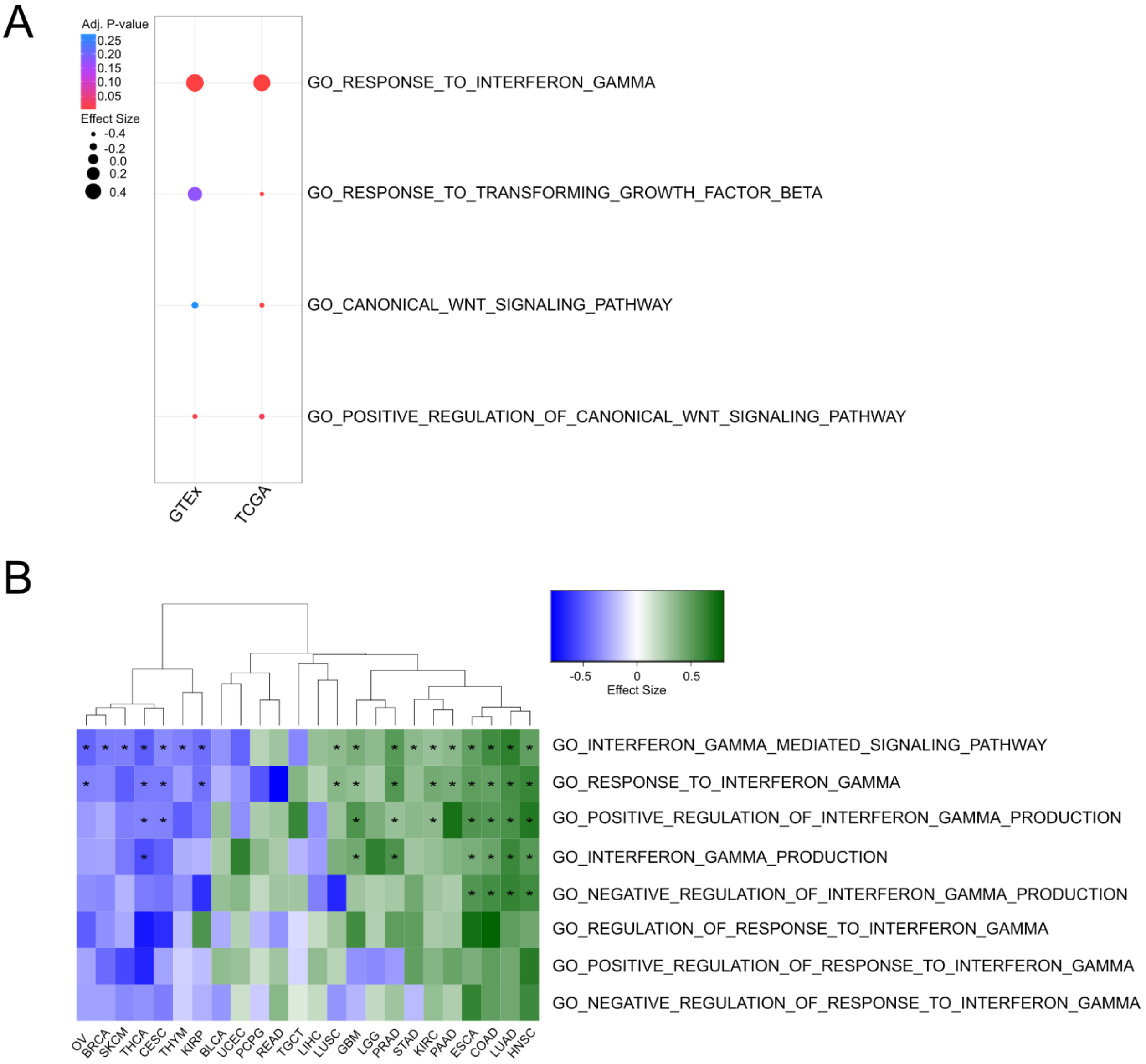


Figure 4.3 - Patient age associates with a more immune stimulatory signaling tumor microenvironment

A Dotplot of gene set enrichment results pan-cancer in TCGA and pan-tissue in GTEx for Interferon Gamma, TGFβ, and canonical WNT pathways. **B** Heatmap of estimated effect sizes for gene set enrichment across TCGA studies for all Interferon Gamma related GO terms. Positive values indicate increased enrichment with increasing age. * indicates a FDR adjusted p-value less than 0.05.

