NORTHWESTERN UNIVERSITY

Data-Driven Strategies for Optimization of Human Megakaryocyte Differentiation

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Biological Sciences

By

Jia J. Wu

EVANSTON, ILLINOIS

March 2019

Abstract

Platelets are circulating anucleate discs derived from megakaryocytes, and play major roles in hemostasis, inflammation, thrombosis, and vascular biology. Multi-phase culture systems for inducing *in vitro* platelet production from mature megakaryocytes have been explored to allow progenitor expansion, megakaryocyte maturation, and promotion of platelet formation and shedding. In this thesis, I describe the development of several methods for identifying influential factors for multi-phase megakaryocyte differentiation. These methods combine both computational and experimental techniques, and build upon existing approaches. After initial experiments in cell-lines, I constructed a method to develop time-course networks for early, middle, and late megakaryopoiesis from transcription factor array data. Validation with prior knowledge and experimental approaches revealed several false positives and false negatives, which led to the development of a windowed Granger causal inference strategy for network discovery. To identify influential culture factors for megakaryocyte differentiation, we screened several strategies and small molecules for improved ex vivo production. I adapted and applied one of the machine learning frameworks embedded in SWING to characterize donor heterogeneity within individual megakaryocyte culture conditions to improve production and build a predictive framework. Finally, I demonstrate a platform for generation of megakaryocytes from valproic acid expanded cells, as well as a computational method to predict culture performance based on observed donor heterogeneity, which provides potential for identification and intervention in *in vitro* megakaryocyte production processes.

Acknowledgements

Graduate school has been an extraordinary experience. I have had the pleasure of building computational models and experimental protocols with brilliant and amicable people. I have learned an enormous amount and grown tremendously as a scientist and as a person. My experience would not have been possible without my world-class colleagues and collaborators.

This work was supported by my advisors, Professors Neda Bagheri and William M. Miller. I thank them for training me, encouraging me to participate in grant writing and peer reviewing activities, teaching me to think critically, introducing me to colleagues, helping me to secure postdoctoral positions, and taking an overall interest in my scientific and personal development. Both mentors stand apart from their peers and I hope to one day grow to fill their shoes; I will remember Neda's lucidity in addressing research questions and enormous capacity to accomplish both research and collaborative tasks. I will remember Bill's sharpness of mind, and tireless consistency in all tasks.

This work would not be possible without my mentors. Prof. Mariano Loza-Coll sharpened my first set of tweezers in the fly room and set me off on an ambitious project; Kaz Higaki, Teresa DeLuca, and Mark Duncan who served as excellent teachers; the members of the Shea and Mahmud lab for showing me how collaborative research is done; my thesis chair Professor Eric Weiss for unforgettable encouragement and life advice. I would also like to thank my other thesis committee members, Professors Luis Amaral and Linda Broadbelt. They have shared their time and advice with me throughout the years in graduate school. I

would like to especially thank Professor Christian Petersen for reading this thesis, and giving helpful feedback and life advice.

Justin Finkle, Andres Martinez, and Jessica Yu are permanent influences for me. I can't wait to see what they do next.

I thank my lab friends who I share many memories with: Albert, Joe, Sebastian, and Narasimhan. I hope our work will be fortified and surpassed by the next generation of graduate students in our lab.

I thank Jenny (my adventure buddy) and Jonathan (my hiking and life partner).

Most importantly, this work was made possible by my family, particularly my parents and my grandparents, who provided unconditional love, encouragement, and support.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction: Understanding platelet generation from megakaryocytes

1.1. The role of platelets in the human body

Platelets, about a tenth of the size of red blood cells, were discovered in the the 19th century when a microscope with sufficient power became available. In 1865, Max Schultze, a medical doctor and professor at the Anatomical Institute at the University of Bonn, first described platelets as "granules" that are associated with the "coagulation of fibrous material" whose "appearances suggest that coagulation begins from these accumulations of granules" [2]. These observations were later cited by Giulio Bizzozero, a professor of General Pathology at the University of Turin, who conducted well-devised experiments into the function of platelets [1]. Bizzozero and several contemporaries recognized the role of platelets

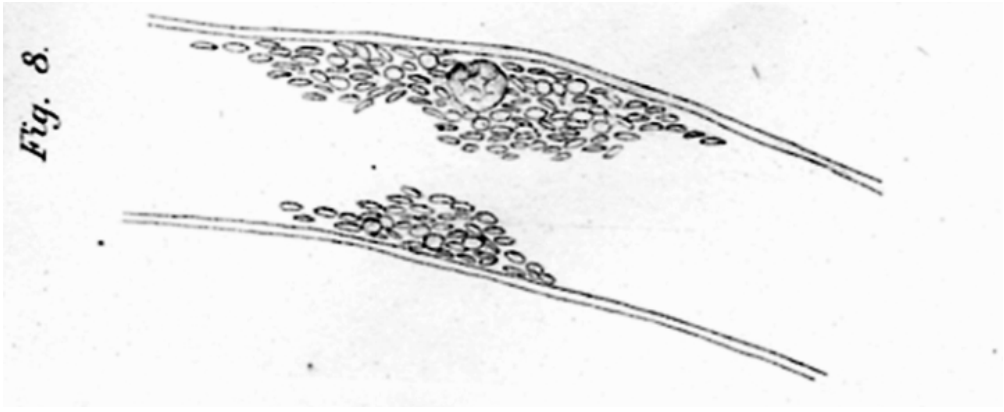

Figure 1.1 | **Guinea pig thrombus (1882).** Two small thrombi formed in the artery of a guinea pig. In the larger thrombus, white blood cells can be seen amongst the platelets. Picture from [1].

in formation of a thrombus in veins of model organisms, such as frogs, guinea pigs, and dogs (Fig. 1.1). Today, platelets are recognized as circulating anucleate discs derived from megakaryocytes (MK), and play major roles in mammalian response to injury or infection, characterized by the interplay of processes such as hemostasis (the process to prevent bleeding or hemorrhage), inflammation (the immune response to injury), and thrombosis (process of coagulation). Platelets have a highly organized cytoskeleton and intracellular stores of proteins, which are secreted at sites of blood vessel injury to initiate and propagate the coagulation response [3]. Specifically, platelets adhere to a von Willebrand factor/collagen matrix at the site of injury, get activated, secrete granules, aggregate via integrin interactions, and produce thrombin to further initiate activation of other platelets and polymerize fibrin (Fig. 1.2). Calcium signaling (Ca^{2+}) plays an important role in the activation process. The formed thrombus activate platelets via GPCR to trigger the release of coagulation factors and the conversion of fibrinogen into fibrin.

Platelets have a short-half life (8-9 days) that requires regular replenishment in the vascular system [5]. In fact, the average human produces 10^{11} new platelets every day. Thus, the hematopoietic system is poised to continuously replenish these lineages by rapid proliferation, differentiation, and maturation of progenitor populations. Platelets are derived from megakaryocytes, rare cells (less than 0.01% of nucleated cells) from the bone marrow [6]. Megakaryocytes are large polyploid cells (up to 128 copies of DNA) which release extensions called proplatelets that elongates, and branches repeatedly [6]. Platelets form selectively at the tips of the extensions and are swept away with blood flow. Biochemically, as platelets develop, granules and organelles are transported by an extended network of membranes called the demarcation membrane system (DMS) to the sites of storage [7]. The time required for MKs to become polyploid and release platelets is estimated to be 5 days in humans [8].

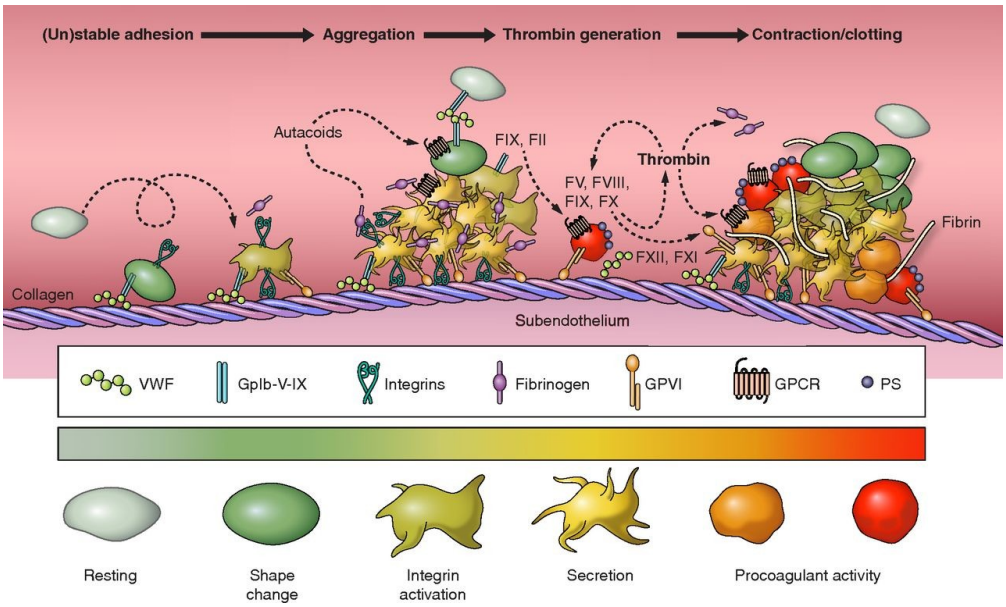

Figure 1.2 | **Thrombosis initation at the site of injury.** Stages of platelet activation and thrombus formation. Platelets form a procoagulant surface and then a contracted thrombus with fibrin. Ca^{2+} signaling ranges from green (low signal) to red (high signal). Image from [4].

1.2. Megakaryocyte differentiation: can we take a systems-level approach to address key questions?

The differentiation process of megakaryocytes is particularly unique; megakaryocytes undergo an extraordinarily dynamic cellular transformation, first becoming polyploid through repeated cycles of DNA synthesis without cell division, then developing proplatelets, and then releasing 2 nm-sized anucleate platelets. To understand parts of the complex life-cycle of the megakaryocyte, I combined approaches in cell biology, engineering, and computational modeling to interrogate a variety of questions. There are several questions that this thesis addresses:

1) On a cell-intrinsic level, how does the core transcription factor (TF) network govern activity during megakaryocyte differentiation?

2) Taking into account external cell signaling and epigenetics, what is the effect of perturbations like histone deacetylase inhibitors on primary human megakaryocyte differentiation and physiology?

3) Finally, on a macroscopic level, what are the key aspects of the endogenous microenviroment that promote and discourage megakaryocyte differentiation?

To address these questions, we employed several *ex vivo* models of megakaryocyte differentiation and a diverse panel of computational approaches. I present a series of studies culminating in multiple biological insights and new techniques. In Chapter 1, I briefly review the history, mechanisms, and current challenges of modeling transcription factor networks of megakaryocytes and other products of the hematopoietic system.

In the first half of the thesis, I describe mechanistic approaches to understand transcriptional and signaling networks of the developing megakaryocyte. In Chapter 2, I employ transcription factor arrays to analyze several aspects of the erythroid-megakaryocyte bifurcation, as well as other transcription factor networks of other systems. In Chapter 3, I describe the development and application of a novel network inference framework to discover temporally-defined transcriptional networks.

In the second half of the thesis, I describe statistical, non-mechanistic approaches for analyzing megakaryocyte differentiation. Chapter 4 introduces several projects related to the characterization and optimization of a novel multi-phase culture protocol to differentiate megakaryocytes from umbilical cord blood-derived CD34+ cells. Chapter 5 describes a novel method for modeling and predicting culture outcomes for megakaryocyte differentiation using time-course flow cytometry. All software described in this thesis will be available under an open-source license. Together, these studies provide the first glimpse into a systems-approach

towards optimizing *ex vivo* megakaryocyte differentiation, which may have broad applications in basic biological science, bioengineering, and medicine.

1.3. *Ex vivo* megakaryocyte production is limited by lack of understanding of coordinated transcription factor network

Recently, genetic studies have provided key insights into the molecular and transcriptional regulation of megakaryocyte (MK) differentiation. Pluripotent human stem cells (HSCs) give rise to MK and erythroid precursor cells, which have the CD34 antigen. These precursor cells are stimulated to begin MK differentiation by a cytokine called thrombopoietin (TPO) [9]. During normal MK differentiation, MKs grow exceptionally large and undergo several rounds of endomitosis, exhibiting high expression of TFs such as GATA1, FOG1, RUNX1, FLI1, SCL/TAL1, and AML1[10] (Fig. 1.3). In this thesis, in addition to using primary cells derived from human patients, we employ several cell lines that can be induced to develop megakaryocytic features such as polyploidization and proplatelet-like-extensions by being induced by phorbol 12-myristate 13-acetate (PMA), such as K562 and CHRF [11, 12]. Disruption of these genes has been shown to decrease ploidy level, and impair differentiation, but their interactions with other TFs have not been systematically studied. As the MK matures, it undergoes proplatelet formation, which is essential for the formation of platelets [6]. A number of genes has been associated with proplatelet formation, such as NFE2, and Myb [6]. As the MK matures, proplatelets eventually extend processes until the cytoplasm is transformed into an extensive network of interconnected proplatelets. Platelets are formed and released at the ends of these pseudopods. During this highly coordinated process, MKs can be identified by their markers CD41/CD61, CD42, and CD62 [13]. Transcription factors coordinately activate numerous genes that function in concert to mediate these processes,

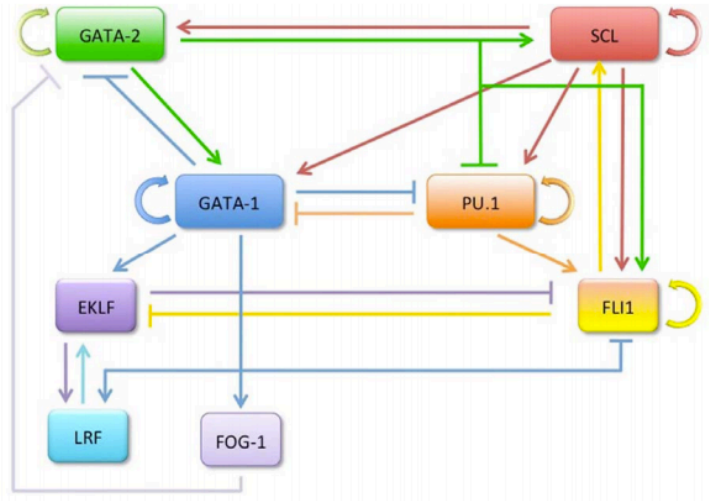

Figure 1.3 | **Commitment of megakaryocytic-erythroid progenitors (MEPs) toward megakaryocytic (MK) and erythroid (E) lineages is orchestrated by a complex network of transcription factors (TF).** Putative transcription factor interactions governing E and MK commitment collected from literature. Figure from Dore, L.C., and Crispino, J.D. (2011) [10].

which are enriched in genetic ontologies such as cytoplasmic reorganization, membrane recruitment, and organelle assembly and transport [10].

Aberrant transcription of transcription factor genes within bipotent progenitor cells of mammalian organisms can lead to dysregulation of proteins involved in pathological progression towards thrombocytosis or thrombocytopenia. Surprisingly, information on how perturbations may affect megakaryocytic-specific transcription factors and overall gene expression is very limited. To understand the maturation program of erythrocytes and megakaryocytes, it is imperative to define how perturbations trigger functional interactions between transcription factors and how these interactions change over time. Insight into the dynamics of the transcription factor network governing cell fate and maturation could reveal new opportunities to drive expression towards the megakaryocyte lineage.

1.3.1. Characterization of transcription factors and transcriptional networks

Transcription factors bind to the promoter region of the gene where they have the ability to recruit other co-factors, and initiate or repress the transcription of the gene. A few studies have investigated transcription factor sequence-specificity on a systematic basis. The primary binding sequence of a transcription factor is called the transcription factor motif. Transcription factor motifs usually consist of 6-10 base pairs, that may or may not be conserved between organisms and cell-types [14]. To identify transcription factor motifs, chromatin immunoprecipitation (ChIP) against the TF of interest is combined with sequencing technologies to map binding sites of the TF within the genome [14]. Additionally, other methods may be employed to establish binding interaction or sequence of binding sites, such as electrophoretic mobility shift assays (EMSA), systematic evolution of ligands by exponential enrichment (SELEX), protein binding microarray (PBM), DNA immunoprecipitation (DIP-CHIP), mechanical trapping (MITOMI), TF-mediated DNA methylation profiling, and surface plasma resonance (BIA-core) [15].

TFs are found to extensively cross-regulate each other through several mechanisms. TFs may enhance or inhibit the function of target TFs by activating or repressing the target promoter by direct binding, binding to a complex, or participating in indirect, complex regulatory mechanisms such as feedback and feedforward inhibition [14]. Additionally, post-translational modifications, such as phosphorylation can alter the binding specificity and combinatorial complexing behavior of a TF [14]. Therefore, TF interactions can readily be represented by an extensive, sometimes ambiguous network, with nodes representing a single or combination of subunits, and edges representing the transfer of information or directed action.

1.3.2. Existing methods of biological network inference are limited by static assumptions

Elucidating gene-gene regulation is a fundamental challenge in molecular biology, and high-throughput technologies continue to provide insight about the underlying organization, or topology, of these interactions. Accurate network models representing genes (nodes) and regulatory interactions (edges) infer information from many observed heterogeneous components while minimizing the effects of noise and hidden nodes. Several methods rely on prior knowledge or databases to generate networks; To illustrate, I have generated a gene regulatory network based on DNAse hypersensitivity maps in K562 cell-lines based on data in Neph et al 2012 (Fig. 1.4). Many methods infer gene regulatory networks (GRNs) from expression profiles [16], but each suffers from limitations—assumptions of linearity, univariate comparisons, or computational complexity —and most ignore temporal information in time-series data. Understanding the temporal dynamics of gene/protein expression is critical to elucidating responses involved in cell cycle, circadian rhythms, DNA damage, and development [17, 18, 19, 20].

The inference of biological networks from high-throughput data has received much attention during the last decade and can be considered an important problem class in systems biology. However, it has been recognized that reliable network inference remains an unsolved problem. Large-scale modeling approaches derive networks directly from 'omics' data. Within the large-scale approach, it is common to handle thousands of unknown parameters, in order to generate a course-grained, genome-wide view of gene regulation. These types of approaches have been shown to perform relatively well in several benchmark studies.

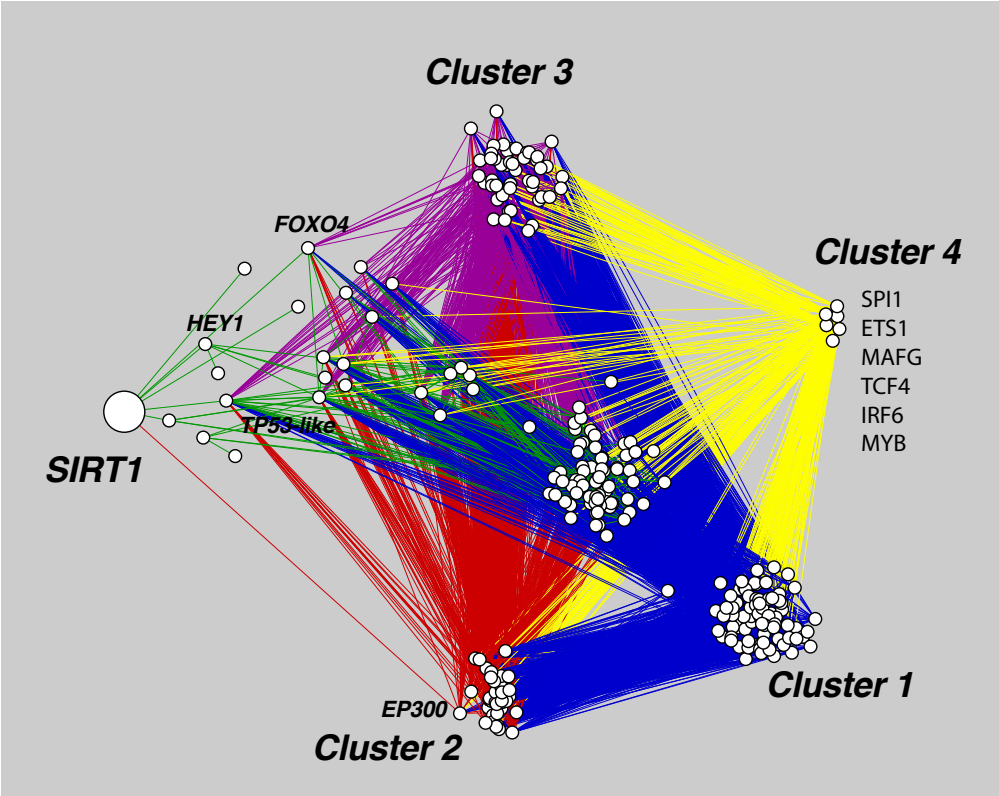

Figure 1.4 | **Example of a gene-regulatory network.** Prior knowledge network was created using a DNAse hypersensitivity K562-specific map. Data available in [21]. Colors of the edges roughly depict the cluster origin. Influential, direct targets of SIRT1 are labeled.

Examples of these approaches include LASSO [22], Inferelator [23], and ARACNE [24]. However, these methods often do not provide high accuracy to realistically predict systems-wide changes in biological systems [25].

Existing methods to infer GRNs from time-series expression profiles include dynamical models, statistical approaches, and hybrids of the two [23, 26, 27, 16]. Dynamical systems models of differential equations can forecast future system behaviors and characterize formal properties such as stability [28], but these models are computationally intractable for large GRNs due to extensive and explicit parameterization requirements [29]. Statistical inference methods—such as regression schemes, mutual information, decision trees, and

Bayesian probability [22, 30, 31]—make no explicit mechanistic assumptions and are often more computationally efficient than dynamical models. However, many implementations of aforementioned algorithms treat time points as independent observations, disregarding time delays associated with transcription, translation, and other processes inherent to gene regulation [25, 32]. Hybrid methods—such as SINDy and Jump3—use statistical methods to optimize the search and parameterization of dynamical models, but they remain computationally expensive and rely on accurate specification of basis functions [33, 34]. Most authors have identified lack of data and deficiencies in the inference algorithms as the main reasons for this situation. During the last decade, many methods have been developed to solve the network-inference (sometimes called reverse-engineering) problems arising in gene expression, signal transduction and metabolic networks. I employed Random Forest methodologies to infer mechanistic factors governing megakaryocytic differentiation, continued to improve on the Random Forest methodology by building a framework for time-series data inference (Chapter 3).

1.4. Testing megakaryocyte differentiation models *ex vivo*

In order to model, perturb, and validate hypothesis pertaining to the life-cycle of the human megakaryocyte, cell or animal models for megakaryocyte differentiation need to be established. The limited understanding of the megakaryocyte lineage development is largely due to the dearth of useful systems in which to conduct experiments. There are several models that others have used to model the megakaryocyte life cycle, from cell-lines to primary cells derived from mouse and human [35]. In particular, the limited number of megakaryocytes and the difficulty of extracting them from the bone marrow, has hindered the procurement of purified primary cell populations. Thus, *ex vivo* or *in vitro* models of

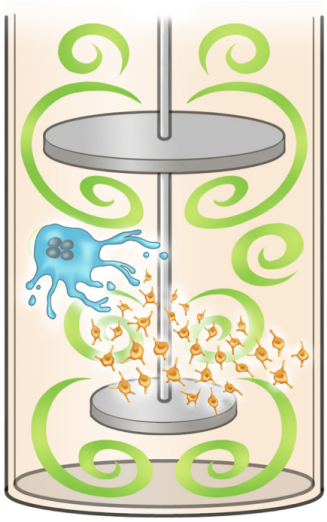

Figure 1.5 | **The dream: a bioreactor for megakaryocyte and platelet generation.** Stimulants to initiate megakaryocyte development are known. Several groups have generated bioreactors for megakaryocyte development and platelet generation. Image from [37].

differentiation of megakaryocytes from primary human stem cells serve as useful tools for understanding megakaryocyte differentiation. Additionally, others have developed systems to generate platelet-like-particles *ex vivo* in near-clinically relevant scales from multiple cell types [36, 37] (Fig. 1.5). These artificial systems derived from iPSC cells, may allow us to manufacture platelets at an industrial scale to replace methods for isolating donor-derived platelets. Chapter 4 is a departure from computational work, where I focus on characterizing differentiating *in vitro* MKs. Detailed summaries of previous work on differentiating *in vitro* MKs can be found in Chapter 4. Finally, collection of this dataset led to another modeling opportunity on a different biological scale; we generated a novel predictive model for heterogenous cellular differentiation on a single cell level for multiple donors (Chapter 5).

Thus, this thesis is an assortment of various computational and experimental analysis of the megakaryocyte life-cycle. In summary, the first half of the thesis mainly describes

computational methods for identifying transcription factor and gene regulatory network interactions, some of which relates to the core megakaryocyte network. The second half of the thesis describe computational and experimental methods to perturb *in vitro* megakaryocyte differentiation process towards additional cell proliferation.

CHAPTER 2

Inferring dynamic activity of key transcription factors during megakaryocyte differentiation and other applications

Work presented in this chapter consists of my work with transcription factor arrays adapted from the following papers:

- Duncan M.T.*, Shin S.*, Wu J.J.*, Mays Z., Weng S., Bagheri N., Miller W.M., Shea L.D. Dynamic transcription factor activity profiles reveal key regulatory interactions during megakaryocytic and erythroid differentiation. Biotechnology and Bioengineering 111:2082–2094. 2014.
- Aguado B.A., Wu J.J., Azarin S.M., Nanavati D., Rao S.S., Bushnell G.G., Chaitanya M.B., Shea, L.D. Secretome identification of immune cell factors mediating metastatic cell homing. Scientific Reports 5, 17566. 2015.

2.1. Abstract

The directed differentiation toward erythroid (E) or megakaryocytic (MK) lineages by the MK–E progenitor (MEP) could enhance the *ex vivo* generation of red blood cells and platelets for therapeutic transfusions. The lineage choice at the MEP bifurcation is controlled in large part by activity within the intracellular signal transduction network, the output of which determines the activity of transcription factors (TFs) and ultimately gene expression. Although many TFs have been implicated, E or MK differentiation is a complex process requiring multiple days, and the dynamics of TF activities during commitment

and terminal maturation are relatively unexplored. Thus my colleagues and I applied a transcription factor array for the large–scale, dynamic quantification of TF activities during MEP bifurcation. A panel of hematopoietic TFs (GATA–1, GATA–2, SCL/TAL1, FLI–1, NF–E2, PU.1, c–Myb) was characterized during E and MK differentiation of bipotent K562 cells. Dynamic TF activity profiles associated with differentiation towards each lineage were identified, and validated with previous reports. From these activity profiles, I show that GATA–1 is an important hub during early hemin–and PMA–induced differentiation, and reveal several characteristic TF interactions for E and MK differentiation that confirm regulatory mechanisms documented in the literature. Additionally, I highlight several novel TF interactions at various stages of E and MK differentiation. Furthermore, I investigated the mechanism by which nicotinamide (NIC) promoted terminal MK maturation using an MK–committed cell line, CHRF–288–11 (CHRF). Concomitant with its enhancement of ploidy, NIC strongly enhanced the activity of three TFs with known involvement in terminal MK maturation: FLI–1, NF–E2, and p53. Dynamic profiling of TF activity represents a novel tool to complement traditional assays focused on mRNA and protein expression levels to understand progenitor cell differentiation.

2.2. Transcription factor arrays in MEP cell line model reveal critical regulation supported by prior knowledge.

Enhancers, which are binding sequences located in non–coding sites and exonic regions, contribute to the regulation of location, timing, and levels of gene transcription of given genes [38]. The genetic reporter assay is a well–established tool for dissecting the activity, or the ability to regulate gene function, of a given enhancer sequence. Enhancers are generally characterized by a reporter assay that links a multimerized sequence binding site

(purportedly for TFs and cofactors) to a minimal reporter sequence and reporter gene (such as lacZ, GFP, or luciferase). These reporter vectors are introduced to cell lines or organisms to examine the binding activity of an enhancer. In this classical method, examination of enhancer activity is low–throughput and time–consuming, since candidate sequences would be introduced to cell lines and examined individually. Recently, the Shea lab has optimized their own methodology for a "living cell array" for the large–scale quantification of dynamic TF activities (Fig. 2.1). This assay directly quantifies the activity of TFs through a reporter assay driven by a minimal CMV promoter expressing luciferase, and allows us to examine time–series identification of TF activity during lineage commitment and differentiation [39, 40].

2.3. TF–Reporter assays are imperfect elements for observing TF activity, so modeled interactions should be experimentally validated

Prepending the discussion of this work, there are several noted caveats with the reporter assay approach to model TF activities. In this "living cell array" approach and other classical enhancer assay designs, the TF binding sequences are removed from their genomic context. Thus the effect of enhancer–promoter distance, looping, and chromosome state is not taken account in the reporter assay. Also, these reporters use a standard minimal promoter instead of the actual promoter targeted by the candidate enhancer constructs.

Specifically with multimeric transcription factor binding sites designed in Fig. 2.1, we note that endogenous TF binding sites found in organisms consist of combinations of sites in a context–dependent manner [38], and not necessarily in a series of concatenated sequences with spacers. Additionally, TF sequences exhibit highly degenerate sequences–that result in a large number of nonspecific interactions or promiscuous binding [41].

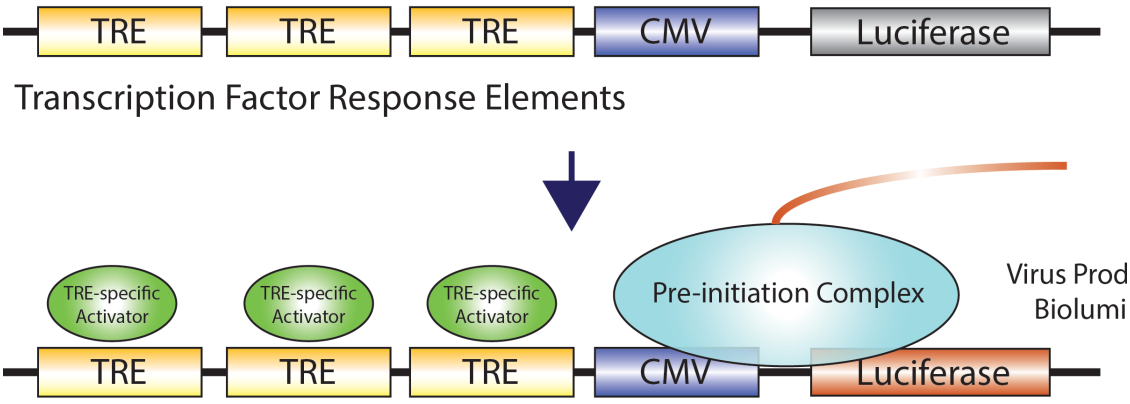

Figure 2.1 | **Schematic showing a transcription factor array construct consisting of three transcriptional response elements (TREs) prepended to a CMV promoter driving expression of luciferase.** The TRE binds transcription factors that recruit the pre–initiation complex (RNA polymerase) to initiate and maintain transcriptional activity.

Ultimately, to characterize the functional effect of an enhancer or TF binding interaction, an experiment needs to be conducted such that the enhancer or TF binding of interest is assayed in an endogenous setting. One potential tool that facilitates this is genome editing via CRISPR/CAS9 technology, which can test the effect of enhancer mutations or knock–in reporter genes at endogenous sequences [42]. In these subsequent experiments, we knock down the TF identified to influence many transcription factors early on. Thus, we caution that the approach described is to be used for screening purposes only.

In this study, I applied the TF activity assay to investigate E versus MK commitment and differentiation using the model cell line K562, which resembles MEPs in that it is bipotent for the E and MK lineages [43]. My colleagues and I selected a panel of seven TFs known to be involved in E/MK differentiation and monitored their dynamic activities throughout the differentiation process. First, I examined the divergence in TF activities associated with the

bifurcation between the E and MK lineages. I then utilized an ensemble tree–based inference algorithm (GENIE3) to infer the TF regulatory network for both lineages and performed a topological analysis of the inferred network [30]. The impact of knocking out the GATA–1 TF on the subsequent response of the TF network was also determined. Finally, I investigated the dynamic TF activity associated with NIC promotion of MK maturation using the CHRF–288–11 (CHRF) cell line, which resembles MK progenitors[44]. The previously established NIC–mediated inhibition of SIRTs, as well as changes in metabolism due to increased NAD+ concentration [45], were expected to influence TF activities. TF activity arrays can provide unique perspectives on cell differentiation, which may ultimately be translated into strategies to more effectively promote production of cells in specific lineages.

2.3.1. TF Activity During E Versus MK Differentiation of K562 Cells show highly transient upregulation and downregulation of critical TFs

The activity profiles of 7 key hematopoietic TFs [10] were quantified over the 5–day culture, during which cells differentiate to either E or MK phenotypes (Fig. 2.2A). Throughout MK differentiation, I noted a gradual reduction (Days 1–3) and recovery (Days 4–5) in the activities of GATA–1, c–Myb, and PU.1. TAL1, had a slight decrease by Day 5. FLI–1 was rapidly increased, due to induction by PMA and remained significantly upregulated with respect to the untreated control ($P < 0.05$) except for a transient decrease at Day 3. NF–E2, important for regulating platelet release from mature MKs also showed an immediate and strong activation, but this activation gradually regressed to the level of untreated cells.

During E differentiation, I observed early, strong activation of both NF–E2 (Day 1) and GATA–2 (Day 2). NF–E2 activity subsequently fluctuated, but remained >2–fold higher than in untreated cells, while GATA–2 activity peaked at Day 3. TAL1 was significantly

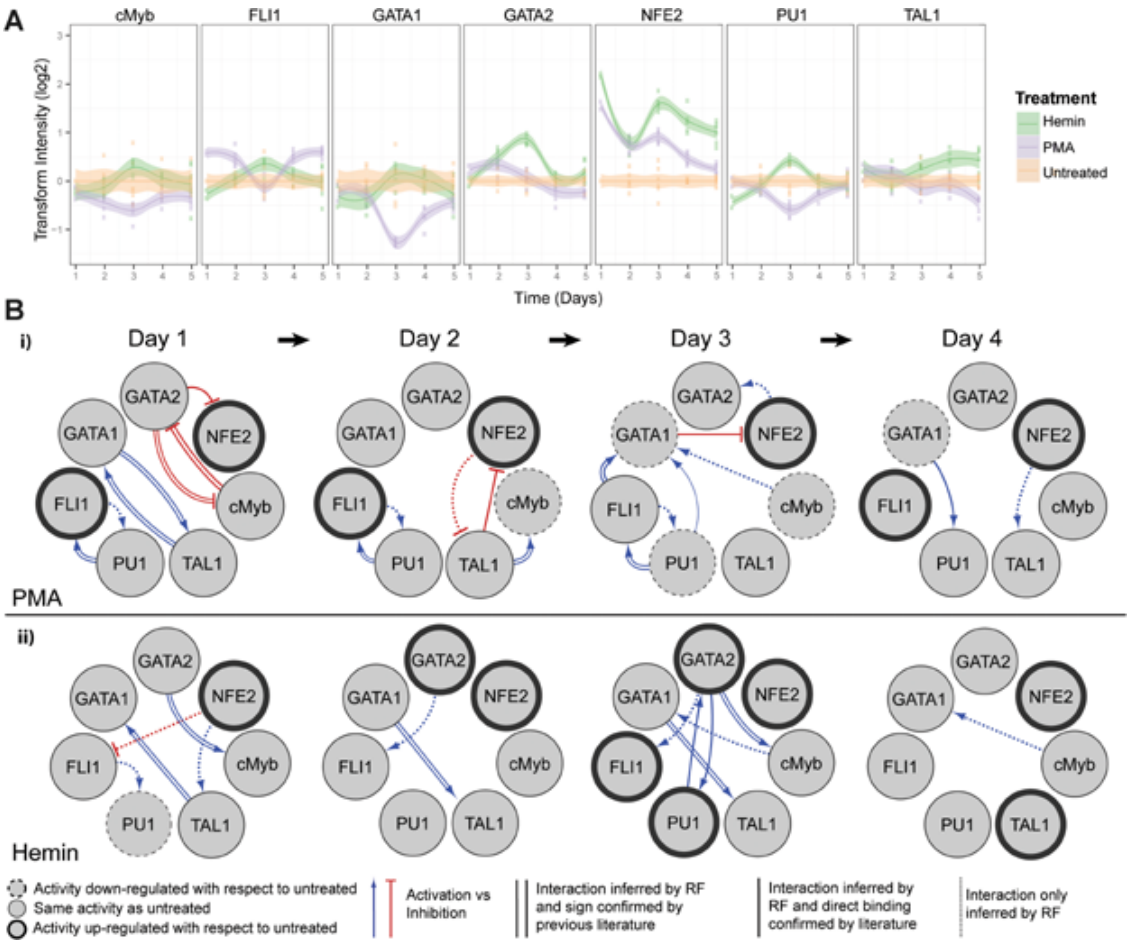

Figure 2.2 | **Dynamic TF activity array and regulatory networks of K562 cells during MK or E differentiation.** A) K562 cells transduced by each TF–activity–reporting lentivirus were treated with 10ng/mL of PMA (MK–differentiation) or 30μM of hemin (E differentiation). TF activity was normalized by TA activity and represented by the ratio to vehicle–treated cells. Shaded regions denote 95% confidence intervals about the mean of six measurements (indicated by data points) from two independent transduction experiments. B) The dynamic TF regulatory networks for PMA–and hemin–treated K562 cells inferred from TF array data. The TF network shows target and regulator TFs (circles), as well as putative direct or indirect interactions between TFs (directed arrows). TF nodes (circles) indicate if TF activity is upregulated (bold outline), downregulated (dashed outline), or unchanged (thin outline) with respect to untreated control on the indicated day. Edge line styles depict whether the inferred interaction is a direct interaction confirmed by previous literature found in GENEGO (parallel lines), a direct interaction confirmed by previous literature but activation (blue lines) or inhibition (red lines) of the target is unknown (solid lines), or inferred interaction is novel (dotted lines).

upregulated with respect to the untreated control by Day 3 and had increased activity throughout the differentiation period. Although an initial slight repression was observed relative to untreated cells, GATA–1 and PU.1 activities trended toward an increase with respect to control until peaking at Day 3 and declining thereafter. The activities of c–Myb and the MK–specific TF FLI–1 remained similar to that of untreated cells beyond Day 3.

The observed trends in our activity data were consistent with literature reports of expression data, and with functional studies describing the specific role of each TF in regulating E/MK differentiation (see below). However, I note that changes in expression level do not always correspond to changes in TF activity level due to posttranslational regulatory mechanisms, and emphasize that TF activity measurements should not be considered as dynamic measurements on the protein level.

2.3.2. Random Forest–based construction of TF regulatory network reveals characteristic edges for E Versus MK differentiation

Next, I utilized the TF activity data to construct a regulatory network that identifies putative positive and negative interactions between TFs. The network was inferred by applying tree ensemble–based models to predict a set of possible regulators for each TF (Fig. 2.2B). We sought to evaluate the number of inferred edges that are supported by evidence of direct binding from literature. Of the 35 edges predicted for E/MK differentiation over 5 days, 20 edges had supporting evidence of direct binding. Although these relationships have been identified by other authors as direct binding, we note that the relationship inferred by the network could be direct or the result of indirect interactions. For several of the edges with evidence of direct binding, sources could not confirm whether the regulatory TF activated or inhibited the target TF. The probability of an edge within the network shown in Figure

PMA (CHRF)

Day 1-2			Day 2-3			Day 3-4			Day 4-5			Day 5-6		
Regulator	Target	p-value	Regulator	Target	p-value	Regulator	Target	p-value	Regulator	Target	p-value	Regulator	Target	p-value
FLI1	NFE2	0.007	p53	FLI1	0.02	GATA1	cMyb	0.007	FLI1	cMyb	0.01	FLI1	cMyb	0.01
cMyb	p53	0.006	GATA1	NFE2	0.06	GATA2	p53	0.006	NFE2	GATA2	0.02	GATA2	NFE2	0.02
p53	GATA2	0.03	FLI1	p53	0.06	p53	GATA2	0.04	FLI1	NFE2	0.04	p53	PU1	0.04
TAL1	FLI1	0.06	GATA1	cMyb	0.07	p53	PU1	0.05	GATA2	FLI1	0.06	GATA1	TAL1	0.06
PU1	cMyb	0.07	FLI1	GATA2	0.08	cMyb	GATA1	0.06	NFE2	TAL1	0.07	p53	GATA2	0.08
cMyb	TAL1	0.08	cMyb	PU1	0.09	GATA2	FLI1	0.06	TAL1	NFE2	0.08	NFE2	GATA1	0.09
PU1	GATA1	0.10	p53	GATA2	0.10	cMyb	NFE2	0.10	PU1	p53	0.09			
			GATA2	p53	0.10				GATA2	p53	0.09			

PMA and NIC (CHRF)

Day 1-2			Day 2-3			Day 3-4			Day 4-5			Day 5-6		
Regulator	Target	p-value	Regulator	Target	p-value	Regulator	Target	p-value	Regulator	Target	p-value	Regulator	Target	p-value
TAL1	NFE2	0.001	FLI1	PU1	0.03	FLI1	p53	0.01	TAL1	NFE2	0.03	TAL1	NFE2	0.02
NFE2	TAL1	0.002	cMyb	GATA1	0.04	PU1	NFE2	0.04	NFE2	FLI1	0.07	NFE2	TAL1	0.04
PU1	FLI1	0.01	GATA1	TAL1	0.08	NFE2	GATA1	0.05	FLI1	TAL1	0.07	PU1	cMyb	0.07
GATA1	cMyb	0.06	p53	GATA2	0.1	FLI1	GATA2	0.07	TAL1	FLI1	0.08	p53	PU1	0.07
FLI1	PU1	0.06				NFE2	FLI1	0.9	GATA2	TAL1	0.1	GATA1	p53	0.08
FLI1	p53	0.07				FLI1	cMyb	0.9	PU1	GATA1	0.1	cMyb	PU1	0.09
PU1	GATA2	0.09				NFE2	PU1	0.1				GATA2	TAL1	0.1
FLI1	cMyb	0.1												

Figure 2.3 | **Leave–one–out table for K562 and CHRF networks.** P–values of inferred edges of CHRF cells during MK maturation with PMA and PMA plus NIC found using the permutation test.

2B was investigated using a leave–one–out (LOO) analysis. All interactions in the resulting network were present in those culminating from LOO analysis, though not all interactions were present in every LOO network in which a TF had been removed. The frequency of a sustained linkage within networks lacking a single TF reflects confidence of that interaction in the resulting model. The ranked list of interactions for the LOO analysis at each time point can be found in 2.3.

In the reconstructed networks, active regulatory interactions, as well as strong regulatory hubs (i.e., nodes with a large number of interactions), have been identified in both PMA (Fig. 2.2Bi) and hemin (Fig. 2.2Bii) networks. Judged by the number of connections, GATA–1 appears to be an influential TF with respect to both hemin–and PMA–mediated differentiation. According to the model, GATA–1 is a target of most TFs that were screened. In both PMA and hemin networks, GATA–1 was found to participate in a mutually activating relationship with TAL1. The activation of these links early in both hemin–and PMA–induced differentiation reflects that these processes are essential to commitment for both E and

MK fate. GATA–1 was also inferred to be activated by c–Myb during Day 3 and 4 of hemin–induced differentiation, and Day 3 for PMA–induced differentiation.

In addition, the inferred networks identified characteristic interactions for hemin–and PMA–induced differentiation of K562 cells. For PMA–induced differentiation, FLI–1 and PU.1 participated in a mutually activating relationship for a majority of the timeline that was observed. During hemin–induced differentiation, GATA–2 and PU.1 were predicted to act cooperatively in Day 3. Additionally, GATA–2 and c–Myb were inferred to participate in a mutually inhibiting relationship in the PMA network, while during hemin–induced differentiation, GATA–2 was predicted to activate c–Myb. Evidence of reciprocal target binding sites has been found in GATA–2 and c–Myb promoters, but it is not known whether this relationship is activating or inhibiting [46, 47]. Our results suggest that both cooperative and antagonistic interplay exist between c–Myb and GATA–2, and that the nature of the relationship may contribute to the specification of either E or MK commitment.

2.3.3. Generation and Characterization of GATA–1 Depleted K562 Cells reveals GATA–1 as a critical regulator of E and MK differentiation

The well–established importance of GATA–1 (confirmed by the preceding network analysis) in regulating normal E and MK differentiation motivated studies with the silencing of GATA–1 during K562 cell differentiation and the measurement of TF activity profiles. Inhibition of GATA–1 is predicted to disrupt differentiation for both E and MK lineages.

Lentivirus encoding shRNA against GATA–1 was delivered, with 90% knockdown of GATA–1 mRNA confirmed by qRT–PCR (Fig. 2.4A). This potent silencing of GATA–1 greatly reduced K562 cell expression of the constitutively expressed E antigens GlyA and CD71 (Fig. 2.4B). In particular, a substantial subpopulation (50%) of GlyACD71 cells

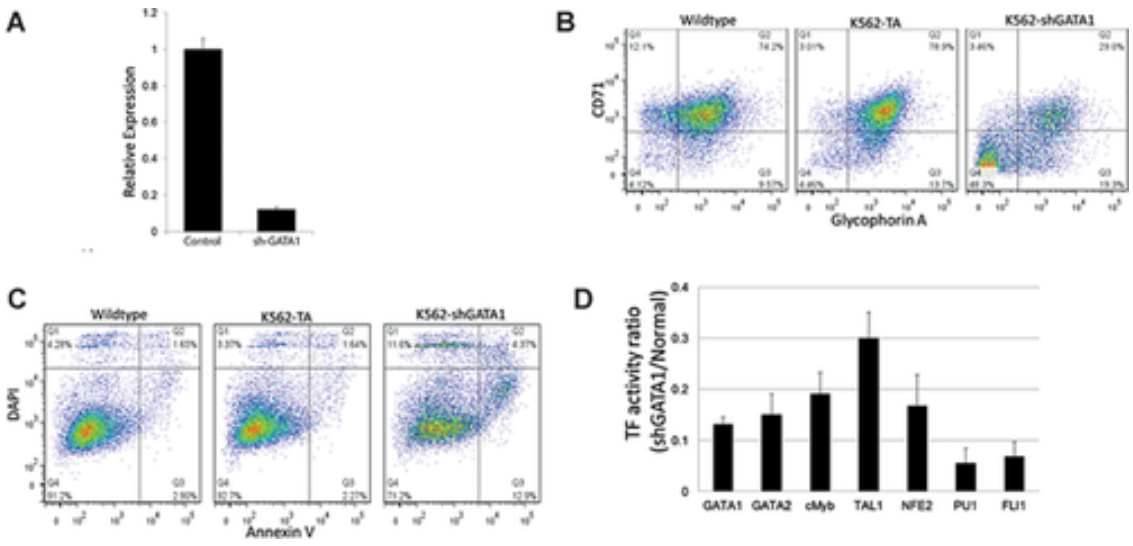

Figure 2.4 | **Characterization of GATA–1 silenced K562 cells in control media.** A: qRT–PCR of GATA–1 mRNA levels in shGATA–1 cells. B: Expression of the erythroid antigens CD71 and Glycophorin A. C: DAPI and Annexin V staining of shGATA–1 K562 cells. Data is representative of two independent transduction experiments. D: TF activity normalized by TA activity and represented by the ratio to normal K562 cells. All tested TF activities were significantly reduced ($P<0.05$) by GATA–1 knockdown. Error bars in (A) and (D) represent the standard deviation (n=3 biological replicates).

emerged, suggesting a definitive regression from the MEP–like phenotype. In addition, GATA–1 silencing increased the fraction of apoptotic (AnV+DAPI) and non–viable cells (DAPI+) (Fig. 2.4C). Further, the basal level of all TF activities was significantly reduced (to 5–30% of wild–type level) by GATA–1 knockdown (Fig. 2.4D). These profound reductions further indicate the importance of GATA–1 for the maintenance of the MEP–like phenotype.

Next, we examined the E and MK differentiation of GATA–1 silenced K562 cells. GATA–1 silenced K562 cells were unresponsive to PMA treatment, failing to acquire CD41 (Fig. 2.5A) or undergo polyploidization (Fig. 2.5B) and typically retaining a small, round, undifferentiated morphology (Fig. 2.5C). Additionally, GATA–1 silenced cells did not respond to hemin, as they failed to produce hemoglobin (Fig. 2.6A and B) or upregulate GlyA expression (Fig.