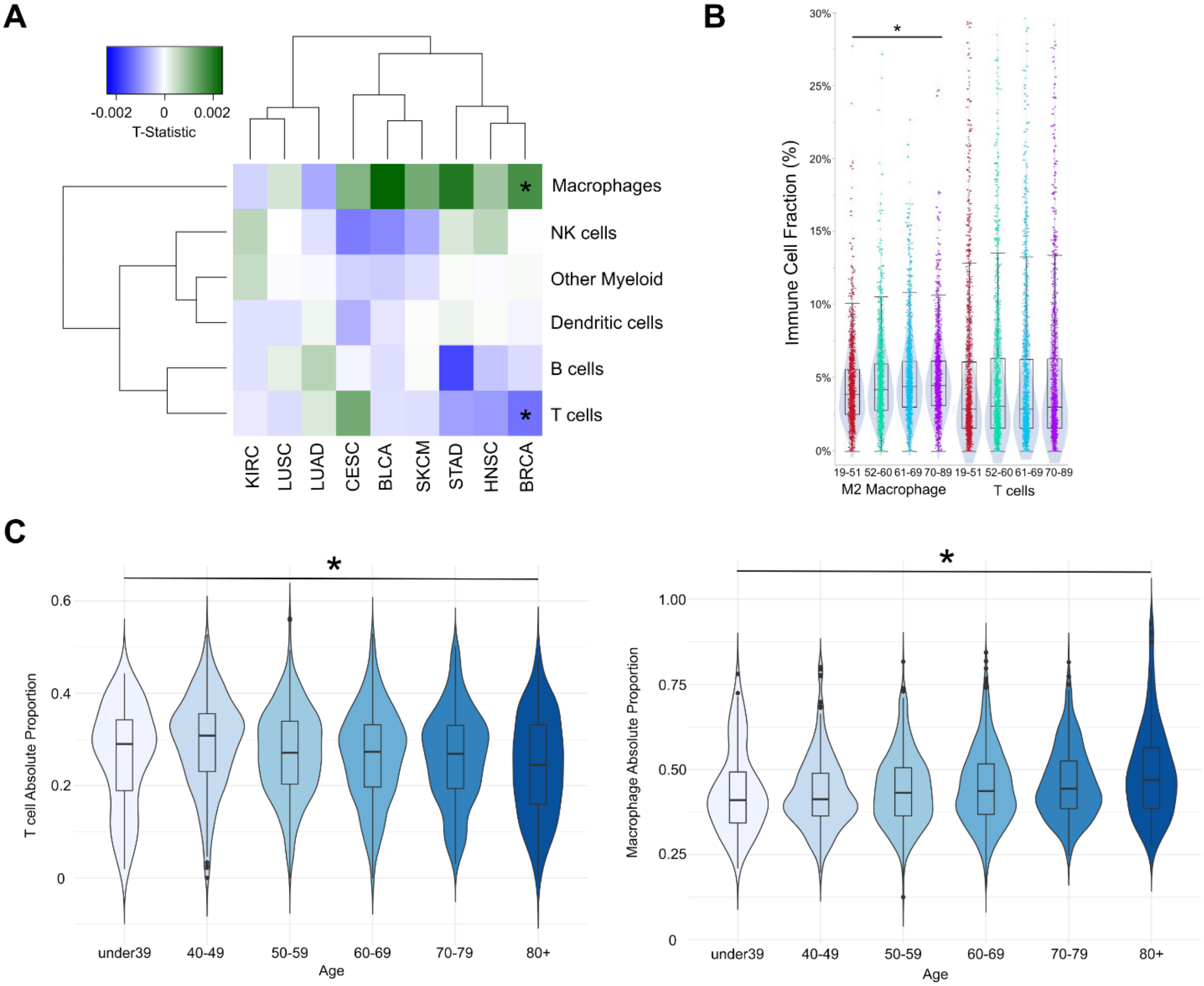


Figure 4.4 - Macrophage infiltration increases with age in patients with breast cancer

A Heatmap displaying the effect size coefficient estimates from linear models fit between immune cell type absolute proportion and patient age in each TCGA cancer type study. Green squares represent an increase in abundance of that immune cell type with increasing age, white represents no change, and blue a decrease. * Indicates a FDR-adjusted p-value < 0.05. **B** Violin plots from the Caris Life Sciences breast cancer cohort (n = 6462) corresponding to tumor infiltrating immune cell fraction across different age groups among M2 Macrophages and T cells. * Indicates a FDR-adjusted p-value < 0.05. **C** Violin plots from 1,818 patients with breast cancer from METABRIC, comparing T cell and macrophage absolute proportion across patient age groups. * Indicates a FDR-adjusted p-value < 0.05.

Tables

Gene	**LogFC (per year)**	**t-statistic**	**p-value**	**q-value**
CXCL9	0.007	3.945	0.001	0.003
PDL2	0.004	3.380	0.001	0.004
LAG3	0.004	3.187	0.001	0.007
CD80	0.004	3.002	0.003	0.012
HAVCR2	0.003	2.642	0.008	0.028
PDL1	0.003	2.612	0.009	0.030
CD86	0.002	2.131	0.033	0.083
TGFB1	-0.002	-2.070	0.039	0.094
CTLA4	-0.001	-0.575	0.566	0.693
JAK2	0.000	-0.563	0.574	0.700
PD1	-0.001	-0.421	0.674	0.779

Table 4.1 – Differential expression of immune checkpoint genes by age in TCGA

Differential expression results for immune checkpoint genes and immune checkpoint related genes pan-cancer in TCGA. The results are shown for the association with patient diagnosis age, including cancer type as a covariate. Note that LogFC is log fold change for each year of age.

CpG	**Gene**	**Estimate (per year)**	**t-statistic**	**p-value**	**q-value**	**CpG island?**
cg01107031	*TGFB1*	0.000218	3.326	0.000881	0.00303	yes
cg16883145	*TGFB1*	-2.73 x 10^{-5}	-0.644	0.519	0.662	yes
cg04387658	*CD86*	-0.000984	-7.101	1.31 x 10^{-12}	2.39 x 10^{-11}	yes
cg08460026	*CTLA4*	-0.000638	-3.888	0.000101	0.000433	no
cg17484237	*HAVCR2*	-0.000810	-5.845	5.19 x 10^{-9}	5.10 x 10^{-8}	no
cg21572897	*CD80*	-0.000526	-4.385	1.17 x 10^{-5}	6.01 x 10^{-5}	no
cg26956535	*LAG3*	-0.000114	-2.468	0.0135	0.0343	no
cg01820374	*LAG3*	-0.000410	-4.219	2.47 x 10^{-5}	0.000119	no
cg02823866	*CD274*	-2.08 x 10^{-5}	-1.512	0.130	0.232	yes
cg19724470	*CD274*	-0.000894	-6.701	2.17 x 10^{-11}	3.27 x 10^{-10}	no

Table 4.2 – Promoter methylation of ICB related genes by age in TCGA

Table of the linear relationships between methylation of all CpGs annotated to ICB related gene promoters to patient diagnosis age in TCGA data. Results are pan-cancer from a multivariate linear model that included cancer type as a covariate. Note that effect size estimates are per year of age at diagnosis.

	Estimate (per year)	**t-statistic**	**p-value**	**q-value**
T cells	-0.0006	-3.44233	0.000585	0.001754
Macrophages	0.001075	3.967909	7.42×10^{-5}	0.000445
B cells	-0.00028	-1.87737	0.060566	0.121131
NK cells	3.82×10^{-6}	0.055398	0.955826	0.955826
Dendritic cells	-0.00015	-1.49017	0.136286	0.204429
Misc. Myeloid	-4.97×10^{-5}	-0.50055	0.616722	0.740066

Table 4.3 – Immune cell type proportion by age in TCGA

Coefficients, statistics, p, and q-values for the diagnosis age term in the linear model fit for each immune cell type in TCGA data pan-cancer. Cancer type and sex were included as covariates for each of these models. Note that estimated coefficients are per additional year of age at diagnosis.

Chapter 5

Conclusions

This thesis explores the extent to which genomics data can be used to predict causal mechanism and future cellular genomics states. A major ideal of computational biology is being able to identify causality, mechanistic hypotheses, and putative therapeutic targets directly from the data. However, such efforts have often not delivered on these promises. I identify fundamental limitations in the application of gene network inference methods to single-cell RNA-seq data sets based on the observation that direct causally linked genes do not necessarily provide more predictive information about their targets than other, non-causal genes provide. I demonstrate this problem using multiple types of simulated single-cell RNA-seq data sets and the results conform strongly to observations regarding the inability of gene network inference methods to reliably distinguish direct mechanism from correlation in biological data sets.

While the predictive information contained in single-cell RNA-seq data does not allow for direct mechanism to be reliably inferred, there is considerable predictive information about the expression of other genes in single-cell RNA-seq data sets. I proposed methodology to predict future RNA expression levels within the same single-cell using metabolic labeling single-cell RNA-seq to generate two time points with which to train a neural network predictor. This method, RNAForecaster, is capable of accurately predicting future expression levels over short time periods, including in the case of perturbations unobserved in the training data and can do so without reliance on a lower dimensional embedding that would require the choice of somewhat arbitrary parameters. RNAForecaster provides a proof of principle for that future RNA expression states of cells, even those that are not observed in the input data, are estimable in the short term. While metabolic labeling single-cell RNA-seq is currently a niche type of data, more data types with temporal resolution seem likely to be necessary to more reliably answer the questions of mechanistic inference that this thesis has discussed, which in turn may increase the applicability of RNAForecaster.

Bibliography

Chapter 1

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086.

Akhmedov, M., Kedaigle, A., Chong, R.E., Montemanni, R., Bertoni, F., Fraenkel, E., and Kwee, I. (2017). PCSF: An R-package for network-based interpretation of high-throughput data. PLoS Comput. Biol. *13*, e1005694.