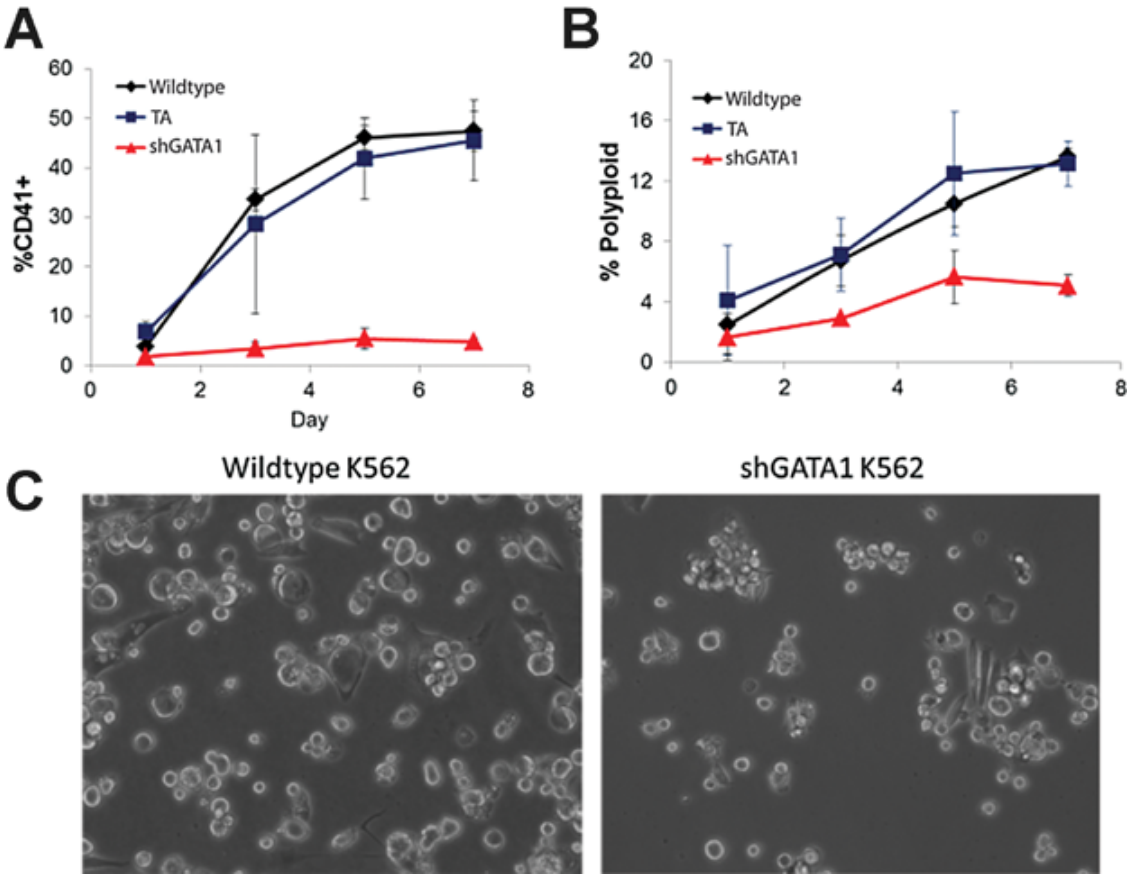

Figure 2.5 | **Characterization of PMA induced MK differentiation of GATA–1 silenced K562 cells.** Ploidy (A) and CD41 expression (B) were monitored for 1 week after PMA induction to assess MK differentiation. Data is from two independent time–course experiments. Error bars represent the standard deviation. C: GATA–1 silenced cells retained a small, rounded morphology 5 days post–PMA treatment.

2.6C). These results were consistent with the preceding network analysis, where GATA–1 was predicted to be essential for both E and MK differentiation.

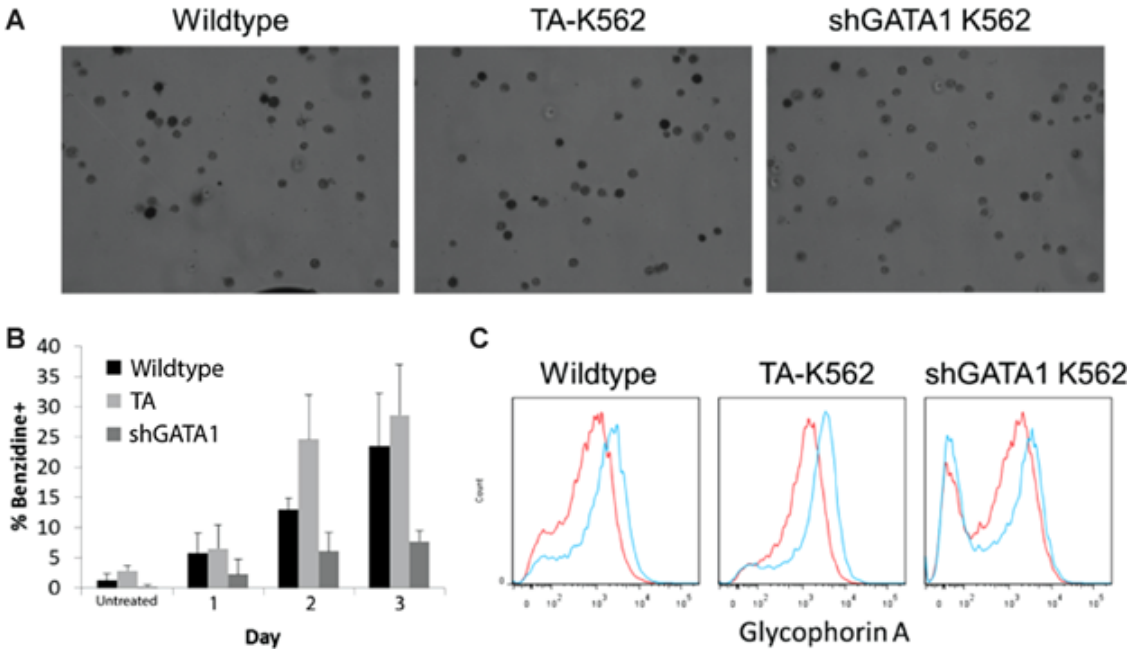

Figure 2.6 | **Characterization of hemin–induced E differentiation of GATA–1 silenced K562 cells.** TA–transduced, sh–GATA1 and wild–type K562 cells were induced with hemin for 3 days in comparison to wild–type and TA transduced cells. Erythroid differentiation was assessed by hemoglobinization using benzidine staining
(A: representative light micrographs on Day 3; B: quantification) and Glycophorin A expression (C). Representative flow cytometry histograms on day 3 for uninduced (red lines) and hemin–induced (blue lines) cells. Data is from two independent time–course experiments. Error bars in (B) represent the standard deviation.

2.3.4. TF Activity Profiles and Regulatory Network of GATA–1 Depleted Cells reveal knockdown of fate–specific TFs

Next, we considered the TF activities of GATA–1 depleted cells during differentiation using the dynamic TF activity array (Fig. 2.7). During MK differentiation, we found that PU.1 and c–Myb activities were no longer repressed by PMA treatment in GATA–1 silenced cells. In addition, although FLI–1 activity was initially elevated, the increase was not sustained. Surprisingly, NF–E2 activation remained robust, indicating that NF-E2 activation is GATA-1 independent.

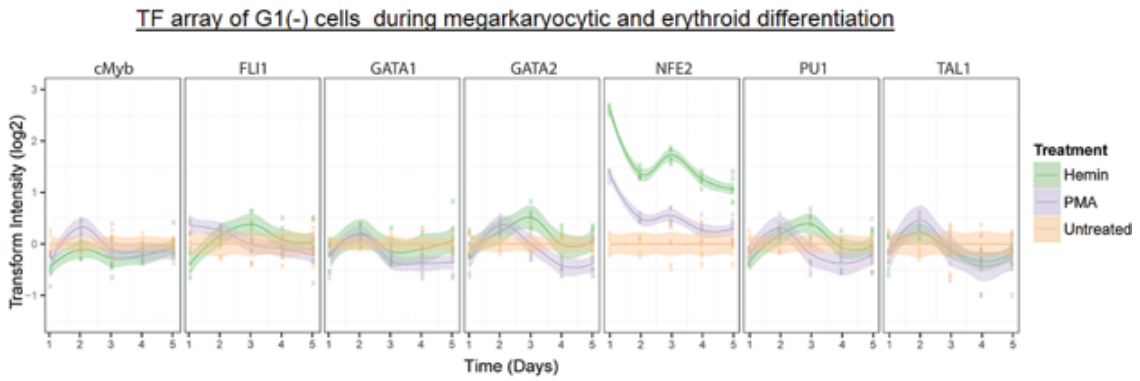

Figure 2.7 | **Dynamic TF activity array and GATA1–silenced–K562 cells during E or MK differentiation.** GATA–1–down–regulated K562 cells were transduced by each TF–activity–reporting lentivirus and treated with PMA or hemin. TF activity was normalized by TA activity and represented by the ratio to vehicle–treated cells. Shaded regions denote the 95% confidence interval about the mean for eight measurements (indicated by data points) from two independent transduction experiments.

For E differentiation, the increase in GATA–2 activity was much less than for wild–type cells and the increase in TAL1 activity observed for wild–type cells did not occur in GATA–1 silenced cells. However, as with PMA treatment, hemin rapidly induced NF–E2 activity despite GATA–1 depletion. Taken together, these results demonstrate that, aside from NF–E2, the TF activity trends shown in Figure 2.2A are GATA–1–dependent and are specifically associated with E and MK differentiation. Finally, we also created an interaction network for GATA–1 silenced cells, similar to that shown in Figure 2.2B. As expected, this network substantially differed from that found during the differentiation of wild–type cells (Fig. 2.8). Of note, a majority of the characteristic interactions identified for PMA–and hemin–induced differentiation in wild–type cells were no longer present. In particular, as may be expected, the mutually activating relationship between GATA–1 and TAL1 was no longer present in early PMA–and hemin–induced differentiation, and GATA–1 no longer appears to be a target of FLI–1, PU.1, and c–Myb in PMA–induced differentiation. Depletion of GATA–1 also

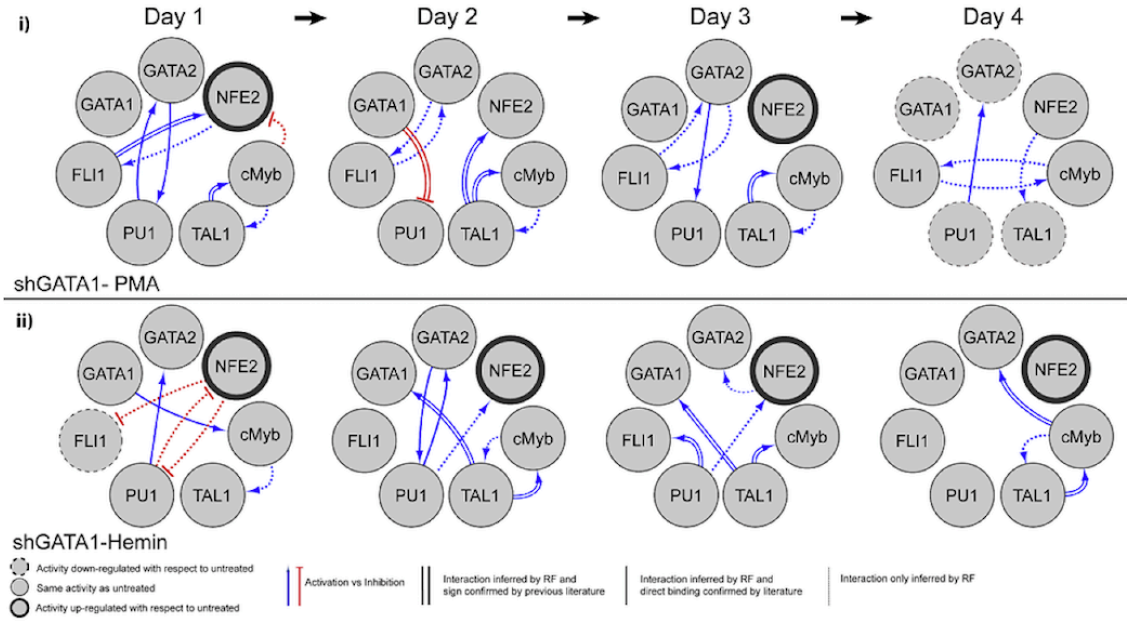

Figure 2.8 | **Dynamic TF regulatory network of GATA–1–downregulated K562 cells during E and MK differentiation.** Inferred TF network indicates that putative interactions are disrupted when GATA–1 is down–regulated. TF networks show target and regulator TFs (circles), as well as putative direct or indirect interactions between TFs (directed arrows). TF nodes indicate if TF activity is up–regulated (bold outline), down–regulated (dashed outline), or unchanged (thin outline) with respect to untreated control on indicated day. Edge weights depict whether the inferred interaction is a direct interaction confirmed by previous literature found in GENEGO (parallel lines), a direct interaction confirmed by previous literature but activation (blue lines) or inhibition (red lines) of the target is unknown (solid lines), or inferred interaction is novel (dotted lines).

resulted in mutually activating relationships between TAL1 and c–Myb (hemin and PMA) and between PU.1 and GATA–2 (PMA only) that were not observed for wild–type cells.

2.3.5. TF Activity and Regulatory Network for CHRF cells identify characteristic interactions influencing cell polyploidization

We investigated the mechanism of NIC action by profiling TF activities and constructing a putative regulatory network using the MK cell line CHRF, which the Miller lab have extensively validated as a model of MK differentiation [48]. CHRF cells are committed to the MK lineage, yet are terminally differentiated by PMA addition, becoming polyploid and forming proplatelet–like extensions. Importantly, as with primary MK cells, CHRF cells substantially increase their ploidy in response to NIC [45].

After establishing CHRF TF–reporter cell lines, dynamic TF activities in response to treatment with either PMA or PMA plus NIC over 6 days of differentiation was quantified. Cells harvested from the array at the final day (Day 6) confirmed that NIC treatment consistently increased polyploidization under the culture conditions (Fig. 2.9A). TF activity analysis (Fig. 2.9B) revealed that, consistent with K562 cells, PMA tended to increase the activities of NF–E2 and FLI–1 versus untreated cells (typically 1.5–to 2–fold induction). Also, similar to K562 cells, GATA–2 and TAL1 activities remained essentially constant after PMA induction. However, the decreases in GATA–1, PU.1, and c–Myb activities characteristic of K562 cells were not observed for CHRF cells. These differences likely result from the greater maturity of untreated CHRF cells (MK progenitor phenotype) relative to K562 cells (MEP phenotype), so that the basal levels of the aforementioned TFs may have already been down–regulated prior to PMA treatment. p53–activity increased throughout MK differentiation. Adding NIC strongly enhanced the activities of FLI–1, PU.1, and NF–E2, while moderately enhancing p53 and GATA–2 activities.

To identify possible mechanisms for the response of CHRF cells to NIC treatment, functional TF networks were created and characteristic relationships were identified. Out of 68

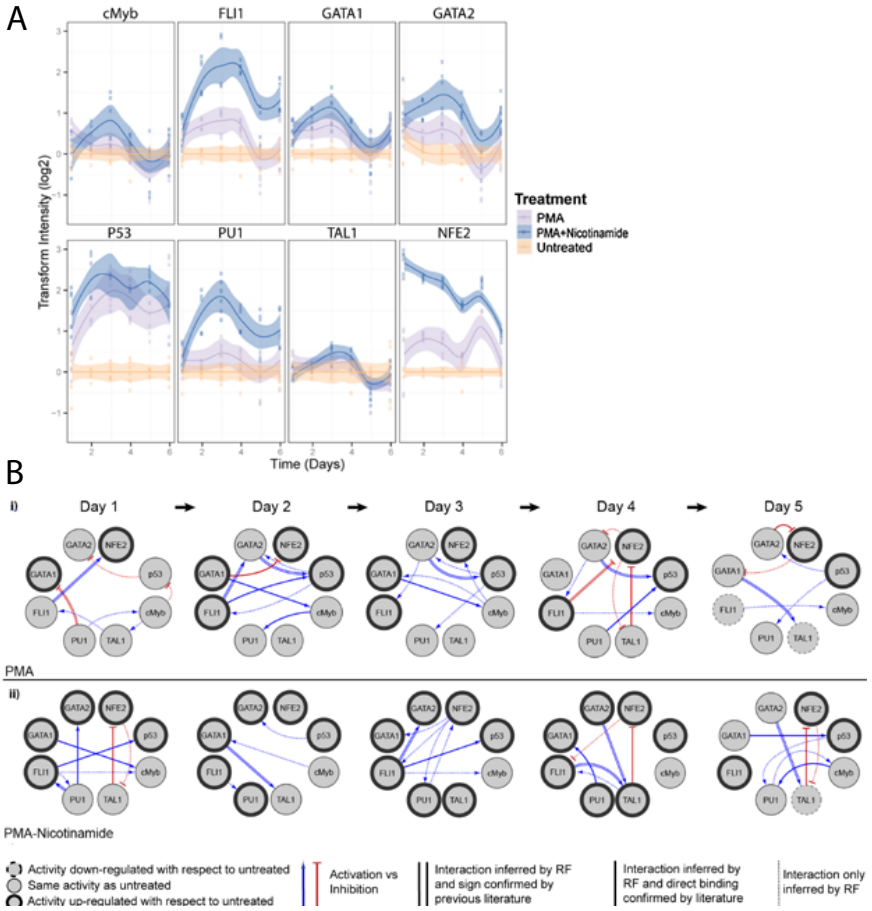

Figure 2.9 | **Dynamic TF activity array and regulatory networks of CHRF cells during MK maturation.** A: Dynamic TF activity profiles. TF activity was normalized by TA activity and represented by the ratio to untreated cells. Shaded regions denote the 95% confidence interval of the mean of six measurements (indicated by data points) from two independent transduction experiments. Error bars indicate standard deviation. B: TF regulatory network was inferred for CHRF cells during MK maturation with PMA, and PMA plus NIC treatments. p53 was added as an additional node to identify potential TFs that regulate its activity. TF networks show target and regulator TFs (circles), as well as putative direct or indirect interactions between TFs (directed arrows). TF nodes indicate if TF activity is upregulated (bold outline), downregulated (dashed outline), or unchanged (thin outline) with respect to untreated control on indicated day. Edge weights depict whether the inferred interaction is a direct interaction confirmed by previous literature found in GENEGO (parallel lines), a direct interaction confirmed by previous literature but activation (blue lines) or inhibition (red lines) of the target is unknown (solid lines), or inferred interaction is novel (dotted lines).

edges inferred, 32 edges were found to have evidence in the literature for direct binding. After NIC stimulation, NF–E2 appears to be a potent cooperator with GATA–1, PU.1, and FLI–1 (Day 3). During PMA treatment without NIC, p53 expression appears to be primarily regulated by GATA–2. With NIC treatment, p53 activation does not strongly correspond to GATA–2 regulation and appears to cooperate with FLI–1 and GATA–1. I observed fewer edges to p53 with NIC treatment. In contrast, there were more edges to PU.1 with NIC. TAL1 and NF–E2 exhibited mutual inhibition, especially in cultures with NIC. Overall, little similarity was observed between the functional TF networks inferred for PMA treatment in the absence or presence of NIC.

2.3.6. Discussion

The distinct TF activity profiles we observed during E versus MK differentiation of K562 cells provide a unique perspective for understanding how these cells develop toward two disparate phenotypes. In particular, measuring TF activities enhances our understanding of previously known expression patterns. Our inferred networks suggest that TAL1, PU.1, and c–Myb are key cooperators with GATA–1, but the precise mechanism by which differentiating K562 cells activate and/or degrade GATA–1 remains to be determined. Recent studies suggests that acetylation may play a key role by promoting GATA–1 transactivation, while also enhancing its degradation by the proteasome [49]. The fact that independent reporters for GATA–1 and TAL1 both exhibited reduced activity after PMA addition suggests that the activity of GATA–1 + TAL1–containing complexes may likewise be reduced. Under PMA treatment, we also observed a steady decrease in the activity of TAL1, which is known to form transcriptional activating complexes with GATA–1 [50]. Conversely, I observed a late

upregulation of TAL1 activity during hemin–mediated differentiation of K562 cells, which is consistent with its reported requirement in terminal erythroid differentiation [50].

Unlike GATA–1, GATA–2 activity transiently increased during the first three days of hemin treatment and recovered to basal level, while there was no significant change under PMA treatment (Fig. 2.2A). Our analysis reveals interesting insights for GATA–2 function in the differentiation of K562 cells and offers an explanation to inconsistencies observed in the literature. These results that GATA–2 activity initially increases and then declines after 4–5 days of hemin treatment, support the idea that ectopic GATA–2 promotes E progenitor cell proliferation, but interferes with their terminal differentiation, and that GATA–2 levels decline during the later stages of erythroid differentiation [51].

We anticipated a decrease in FLI–1 activity in response to hemin, since FLI–1 antagonizes commitment of MEPs to the erythroid lineage in favor of MK lineage [52]. However, hemin did not alter basal FLI–1 activity. Since K562 cells already appear to exhibit a bias toward the erythroid lineage (e.g., express the erythroid–specific antigens CD71 and GlyA), this suggests that K562 cells may have already down–regulated FLI–1 expression and/or activity prior to treatment.

TF analysis also provides interesting insights into the activity of NF–E2. In particular, I found that NF–E2 is strongly and rapidly activated by the addition of hemin or PMA. Furthermore, this activation does not depend on the presence of GATA–1 (Fig. 2.8) or the ability to differentiate to the E and MK lineages as normally occurs in K562 cells. This may indicate that NF–E2 activation is not far downstream in the intracellular signaling that occurs upon K562 cell stimulation with either hemin or PMA.

The dynamic nature of the TF activity profiling assay provides the ability to capture how TF networks functionally rewire during differentiation. Novel and known characteristic

relationships were inferred for PMA–and hemin–mediated differentiation, which illustrates changes in TF regulation during MK and E differentiation. Activities of some TFs were consistent with and expected from previous reports, which verifies the dynamic TF activity analysis. A tree–ensemble–based inference method was employed to create characteristic networks describing active regulatory interactions, and this highlighted several functional relationships previously identified in the literature. In addition, my colleagues created a GATA–1 silenced model and verified GATA–1 as an early, essential regulator for both E and MK differentiation in K562 cells, as indicated by the network analysis and previous reports. Directed edges represent causal influences, but such predicted influences may be indirect and not be mediated by a direct binding relationship. The causal influence (assume inferred activation) of TF B by TF A could be explained by TF A binding to the promoter sequence of the gene controlling the transcription of TF B. This interaction could also be explained by a more complicated process, such as TF A binding to a gene that encodes a metabolic enzyme producing a metabolite which in turn regulates the transcription of gene B. These detailed biochemical events are hidden in the observed set of variables. Although likely biological mechanisms have been described for known TF interactions, further experimental verification is needed to uncover the biochemical mechanisms involved in novel relationships.

2.3.7. Methods

2.3.7.1. Cell Culture, Differentiation, and Transduction of Cell Lines. K562 cells were maintained in exponential growth in RPMI 1640 media, supplemented with 10% fetal bovine serum (FBS) from Hyclone (Logan, UT) and 1% penicillin/streptomycin (pen/strep). CHRF cells were maintained in Iscoves Modified Dulbeccos Medium (IMDM) with 10% FBS

and 1% pen/strep. The MK differentiation of either cell line was stimulated by the addition of 10 ng/mL PMA. For E differentiation, K562 cells were treated with 30 μM hemin.

Phorbol 12–myristate 13–acetate (PMA), hemin, benzidine dihydrochloride, and nicotinamide were obtained from Sigma–Aldrich (St. Louis, MO). CD41a and GlyA antibodies were from BD Biosciences (San Jose, CA). Annexin V antibody was from ebiosciences (San Diego, CA).

Lentivirus was produced by co–transfecting HEK–293T cells with previously described lentiviral packaging vectors (pMDL–GagPol, pRSV–Rev, pIVS–VSV–G) (Dull et al., 1998) and lentiviral vectors such as pLenti–TRE–dsGFP–ffluc and GATA–1–directed TRC lentiviral short hairpin RNA vectors (Open Biosystems, Huntsville, AL) using Lipofectamine 2000 (Life Technologies, Carlsbad, CA). After 48 h, supernatants were collected and cell debris was spun down and removed. Viruses were concentrated using PEG–it (Systems Biosciences, Mountain View, CA) and re–suspended in phosphate buffered saline (PBS). Lentivirus titers were determined by HIV–1 p24 Antigen ELISA Kit (ZeptoMetrix Co., Buffalo, NY).

Transduction of GATA–1–targetting lentiviral shRNA into K562 cells was performed by spinoculation. Two days after transduction, cells were cultured with media containing 2 μg/mL puromycin (Life Technologies) for 7 days. Cells were then cultured in regular media for 3 days before being used for experiments. GATA–1 down–regulation was confirmed by real–time PCR.

2.3.7.2. Transcription Factor Activity Reporter. TF reporters consist of a specific TF response element (TRE) cloned upstream of a minimal cytomegalovirus (CMV) promoter (TA) driving the gene for firefly luciferase (FLUC) and destabilized GFP packaged in self–inactivating lentiviral vectors (pGreenFire, System Biosciences). Increased binding on TRE by TFs results in increased luciferase production and a proportional increase in

luminescence when an excess of substrate is added during imaging, thus providing a quantitative measure of relative transactivation. TF reporter specificity and sensitivity studies are referenced on the TRANSFAC [53] and Panomics database (Affymetrix, Redwood, CA). TF reporters were prepared by cloning specific binding elements into the pGreenFire lentiviral backbone. Each lentiviral reporter consists of three repeats of a TF–specific binding element driving expression of FLUC, and a puromycin resistance cassette. K562 or CHRF cells were mixed with lentiviral vectors bearing TF reporter constructs at a multiplicitiy of infection (MOI) of approximately 10 virions per cell and centrifuged at 800g for 45 min at 32 C. After removing the supernatant, cell pellets were resuspended and treated with medium containing 1–2 μg/mL puromycin to select transduced cells. K562 cells bearing reporter vectors were plated at 2 104/well in black 96–well plates (Greiner Bio–One, Monroe, NC) and treated with hemin (30 μM) or PMA (10 ng/mL) to induce E or MK differentiation, respectively. CHRF cells bearing reporter vectors were plated at 1 104/well and treated with PMA (10 ng/mL) or PMA + NIC (12.5 mM) to induce MK differentiation. To measure TF–activity–dependent luciferase production, d–luciferin (Molecular Imaging Products, Bend, OR) was added to wells to a final concentration of 1 mM, which had been previously determined to be well in excess of a limiting concentration. Following a 20–min equilibration period, luminescence in each well was measured using an IVIS Lumina LTE camera system (Caliper Life Sciences, Hopkinton, MA). Untransduced cells in arrays served as controls for non–enzymatic d–luciferin breakdown. Cells transduced with a minimal CMV–FLUC (denoted TA–FLUC) reporter construct without additional TF response elements served as controls for any differences in basal promoter activity between conditions. The media and the inducing agent (hemin or PMA) were exchanged for fresh media containing the inducing agent every other day.

2.3.7.3. TF Activity Pre–processing. Luminescence values for each well on each day were divided by the average of three luminescence readings from corresponding TA–FLUC control wells to control for differences in basal TA promoter activity. Luminescence read–outs for a reporter TF(J)–r in cells treated with treatment Tx on day Dx after adjustment for basal transactivation from control reporter TA–r is therefore represented by the formula.

Normalized luminescence values for each treatment on each day were then divided by the average of the normalized values for the respective untreated control (Veh) to correct for TF activity changes due to continued cell growth in arrays that cannot be attributed to differentiation.

Each TF activity was subsequently log–transformed to normalize the variance of TF(J)–r. Each array had three replicates per TF reporter and complete array experiments were repeated two times on different days. Plate position of cells expressing each TF reporter was varied between experiments beginning on different days.

2.3.7.4. Random Forest Inference. The GENIE3 algorithm was used to generate a network model for each set of time–points[30]. For the random forest (RF) algorithm, parameters were set using the following criteria: K (the number of features selected at random to generate each regression tree) was set to the number of TFs minus one. Ntrees (the total number of trees generated for the ensemble) was set to 1,000. The sign of the interaction (activating or inhibiting relationship) was determined by the sign of the correlation coefficient between the putative regulator and target TFs.

GENIE3 receives a TF activity matrix as an input, and outputs a ranked list of edges and importance scores associated with each edge. For the confidence estimation procedure, each TF importance score was compared to a randomized score from a null model obtained by using internal sampling (randomly shuffling initial activity values by 10,000 iterations).

By randomly shuffling the data, any association between TFs as predicted by the algorithm is removed. If a predicted interaction is observed in 4.99% of the null model predictions, then a P–value of 0.0499 is implied. The importance score is plotted against the number of false positives, and an importance score cut–off of 0.13 corresponding to a P–value of 0.1 was set for screening purposes.

In the absence of a reference hematopoietic TF network, I derived a network of experimentally validated regulatory interactions from GeneGO. Direct interactions were downloaded from the GeneGO database, yielding a network of 34 interactions among seven TFs. The resulting networks from both GeneGO and RF inference were visualized by Cytoscape [54].

Results of experiments are presented as the mean standard deviation, unless otherwise indicated. Analysis of Living Cell Arrays (ALCA), an R package that was previously developed and I subsequently worked to improve specifically for TF activity arrays, was used to visualize and analyze the data [40]. Unless noted otherwise differences in means were evaluated by a paired moderated t–test using false discovery rate correction [55, 56]. A P ¡0.05 was considered to be statistically significant. I performed leave–one–out (LOO) cross validation on our network models to evaluate network sensitivity with respect to the presence/absence of transcription factors.

Transcription factor regulatory networks were generated by aggregating the most likely interactions identified by GENIE3, a random–forest algorithm. I performed a leave–one–out (LOO) cross validation to explore whether interactions were substantially impacted when an individual TF was removed from the analysis. To do this, seven transcription factor regulatory networks were generated by sequentially omitting data corresponding to one of the seven observed transcription factors. For example, to identify possible regulators of the

transcription factor cMyb, one regression tree was created with data from 5 of the other available transcription factors (e.g., FLI–1, GATA1, GATA2, NF–E2, PU.1, leaving out TAL1) instead of 6. For each target transcription factor, the importance score of each candidate regulator was identified and ranked to generate a list of interactions [30]. Only interactions that were identified as significant were shown in networks such as Figure 2B. Thus, I systematically generated 7 LOO networks with a different transcription factor omitted from the analysis. To identify significant interactions within these networks, I used the permutation test procedure to generate null models [57, 58]. In the permutation test, the data is shuffled and random networks are generated, and this was performed 10,000 times. For these random networks, I determined the number of occurrences that an interaction had a similar or higher rank than what was obtained with the original data set. The number of occurrences divided by the total number of random network permutations (10,000 in our studies) reflects the probability (p–value) that an interaction would be a false positive and thus lead to inappropriately rejecting the null hypothesis. I utilized a maximum p–value of 0.1 to identify significant interactions in the leave–one–out models. This cut–off value was used as it corresponds approximately to the elbow when the importance score is plotted against the false positive rate, and thus further increasing the p–value cut–off would significantly increase the false positive rate.

2.4. Transcription factor arrays and application to identification of critical secreted factors mediating metastatic cell homing

Metastatic cell homing is a complex process mediated in part by diffusible factors secreted from immune cells found at a pre–metastatic niche. I discuss another application and analysis pipeline for the TF array to identify functional paracrine interactions between immune cells

and metastatic cells as novel mediators of homing. Metastatic breast cancer mouse models were used to generate a diseased splenocyte conditioned media (D–SCM) containing immune cell secreted factors. MDA–MB 231 metastatic cell activity including cell invasion, migration, transendothelial migration, and proliferation were increased in D–SCM relative to control media. D–SCM secretome analysis yielded 144 secreted factor candidates that contribute to increased metastatic cell activity. The functional mediators of homing were identified using MetaCore software to determine interactions between the immune cell secretome and the TRACER–identified active transcription factors within metastatic cells. Among the 5 candidate homing factors identified, haptoglobin was selected and validated *in vitro* and *in vivo* as a key mediator of homing. I demonstrate a novel systems biology approach to identify functional signaling factors associated with a cellular phenotype, which provides an enabling tool to complement large–scale protein identification provided by proteomics.

2.4.1. Bridging the gap between heterogeneous data–types reveals critical factors for metastatic cell homing

My colleague, Brian Aguado, stimulated MDA–MB 231 breast tumor cells using a splenocyte conditioned media (SCM) containing a complex mixture of immune cell secreted factors and induced phenotypic changes in metastatic cell activity. Using a secretomics approach, the immune cell secretome was analyzed to identify the secreted factors involved in activating the phenotypic changes in cancer cells. In parallel, I used TF array data to identify active transcription factors (TFs) involved with the increased MDA–MB 231 metastatic activity in response to the secreted factors. Upon connecting the two data sets, the generated network connected the SCM secreted factors to the activated TFs. The network was utilized

to identify functional secreted factors that contribute to metastatic cell homing. One candidate secreted factor, haptoglobin, was validated *in vitro* and *in vivo* to confirm its role in metastatic cell homing.

2.4.2. Isolation of secreted factors induces phenotypic changes in MDA–MB 231 cell lines

Leukocytes were harvested from spleens of diseased mice (inoculated with breast cancer cells) and healthy mice (not inoculated with breast cancer cells), which are referred to as diseased and healthy spleens, respectively. Splenocyte conditioned media from healthy (H–SCM) and diseased (D–SCM) splenocyte populations were generated. Differences in metastatic cell activity in H–SCM and D–SCM were evaluated using transwell culture assays [59]. Representative images from the transwell assays are provided in the publication, which had increased MDA–MB 231 invasion, migration, and transendothelial migration when cultured in D–SCM compared to H–SCM and RPMI controls [59].

2.4.3. Merging proteomics data and TF activity data yields candidate homing targets

Secretomics techniques were employed to identify the candidate immune cell secreted factors in D–SCM that increase metastatic activity of MDA–MB 231 cells *in vitro*. A total of 615 proteins were identified in both D–SCM and H–SCM, with 101 proteins identified exclusively in D–SCM and 139 proteins identified exclusively in H–SCM. Out of the 375 proteins identified in both media, 115 of those proteins were identified ontologically as secreted factors (Fig. 2A). From this secreted factor pool, 23 proteins in D–SCM and 16 proteins in H–SCM had a log2 fold change greater than 1.5, indicating increased protein abundance in the sample

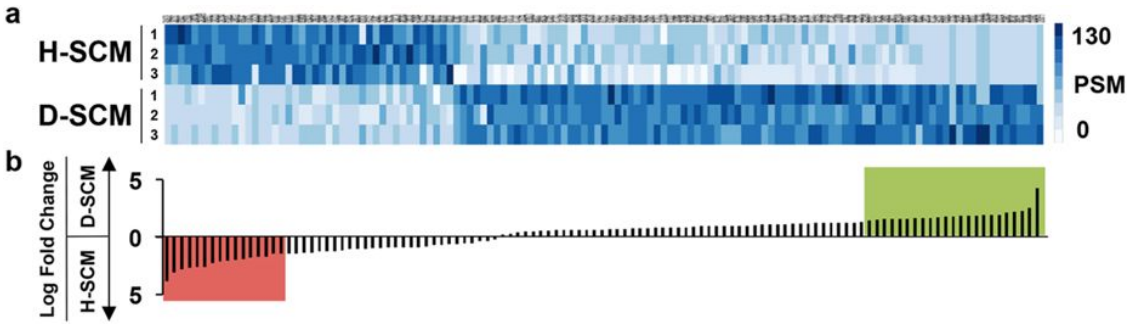

Figure 2.10 | **Secretome analysis of SCM.** (a) Heat–map indicating peptide spectral matches (PSM) of 115 identified secreted factors for three replicates of H–SCM and D–SCM. (b) Log–fold change of corresponding secreted factor peptide hits. The green shaded region indicates secreted factors with a log–fold change greater than 1.5 in D–SCM and the red shaded region indicates secreted factors with a log–fold change greater than 1.5 in H–SCM.

(Fig. 2.10B. In addition to the 115 secreted factors identified in both media, 29 of the 101 proteins exclusively present in D–SCM were identified as secreted factors and included in the secreted factor list, bringing the total to 144 secreted factors in D–SCM (Fig. 2.11).

The transcription factors (TFs) activated in response to D–SCM were subsequently measured using TF arrays. The transactivation profiles of 52 TF reporter constructs over a time period of 8hours were determined by measuring TF activity of cells cultured in D–SCM (Fig. 2.13). Of the 52 TF reporters, 35 reporters had significantly altered TF activity (adjusted p–value¡0.05) for cells cultured in D–SCM. Using k–means clustering, TF activity profiles were grouped into 7 clusters, revealing TFs with similar temporal activation over the 8–hour period (Fig. 2.12A). The 7 clusters were organized from most to least active, allowing visualization of TF clusters that are most active in response to D–SCM secreted factors (Fig. 2.12B). The cluster with the greatest increase in reporter activation contained 10 TFs involved in metastatic cell processes including migration, proliferation, and invasion (Fig. 2.14).

Accession	Description	MW [kDa]	Length	#PSM - D-SCM 1	#PSM - D-SCM 2	#PSM - D-SCM 3	Average D-SCM	#PSM - H-SCM 1	#PSM - H-SCM 2	#PSM - H-SCM 3	Average H-SCM	Log 2 Fold Change	P-Value
P11247	Myeloperoxidase OS=Mus musculus GN=Mpo PE=2 SV=2 - [PERM_MOUSE]	81.1	718	37	28	52	39.00	1	1	4	2.00	4.3	0.026
Q9ET01	Glycogen phosphorylase, liver form OS=Mus musculus GN=Pygl PE=1 SV=4 - [PYGL_MOUSE]	97.4	850	44	48	49	47.00	8	10	6	8.00	2.6	0.000
P05201	Aspartate aminotransferase, cytoplasmic OS=Mus musculus GN=Got1 PE=1 SV=3 - [AATC_MOUSE]	46.2	413	3	8	8	6.33	1	2	1	1.33	2.2	0.042
Q9D154	Leukocyte elastase inhibitor A OS=Mus musculus GN=Serpinb1a PE=1 SV=1 - [ILEUA_MOUSE]	42.5	379	67	65	65	65.67		18	11	14.50	2.2	0.000
P30681	High mobility group protein B2 OS=Mus musculus GN=Hmgb2 PE=1 SV=3 - [HMGB2_MOUSE]	24.1	210	33	42	33	36.00	3	9	13	8.33	2.1	0.003
E9Q3W4	Plectin OS=Mus musculus GN=Plec PE=4 SV=1 - [E9Q3W4_MOUSE]	498.8	4386	9	9	9	9.00	1	1	5	2.33	1.9	0.007
Q9R111	Guanine deaminase OS=Mus musculus GN=Gda PE=1 SV=1 - [GUAD_MOUSE]	51.0	454	25	24	30	26.33		9	5	7.00	1.9	0.006
O88593	Peptidoglycan recognition protein 1 OS=Mus musculus GN=Pglyrp1 PE=2 SV=1 - [PGRP1_MOUSE]	20.5	182	18	16	22	18.67	4		6	5.00	1.9	0.011
P41245	Matrix metalloproteinase-9 OS=Mus musculus GN=Mmp9 PE=2 SV=2 - [MMP9_MOUSE]	80.5	730	17	22	16	18.33		4	6	5.00	1.9	0.013
P29351	Tyrosine-protein phosphatase non-receptor type 6 OS=Mus musculus GN=Ptpn6 PE=1 SV=2 - [PTN6_MOUSE]	67.5	595	7	4	7	6.00	1	3	1	1.67	1.8	0.023
Q9QUH0	Glutaredoxin-1 OS=Mus musculus GN=Glrx PE=1 SV=3 - [GLRX1_MOUSE]	11.9	107	8	4	6	6.00	1	3	1	1.67	1.8	0.031
D3Z2P2	Chitinase-3-like protein 1 OS=Mus musculus GN=Chi3l1 PE=2 SV=1 - [D3Z2P2_MOUSE]	42.8	389	19	16	14	16.33	4	4	6	4.67	1.8	0.002
Q91Z98	Chitinase-3-like protein 4 OS=Mus musculus GN=Chi3l4 PE=1 SV=2 - [CH3L4_MOUSE]	44.9	402	19	19	20	19.33	1	15	1	5.67	1.8	0.043
Q00612	Glucose-6-phosphate 1-dehydrogenase X OS=Mus musculus GN=G6pdx PE=1 SV=3 - [G6PD1_MOUSE]	59.2	515	44	37	50	43.67	11	15	13	13.00	1.7	0.001
Q3TRM8	Hexokinase-3 OS=Mus musculus GN=Hk3 PE=1 SV=2 - [HXK3_MOUSE]	100.0	922	17	12	25	18.00	9	5	3	5.67	1.7	0.042
P31786	Acyl-CoA-binding protein OS=Mus musculus GN=Dbi PE=1 SV=2 - [ACBP_MOUSE]	10.0	87	6	6	10	7.33	1	5	1	2.33	1.7	0.057
Q91YR9	Prostaglandin reductase 1 OS=Mus musculus GN=Ptgr1 PE=2 SV=2 - [PTGR1_MOUSE]	35.5	329	14	12	16	14.00	3		6	4.50	1.6	0.015
A3KGU5	Spectrin alpha chain, non-erythrocytic 1 OS=Mus musculus GN=Sptan1 PE=2 SV=1 - [A3KGU5_MOUSE]	282.7	2457	10	7	6	7.67	2		3	2.50	1.6	0.048
P01027	Complement C3 OS=Mus musculus GN=C3 PE=1 SV=3 - [CO3_MOUSE]	186.4	1663	27	17	26	23.33	5	11	7	7.67	1.6	0.013
Q61166	Microtubule-associated protein RP/EB 1 OS=Mus musculus GN=Mapre1 PE=1 SV=3 - [MARE1_MOUSE]	30.0	268	7	7	10	8.00	1	1	6	2.67	1.6	0.052
P24527	Leukotriene A-4 hydrolase OS=Mus musculus GN=Lta4h PE=1 SV=4 - [LKHA4_MOUSE]	69.0	611	51	54	68	57.67	15	22	21	19.33	1.6	0.003
Q61646	Haptoglobin OS=Mus musculus GN=Hp PE=1 SV=1 - [HPT_MOUSE]	38.7	347	22	19	23	21.33		6	9	7.50	1.5	0.005
Q9DCD0	6-phosphogluconate dehydrogenase, decarboxylating OS=Mus musculus GN=Pgd PE=2 SV=3 - [6PGD_MOUSE]	53.2	483	53	43	53	49.67	9	23	22	18.00	1.5	0.005
P08071	Lactotransferrin OS=Mus musculus GN=Ltf PE=2 SV=4 - [TRFL_MOUSE]	77.8	707	125	128	129	127.33	22	58	75	51.67	1.3	0.008
Q921M7	Protein FAM49B OS=Mus musculus GN=Fam49b PE=2 SV=1 - [FA49B_MOUSE]	36.8	324	7	6	11	8.00	2	5	3	3.33	1.3	0.057
Q922B2	Aspartate--tRNA ligase, cytoplasmic OS=Mus musculus GN=Dars PE=2 SV=2 - [SYDC_MOUSE]	57.1	501	3	4	5	4.00	1	1	3	1.67	1.3	0.057
P06745	Glucose-6-phosphate isomerase OS=Mus musculus GN=Gpi PE=1 SV=4 - [G6PI_MOUSE]	62.7	558	74	76	76	75.33	27	31	38	32.00	1.2	0.000
P10649	Glutathione S-transferase Mu 1 OS=Mus musculus GN=Gstm1 PE=1 SV=2 - [GSTM1_MOUSE]	26.0	218	27	22	28	25.67	8	11	14	11.00	1.2	0.004
Q60865	Caprin-1 OS=Mus musculus GN=Caprin1 PE=1 SV=2 - [CAPR1_MOUSE]	78.1	707	5	4	5	4.67	1	4	1	2.00	1.2	0.065
P07356	Annexin A2 OS=Mus musculus GN=Anxa2 PE=1 SV=2 - [ANXA2_MOUSE]	38.7	339	13	22	16	17.00	6	10	6	7.33	1.2	0.031
Q9R0P5	Destrin OS=Mus musculus GN=Dstn PE=1 SV=3 - [DEST_MOUSE]	18.5	165	6	8	10	8.00	3		4	3.50	1.2	0.061
P05064	Fructose-bisphosphate aldolase A OS=Mus musculus GN=Aldoa PE=1 SV=2 - [ALDOA_MOUSE]	39.3	364	21	33	33	29.00	11	15	13	13.00	1.2	0.018
P47791	Glutathione reductase, mitochondrial OS=Mus musculus GN=Gsr PE=2 SV=3 - [GSHR_MOUSE]	53.6	500	15	15	19	16.33	11	5	6	7.33	1.2	0.017
Q9R1P1	Proteasome subunit beta type-3 OS=Mus musculus GN=Psmb3 PE=1 SV=1 - [PSB3_MOUSE]	22.9	205	10	9	11	10.00	5	5	4	4.67	1.1	0.001
P52480	Pyruvate kinase PKM OS=Mus musculus GN=Pkm PE=1 SV=4 - [KPYM_MOUSE]	57.8	531	36	36	39	37.00	8	20	24	17.33	1.1	0.016
P63028	Translationally-controlled tumor protein OS=Mus musculus GN=Tpt1 PE=1 SV=1 - [TCTP_MOUSE]	19.4	172	14	19	16	16.33	6	7	10	7.67	1.1	0.010
Q920P5	Twinfilin-2 OS=Mus musculus GN=Twf2 PE=1 SV=1 - [TWF2_MOUSE]	39.4	349	6	6	7	6.33		3	3	3.00	1.1	0.004
Q61599	Rho GDP-dissociation inhibitor 2 OS=Mus musculus GN=Arhgdib PE=1 SV=3 - [GDIR2_MOUSE]	22.8	200	31	39	35	35.00	11	21	19	17.00	1.0	0.009
Q60864	Stress-induced-phosphoprotein 1 OS=Mus musculus GN=Stip1 PE=1 SV=1 - [STIP1_MOUSE]	62.5	543	9	8	10	9.00		4	5	4.50	1.0	0.012
P13020	Gelsolin OS=Mus musculus GN=Gsn PE=1 SV=3 - [GELS_MOUSE]	85.9	780	10	9	11	10.00		6	4	5.00	1.0	0.018
G3UVV4	Hexokinase 1, isoform CRA_f OS=Mus musculus GN=Hk1 PE=3 SV=1 - [G3UVV4_MOUSE]	101.8	917	6	7	8	7.00	3	4		3.50	1.0	0.025
P11672	Neutrophil gelatinase-associated lipocalin OS=Mus musculus GN=Lcn2 PE=1 SV=1 - [NGAL_MOUSE]	22.9	200	26	20	20	22.00	5	13	15	11.00	1.0	0.160
Q9JKF1	Ras GTPase-activating-like protein IQGAP1 OS=Mus musculus GN=Iqgap1 PE=1 SV=2 - [IQGA1_MOUSE]	188.6	1657	24	22	31	25.67	9	17	13	13.00	1.0	0.024
P40142	Transketolase OS=Mus musculus GN=Tkt PE=1 SV=1 - [TKT_MOUSE]	67.6	623	85	86	93	88.00	34	53	49	45.33	1.0	0.002
O89053	Coronin-1A OS=Mus musculus GN=Coro1a PE=1 SV=5 - [COR1A_MOUSE]	51.0	461	36	35	31	34.00	14	17	22	17.67	0.9	0.004
B7FAU9	Filamin, alpha OS=Mus musculus GN=Flna PE=4 SV=1 - [B7FAU9_MOUSE]	280.3	2639	96	93	98	95.67	28	58	65	50.33	0.9	0.017
P08905	Lysozyme C-2 OS=Mus musculus GN=Lyz2 PE=1 SV=2 - [LYZ2_MOUSE]	16.7	148	20	18	25	21.00	11	11	13	11.67	0.8	0.013
Q6DBJ1	Phosphoglucerate mutase 1 OS=Mus musculus GN=Pgam1 PE=1 SV=3 - [PGAM1_MOUSE]	28.8	254	28	39	32	33.00	14	18	23	18.33	0.8	0.024
P10810	Monocyte differentiation antigen CD14 OS=Mus musculus GN=Cd14 PE=1 SV=1 - [CD14_MOUSE]	39.2	366	10	10	14	11.33	6	7	6	6.33	0.8	0.022
P31725	Protein S100-A9 OS=Mus musculus GN=S100a9 PE=1 SV=3 - [S10A9_MOUSE]	13.0	113	44	32	31	35.67	14	25	21	20.00	0.8	0.041
P14211	Calreticulin OS=Mus musculus GN=Calr PE=1 SV=1 - [CALR_MOUSE]	48.0	416	21	28	27	25.33	20	10	13	14.33	0.8	0.040
Q9CQ60	6-phosphogluconolactonase OS=Mus musculus GN=Pgls PE=2 SV=1 - [6PGL_MOUSE]	27.2	257	15	14	14	14.33	6	11	8	8.33	0.8	0.016
Q9CR16	Peptidyl-prolyl cis-trans isomerase D OS=Mus musculus GN=Ppid PE=1 SV=3 - [PPID_MOUSE]	40.7	370	5	6	7	6.00		3	4	3.50	0.8	0.058
Q9CVB6	Actin-related protein 2/3 complex subunit 2 OS=Mus musculus GN=Arpc2 PE=1 SV=3 - [ARPC2_MOUSE]	34.3	300	29	24	27	26.67	14	22	11	15.67	0.8	0.038
Q00519	Xanthine dehydrogenase/oxidase OS=Mus musculus GN=Xdh PE=1 SV=5 - [XDH_MOUSE]	146.5	1335	17	16	16	16.33	15	8	7	10.00	0.7	0.067
Q91VW3	SH3 domain-binding glutamic protein 3 OS=Mus musculus GN=Sh3bgrl3 PE=1 SV=1 - [SH3L3_MOUSE]	10.5	93	6	7	8	7.00	3	5	5	4.33	0.7	0.039
P17742	Peptidyl-prolyl cis-trans isomerase A OS=Mus musculus GN=Ppia PE=1 SV=2 - [PPIA_MOUSE]	18.0	164	33	31	34	32.67	22	17	22	20.33	0.7	0.003
P27005	Protein S100-A8 OS=Mus musculus GN=S100a8 PE=1 SV=3 - [S10A8_MOUSE]	10.3	89	77	89	84	83.33	32	71	57	53.33	0.6	0.066
P68254	14-3-3 protein theta OS=Mus musculus GN=Ywhaq PE=1 SV=1 - [1433T_MOUSE]	27.8	245	15	13	17	15.00	11	9	9	9.67	0.6	0.016
P68037	Ubiquitin-conjugating enzyme E2 L3 OS=Mus musculus GN=Ube2l3 PE=2 SV=1 - [UB2L3_MOUSE]	17.9	154	6	5	6	5.67	3	5	3	3.67	0.6	0.055
H7BWZ3	Actin-related protein 2/3 complex subunit 3 OS=Mus musculus GN=Arpc3 PE=2 SV=1 - [H7BWZ3_MOUSE]	19.6	170	7	8	8	7.67	6	5	4	5.00	0.6	0.016
O35744	Chitinase-3-like protein 3 OS=Mus musculus GN=Chi3l3 PE=1 SV=2 - [CH3L3_MOUSE]	44.4	398	62	59	68	63.00	33	42	50	41.67	0.6	0.019
P17182	Alpha-enolase OS=Mus musculus GN=Eno1 PE=1 SV=3 - [ENOA_MOUSE]	47.1	434	56	44	48	49.33	27	33	38	32.67	0.6	0.025
P16110	Galectin-3 OS=Mus musculus GN=Lgals3 PE=1 SV=3 - [LEG3_MOUSE]	27.5	264	27	24	23	24.67	10	20	19	16.33	0.6	0.070

Figure 2.11 | **Summary of identified proteins from secretomics analysis of H–SCM and D–SCM.** List of 144 protein matches identified as secreted proteins (115 identified in both D–SCM and H–SCM and 29 identified exclusively in D–SCM), with significance cut–off of p ¡ 0.1. Matched proteins highlighted in green have peptide spectral matches with a log2 fold difference greater than 1.5 in D–SCM. Matched proteins highlighted in red have peptide spectral matches with a log2 fold difference greater than 1.5 in H–SCM.