Arneson, D., Zhang, G., Ying, Z., Zhuang, Y., Byun, H.R., Ahn, I.S., Gomez-Pinilla, F., and Yang, X. (2018). Single cell molecular alterations reveal target cells and pathways of concussive brain injury. Nat. Commun. *9*, 3894.

Baghban, R., Roshangar, L., Jahanban-Esfahlan, R., Seidi, K., Ebrahimi-Kalan, A., Jaymand, M., Kolahian, S., Javaheri, T., and Zare, P. (2020). Tumor microenvironment complexity and therapeutic implications at a glance. Cell Commun. Signal. *18*, 59.

Basson, M.A. (2012). Signaling in cell differentiation and morphogenesis. Cold Spring Harb. Perspect. Biol. *4*.

Bayat Mokhtari, R., Homayouni, T.S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., and Yeger, H. (2017). Combination therapy in combating cancer. Oncotarget *8*, 38022–38043.

Belyaeva, A., Squires, C., and Uhler, C. (2020). DCI: Learning Causal Differences between Gene Regulatory Networks. BioRxiv.

Bhan, A., and Mandal, S.S. (2014). Long noncoding RNAs: emerging stars in gene regulation, epigenetics and human disease. ChemMedChem *9*, 1932–1956.

Browaeys, R., Saelens, W., and Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. Nat. Methods *17*, 159–162.

Brunk, E., Sahoo, S., Zielinski, D.C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G.A., Aurich, M.K., et al. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat. Biotechnol. *36*, 272–281.

Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., and Collins, J.J. (2018). Next-Generation Machine Learning for Biological Networks. Cell *173*, 1581–1592.

Cang, Z., and Nie, Q. (2020). Inferring spatial and signaling relationships between cells from single cell transcriptomic data. Nat. Commun. *11*, 2084.

Cech, T.R., and Steitz, J.A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. Cell *157*, 77–94.

Chan, T.E., Stumpf, M.P.H., and Babtie, A.C. (2017). Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. Cell Syst. *5*, 251-267.e3.

Chen, S., and Mar, J.C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinformatics *19*, 232.

Cherry, C., Maestas, D.R., Han, J., Andorko, J.I., Cahan, P., Fertig, E.J., Garmire, L.X., and Elisseeff, J.H. (2021). Computational reconstruction of the signalling networks surrounding implanted biomaterials from single-cell transcriptomics. Nat. Biomed. Eng.

Choi, H., Sheng, J., Gao, D., Li, F., Durrans, A., Ryu, S., Lee, S.B., Narula, N., Rafii, S., Elemento, O., et al. (2015). Transcriptome analysis of individual stromal cell populations identifies stroma-tumor crosstalk in mouse lung cancer model. Cell Rep. *10*, 1187–1201.

Davis-Marcisak, E.F., Deshpande, A., Stein-O'Brien, G.L., Ho, W.J., Laheru, D., Jaffee, E.M., Fertig, E.J., and Kagohara, L.T. (2021). From bench to bedside: Single-cell analysis for cancer immunotherapy. Cancer Cell *39*, 1062–1080.

Deshpande, A., Chu, L.-F., Stewart, R., and Gitter, A. (2019). Network Inference with Granger Causality Ensembles on Single-Cell Transcriptomic Data. BioRxiv.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell *167*, 1853-1866.e17.

Dugourd, A., Kuppe, C., Sciacovelli, M., Gjerga, E., Gabor, A., Emdal, K.B., Vieira, V., Bekker-Jensen, D.B., Kranz, J., Bindels, E.M.J., et al. (2021). Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. Mol. Syst. Biol. *17*, e9730.

Efremova, M., Vento-Tormo, M., Teichmann, S.A., and Vento-Tormo, R. (2020). CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat. Protoc. *15*, 1484–1506.

Elyanow, R., Dumitrascu, B., Engelhardt, B.E., and Raphael, B.J. (2020). netNMF-sc: leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis. Genome Res. *30*, 195–204.

Fernandez, D.M., Rahman, A.H., Fernandez, N.F., Chudnovskiy, A., Amir, E.-A.D., Amadori, L., Khan, N.S., Wong, C.K., Shamailova, R., Hill, C.A., et al. (2019). Single-cell immune landscape of human atherosclerotic plaques. Nat. Med. *25*, 1576–1588.

Fertig, E.J., Favorov, A.V., and Ochs, M.F. (2013). Identifying context-specific transcription factor targets from prior knowledge and gene expression data. IEEE Trans. Nanobioscience *12*, 142–149.

Gosline, S.J.C., Spencer, S.J., Ursu, O., and Fraenkel, E. (2012). SAMNet: a network-based approach to integrate multi-dimensional high throughput datasets. Integr Biol (Camb) *4*, 1415–1427.

Greenfield, A., Madar, A., Ostrer, H., and Bonneau, R. (2010). DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models. PLoS ONE *5*, e13397.

Harmston, N., and Lenhard, B. (2013). Chromatin and epigenetic features of long-range gene regulation. Nucleic Acids Res. *41*, 7185–7199.

Hill, S.M., Heiser, L.M., Cokelaer, T., Unger, M., Nesser, N.K., Carlin, D.E., Zhang, Y., Sokolov, A., Paull, E.O., Wong, C.K., et al. (2016). Inferring causal molecular networks: empirical assessment through a community-based effort. Nat. Methods *13*, 310–318.

Hou, R., Denisenko, E., Ong, H.T., Ramilowski, J.A., and Forrest, A.R.R. (2020). Predicting cell-to-cell communication networks using NATMI. Nat. Commun. *11*, 5011.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS ONE *5*.

Kirouac, D.C., Ito, C., Csaszar, E., Roch, A., Yu, M., Sykes, E.A., Bader, G.D., and Zandstra, P.W. (2010). Dynamic interaction networks in a hierarchically organized tissue. Mol. Syst. Biol. *6*, 417.

Kivela, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., and Porter, M.A. (2014). Multilayer networks. J. Complex Netw. *2*, 203–271.

Kuijjer, M.L., Fagny, M., Marin, A., Quackenbush, J., and Glass, K. (2020). PUMA: PANDA using microrna associations. Bioinformatics *36*, 4765–4773.

Kumar, M.P., Du, J., Lagoudas, G., Jiao, Y., Sawyer, A., Drummond, D.C., Lauffenburger, D.A., and Raue, A. (2018). Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. Cell Rep. *25*, 1458-1468.e4.

Lê Cao, K.-A., Abadi, A.J., Davis-Marcisak, E.F., Hsu, L., Arora, A., Coullomb, A., Deshpande, A., Feng, Y., Jeganathan, P., Loth, M., et al. (2021). Community-wide hackathons to identify central themes in single-cell multi-omics. Genome Biol. *22*, 220.

Liu, A., Trairatphisan, P., Gjerga, E., Didangelos, A., Barratt, J., and Saez-Rodriguez, J. (2019). From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. NPJ Syst. Biol. Appl. *5*, 40.

Liu, X., Maiorino, E., Halu, A., Glass, K., Prasad, R.B., Loscalzo, J., Gao, J., and Sharma, A. (2020). Robustness and lethality in multilayer biological molecular networks. Nat. Commun. *11*, 6043.

Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., DREAM5 Consortium, Kellis, M., Collins, J.J., et al. (2012). Wisdom of crowds for robust gene network inference. Nat. Methods *9*, 796–804.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics *7 Suppl 1*, S7.

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M.S.H., Ko, S.B.H., Gouda, N., Hayashi, T., and Nikaido, I. (2017). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics *33*, 2314–2321.

Mercatelli, D., Scalambra, L., Triboli, L., Ray, F., and Giorgi, F.M. (2020). Gene regulatory network inference resources: A practical overview. Biochim. Biophys. Acta Gene Regul. Mech. *1863*, 194430.

Nguyen, H., Tran, D., Tran, B., Pehlivan, B., and Nguyen, T. (2021). A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. Brief. Bioinformatics *22*.

Osorio, D., Zhong, Y., Li, G., Huang, J.Z., and Cai, J.J. (2020). scTenifoldNet: A Machine Learning Workflow for Constructing and Comparing Transcriptome-wide Gene Regulatory Networks from Single-Cell Data. Patterns (N Y) *1*, 100139.

Osorio, D., Zhong, Y., Li, G., Xu, Q., Hillhouse, A.E., Chen, J., Davidson, L.A., Tian, Y., Chapkin, R.S., Huang, J.Z., et al. (2021). scTenifoldKnk: a machine learning workflow performing virtual knockout experiments on single-cell gene regulatory networks. BioRxiv.

Otasek, D., Morris, J.H., Bouças, J., Pico, A.R., and Demchak, B. (2019). Cytoscape Automation: empowering workflow-based network analysis. Genome Biol. *20*, 185.

Page, L., Brin, S., Motwani, R., and Winoigrad, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web.

Papili Gao, N., Ud-Dean, S.M.M., Gandrillon, O., and Gunawan, R. (2018). SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. Bioinformatics *34*, 258–266.

Pearl, J. (1995). Causal diagrams for empirical research. Biometrika *82*, 669–688.

Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., and Murali, T.M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat. Methods *17*, 147–154.

Qiu, X., Rahimzamani, A., Wang, L., Ren, B., Mao, Q., Durham, T., McFaline-Figueroa, J.L., Saunders, L., Trapnell, C., and Kannan, S. (2020). Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. Cell Syst. *10*, 265-274.e11.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554.

Schaffer, L.V., and Ideker, T. (2021). Mapping the multiscale structure of biological systems. Cell Syst. *12*, 622–635.

Schaffter, T., Marbach, D., and Floreano, D. (2011). GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics *27*, 2263–2270.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Sonawane, A.R., Weiss, S.T., Glass, K., and Sharma, A. (2019). Network medicine in the age of biomedical big data. Front. Genet. *10*, 294.

Stoeger, T., Gerlach, M., Morimoto, R.I., and Nunes Amaral, L.A. (2018). Large-scale investigation of the reasons why potentially important genes are ignored. PLoS Biol. *16*, e2006643.

Stone, M., McCalla, S.G., Fotuhi Siahpirani, A., Periyasamy, V., Shin, J., and Roy, S. (2021). Identifying strengths and weaknesses of methods for computational network inference from single cell RNA-seq data. BioRxiv.

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science *302*, 249–255.

Su, A.I., and Hogenesch, J.B. (2007). Power-law-like distributions in biomedical publications and research funding. Genome Biol. *8*, 404.

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. Nat. Biotechnol. *32*, 381–386.

Tunnacliffe, E., and Chubb, J.R. (2020). What is a transcriptional burst? Trends Genet. *36*, 288–297.

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat. Methods *13*, 966–967.

Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J.M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics *26*, i237-45.

Wang, L., Yu, P., Zhou, B., Song, J., Li, Z., Zhang, M., Guo, G., Wang, Y., Chen, X., Han, L., et al. (2020). Single-cell reconstruction of the adult human heart during heart failure and recovery reveals the cellular landscape underlying cardiac function. Nat. Cell Biol. *22*, 108–119.

Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., Tang, Q., Meyer, C.A., Zhang, Y., and Liu, X.S. (2013). Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nat. Protoc. *8*, 2502–2515.

Wang, S., Karikomi, M., MacLean, A.L., and Nie, Q. (2019). Cell lineage and communication network inference via optimization for single-cell transcriptomics. Nucleic Acids Res. *47*, e66.

Weighill, D.A., Ben Guebila, M., Glass, K., Quackenbush, J., and Platig, J. (2021). Predicting genotype-specific gene regulatory networks. BioRxiv.

Ximerakis, M., Lipnick, S.L., Innes, B.T., Simmons, S.K., Adiconis, X., Dionne, D., Mayweather, B.A., Nguyen, L., Niziolek, Z., Ozek, C., et al. (2019). Single-cell transcriptomic profiling of the aging mouse brain. Nat. Neurosci. *22*, 1696–1708.

Yao, R.-W., Wang, Y., and Chen, L.-L. (2019). Cellular functions of long noncoding RNAs. Nat. Cell Biol. *21*, 542–551.

Yuan, B., Shen, C., Luna, A., Korkut, A., Marks, D.S., Ingraham, J., and Sander, C. (2021). CellBox: Interpretable Machine Learning for Perturbation Biology with Application to the Design of Cancer Combination Therapy. Cell Syst. *12*, 128-140.e4.

Zhou, S., Huang, Y.-E., Liu, H., Zhou, X., Yuan, M., Hou, F., Wang, L., and Jiang, W. (2021). Single-cell RNA-seq dissects the intratumoral heterogeneity of triple-negative breast cancer based on gene regulatory networks. Mol. Ther. Nucleic Acids *23*, 682–690.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*, 481–487.

Chapter 2

Bastidas-Ponce, Aimée, Sophie Tritschler, Leander Dony, Katharina Scheibner, Marta Tarquis-Medina, Ciro Salinno, Silvia Schirge, et al. 2019. "Comprehensive Single Cell MRNA Profiling Reveals a Detailed Roadmap for Pancreatic Endocrinogenesis." *Development* 146 (12). https://doi.org/10.1242/dev.173849.

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

Bergen, Volker, Marius Lange, Stefan Peidli, F Alexander Wolf, and Fabian J Theis. 2020. "Generalizing RNA Velocity to Transient Cell States through Dynamical Modeling." *Nature Biotechnology* 38 (12): 1408–14. https://doi.org/10.1038/s41587-020-0591-3.

Bunne, Charlotte, Laetitia Papaxanthos, Andreas Krause, and Marco Cuturi. 2022. "Proximal Optimal Transport Modeling of Population Dynamics," May.

Campbell, Kieran R, and Christopher Yau. 2018. "Uncovering Pseudotemporal Trajectories with Covariates from Single Cell and Bulk Expression Data." *Nature Communications* 9 (1): 2442. https://doi.org/10.1038/s41467-018-04696-6.

Cao, Junyue, Wei Zhou, Frank Steemers, Cole Trapnell, and Jay Shendure. 2020. "Sci-Fate Characterizes the Dynamics of Gene Expression in Single Cells." *Nature Biotechnology* 38 (8): 980–88. https://doi.org/10.1038/s41587-020-0480-9.

Cawte, Adam D, Peter J Unrau, and David S Rueda. 2020. "Live Cell Imaging of Single RNA Molecules with Fluorogenic Mango II Arrays." *Nature Communications* 11 (1): 1283. https://doi.org/10.1038/s41467-020-14932-7.