Candidate homing factors were subsequently identified through the intersection of TFs downstream from D–SCM proteins and TFs identified with TRACER. First, 47 TFs were identified to be downstream of the 144 D–SCM secreted factors using public data sources that curate experimentally verified interactions. MetaCore network analysis software was used to generate a network containing 144 D–SCM secreted factors, all known human receptors, and all known human TFs as nodes. The group of 47 TFs predicted to be downstream of the secreted factors were compared to the 10 TFs within the cluster with the greatest increase in

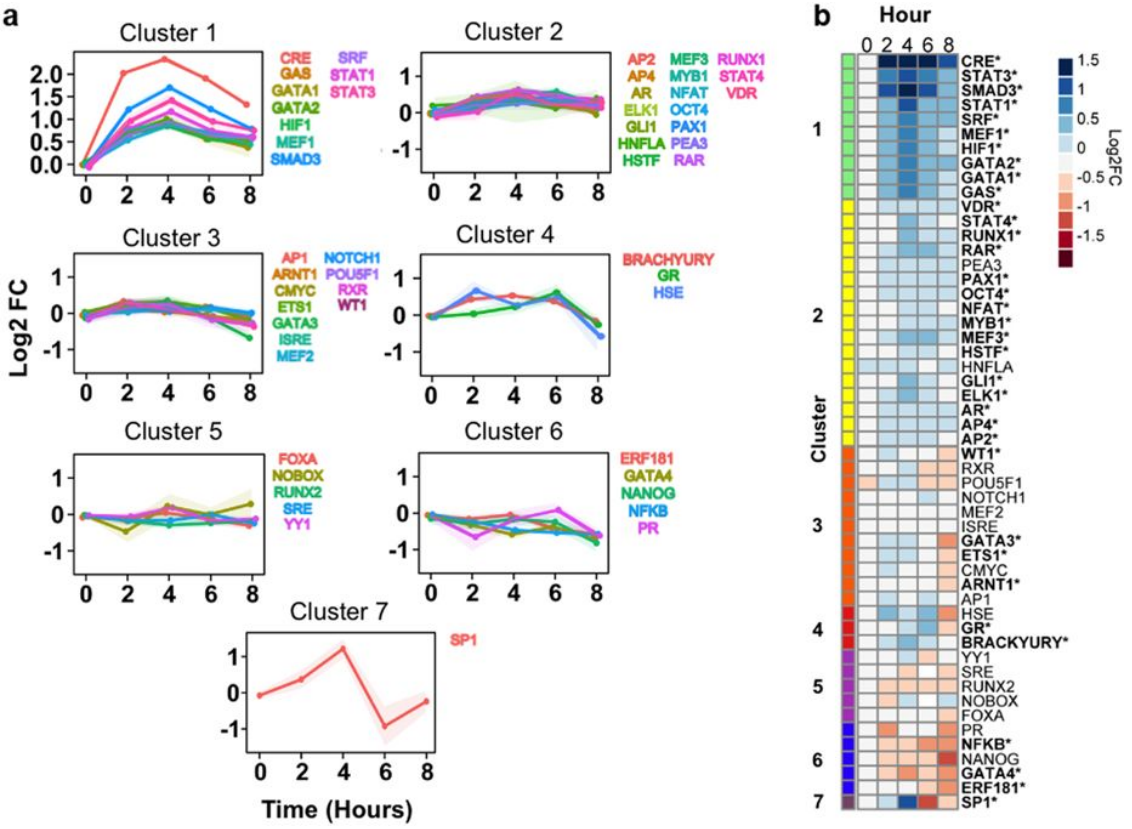

Figure 2.12 | **TF activity of MDA–MB 231 cells cultured in D–SCM.** (a) K–cluster line–graphs for TF reporter activity of MDA–MB 231 cells cultured in D–SCM. (b) Heat map of normalized TF reporter activity values for MDA–MB 231 cells cultured in D–SCM over 8hours. RPMI was chosen as a control for basal TF reporter activity since there was no observable difference in cell phenotype between MDA–MB 231 cells cultured in RPMI or H–SCM. Activity values are organized using k–means into 7 clusters (n=6 arrays). Significance in TF activity for at least one time point indicated with an asterisk (*p¡0.05).

reporter activation identified with TRACER, revealing 6 common TF targets (Fig. 2.15A). We determined that the highly activated cluster of TFs was significantly enriched with TFs predicted to be downstream of 144 secreted factors.

Next, an interaction network was generated to determine functional connections between secreted factors and active TFs. The 144 secreted factors, 35 significantly active TFs, and

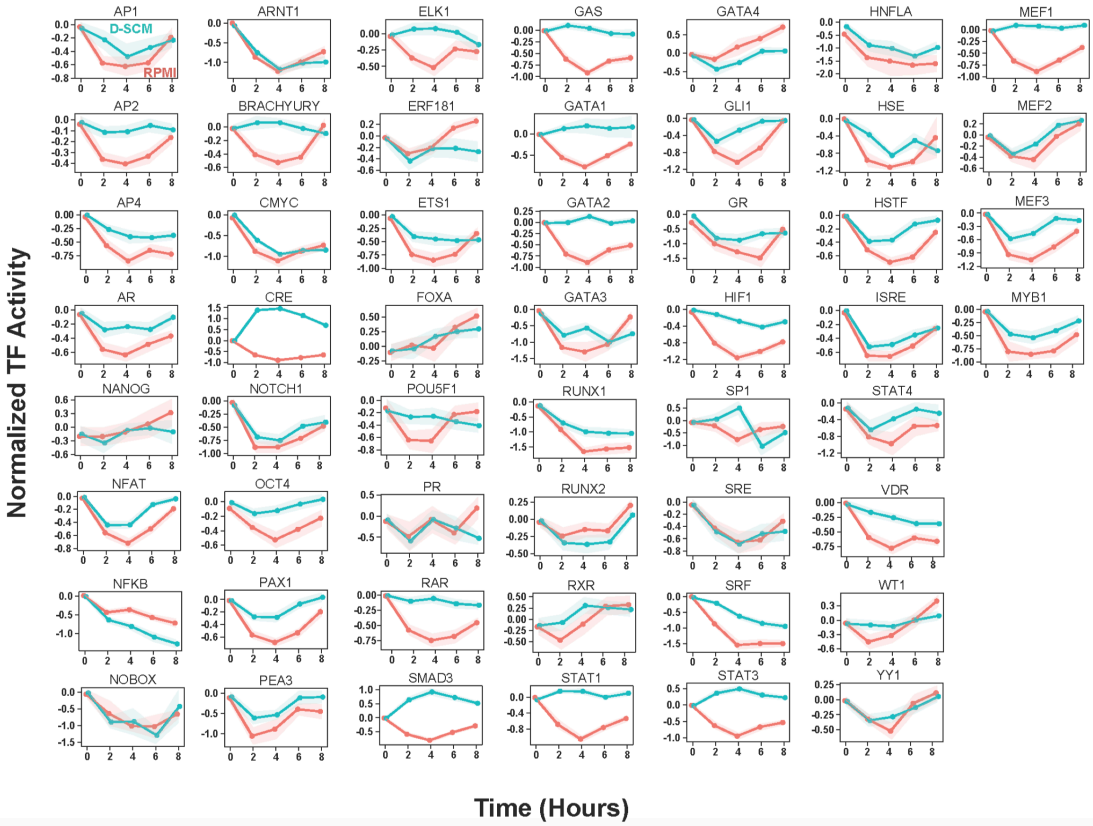

Figure 2.13 | **TF activity of MDA–MB 231 cells measured cultured in D–SCM.** List of 144 protein matches identified as secreted proteins (115 identified in both D–SCM and H–SCM and 29 identified exclusively in D–SCM), with significance cut–off of p ¡ 0.1. Matched proteins highlighted in green have peptide spectral matches with a log2 fold difference greater than 1.5 in D–SCM. Matched proteins highlighted in red have peptide spectral matches with a log2 fold difference greater than 1.5 in H–SCM.

a list of all known human receptors obtained from MetaCore were connected as nodes to generate a network to link our secretomics and TF actvity results (Fig. 2.15B). Edges between nodes represent experimentally verified protein–protein or gene–gene interactions. Additionally, the initial receptor and TF nodes were used as seed nodes and expanded by one degree to include additional signaling components. The final network consisted of 3562 known interactions. The 6 common TF targets were predicted to interact with receptors

Cluster	Reporter	TF Name	Category	Associated GO Processes
1	CRE	cAMP Response Element Binding Protein 1	Canonical Pathways	cAMP response, cell proliferation, migration, invasion
1	STAT3	Signal Transducer and Activator of Transcription 3	Inflammatory Response	Acute phase response, proliferation, cytokine response
1	SMAD3	Smad3	Canonical Pathways	TGFβ pathway response, cell cycle, cell invasion
1	STAT1	Signal Transducer and Activator of Transcription 1	Inflammatory Response	Interferon response, proliferation, cytokine response
1	SRF	Serum Response Factor	Canonical Pathways	MAPK pathway, cell cycle, growth, migration, apoptosis
1	MEF1	Myocyte-specific enhancer factor 1	Differentiation/Development	Mesodermal differentiation, cell adhesion
1	HIF1	Hypoxia Inducible Factor 1A	Differentiation/Development	Hypoxia response, angiogenesis
1	GATA2	Gata binding protein 2	Differentiation/Development	Endothelial; adipocyte differentiation, angiogenesis
1	GATA1	Gata binding protein 1	Differentiation/Development	Hematopoietic differentiation
1	GAS	Interferon-Gamma Activated Sequence	Inflammatory Response	STAT family promoter

Figure 2.14 | **List of significantly active transcription factor reporters in Cluster 1 of D–SCM TF activity screen.** TFs are listed by TF reporter category and associated gene ontology (GO) processes. Sources: TRANSFAC, MetaCore, NCBI databases.

known to respond to 5 secreted factors identified in the network (calgranulin A, calgranulin B, haptoglobin, heme binding protein, and myeloperoxidase) (Fig. 2.15C). By narrowing the list of TFs included in the network with TF activity, we could objectively identify secreted factor candidates that have a downstream effect on the network.

2.4.4. Validation of haptoglobin as a secreted factor that mediates tumor cell recruitment *in vitro*

Among the list of 5 candidate secreted factors, haptoglobin was chosen for validation. In the generated network, haptoglobin interacts with the CCR2 receptor and activates multiple downstream TFs. Next, the role of haptoglobin on the *in vitro* metastatic activity of

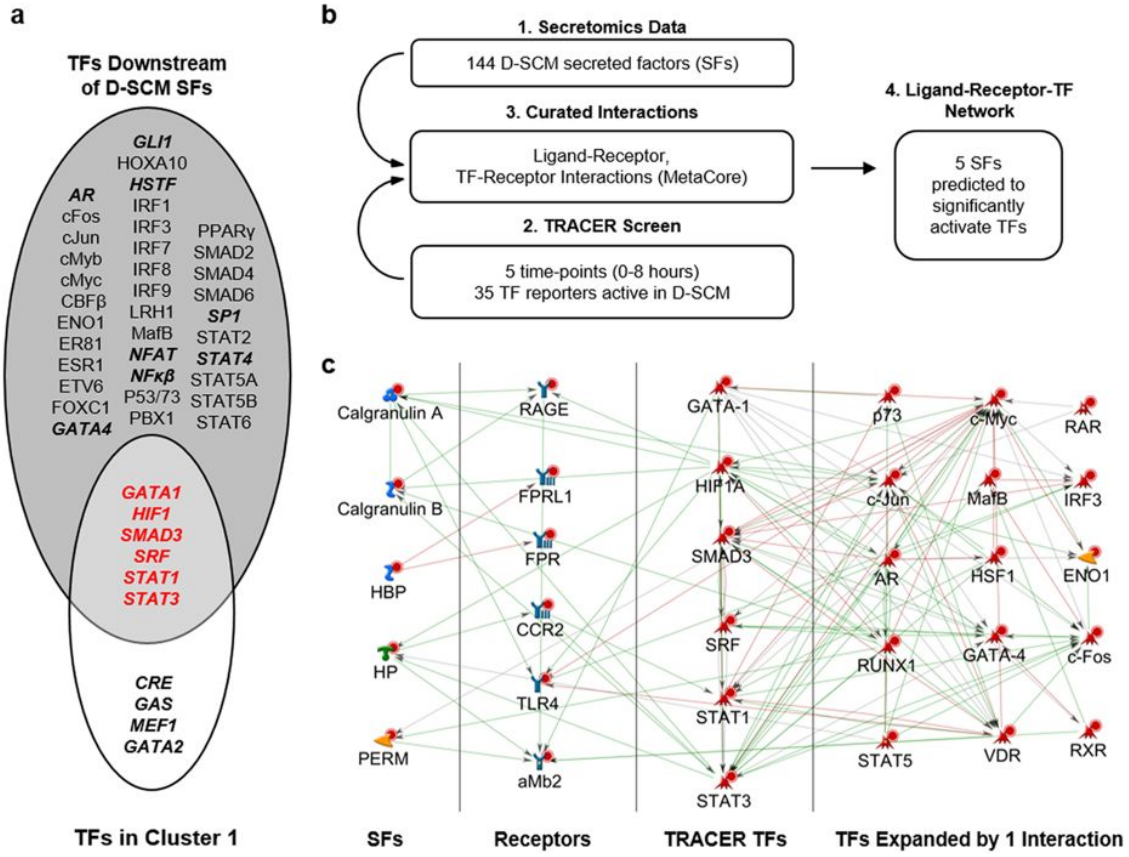

Figure 2.15 | **Identification of secreted factors and transcription factors mediating metastatic cell homing.** (a) Venn diagram summarizing predicted TFs downstream of D–SCM. The white circle identifies the most active TFs (Cluster 1) in the D–SCM TF activity screen, and the grey circle identifies TFs that were predicted to be downstream of the secreted factors (SFs) identified using MetaCore. A Fisher test was used to test the significance of the overlapping TFs in red (p¡0.01). Bolded TFs indicate reporters that were significantly active in the D–SCM TF activity screen. (b) Overview of experimental approach and computational analysis to narrow down secreted factor list of candidates. (c) Network generated in MetaCore summarizing interactions between 5 secreted factor candidates, receptors, and downstream TFs from TF activity in Cluster 1. The network includes TFs expanded by one interaction.

MDA–MB 231 cells was investigated using two approaches: i) RPMI supplemented with recombinant haptoglobin (rHp) relative to control RPMI media, and ii) D–SCM supplemented with a haptoglobin antibody (HpAb) relative to control D–SCM. MDA–MB 231

migration increased almost three–fold to 295.916.9 cells in rHp supplemented medium compared to 100.510.4 cells in control RPMI. Additionally, MDA–MB 231 migration was decreased almost two–fold in D–SCM supplemented with HpAb, with 118.66.2 cells compared to 219.916.9 cells in D–SCM (Shown in the publication [59]). MDA–MB 231 invasion also decreased in D–SCM supplemented with HpAb compared to D–SCM; however, invasion in RPMI supplemented with rHp was similar to control RPMI. The addition of rHp to RPMI and HpAb to D–SCM had no effect on MDA–MB 231 transendothelial migration. MDA–MB 231 cells cultured in rHp had a 1.5–fold increase in proliferation compared to RPMI by Day 5 of culture. From these results, rHp showed differing effects on metastatic cell phenotype, suggesting other factors in the conditioned media contribute to the phenotypic effects.

Given these phenotypic changes in MDA–MB 231 metastatic processes, TF activity arrays were employed to confirm the activity of TFs downstream of haptoglobin identified in the network. Two TRACER arrays were performed to compare TF activity between i) RPMI vs. RPMI+rHp (rHp TRACER) and ii) D–SCM vs. D–SCM+HpAb (HpAb TRACER). Reporters for the rHp and HpAb TF activity arrays were selected based on the TFs located downstream of the Hp–CCR2 interaction in the network. Reporters with increases in TF activity identified by TF actvity for cells cultured in D–SCM were also selected. The transactivation profiles of 16 TF reporters were measured over a period of 8hours. K–means clustering of the TRACER profiles characterized the response to rHp and HpAb treatments into 6 temporally distinct activity profiles (Fig. 2.16C,D). TF activity from 5 of 16 reporters (STAT3, NF, PAX1, CRE, and SRF) correlated with recombinant haptoglobin addition, and displayed increased activity with rHp treatment while having decreased activity with HpAb treatment (Fig. 2.16C). Other reporters deviated from this trend, including SMAD3, SP1, STAT1, and MEF1, suggesting that other secreted factors in D–SCM may contribute to

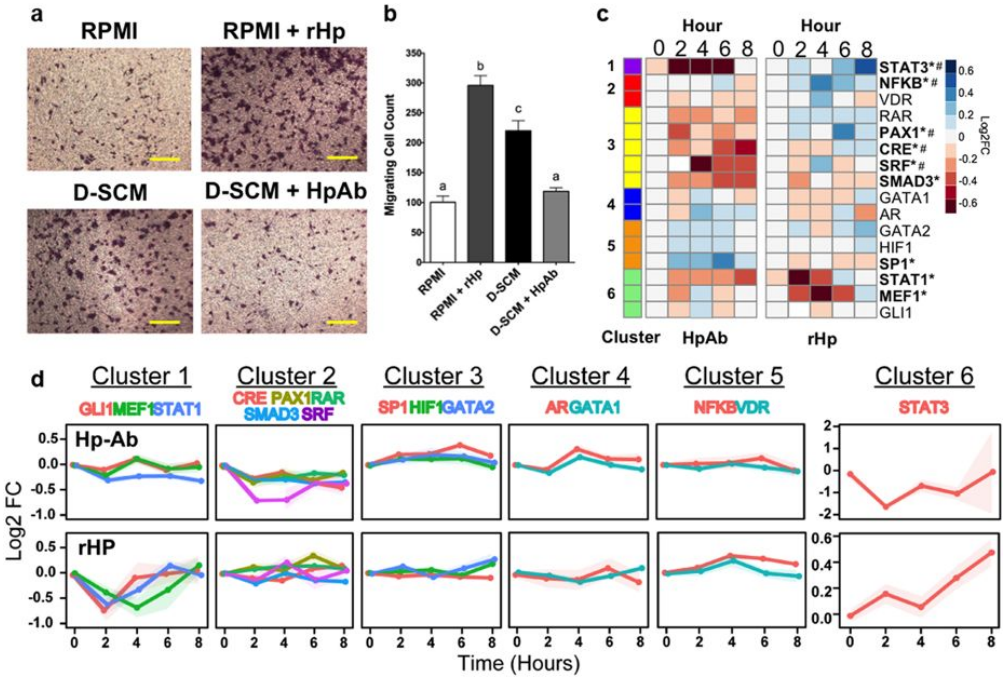

Figure 2.16 | ***in vitro* validation of haptoglobin as a secreted factor mediating MDA–MB 231 migration.** (a) Representative migration assay images of MDA–MB 231 cells cultured in RPMI, RPMI supplemented with recombinant haptoglobin (RPMI+rHp), D–SCM, and D–SCM supplemented with haptoglobin antibody (D–SCM+HpAb) (scale bars=300μm). (b) Migrating cell count of MDA–MB 231 cells cultured with RPMI, RPMI+rHp, D–SCM, and D–SCM+HpAb (n=8). Letters above each data column indicate statistical significance, with different letters signifying distinct statistical groups (p¡0.05). (c) TF activity array data showing MDA–MB 231 TF activity for cells cultured in D–SCM+HpAb normalized to activity for cells cultured in D–SCM, and activity for cells cultured in RPMI+rHp normalized to activity for cells cultured in RPMI (n=6 arrays). Significant changes in TF activity for at least one time point indicated with an asterisk (*p¡0.05). TF reporters showing both increased activity in rHp and decreased activity in HpAb are indicated with a pound sign (#). (d) K–cluster line–graphs comparing TF activity clusters of MDA–MB 231 cells cultured in RPMI vs. rHp and D–SCM vs. D–SCM+HpAb.

altering the activity of these TF reporters. These TRACER results demonstrated that haptoglobin activated multiple TFs associated with increased metastatic activity, and validated key TFs identified as downstream of Hp–CCR2 signaling from the network.

2.4.5. A novel systems biology pipeline combining secretomic and TF data identifies influential factors mediating cell homing

Our novel systems biology approach of connecting secretomics data with functional TF activity data from TRACER allows narrowing of candidate factors to identify functional secreted protein candidates involved in paracrine signaling. Immune cells reflective of the cell types that recruit tumor cells to the pre–metastatic niche were used to generate the D–SCM containing a mixture of functional secreted factors. My colleagues observed phenotypic changes in MDA–MB 231 such as invasion, migration, transendothelial migration, and proliferation in response to the immune cell secreted factors. Secretomics analysis of the media identified 144 candidate factors that mediate the phenotypic changes, and this list was shortened through determining activated TFs within the cells. Using curated factor–receptor interactions to connect secretomics and the most active TFs identified by TF activity, I identified 5 secreted factors that are predicted to induce the observed TF activity activation profiles and MDA–MB 231 phenotype changes. D–SCM TF activity screens identified a cluster of TFs that may be immediately downstream of secreted factors, given increases in activity on a short 8–hour time span. The 6 enriched TFs targets, including GATA1, HIF1, SMAD3, SRF, and STAT1/STAT3 proteins, have been associated with metastasis, and herein, I demonstrate their role in mediating metastatic cell homing. TF activity allowed for a broad evaluation of paracrine signaling between immune cells and metastatic cells, and connecting together secretomics with TF activity data sets enabled the identification of functional secreted factors associated with metastatic cell homing.

My systems biology pipeline allows for the identification of functional paracrine signaling factors within a secretome and is poised to uncover new candidates for targeted therapies against metastatic cell homing. More generally, connecting together secretomics and TF

activity data may be employed to identify secreted factors relevant for paracrine signaling that may underlie a variety of cell phenotypes.

2.4.6. Methods

2.4.6.1. MDA–MB 231 cell culture. The human breast adenocarcinoma cell line MDA MB 231 was used for all *in vitro* experiments. MDA–MB 231 cells were routinely cultured on tissue culture polystyrene flasks in RPMI 1640 media supplemented with 10% fetal bovine serum (FBS), 1% penicillin–streptomycin solution, 1% non–essential amino acids, and 1% sodium pyruvate (Life Technologies). Media was exchanged every other day. Once 80% confluent, cells were harvested with TrypLE Express (Life Technologies) solution and counted using a Trypan blue stain (Sigma Aldrich) and a Cell Countess automated hemocytometer (Life Technologies). All cells were cultured in a humidified 5% CO2 incubator at 37C.

2.4.6.2. Splenocyte–conditioned medium. Splenocyte conditioned media preparation Animal studies were performed in accordance with and approved by the Northwestern University Institutional Animal Care and Use Committee (IACUC). Immunodeficient NOD–scid IL2Rgammanull (NSG) female mice (Jackson Labs) were injected with 2.0106 MDA MB 231BR cells labeled with tdTomato and F–luciferase reporters in the right mammary fat pad.

Spleens were harvested 28 days later from tumor–bearing and tumor–free mice, ballooned using injections of 0.38mg/mL solutions of liberase LT (Roche Diagnostics), and minced using micro–scissors. Minced tissue was incubated at 37C for 20minutes, neutralized with 0.125M EDTA, and processed into cell suspensions using FACS buffer (1X PBS, 0.5% BSA, 2mM EDTA) and a 70μm cell sieve. Cell suspensions were centrifuged using a swinging bucket rotor at 500g for 5minutes at 4C. The supernatant was removed and the cells was

re–suspended in ACK buffer (Life Technologies) for 2minutes, neutralized with PBS and re–centrifuged. The splenocyte pellet was re–suspended in RPMI 1640 and cells were counted using a Trypan blue stain and a Cell Countess automated hemocytometer (Life Technologies). Cells were plated at 106cells/mL of serum free RPMI 1640 media supplemented with 1% penicillin–streptomycin, 1% non–essential amino acids, and 1% sodium pyruvate. All media was conditioned for 48hours, filtered through a 0.22μm filtration unit (Millipore), and stored at 80C until use.

2.4.6.3. Proteomics analysis. Proteins in SCM samples were concentrated for secretomics analysis using a 3kDa Amicon cellulose centrifugal filter unit (Millipore). For each concentrated conditioned media sample (three biological replicates), 5μg of protein was solubilized by adding 8M urea and incubating at 50C for 60min. Following denaturation, proteins were solubilized and reduced by adding 10mM DTT (final concentration 1mM) and incubating at 50C for 15min. After reduction, proteins were alkylated by adding 100mM iodoacetamide (final concentration 10mM) and incubated in the dark at room temperature for 15min. Protein samples were digested by diluting the 8M urea solution to 1M by adding 100mM ammonium bicarbonate and trypsin. Samples were digested at 37C overnight. The digested samples were desalted using reverse phase C18 spin columns (Thermo Fisher Scientific). After desalting, the peptides were concentrated in vaccuo until dry. After drying, peptides were suspended in 5% acetonitrile and 0.1% formic acid. The samples were loaded directly onto a 15cm long, 75μM reversed phase capillary column (ProteoPep II C18, 300, 5μm size, New Objective) and separated using a 200–minute gradient from 5% acetonitrile to 100% acetonitrile on a Proxeon Easy n–LC II (Thermo Scientific). The peptides were eluted into an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific) with electrospray

ionization at a 350nL/minute flow rate. The mass spectrometer was operated in data dependent mode. For each MS1 precursor ion scan, the ten most intense ions were selected for fragmentation by CID (collision induced dissociation). Additional parameters for mass spectrometry analysis included setting the resolution of MS1 at 60,000, the normalized collision energy at 35%, the activation time at 10ms, and the isolation width at 1.5. Charge states +4 and higher were rejected.

The data were processed using Proteome Discoverer (version 1.4, Thermo Scientific) and searched using an embedded SEQUEST HT search engine. The data were searched against a mouse reference proteome (September 2013, uniprot.org). Additional search parameters were as follows: (i) enzyme specificity: trypsin, (ii) fixed modification: cysteine carbamidomethylation, (iii) variable modification: methionine oxidation and N–terminal acetylation, (iv) precursor mass tolerance was10ppm, and (v) fragment ion mass tolerance was0.8Da. All the spectra were searched against target/decoy databases and results were used to estimate the q values with the Percolator algorithm embedded in Proteome discoverer 1.4. The peptide identification was considered valid at q value¡0.1 and were grouped for protein inference to satisfy the rule of parsimony. Further, each protein in the final identification list was considered valid if supported with a minimum of one unique peptide.

Proteins were quantified using spectral counting42 and normalized spectral abundance factors (NSAF)43,44. The NSAF normalization takes into consideration of the length of the protein, which may result into higher spectral count per protein. Initially, the total number of spectral counts (Spc) per protein was divided by the peptide length (L), and then divided by the sum (spc/L) of all the values in the sample. Proteins were determined significantly changed if t–test on a significance level of 90% and if log2 fold change was greater than or

equal to 1.5. The mass spectrometry data have been deposited to the ProteomeXchange Consortium45 via the PRIDE partner repository with the dataset identifier PXD002051.

2.4.6.4. Transcription factor array setup. Cell arrays were performed as previously described [39, 40]. Harvested MDA–MB 231 cells were suspended in RPMI media to a final concentration of 50cells/μL. 400μL of this suspension was aliquoted into separate 1.5mL Eppendorf tubes for viral infection. The aliquot was mixed with lentiviral vectors containing TF reporter constructs16 at a multiplicity of infection (MOI) of approximately 10 virions per cell. Cells and virus were mixed and plated at 2000 cells/well in a black, clear bottom, 384–well plate (Greiner Bio–One). Each TF reporter is represented with n=4 measurements per array plate, and arrays were repeated a total of 6 times. After infection, cells were incubated for 48hours.

To measure TF activity, D–luciferin (DLuc, RR Labs, Inc.) diluted in the appropriate media was added to wells in excess at a final concentration of 2mM. After a 45–minute incubation period with the DLuc, the luminescence was quantified using an IVIS Lumina LTE imaging system (Caliper Life Sciences). Cells plated without virus infection served as negative controls for non–enzymatic DLuc degradation. A positive control consisted of a TA–FLuc reporter construct without any additional TF binding elements, which was used to determine basal promoter activity. Luminescence was quantified using an IVIS Lumina LTE imaging system (Caliper Life Sciences). All luminescence readings, measured in photon flux (photons/second), were normalized to the TA luminescence. On Day 0, cells were treated with either RPMI or D–SCM containing 2mM of DLuc and 10% FBS. Bioluminescence imaging was conducted every 2hours, and 5 reads were taken in one day. Each TF reporter is represented with n=4 measurements per array plate, and arrays were repeated a total of 6 times.

Initial methodology to normalize and determine statistical significance was slightly modified. Array data was log2 transformed and filtered to eliminate all intensities below background (p¡0.05). The background was defined as the mean measured intensity in non–infected cells subject to the same treatment at the same time and plate. At each time–point, the TA control reporter and the control condition were used to normalize reporter activity to calculate the fold–change between D–SCM vs. RPMI, rHp vs. RPMI, and HpAb vs. D–SCM, respectively. Normalized values that were identified to be outliers (p¡0.003) for each reporter were removed.

Normalized log2 TF activity fold–change of SCM, rHp, and HpAb were compared directly to control conditions using the limma package in R [56]. A linear model was fit to the normalized log2 values for each TF and was used to generate estimated coefficients and squared errors for each time point of the compared samples. The estimated coefficients and squared errors were then used to compute moderated t–statistics, moderated F–statistics, and log–odds of differential expression. Adjusted p–values were computed using the Benjamini–Hochberg procedure to correct for multiple comparisons. TFs identified to be differentially active had an adjusted p–value of less than 0.05. To generate heat–maps, the replicate log2 fold–change for each condition and time–point was averaged. Normalized values were then clustered by k–means clustering with random starts. The sum–of–square error was computed for each cluster using the group mean. The optimal number of clusters was determined by maximizing the difference between the sum–of–square error of the computed k–means model and permuted null models.

2.4.6.5. MetaCore analysis. All interaction networks were generated using MetaCore (Thomson Reuters). The network to predict downstream TFs included the 144 identified secreted factors in D–SCM and lists of all known human receptors and transcription factors

obtained from MetaCore. In another network to determine functional interactions, we included the D–SCM secreted factors, a list of all known human receptors, and significantly active TFs from our TRACER screen. Secreted factors candidates were identified as ligands showing direct interactions with a receptor and downstream TFs. The TFs were expanded by one interaction to generate a list of downstream TFs.

2.4.6.6. *in vitro* validation of haptoglobin. Invasion, migration, transendothelial migration, MTS assays, and TRACER arrays were performed as described above. Recombinant haptoglobin (rHp, ProSpec) was diluted in serum–free RPMI at a concentration of 2μg/mL. Haptoglobin antibody (HpAb, Abcam) was added to SCM at a dilution of 1:250, and the solution was incubated on ice for 30min before use. For arrays, rHp and HpAb solutions were supplemented with 2mM of DLuc and added prior to bioluminescence imaging.

CHAPTER 3

Improved methods for temporally-defined generation of transcription factor networks

Work presented in this chapter is adapted from the following paper:

- Finkle J.D*., Wu J.J.*, Bagheri N.B. Windowed Granger Causal Inference Strategy Improves Discovery of Gene Regulatory Networks. Proceedings of the National Academy of Sciences. 2018.

3.1. Abstract

Discovery of gene regulatory networks (GRNs) is crucial for gaining insights into biological processes involved in development or disease. Accurate inference of regulatory networks from experimental data facilitates the rapid characterization and understanding of biological systems. High-throughput technologies can provide a wealth of time-series data to better interrogate the complex regulatory dynamics inherent to organisms, but many network inference strategies do not effectively use temporal information. We address this limitation by introducing Sliding Window Inference for Network Generation (SWING), a generalized framework that incorporates multivariate Granger causality to infer network structure from time-series data. SWING moves beyond existing Granger methods by generating windowed models that simultaneously evaluate multiple upstream regulators at several potential time delays. We demonstrate that SWING elucidates network structure with greater accuracy in both *in silico* and experimentally-validated *in vitro* systems. We estimate the apparent time

delays present in each system and demonstrate that SWING infers time-delayed, gene-gene interactions that are distinct from baseline methods. By providing a temporal framework to infer the underlying directed network topology, SWING generates testable hypotheses for novel gene-gene influences.

3.2. Challenges of gene regulatory network inference

If the experimental sampling interval is less than or equal to the time delay between a regulator and its downstream target, it is possible to employ Granger causality to incorporate intrinsic delays that are often hidden from measurement [60]. Current implementations of Granger causal network inference methods are limited; the inference (*i*) is conducted pairwise, prohibiting simultaneous assessment of multiple upstream regulators, (*ii*) has a single user-defined delay, which assumes a uniform delay between all regulators and their targets, or (*iii*) requires each explanatory variable, assessed at multiple delays, to be selected as a group [61, 62, 63, 64, 65]. Thus, their implementation has limited broad utility in biological systems with heterogeneous time delays.

To allow for multiple time delays to affect downstream target nodes, we introduce an extensible framework to infer GRNs from time-series data, termed Sliding Window Inference for Network Generation (SWING). SWING embeds existing multivariate methods, both linear and nonlinear, into a Granger causal framework that concurrently considers multiple time delays to infer causal regulators for each node. SWING also uses sliding windows to create many sensitive, but noisy, inference models that are aggregated into a more stable and accurate network. We validate the efficacy of SWING on several *in silico* time-series data sets, and existing *in vitro* data sets with corresponding gold standard networks. We show that SWING performs network reconstruction more accurately than baseline methods,

and demonstrate that this performance boost is partly attributed to inferring edges that involve an identifiable time delay between upstream regulators and targets. In validation studies analyzing networks derived from *E. coli* and *S. cerevisiae*, SWING infers networks with distinct topologies, and can therefore be combined with other methods to improve consensus models. The SWING framework is available for use and can be found on GitHub (*https://github.com/bagherilab/SWING*).

3.3. Problem setup for inferring regulatory networks

SWING addresses the challenge of inferring regulatory networks from gene expression data. Gene regulatory networks are directed graphs with N nodes, where each node represents a gene. An edge from gene g_i to gene g_j indicates that g_i regulates the expression of g_j.

3.3.1. Time-series data for biological data is stacked

The time-series measurement of expression for gene, i, with T time points, is defined as $G_i = [g_i^1, g_i^2, \ldots, g_i^T]^\top$. Thus, a time-series experiment is defined as $\boldsymbol{T} = [G_1, \ldots, G_N]$. $\boldsymbol{T}$ is a $T \times N$ matrix which provides an ordered sequence of values for each observed gene (columns) at each time point (rows).

$$\boldsymbol{T} = \begin{bmatrix} g_1^1 & \cdots & g_N^1 \\ \vdots & \ddots & \vdots \\ g_1^T & \cdots & g_N^T \end{bmatrix} \tag{3.1}$$

For simplicity we describe the case where there are no replicates. However, if there are multiple time series, P, of the same length for each gene, such as experiments with multiple

biological replicates or experimental perturbations, they are stacked into a $(T \cdot P) \times N$ matrix such that $\boldsymbol{T} = [\boldsymbol{T}_1, ..., \boldsymbol{T}_p]^\top$.

3.3.2. SWING divides time-series using sliding window

SWING employs a fixed-length sliding window to divide time-series observations into ensembles of training data with the same measured features within each time series.

Given a time-series data set $\boldsymbol{T}$, SWING creates Q consecutive windows. Q is defined as

$$Q = (T - w + 1)/s, \tag{3.2}$$

where w is the window width, such that $w \leq T$, and s is the step size between windows. Both w and s are specified by the user. Each window $\boldsymbol{W}_q$, where $q \in \{1, \ldots, Q\}$, is a subset of rows from the time-series data $\boldsymbol{T}$, such that:

$$\boldsymbol{W_q} = \begin{bmatrix} g_1^{s(q-1)+1} & \cdots & g_N^{s(q-1)+1} \\ g_1^{s(q-1)+2} & \cdots & g_N^{s(q-1)+2} \\ \vdots & \ddots & \vdots \\ g_1^{s(q-1)+w} & \cdots & g_N^{s(q-1)+w} \end{bmatrix} \tag{3.3}$$

If $w = T$ then there is only one window and SWING performs network inference equivalent to the base method. Additional parameters for window creation are described in the *SWING parameter selection*.

3.3.3. Edges are partitioned into several sub-edges defined by minimum and maximum lag

Once the temporal windows are delimited, we apply multivariate Granger causality to generate training sets for inference algorithms. Traditional Granger causality models assess pairwise predictions with a set delay between the variables. Previous methods expanded the Granger models to be multivariate, but do not simultaneously compare multiple delays between explanatory and response variables. Here we describe the formulation of a Granger model that is both multivariate and includes multiple delays.

SWING utilizes a general statistical framework where weights between explanatory variables and a response variable are calculated using supervised learning algorithms. For each window, W_q, we sequentially define a response vector for each gene, j, as $\boldsymbol{y}_j = \boldsymbol{W}_{q,j}$, which is the jth column of window $\boldsymbol{W}_q$. The explanatory data is created based on two user-specified parameters. The maximum lag, k_{max}, and minimum lag, k_{min}, define the number of time points that can exist between the explanatory variables and the response. They are used to define the user-allowed set of delays, $L = \{k_{min}, k_{min}+1, \ldots, k_{max}\}$.

$|L|$ is the cardinality of the set L, and is used to calculate the maximum number of explanatory variables. For most windows the number of user-allowed delays is $|L| = k_{max} - k_{min} + 1$, but there will be fewer when $q \leq k_{max}$. The explanatory data matrix for each response vector is constructed by concatenating data from the delayed windows, and is defined as

$$\boldsymbol{X} = \begin{cases} [\boldsymbol{W}_{q-k_{min}}, \ldots, \boldsymbol{W}_{q-k_{max}}] & q > k_{max} \\ [\boldsymbol{W}_{q-k_{min}}, \ldots, \boldsymbol{W}_1] & k_{min} < q \leq k_{max} \end{cases} \tag{3.4}$$

To maintain consistency between SWING and existing methods, if $k_{min} = 0$, the response variable is excluded from the explanatory data, prohibiting self-edges within the same window. $\boldsymbol{X}$ has an augmented number of explanatory variables, corresponding to an explanatory variable for each gene at each delay. The number of columns in $\boldsymbol{X}$ is $N \cdot |L|$ if $k_{min} > 0$, or $N \cdot |L|-1$ if $k_{min} = 0$. We did not include any self edges, regardless of delay, during our testing, because the *in silico* and *in vitro* data was collected in a way that does not account for self-edges.

3.3.4. Edge rank is aggregated using group average between different windows

SWING aggregates the results from several weak, but sensitive, windowed models to generate a ranked list of edges. Each window generates an $N \times (N \cdot |L|)$ adjacency matrix, $\boldsymbol{A}$, of edge scores where $A^k_{i,j}$ is the inferred score for gene i as the upstream regulator of gene j with delay k.

The time-series data are naturally left censored, as we cannot know measurements before the experiment occurs. As such, depending on the user specified k_{min} and k_{max}, some windows, particularly the earlier ones, will not infer interactions for larger values of k (*e.g.* $g_i^{q-2} \rightarrow g_j^q$ cannot be inferred if $q < 2$). Therefore, each window $\boldsymbol{W}_q$ infers at most $|L|$ scores, for each gene pair.

In order to combine scores across multiple windows and different delays into a single score $g_i \rightarrow g_j$, SWING does two aggregations. Confidence values from windowed subsets are aggregated into a single network by taking the mean rank of the edge at each delay k, and then taking the mean rank of the edge across all delays. Additionally, community networks estimated from multiple classifiers are built by computing the mean rank of edges outputted from RF, LASSO, and PLSR. We use the edge rank because scores between window models

and methods may not have equivalent distributions. The median of edge ranks may also be used, but in preliminary testing it did not significantly change the results.

3.3.5. SWING graph generation uses resulting adjacency matrix and user-defined cutoff

A directed SWING graph shows causal relationships between N nodes in a system and can be represented by the adjacency matrix $\boldsymbol{A}$ in which each element $A_{i,j}$ is the confidence that an edge exists between parent node g_i and child node g_j. Given Q user-defined windows, for each window, W_q, there are at most $N^2|L|-N$ possible edges that exist in the inferred model. Therefore, the adjacency matrix for each window is

$$\boldsymbol{A_q} = \begin{bmatrix} A_{1,1}^{k_{min}} & \dots & A_{1,N}^{k_{max}} \\ \vdots & \ddots & \vdots \\ A_{N,1}^{k_{min}} & \dots & A_{N,N}^{k_{max}} \end{bmatrix}, \tag{3.5}$$

where $A_{i,j}^{k}$ is the confidence of the interaction whereby the parent node g_i is said to be Granger causal of the child node g_j with a delay of k time points. Self edges within the same window are prohibited, and therefore values $A_{i,i}^{0}$ are set to 0. In this way, a network model with N targets and at most $N \cdot |L|$ regulators is created for each window.

For each window, SWING estimates the confidence of each edge and generates a ranked list of edges based on method-specific criteria. Specifically, RF uses the importance score calculated with the mean squared error [66]; LASSO uses a stability selection metric [32], and PLSR uses the variable importance in projection (VIP) score [67]. The rank of an edge in each windowed model can be used as the confidence metric to compare across methods. We compute a consensus model (SWING-Community) by calculating the mean rank across

methods for each possible edge:

$$\overline{A}_{i,j} = \frac{R_{i,j}^{SWING-RF} + R_{i,j}^{SWING-LASSO} + R_{i,j}^{SWING-PLSR}}{3}, \tag{3.6}$$

where $R_{i,j}$ are the ranks of the edge for each of the tested methods, and $\overline{A}_{i,j}$ is the average rank of the edge $g_i \rightarrow g_j$ used as the confidence metric in the consensus network.

3.3.6. SWING parameter selection is defined by embedded inference methods

SWING is a generalized framework that can be used with any multivariate machine learning inference method. In developing and testing SWING, we implemented three different existing methods: RF, LASSO, and PLSR. Each algorithm requires different tuning parameters. When using RF, we selected the number of trees, the maximum depth of the tree, and the number of trees based on guidelines from the GENIE3 manuscript [66]. For LASSO, we utilized two methods to select the regularization parameter [32]: for *in silico* studies, we selected the regularization parameter based on the cross-validation score; for *in vitro* data sets with comparatively less data, we selected the regularization parameter based on sensitivity analysis for a single random subnetwork and evaluated all subnetworks with the subsequent parameter. For PLSR, we selected the number of principal components to use based on the elbow criterion [67].

In addition to the base methods specific parameters, SWING has user-selected parameters that require knowledge of the system and data. For optimal performance, we suggest the window size be selected such that $T/2 \leq w \leq T$, where T is the number of time points in the time series. If $w < T/2$, increased noise can lead to inference of more false-positive edges. In general, the step size can be set to $s = 1$, unless the user has an abundance of time points and wishes to train on only a subset of the data.

The *in silico* data from GNW is generated such that the perturbation is applied before the simulation and removed at $T/2$. We therefore used $w = 0.5T \approx 10$, to capture the change in dynamics based on the perturbation. For consistency, we also used $w = 0.5T = 7$ for the *in vitro E. coli* SOS network inference.

The allowed delay range is specified by the user in setting k_{max} and k_{min}. We recommend the user set these values based on the range of dynamics expected in the system, or by prior delay analysis such as cross-correlation. Since k_{max} and k_{min} are integer values, they also depend on the sampling interval of the experimental data. Specifying $k_{min} = 0$ allows SWING to infer edges with no delay, as many existing methods do. When testing the *in silico* networks we used $k_{max} = 3$ and $k_{min} = 1$, corresponding to an allowed delay range of $50 \leq k \leq 150$ minutes based upon the *in silico* sampling strategy. This range is consistent with the delays in the *in silico* data estimated using cross-correlation. If, however, the user specifies null SWING parameters—specifically, $w = T$, $k_{max} = 0$, $k_{min} = 0$, and $s = 1$—there is only a single window with no delays between the explanatory and response variables. This condition corresponds to running the base methods independent of SWING.

3.3.7. *In silico* data generation by time-delayed SDEs emulating transcriptional-translation delays

All *in silico* networks were created using GeneNetWeaver [68]. GNW creates a stochastic differential equation (SDE) model from which time-series data are sampled. The kinetic models incorporate Hill kinetics and include both transcriptional and translational components. We generated time-series perturbations for 20 non-isomorphic, 10-node and 100-node subnetworks from the curated *E. coli* and *S. cerevisiae* networks. Simulated data includes

ten random combinations of perturbations which are uniformly sampled at 21 time points with a maximum time of 1000 in arbitrary units.

3.3.8. Parameters for GNW subnetwork extraction

GeneNetWeaver (GNW) is designed to provide synthetic benchmarking data sets for the assessment of network inference methods. GNW includes the networks used for assessment in the DREAM4 challenge, as well as *E. coli*-derived and *S. cerevisiae*-derived gene regulatory networks, which can be used to extract testable subnetworks [69]. These features make GNW ideal for generating *in silico* gene expression data paired with an unambiguous gold standard.

We extracted subnetworks from curated *E. coli*-derived and *S. cerevisiae*-derived networks included in GNW. For each model organism we extracted 20 non-isomorphic networks with 10 and 100 nodes. All subnetworks were extracted with neighbors chosen via greedy selection. The *S. cerevisiae*-derived subnetworks were extracted with 50% of the nodes chosen from the strongly connected component. The curated *E. coli*-derived network does not have one strongly connected component, and therefore *E. coli*-derived subnetworks were extracted starting with a randomly selected vertex. To ensure uniqueness of subnetworks, each sequential network is randomly extracted and preserved only if it is non-isomorphic to all previously extracted networks.

Time-series perturbation data was generated for each of the extracted subnetworks using the default DREAM4 challenge parameters included in GNW. Simulated data includes ten random combinations of perturbations. Simulated experimental perturbations are applied immediately before the time-series data is sampled, and removed halfway through the simulation.

3.3.9. *In silico* predictions and scoring

We scored the inferred networks by calculating the mean increase in the area under the precision-recall (AUPR) curve and the area under the receiver operator characteristic (AUROC) curve for N networks as follows:

$$\overline{S}_{increase} = \frac{\sum_{n=1}^{N} S_{n,SWING} - S_{n,base}}{N}, \tag{3.7}$$

where S_n is the AUPR or AUROC for an individual network, n. For a stochastic method, like RF, S_n, is the mean AUPR/AUROC over several trials, for an individual network.

3.3.10. Cross-correlation and lag analysis identifies time-delayed edges

Temporal cross-correlation has been used by multiple studies to describe how well two signals are correlated when one is shifted in time relative to the other [70][71]. Let $G_i = [g_i^1, g_i^2, \ldots, g_i^T]$ represent measurements of a single gene in a time-series data set. We calculated the pairwise cross-correlation, R, between a pair of signals, G_i and G_j, for a delay k as:

$$R^k_{G_i,G_j}(t) = \frac{\sigma_{G_iG_j}(t)}{\sigma_{G_i}\sigma_{G_j}} \tag{3.8}$$

σ_{G_i}, σ_{G_j}, and $\sigma_{G_iG_j}(t)$ refer to the standard deviation of G_i, standard deviation of G_j, and cross-covariance of G_i and G_j at time t, respectively. The cross-covariance is defined by:

$$\sigma_{G_iG_j}(t) = \frac{1}{N-1}\sum_{t=1}^{N}(G_{i,t-k} - \mu_{G_i})(G_{j,t} - \mu_{G_j}), \tag{3.9}$$

where μ_{G_i} and μ_{G_j} define the mean values of each time series.

We applied several stringent criteria to evaluate time-delayed edges. We calculated the two-sided p-value using the t-distribution equation and subsequently corrected the p-value using the Bonferroni correction (the significant p-values were those less than $\alpha \leq \frac{0.05}{m}$ where m is the total number of edges evaluated) [72]. Since multiple experiments were evaluated for each pairwise comparison, we filtered noisy lagged edges by removing edges in which the sign of the lag differed in more than 10% of experimental perturbations. For *E. coli* and *S. cerevisiae in vitro* data, we also incorporated prior knowledge regarding the sign of the interaction into the lag selection. If multiple delays were significant, depending on whether the parent positively or negatively regulated the target in the gold standard, we selected the lag with the smallest p-value that maximized ($0 < R < 1$) or minimized ($-1 < R < 0$) cross-correlation, respectively. We evaluated cross-correlation at $k = \{0, 10, 20, 30, 60, 90\}$ in *E. coli* and *S. cerevisiae* data sets.

3.3.11. *In vitro* data aggregation

We extracted *in vitro* gold standard networks for *E. coli* and *S. cerevisiae* from RegulonDb and DREAM5 Yeast gold standards (Network4) respectively [25]. For *E. coli*, we extracted the known set of TF and gene interactions from RegulonDb 9.0 [73]. To derive subnetworks from parent gold standards, we performed MCODE clustering using modularity parameters of 0.25 (*E. coli*) and 0.5 (*S. cerevisiae*), resulting in subnetworks where the number of nodes in each module is between 3 and 145 (Tables 3.4 and 3.7). Gene ontology enrichment analysis was performed using a cutoff for false discovery rate-corrected p <0.05 and the *goatools* package [74].

Sources of time-series data sets for *E. coli* and *S. cerevisiae* are described in Tables 3.2 and 3.6. To run SWING, 10 minute time points were generated using cubic spline interpolation

and this data was used to train both SWING and baseline methods [75]. Data interpolation was not needed for lag analysis in Figs. 3.12A and 3.11. Time-series data sets were mean centered.

3.3.12. Computational development

The SWING package was developed in Python 3.4.5 using the following major packages: *NumPy and SciPy* [76], *pandas* [77], and *NetworkX* [78]. The RF, LASSO, and PLSR algorithms use implementations available in *scikit-learn* [79]. Figures were generated using *seaborn* and *matplotlib* [80]. The code for SWING can be found on GitHub: (*https://github.com/bagherilab/SWING*).

3.4. In silico validation and parameter sweep of SWING

SWING integrates multivariate Granger causality and ensemble learning to infer interactions from gene expression data. First, SWING subdivides time-series data into several temporally-spaced windows based on user-specified parameters (Fig. 3.2A). For each window, edges are inferred from the selected window and previous windows, representing interactions with specific delays. This inference results in a ranked list of time-delayed, gene-gene interactions for each window. (Fig. 3.2B). The ensemble of models is aggregated based on edge rank into a static GRN (Fig. 3.2C). *In silico* and *in vitro* validation confirm notable performance improvements.

3.5. SWING improves the inference of *in silico* GRNs

We applied SWING to reconstruct *in silico* GRNs simulated by GeneNetWeaver (GNW) [68]. 20 subnetworks with 10 nodes and non-isomorphic topologies were extracted from *E. coli* and *S. cerevisiae* networks included in GNW to use as gold standards. Networks

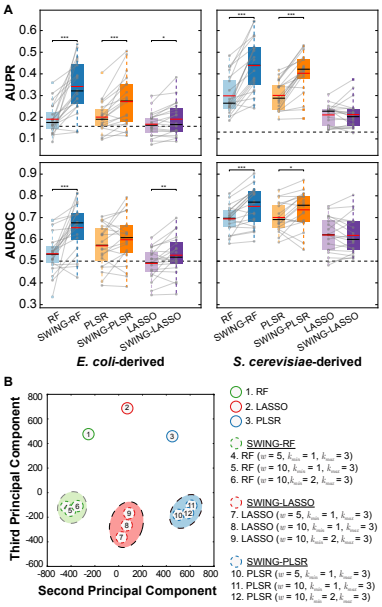

Figure 3.1 | **SWING improves inference of 10-node *in silico* networks.** (A) Changes in AUPR and AUROC in GNW networks. Score changes to individual networks are shown in grey. The mean (red) and median (black) of each score distribution is shown. AUPR and AUROC increase when using SWING-RF or SWING-PLSR compared to their respective base method. SWING-LASSO outperforms LASSO in the *E. coli*-derived networks. The expected score based on random for each metric is shown as a dashed line. n=20 networks, $k_{min} = 1$, $k_{max} = 3$, and $w = 10$ for all networks. p-values were calculated using the Wilcoxon signed-rank test, $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. (B) SWING and non-SWING methods are grouped according to similarity of ranked predictions for 40 10-node *in silico* networks via PCA. PC1 largely separates inference methods based on performance (Fig. 3.4), while PC2 separates methods based on underlying base method. Networks inferred by various SWING parameter selections cluster together according to inference type, with SWING methods forming clusters distinct from corresponding base methods.

were inferred from the generated time-series data using existing multivariate methods as a basis for comparison. We employed RandomForest (RF), Least Absolute Shrinkage and Selection Operator (LASSO), and Partial Least Squares Regression (PLSR) [22, 30, 67], which represent the areas of sparse, nonlinear, and PLS-based regression. We implemented

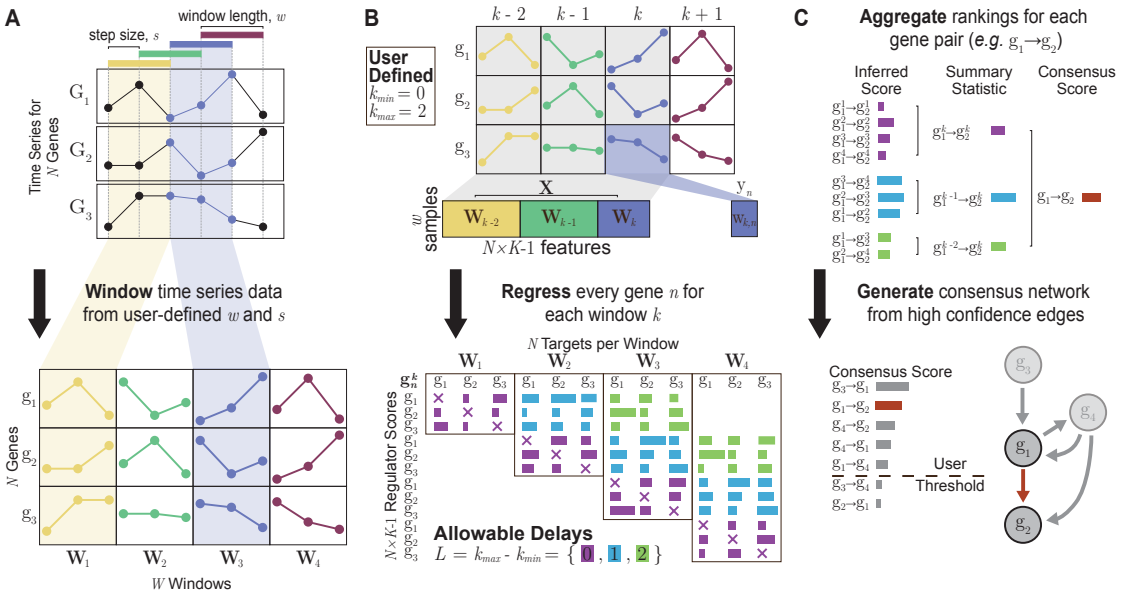

Figure 3.2 | **Overview of the SWING framework.** (A) Time-series data is divided into windows with a user-specified width, w. (B) For each window, inference is performed by iteratively selecting response and explanatory genes. The subset of available explanatory genes is defined by the minimum and maximum user-allowed time delays. (C) Edges from each window model are aggregated into a single network representation of the biological interactions between measured variables.

the SWING chassis and compared the performance of each SWING frontline method with its base method: SWING-RF vs. RF, SWING-LASSO vs. LASSO, SWING-PLSR vs. PLSR.

To capture short-term dynamics consistent with simulated perturbations, we set the window size to roughly half the duration of the time series. The minimum and maximum lags were set to $k_{min} = 1$ and $k_{max} = 3$, which correspond to 50 and 100min. We compared the group of inferred networks by calculating the mean increase in the area under the precision-recall (AUPR) and area under the receiver operating characteristic (AUROC) curves of 40 *in silico* networks. Compared to respective baseline methods, SWING shows a statistically significant increase in AUROC and AUPR for many of the 10-node networks (Fig. 3.1A and Table 3.1) and across all of the 100-node networks (Fig. 3.3, Table 3.1). In particular, RF

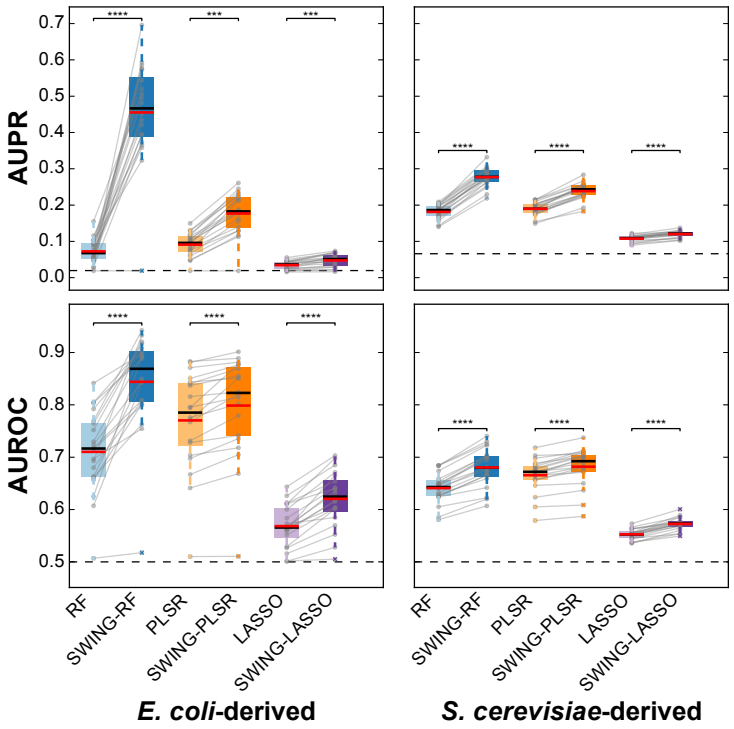

Figure 3.3 | **Changes in AUPR and AUROC curve distributions for 100-node GNW networks.** All networks and methods show a significant improvement in both the AUPR and AUROC when using SWING. Score changes to individual networks are shown as grey lines. The mean (red) and median (black) of each score distribution is shown. The expected score based on random for each metric is shown as a dashed line. n=20 networks, k_{min} =1, k_{max} =3, and w =10 for all networks. p-values are calculated using the Wilcoxon signed-rank test, **** $p < 0.0001$, *** $p < 0.001$.

receives the most notable benefit from SWING; SWING-RF outperforms RF in 39 out of 40 *in silico* networks and application of SWING-RF results in the highest mean AUROC and AUPR for *in silico* networks among tested methods.

3.6. SWING infers distinct edges in networks

No single method performs optimally across all data sets, partially due to biases in predicting different network topologies. For example, *E. coli*-derived networks predominately

Table 3.1 | **Summary of SWING performance on *in silico* networks.** The change in mean AUPR and AUROC for 20 *in silico* 10 and 100-node networks. p-values are calculated using the Wilcoxon signed-rank test.

		AUPR			AUROC		
Dataset	**Method**	Δ	Δ (%)	p-value	Δ	Δ (%)	p-value
Ecoli10	SWING-RF	0.151	98.1	1.20E-04	0.120	25.5	2.93E-04
Ecoli10	SWING-PLSR	0.072	35.7	3.90E-04	0.028	5.4	1.35E-01
Ecoli10	SWING-LASSO	0.023	13.1	1.37E-02	0.037	7.7	6.42E-03
Scerevisiae10	SWING-RF	0.139	50.4	1.20E-04	0.058	8.3	6.81E-04
Scerevisiae10	SWING-PLSR	0.102	35.1	1.40E-04	0.035	5.1	1.37E-02
Scerevisiae10	SWING-LASSO	0.001	0.6	6.54E-01	-0.002	0.0	7.94E-01
Ecoli100	SWING-RF	0.383	647.9	8.86E-05	0.134	19.1	8.86E-05
Ecoli100	SWING-PLSR	0.086	96.2	1.03E-04	0.028	3.7	8.86E-05
Ecoli100	SWING-LASSO	0.013	37.3	8.86E-05	0.052	9.1	8.86E-05
Scerevisiae100	SWING-RF	0.096	53.3	8.86E-05	0.040	6.1	8.86E-05
Scerevisiae100	SWING-PLSR	0.049	25.7	8.86E-05	0.016	2.4	8.86E-05
Scerevisiae100	SWING-LASSO	0.012	11.6	8.86E-05	0.020	3.6	8.86E-05

feature fan-out motifs, which RF infers with greater sensitivity. In contrast, *S. cerevisiae*-derived networks contain more cascade motifs, which are inferred with greater sensitivity by linear methods [25].

To determine if SWING methods provide distinct information from RF, LASSO, and PLSR, we ran principal component analysis (PCA) on ranked edge lists predicted by SWING and the corresponding base methods (Fig. 3.1B). We discarded PC1 because it largely explains the overall performance of each inference method (58% variance explained; Fig. 3.4). Clustering of results in PC2 and PC3 seems to explain biases toward specific network motifs [25]. Along PC2, edge rankings appear to separate based on the internal base method (15% variance explained), while along PC3, SWING edge rankings appear to separate from those of their base methods (5% variance explained). These results suggest that SWING recovers connectivities that are distinct from those recovered from RF, LASSO, and PLSR.

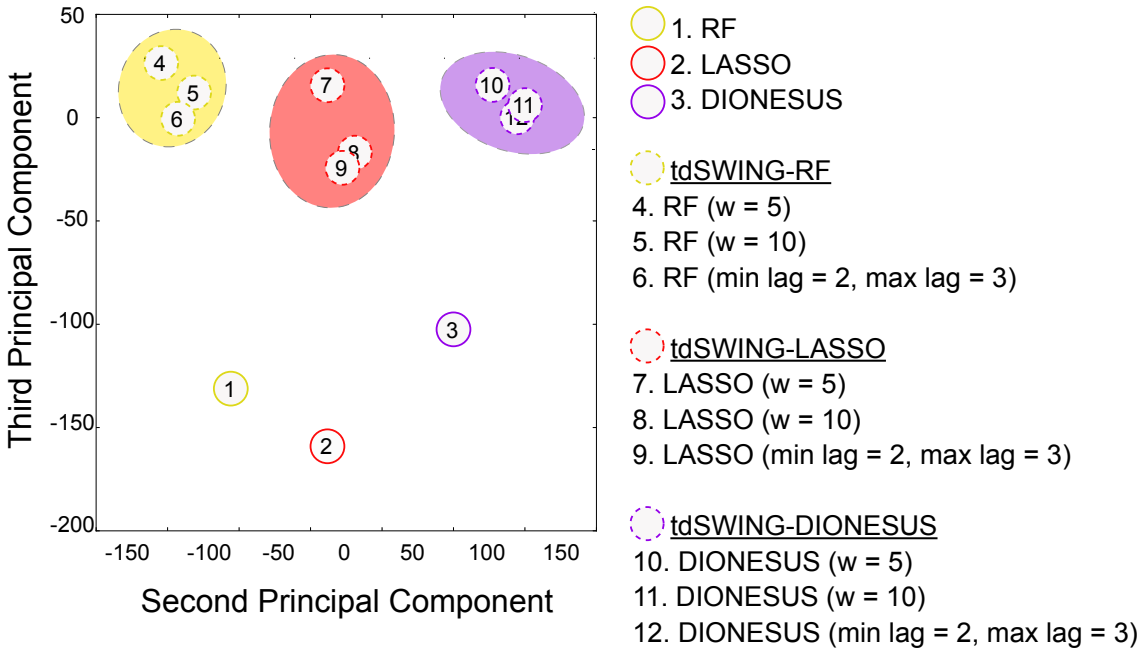

Figure 3.4 | **SWING and non-SWING methods are grouped according to similarity of ranked predictions for 40 10-node *in silico* networks via principal component analysis.** The first (58% variance explained) and second (15% of variance explained), and the second and third (5% of variance explained) principal components are shown. The first principal component largely separates inference methods based on performance, while the second component separates methods based on underlying base method. The second and third principal components seem to explain motif biases. Networks inferred by various SWING parameter selections cluster together according to inference type, with SWING methods forming clusters distinct from corresponding base methods.

Given that it is difficult to determine *a priori* which methods perform optimally in different contexts, deriving a community network is a good strategy for robustly improving predictions [25]. We evaluated the performance of SWING-Community, which combines SWING-RF, SWING-LASSO, and SWING-PLSR predictions by calculating the mean rank across all methods for each possible edge. We note that SWING-Community outperforms RF, resulting in a 52% and 8% mean increase in AUPR and AUROC, respectively, suggesting that SWING infers distinct and complementary networks (Fig. 3.5).

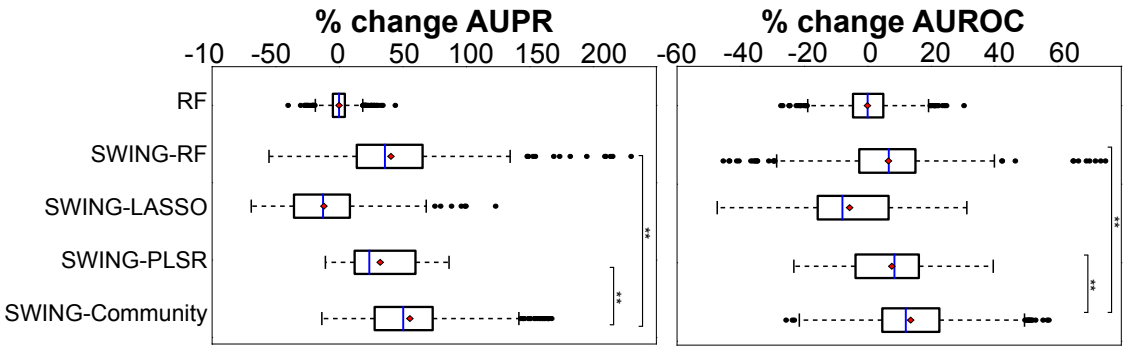

Figure 3.5 | **Boxplots show the percent change in performance of RF vs SWING-RF, SWING-LASSO, SWING-PLSR, SWING-Community prediction for 40 10-node networks.** The percent change of AUPR and AUROC from RF was calculated for each network (500 trials). For SWING methods, the following parameters were used: $w = 10$, $k_{min} = 1$, and $k_{max} = 3$. p-values are calculated using the Mann-Whitney U test, ** p <0.01.

3.7. SWING improves network inference by promoting time-delayed edges

Endogenous reactions, such as protein translation, post-translational modifications, translocation, or oligomerization are often not accounted for in the inference model. However, even if underlying network kinetics are linear (or approximately linear), the resulting dynamics can appear delayed when not all nodes are observed (Fig. 3.6A). Delayed behavior in gene expression and protein translation has been established in several studies [71, 81].

We estimated the apparent time delay of each interaction in a 10-node GNW network by calculating the pairwise peak cross-correlation between time series of all true regulator and target combinations. The majority of true interactions within GNW networks have a time delay between 0 and 150min (Fig. 3.6B). We observe that SWING is more likely to promote edges with an identifiable delay within the range of user-specified parameters (Fig. 3.7A). Across all *in silico* networks, SWING-RF promotes 65.8% of true edges with a delay versus 55.4% of true edges without a delay (p =0.018), and SWING-PLSR promotes 67.0% of true edges with a delay versus 47.1% of true edges without a delay (p =6e-6)(Fig. 3.7B).

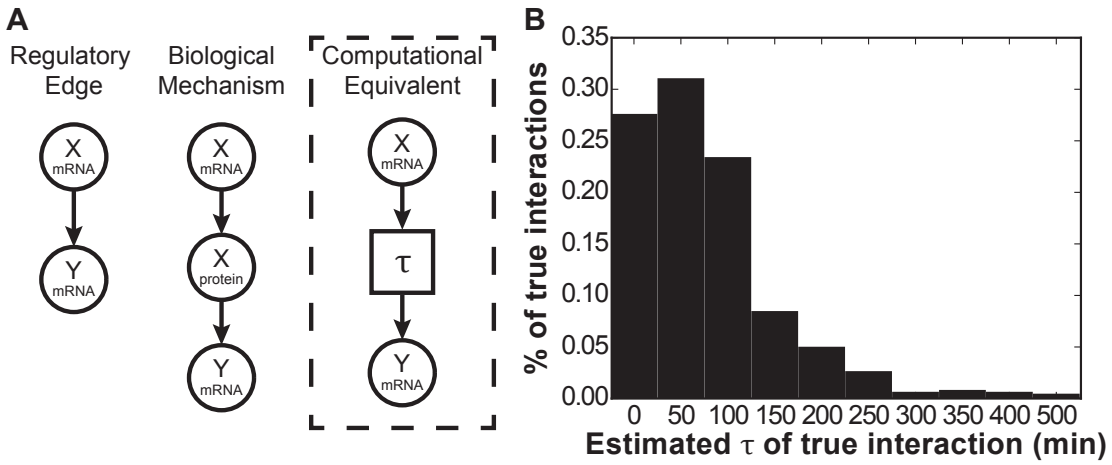

Figure 3.6 | **Identification of delays in DREAM 4 *in silico* networks.** (A) Real biological systems require additional steps, such as translation, whose kinetics determine the delay between upstream gene expression and downstream nodes. The physical time required for these steps can be expressed as a delay, τ, and must be accounted for when inferring GRNs from gene expression data. (B) The τ for each edge in five DREAM 4 *in silico* networks was calculated using the cross-correlation function. $\tau = 0$ minutes was calculated for around 25% of the interactions, indicating no delay could be identified. A large fraction of interactions have $50 \leq \tau < 150$ min, indicating that the kinetics of the model result in a delay. Larger values of τ may be due to the kinetics, but very high values are likely due to noise.

Many of the promoted edges with an identifiable delay are highly ranked by base methods RF and PLSR. In general, delayed true edges ranked in the first quartile by the base method are likely to be promoted, while those ranked lower are no more likely to be promoted than nondelayed true edges (Fig. 3.7B). While SWING is more likely to promote true edges with a delay, the magnitude of this promotion is not consistent across the different base methods or networks. SWING-RF promotes true edges with an apparent time delay by an average of 7.50 ranks relative to true edges without an apparent time delay (p =4.75e-3) for *S. cerevisiae*-derived networks. In contrast, SWING-PLSR promotes true edges with an apparent delay by an average of 7.78 ranks relative to true edges without an apparent time delay (p =6.89e-5) for *E. coli*-derived networks (Fig. 3.7B). In one example, *S. cerevisiae*

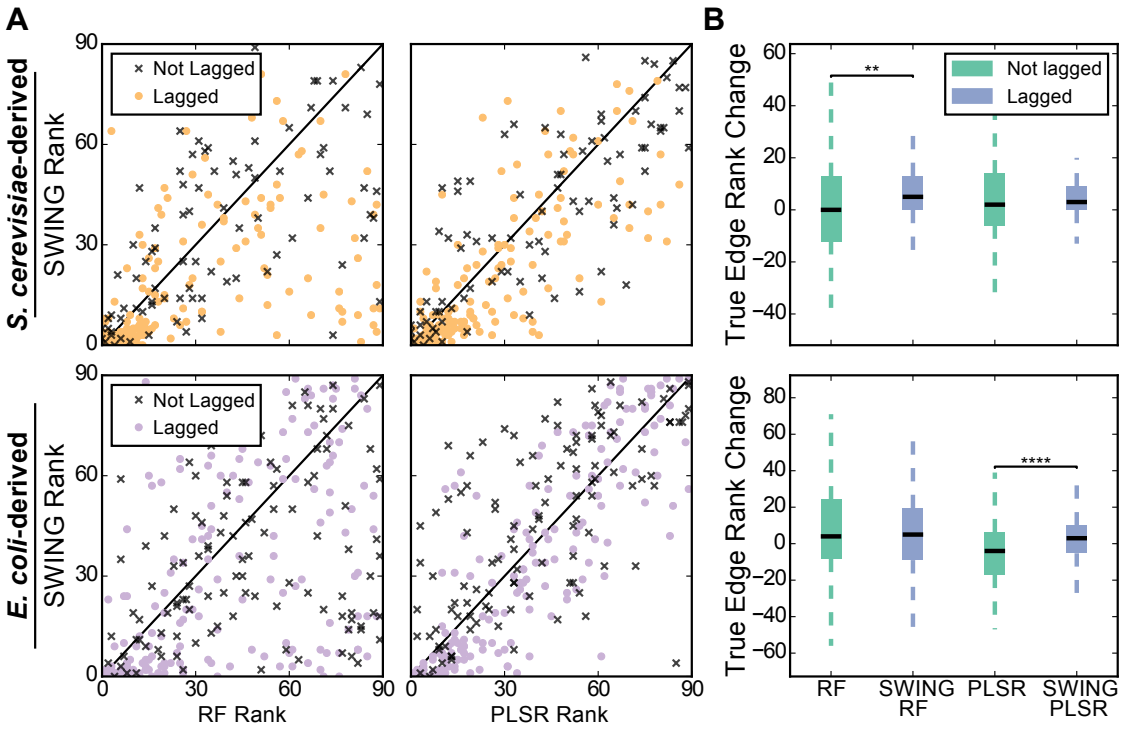

Figure 3.7 | **SWING promotes edges with apparent time delays between genes.** (A) The inferred rank using SWING versus the respective control methods. All true edges for the 20 networks are plotted. Edges that present below the diagonal line are promoted by SWING, and those that present above are demoted. For all methods and networks, SWING is significantly more likely to promote an edge with apparent lag. (B) Distribution of true edge rank changes when using SWING for lagged and not lagged edges. The median true edge promotion of lagged edges is significantly greater for *E. coli* networks when SWING is run using PLSR, but significantly greater for *S. cerevisiae* networks when SWING is run using RF. p-values are calculated using the Mann-Whitney U test, **** $p < 0.0001$, ** $p < 0.01$. n=292 for *E. coli* and n=257 for the *S. cerevisiae*.

Network 12, SWING-RF improves the AUROC from 0.539 to 0.872, a 61.7% increase relative to the base method. Compared to RF the edge ranking for SWING-RF promotes many true edges, and all of the true edges with a delay are promoted by SWING (Fig. 3.8A).

To demonstrate how SWING promotes delayed edges, we highlighted the true edge between Gene 2 (G2) and Gene 1 (G1) in *S. cerevisiae* Network 12. G2 is the only node

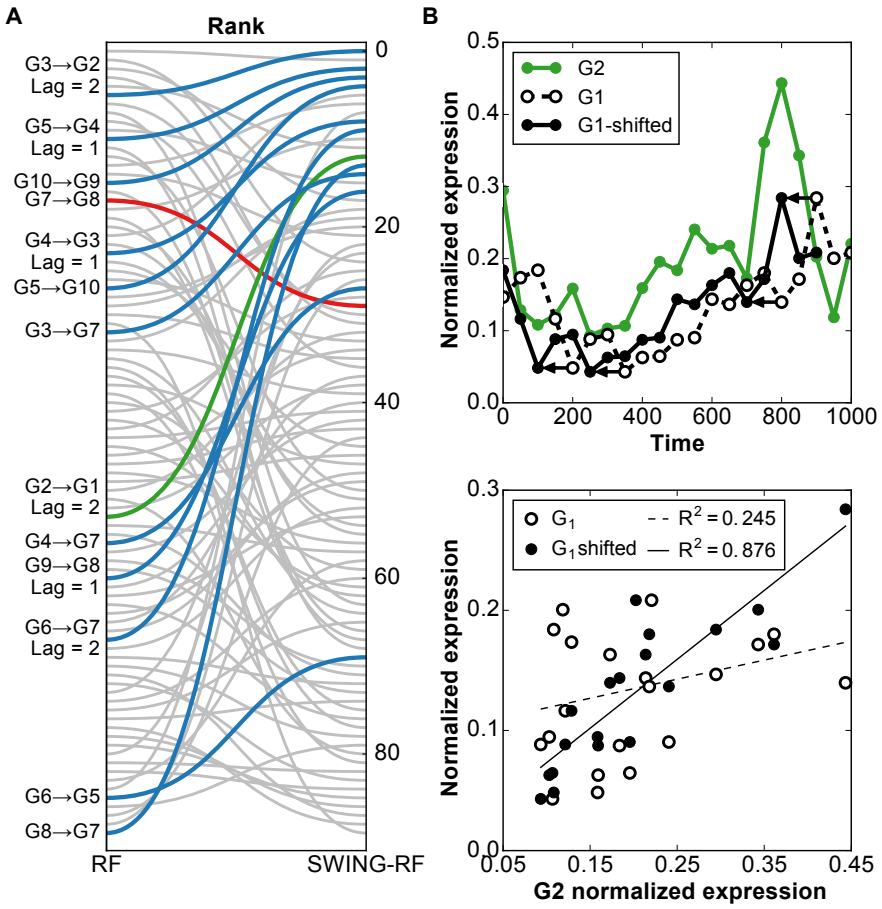

Figure 3.8 | **SWING promotes time-delayed edges and increases correlation between genes.** (A) Edge rank comparison for *S. cerevisiae* network 12 when using SWING-RF compared to RF (blue = promoted edges, red = demoted edges, grey = false edges, green = $G2 \rightarrow G1$ analyzed in panel B). Edges for which a time delay could be estimated are labeled. (B) Improved correlation between G2 and G1 when a lag is artificially introduced.

upstream of G1, and the input data includes an experiment where only G2 is perturbed, thus the delay between G2 stimulation and G1 response is unambiguously isolated (Fig. 3.9A). We estimated the delay between G2 and G1 as two time points, or 100min. We shifted the G1 time series by two time points to show that the Pearson correlation of the resulting time series notably increases (Fig. 3.8B).

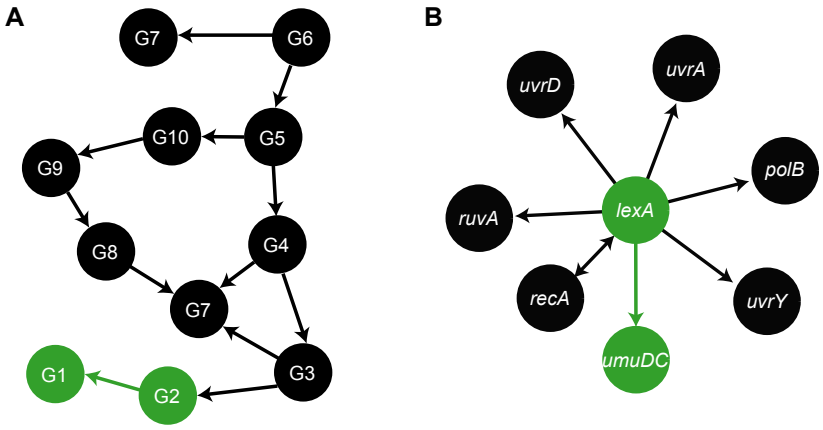

Figure 3.9 | **Graph representations of (A) *S. cerevisiae*-derived network 12 and (B) *E. coli* SOS network.** Nodes and edges highlighted in green are specifically interrogated in Figs. 3.8 and 3.10.

3.8. SWING infers apparent time-delayed edges with greater sensitivity in the *E. coli* SOS network

We applied SWING to an *in vitro* 8-node *E. coli* GRN that activates with DNA damage [82, 62]. The SOS network contains several complex interactions, including multiple cascades and feedback loops generated by a combination of transcriptional activators and repressors. We computed the mean of three replicates for each time point following DNA damage inducing Norfloxacin treatment [83].

The sampling strategy for the *in vitro* SOS data is different from that of the *in silico* GNW data. Due to fewer time points, we were restricted to assessing interactions with shorter possible time delays. Using $w = 0.5T = 7$, $k_{min} = 0$, and $k_{max} = 1$, SWING-RF infers the network more accurately than other reported inference algorithms including RF, LASSO, TSNI [83], and BANJO [84]. Because RF is a stochastic method, we ran both RF and SWING-RF 50 times on the SOS network. On average, SWING-RF increases the AUPR from 0.286 to 0.356 (24.6%, p =1.41e-13) and the AUROC from 0.756 to 0.819 (8.3%, p =5.28e-34). To assess promotion of time-delayed edges, we calculated the mean edge ranks

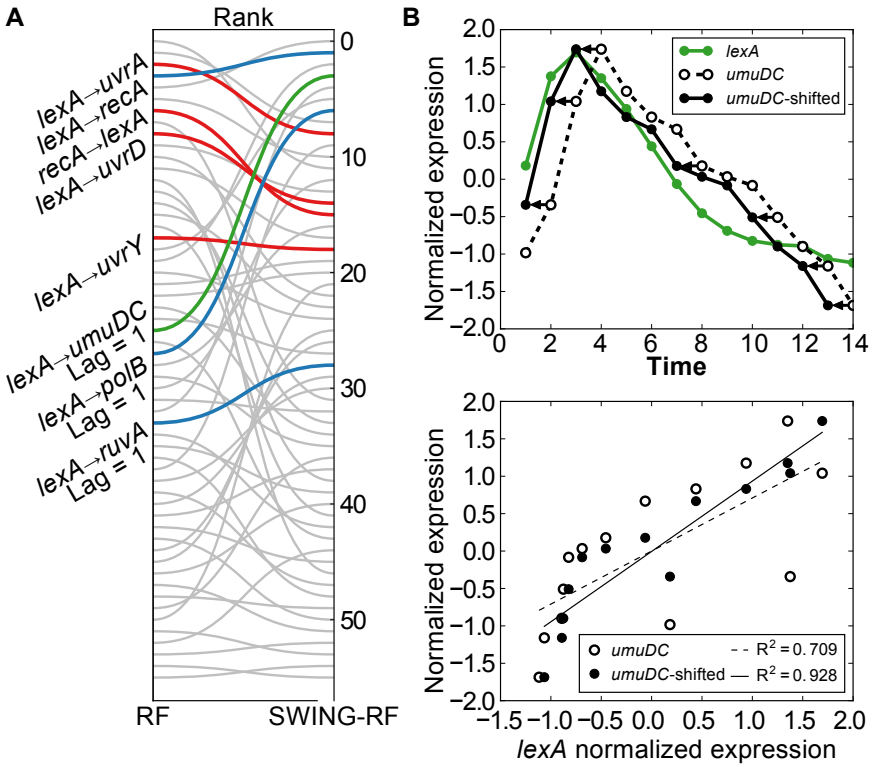

Figure 3.10 | **SWING promotes edges with apparent time delays and increases correlation between genes.** The true network structure is provided in SI Appendix, Fig. 3.9B. (A) Edge rank comparison for *E. coli* SOS network when using RF and SWING-RF (blue=promoted edges, red=demoted edges, black=no change, grey=false edges, green=lexA → *umuDC* analyzed in panels B and C). We report the lag for edges with an apparent time delay. (B) Time series for *lexA* and *umuDC* show better alignment when *umuDC* is shifted by one time period. (C) Improved correlation between *lexA* and *umuDC* the time series of *umuDC* is shifted by one time period.

across all 50 runs and compared the resulting lists. Though SWING-RF demotes some true edges, it promotes all three edges that exhibit a time delay (Fig. 3.10A). We highlight the edge between *lexA* and *umuDC* (Fig. 3.9B), which has an estimated lag of 6min. When the *umuDC* time series is shifted by this amount the correlation between *lexA* and *umuDC* increases from 0.709 to 0.928 (Fig. 3.10B). These findings reaffirm that SWING improves network inference, in part, by promoting edges with identifiable delays.

Table 3.2 | ***E. coli* data set for RegulonDB lag analysis.** References, time-points, and conditions of datasets used for aggregate gene regulatory network analysis.

Strain/Condition	Time Points	Citation
MG1655 control	t=10,20,30,40,50	[85]
MG1655 cold stress	t=10,20,30,40,50	[85]
MG1655 heat stress	t=10,20,30,40,50	[85]
MG1655 oxidative stress	t=10,20,30,40,50	[85]
EMG2 LB 0pt02percent glucose	t=150,180,210,240,270,300,330,360,480	[25]
EMG2 LB 0pt04percent glucose	t=150,180,210,240,270,300,330,360,480	[25]
MG1655 wt untreated	t=0,30,60,90,120	[25]
MG1655 wt MMC 2pt5ug	t=0,30,60,90,120	[25]
MG1655 wt UV 500en	t=0,30,60,90,120	[25]
BW25113 uninduced	t=0,30,60,120,180	[25]
BW25113 norflaxacin	t=0,30,60,120,180	[25]
BW25113 D recA	t=0,30,60,120,180	[25]
BW25113 U ccdB	t=0,30,60,120,180	[25]
BW25113 D recA norflaxacin	t=0,30,60,120,180	[25]
BW25113 D recA U ccdB	t=0,30,60,120,180	[25]
MG1655 U lacZ	t=0,30,60,90	[25]
MG1655 U ccdB	t=0,30,60,90	[25]
EMG2 LB norf 25ng	t=0,12,24,36,48,60	[86]

3.9. SWING accurately infers RegulonDB modules with time-delayed edges

We curated microarray data to infer time-delayed edges from experimentally validated GRNs in *E. coli* (Fig. 3.12A) and *S. cerevisiae* (Fig. 3.11). This curated data was aggregated across 18 data sets for *E. coli* and 8 data sets for *S. cerevisiae*, where data was unevenly sampled for time intervals that range from 5 to 120min (Table 3.2). To assess the landscape of apparent time delays present in these gene expression data, we performed pairwise cross-correlation lag selection between experimentally-confirmed edges [70]. We reveal that of 2870 experimentally confirmed edges, only 23.7% exhibit an apparent time delay of 0 and 13.7% exhibit a time delay of at least 10min. Surprisingly, only 37.4% of confirmed edges exhibited pairwise correlation (R >0.7, p <1e-5; Fig. 3.12A).

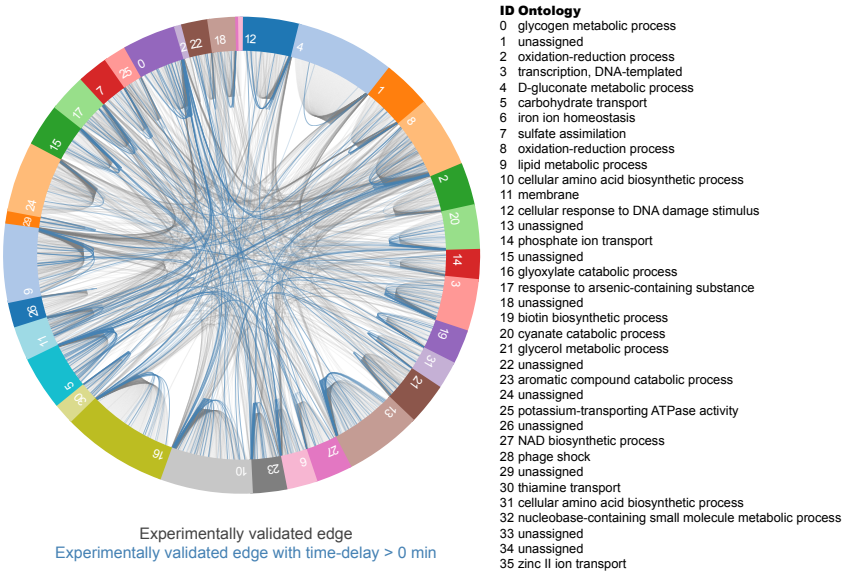

Figure 3.11 | **Cross-correlation analysis of time-delayed interactions derived in *S. cerevisiae*.** Circular diagram depicts experimentally validated interactions and gene ontologies present in each module. Blue edges depict edges displaying time-delayed interactions (time delay of 10 minutes or greater) inferred using pairwise cross-correlation from curated microarray data.

To determine whether lag is associated with modularity and function, we clustered the *E. coli* and *S. cerevisiae* network into smaller modules using MCODE [87] and performed gene ontology enrichment analysis. Several modules, such as those associated with catabolic processes and metal ion binding, are enriched with time-delayed edges of at least 10min (Tables 3.3 and 3.4). Transcription factors are known to regulate genes on a global or combinatorial scale tend to exhibit similar time delays (Table 3.5).

To determine if SWING more accurately infers network structure in diverse contexts, we performed cubic spline interpolation to generate evenly sampled time-series gene expression at 10min intervals and benchmarked SWING-Community performance against an ensemble model of RF, LASSO and PLSR (R/L/P) base for each clustered module using this dataset. SWING-Community outperformed R/L/P in subnetworks in which more than 10% of edges

Table 3.3 | **Lagged edge analysis of 35 *E. coli* subnetworks from RegulonDB.** We highlight lagged edges with apparent time delays of 10 minutes or greater.

Cluster ID	**Total # Edges**	**# of Lagged Edges ($k \geq 10m$)**	**% Lagged Edges**
0	125	8	6%
1	542	138	25%
2	193	14	7%
3	113	24	21%
4	13	4	31%
5	299	87	29%
6	55	5	9%
7	132	11	8%
8	65	8	12%
9	71	2	3%
10	124	29	23%
11	106	8	8%
12	76	6	8%
13	203	17	8%
14	36	3	8%
15	56	2	4%
16	18	8	44%
17	102	8	8%
18	141	15	11%
19	36	3	8%
20	30	1	3%
21	32	7	22%
22	86	10	12%
23	27	12	44%
24	204	25	12%
25	167	20	12%
26	94	14	15%
27	45	3	7%
28	43	3	7%
29	34	2	6%
30	271	33	12%
31	114	19	17%
32	29	8	28%
33	29	1	3%
34	23	0	0%

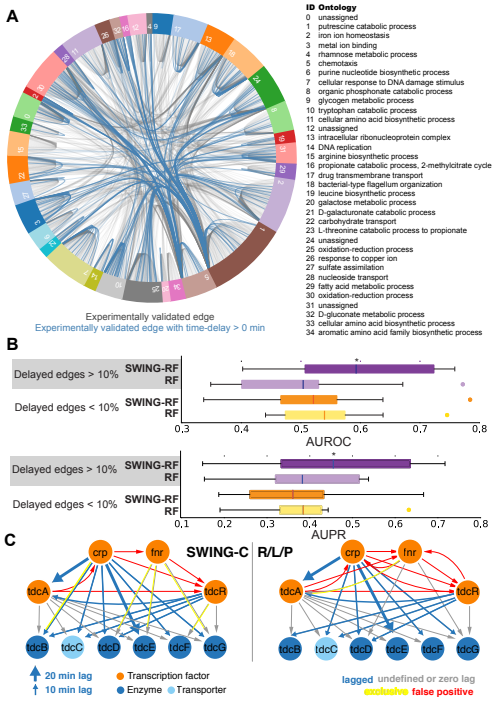

Figure 3.12 | **Application of SWING on time-delayed gene regulatory network modules in *E. coli*.** (A) Circular diagram depicts experimentally validated interactions and gene ontologies present in each module (RegulonDb). Blue edges depict time-delayed interactions inferred using pairwise cross-correlation from curated microarray data. (B) SWING-Community, with $w = 4$, $k_{min} = 1$, $k_{max} = 1$ applied to RegulonDb subnetworks that are and are not enriched with time-delayed edges (fraction of delayed edges is greater than 10%, n =12 subnetworks; fraction of delayed edges is less than 10%, n =14 subnetworks). (C) SWING-Community and RF/LASSO/PLSR (R/L/P) ensemble method applied to *tdcABC* regulon, which is the module found to have the highest enrichment of time-delayed edges (44% edges with a time delay of 10min or greater).

are time-delayed (N=12 clusters, 9 clusters with fewer then 10 genes, or fewer than 3 transcription factors were removed from analysis, p =0.031; Fig. 3.12B). As an example, we identified time-delayed properties of key regulators of the *tdcABC E. coli* operon that are responsible for the transport of threonine and serine during anaerobic growth [91]. In particular, our analysis identifies two global transcription factors that bind combinatorially to

Table 3.4 Gene ontological analysis of *E. coli* subnetworks in RegulonDB

Cluster ID	Gene Ontology	Significance (Corrected P-Value)	# of Genes in GO	Size of GO Category
1	putrescine catabolic process	4.86E-05	7	110
2	iron ion homeostasis	9.24E-25	20	74
3	metal ion binding	0.008352298	20	41
3	oxidation-reduction process	0.002072459	18	41
4	rhamnose metabolic process	1.72E-10	5	6
5	ATP-binding cassette (ABC) transporter complex	0.004500043	8	43
5	chemotaxis	4.42E-05	7	43
6	purine nucleotide biosynthetic process	5.01E-20	12	25
7	cellular response to DNA damage stimulus	2.28E-16	30	63
8	organic phosphonate catabolic process	2.20E-12	8	40
9	glycogen metabolic process	0.000432348	4	29
10	tryptophan catabolic process	0.027249226	3	37
11	cellular amino acid biosynthetic process	0.028322085	9	50
13	translation	2.29E-06	12	40
13	intracellular ribonucleoprotein complex	7.77E-08	11	40
14	DNA replication	0.000287344	6	16
15	arginine biosynthetic process	1.17E-18	11	33
16	propionate catabolic process, 2-methylcitrate cycle	3.84E-06	4	10
17	drug transmembrane transport	0.000254503	7	47
18	bacterial-type flagellum organization	1.70E-06	8	53
19	leucine biosynthetic process	1.16E-06	5	16
20	galactose metabolic process	1.20E-09	5	10
21	D-galacturonate catabolic process	0.009971194	3	13
22	carbohydrate transport	5.04E-12	16	37
23	L-threonine catabolic process to propionate	1.47E-12	6	8
25	oxidation-reduction process	1.07E-05	24	54
26	response to copper ion	1.32E-07	6	20
27	sulfate assimilation	8.24E-09	7	29
28	nucleoside transport	0.005095171	3	14
29	fatty acid metabolic process	1.94E-27	16	21
30	oxidation-reduction process	2.21E-11	32	59
32	D-gluconate metabolic process	2.80E-12	7	13
33	cellular amino acid biosynthetic process	1.90E-07	10	19
34	aromatic amino acid family biosynthetic process	8.37E-23	13	20

induce activity in the *tdcABC* operon. *Crp* and *fnr* are global regulators that respond to glucose starvation and anaerobic growth respectively [92, 93].

Interestingly, lag analysis identifies 10 and 20min time delays between *crp* and target genes in the *E. coli tdcABC* operon. While the precise delay identified by our analysis is not consistent with that observed in experiments, studies confirm that a delay exists between *crp* induction and the induction of several target genes due to post-translational modification [94, 95]. Of 32 edges in the gold standard, SWING identifies 27 true-positive (TP) edges and 5 false-positive (FP) edges (85% TP) while the ensemble model predicts 24 true-positive edges and 8 false-positive edges (75% TP). In this example, SWING-Community infers both time-delayed and non time-delayed edges more sensitively than the R/L/P ensemble model.

Table 3.5 Lagged edge analysis of *E. coli* transcription factors from RegulonDB. We highlight lagged edges with apparent time delays of 10 minutes or greater.

Transcription Factor	Total # Edges	# of Lagged Edges ($k \geq 10m$)	% Lagged Edges
crp	466	129	28%
fnr	286	34	12%
csgd	24	7	29%
fis	166	6	4%
torr	11	6	55%
lexa	52	5	10%
iscr	27	5	19%
arca	155	4	3%
gntr	11	4	36%
narl	120	3	3%
soxs	32	3	9%
fliz	19	3	16%
oxyr	32	2	6%
cysb	28	2	7%
argp	13	2	15%
fur	117	1	1%
cpxr	55	1	2%
narp	50	1	2%
pdhr	35	1	3%
purr	29	1	3%
rcsb	20	1	5%
fadr	18	1	6%
evga	15	1	7%
mraz	15	1	7%
arac	13	1	8%
leuo	12	1	8%
basr	12	1	8%
lrp	69	0	0%
phob	54	0	0%
mode	46	0	0%
phop	39	0	0%
argr	35	0	0%
mara	31	0	0%
nagc	30	0	0%
fhla	29	0	0%
gade	26	0	0%
gadx	26	0	0%
rob	20	0	0%
nac	18	0	0%
metj	15	0	0%
ydeo	15	0	0%

The false-positive edges inferred by SWING-Community are also within the subset of false-positive edges inferred by the base community method.