Chen, Huidong, Luca Albergante, Jonathan Y Hsu, Caleb A Lareau, Giosuè Lo Bosco, Jihong Guan, Shuigeng Zhou, et al. 2019. "Single-Cell Trajectories Reconstruction, Exploration and Mapping of Omics Data with STREAM." *Nature Communications* 10 (1): 1903. https://doi.org/10.1038/s41467-019-09670-4.

Chen, Tian Qi, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. 2018. "Neural Ordinary Differential Equations." *ArXiv*, June.

Chen, Zhanlin, William C King, Aheyon Hwang, Mark Gerstein, and Jing Zhang. 2022. “DeepVelo: Single-Cell Transcriptomic Deep Velocity Field Learning with Neural Ordinary Differential Equations.” *Science Advances* 8 (48): eabq3745. https://doi.org/10.1126/sciadv.abq3745.

Dixit, Atray, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, et al. 2016. “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens.” *Cell* 167 (7): 1853-1866.e17. https://doi.org/10.1016/j.cell.2016.11.038.

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

Fertig, Elana J, Elizabeth M Jaffee, Paul Macklin, Vered Stearns, and Chenguang Wang. 2021. “Forecasting Cancer: From Precision to Predictive Medicine.” *Med (New York, N.Y.)* 2 (9): 1004–10. https://doi.org/10.1016/j.medj.2021.08.007.

Fertig, Elana J., John Harlim, and Brian R. Hunt. 2007. “A Comparative Study of 4D-VAR and a 4D Ensemble Kalman Filter: Perfect Model Simulations with Lorenz-96.” *Tellus A: Dynamic Meteorology and Oceanography* 59 (1): 96–100. https://doi.org/10.1111/j.1600-0870.2006.00205.x.

French, R M. 1999. “Catastrophic Forgetting in Connectionist Networks.” *Trends in Cognitive Sciences* 3 (4): 128–35. https://doi.org/10.1016/s1364-6613(99)01294-2.

Gorin, Gennady, Meichen Fang, Tara Chari, and Lior Pachter. 2022. “RNA Velocity Unraveled.” *BioRxiv*, February. https://doi.org/10.1101/2022.02.12.480214.

Gorin, Gennady, Valentine Svensson, and Lior Pachter. 2020. “Protein Velocity and Acceleration from Single-Cell Multiomics Experiments.” *Genome Biology* 21 (1): 39. https://doi.org/10.1186/s13059-020-1945-3.

Hendriks, Gert-Jan, Lisa A Jung, Anton J M Larsson, Michael Lidschreiber, Oscar Andersson Forsman, Katja Lidschreiber, Patrick Cramer, and Rickard Sandberg. 2019. “NASC-Seq Monitors RNA Synthesis in Single Cells.” *Nature Communications* 10 (1): 3138. https://doi.org/10.1038/s41467-019-11028-9.

Kostelich, Eric J, Yang Kuang, Joshua M McDaniel, Nina Z Moore, Nikolay L Martirosyan, and Mark C Preul. 2011. “Accurate State Estimation from Uncertain Data and Models: An Application of Data Assimilation to Mathematical Models of Human Brain Tumors.” *Biology Direct* 6 (December): 64. https://doi.org/10.1186/1745-6150-6-64.

La Manno, Gioele, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, et al. 2018. “RNA Velocity of Single Cells.” *Nature* 560 (7719): 494–98. https://doi.org/10.1038/s41586-018-0414-6.

Lotfollahi, Mohammad, F Alexander Wolf, and Fabian J Theis. 2019. “ScGen Predicts Single-Cell Perturbation Responses.” *Nature Methods* 16 (8): 715–21. https://doi.org/10.1038/s41592-019-0494-8.

Pratapa, Aditya, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and T M Murali. 2020. “Benchmarking Algorithms for Gene Regulatory Network Inference from Single-Cell Transcriptomic Data.” *Nature Methods* 17 (2): 147–54. https://doi.org/10.1038/s41592-019-0690-6.

Qiu, Qi, Peng Hu, Xiaojie Qiu, Kiya W Govek, Pablo G Cámara, and Hao Wu. 2020. “Massively Parallel and Time-Resolved RNA Sequencing in Single Cells with ScNT-Seq.” *Nature Methods* 17 (10): 991–1001. https://doi.org/10.1038/s41592-020-0935-4.

Qiu, Xiaojie, Yan Zhang, Jorge D Martin-Rufino, Chen Weng, Shayan Hosseinzadeh, Dian Yang, Angela N Pogson, et al. 2022. “Mapping Transcriptomic Vector Fields of Single Cells.” *Cell* 185 (4): 690-711.e45. https://doi.org/10.1016/j.cell.2021.12.045.

Reid, John E, and Lorenz Wernisch. 2016. “Pseudotime Estimation: Deconfounding Single Cell Time Series.” *Bioinformatics* 32 (19): 2973–80. https://doi.org/10.1093/bioinformatics/btw372.

Saelens, Wouter, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. 2019. “A Comparison of Single-Cell Trajectory Inference Methods.” *Nature Biotechnology* 37 (5): 547–54. https://doi.org/10.1038/s41587-019-0071-9.

Schiebinger, Geoffrey, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, et al. 2019. “Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming.” *Cell* 176 (4): 928-943.e22. https://doi.org/10.1016/j.cell.2019.01.006.

Stein-O’Brien, Genevieve L, Michaela C Ainsile, and Elana J Fertig. 2021. “Forecasting Cellular States: From Descriptive to Predictive Biology via Single-Cell Multiomics.” *Current Opinion in Systems Biology* 26 (June): 24–32. https://doi.org/10.1016/j.coisb.2021.03.008.

Tong, Alexander, Jessie Huang, Guy Wolf, David van Dijk, and Smita Krishnaswamy. 2020. “Trajectorynet: A Dynamic Optimal Transport Network for Modeling Cellular Dynamics.” *Proceedings of Machine Learning Research* 119 (July): 9526–36. https://doi.org/10.48550/arxiv.2002.04461.

Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. 2014. “The Dynamics and Regulators of Cell Fate Decisions Are Revealed by

Pseudotemporal Ordering of Single Cells.” *Nature Biotechnology* 32 (4): 381–86. https://doi.org/10.1038/nbt.2859.

Tunnacliffe, Edward, and Jonathan R Chubb. 2020. “What Is a Transcriptional Burst?” *Trends in Genetics* 36 (4): 288–97. https://doi.org/10.1016/j.tig.2020.01.003.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” *Advances in Neural Information Processing Systems*.

Wang, Danyang, Ana Shalamberidze, A Emilia Arguello, Byron W Purse, and Ralph E Kleiner. 2022. “Live-Cell RNA Imaging with Metabolically Incorporated Fluorescent Nucleosides.” *Journal of the American Chemical Society* 144 (32): 14647–56. https://doi.org/10.1021/jacs.2c04142.

Yu, Hengshi, and Joshua D Welch. 2022. “PerturbNet Predicts Single-Cell Responses to Unseen Chemical and Genetic Perturbations.” *BioRxiv*, July. https://doi.org/10.1101/2022.07.20.500854.