Table 3.6 *S. cerevisiae* data set for DREAM5 lag analysis

Strain/Condition	Time Points	Citation
Y262 Wild type cells, oxidative stress	t=0,30,60,100,140,180	[88]
Y262 Wild type cells, DNA damage stress	t=0,30,60,100,140,180	[88]
Y262 Wild type cells, oxidative decay	t=0,5,10,15,20,30,40,50,60	[88]
Y262 Wild type cells, DNA damage decay	t=0,5,10,15,20,30, 40, 50, 60	[88]
IFO0233 Wild type cells, control	t = 830,834,838,842,846,850,854,858,862,866,870	[89]
IFO0233 Wild type cells, phenelzine treatment	t = 874,878,882,886,890,894,898,902,906,910,914,918,922,926,930,934,938,942,946,950,954,958,962,966,970,974,978,982,986,990,994,998,1002,1006,1010,1014,1018	[89]
BF264-15Dau Wild type cells, YEP medium	t = 30,38,46,54,62,70,78,86,94, 102,110,118,126,134,142,150,158,166,174,182,190,198,206,214,222,230,238,246,254,262	[90]
BF264-15Dau D CLB1 cells, YEP medium	t = 30,38,46,54,62,70,78,86,94, 102,110,118,126,134,142,150,158,166,174,182,190,198,206,214,222,230,238,246,254,262	[90]

Table 3.7 Gene ontological analysis of *S. cerevisiae* subnetworks

Cluster ID	Gene Ontology	Significance (Corrected P-Value)	# of Genes in GO	Size of GO Category
0	glycogen metabolic process	0.008232665	4	59
2	oxidation-reduction process	2.07E-22	87	223
3	transcription, DNA-templated	0.032096568	23	93
4	D-gluconate metabolic process	2.80E-12	7	13
5	carbohydrate transport	9.24E-29	53	299
5	carbohydrate metabolic process	3.18E-25	59	299
5	cytoplasm	0.034127985	94	299
6	iron ion homeostasis	8.10E-23	21	111
6	ion transport	4.93E-09	22	111
6	transport	6.25E-06	46	111
7	sulfate assimilation	6.26E-09	7	28
7	sulfur compound metabolic process	2.51E-08	7	28
8	oxidation-reduction process	0.016750762	22	68
9	lipid metabolic process	7.45E-20	16	22
10	cellular amino acid biosynthetic process	3.55E-08	19	95
11	membrane	2.92E-07	20	173
12	cellular response to DNA damage stimulus	6.90E-17	30	61
14	phosphate ion transport	1.32E-10	8	54
16	glyoxylate catabolic process	0.001176962	3	9
17	response to arsenic-containing substance	1.40E-05	3	3
19	biotin biosynthetic process	3.46E-12	6	6
20	cyanate catabolic process	1.40E-05	3	4
21	glycerol metabolic process	0.00127654	3	3
23	aromatic compound catabolic process	1.91E-08	5	6
25	potassium-transporting ATPase activity	1.83E-08	4	5
27	NAD biosynthetic process	1.15E-06	4	5
28	phage shock	0.000280394	3	6
30	thiamine transport	7.01E-05	3	6
31	cellular amino acid biosynthetic process	2.49E-12	13	20
32	nucleobase-containing small molecule metabolic process	0.013445757	2	3
35	zinc II ion transport	0.003916298	3	6

3.10. SWING performance is robust across parameters

SWING adds user-defined parameters to baseline methods, which are necessary for window creation and time-delay inference. The selection of these parameters is both context and data specific. We conducted parametric sensitivity analysis of SWING as a function of window size, combinations of k_{min} and k_{max}, and experimental sampling interval in context of the *in silico* networks and the *E. coli* SOS network (Figs. 3.14-3.13). While SWING outperforms baseline methods over a wide range of window sizes (Fig. 3.14), the performance of a single network may differ from other networks, suggesting that the optimal window size is

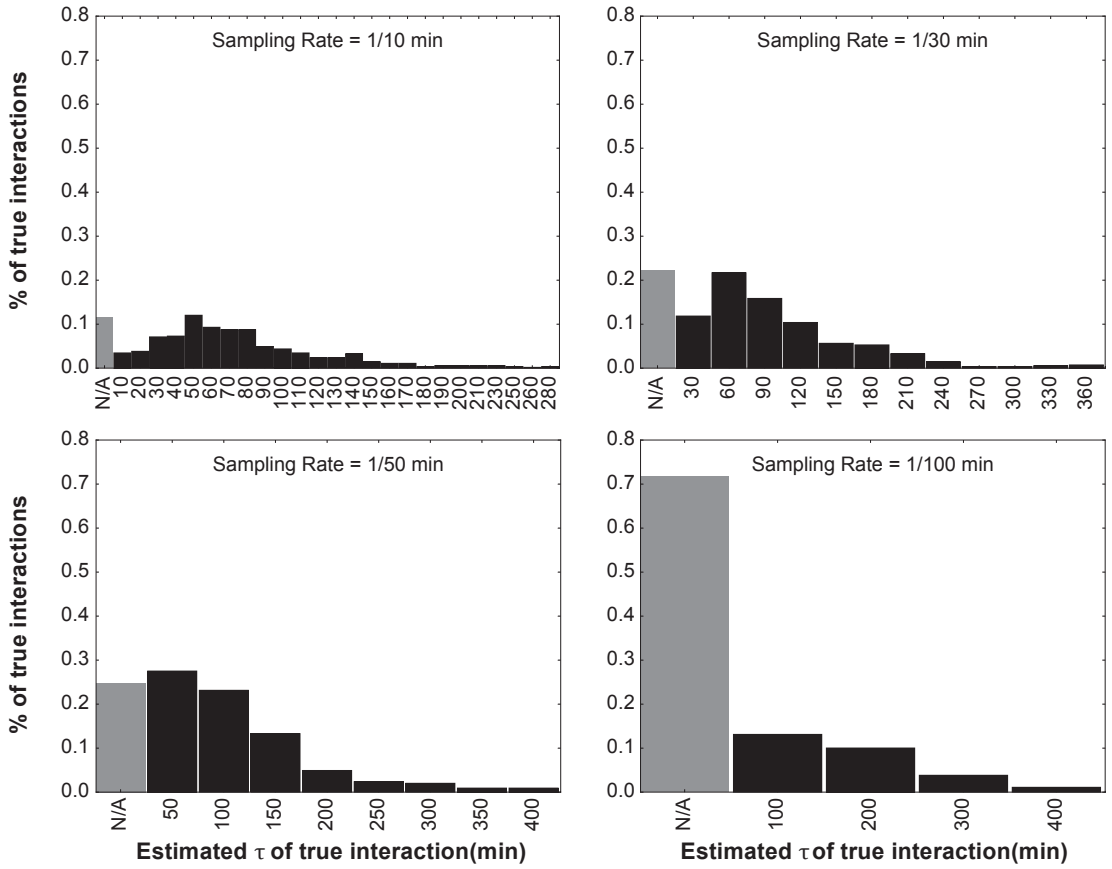

Figure 3.13 | **Barplots show the lag distribution each sampling interval for each aggregated *in silico* data sets.** For example, the sampling interval 333 indicates that samples were taken at t=0, 333, 666, and 999. Lag was calculated using cross-correlation for 40 10-node networks, 20 *E coli*-derived and 20 *S. cerevisiae*-derived. True edges for which no apparent lag was calculated are labeled as "N/A".

partially dependent on the underlying inference method and network structure. Therefore, user-specified SWING parametersk_{min}, k_{max}, and $w$should be chosen based on the data, and are discussed in detail in Supporting Information: Sensitivity Analysis. Overall SWING outperforms baseline methods for a wide range of possible parameters (Figs. 3.14-3.15).

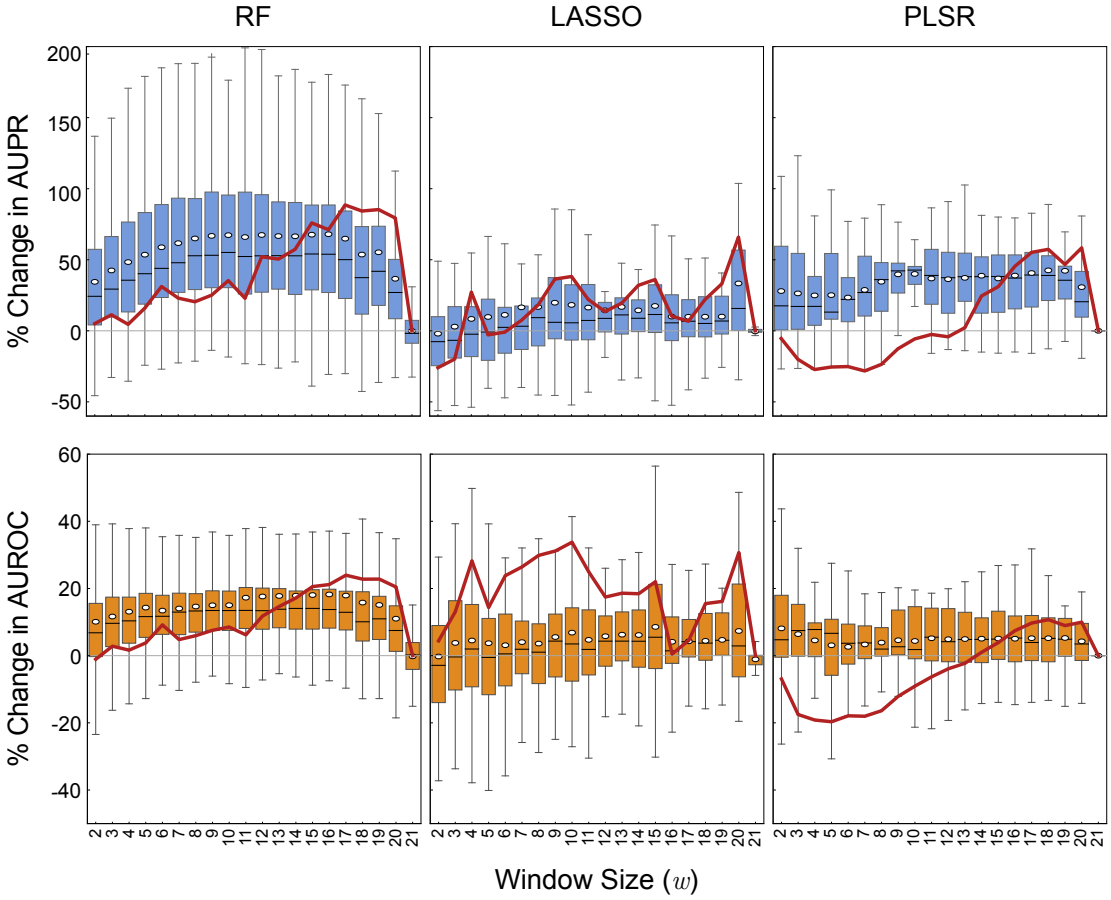

Figure 3.14 | **Boxplots show the percent change in performance of SWING-RF, SWING-LASSO, SWING-PLSR compared to baseline methods (RF, LASSO, PLSR respectively) for each window size change (white dot = mean, black bar = median).** The percent change of AUPR and AUROC from non-SWING methods was calculated for each network (100 trials, 40 10-node networks). The red line indicates the AUROC/AUPR for one example network as w changes. For SWING methods, the following parameters were used: $k_{min} = 1$, and $k_{max} = 3$ (k_{max} was adjusted accordingly to be the largest allowed value when w was 19, 20, and 21).

3.11. Discussion

Tight regulation of gene expression is critical to maintaining robust responses to perturbations and environmental disturbances, and misregulation of intracellular signaling dynamics can lead to a wide variety of diseases. For this reason, uncovering the topology of GRNs is

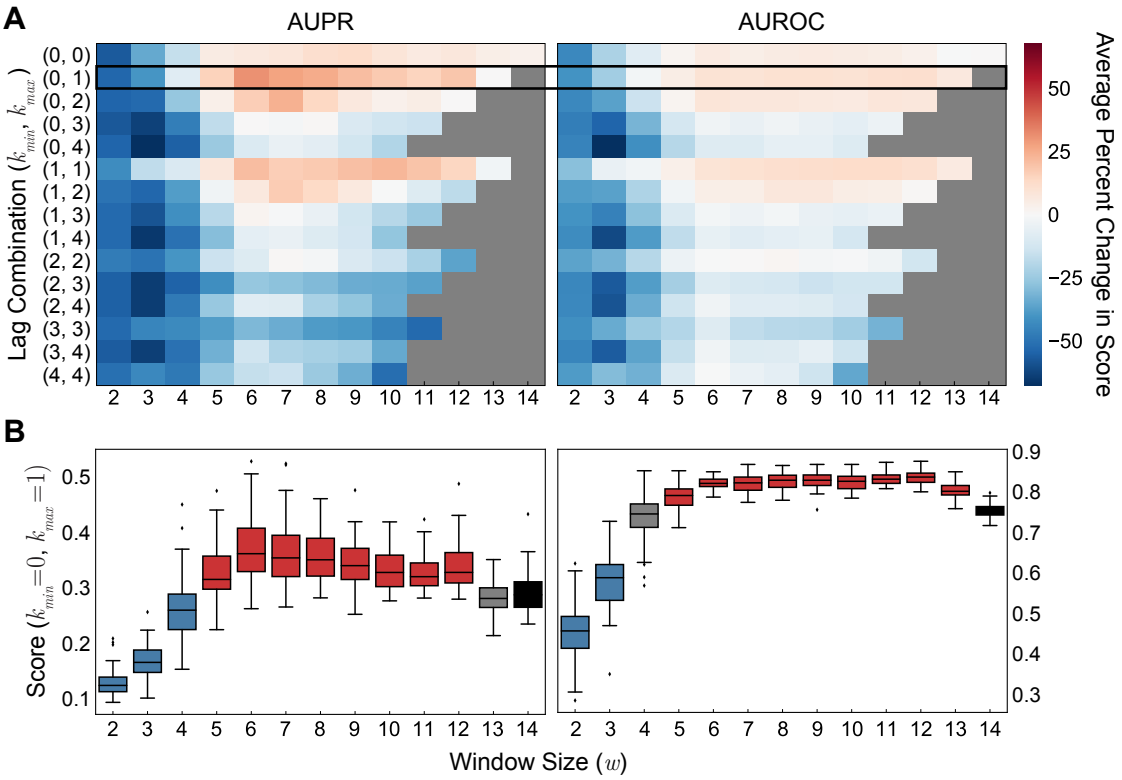

Figure 3.15 | **Results of sensitivity analysis on *in vitro* SOS data using SWING-RF.** (A) Heatmaps show performance change of a three parameter scan (w, k_{min}, and k_{max}) of the *in vitro* SOS network. Red denotes that SWING-RF performed better than the RF baseline method with the indicated parameters while blue denotes that SWING-RF performed worse than the RF baseline method. Parameter combinations that are not possible are shown in grey. (B) Boxplots show AUPR and AUROC distributions for 50 trials of SWING-RF at each window size, w, with $k_{min} = 0$, $k_{min} = 1$. These plots show an example of the variance AUPR/AUROC scores for each RF realization, for the row outlined in A (black = baseline distribution using RF; blue = distributions with a significantly lower score than the baseline; red = distributions with a significantly higher score than the baseline; grey = distributions with no significant score difference than the baseline. p-values are calculated with a paired t-test. Values are considered significant with $p < 0.05$).

of fundamental interest to the scientific community, since the resulting maps can be used to identify interventions to control cellular phenotypes. Many current methods disregard temporal information and are limited in their ability to accurately infer network topology.

Indifference to time delays will be the Achilles heel of many systems biology strategies. We developed a general temporal framework for network inference that accurately uncovers the regulatory structures governing complex biological systems by accounting for these fundamental delays. SWING improves upon existing Granger methods by generating an ensemble of windowed models that simultaneously evaluate multiple upstream regulators at several potential time delays. We validate its utility and performance in several *in silico* (Fig. 3.1A) and *in vitro* (Figs. 3.10 and 3.12B) systems.

3.11.1. Consideration of time delays improves SWING performance and should be integrated in experimental design

Our *in silico* and *in vitro* results demonstrate that promoted edges were enriched for those with apparent time delays (Fig. 3.7B), suggesting that network inference is improved, in part, by accounting for temporal information. We support this finding by demonstrating that SWING-RF promotes an edge with a distinct and singular delay (Fig. 3.8A). We also used SWING to predict directed edges of several *E. coli* sub-networks using cubic spline interpolated microarray datasets. Through cross-correlation analysis, we estimate time-delayed interactions in *in silico*, *E. coli*, and *S. cerevisiae* networks, and show that SWING performs better than baseline methods in modules with more frequent time-delayed edges, such as the *tdcABC* regulon.

Interestingly, the apparent time delay only partially explains improved performance, as SWING also promotes edges without apparent time delays in *in silico* and *in vitro* networks. This discrepancy may arise from our conservative approach for identifying time delays; a more liberal approach could assign time delays to a greater fraction of the promoted edges.

However, it is particularly challenging to estimate time delays for genes with multiple regulators using cross-correlation. More complex algorithms that incorporate additional information (*i.e.*, nonlinearity and partial correlation) could improve time delay estimation between regulators and targets [96].

An additional consideration involves interactions that occur faster than the sampling interval. These interactions will not exhibit a delay in the time series, and will resist inference and estimation of time delay regardless of methodology. This bottleneck can be managed by designing experiments with shorter sampling intervals. The choice of sampling interval is context specific, and we recommend sampling with sufficient frequency to capture dynamics of interest.

3.11.2. SWING outperforms common network inference algorithms across scales

SWING outperforms common network inference algorithms–RF, LASSO, and PLSR–but is limited by computational expense. Since SWING constructs a larger explanatory matrix and executes multivariate comparisons between multiple time delays, it is more expensive than the aforementioned methods. Fortunately, SWING is trivially parallelizable and can be implemented on any multicore processing system. We conducted similarly derived 100-node *in silico* networks and found that SWING increased the AUPR and AUROC for all three methods (Fig 3.3), including SWING-LASSO, which had no significant difference for the 10-node networks (Fig. 3.1A). Remarkably, every single network was inferred with greater accuracy, indicating that SWING has notable benefits for larger inference tasks (Fig 3.3, Table 3.1).

3.11.3. SWING is an extensible framework

Compared to other time-delayed inference algorithms, SWING is a flexible and extensible framework that is not limited to using a single statistical method. The SWING framework was implemented with RF, LASSO, and PLSR; it can be easily expanded to use other multivariate inference algorithms, including those that utilize prior information and heterogeneous data types [97]. Additional improvements can be made by incorporating complex weighting of methods for consensus analysis that leverage known weaknesses and biases of inference methods. Methods that involve empirical optimization of combination weights, such as those assessed in the DREAM challenge, are expected to substantially improve SWING performance [98].

Although we implemented SWING to infer interactions from gene expression data, the same Granger causality principles can be applied to a wide variety of contexts with temporal dynamics. Provided sufficient time-series data, we expect SWING to identify regulatory relationships in related intracellular signaling pathways, as well as broader fields such as ecology, social sciences, and economics. As the sensitivity/specificity of experimental tools increases and the cost of implementation decreases, we expect longer and higher resolution time-series data to become widely available. We expect this increase in time resolution to further improve the accuracy of SWING-based network inference, especially as the community continues to build on the SWING chassis. The SWING framework, with currently implemented methods, is available on GitHub:
(*https://github.com/bagherilab/SWING*).

CHAPTER 4

Flow cytometry-based characterization and screening of several megakaryocyte culture conditions from CD34$^+$ cells derived from umbilical cord blood

Work presented in this chapter is adapted from the following paper under review:

- Wu, J.J., Abbott, D.A., Terzioglu, M.K., Ranjan, R., Mahmud, D., Issa, H., Bagheri, N., Mahmud, N., and Miller, W.M. Multi-Phase Ex Vivo Generation of Platelet-like Particles from CD34$^+$ Cord Blood Cell-Derived Megakaryocytes.

4.1. Abstract

Generating platelets in culture from alternative cell sources may provide a platform toward donor-independent platelet transfusions. We present an augmented differentiation protocol that pre-expands the number of hematopoietic stem and progenitor cells (HSPCs) toward the megakaryocyte (MK) lineage using valproic acid (VPA), before generating platelet-like particles (PLPs). We show that the length of primary pre-expansion culture (P0) affects CD41a+CD42b+ expression in secondary culture (P1) and that VPA significantly increases the numbers of MKs and PLPs produced compared to pre-expanded cells without VPA. Our strategy generated 500 MKs and 10^4 CD41a+CD42b+ PLPs per input CD34$^+$ cell. We found that increasing pre-expansion time from 6 to 8 days and initial HSPC expansion concomitantly resulted in upregulation of p21Cip/Waf1 and p16INK4 protein levels, while VPA treatment decreased the extent of upregulation. We demonstrate that the resulting

PLPs exhibit functional activity. Notably, we report substantial donor-to-donor variability across expanded and unexpanded cultures. We investigated the extent to which total MK production could be predicted by early culture characteristics. Correlation analysis showed that %CD41a+ and %CD34$^+$ CD41a+ early in culture, as well as early indicators of total and CD34$^+$ cell expansion, are predictive factors of MK yields in secondary culture. Our multi-phase culture strategy provides the basis for additional experimentation to further enhance MK and PLP yields.

4.2. Multi-step production of CD41a$^+$CD42b$^+$ cells from CB CD34$^+$ cells.

Platelets are derived from megakaryocytes (MKs) and play major roles in hemostasis, inflammation, thrombosis, and vascular biology[3, 99, 100, 101]. Patients who undergo radiation, myelosupressive chemotherapy, or stem cell transplantation often suffer from thrombocytopenia and rely on platelet transfusions as supportive therapy. More than 2 million platelet transfusions are performed each year in the US, however platelet transfusions are expensive, have the potential for adverse reactions, and are a limited resource dependent on blood bank infrastructure for supply and distribution, as units are solely derived from volunteer donors [102]. Generating platelets from alternative sources such as banked stem cells would provide a platform towards donor-independent platelet transfusions.

Multiple strategies have been explored to generate platelet-like particles (PLPs) from human stem cell sources such as umbilical cord blood (CB), adult growth-factor-mobilized peripheral blood, embryonic stem cells, and induced pluripotent stem cells (iPSCs) [103, 104, 105]. Currently, strategies using unexpanded primary cells as starting material have produced insufficient numbers of PLPs for potential clinical use. Augmenting differentiation protocols by pre-expanding the number of hematopoietic stem and progenitor cells (HSPCs)

and directing them toward the MK lineage could increase PLP yield. However, whether expanded HSPCs differ in terms of MK production and PLP generation has not been evaluated. Strategies involving co-culture, inflammatory cytokines, or small molecule inhibitors have been shown to improve proliferation of donor HSPCs, yet many of these agents reduce MK differentiation of expanded cells and were not shown to improve MK or PLP yields [9, 106]. Adding histone deacetylase inhibitors, such as valproic acid (VPA), has been shown to increase the numbers of $CD34^+$ and $CD34^+$ CD90+ cells in a xenotransplant model, but whether augmenting ex vivo PLP-producing cultures with a pre-expansion phase improves MK and PLP yields has not been explored [107].

An essential, but often overlooked, component of generating PLPs ex vivo is characterization of donor heterogeneity for MK maturation and terminal stages of PLP release. Due in part to asynchronous maturation of hematopoietic cells in culture, differences in the peak output of various donor samples remains an important consideration for developing processes that yield meaningful levels of PLP production. Donor heterogeneity of CB units in terms of cell recovery and graft potency has been reported by several others [108]. For instance, transplant data suggests that $CD34^+$ progenitor graft content correlates with speed of engraftment and long-term survival [109, 110, 111, 112]. However, no studies to date have characterized the effect of CB sample heterogeneity on subsequent MK and PLP yields ex vivo. Identifying CB units with high and low MK potential early in the 20-day culture process can save expensive resources and provides the potential to intervene during culture.

We characterized several aspects of a novel serum-free, stroma-free, multi-phase process to produce PLPs from CD34-selected CB cells (Fig. 4.1). Our multi-phase culture strategy expands the number of $CD34^+$ cells towards the MK lineage, thus increasing numbers of mature MKs and PLPs per input $CD34^+$ cell several-fold in a donor-dependent manner.

We showed that pre-expansion with VPA increases MK production and ploidy, although with lower MK purity. We also characterized the synchrony, quantity, and functionality of PLPs collected over several days. Our results suggest that VPA increases MK production in part by decreasing levels of p21Cip/Waf1 and p16INK4 in pre-expanded cells. Finally, we demonstrate the ability to distinguish donor samples with high- versus low-MK potential, and show that high potential units generate up to 4 times more MKs per input $CD34^+$ cell than low potential units in secondary culture. Overall, this work addresses several challenges of generating MKs and PLPs ex vivo.

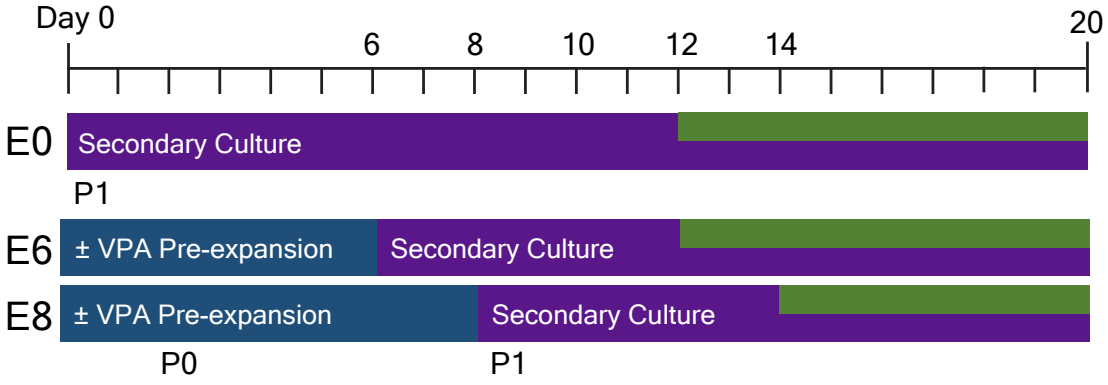

Figure 4.1 | **Timeline of ex vivo MK culture process illustrating durations of pre-expansion and secondary culture phases of E0, E6, E8 culture conditions.** Green bar represents the duration of fragmentation step when orbital shaking is applied to MK cultures.

4.3. Valproic acid (VPA) pre-expansion increases the number of $CD34^+$ cells for subsequent culture steps.

CB-derived CD34-selected cells were pre-expanded using cytokines Flt3-ligand, TPO, SCF, IL-3 in combination with VPA, which has been previously reported to expand $CD34^+$ CD90+ primitive, transplantable HSPCs in culture[107]. We compared the HSPC output of pre-expansion protocols with VPA (VPA+) to E6 and E8, in which the P0 period spans

6 and 8 days, respectively to E6 and E8 vehicle-expanded controls (VPA) and E0 controls without pre-expansion (Fig. 4.1). We monitored surface expression of CD34, CD41a, and CD42b throughout pre-expansion (P0) and secondary (P1) culture stages. On day 8, a higher proportion of cells was $CD34^+$ in VPA+ cultures compared to VPA and E0 controls (Fig. 4.2). E6 and E8 VPA treatment maintained similar proportions of $CD34^+$ CD41a+ cells, while pre-expansion without VPA significantly decreased the proportion of $CD34^+$ CD41a+ cells compared to E0 control (Fig. 4.3). E6 pre-expanded cultures yielded significantly higher peak $CD34^+$ cell levels compared to E0 cultures, while E8 trended towards yielding higher peak $CD34^+$ cell levels (Fig. 4.3). VPA treatment also increased the mean proportion and absolute production of $CD34^+$ cells compared to E0 culture, suggesting that a larger pool of primitive HSPCs persist throughout the culture (Fig. 4.4). With respect to vehicle controls, E6 and E8 VPA treatment significantly increased production of $CD34^+$ cells by day 8 (Fig. 4.5) and expansion of the primitive $CD34^+$ CD90+ cell subpopulation (Fig. 4.5). E6 and E8 pre-expansion increased peak production of $CD34^+$ cells (Fig. 4.4) and total nucleated cells (TNC) (Fig. 4.5) compared to E0, though significant differences between VPA and VPA+ treatments were only found for E8 $CD34^+$ cells. These results suggest that subsets of $CD34^+$, $CD34^+$ CD41a+, and $CD34^+$ CD90+ cells derived from $CD34^+$ CB cells can be greatly expanded and maintained with VPA pre-expansion.

4.4. Length of primary pre-expansion culture (P0) affects CD41a+CD42b+ expression in secondary culture (P1) under static and shear conditions.

The pre-expanded cell cultures were re-suspended in secondary culture (P1) geared towards MK progenitor proliferation and maturation as described [9]. In P1, numbers of CD41a+CD42b+ MKs per P0 $CD34^+$ input cell and %CD41a+CD42b+ cells progressively

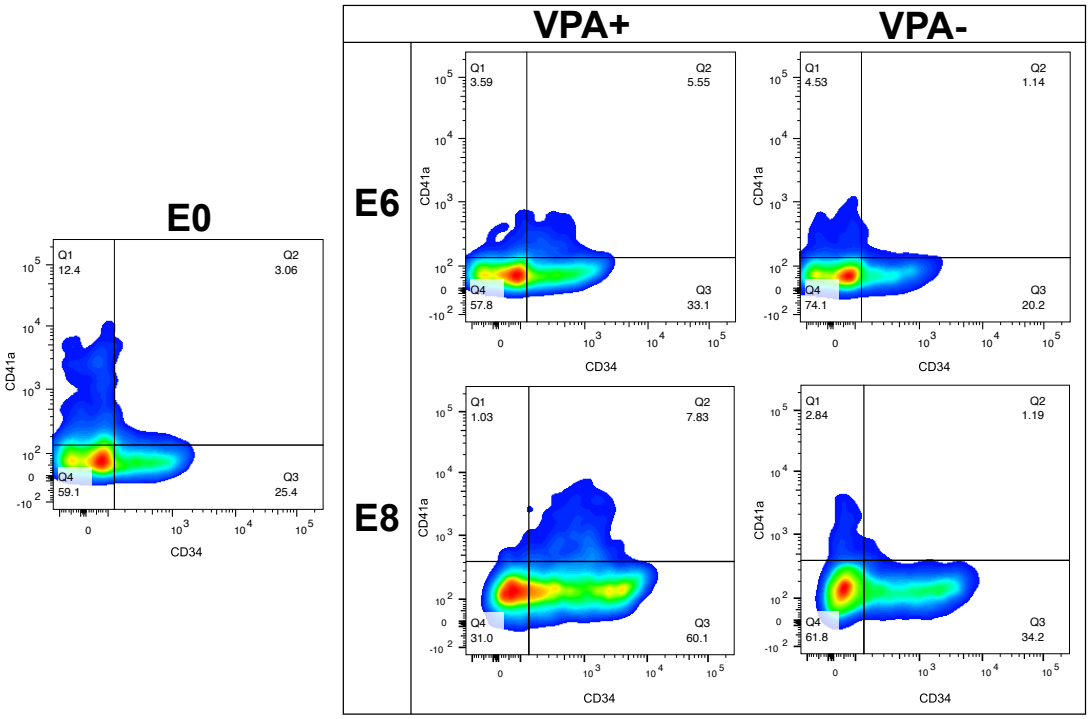

Figure 4.2 | **Flow cytometry density plots showing representative expression.** CD34 (x-axis) and CD41a (y-axis) on Day 8 (summative P0+P1) of culture for E0, E6, E8 and VPA+, VPA- treatments.

increased in all cultures until a peak day, then subsequently declined (Fig. 4.6). On average, E0 control cultures peaked on day 14 and quickly declined, while E6 and E8 VPA+ cultures tended to peak 2-3 days later and maintain high levels for extended periods in P1. E8 pre-expansion without VPA treatment decreased MK production and peak %CD41a+CD42b+ relative to E0 controls and E6 pre-expansion (Fig. 4.7), suggesting that pre-expanding HSPCs without VPA is detrimental to MK production. In contrast, E8 VPA+ pre-expansion significantly increased production of CD41a+CD42b+ MKs per $CD34^+$ input cell compared to E0 and E8 VPA- and exhibited peak %CD41a+CD42b+ cells intermediate between E0 and E8 VPA, suggesting that VPA partially rescues the effect of extended pre-expansion on MK commitment and proliferation. While E6 pre-expansion without VPA significantly

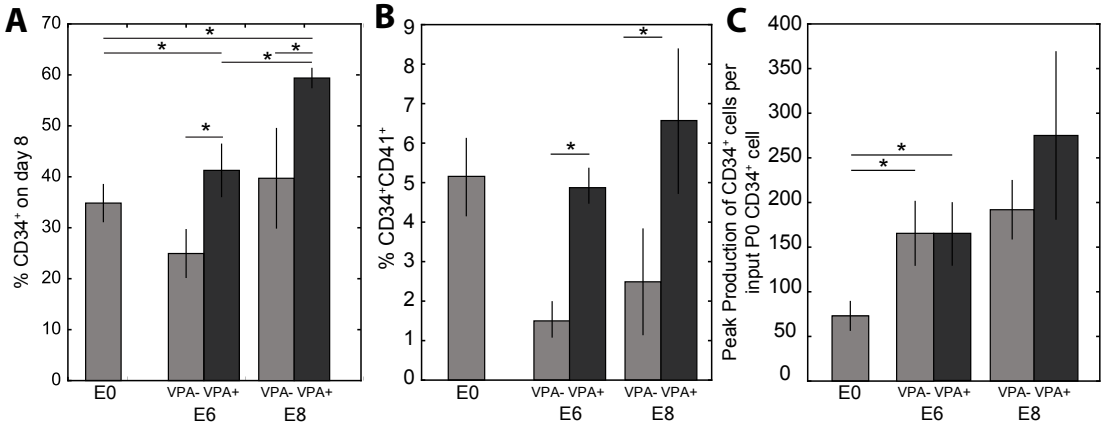

Figure 4.3 | **VPA enhances proportion of CD34$^+$ cells.** A. Bar graphs represent the mean (± SEM) percentage of CD34$^+$ cells for E0, E6, E8 and VPA+, VPA- treatments on day 8 (P0+P1) as measured by flow cytometry. E0: 35 ± 4; E6 VPA+: 41 ± 5; E6 VPA-: 25 ± 5; E8 VPA+: 60 ± 2; E8 VPA-: 40 ± 10; E0: n=6, E6: n=3, E8: n=6; (E0 vs. E6 VPA+)*, (E0 vs. E8 VPA+)*, (E6 VPA+ vs. E6 VPA-)**, E8 VPA+ vs. E8 VPA-: p = 0.12, (E6 VPA+ vs. E8 VPA+)**, (E0 vs E6 VPA-)**, (E0 vs E8 VPA-)**. ($p < 0.05$)*, ($p < 0.01$)**. B. Bar graphs represent the mean (± SEM) percentage of CD34$^+$CD41$^+$ cells for E0, E6, E8 and VPA+, VPA- treatments on Day 8 (P0+P1) as measured by flow cytometry. E0: 5.1 ± 1.0; E6 VPA+: 4.9 ± 0.5; E6 VPA-: 1.5 ± 0.5; E8 VPA+: 6.6 ± 1.8; E8 VPA-: 2.5 ± 1.4; E0: n=6, E6: n = 3; E8: n = 6; (E6 VPA+ vs. E6 VPA-)***, (E8 VPA+ vs. E8 VPA-)*, E0 vs E6 VPA-: p=.13, (E0 vs E8 VPA-)*. (p <0.05)*, (p <0.01)**, (p <0.001)***. C. Bar graphs represent the mean (± SEM) peak production of CD34$^+$ cells per input P0 CD34$^+$ cell for E0, E6, E8 and VPA+, VPA- treatments (P0+P1) as measured by flow cytometry and cell counts. E0: 73.0 ± 17; E6 VPA+: 165 ± 36; E6 VPA-: 166 ± 36; E8 VPA+: 275 ± 94; E8 VPA-: 192 ± 33; E0 and E6: n=9, E8: n = 6; (E0 vs. E6 VPA+)*, E0 vs. E8 VPA+: p=0.13, (E0 vs E6 VPA-)*, (E0 vs E8 VPA-)*. (p <0.05)* (p <0.01)** (p <0.001)***

reduced peak %CD41a+CD42b+ cells, MK production per P0 CD34$^+$ input cell was similar for E6 VPA+ and VPA (Fig. 4.7). As an added effect, we observed that placing cultures under shear to induce PLP release in P1 culture also improved or sustained MK purity (Fig. 4.8). Orbital shaking increased mean peak %CD41a+CD42b+ for E8 VPA+ without significantly affecting overall MK production (Fig. 4.8 and 4.9). In contrast, culturing E0 cells

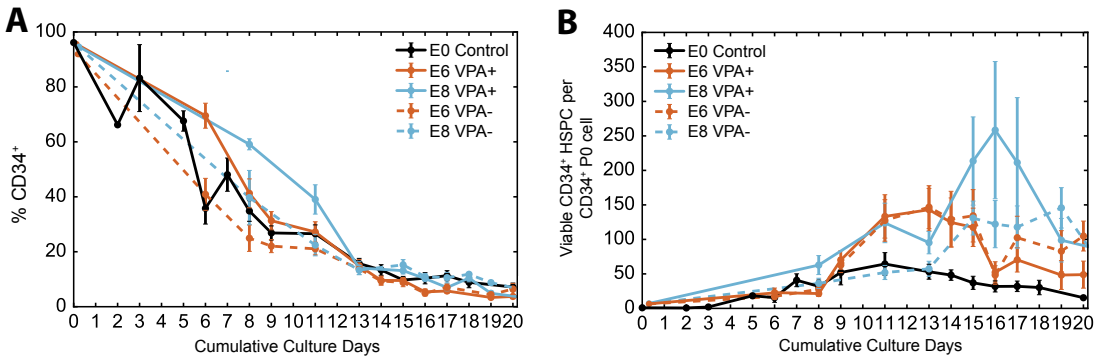

Figure 4.4 | **VPA maintains larger pool of CD34$^+$ cells in UCB.** A.Line graphs represent time-course of flow cytometry analysis of mean (± SEM) %CD34$^+$ cells during 20 days (P0 + P1) of culture (n=9 for E0 and E6; n=6 for E8). Line graphs represent time-course of mean (± SEM) cumulative production of CD34$^+$ cells per CD34$^+$ P0 input cell during 20 days (P0 + P1) of culture (n=9 for E0 and E6; n=6 for E8).

under orbital shear decreased MK production and did not increase peak %CD41a+CD42b+. Taken together, the results suggest that culturing VPA-pre-expanded cells under shear would be permissive towards continuous PLP harvest.

We assessed several qualities of P1 cells related to MK maturity, such as the degree of polyploidization and proplatelet-forming ability. MKs from all conditions exhibited increased cytoplasmic size, multi-lobular nuclei, and the ability to form -tubulin-positive long filaments with bulbous tips at terminal ends (Fig. 4.10). At the peak of maturity, we found no qualitative difference between conditions in terms of proplatelet morphology (Fig. 4.10), suggesting that pre-expansion and VPA treatment do not decrease MK proplatelet formation. Regarding polyploidization, we found that E6 VPA+ and E8 VPA+ pre-expansion yielded a significantly greater proportion of CD41a+ cells with >4N ploidy compared to E0 CD41a+ cells (Fig. 4.11). E8 VPA+ pre-expanded cultures also produced a greater proportion of polyploid CD41a+ MKs than E6 VPA+, despite a shorter duration in P1 culture. Thus, CB

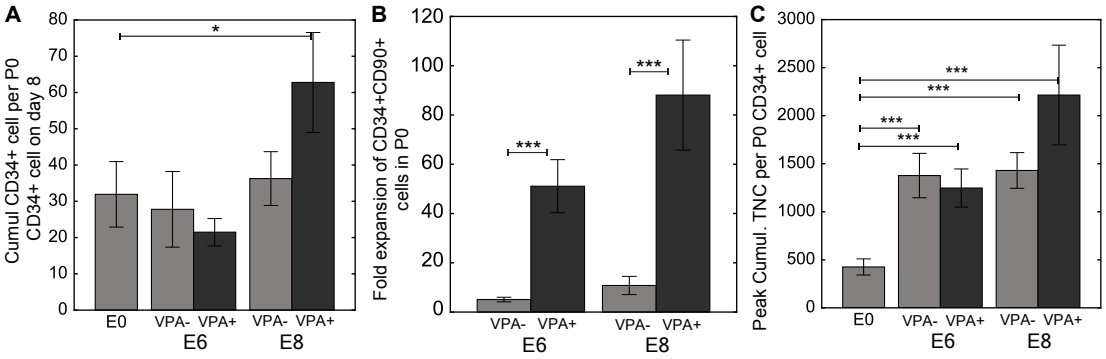

Figure 4.5 | **Cultures pre-expanded with VPA for 8 days showed greater expansion of CD34$^+$, CD34$^+$ CD90+, and total nucleated cells.** A. Bar chart shows significant changes in Day 8 CD34$^+$ cell production per P0 input CD34$^+$ cell for VPA+ and VPA- conditions for E0, E6, and E8 pre-expansions. E0: 40 ± 9; E6 VPA+: 22 ± 4; E6 VPA-: 28 ± 10; E8 VPA+: 63 ± 14; E8 VPA-: 36 ± 7. (E0 vs E8 VPA+)*, n=6, n=6, (E8 VPA+ vs E8 VPA-) p=0.16, n=6, n=6. (p <0.05)*. B. Bar chart shows significant difference in fold expansion of CD34$^+$ CD90+ cells in P0 for VPA+ vs. VPA- conditions for E6 and E8 pre-expansions. E6 VPA+: 51 ± 11; E6 VPA-: 5 ± 1; E8 VPA+: 88 ± 22; E8 VPA-: 11 ± 4. (E6 VPA+ vs E6 VPA-)**, n=9, n=8, (E8 VPA+ vs E8 VPA-)**, (E6 VPA+ vs E8 VPA+)*, (E6 VPA- vs E8 VPA-) p=0.12, n=8, n=8. (p <0.05)*, (p <0.01)**. C. Bar chart shows significant changes in peak cumulative TNC per P0 input CD34$^+$ cell for VPA+ and VPA- conditions for E0, E6, and E8 pre-expansions. E0: 426 ± 84; E6 VPA+: 1247 ± 198; E6 VPA-: 1377 ± 231; E8 VPA+: 2215 ± 519; E8 VPA-: 1430 ± 186 viable TNC per CD34$^+$ cell. (E0 vs E6 VPA+)**, (E0 vs E8 VPA+)**, (E6 VPA+ vs E8 VPA+) p = 0.17, (E0 vs E6 VPA-)**, (E0 vs E8 VPA-)**. (p <0.05)*, (p <0.01)**, (p <0.001)***.

CD34$^+$ cells pre-expanded either in the presence or absence of VPA display characteristic hallmarks of MK maturity.

4.5. E8 VPA+ pre-expanded cultures produce greater numbers of PLPs than VPA- conditions.

We noticed a great deal of variability in the timing of the peak and the extent of PLP release with E0 and VPA ± pre-expanded cultures (Fig. 4.12). We sampled PLP release every 24 hours from continuously shaken E0, E6, and E8 cultures and found that E8 VPA+

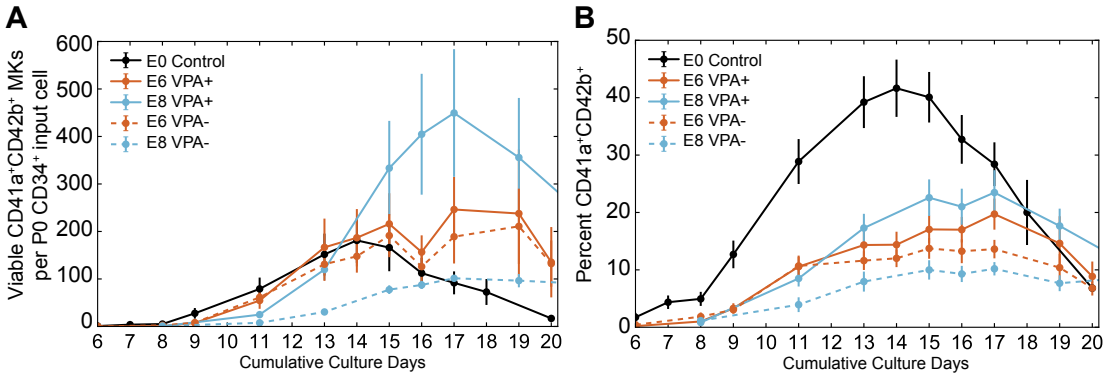

Figure 4.6 | **Pre-expansion with VPA increases CD41a+CD42b+ expression and cell production during P1 culture.** A.Line graphs represent time-course of mean (± SEM) CD41a+CD42b+ MKs generated per input P0 $CD34^+$ cell (n=9 for E0 and E6; n=6 for E8). B. Line graphs represent time-course for mean (± SEM) %CD41a+CD42b+ of live cell population quantified by flow cytometry (n=9 for E0 and E6; n=6 for E8).

cultures trended towards producing greater numbers of PLPs per P0 $CD34^+$ input cell than E6 pre-expansion or E8 VPA cultures (Fig. 4.12). The peak number of PLPs collected per MK counted on the previous day in culture tended to be greater for E0 than for E6 and E8 pre-expansions, although with substantial donor variation (Fig. 4.13). Additionally, the E8 VPA+ condition trended towards increased PLP release efficiency compared to E8 VPA- (Fig. 4.13). This suggests that VPA treatment partially rescues the negative effect of pre-expansion on PLP release.

The number of CD41a+CD42b+ PLPs in culture represents the combined effects of production and degradation. To measure PLP degradation, we separated MKs from PLPs via centrifugation on days 15 and 17 for E0/E6 and E8 cultures, respectively, and placed isolated PLPs in fresh media supplemented with P1 cytokines in orbital shear conditions. After 24 hours, less than 10% of CD41a+CD42b+ PLPs originally isolated were observed to be CD41a+CD42b+ for all conditions (Fig. 4.13). This suggests that the PLPs measured each day were not carried over to the next day. Thus, we estimated the cumulative number

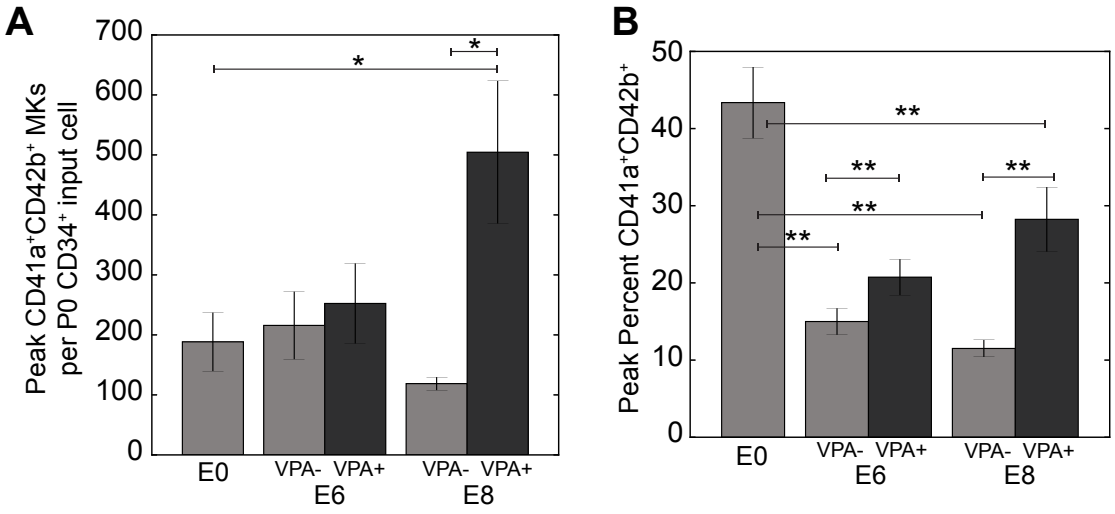

Figure 4.7 | **Pre-expansion with VPA increases peak CD41a+CD42b+ expression and cell production during P1 culture.** A. Bar graphs represent mean (± SEM) peak CD41a+CD42b+ MKs generated per input P0 $CD34^+$ cell for each culture condition (n=9 for E0 and E6; n=6 for E8). E0: 189 ± 49; E6 VPA+: 250 ± 66; E6 VPA-: 216 ± 56; E8 VPA+: 505 ± 119; E8 VPA-: 119 ± 11. (E0 vs E8 VPA+) p=.14, (E8 VPA+ vs E8 VPA-)*, (E6 VPA+ vs E8 VPA+)*, (E0 vs E8 VPA-) p=.14. (p <0.05)* (p <0.01)** (p <0.001)*** B. Bar graphs represent mean (± SEM) peak %CD41a+CD42b+ cells as measured by flow cytometry (n=9 for E0 and E6; n=6 for E8). E0: 43 ± 5; E6 VPA+: 21 ± 2; E6 VPA-: 15 ± 2; E8 VPA+: 28 ± 4; E8 VPA-: 12 ± 1. (E0 vs E6 VPA+)***, (E6 VPA+ vs E6 VPA-)***, (E0 vs E8 VPA+)***, (E8 VPA+ vs E8 VPA-)***, (E0 vs E6 VPA-)***, (E0 vs E8 VPA-)**. (p <0.05)* (p <0.01)** (p <0.001)***

of PLPs we could harvest per P0 $CD34^+$ cell by assuming harvest every 24 hours with complete recovery of MKs and progenitor cells after each harvest. Cumulatively, E8 VPA+ pre-expanded cultures produced around 104 CD41a+CD42b+ PLPs per P0 input $CD34^+$ cell, which is significantly greater than E8 VPA- cells and trends towards being greater than E0 cultures (Fig. 4.13). Similar PLP production per P0 $CD34^+$ cell in E0 and E8 VPA+ cultures is surprising because E0 cultures generally produced substantially fewer MKs overall, suggesting trade-offs between numbers of resulting MKs produced and PLPs released. Greater PL3P release in E0 and E8 VPA+ cultures than E6 VPA+ cultures was

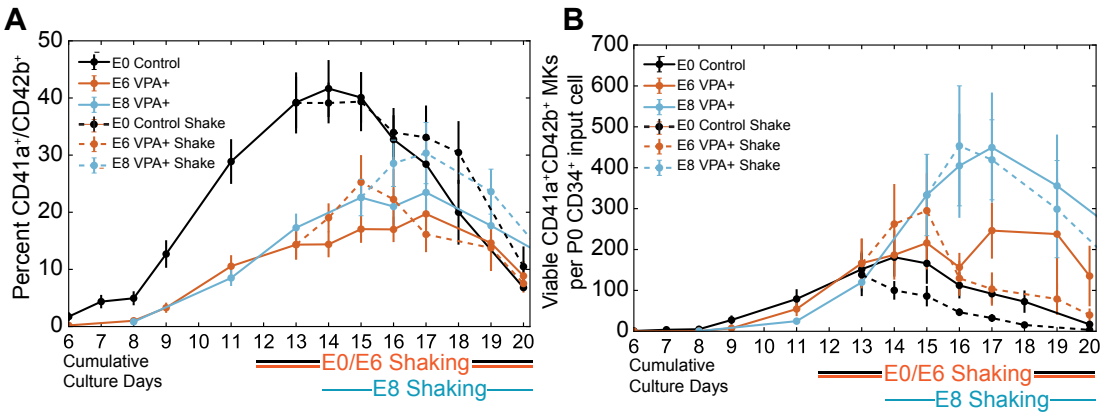

Figure 4.8 | **Effect of shear on pre-expanded cells.** A. Line graphs represent time-course of mean (± SEM) %CD41a+CD42b+ cells (n=9 for E0 and E6; n=6 for E8) comparing shear vs. static VPA+ conditions. B. Line graphs represent time-course of mean (± SEM) viable CD41a+CD42b+ MKs per P0 $CD34^+$ input cell comparing shear vs. static VPA+ conditions.

also observed using a microfluidic reactor (Fig. 4.14). Video analysis revealed that E8 VPA+ conditions released more PLPs per 5-minute interval than E6 VPA+ MKs, but less than E0 MKs for the same donor (Fig. 4.14). Taken together, VPA has a positive effect on overall PLP release compared to pre-expanded VPA- cultures and potentially unexpanded E0 cultures.

4.6. PLPs derived from VPA+ pre-expanded cultures exhibit functional activity.

Thrombin-activated PLPs from VPA+ pre-expanded cultures exhibited characteristic spreading of -tubulin and F-actin on fibrinogen-coated surfaces similar to E0 PLPs and donor platelets (Fig. 4.15). Activated PLPs were on average, much larger than donor platelets. To quantitatively determine whether culture-derived PLPs had functional activity comparable to donor platelets, we stained PLPs for PAC1 and CD62P before and after activation with thrombin (Fig. 4.16). Overall, CD41a+CD42b+ PLPs derived from VPA+

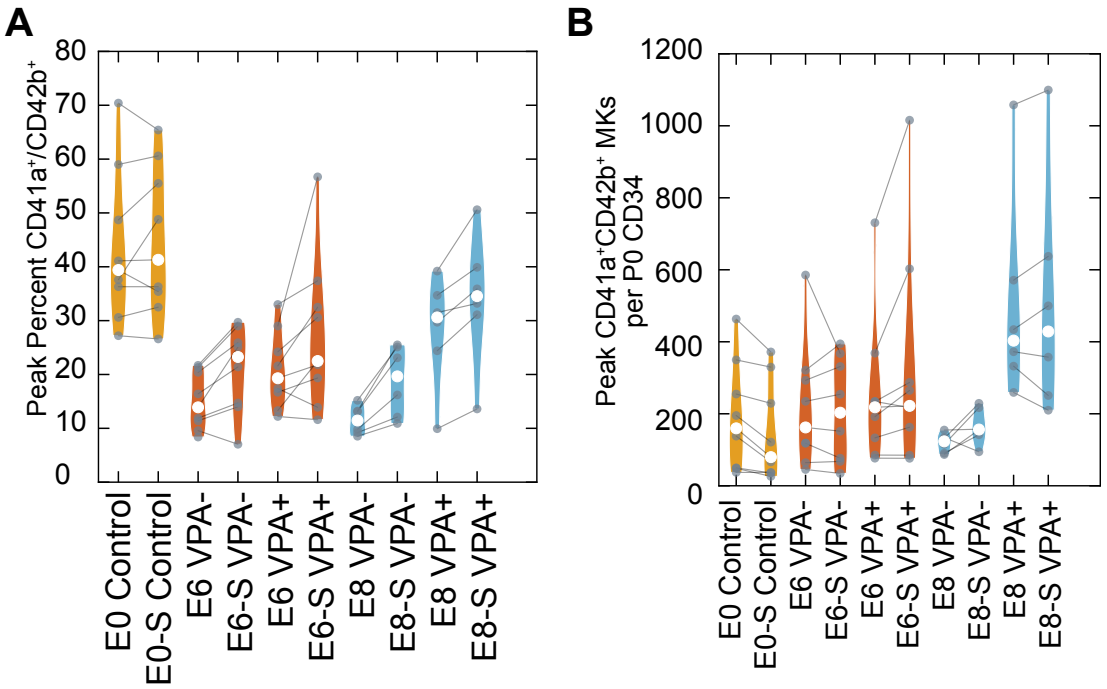

Figure 4.9 | **Effect of shear on individual pre-expanded donors.** A. Violin plots represent peak %CD41a+CD42b+ of each donor sample for E0, E6, E8 conditions under static and orbital shear conditions (shear stress = S, white dot = median, black bars = 25/75 quantiles, gray lines = same donor). Median (Interquartile Range or IQR): E0: 39 (36-49); E0-S: 41 (35-56); E6 VPA+: 19 (17-24); E6-S VPA+: 23 (19-33); E6 VPA-: 14 (11-20); E6-S VPA-: 26 (15-30); E8 VPA+: 31 (26-34); E8-S VPA+: 35 (32-39); E8 VPA-: 11 (9-13); E8-S VPA-: 20 (13-25). (E6 VPA+ vs E6-S VPA+) p=6.03E-02, n=9, n=9, (E6 VPA- vs E6-S VPA-)*, n=9, n=8, (E8 VPA+ vs E8-S VPA+)**, n=6, n=6, (E8 VPA- vs E8-S VPA-)**, n=6, n=6. B. Violin plots represent peak CD41a+CD42b+ MKs per P0 $CD34^+$ input cell (production) for each donor sample for E0, E6, E8 conditions under static and orbital shear conditions (shear stress = S, white dot = median, black bar = 25/75 quantiles, gray lines = same donor). E0: 160 (50-255); E0-S: 80 (37-229); E6 VPA+: 218 (133-234); E6-S VPA+: 222 (163-287); E6 VPA-: 162 (118-293); E6-S VPA-: 203 (74-340); E8 VPA+: 403 (342-537); E8-S VPA+: 429 (278-603); E8 VPA-: 123 (97-133); E8-S VPA-: 156 (145-202).

conditions did not differ significantly from E0 PLPs or donor platelets in terms of PAC1 and CD62P binding (Fig. 4.17). To measure ADP-dependent functional activity under shear conditions, we flowed PLPs through a fibrinogen-coated open channel microreactor

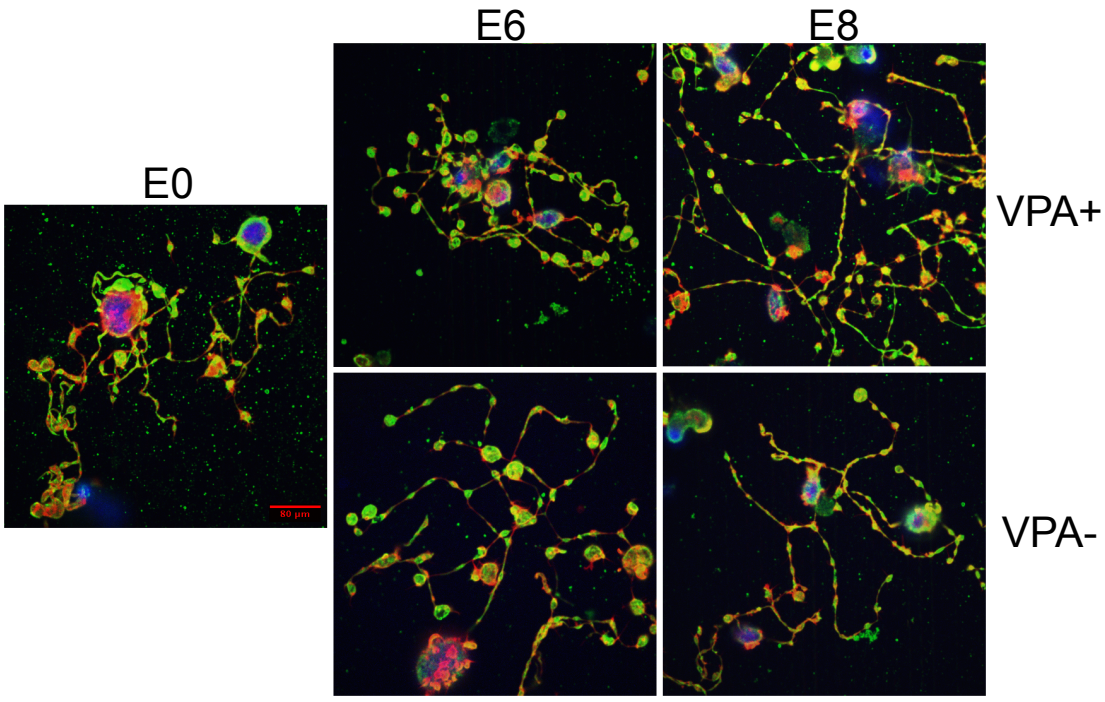

Figure 4.10 | **VPA pre-expansion does not appear to have an effect on ability to form proplatelets.** Representative fluorescence confocal microscopy images of proplatelet formation by MKs cultured from all conditions for donor CB560 (40X). Colors used: Blue—DAPI, Red—F-actin, Green—B-tubulin. The scale bar represents 80 μ m.

with and without ADP. PLPs isolated from all culture conditions produced calcein-labeled aggregates after 15 minutes (Fig. 4.18). We observed much less aggregation without ADP in a similar time-frame, confirming stimulus-specific aggregation. Altogether, this suggests that VPA pre-expansion primarily affects early-to-mid processes HSPC proliferation, MK differentiation, and PLP release and has limited effects on the functionality of released PLPs.

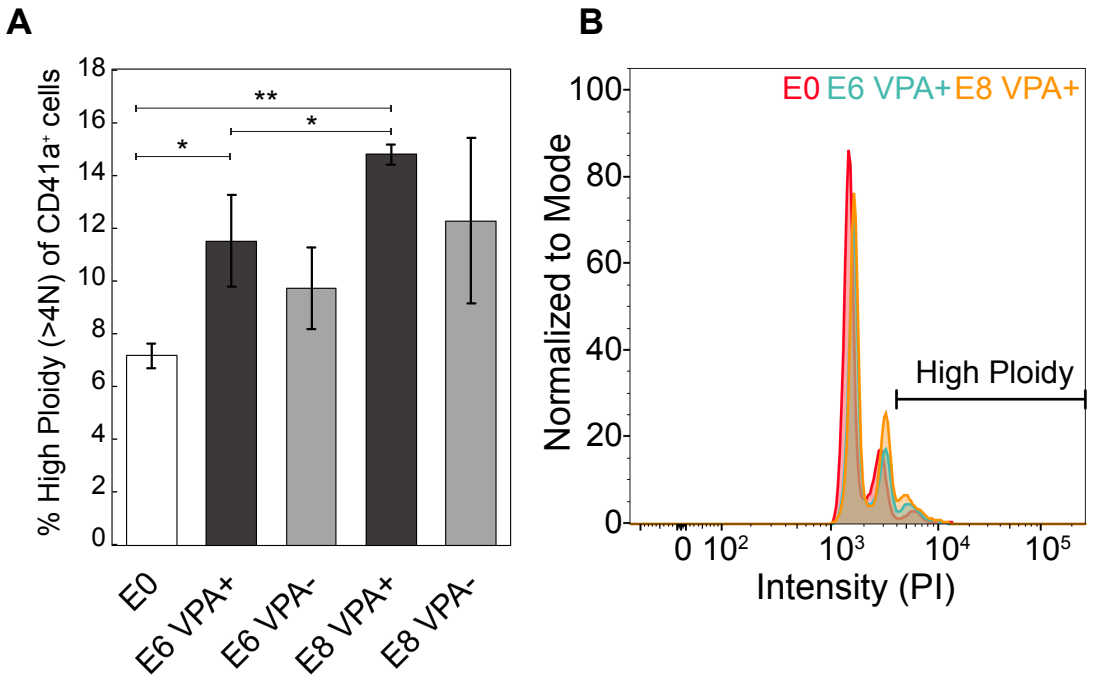

Figure 4.11 | **VPA pre-expansion increases polyploidization of CD41$^+$ cells.** A. Bar graphs represent the mean (± SEM) percent of CD41$^+$ cells that are high ploidy (>4N) on day 16 (P0+P1) for all conditions (N=6). Polyploidization quantified using flow cytometry and p-values were calculated using paired t-test. E0: 7.2 ± 0.4 (n=6), E6 VPA+: 12 ± 1 (n=6), E6 VPA- 10 ± 2 (n=3), E8 VPA+ 14.8 ± 0.3 (n=6), E8 VPA- 12 ± 4 (n=2). (E0 vs E6 VPA+)*, (E0 vs E8 VPA+)**, (E6 VPA+ vs E8 VPA+)**. B. Smoothed histogram shows gating and representative distribution of PI flow cytometry staining of CD41$^+$ cells for E0, E6 VPA+, E8 VPA+ for donor CB560.

4.7. p16INK4 and p21Cip/Waf1 are upregulated in pre-expansion conditions and downregulated with VPA treatment.

We hypothesized that VPA affects MK proliferation, differentiation, and polyploidization in processes related to cell cycle and senescence, potentially targeting endogenous inhibitors of cyclin-dependent kinases, p16INK4 and p21Cip/Waf1. We performed RT-qPCR comparing P0 input CD34$^+$ cells prior to culture with VPA+ and VPA pre-expanded cells at

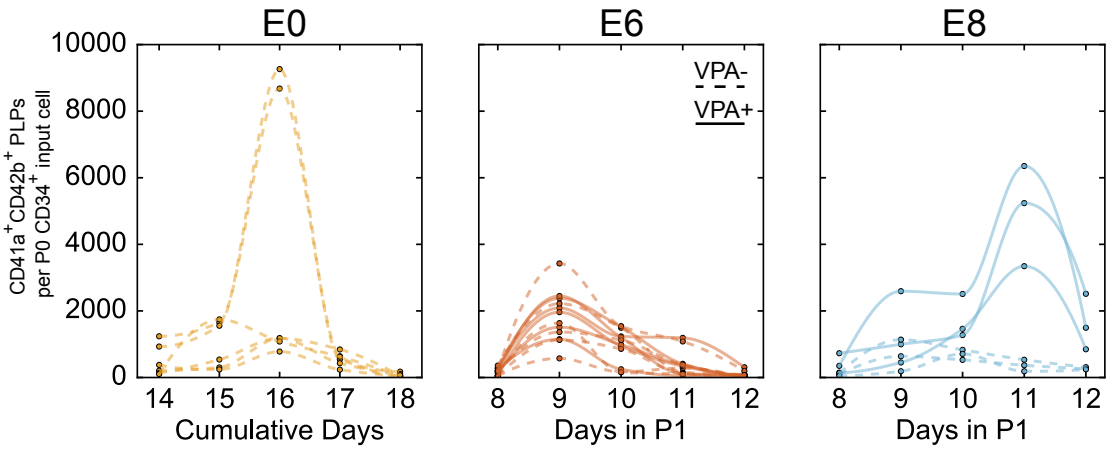

Figure 4.12 | **PLPs were collected from all conditions over multiple days.** A. Line graphs represent number of CD41a+CD42b+ platelet-like-particles (PLPs) quantified by flow cytometry per P0 $CD34^+$ input cell collected over the course of the orbital shear phase for E0 (n=9), E6 (n=9), E8 (n=6). Each line represents a separate CB unit.

day 20 and showed that overall transcript levels of p16INK4 and p21Cip/Waf1 were significantly decreased with VPA+ treatment on day 20 (Fig. 4.19). We then measured p16INK4 and p21Cip/Waf1 using intracellular flow cytometry on day 11 and found that longer pre-expansion substantially increased p16INK4 and p21Cip/Waf1 protein levels in CD41 cells and CD41-low cells (Fig. 4.20), which contain HSPCs and early MK populations, respectively, while VPA treatment generally decreased the proportion of cells expressing p16INK4 and p21Cip/Waf1 in these compartments early in P1 culture.

Histograms of p16INK4 and p21Cip/Waf1 for a single donor show that average expression of these proteins is substantially increased with pre-expansion, but reduced with VPA addition (Fig. 4.21 and 4.22). Together, this suggests that pre-expansion of HSPCs without VPA causes a proportion of cells to become senescent by increasing cyclin-dependent kinase (CDK) inhibitors p16INK4 and p21Cip/Waf1, thus lowering MK and cell division potential, and that VPA partially offsets the increase in p16INK4 and p21Cip/Waf1 expression.

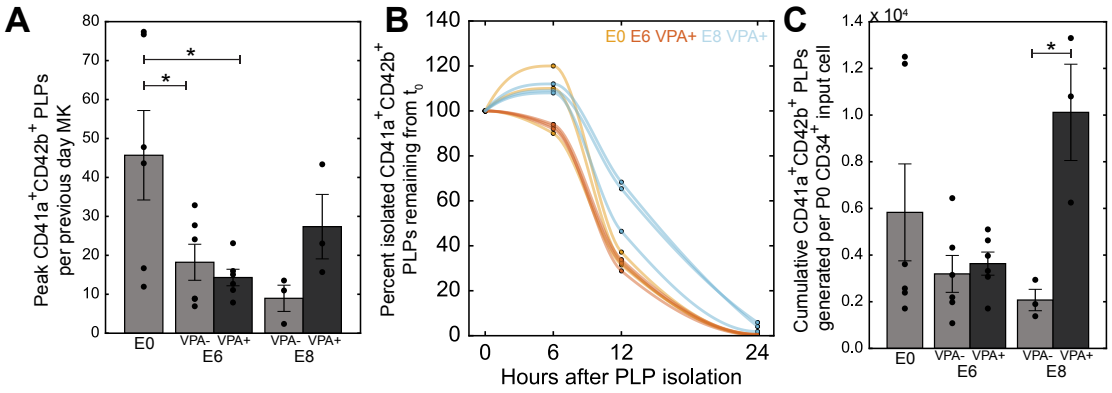

Figure 4.13 | **VPA increases PLP output compared to pre-expanded control.** A. Bar graphs represent the mean (± SEM) peak CD41a+CD42b+ PLPs collected per input MK (number of PLPs at t / number of MKs at t-1) quantified by flow cytometry. E0: 46 ± 12 (n=6); E6 VPA+: 14 ± 2 (n=6); E6 VPA-: 18 ± 5 (n=6); E8 VPA+: 21 ± 6 (n=3); E8 VPA-: 10 ± 3 (n=3). (E0 vs E6 VPA+)*, (E0 vs E6 VPA-)* B. Line graph shows percent isolated CD41a+CD42b+ PLPs remaining as a function of incubation time (PLPs at t / PLPs at t0). Each line represents PLPs generated from a separate donor sample (n=3) quantified by flow cytometry. C. Bar graph shows cumulative CD41a+CD42b+ PLPs collected per 104 P0 $CD34^+$ input cell (scale: 1x104). E0: 0.6 ± 0.2 (n=6); E6 VPA+: 0.4 ± 0.5 (n=6); E6 VPA-: 0.3 ± 0.8 (n=6); E8 VPA+: 1.0 ± 0.2 (n=3); E8 VPA-: 0.2 ± 0.1 (n=3). (E8 VPA+ vs E8 VPA-)*.

4.8. Substantial variability in P1 MK production can be predicted via early culture characteristics.

To improve efficiency of the culture process, we investigated whether the extensive variability observed in MK production for different CB units could be predicted using factors observed at earlier time-points. We observed that production of $CD34^+$ cells and MKs for different CB units varied greatly across multiple conditions, affecting both HSPC expansion and MK differentiation rates (Fig. 4.23). To delineate whether MK cultures belonged to a

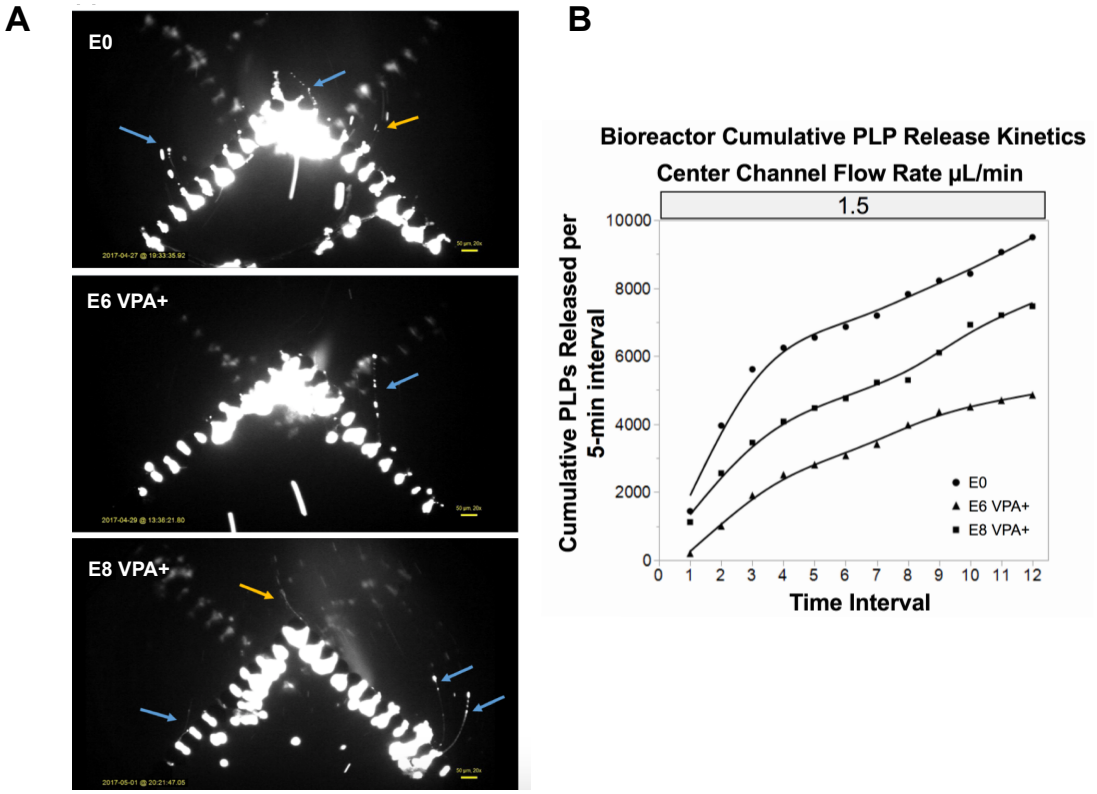

Figure 4.14 | **E8 VPA condition releases more PLPs per interval than that of E6 VPA, though unexpanded cells releases most PLPs per interval.** A. Fluorescence microscopy images show calcein-stained CD61+ selected MKs extending proplatelets under uniform shear conditions. Colored arrows point to extended proplatelets. B. Line graphs represent number of cumulative calcein-stained PLPs released and counted per 5-minute time interval in a microfluidic bioreactor for one donor comparing E0 and E6/E8 VPA+ pre-expanded conditions.

high-performing or low-performing group, MK production, purity, and TNC production trajectories for each CB unit were clustered using K-means. Clustering revealed two phenotypes that we named High-MK and Low-MK (Fig. 4.24).

A linear mixed-effect model was implemented to determine whether CB units belonging to the High-MK group displayed significantly greater MK production than those in the Low-MK group. Within the E6 VPA+ condition, High-MK cultures produced 4.6-fold more MKs

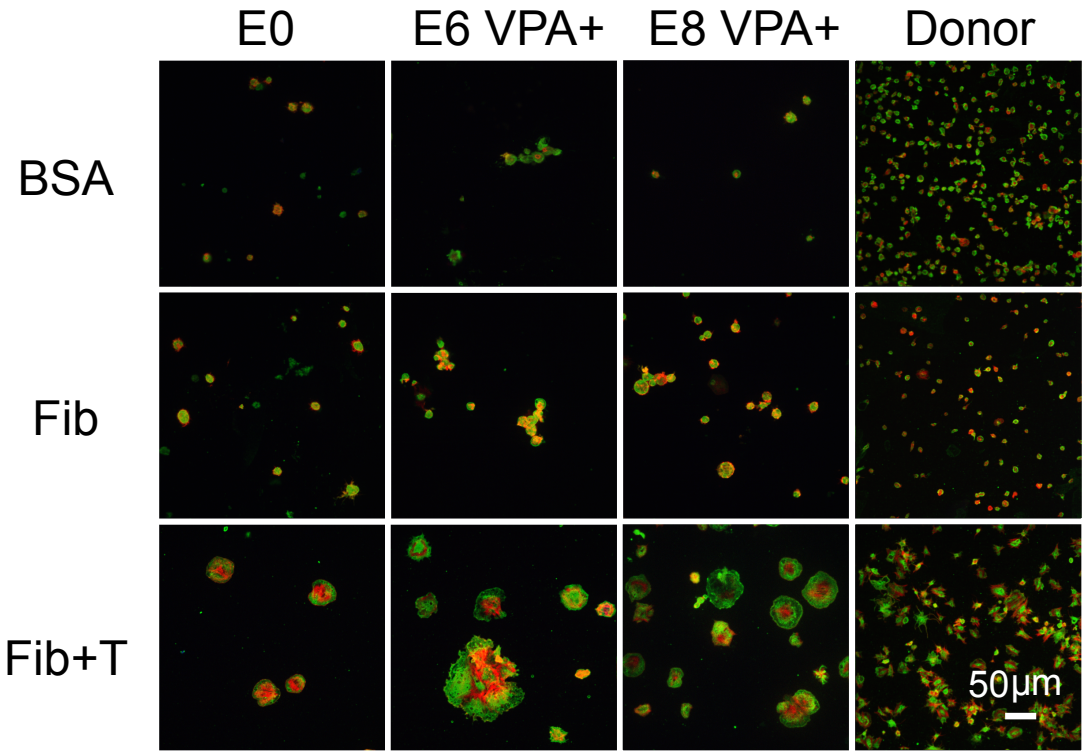

Figure 4.15 | **PLPs derived from VPA pre-treated cells display characteristic spreading.** Fluorescence confocal microscopy images (40x) show morphology of PLPs isolated from cultured MKs (E0, E6 VPA+, E8 VPA+) compared to that of donor platelets on coated surfaces: BSA, fibrinogen (Fib), fibrinogen with thrombin activation (Fib+T). Scale bar represents 50 μ m.

in P1 culture than Low-MK cultures (Fig. 4.25 p=0.003). The effect of donor variability was attenuated within the E8 VPA+ condition for which High-MK cultures produced 1.9-fold more MKs in P1 than Low-MK cultures (Fig. 4.25, Fig. 4.24; trending p=0.08). Differences between High-MK and Low-MK cultures were still seen when taking into account overall production of MKs from P0 $CD34^+$ cells (Fig. 4.25; E6, p=0.009; E8, p=0.2). We investigated whether paired, corresponding E0 cultures in the same Low-MK and High-MK groups displayed the same distinctions. E0 cultures of the High-MK CB group produced twice as many MKs as the Low-MK group (Fig. 4.26; trending p=0.07). Thus, clustering of

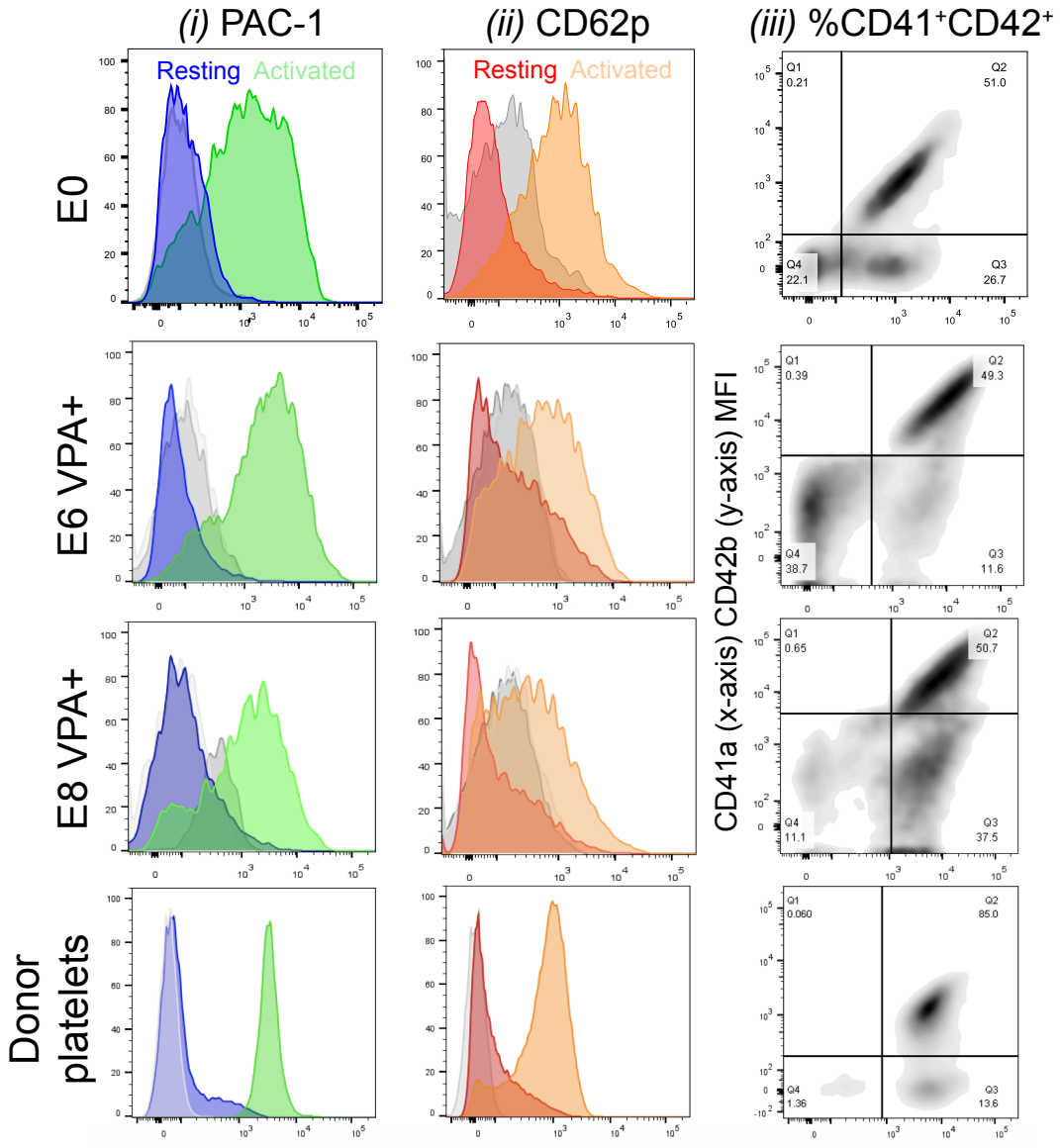

Figure 4.16 | **PAC1 and CD62P activation of VPA-PLPs appear to be similar to E0.** Histograms of (i) PAC-1, and (ii) CD62P fluorescence intensity for resting and activated (with thrombin) PLPs from E0, E6 VPA+, and E8 VPA+ for a representative culture compared to donor platelets. Gray histograms represent isotype staining. (iii) Dot plot for CD42b (y-axis) vs. CD41a (x-axis) fluorescence from E0, E6 VPA+, E8 VPA+ PLPs, and donor platelets. Gates were set based on isotype staining.

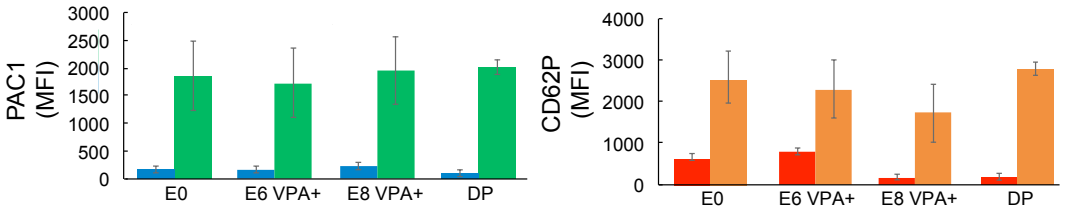

Figure 4.17 | **PAC1 and CD62P mean activation appears to be similar.** A. Bar graphs represent quantification of PAC1 and CD62P mean (± SEM) fluorescence intensity of resting and thrombin-activated states for E0, E6 VPA+, E8 VPA+ culture-derived PLPs and donor platelets.

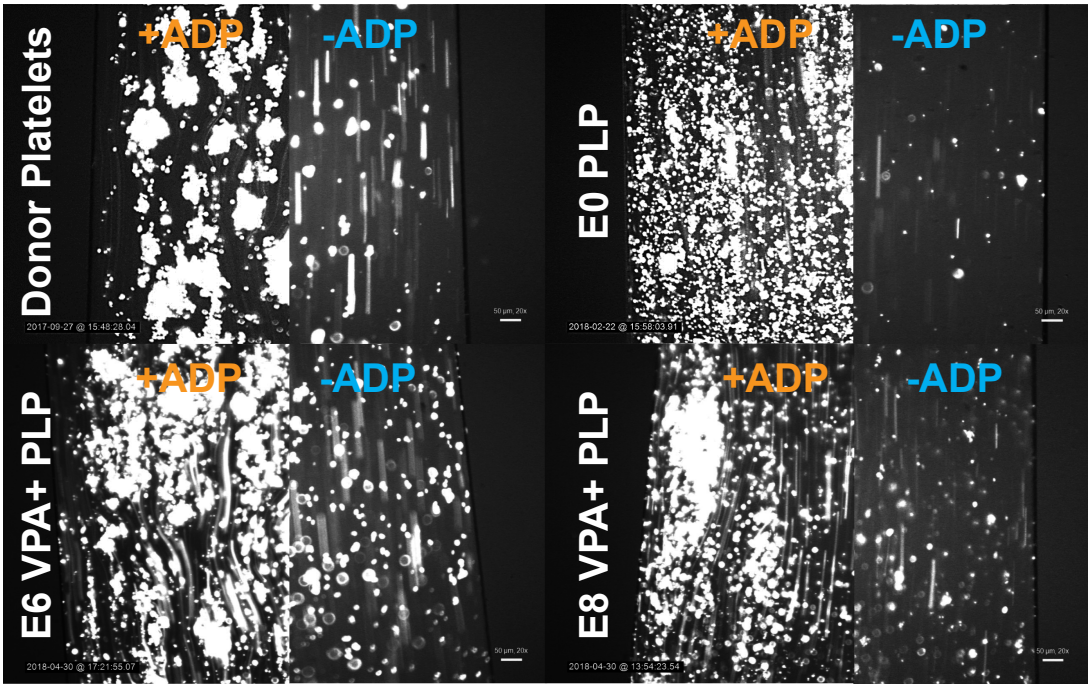

Figure 4.18 | **Aggregation assay reveals that all derived PLPs aggregate in response to ADP agonist.** A. In situ fluorescence microscopy shows culture-derived PLPs and donor platelets forming clots after 15 minutes of directional flow in an open-channel reactor coated with 60 ng/ml fibrinogen with 20 uM ADP (+ADP; left) and without ADP (ADP; right). Scale bar represents 50 μm.

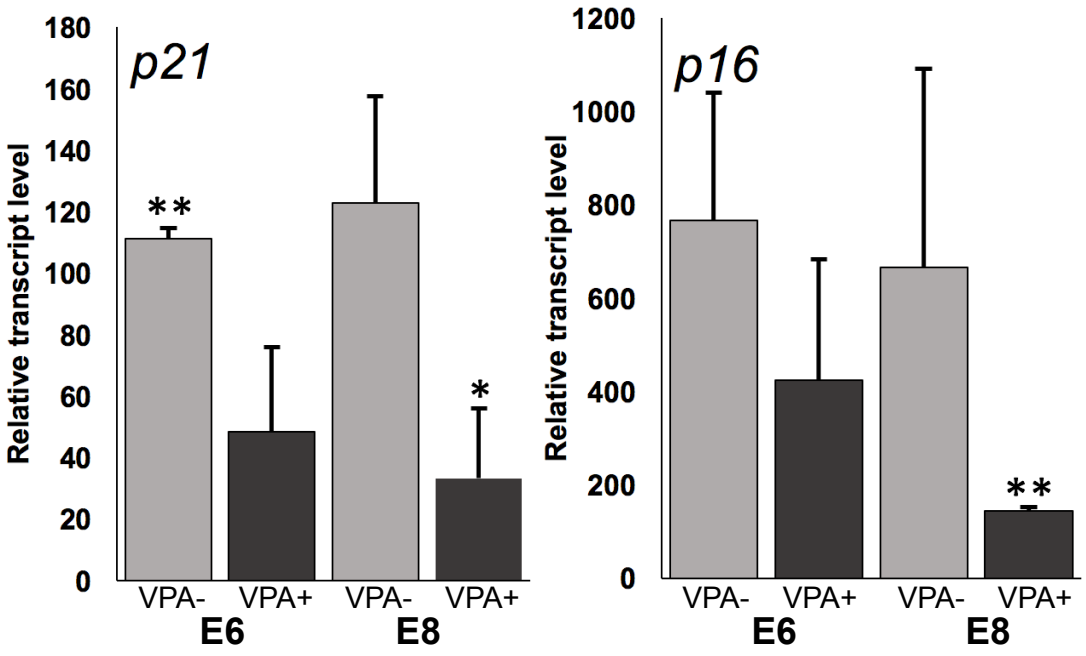

Figure 4.19 | **Transcript levels of p21 and p16 decrease with VPA treatment.** Bar chart shows mean (± SEM) relative transcript levels of p21 and p16 measured by real-time quantitative PCR for samples of E6 and E8 cultures pre-expanded with or without VPA on day 20 (summative P0+P1) of culture (n=3 for E6-derived cells, n=2 for E8-derived cells). Statistical significance *$p<0.05$ vs. day 0 control. **$p<0.01$ vs. day 0 control

production trajectories for individual CB units is similar for different culture processes and appears to be a cell-intrinsic property.

We investigated to what extent MK production and purity could be predicted by donor early culture characteristics. Correlation analysis for E0, E6 VPA+, and E8 VPA+ conditions in which we aggregated pairwise correlations between 38 observed variables revealed multiple factors measured in P1 and P0 that correlated with overall MK production (E0: n=10, E6: n=20, E8: n=17, Fig. 4.27).

Significantly, we observed a positive correlation between CD41a+ cell production early in P1 culture and peak MKs subsequently produced in culture (Fig. 4.28). Additionally,

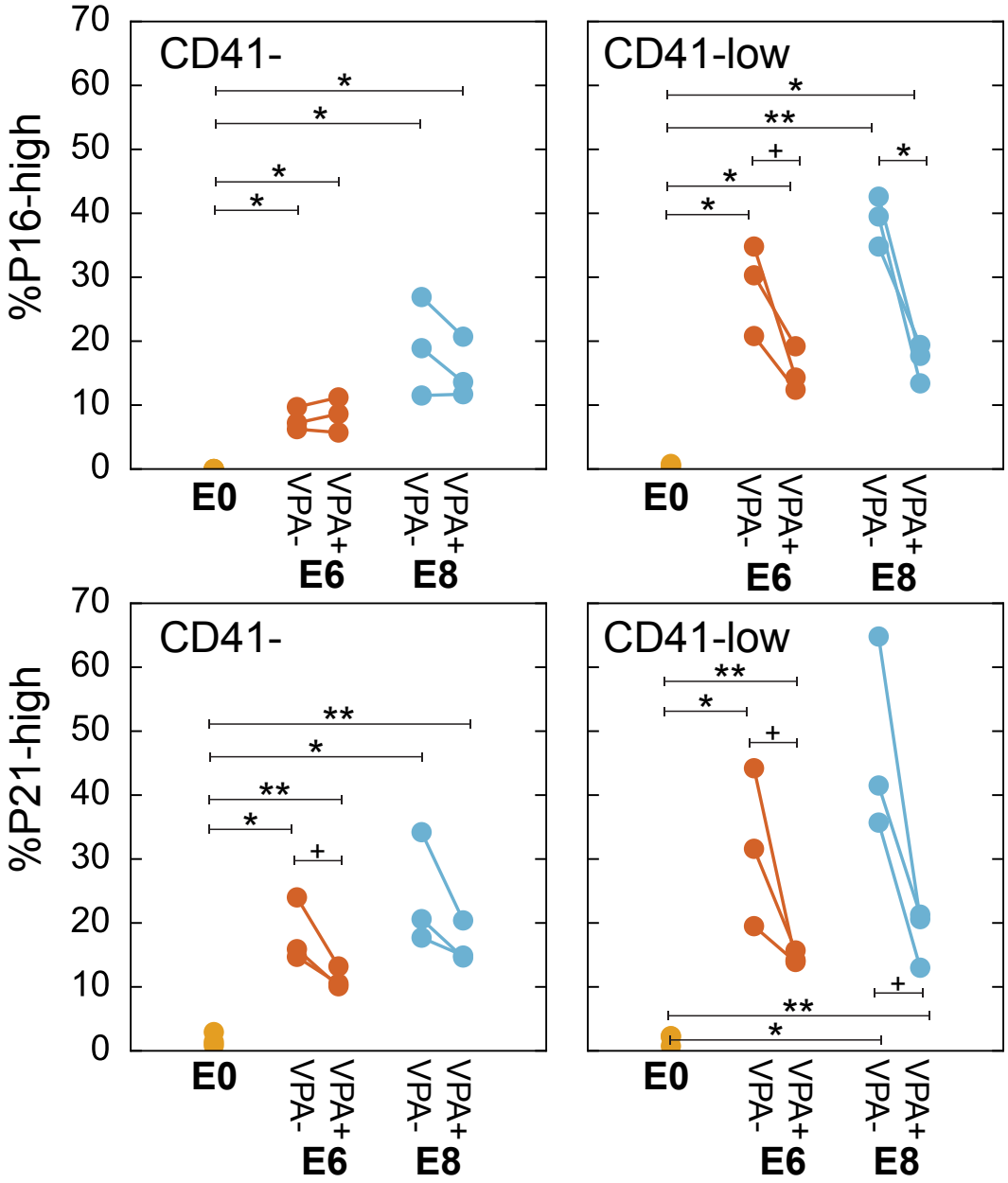

Figure 4.20 | **Protein levels of p21 and p16 increase with greater pre-expansion and decrease with treatment of VPA.** Paired plots show flow cytometry measurements of intracellular p16INK4 (p16) and p21Cip/Waf1 (p21) protein levels. Percent p16+ or p21+ of CD41 cells and CD41-low cells for E6 and E8 VPA+ and VPA- treatments, and E0 for cumulative day 11 of culture (n=3 donors).

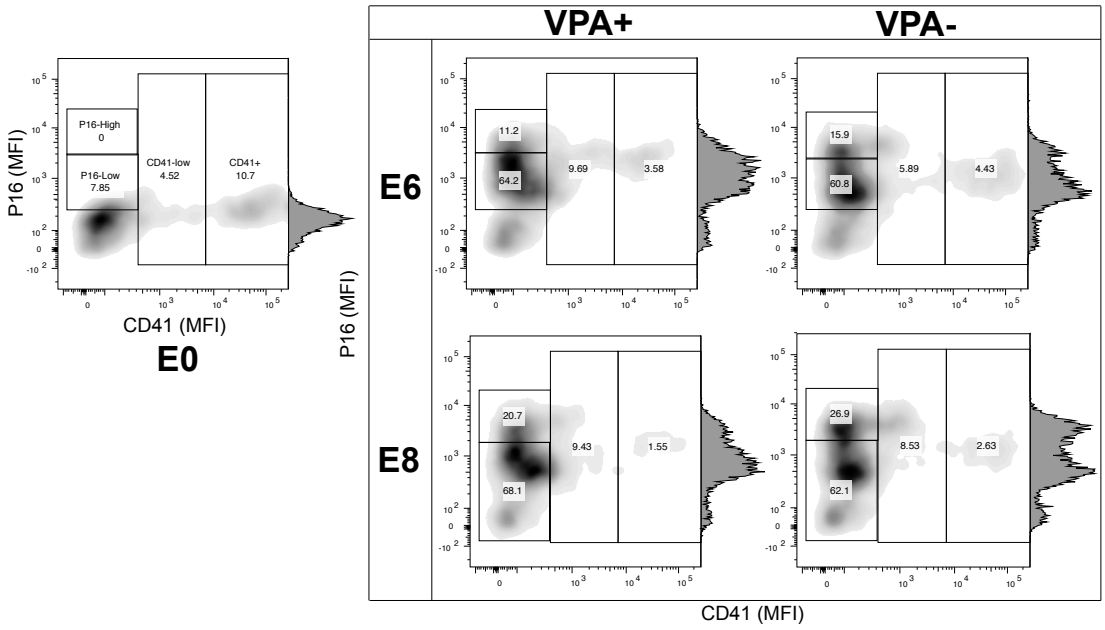

Figure 4.21 | **P16 gating strategy showing subpopulations of CD41$^+$ cells that are p16+.** Representative flow cytometry plots of intracellular p16INK4 (p16) vs surface CD41a for one donor. Gates show partitions for CD41, CD41-low, and CD41$^+$ cells, as well as p16-low, p16-high. Adjacent histogram shows median levels of p16.

peak purity of CD41a+CD42b+ cells correlated with initial %CD41$^+$ cells as early as day 0 of P1 (Fig. 4.28B) with greater correlation on days 3 and 5 in P1 (Fig. 4.28C and Fig. 4.27). This suggests that cell populations are pre-committed towards their current trajectory of producing more or less MKs regardless of VPA treatment or pre-expansion. Interestingly, peak TNC production is negatively correlated with MK commitment (Fig. 4.28D), suggesting that greater peak TNC production is indicative of greater rates of non-MK cell differentiation (Fig. 4.27). Thus, we identified multiple variables related to MK commitment and growth capability which are also correlated to MK production (Fig. 4.29).

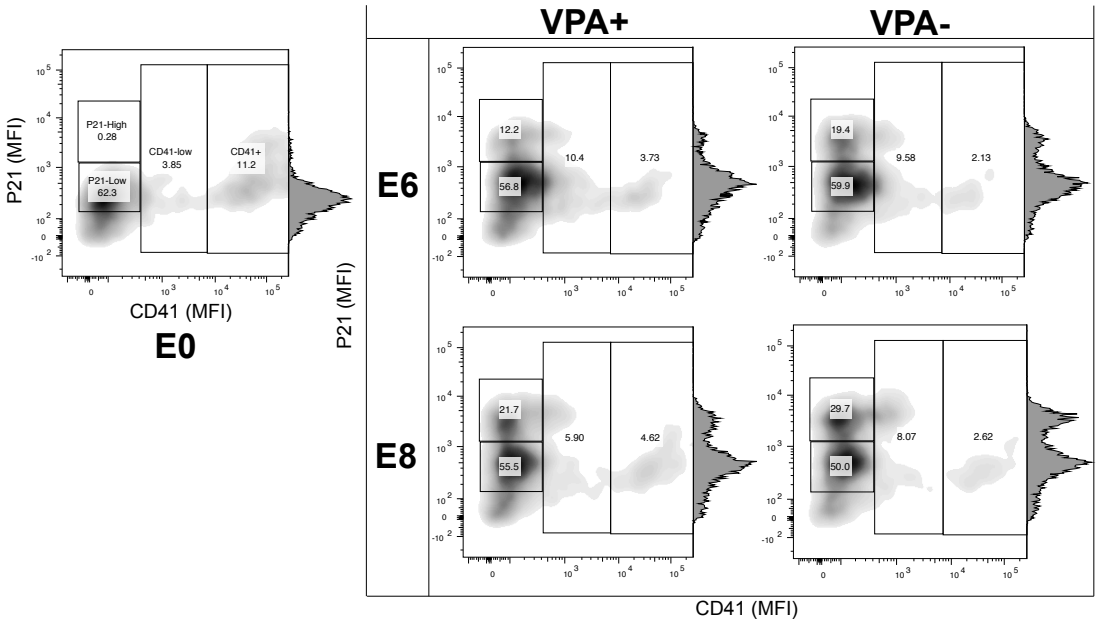

Figure 4.22 | **P21 gating strategy showing subpopulations of CD41$^+$ cells that are p21+.** A. Representative flow cytometry plots of intracellular p21Cip/Waf1 (p21) vs surface CD41a for one donor. Gates show partitions for CD41, CD41-low, and CD41$^+$ cells, as well as p21-low, p21-high. Adjacent histogram shows median levels of p21.

4.9. Discussion

In this study, we demonstrate a multi-phase culture system to generate PLPs from CB-derived CD34$^+$ cells. In particular, we tested two pre-expansion periodswith and without VPAto investigate and highlight relative improvements in fold-expansion with VPA and the subsequent improvement in MK and PLP yields. Many aspects of MKs and PLPs were extensively characterized during culture to show that VPA pre-expansion primarily affected early cell division, MK differentiation, and proplatelet extension and release, while minimally affecting PLP production and functionality. This reflects previous observations

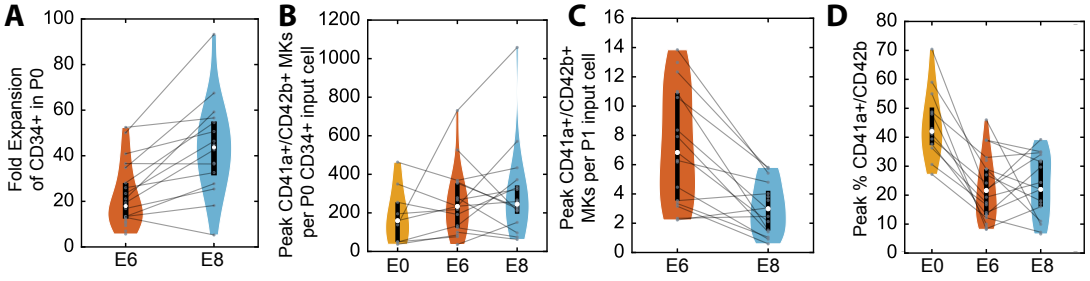

Figure 4.23 | **Extensive donor heterogeneity can be clustered into high-MK and low-MK groups.** A. Violin plots represent CD34$^+$ cell fold change on cumulative day 6 (E6) or day 8 (E8) of culture of VPA-treated conditions for each CB unit (E6: n=20, E8: n=16). White dot = median, black bars = 25/75 quantiles, gray lines = same donor. E6 VPA+: 18 (12-28); E8 VPA+: 43 (28-55). B. Violin plots represent peak CD41a+CD42b+ MK production in P1 culture per P0 input CD34$^+$ cell for E0 and VPA-treated conditions for each CB unit (E0: n=9, E6: n=16, E8: n=16); E0: 160 (50-255); E6 VPA+: 234 (122-354); E8 VPA+: 245 (195-342). C. Violin plots represent peak CD41a+CD42b+ MK production in P1 culture per P1 input cell for E0 and VPA-treated conditions for each CB unit (E6: n=16, E8: n=16). Median (IQR); E6 VPA+: 7 (4-11); E8 VPA+: 3 (2-4).