Zheng, Shijie C., Genevieve Stein-O’Brien, Jonathan J. Augustin, Jared Slosberg, Giovanni A. Carosso, Briana Winer, Gloria Shin, Hans T. Bjornsson, Loyal A. Goff, and Kasper D. Hansen. 2021. “Universal Prediction of Cell Cycle Position Using Transfer Learning.” *BioRxiv*, April. https://doi.org/10.1101/2021.04.06.438463.

Chapter 3

Alpert, Tara, Lydia Herzel, and Karla M Neugebauer. 2017. “Perfect Timing: Splicing and Transcription Rates in Living Cells.” *Wiley Interdisciplinary Reviews. RNA* 8 (2). https://doi.org/10.1002/wrna.1401.

Cambridge, Sidney B, Florian Gnad, Chuong Nguyen, Justo Lorenzo Bermejo, Marcus Krüger, and Matthias Mann. 2011. “Systems-Wide Proteomic Analysis in Mammalian Cells Reveals Conserved, Functional Protein Turnover.” *Journal of Proteome Research* 10 (12): 5275–84. https://doi.org/10.1021/pr101183k.

Chan, Thalia E, Michael P H Stumpf, and Ann C Babtie. 2017. “Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures.” *Cell Systems* 5 (3): 251-267.e3. https://doi.org/10.1016/j.cels.2017.08.014.

Chen, Shuonan, and Jessica C Mar. 2018. “Evaluating Methods of Inferring Gene Regulatory Networks Highlights Their Lack of Performance for Single Cell Gene Expression Data.” *BMC Bioinformatics* 19 (1): 232. https://doi.org/10.1186/s12859-018-2217-z.

Das, Barnali, and Pralay Mitra. 2021. “High-Performance Whole-Cell Simulation Exploiting Modular Cell Biology Principles.” *Journal of Chemical Information and Modeling* 61 (3): 1481–92. https://doi.org/10.1021/acs.jcim.0c01282.

Deshpande, Atul, Li-Fang Chu, Ron Stewart, and Anthony Gitter. 2019. “Network Inference with Granger Causality Ensembles on Single-Cell Transcriptomic Data.” *BioRxiv*, January. https://doi.org/10.1101/534834.

Erbe, Rossin, Jessica Gore, Kelly Gemmill, Daria A Gaykalova, and Elana J Fertig. 2022. “The Use of Machine Learning to Discover Regulatory Networks Controlling Biological Systems.” *Molecular Cell* 82 (2): 260–73. https://doi.org/10.1016/j.molcel.2021.12.011.

Fertig, Elana J, Ludmila V Danilova, Alexander V Favorov, and Michael F Ochs. 2011. “Hybrid Modeling of Cell Signaling and Transcriptional Reprogramming and Its Application in C. Elegans Development.” *Frontiers in Genetics* 2 (November): 77. https://doi.org/10.3389/fgene.2011.00077.

Gorin, Gennady, John J Vastola, Meichen Fang, and Lior Pachter. 2022. “Interpretable and Tractable Models of Transcriptional Noise for the Rational Design of Single-Molecule Quantification Experiments.” *Nature Communications* 13 (1): 7620. https://doi.org/10.1038/s41467-022-34857-7.

Gry, Marcus, Rebecca Rimini, Sara Strömberg, Anna Asplund, Fredrik Pontén, Mathias Uhlén, and Peter Nilsson. 2009. “Correlations between RNA and Protein Expression Profiles in 23 Human Cell Lines.” *BMC Genomics* 10 (August): 365. https://doi.org/10.1186/1471-2164-10-365.

Herbach, Ulysse, Arnaud Bonnaffoux, Thibault Espinasse, and Olivier Gandrillon. 2017. “Inferring Gene Regulatory Networks from Single-Cell Data: A Mechanistic Approach.” *BMC Systems Biology* 11 (1): 105. https://doi.org/10.1186/s12918-017-0487-0.

Hill, Steven M, Laura M Heiser, Thomas Cokelaer, Michael Unger, Nicole K Nesser, Daniel E Carlin, Yang Zhang, et al. 2016. “Inferring Causal Molecular Networks: Empirical Assessment through a Community-Based Effort.” *Nature Methods* 13 (4): 310–18. https://doi.org/10.1038/nmeth.3773.

Huynh-Thu, Vân Anh, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. 2010. “Inferring Regulatory Networks from Expression Data Using Tree-Based Methods.” *Plos One* 5 (9). https://doi.org/10.1371/journal.pone.0012776.

Margolin, Adam A, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. 2006. “ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context.” *BMC Bioinformatics* 7 Suppl 1 (March): S7. https://doi.org/10.1186/1471-2105-7-S1-S7.

Matsumoto, Hirotaka, Hisanori Kiryu, Chikara Furusawa, Minoru S H Ko, Shigeru B H Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. 2017. “SCODE: An Efficient Regulatory Network Inference Algorithm from Single-Cell RNA-Seq during Differentiation.” *Bioinformatics* 33 (15): 2314–21. https://doi.org/10.1093/bioinformatics/btx194.

Osorio, Daniel, Yan Zhong, Guanxun Li, Jianhua Z Huang, and James J Cai. 2020. “ScTenifoldNet: A Machine Learning Workflow for Constructing and Comparing Transcriptome-Wide Gene Regulatory Networks from Single-Cell Data.” *Patterns (New York, N.Y.)* 1 (9): 100139. https://doi.org/10.1016/j.patter.2020.100139.

Papili Gao, Nan, S M Minhaz Ud-Dean, Olivier Gandrillon, and Rudiyanto Gunawan. 2018. “SINCERITIES: Inferring Gene Regulatory Networks from Time-Stamped Single Cell Transcriptional Expression Profiles.” *Bioinformatics* 34 (2): 258–66. https://doi.org/10.1093/bioinformatics/btx575.

Pratapa, Aditya, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and T M Murali. 2020. “Benchmarking Algorithms for Gene Regulatory Network Inference from Single-Cell Transcriptomic Data.” *Nature Methods* 17 (2): 147–54. https://doi.org/10.1038/s41592-019-0690-6.

Qiu, Xiaojie, Arman Rahimzamani, Li Wang, Bingcheng Ren, Qi Mao, Timothy Durham, José L McFaline-Figueroa, Lauren Saunders, Cole Trapnell, and Sreeram Kannan. 2020. “Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe.” *Cell Systems* 10 (3): 265-274.e11. https://doi.org/10.1016/j.cels.2020.02.003.

Schwanhäusser, Björn, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. 2011. “Global Quantification of Mammalian Gene Expression Control.” *Nature* 473 (7347): 337–42. https://doi.org/10.1038/nature10098.

Sousa Abreu, Raquel de, Luiz O Penalva, Edward M Marcotte, and Christine Vogel. 2009. “Global Signatures of Protein and MRNA Expression Levels.” *Molecular Biosystems* 5 (12): 1512–26. https://doi.org/10.1039/b908315d.

Stone, Matthew, Sunnie Grace McCalla, Alireza Fotuhi Siahpirani, Viswesh Periyasamy, Junha Shin, and Sushmita Roy. 2021. “Identifying Strengths and Weaknesses of Methods for Computational Network Inference from Single Cell RNA-Seq Data.” *BioRxiv*, June. https://doi.org/10.1101/2021.06.01.446671.