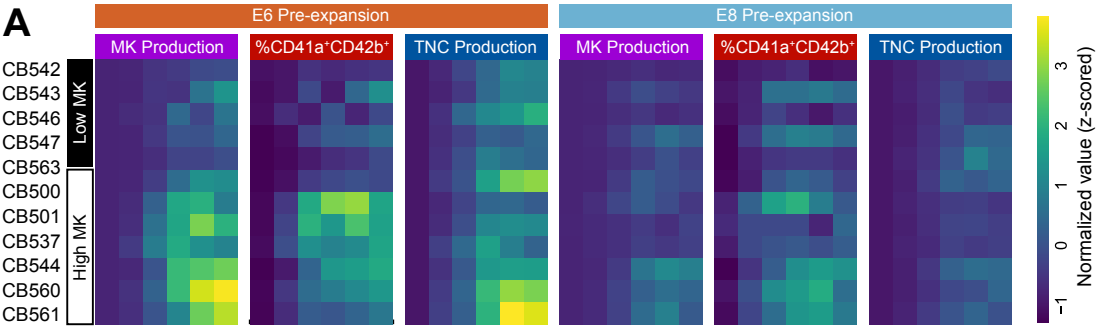

Figure 4.24 | **K-means clustering of pre-expanded growth.** Heatmap shows the aggregated data matrix that clustered 3 variables from 6 time-points and 2 conditions for N = 12 cultures from different donor samples.

that a proportion of functional CD42b+ particles released from culture-derived MKs are pre-PLPs (larger than donor platelets) rather than PLPs (comparable in size to donor platelets) [113].

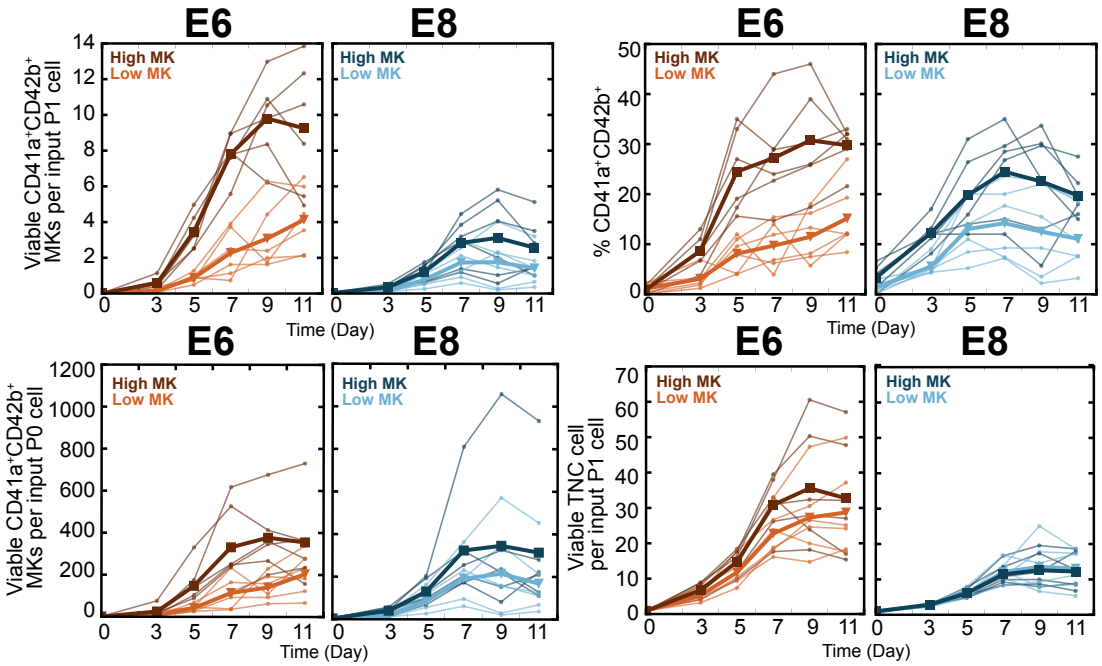

Figure 4.25 | **Linear mixed effect modeling shows significant difference between high and low MK groups in E6 and E8 expansions.** Line graphs shows clustering and significant changes between high- and low-MK groups during P1 culture for E6 and E8 VPA+ conditions. Thick lines represent the means of the High-MK (square) and Low-MK (triangle) group. N=5 (low-MK), N=7 (high-MK).

Compared to other studies using CB or adult peripheral-blood-derived $CD34^+$ cells, we demonstrate several-fold higher MK/PLP yields. Our modified E8 VPA+ pre-expansion process generated on average 500 mature CD41a+CD42b+ MKs per starting input $CD34^+$ cell, which is several-fold more MKs than for respective VPA and unexpanded controls. Others using CB or adult $CD34^+$ cells have demonstrated at most 130 MKs per input CB-derived cell and 50 MKs per adult-derived cell respectively [9, 114, 115, 116, 117, 13]. Additionally, our method generated PLPs at efficiencies similar to, or better than, other reported methods utilizing orbital shaking. We generated 20 PLPs/MK and about 104 CD41a+CD42b+ PLPs cumulatively per P0 $CD34^+$ input cell with continuous orbital

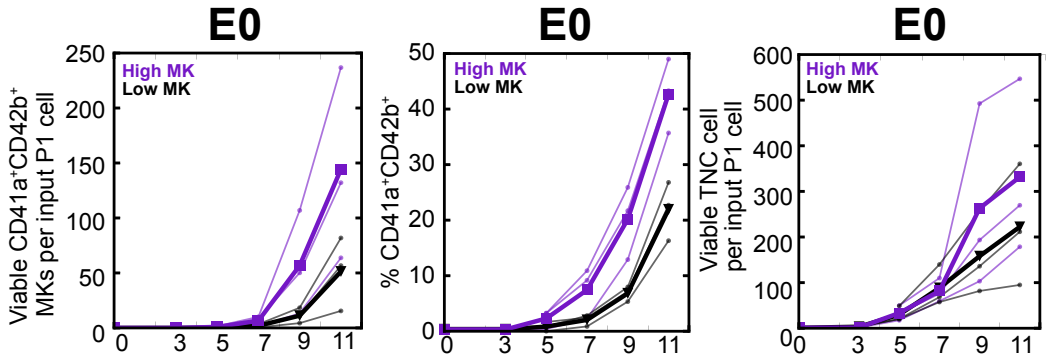

Figure 4.26 | **Linear mixed effect modeling shows significant difference between high and lower MK groups in E0 expansion.** Line graph shows clustering and significant changes between high- and low-MK groups for cultures that have a paired E0 condition. Thick lines represent the means of the High-MK (square) and Low-MK (triangle) group. N=3 (high-MK), N=3 (low-MK).

shaking, highlighting the potential of these pre-expanded CB cells to produce PLPs. The highest reported yield of PLPs per input MK or HSPC is 3.4x10ˆ4 PLPs per input HSPC, but this reflects the number of $CD41^+$, rather than CD41a+CD42b+, PLPs [103]. The co-expression of CD42b in $CD41^+$ PLPs is a critical attribute of functional platelets. Studies utilizing iPSCs or immortalized MK cell lines have generated 5-18 PLPs per input MK progenitor [118, 36, 119, 37]. In a recent study, Ito and colleagues improved PLP yields to 70-80 PLPs per input MK progenitor with a turbulence-based bioreactor, which indicates that downstream PLP collection may be greatly improved with better downstream processing. VPA pre-expansion can be used in conjunction with perfusion or turbulent bioreactors that recapitulate physical forces in endogenous blood flow, thus increasing numbers of PLPs per input MK [37].

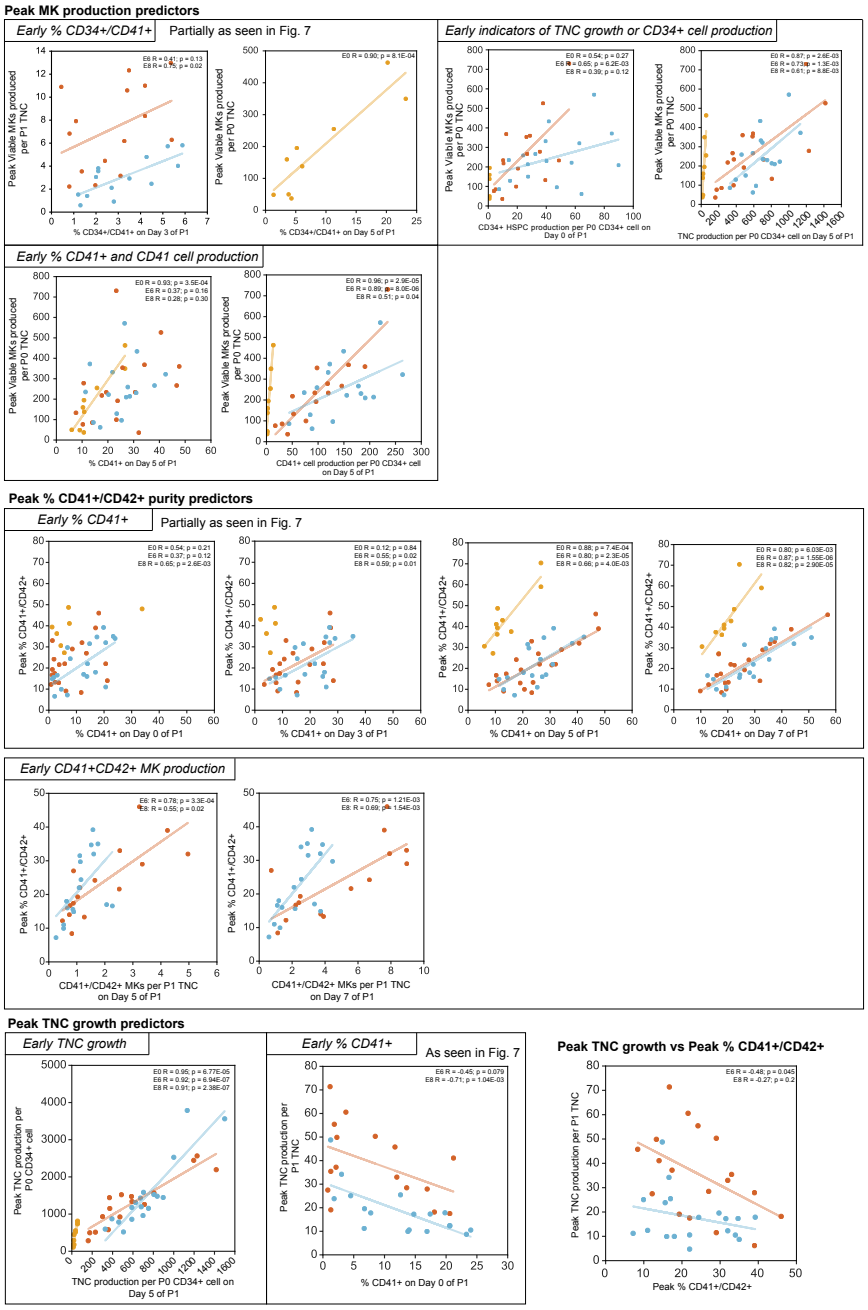

Figure 4.27 | **Correlation analysis between culture response variables.** Panel of scatterplots showing correlations between several culture response variables for different culture conditions.

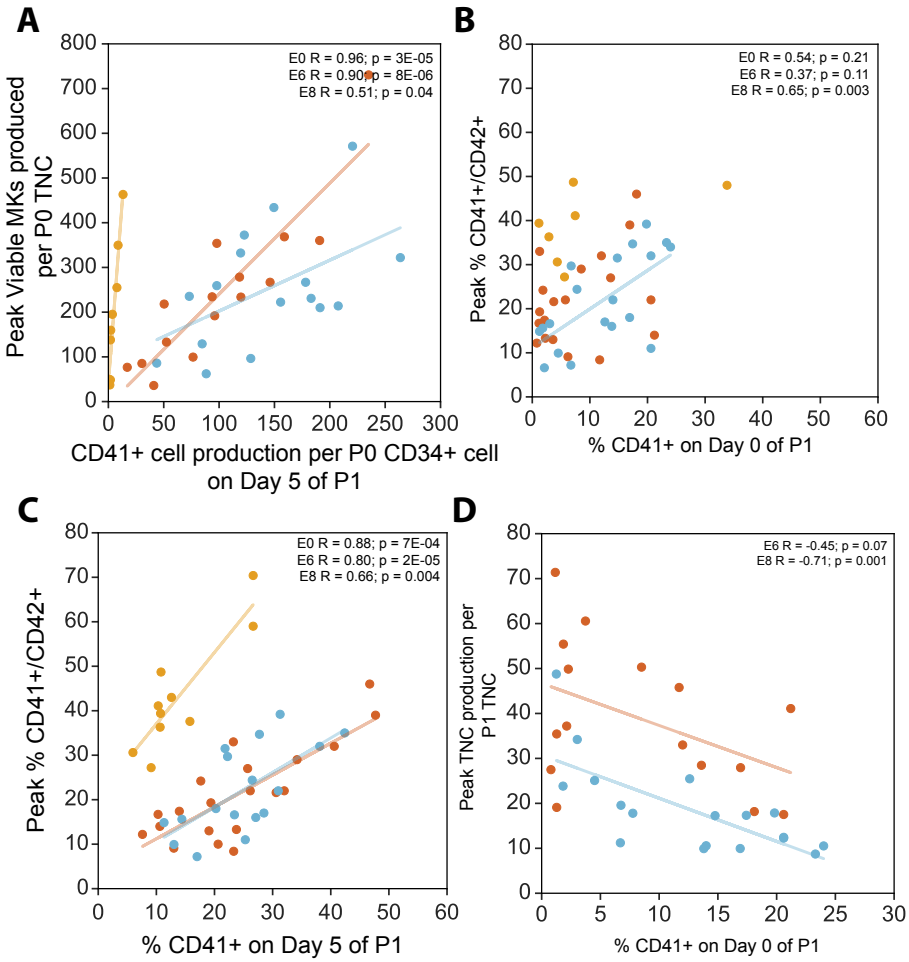

Figure 4.28 | **Correlation analysis between culture response variables.** A. Scatterplot shows correlations between peak MK production and $CD41^+$ cell production on day 5 in P1 culture for E0, E6 VPA+, E8 VPA+ culture conditions. B. Correlation between peak %CD41a+CD42b+ and $\%CD41^+$ on day 0 of P1 culture. C. Correlation between peak %CD41a+CD42b+ and $\%CD41^+$ on day 5 of P1 culture. D. Correlation between peak TNC production per P1 input cell and $\%CD41^+$ on day 0 of P1 culture.

Previous studies show that HDAC inhibitors (5-Azacitidine, trichostatin A, suberoylanilide hydroxamic acid, VPA, and combinations thereof) broadly affect global gene expression of pre-expanded $CD34^+$ cells, especially affecting genes related to HSPC maintenance [107]. Since our study established that early stages of culture involving HSPC maintenance

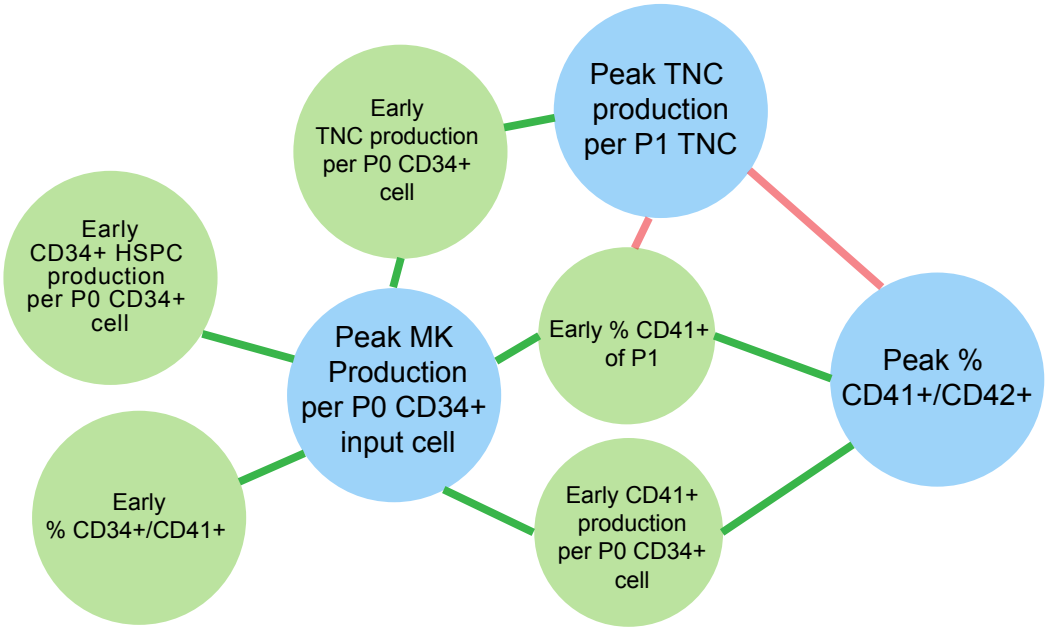

Figure 4.29 | **Correlation network between culture response variables shows influential factors of MK culture.** Correlation network shows potentially influential factors of MK culture between explanatory variables (green nodes) and response variables (blue nodes). Lines indicate that the significance of the Pearson correlation is $p < 0.05$ and the color of the line indicates whether the correlation is positive (green) or negative (red).

tend to be affected rather than later stages that involve PLP potency, we hypothesized that VPA primarily affected processes such as cell-cycle and recognition of senescence. In the context of MK differentiation, p21Cip/Waf1 is considered critical for establishing senescence, whereas p16INK4 is involved more in the maintenance of senescence [120]. We found that increasing pre-expansion time and thus initial HSPC expansion concomitantly caused upregulation of p21Cip/Waf1 and p16INK4 protein levels, while VPA treatment with pre-expansion decreased the extent of upregulation. Thus, VPA treatment during a critical early window has enduring effects on p16INK4 and p21Cip/Waf1 expression in resulting progeny, which facilitates higher MK and PLP yields. p16INK4 belongs to a family of proteins that

function as antagonists of cyclin D-Cdk4/6, blocking phosphorylation of Rb family members and subsequent entry into S phase, while p21Cip/Waf1 restrains entry into S phase by inhibiting cyclin E-Cdk2 [121]. Studies in human diploid fibroblasts show that p21Cip/Waf1 accumulates progressively in aging cells and decreases when senescence is achieved [121, 122]. The effect of VPA towards reducing levels of p21Cip/Waf1 and p16INK4 has been corroborated in other contexts; In the context of cellular reprogramming, adding VPA increases the reprogramming efficiency of iPSC cells by attenuating the effect of senescence pathways [123, 124]. We found that p21Cip/Waf1 and p16INK4 were upregulated during P1 culture after substantial pre-expansion, which suggests that HSPC pre-expansion in an ex vivo environment, especially in the absence of VPA, may lead to early induction of senescence due to extrinsic factors and imperfect recapitulation of the microenvironment and HSPC niche. Our study suggests that targeting p21Cip/Waf1 and p16INK4 using small molecules in pre-expansion and secondary culture protocols may substantially improve yields of MKs and/or PLPs.

Understanding donor variability and skewing towards the MK lineage of CB-derived cells is important for developing more efficient MK differentiation strategies and could also inform other studies that assess graft potency and subsequent in vivo platelet production after transplantation. A number of studies have analyzed maternal and neonatal factors associated with differences in CB $CD34^+$ cells [3, 99] but did not analyze subsequent effects on differentiation especially towards the MK lineage. Our results show that donor samples can be divided into high- and low-MK-producers, and that these divisions appear to be cell-intrinsic since unexpanded CB cells and pre-expanded CB cells display similar donor clustering of heterogeneity in MK yields. Further, we showed that high %CD41a+ and %$CD34^+$ CD41a+ early in culture are predictive of high peak MK purity and production.

This provides a platform for future studies into donor heterogeneity, but it is currently unclear what differences may be driving high MK production in the high-production donors. Single-cell sequencing studies have found a cluster of unipotent cells, identified by CD41 within $CD34^+$ CD38+ IL-3RdimCD45RA myeloid progenitors in adult mobilized peripheral blood, that exhibit robust differentiation into MKs while lacking potential towards other lineages [125]. Our study suggests that this population might be overrepresented in certain CB donor samples, thus giving these samples higher potential to produce MKs and PLPs.

Our protocols provide the basis for additional experimentation with each step to further augment MK and PLP yields. For example, building on the protocol by combinatorically adding other small molecules such as UM171 or SR-1 may further promote HSPC proliferation and differentiation in culture [126, 127]. Downstream PLP collection can also be improved. VPA pre-expansion can be used in conjunction with perfusion bioreactors that recapitulate shear forces in endogenous blood flow, thus increasing the number of PLPs per input MK [128, 129, 130] Further studies are needed to ascertain whether higher MK and PLP production can be achieved in large scale with other forms of shear-inducing technologies. Nevertheless, these results demonstrate the feasibility of a scalable process to generate PLPs from HSPCs using a modified pre-expansion protocol.

4.10. Methods

4.10.1. Cell culture

Fresh human CB collections were obtained from the Placental Blood Program of the New York Blood Center (New York, NY) according to guidelines established by the University of Illinois at Chicago Institutional Review Board. Low density CB cells were isolated using Ficoll-Paque (1.077 g/ml) (Amersham Biosciences, Uppsala, Sweden). $CD34^+$ cells were

immunomagenetically enriched using the MACS CD34 progenitor kit (Miltenyi Biotech, Auburn, CA) as described previously1. Purified $CD34^+$ cells (90% $CD34^+$) were seeded at 4 10ˆ4cells/ml in tissue culture-treated (TC) well plates in 6 well plates in 2.5 ml of Iscoves modified Dulbecco's medium (IMDM) (Biochrom) in serum-free medium (Sigma, St. Louis, MO) and supplemented with 100 ng/ml stem cell factor (SCF), 100 ng/ml FLT-3 ligand (FL), 100 ng/ml thrombopoietin, and 50 ng/ml interleukin 3 (IL-3). All cytokines were purchased from Peprotech (Rocky Hill, NJ). Cells were treated with 1 mM valproic acid and added media and cytokine supplements as above except IL-3 at 16 hours and incubated for another 6 or 8 days at 37C in a fully humidified atmosphere of 5% CO2, then transferred to secondary culture. In secondary culture, unselected cells were suspended in cytokine cocktails as described2. Briefly, cells were resuspended in 78% IMDM (Gibco, Carlsbad, CA) + 20% BIT 9500 Serum Substitute (STEMCELL, Vancouver, BC, Canada) + 1% Glutamax (Gibco) + 1 μg/mL low-density lipoproteins (Calbiochem) + 100 U/mL (Pen/Strep), supplemented with 100 ng/mL thrombopoietin (Tpo), 100 ng/mL stem cell factor (SCF), 2.5 ng/mL interleukin (IL)-3 (R&D Systems, Minneapolis, MN), 10 ng/mL IL-6, and 10 ng/mL IL-11 and seeded in TC-treated T-flasks at 50,000 cells/mL. Cells were cultured in a fully humidified chamber at 37 C, 5% CO2, and 5% O2 for 5 days. On day 5 of secondary culture, cells were pelleted and resuspended in fresh IMDM + 20% BIT supplemented with 100 ng/mL Tpo, 100 ng/mL SCF, 10 ng/mL IL-3, 10 ng/mL IL-9, and 10 ng/mL IL-11. Cells were cultured at 20% O2 thereafter. On day 7 cells were pelleted and resuspended in fresh IMDM+20% BIT supplemented with 100 ng/mL Tpo, 100 ng/mL SCF, and 6.25 mM nicotinamide (Sigma) and seeded at 200,000/mL on TC-treated dishes.

4.10.2. Cell Counting

Cells were counted using cetrimide and the Multisizer 3 (BD). The absolute number of MKs was determined as the product of the live cell count and the percentage of cells that were CD41a+. The absolute number of mature MKs was calculated as the product of the live cell count and the percentage of cells that were CD41a+CD42b+. Cell counts were normalized to input $CD34^+$ cell population (Day 0) to determine the fold increase of specific cell populations. MK production was determined as the number of CD41a+CD42b+ cells normalized to the number of starting $CD34^+$ cells.

4.10.3. Polyploidization analysis

MK ploidy was analyzed on Day 16 of cumulative culture. Cells were labeled with FITC-conjugated anti-CD41, incubated for 20 minutes at 4C, fixed with 0.5% paraformaldehyde (Fisher Scientific, Waltham, MA) for 15 minutes and then washed with PBS. Cells were then permeabilized with 75% methanol (Sigma Aldrich) for 1.5 hours at 4C and washed with PBS+2% BSA(Fisher Scientific). Finally, cells were treated with 1 μg/ml RNAse (Sigma) and propidium iodide/RNase solution (Becton Dickinson), and then incubated for 30 minutes in the dark at room temperature before flow cytometry analysis.

4.10.4. Flow cytometry analysis for MK differentiation

20,000 cells were collected and centrifuged at 340xg for 3 minutes, then resuspended in phosphate buffered saline (PBS) + 2% BSA (Fisher Scientific). Cells were stained with fluorescein isothiocyanate (FITC)-conjugated mouse monoclonal (clone 581) anti-human-CD41 antibody, phycoerythrin (PE)-conjugated mouse monoclonal (clone HIP8) anti-human-CD34 antibody, and APC-conjugated mouse anti-human-CD42b antibody (clone HIP1) at 4C, in

the dark for 20 minutes. All antibodies for flow cytometry stains were obtained from BD Biosciences (Franklin Lakes, NJ). Cells were then washed twice and stained with 5 μg/ml DAPI for 10 minutes before acquisition.

4.10.5. Aggregation assay

A single-channel reactor was coated with 60 ng/ml fibrinogen for 1 hour at 37C, 5% CO2. PLPs in HT buffer were labeled with 1 μM Calcein AM for 15 min at 37C and then microinjected upstream of the viewing area at a concentration of 40x106 PLPs/ml. CaCl2 (to 2mM) was added to PLPs, expired platelets, and HT buffer running solution before injecting with pump controlled syringes (NE-300, New Era Pump Systems Inc.) at a rate of 1.5 μl/second with or without addition of ADP (to 25 μM).

4.10.6. Aggregation assay open channel reactor fabrication

A single channel polydimethylsiloxane reactor was fabricated similarly to methods previously described with only the channel present and the slits removed.

4.10.7. Microfluidic shear analysis

CD61+ selected MKs were seeded on a polydimethylsiloxane USRB reactor with 5μm slits as previously described. Briefly, MKs at a density of 50,000/mL were stained for 15 min with 1 μM Calcein AM at 37C. 25,000 stained MKs were microinjected into the tubing upstream of slits and subsequently exposed to a center channel flow rate of 1.5 μL/min and a combined outer channel flow rate of 0 μL/min. Proplatelet and PLP formation were observed and recorded for 30 minutes inside the incubator using a Lumascope v500 microscope (Etaluma).

4.10.8. qRT-PCR

Total RNA was extracted using Trizol (Invitrogen, Carlsbad, CA) from CB $CD34^+$ cells (Day 0) or the culture product of $CD34^+$ cells pre-expanded with or without VPA and obtained on cumulative day 20 of culture in the presence of cytokines. Relative transcript levels from pre-culture (Day 0) and Day 20 (E6 and E8) culture were determined by real-time quantitative PCR. Transcription into cDNA was performed using oligo (dT)18 primers and Superscript II reverse transcriptase (Invitrogen) according to the manufacturers instructions. (C1000 BIO RAD). Briefly, 2.5 μg of total RNA from all samples were used to reverse transcribe into cDNA. All PCR reactions used SYBR green PCR Master Mix (Applied Bio systems, Foster City, CA, USA) to a final volume of 20 μl as described previously4. PCR cycling conditions were standard except for annealing/elongation temperature, which ranged between 57C and 62C and was chosen based on preliminary primer optimization experiments. To quantitate the expression level, each cDNA sample was analyzed in triplicate in the ABI 7500 Fast System (Applied Bio systems) and a negative control (lacking cDNA template) was included in each assay. GAPDH was used an internal calibrator (control gene), the standard curve method was used for relative mRNA quantitation and final values are presented relative to the pre-culture sample (D0). The primer sequences used in the real-time PCR assays are as follows: GAPDH forward primer:

5-TGCACCACCAACTGCTTAGC-3,

reverse primer: 5-TCTTCTGGGTGGCAGTGATG-3,

p21 forward primer: 5-GGTCTGACCCCAAACACCTTC-3,

reverse primer: 5-AACGGGAACCAGGACACACATG-3,

p16 Forward primer: 5-CATAGATGCCGCGGAAGGT-3,

reverse primer 5-CTAAGTTTCCCGAGGTTTCTCAGA-3.

4.10.9. Intracellular flow cytometry

Cultures were sampled on cumulative day 11 of culture. 105 cells were washed, then incubated with PE-conjugated anti-CD34, APC-conjugated anti-CD41 for 20 minutes, then washed twice. Samples were fixed with 0.5% formaldehyde and incubated at RT for 15 minutes, washed twice, then incubated with 70% cold methanol for 30 minutes at 4C then washed twice. To block non-specific protein interactions, cells were incubated with 1X PBS containing 10% normal goat serum with 0.3M glycine along with anti-p21Cip/Waf1 (1/100; ab109520; Abcam) or anti-p16INK4 antibody (1/270; ab108349; Abcam) for 30 minutes at RT. Cells were washed twice then incubated with secondary antibody anti-rabbit IgG-Alexa488 (1/2000; Abcam) for 30 minutes at RT. Controls were created by staining fixed samples with secondary antibody only. Samples were then washed, resuspended, and acquired via flow cytometry.

4.10.10. Platelet-like particle (PLP) preparation and analysis

On Day 7 of secondary culture, MK cultures were placed on an orbital shaker (Scilogex, model SK-0180-E) set to 50rpm in an incubator at 5% CO2and 20% O2 at a density of 600,000 cells/ml in a 100 x 20 mm polystyrene tissue culture dish (Corning). At each timepoint, a 500 μl sample was transferred to a conical tube. 100 μl of the sample was added to cetrimide and counted with a Multisizer 3 (BD) to determine live cell concentration (nuclei 3 um). 400 μl of sample was labeled with FITC-conjugated anti-CD41, APC-conjugated anti-CD42, and DAPI (5 μg/ml) to assess the ratio of CD41a+CD42b+ PLPs to live cells.

PLPs collected from the orbital shaker were tested for functional activity based on the externalization of P-selectin and PAC1 binding after activation with thrombin using flow cytometry. Briefly, PLP samples were centrifuged for 300xg for 10 minutes at RT to pellet

large MKs. Prostaglandin E1 (Cayman Chemicals) was added to a final concentration of 140 uM to the upper phase then transferred into a Falcon tube. The supernatant was centrifuged for 2200 xg for 20 minutes at RT to pellet platelets, then rested for 20 minutes, before adding CaCl2 to 2 mM. PLPs were then stained, fixed, and analyzed by flow cytometry for P-selectin externalization and PAC-1 binding. Anti-Y1-PE (Becton Dickinson) and iso-FITC were used as isotype controls. PLPs were identified as $CD41^+$ CD42+events that fell in the side scatter versus forward scatter gate corresponding to fresh platelets isolated from whole blood or recently expired platelets from apheresis units. Expression of activation markers was compared to an unactivated, stained sample.

4.10.11. PLP degradation analysis

PLP samples in media containing at least 4x106 CD41a+CD42b+ PLPs and MKs were isolated on day 14 of overall culture. Samples were spun down at 300xg to pellet MKs. The upper phase was transferred to a 6-well TC plate (Corning) then adjusted to a concentration of 2x106 CD41a+CD42b+ PLPs/3 ml of media, and placed on a 50-rpm orbital shaker for 24 hours at 37C, 5% CO2. Flow cytometry was performed by staining PLP samples with anti-CD41a+-FITC, anti-CD42b+-APC, and 5 μg/ml DAPI, then CountBright beads were added (Thermofisher) to determine the ratio of CD41a+CD42b+ PLPs to beads, and confirm that all MKs were removed from the shaking culture.

4.10.12. Confocal microscopy

MKs and PLPs were seeded on glass chamber slides (Nunc) that were coated with 1% BSA or 60 ng/ml fibrinogen (Innovative Research). To induce activation, PLPs were rested for 20 minutes then mixed with thrombin for a final concentration of 0.5 U/mL. To prepare

slides for immunocytochemistry, the cell solution was removed from each well and cells were fixed with 3.7% formaldehyde solution for 10 minutes, washed, then permeabilized with 0.3% Triton X. Image iT-FX signal enhancer (Thermofisher Scientific) was added to each well for 30 minutes, washed twice, then cells were stained with mouse anti- tubulin primary antibody overnight. After washing twice, cells were incubated with 1% Alexa 488-goat-anti-mouse secondary antibody (Jackson Immunoresearch) and 2% normal goat serum for 1 hour, washed twice, then incubated with TRITC phalloidin, washed twice, then incubated with 1 ug/ml DAPI. Fluoromount-G (Thermofisher Scientific) was applied and the slide was allowed to cure for 24 hours. Slides were imaged using a 40X oil objective on a SP5 II Laser Scanning Confocal Microscope (Leica).

4.10.13. Immunofluoresence staining and microscopy

For analysis of platelet activation, glass chamber slides (Nunc) were coated with BSA or fibrinogen. Cell and PLP suspensions added to each well in the presence or absence of 3U/mL thrombin for 1h at 37 C, 5% CO2. Slides were fixed with 3.7% paraformaldehyde and permeabilized with 0.3% Triton X-100 before staining with 5μg/mL mouse anti--tubulin primary antibody, and 140μg/mL of FITC-conjugated goat anti-mouse secondary antibody. After removing the secondary antibody, cells were incubated with TRITC-phalloidin, washed, and stained with DAPI nuclear stain (Invitrogen) to confirm the presence or absence of DNA. Proplatelet and PLP slides were imaged using a 40X objective, respectively, on an Spinning Disk Confocal Microscope (Leica).

4.10.14. k-means clustering

A combined data matrix was generated for (3) measured variables at each day (6 time-points) for both E6 and E8 cultures (2 conditions): %CD41a+CD42b+, Viable MK production per input P1 TNC cell, and Viable TNC production per input P1 TNC. Each measured variable was z-scored within each category across all cultures and time-points to weight variables equally. Measured variables were concatenated into a 12x36 (6 time-points * 2 conditions * 3 variables = 36 columns) matrix, where the rows represent each culture, and columns represent the measured variables on each day for E6 and E8 cultures. K-means clustering was performed to determine whether the culture belonged to a high-MK group or low-MK group using k=110 centroids, n˙init=1000 (number of centroid initializations), max˙iter=3000 (number of iterations in a single run) and the scikit learn package (python version 2.7). The optimal number of clusters (k=2) was selected based on the elbow rule.

4.10.15. Statistics

Data representing cell viability; and CD34, CD41a, and CD42b expression are presented as mean±standard error. Two-tailed paired t-tests were performed to determine whether treatments had a statistically significant effect on paired donor samples. and a value of $p < 0.05$ was considered statistically significant. Pearson correlation coefficient (R) was used to measure the association between two variables. We used R (version 3.5) and lme4 to perform a linear mixed effects analysis modeling the relationship between %CD41a+CD42b+, Viable MK production per input P1 TNC cell, and Viable TNC production per input P1 TNC and whether the culture belonged to a high-MK group or low-MK group. As fixed effects, we entered one of the three measured variables, such as Viable MK production per input P1 TNC cell, and day into the model. As random effects, we had intercepts for each culture

unit for the effect on the measured variable. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. P-values were obtained by likelihood ratio tests of the full model with the effect (high-MK or low-MK assignment taken into account) against the model without the effect (high-MK or low-MK assignment removed).

CHAPTER 5

Ex vivo cultures for megakaryocyte differentiation described by time-course models

Work presented in this chapter is adapted from the following paper in preparation:

- Wu, J.J., Abbott, D.A., Mahmud, N., Miller, W.M., and Bagheri, N. Gaussian Mixture Models and Machine Learning Predict Megakaryocytic Growth and Differentiation Potential Ex Vivo.

5.1. Abstract

The ability to analyze single cells via flow cytometry has impacted a wide range of biological and medical applications, such as enabling the classification of immune and hematopoietic cells during cellular differentiation. Manual analysis of temporal trends is time-consuming and subjective for flow cytometry datasets. Thus we have developed a novel computational algorithm to quantify and predict temporal trends of developing cell subpopulations in flow cytometry data, called Topographical Extraction of Gaussian Models (TEGM). We propose a way to segment time-series flow cytometry data and extract additional features from samples. We applied our method to a novel gold standard dataset comprised of 720 features from 80 perturbations of 54 samples, with 7 time-point measurements, capturing ex vivo MK differentiation and maturation of hematopoietic cells from donors with varying potential to generate $CD41^{+}CD42^{+}$ cells. We demonstrate the ability for TEGM to identify latent donor heterogeneity using the training set. TEGM predicted peak $CD41^{+}CD42^{+}$

maturation purities of megakaryocyte cultures derived from diverse individual donors that were highly concordant with actual peak culture outcomes (p=7.4e-09, R = 0.87). The resulting model captured continuous progression of cell development processes represented by cell surface markers (CD34, CD41, and CD42), and predicts MK differentiation and maturation potential in culture of a given CB unit in a separate validation set. We applied the TEGM computational framework to generate additional quantitative metrics capturing subtle aspects of megakaryocyte differentiation.

5.2. Introduction to flow cytometry analysis

Flow cytometers provide high-dimensional quantitative measurements of light scatter and fluorescence emission properties of thousands of single cells in each sample [131, 132]. Flow cytometry is routinely used in both research and clinical settings to study abnormal cell structure and function, and diagnose and monitor human disease [133, 134]. A key step in the traditional analysis of flow cytometry data is the grouping of individual cell events into discrete populations on the basis of similarities in light scattering and fluorescence [135]. This analysis is usually accomplished by sequential manual partitioning or "gating" of cell events into populations through visual inspection of plots in one or two dimensions at a time. Manual flow gating and analysis is subjective, time-consuming, and is ill-suited for finding relationships in high-dimensional data [132, 136].

Several researchers have developed automatic clustering techniques to assist in flow cytometry analysis, based on K-means, T-SNE models, and statistical mixture models [134, 137, 138]. These applications have been primarily focused on using statistical clustering to identify rare subpopulations within a single patient sample. For example, viSNE and SPADE project high dimensional measurements into a two dimensional map, allowing

several cell subpopulations to be compared and contrasted without gating [138, 139]. Pseudotime algorithms, such as Wanderlust and Monocle, identify the underlying developmental trajectory from single flow cytometry samples [140, 141]. There is currently no algorithm that incorporates multiple time series flow cytometry measurements. Additionally, very few algorithms couple gating frameworks to machine-learning models that learn patterns predictive of response variables. We have developed a framework that incorporates time-series flow cytometry datasets into Gaussian mixture models (GMMs) for automatic gating and quantification. We demonstrate the application of this framework by predicting endpoint surface markers and $CD41^{+}CD42^{+}$ cell production using early surface markers and metadata in the context of *ex vivo* MK differentiation. We demonstrate that the resulting model automatically gates relevant populations represented by cellular surface markers, CD34, CD41, and CD42, and accurately predicts megakaryocytic culture potential of a given donor.

5.3. TEmporal Gaussian Models (TEGM) is a framework for time-series flow cytometry analysis and identifies populations using unbiased, automated segmentation of each time-point data using Gaussian mixtures

First, we describe how our hierarchical feature extraction approach extracts quantitative information from time-series flow cytometry data. This process is divided into two main parts: segmentation (Fig. 5.3A, 5.3B) and feature extraction (Fig. 5.3C, 5.3D).

Before running the segmentation algorithm, samples encompassing time-series flow cytometry measurements of 54 cultures are collected as described previously in Chapter 4 (see subsection: "Flow cytometry analysis for MK differentiation"). Time-points were taken on several days (Days 0, 3, 5, 7, 9, 11) for 54 cultures for several surface markers (CD34, CD41, CD42) and a viability marker (4',6-diamidino-2-phenylindole; DAPI), and adjusted

for compensation. Compensation normalization was applied to several variables (CD34, CD41, CD42) based on multiple voltage settings as previously described [142].

5.3.1. Development of GMM at each time-point organizes data into representative populations

To segment the point-clouds into discrete cell populations (Fig. 5.3B), we first identify the number of cell populations and also the boundary that encapsulates each cell population. All distributions of cells within a given flow cytometry sample are assumed to be effectively modeled by a mixture of Gaussian multivariate probability distributions. A summary of the algorithm is described in Algorithm 5.1.

Formally, given an input which is a M by d matrix representing the aggregate point-cloud for a single flow cytometry dataset at a given time-point, where M is the number of cells, and d is the number of measurements per single cell (Algorithm 5.1), the output of our algorithm is a C by F matrix where C is the number of conditions and F is the number of extracted features. First, the algorithm assigns a cluster to each cell by finding parameters such that the goodness of fit of the estimated probability distribution (described by parameters $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\alpha}, K$) against the observation dataset X is maximized ($P(X|\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\alpha}, K)$). The previously described EM algorithm is used to estimate the parameters for the mixture of Gaussian distributions [143]. In a mixture of Gaussian distributions where there are K number of mixture components (indexed by k), $\mathcal{N}$ is the Gaussian probability density function of a random variable and $\boldsymbol{\alpha} = \alpha_1, ..., \alpha_K$ are mixing probabilities where $\sum_{k=1}^{K} \alpha_k = 1$. Thus the posterior probability of the generative model representing the probability of the data given the fitted model parameters is:

Algorithm 5.1: TEGM Segmentation and Feature Extraction

Input: A **M** by **d** matrix representing a time-series flow cytometry dataset where **M** number of cells, **d** single cell measurements, for each **C** conditions, and **T** time-points,

Output: A **C** by **F** matrix representing the features extracted from input matrix.

```
1: For c = 1 to C {
2:    For t = 1 to T {
          \\ Estimate K number of components of GMM for each M by d matrix by:
3:       For k = 1 to 20:
4:          Calculate log P(X | μ, σ, α, K=k) of GMM using EM algorithm
5:          Calculate model fit using Bayesian Information Criterion (BIC)
6:       Estimate best K using elbow rule of BIC.
7:       Fit GMM by maximizing log P(X| μ, σ, α, K) using EM algorithm  \\ each point now has a posterior
probability of belonging to one of the components in k
8:       Assign component label of each point using posterior probability cutoff. A point x is assigned to a
component if P(x| μ, σ, α, K) > 0.95 \\ the model now has several components with a point cloud belonging to
each component
9:       Extract features (μ, σ, α, Σ, %within) for each P  \\ μ(mean), σ(sd), α (mixture weight), Σ (covariance),
P is a population
10:         Derive eigenfeatures for each P using Σ  \\ see eigenfeatures table
11:   }
12: }
13: Assign population labels by hierarchically clustering each P for each dataset (P_L) and setting max d = 0.3
14: For c = 1 to C {
15:    Derive betweenPopulation features
16:    Fill each row of a C by F matrix with features from each P_L with each column as a feature,
eigenfeature, or betweenPopulation feature.
17: }

end
```

Figure 5.1 | **Summary of TEGM segmentation and feature extraction algorithm.** Feature extraction algorithm is summarized for each condition and time-point. Features, eigenfeatures, and betweenPopulation features are extracted from the dataset and stored for learning.

$$P(X|\mu, \sigma, \alpha, K) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(X|\mu_k, \sigma_k^2)$$

The result of fitting the GMM is that each datapoint now has a posterior probability associated with belonging to any of the mixture components with an estimated μ_k and σ_k.

Means of mixture components across multiple timepoints and conditions are aggregated so that features extracted from each "type" of population can be compared. To assign a population "type", populations across multiple timepoints and conditions are grouped by location of the mean centroid. This classification allows us to compare features from

populations of the same "type", for example, all populations that have a mean centroid in the high CD41/high CD42/low CD34/high FSC/mid SSC/low DAPI region will be classified as Population 1, which corresponds to the mature MK populations, allowing all features of mature MK populations across from different conditions and time-points to be compared to one another. Formally, the points belonging to each mixture component are denoted as $\mathcal{P}_L$ which is a subset of points $\mathcal{P}$ indexed by a label L. We performed hierarchical clustering on the mean centroid of each cluster with a cut-off distance of H_d=0.3 to determine the similarity of each cluster in Euclidean space and subsequently population labels for each cluster. Points are only assigned to a population if the posterior probability of belonging to the cluster is greater than 0.99. If the posterior probability is greater than 0.99 for multiple mixture components, the point is assigned to the component that scored the highest posterior probability.

5.3.2. Eigenfeatures define quantitative features of cellular differentiation

Eigenfeatures are extracted from each population based on a covariance matrix σ of the given points (Fig. 5.3C).

$$\Sigma_i = \begin{bmatrix} cov(\overline{x}_1, \overline{x}_1) & cov(\overline{x}_1, \overline{x}_2) & cov(\overline{x}_1, \overline{x}_3) & \dots & cov(\overline{x}_1, \overline{x}_d) \\ cov(\overline{x}_2, \overline{x}_1) & cov(\overline{x}_2, \overline{x}_2) & cov(\overline{x}_2, \overline{x}_3) & \dots & cov(\overline{x}_2, \overline{x}_d) \\ \dots & \dots & \dots & \dots & \dots \\ cov(\overline{x}_d, \overline{x}_1) & cov(\overline{x}_d, \overline{x}_2) & cov(\overline{x}_d, \overline{x}_3) & \dots & cov(\overline{x}_d, \overline{x}_d) \end{bmatrix}$$

The covariance matrix Σ for any Gaussian is symmetric and positive definite, and can be decomposed using eigendecomposition (eig, python):

$$\Sigma^{-1} = (Q\Lambda Q)^{-1}$$

where Q is an orthogonal matrix whose columns are eigenvectors, and Λ is a diagonal matrix with diagonal entries denoting the eigenvalues. These values are used for subsequent calculations for machine-learning as described below.

Additionally, other features that recapitulate manual gating and betweenPopulation features are extracted. All extracted features are organized in a **C** by **F** matrix where **C** is the total number of conditions and **F** is the total number of features.

5.3.3. Machine learning identifies relative importance of extracted features

gradient–boosted regression tree ensembles combine weak classifiers to produce strong predictions for continuous variables (Fig. 5.3D). To determine the variable importance of each feature, we minimize the squared-error loss function:

$$L(r, f(Q)) = \sum_{j=1}^{F} (r_j - f(W_j))^2$$

where r is the response variable (peak $CD41^+CD42^+$), f is a regression function of trees with features $W = (W_1, W_2, ...W_F)$. The gradient boost algorithm aims to iteratively minimize the expected square error loss, with respect to f, on weighted versions of the training data. gradient–boosted regression trees models were fitted using the GBM R package (https://CRAN.R-project.org/package=gbm) with a squared error loss function.

We performed a grid search on parameters that optimized the loss function on a training dataset consisting of 75% of the donors: A total of 2,000-10,000 trees were fitted with an interaction depth from 2 to 5, a shrinkage parameter from 0.001 to 0.01, and a bag fraction from 0.5 to 0.9. The permutational variable importance, which is proportional to the amount of variance reduced in the response variable when a factor is selected as a node in a gradient–boosted Tree, was calculated. We reserved 25% of the data as a test set to

assess model accuracy. Model accuracy was assessed by the Qc^2 score and p-value comparing actual versus predicted peak percent $CD41^+CD42^+$. This is calculated by:

$$Qc^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y} - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where y_i is the actual value of peak percent $CD41^+CD42^+$ in the test set, $\hat{y}$ is the predicted value of the statistic in the test set, and $\bar{y}$ is the mean of the actual values in the test set.

5.4. Results

We describe a model framework to segment a time-series flow cytometry dataset, extract novel features that incorporate temporal trends, and predict selected response variables. We tested model predictions on a novel dataset comprised of 720 measurements from 54 perturbed samples, with 7 time-point measurements, capturing *ex vivo* MK differentiation and maturation of hematopoietic cells from donors with varying potential to generate $CD41^+CD42^+$ cells. The number of events for each sample ranges from 5,000 to 12,000.