Thornburg, Zane R, David M Bianchi, Troy A Brier, Benjamin R Gilbert, Tyler M Earnest, Marcelo C R Melo, Nataliya Safronova, et al. 2022. “Fundamental Behaviors Emerge from Simulations of a Living Minimal Cell.” *Cell* 185 (2): 345-360.e28. https://doi.org/10.1016/j.cell.2021.12.025.

Vogel, Christine, and Edward M Marcotte. 2012. “Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses.” *Nature Reviews. Genetics* 13 (4): 227–32. https://doi.org/10.1038/nrg3185.

Chapter 4

Asgarova, A., Asgarov, K., Godet, Y., Peixoto, P., Nadaradjane, A., Boyer-Guittaut, M., Galaine, J., Guenat, D., Mougey, V., Perrard, J., et al. (2018). PD-L1 expression is regulated by both DNA methylation and NF-kB during EMT signaling in non-small cell lung carcinoma. Oncoimmunology *7*, e1423170.

Aspinall, R., and Andrew, D. (2000). Thymic involution in aging. J. Clin. Immunol. *20*, 250–256.

Aw, D., Silva, A.B., and Palmer, D.B. (2007). Immunosenescence: emerging challenges for an ageing population. Immunology *120*, 435–446.

Bolotin, D.A., Shugay, M., Mamedov, I.Z., Putintseva, E.V., Turchaninova, M.A., Zvyagin, I.V., Britanova, O.V., and Chudakov, D.M. (2013). MiTCR: software for T-cell receptor sequencing data analysis. Nat. Methods *10*, 813–814.

Braun, D.A., Hou, Y., Bakouny, Z., Ficial, M., Sant’ Angelo, M., Forman, J., Ross-Macdonald, P., Berger, A.C., Jegede, O.A., Elagina, L., et al. (2020). Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma. Nat. Med. *26*, 909–918.

Britanova, O.V., Putintseva, E.V., Shugay, M., Merzlyak, E.M., Turchaninova, M.A., Staroverov, D.B., Bolotin, D.A., Lukyanov, S., Bogdanova, E.A., Mamedov, I.Z., et al. (2014). Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. J. Immunol. *192*, 2689–2698.

Chalmers, Z.R., Connelly, C.F., Fabrizio, D., Gay, L., Ali, S.M., Ennis, R., Schrock, A., Campbell, B., Shlien, A., Chmielecki, J., et al. (2017). Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Med. *9*, 34.

Chatsirisupachai, K., Lesluyes, T., Paraoan, L., Van Loo, P., and de Magalhães, J.P. (2021). An integrative analysis of the age-associated multi-omic landscape across cancers. Nat. Commun. *12*, 2345.

Conway, J.R., Kofman, E., Mo, S.S., Elmarakeby, H., and Van Allen, E. (2018). Genomics of response to immune checkpoint therapies for cancer: implications for precision medicine. Genome Med. *10*, 93.

Cristescu, R., Mogg, R., Ayers, M., Albright, A., Murphy, E., Yearley, J., Sher, X., Liu, X.Q., Lu, H., Nebozhyn, M., et al. (2018). Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. Science *362*.

Daste, A., Domblides, C., Gross-Goupil, M., Chakiba, C., Quivy, A., Cochin, V., de Mones, E., Larmonier, N., Soubeyran, P., and Ravaud, A. (2017). Immune checkpoint inhibitors and elderly people: A review. Eur. J. Cancer *82*, 155–166.

Easwaran, H., and Baylin, S.B. (2019). Origin and mechanisms of DNA methylation dynamics in cancers. In The DNA, RNA, and Histone Methylomes, S. Jurga, and J. Barciszewski, eds. (Cham: Springer International Publishing), pp. 27–52.

Easwaran, H., Johnstone, S.E., Van Neste, L., Ohm, J., Mosbruger, T., Wang, Q., Aryee, M.J., Joyce, P., Ahuja, N., Weisenberger, D., et al. (2012). A DNA hypermethylation module for the stem/progenitor cell signature of cancer. Genome Res. *22*, 837–849.

Egorov, E.S., Kasatskaya, S.A., Zubov, V.N., Izraelson, M., Nakonechnaya, T.O., Staroverov, D.B., Angius, A., Cucca, F., Mamedov, I.Z., Rosati, E., et al. (2018). The changing landscape of naive T cell receptor repertoire with human aging. Front. Immunol. *9*, 1618.

Elias, R., Giobbie-Hurder, A., McCleary, N.J., Ott, P., Hodi, F.S., and Rahma, O. (2018). Efficacy of PD-1 & PD-L1 inhibitors in older adults: a meta-analysis. J. Immunother. Cancer *6*, 26.

Fernández, E.A., Mahmoud, Y.D., Veigas, F., Rocha, D., Miranda, M., Merlo, J., Balzarini, M., Lujan, H.D., Rabinovich, G.A., and Girotti, M.R. (2020). Unveiling the immune infiltrate modulation in cancer and response to immunotherapy by MIXTURE-an enhanced deconvolution method. Brief. Bioinformatics.

Finotello, F., Mayer, C., Plattner, C., Laschober, G., Rieder, D., Hackl, H., Krogsdam, A., Loncova, Z., Posch, W., Wilflingseder, D., et al. (2019). Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. *11*, 34.

Franceschi, C., Bonafè, M., Valensin, S., Olivieri, F., De Luca, M., Ottaviani, E., and De Benedictis, G. (2000). Inflamm-aging. An evolutionary perspective on immunosenescence. Ann. N. Y. Acad. Sci. *908*, 244–254.

Frankel, T., Lanfranca, M.P., and Zou, W. (2017). The role of tumor microenvironment in cancer immunotherapy. Adv. Exp. Med. Biol. *1036*, 51–64.

Freeman, A.J., Vervoort, S.J., Ramsbottom, K.M., Kelly, M.J., Michie, J., Pijpers, L., Johnstone, R.W., Kearney, C.J., and Oliaro, J. (2019). Natural Killer Cells Suppress T Cell-Associated Tumor Immune Evasion. Cell Rep. *28*, 2784-2794.e5.

Fulop, T., Larbi, A., Dupuis, G., Le Page, A., Frost, E.H., Cohen, A.A., Witkowski, J.M., and Franceschi, C. (2017). Immunosenescence and Inflamm-Aging As Two Sides of the Same Coin: Friends or Foes? Front. Immunol. *8*, 1960.

Gardner, I.D. (1980). The effect of aging on susceptibility to infection. Rev. Infect. Dis. *2*, 801–810.

Goodman, A.M., Kato, S., Bazhenova, L., Patel, S.P., Frampton, G.M., Miller, V., Stephens, P.J., Daniels, G.A., and Kurzrock, R. (2017). Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. Mol. Cancer Ther. *16*, 2598–2608.

Gounder, S.S., Abdullah, B.J.J., Radzuanb, N.E.I.B.M., Zain, F.D.B.M., Sait, N.B.M., Chua, C., and Subramani, B. (2018). Effect of aging on NK cell population and their proliferation at ex vivo culture condition. Anal Cell Pathol (Amst) *2018*, 7871814.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Han, J., Duan, J., Bai, H., Wang, Y., Wan, R., Wang, X., Chen, S., Tian, Y., Wang, D., Fei, K., et al. (2020). TCR Repertoire Diversity of Peripheral PD-1+CD8+ T Cells Predicts Clinical Outcomes after Immunotherapy in Patients with Non-Small Cell Lung Cancer. Cancer Immunol. Res. *8*, 146–154.

Higgs, B.W., Morehouse, C.A., Streicher, K., Brohawn, P.Z., Pilataxi, F., Gupta, A., and Ranade, K. (2018). Interferon Gamma Messenger RNA Signature in Tumor Biopsies Predicts Outcomes in Patients with Non-Small Cell Lung Carcinoma or Urothelial Cancer Treated with Durvalumab. Clin. Cancer Res. *24*, 3857–3866.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol. *14*, R115.

Hurez, V., Padrón, Á., Svatek, R.S., and Curiel, T.J. (2018). Considerations for successful cancer immunotherapy in aged hosts. Exp. Gerontol. *107*, 27–36.

Jain, V., Hwang, W.-T., Venigalla, S., Nead, K.T., Lukens, J.N., Mitchell, T.C., and Shabason, J.E. (2019). Association of Age with Efficacy of Immunotherapy in Metastatic Melanoma. Oncologist.

Jhunjhunwala, S., Hammer, C., and Delamarre, L. (2021). Antigen presentation in cancer: insights into tumour immunogenicity and immune evasion. Nat. Rev. Cancer.

Klutstein, M., Moss, J., Kaplan, T., and Cedar, H. (2017). Contribution of epigenetic mechanisms to variation in cancer risk among tissues. Proc Natl Acad Sci USA *114*, 2230–2234.

Kovtonyuk, L.V., Fritsch, K., Feng, X., Manz, M.G., and Takizawa, H. (2016). Inflamm-Aging of Hematopoiesis, Hematopoietic Stem Cells, and the Bone Marrow Microenvironment. Front. Immunol. *7*, 502.

Kugel, C.H., Douglass, S.M., Webster, M.R., Kaur, A., Liu, Q., Yin, X., Weiss, S.A., Darvishian, F., Al-Rohil, R.N., Ndoye, A., et al. (2018). Age Correlates with Response to Anti-PD1, Reflecting Age-Related Differences in Intratumoral Effector and Regulatory T-Cell Populations. Clin. Cancer Res. *24*, 5347–5356.

La, J., Cheng, D., Brophy, M.T., Do, N.V., Lee, J.S.H., Tuck, D., and Fillmore, N.R. (2020). Real-World Outcomes for Patients Treated With Immune Checkpoint Inhibitors in the Veterans Affairs System. JCO Clin. Cancer Inform. *4*, 918–928.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Lo, H.C., Xu, Z., Kim, I.S., Pingel, B., Aguirre, S., Kodali, S., Liu, J., Zhang, W., Muscarella, A.M., Hein, S.M., et al. (2020). Resistance to natural killer cell immunosurveillance confers a selective advantage to polyclonal metastasis. Nat. Cancer.

Micevic, G., Thakral, D., McGeary, M., and Bosenberg, M.W. (2019). PD-L1 methylation regulates PD-L1 expression and is associated with melanoma survival. Pigment Cell Melanoma Res. *32*, 435–440.

Morley, A. (1998). Somatic mutation and aging. Ann. N. Y. Acad. Sci. *854*, 20–22.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.

Patel, S.P., and Kurzrock, R. (2015). PD-L1 Expression as a Predictive Biomarker in Cancer Immunotherapy. Mol. Cancer Ther. *14*, 847–856.

Pereira, B., Chin, S.-F., Rueda, O.M., Vollan, H.-K.M., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat. Commun. *7*, 11479.

Qing, T., Mohsen, H., Marczyk, M., Ye, Y., O'Meara, T., Zhao, H., Townsend, J.P., Gerstein, M., Hatzis, C., Kluger, Y., et al. (2020). Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden. Nat. Commun. *11*, 2438.

Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J.E., Ashenberg, O., Cerami, E., Coffey, R.J., Demir, E., et al. (2020). The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell *181*, 236–249.

Sergushichev, A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. BioRxiv.

Shah, Y., Verma, A., Marderstein, A., Bhinder, B., and Elemento, O. (2020). Pan-cancer analysis reveals unique molecular patterns associated with age. MedRxiv.

Shimasaki, N., Jain, A., and Campana, D. (2020). NK cells for cancer immunotherapy. Nat. Rev. Drug Discov. *19*, 200–218.

Solana, R., and Mariani, E. (2000). NK and NK/T cells in human senescence. Vaccine *18*, 1613–1620.

Tabula Muris Consortium (2020). A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. Nature *583*, 590–595.

Taube, J.M. (2014). Unleashing the immune system: PD-1 and PD-Ls in the pre-treatment tumor microenvironment and correlation with response to PD-1/PD-L1 blockade. Oncoimmunology *3*, e963413.

Tauriello, D.V.F., Palomo-Ponce, S., Stork, D., Berenguer-Llergo, A., Badia-Ramentol, J., Iglesias, M., Sevillano, M., Ibiza, S., Cañellas, A., Hernando-Momblona, X., et al. (2018). TGFβ drives immune evasion in genetically reconstituted colon cancer metastasis. Nature *554*, 538–543.

Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.-H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al. (2018). The immune landscape of cancer. Immunity *48*, 812-830.e14.

Tomasetti, C., Li, L., and Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science *355*, 1330–1334.

Wu, Y., Wei, J., Chen, X., Qin, Y., Mao, R., Song, J., and Fan, Y. (2019). Comprehensive transcriptome profiling in elderly cancer patients reveals aging-altered immune cells and immune checkpoints. Int. J. Cancer *144*, 1657–1663.

Xie, W., Kagiampakis, I., Pan, L., Zhang, Y.W., Murphy, L., Tao, Y., Kong, X., Kang, B., Xia, L., Carvalho, F.L.F., et al. (2018). DNA Methylation Patterns Separate Senescence from Transformation Potential and Indicate Cancer Risk. Cancer Cell *33*, 309-321.e5.

Yager, E.J., Ahmed, M., Lanzer, K., Randall, T.D., Woodland, D.L., and Blackman, M.A. (2008). Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus. J. Exp. Med. *205*, 711–723.

Yarchoan, M., Hopkins, A., and Jaffee, E.M. (2017). Tumor Mutational Burden and Response Rate to PD-1 Inhibition. N. Engl. J. Med. *377*, 2500–2501.

Permissions

Erbe et al. Evaluating the impact of age on immune checkpoint therapy biomarkers. Cell Reports. August, 2021.
This is an open access article distributed under the terms of the Creative Commons CC-BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Erbe et al. The use of machine learning to discover regulatory networks controlling biological systems. Molecular Cell. January 2022.
Retained author rights for Elselvier authors allow articles to be published in a thesis or dissertation. See https://www.elsevier.com/about/policies/copyright#Author-rights.

Erbe et al. Transcriptomic forecasting with neural ODEs. BioRXiv. August, 2022.
This is an open access article distributed under the terms of the Creative Commons CC-BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.