5.4.1. TEGM gating identifies primary cell populations in well-defined *ex vivo* data-set

To determine if our Gaussian Mixture Model approach captured well-defined populations, we first identified the mean and covariance matrices of fluorescent bead populations. TEGM accurately separated bead populations from mixed cell and debris solutions, showing that well-defined clusters may be identified by the approach (Fig. 5.4A). Next, we applied the TEGM framework to identify megakaryocyte cells in a mixed population and generated a quad gating framework to recapitulate manual gating (Fig 5.4B). Even when encountering continuous and discrete flow cytometry profiles, TEGM identified 2 populations via BIC

Notation	Explanation
M	Number of cells
d	Number of measurements per cell
C	Number of conditions
F	Number of extracted features
X	Matrix containing data that is size M by d
$\mu = \{ \mu_1, \ldots, \mu_k \}$	Vector containing expected values of mixture of Gaussians
$\sigma = \{ \sigma_1, \ldots, \sigma_k \}$	Vector containing standard deviation of mixture of Gaussians
$\alpha = \{ \alpha_1, \ldots, \alpha_k \}$	Vector containing weights of mixture of Gaussians
K	Number of components/Gaussians in a single FCS file
$\mathcal{N}(X \mid \mu_k, \sigma_k)$	Gaussian probability density function of random variable X with expected value μ_k and standard deviation σ_k
$\mathcal{P}_L$, L	The set of points belonging to a population indexed by L
$\mathcal{H}_d$	Cophentic distance threshold used for determining clusters; distance between any two observations in a cluster can be no larger than $\mathcal{H}_d$
Σ	Covariance matrix
$W = \{ w_1, \ldots, w_F \}$	Matrix containing columns of features w_F
r	Response variable for gradient boosted regression
Qt^2	Correlation between predicted and actual values of the response variable in the test set
y_i	Actual value of the response variable in the test set
$\hat{y}$	Predicted value of the response variable in the test set
$\bar{y}$	Mean value of the response variable in the test set

Figure 5.2 | **Notation table of TEGM.** Table describing list of notations used to describe TEGM steps.

model comparison. The Bayesian information criteron (BIC) was used to identify the optimal number of parameters for each multivariate GMM fit. We selected the smallest number of

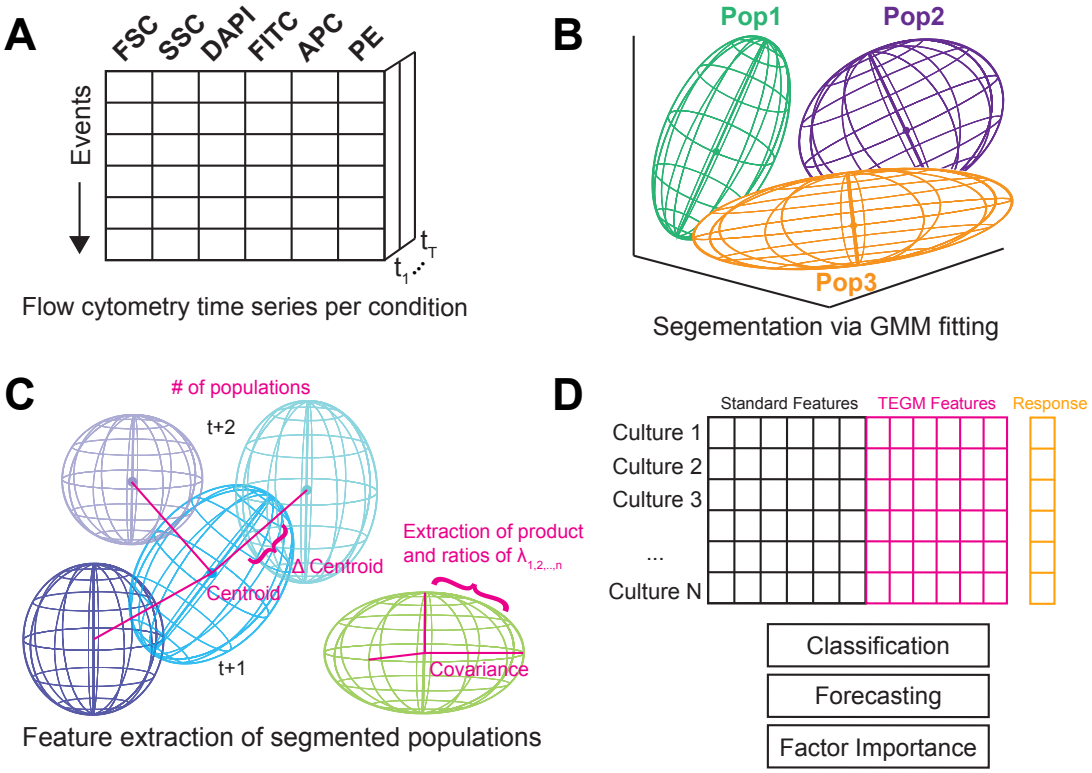

Figure 5.3 | **Major steps implemented by TEGM package integrate automated gating and feature extraction of time-series flow cytometry data and machine learning prediction versus manual analysis.** A) Time-series flow cytometry data is pre-processed and compensated. B) Populations of flow cytometry data are segmented using Gaussian Mixture Modeling. C) Features are extracted from segmented populations. D) Features are placed in a machine-learning framework for prediction and evaluation of factor importance.

components for the GMM that minimized the BIC. To define the quad gating strategy, for each time-series, we identified populations that were CD34-, CD41-, and CD42- and fitted quad gates to the corner of the negative populations, then we propagated the gating boundaries to the entire time-series for each condition.

To determine if our approach was comparable to manual analysis, we tested whether TEGM and the proxy quad gating strategy was able to identify comparable percentages of key populations of cells (CD34+/-, CD41+/-, CD42+/-). We then calculated the percentage

of events in each TEGM-fitted quadrant and compared them to expert-defined manual quad gating (Fig. 5.4C). Gating of CD41+ cells is the most dissimilar between automated and manual forms of gating, and percentage of CD41+ in the automated case is often a slight underestimate of true CD41 percentage. This may be due to the continuous distribution of CD41% commitment in some cases. Notably, the estimation of the percentage of CD34+ and CD42+ cells is within 1% of manual gating estimates. We then used TEGM to define levels of CD34%, $CD41^{+}CD42^{+}$, and Viability for all conditions in the dataset (7 donors each for E5 and E8 conditions pictured in Fig. 5.4D-F) over time. TEGM identified several trajectories that reflect the donor heterogeneity of the overall samples in terms of loss of CD34+, and gain and loss of $CD41^{+}CD42^{+}$. TEGM also identified the lack of donor heterogeneity for Viability, which did not vary much between donors over time (within 15% CV). Thus, TEGM is an automated segmentation procedure based on GMMs and hierarchical clustering, which extracts subpopulation information within aggregated sets of timepoints, with one of the subpopulations being live DAPI-low/CD34-/CD41+/CD42+ megakaryocytes.

Next, features are constructed based on eigenfeatures for each flow cytometry time-course dataset using resulting GMM models and clustering. Features that can be obtained by manual gating, or apparent descriptors, are extracted and compared to gold standards to demonstrate accurate gating. Derived descriptors, such as eigenfeatures describing dispersion of all subpopulations, and rate of differentiation of subpopulations, are used to construct an explanatory matrix for a compendium of machine–learning algorithms. Overall, the final learning matrix for the gradient–boosted tree consists of 730 columns (features) and 65 rows (samples).

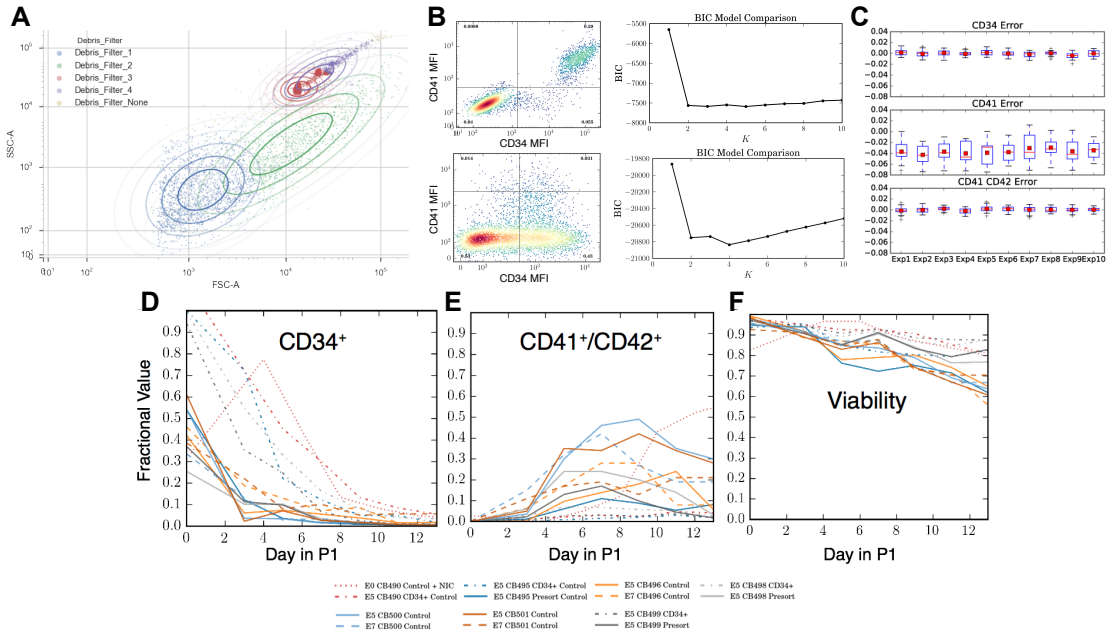

Figure 5.4 | **TEGM automated gating identifies bead and cell populations in well-defined cell populations.** A) TEGM clustering and covariance measurements of fluorescent beads projected onto 2D, demonstrating that TEGM identifies clusters of multiple populations. The cluster of beads, fluorescing in all channels is identified by cluster Debris_Filter_2 (red). Each color represents a different population defined by TEGM. Each concentric circle represents 1, 2, 3, 4 standard deviations from the mean of the population. B) Example of automated quad-gating behavior of discrete and continuous populations emulating expert-defined manual gating. GMMs are able to identify separate and continuous populations using Bayesian Information Criterion (BIC) model comparison. C) TEGM automated gating is comparable to expert-defined manual gating. Error = TEGM Marker positive divided by Expert Determined Marker positive (n=359). Automated feature extraction based on TEGM quad gating of D) CD34+ human stem and progenitor cell (HSPC) primitive lineage marker E) $CD41^{+}CD42b^{+}$ maturation markers and F) viability (DAPI) markers showing dynamic differences in differentiation kinetics for 14 culture conditions of 7 donor samples (n=7 for E5 and E7).

5.4.2. TEGM introduces a new method to use FC data for prediction of cell responses

We evaluated the effectiveness of the integrated pipeline in predicting culture outcome by training a gradient–boosted tree model (GBM) with features derived from automated gating

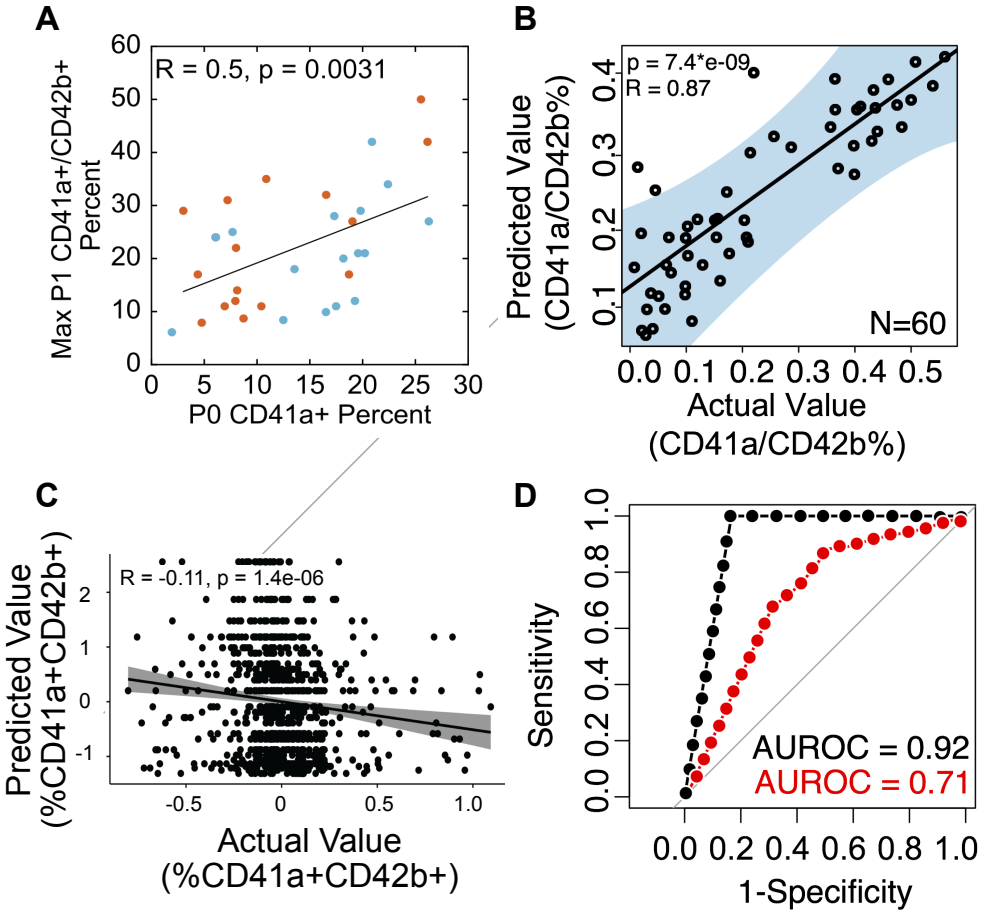

Figure 5.5 | **Gradient–boosted trees identify top TEGM features.** Tree model trained on TEGM–extracted features improves prediction as measured by correlation for 3 non-overlapping 20–test cultures. A) Simple correlation and linear regression model shows that final CD41+CD42+ percentage cannot be predicted by early CD41+ percent alone. B) GBM model for predicting maximum CD41+CD42+ of each donor shows that correlation between prediction and actual values is improved with TEGM approach (N=20x3 iterations of testing data). C) Permuted null model shows that TEGM–extracted features and performance are not due to random noise (N=20x500 iterations of permuted null models) D) AUROC curve for GBM trained to predict binary outcomes (greater or less than 20% CD41+CD42+) show that TEGM approach incorporating eigenfeatures (black) is more accurate than TEGM approach without eigenfeatures (red).

and feature extraction as described above. We extracted several features that were correlative of peak percent $CD41^+CD42^+$, that alone cannot be used to predict culture outcome with high specificity, but are more effective in combination (Fig. 5.5A and 5.5B). To determine whether TEGM is able to predict peak percent $CD41^+CD42^+$ commitment based on early factors, we evaluated the GBM model based on 3 independent selections of training and test sets to predict the response value (peak percent $CD41^+CD42^+$). The correlation between predicted and actual response was significantly positively correlated (p=7.4e-09, R = 0.87) compared to a shuffled null model which had a negatively correlated trend (Fig. 5.5B). We compared this to a null model with permuted rows to show that noise in the dataset does not significantly bias the correlation of the results (Fig. 5.5C; p=1.4e-06, R=-0.11). To determine if TEGM is able to leverage the features to predict binary outcomes (greater than or less than 20% peak $CD41^+CD42^+$), we compared two feature sets derived from TEGM, a feature set with and without eigenfeatures (Fig. 5.5D). We show that while TEGM extraction of manually derived features (without eigenfeatures) has high accuracy in predicting binary outcomes (AUROC = 0.71), TEGM with eigenfeatures significantly increases the performance of the algorithm (AUROC = 0.92).

5.4.3. TEGM identifies culture factors that correspond to and potentially govern cell responses

Overall, we identified several influential early culture factors that are predictive of peak $CD41^+CD42^+$ commitment. We calculated the Variable Importance of each selected feature to determine which extracted features may improve performance. As previously described, Variable Importance of a feature is proportional to the amount of variance reduced in the response variable when the feature is selected as a decision tree node. Early $CD41^+$ percentage

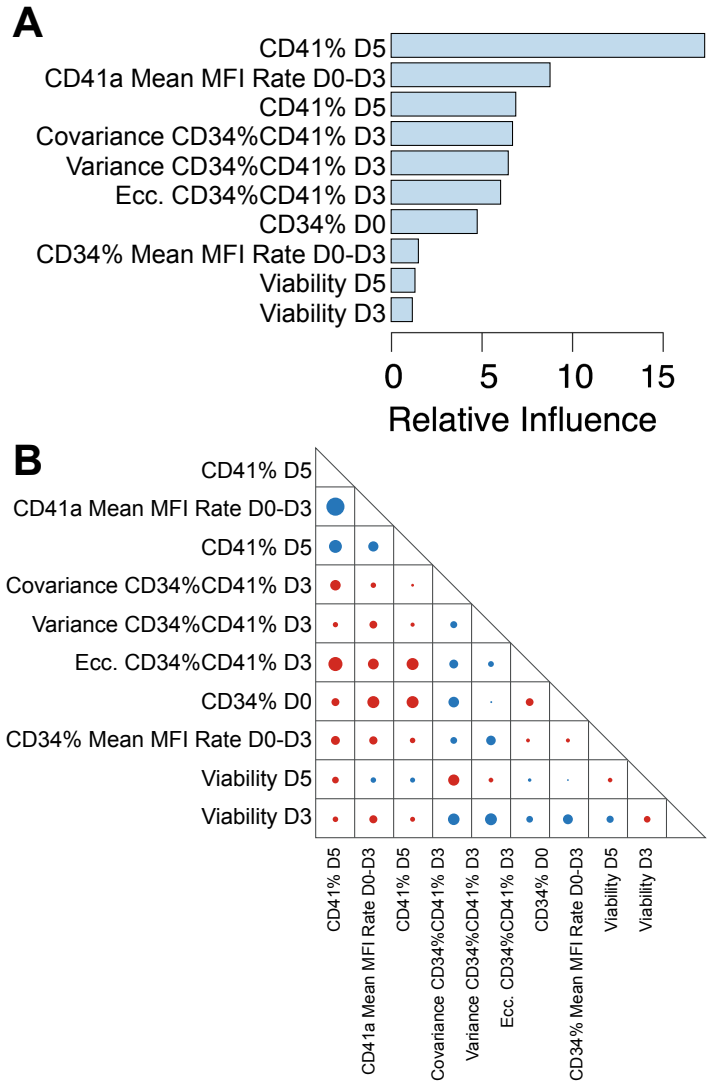

Figure 5.6 | **TEGM reveals influential culture factors using relative influence.** A) Influential culture factors calculated using the Variable Importance statistic. B) Dot plot shows the size and direction of variable interaction using partial dependence plots. Red indicates a negative interaction and blue indicates a positive interaction. The size of the circle reflects the strength of interaction effects.

on Day 5 is highly predictive of peak $CD41^{+}CD42^{+}$ commitment and is the most influential feature for performance as determined by Variable importance (Fig. 5.6A). We note that this is fairly late in culture, because peak $CD41^{+}CD42^{+}$ commitment often occurs on Day

9 to Day 11, so this feature may have limited potential to provide the user with opportunities for intervention. Interestingly, the rate of differentiation within the first three days is also highly influential. Viability and CD34 percentage are comparatively less influential for peak CD42+ percentage (Fig. 5.4A). Eigenfeatures such as Covariance, Variance, and Eccentricity of related populations (particularly the CD34+CD41+ population) also appear to be influential in the top 10 influential variables for prediction.

Combinatorial effects of various culture factors are known to perturb performance. Thus, many anticipated predictor variables tend to be involved in a variety of interaction effects of varying types and strength. TEGM is able to uncover two-way non-linear and linear interaction effects within the constructed GBM. Partial dependence plots were used to graphically examine the dependence of predictive models on subsets of selected features. To calculate partial dependence, we performed a previously described method that varies two selected variables within a GBM while averaging the effects of other variables. Fig. 5.4B displays the strengths of the interaction effects involving each of the 10 predictor variables. These interaction effects suggest that CD41%+ D5, Mean MFI Rate CD41 D0-D3, and CD34%+ D3 substantially interact with other variables. Strongly positive interactions between Mean MFI Rate CD41 Day 0-3 and CD41%+ on Day 5 suggest that the strong presentation of both factors is highly indicative of strong peak CD42%+ maturation. Interestingly, variables that account for maturation, such as factors related to CD41+%, negatively interact with factors related to progenitor status (CD34 percentage). A high CD34%+ early in culture combined with an high differentiation rate dampens the effect of peak CD42%+ maturation. Additionally, early viability percentage did not have substantial interaction effects with MK maturation and progenitor status.

5.5. Discussion

We provide a flow cytometric analysis method that can be used as a tool to predict responses (i.e. the peak percent commitment of a differentiating culture) from explanatory variables. This method compares both temporal trends of eigenfeatures and variability of metadata (such as population counts) with integrated eigenfeatures. This method can be used with different sets of related flow-cytometry data, providing versatile and applicable analysis to other cell differentiation protocols. We note a few important considerations required to successfully implement this method: several variables involving experimental technique can affect data clarity, such as cellular toxicity related to the protocol, the time of exposure to antibodies, the number of flow cytometry machines on which the cells are assayed, and the type of cellular differentiation. As several have noted, these factors can greatly affect the consistency and presentations of the results [144]. We did not test the robustness of our computational methods to perturbations in the form of noise (either chemically-derived or instrumentation) or by using different types of flow cytometry machines (benchtops or different styles). Thus, we recommend benchmarking this technique with an internal gold standard before utilizing this method to identify novel classes or subsets of any response variable.

Flow cytometry is an important tool in molecular biology that is applicable for assessing cellular differentiation [145, 146, 147]. One of the biggest challenges to investigate the surface marker function in cell differentiation studies is to have reliable methods for detection of population events, especially tracking populations over time. The ideal approach should be selected depending on cell type and experimental needs, and should have high efficiency, low bias, and be easy to use and reproducible. A rapid, robust and reliable method that can be used to efficiently quantify donor heterogeneity based on time-series flow cytometry

is not only valuable for data analysis, but also can be an important tool in designing new interventions to improve cellular differentiation.

Automated gating techniques have been shown to overcome some of the major limitations of manual analysis, but several challenges remain. A major challenge is the consistent labelling of cell populations across multiple samples in the presence of biological variability and heterogeneity [148]. If the variability is small, cells from different samples can be clustered as one dataset. However, when the variability increases, it becomes necessary to map corresponding clusters from different samples. While some techniques, such as BayesFlow [149], FLAME [150], HDPGMM [148], and ASPIRE [147] already incorporate such a mapping, many other techniques (including TEGM presented in this thesis) do not model this step. TEGM identifies populations that are assumed to be consistently Gaussian, which suffices for relatively well-defined populations such as *ex vivo* cell differentiation, but suffers in contexts where whole blood samples across multiple centers are processed. The FlowMapFriedman-Rafsky algorithm was developed as a post-processing step to fill this gap [151], but mapping cell types across samples still remains a challenging issue. Again, standardization will be crucial, especially when comparing data between multiple centres or over time. Additionally, TEGM needs to be augmented with additional flow cytometry analysis tools that enable the analysis of flow cytometry data with greater complexity.

One challenge of this algorithm concerns the identification of very rare cell types, which are easily mistaken for noise by many clustering algorithms. TEGM utilizes a tree-based approach to identify features that are and are not helpful for classification using the Variable Importance metric. With the traditional clustering algorithms, it is recommended to ensure that only relevant markers are used for the clustering —thus, TEGM represents a substantial improvement. Markers that vary little or that indicate properties not relevant for cell type

identification (for example, viability markers) are best left out, as these will only contribute noise to the feature extraction. Others have utilized an adapted distance measure, which assigns different importance scores to different markers, which can in some cases also be helpful in removing variables that are not important [148]. We have not yet benchmarked our framework against other techniques, as this is ongoing work.

5.6. Conclusion and Summary

We have developed a novel computational algorithm to quantify and predict temporal trends of developing cell subpopulations in flow cytometry data, called Topographical Extraction of Gaussian Models (TEGM). We applied our method to a novel time-series flow cytometry dataset comprised of 720 features from 80 perturbations of 54 samples, with 7 time-point measurements, capturing *ex vivo* MK differentiation and maturation of hematopoietic cells from donors with varying potential to generate $CD41^{+}CD42^{+}$ cells. We demonstrate the ability for TEGM to identify latent donor heterogeneity using the training set. TEGM predicted peak $CD41^{+}CD42^{+}$ maturation purities of megakaryocyte cultures derived from diverse individual donors that were highly concordant with actual peak culture outcomes (p=7.4e-09, R = 0.87). The resulting model captured continuous progression of cell development processes represented by cell surface markers (CD34, CD41, and CD42), and predicts MK differentiation and maturation potential in culture of a given CB unit in a separate validation set. We applied the TEGM computational framework to generate additional quantitative metrics capturing subtle aspects of megakaryocyte differentiation. This method overcomes limitations of prior established methods such as those that employ non-automated gating, to evaluate the effects of differentiation, and provides substantial improvements in automated flow cytometry detection methods.

CHAPTER 6

Conclusions and Future Outlook

The bottlenecks for producing clinically relevant numbers of PLPs include the limited understanding of transcription factor networks controlling MK commitment, limited expansion and differentiation of input cells, and insufficient platelet production from differentiated megakaryocytes. I have elucidated enhanced methods for megakaryocyte production from $CD34^+$ umbilical cord blood cells (CB) and platelet release that takes a step towards addressing the aforementioned bottlenecks. While these methods enable improved generation of megakaryocytes and platelet–like–particles with comparable morphology and function, we still fall short of generating clinically relevant amounts of platelets for transfusion.

Overall this thesis presents a novel culture strategy, as well as possible mechanisms of MK differentiation. A key finding is that we can transiently expand the number of $CD34^+$ cells and increase the total number of $CD41a^+CD42b^+$ mature MKs and PLPs per input $CD34^+$ cell by several fold compared to cultures without pre–expansion and previous reports by others. I showed that extending the pre–expansion culture to 8 days in the presence valproic acid (VPA) greatly increased MK and PLP production in secondary culture. To provide mechanistic insights, I hypothesized that VPA affects cell cycle and senescence pathways that involve cyclin–dependent kinases, p16INK4 and p21Cip/Waf1, and showed that p16INK4 and p21Cip/Waf1 levels are substantially increased with pre–expansion, but to a lesser extent with VPA addition. I also demonstrated that the resulting $CD41a^+CD42b^+$ PLPs aggregate under *in vitro* shear conditions in an ADP–dependent manner. Uniquely, I was able to build

a model capable of distinguishing between CB units with high–and low–MK–potential using statistical modeling and show that variability in culture performance can be predicted using factors observed earlier in culture.

More generally, this culture strategy and associated insights provide unique perspectives towards several important aspects of *ex vivo* platelet generation, such as the effects of stem and progenitor cell pre–expansion and donor heterogeneity. I expect that this combination of hematopoietic progenitor cell expansion and the generation of megakaryocytes and platelet–like–particles—in combination with computational analysis to identify sources of heterogeneity will have an impact on several in the field that have noticed donor–to–donor variation in mouse and human MKs. My work has the potential to be translated into strategies that effectively generate culture–derived platelets at a clinically relevant scale.

In the future, I would like to identify cytokine combinations that optimize the commitment and proliferation of MKs (expanded or unexpanded) in culture. I have started this work with a preliminary screen by systematically varying several factors explored in this thesis such as O_2 and cytokine level.

I screened 30 conditions probing fed–batch media dilution and explored various dosages of our previously published cytokine treatment regimens at various lengths of low O_2 incubation. Several promising conditions that increase MK purity or production level were identified systematically (Fig. 6.1). For example, it appears that longer incubation periods in low O_2 perturbs both the purity and production of $CD41a^+CD42b^+$ (Fig. 6.1). Regime 2, which utilizes higher doses of IL–3, IL–6, and IL–11 greatly improves TNC production (Fig. 6.2) at the expense of MK purity (Fig. 6.1).

As noted by others [152], fed–batch media dilution in general enables rapid expansion of HSPCs within our culture (Fig. 6.2). Titrating cytokines in fed–batch dilution schemes

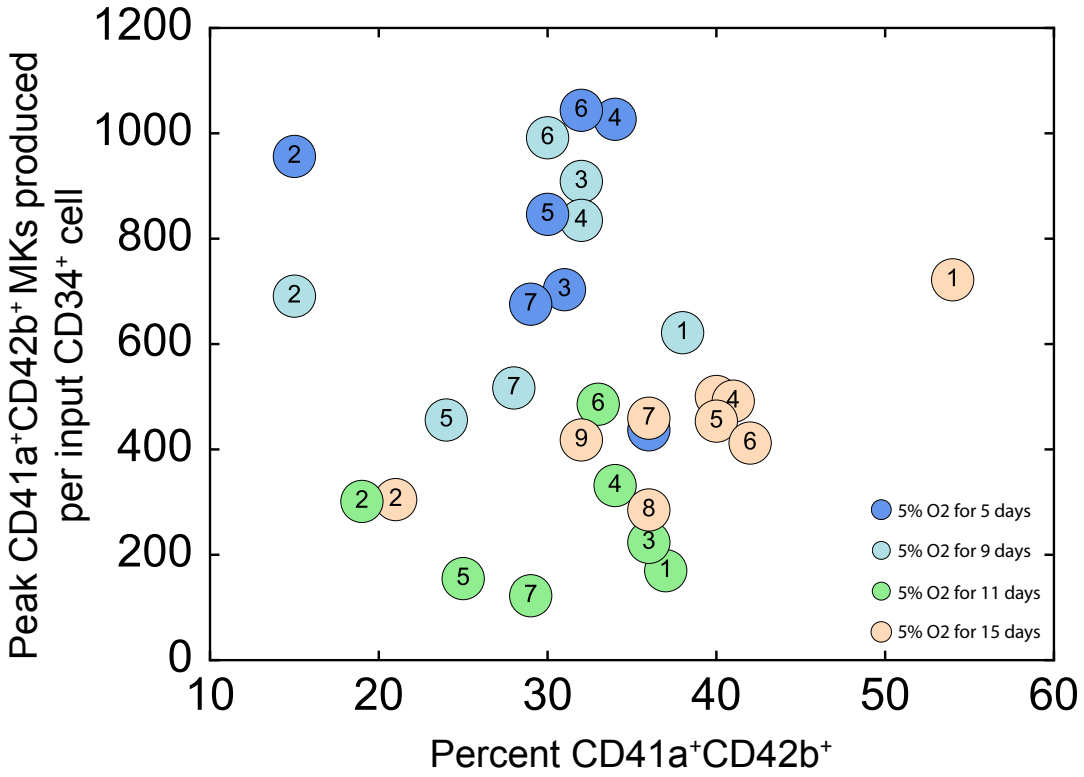

Figure 6.1 | **Screen of conditions of UCB not included in published work.** Screen of 30 conditions varying length of low O_2 period and cytokine regimes 1–7 identifies desirable culture conditions in terms of MK purity and peak MK production.

could result in higher MK production (Fig. 6.3). Thus, our goal is to optimize cultures along the axis of MK purity and overall culture growth to consistently generate 1000 MKs per input CD34+ cell at high (> 50%) purity. It is possible to screen several conditions involving cytokine–dosage optimization, fed–batch vs. full resuspension, extended low O_2 screen to identify conditions that optimize MK production.

Studies exploring optimization of *ex vivo* megakaryocyte cultures in the future should aim to:

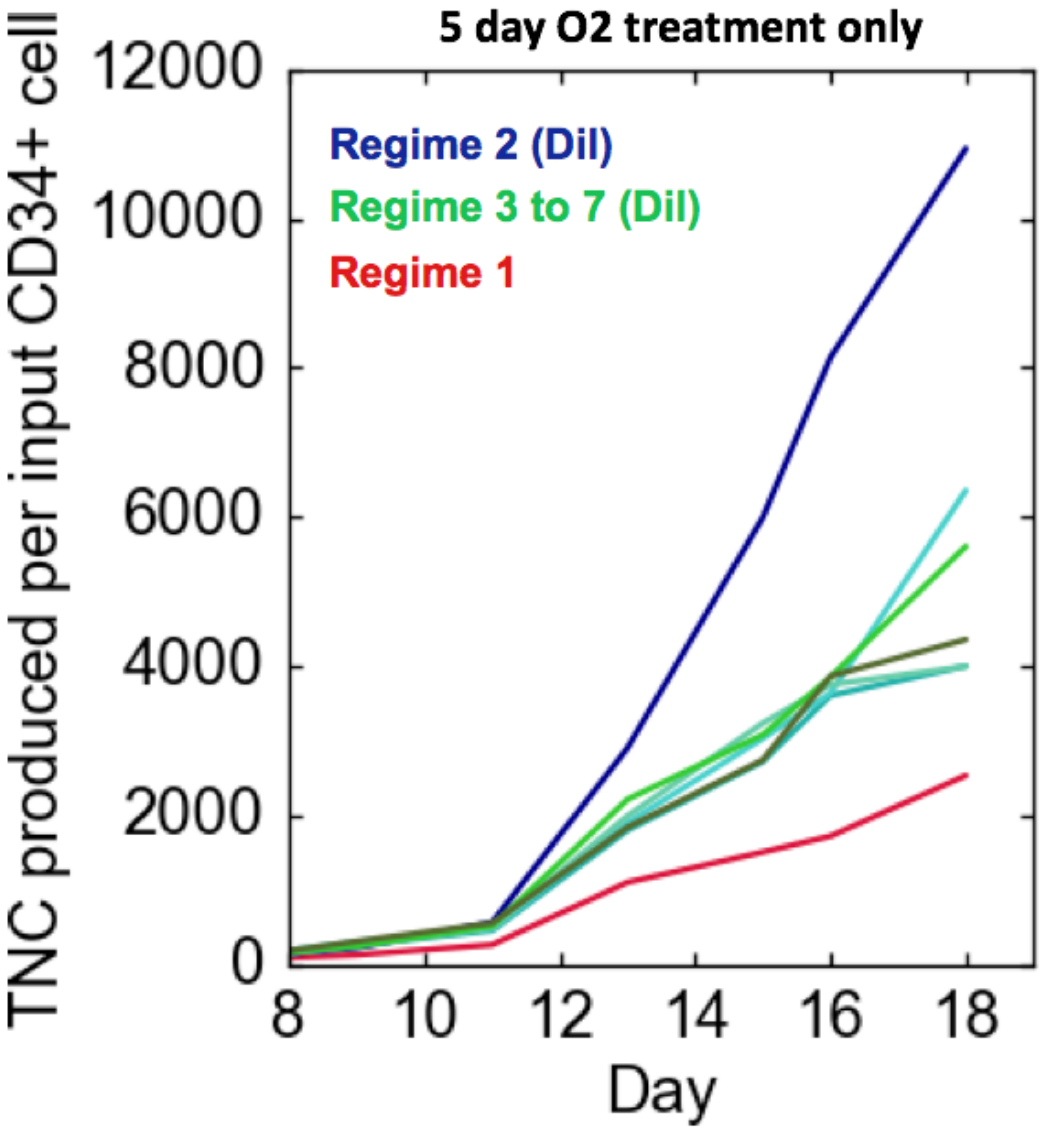

Figure 6.2 | **Fed–batch cytokine regimes enable greater proliferation of TNCs.** Cytokine Regimes 2–7 which utilize fed–batch method of feeding of various levels of cytokines results in greatly increased TNC growth compared to Regime 1 (control with no dilution and resuspension) at 5 day O_2 level.

1) Develop a flow cytometry–based screening approach focused on screening megakaryocytic features, such as polyploidization or CD42b expression.

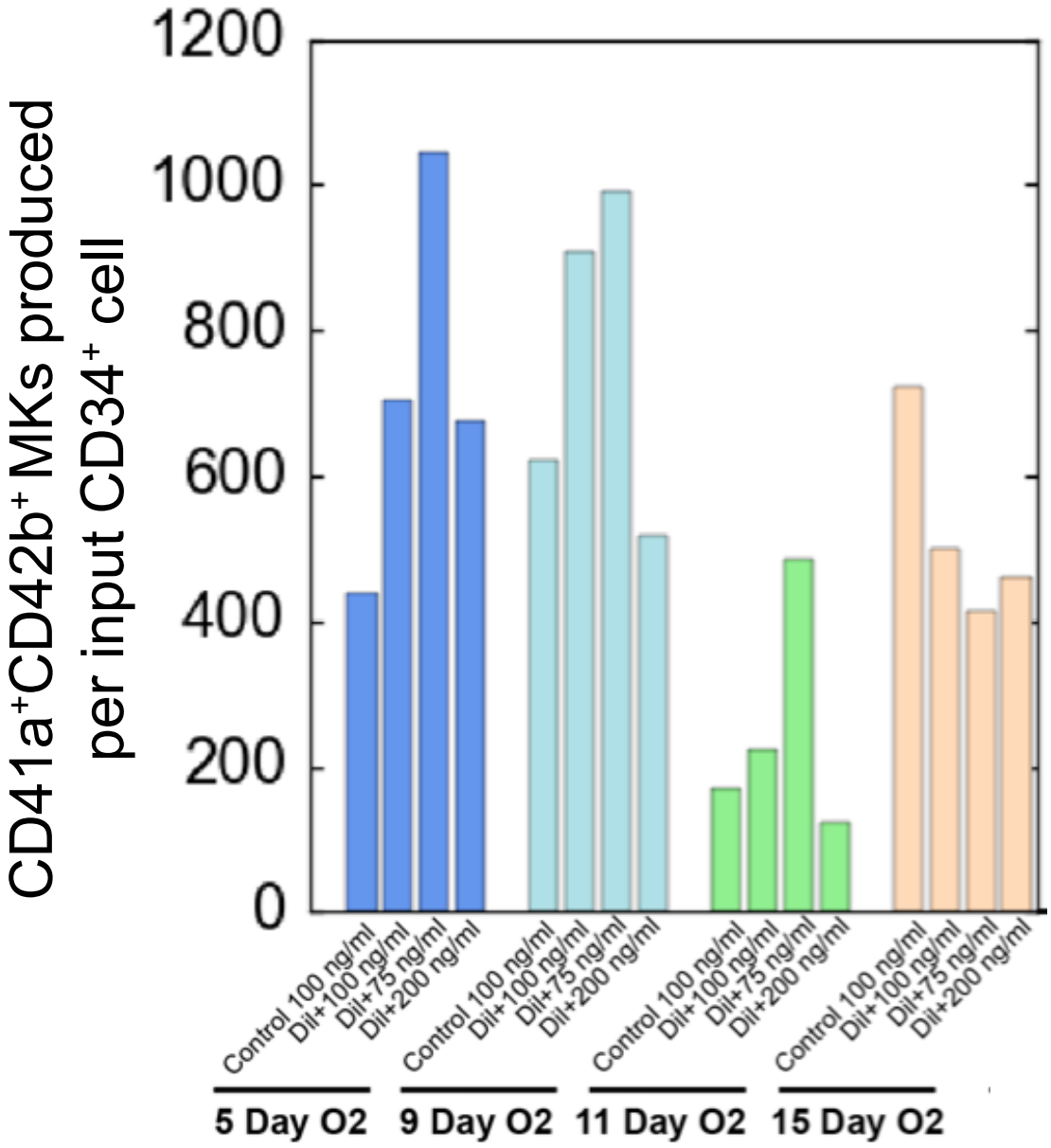

Figure 6.3 | **Peak MK production under explored conditions is several fold greater than control.** Peak MKs produced per input CD34+ cell tends to be highest with dilution of lower concentration of cytokines except in 15–day extended low O_2.

2) Conduct small scale screens perturbing oxygen level and small molecule regimen to affect signaling pathways and epigenetic regulators, and measure the outcome on downstream megakaryocyte development and cell cycle.

3) Employ high–throughput flow cytometry analysis tools to identify hits and experimentally validate results in focused culture scenarios where standard well–plates or flasks are used.

This involves greatly scaling down current approaches to culture MKs into 96–well plates. Screening requires the consideration of certain factors such as optimizing cell density, timing of incubations, reagent additions, media changes, and toxicity of small molecule conditions. Additionally, mechanosensory modalities may be applied, such as using plate vortexing to stimulate platelet–like–particle (PLP) production in microplates. Various intensities and durations that promote PLP release and maintain platelet viability under these modalities would need to be optimized. Thus, exploration of factors perturbing a variety of gene regulatory networks, and careful observation of response under various donor backgrounds has the potential of enabling clinically significant production of donor–independent platelets, which would broadly impact cancer treatment and other forms of care.

Bibliography

[1] Julius Bizzozero. Ueber einen neuen Formbestandtheil des Blutes und dessen Rolle bei der Thrombose und der Blutgerinnung. *Archiv fr pathologische Anatomie und Physiologie und fr klinische Medicin*, 90(2):261–332, November 1882. ISSN 0720-8723, 1432-2307. doi: 10.1007/BF01931360. URL `https://link.springer.com/article/10.1007/BF01931360`.

[2] Max Schultze. Ein heizbarer Objecttisch und seine Verwendung bei Untersuchungen des Blutes. *Archiv fr mikroskopische Anatomie*, 1(1):1–42, December 1865. ISSN 0176-7364. doi: 10.1007/BF02961404. URL `https://link.springer.com/article/10.1007/BF02961404`.

[3] Kerstin Jurk and Beate E. Kehrel. Platelets: physiology and biochemistry. *Seminars in Thrombosis and Hemostasis*, 31(4):381–392, 2005. ISSN 0094-6176. doi: 10.1055/s-2005-916671.

[4] Henri H. Versteeg, Johan W. M. Heemskerk, Marcel Levi, and Pieter H. Reitsma. New Fundamentals in Hemostasis. *Physiological Reviews*, 93(1):327–358, January 2013. ISSN 0031-9333. doi: 10.1152/physrev.00016.2011. URL `https://www.physiology.org/doi/full/10.1152/physrev.00016.2011`.

[5] Manasa K. Nayak, Paresh P. Kulkarni, and Debabrata Dash. Regulatory role of proteasome in determination of platelet life span. *The Journal of Biological Chemistry*, 288(10):6826–6834, March 2013. ISSN 1083-351X. doi: 10.1074/jbc.M112.403154.

[6] Kellie R. Machlus and Joseph E. Italiano. The incredible journey: From megakaryocyte development to platelet formation. *The Journal of Cell Biology*, 201(6):785–796, June 2013. ISSN 0021-9525, 1540-8140. doi: 10.1083/jcb.201304054. URL http://jcb.rupress.org/content/201/6/785.

[7] Anita Eckly, Harry Heijnen, Fabien Pertuy, Willie Geerts, Fabienne Proamer, Jean-Yves Rinckel, Catherine Lon, Franois Lanza, and Christian Gachet. Biogenesis of the demarcation membrane system (DMS) in megakaryocytes. *Blood*, 123(6):921–930, February 2014. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2013-03-492330. URL http://www.bloodjournal.org/content/123/6/921.

[8] Laurence A. Harker and Clement A. Finch. Thrombokinetics in man. *The Journal of Clinical Investigation*, 48(6):963–974, June 1969. ISSN 0021-9738. doi: 10.1172/JCI106077. URL https://www.jci.org/articles/view/106077.

[9] Swapna Panuganti, Alaina C. Schlinker, Paul F. Lindholm, Eleftherios T. Papoutsakis, and William M. Miller. Three-Stage Ex Vivo Expansion of High-Ploidy Megakaryocytic Cells: Toward Large-Scale Platelet Production. *Tissue Engineering. Part A*, 19(7-8):998–1014, April 2013. ISSN 1937-3341. doi: 10.1089/ten.tea.2011.0111. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3592379/.

[10] L. C. Dore and J. D. Crispino. Transcription factor networks in erythroid cell and megakaryocyte development. *Blood*, 118(2):231–239, July 2011. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2011-04-285981. URL http://www.bloodjournal.org/cgi/doi/10.1182/blood-2011-04-285981.

[11] M. W. Long, C. H. Heffner, J. L. Williams, C. Peters, and E. V. Prochownik. Regulation of megakaryocyte phenotype in human erythroleukemia cells. *The Journal of Clinical Investigation*, 85(4):1072–1084, April 1990. ISSN 0021-9738. doi: 10.1172/JCI114538.

[12] X. Y. Yang, M. Kimura, E. Jeanclos, and A. Aviv. Cellular proliferation and telomerase activity in CHRF-288-11 cells. *Life Sciences*, 66(16):1545–1555, 2000. ISSN 0024-3205.

[13] Chantal Proulx, Lucie Boyer, Darin R. Hurnanen, and Ral Lemieux. Preferential ex vivo expansion of megakaryocytes from human cord blood CD34+-enriched cells in the presence of thrombopoietin and limiting amounts of stem cell factor and Flt-3 ligand. *Journal of Hematotherapy & Stem Cell Research*, 12(2):179–188, April 2003. ISSN 1525-8165. doi: 10.1089/152581603321628322.

[14] Franois Spitz and Eileen E. M. Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews. Genetics*, 13(9):613–626, September 2012. ISSN 1471-0064. doi: 10.1038/nrg3207.

[15] Marcel Geertz and Sebastian J. Maerkl. Experimental strategies for studying transcription factorDNA binding specificities. *Briefings in Functional Genomics*, 9(5-6):362–373, December 2010. ISSN 2041-2649. doi: 10.1093/bfgp/elq023. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3080775/`.

[16] Mark F Ciaccio, Justin D Finkle, Albert Y Xue, and Neda Bagheri. A systems approach to integrative biology: an overview of statistical methods to elucidate association and architecture. *Integrative and comparative biology*, page icu037, 2014.

[17] Emi Nagoshi, Camille Saini, Christoph Bauer, Thierry Laroche, Felix Naef, and Ueli Schibler. Circadian Gene Expression in Individual Fibroblasts: Cell-Autonomous and Self-Sustained Oscillators Pass Time to Daughter Cells. *Cell*, 119(5):693–705, November 2004. ISSN 0092-8674. doi: 10.1016/j.cell.2004.11.015. URL `http://www.sciencedirect.com/science/article/pii/S0092867404010542`.

[18] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher. Comprehensive Identification of Cell Cycleregulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, December 1998. ISSN 1059-1524, 1939-4586. doi: 10.1091/mbc.9.12.3273. URL http://www.molbiolcell.org/content/9/12/3273.

[19] Naama Geva-Zatorsky, Nitzan Rosenfeld, Shalev Itzkovitz, Ron Milo, Alex Sigal, Erez Dekel, Talia Yarnitzky, Yuvalal Liron, Paz Polak, Galit Lahav, and Uri Alon. Oscillations and variability in the p53 system. *Molecular Systems Biology*, 2:2006.0033, 2006. ISSN 1744-4292. doi: 10.1038/msb4100068.

[20] Yun-Jin Jiang, Birgit L. Aerne, Lucy Smithers, Catherine Haddon, David Ish-Horowicz, and Julian Lewis. Notch signalling and the synchronization of the somite segmentation clock. *Nature*, 408(6811):475–479, November 2000. ISSN 0028-0836. doi: 10.1038/35044091. URL https://www.nature.com/nature/journal/v408/n6811/full/408475a0.html.

[21] Shane Neph, Jeff Vierstra, Andrew B. Stergachis, Alex P. Reynolds, Eric Haugen, Benjamin Vernot, Robert E. Thurman, Sam John, Richard Sandstrom, Audra K. Johnson, and others. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, 2012. URL http://www.nature.com/nature/journal/v489/n7414/abs/nature11212.html.

[22] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[23] Aviv Madar, Alex Greenfield, Harry Ostrer, Eric Vanden-Eijnden, and Richard Bonneau. The Inferelator 2.0: a scalable framework for reconstruction of dynamic regulatory network models. *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2009:5448–5451, 2009. ISSN 1557-170X. doi: 10.1109/IEMBS.2009.5334018.

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

[25] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, Gustavo Stolovitzky, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

[26] E. P. van Someren, B. L. T. Vaes, W. T. Steegenga, A. M. Sijbers, K. J. Dechering, and M. J. T. Reinders. Least absolute regression network analysis of the murine osteoblast differentiation network. *Bioinformatics (Oxford, England)*, 22(4):477–484, February 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti816.

[27] Tarmo ij, Kirsi Granberg, and Harri Lhdesmki. Sorad: a systems biology approach to predict and modulate dynamic signaling pathway response from phosphoproteome time-course measurements. *Bioinformatics (Oxford, England)*, 29(10):1283–1291, May 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt130.

[28] Daniel E. Zak, Gregory E. Gonye, James S. Schwaber, and Francis J. Doyle. Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network. *Genome Research*, 13(11):2396–2405, November 2003. ISSN 1088-9051. doi: 10.1101/gr.1198103.

[29] A. Raue, V. Becker, U. Klingmller, and J. Timmer. Identifiability and observability analysis for experimental design in nonlinear dynamical models. *Chaos (Woodbury, N.Y.)*, 20(4):045105, December 2010. ISSN 1089-7682. doi: 10.1063/1.3528102.

[30] Vn Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE*, 5(9):e12776, September 2010. doi: 10.1371/journal.pone.0012776. URL `http://dx.doi.org/10.1371/journal.pone.0012776`.

[31] Neil D. Lawrence, Guido Sanguinetti, and Magnus Rattray. Modelling transcriptional regulation using Gaussian Processes. In P. B. Schlkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 785–792. MIT Press, 2007. URL `http://papers.nips.cc/paper/3119-modelling-transcriptional-regulation-using-gaussian-processes.pdf`.

[32] Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6(1):145, 2012.

[33] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, April 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1517384113. URL `http://www.pnas.org/`

content/113/15/3932.

[34] Vn Anh Huynh-Thu and Guido Sanguinetti. Combining tree-based and dynamical systems for the inference of gene regulatory networks. *Bioinformatics (Oxford, England)*, 31(10):1614–1622, May 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu863.

[35] George J. Murphy and Andrew D. Leavitt. A model for studying megakaryocyte development and biology. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6):3065–3070, March 1999. ISSN 0027-8424. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC15895/.

[36] Thomas Moreau, Amanda L. Evans, Louella Vasquez, Marloes R. Tijssen, Ying Yan, Matthew W. Trotter, Daniel Howard, Maria Colzani, Meera Arumugam, Wing Han Wu, Amanda Dalby, Riina Lampela, Guenaelle Bouet, Catherine M. Hobbs, Dean C. Pask, Holly Payne, Tatyana Ponomaryov, Alexander Brill, Nicole Soranzo, Willem H. Ouwehand, Roger A. Pedersen, and Cedric Ghevaert. Large-scale production of megakaryocytes from human pluripotent stem cells by chemically defined forward programming. *Nature Communications*, 7:11208, April 2016. ISSN 2041-1723. doi: 10.1038/ncomms11208. URL http://www.nature.com/ncomms/2016/160407/ncomms11208/full/ncomms11208.html.

[37] Yukitaka Ito, Sou Nakamura, Naoshi Sugimoto, Tomohiro Shigemori, Yoshikazu Kato, Mikiko Ohno, Shinya Sakuma, Keitaro Ito, Hiroki Kumon, Hidenori Hirose, Haruki Okamoto, Masayuki Nogawa, Mio Iwasaki, Shunsuke Kihara, Kosuke Fujio, Takuya Matsumoto, Natsumi Higashi, Kazuya Hashimoto, Akira Sawaguchi, Ken-ichi Harimoto, Masato Nakagawa, Takuya Yamamoto, Makoto Handa, Naohide Watanabe, Eiichiro Nishi, Fumihito Arai, Satoshi Nishimura, and Koji Eto. Turbulence Activates Platelet Biogenesis to Enable Clinical Scale ExVivo Production. *Cell*, 174

(3):636–648.e18, July 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.06.011. URL http://www.sciencedirect.com/science/article/pii/S0092867418307360.

[38] Amartya Sanyal, Bryan Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, September 2012. ISSN 0028-0836. doi: 10.1038/nature11279. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3555147/.

[39] Abigail D Bellis, Beatriz Pealver-Bernab, Michael S Weiss, Michael E Yarrington, Maria V Barbolina, Angela K Pannier, Jacqueline S Jeruss, Linda J Broadbelt, and Lonnie D Shea. Cellular arrays for large-scale analysis of transcription factor activity. *Biotechnology and bioengineering*, 108(2):395–403, February 2011. ISSN 1097-0290. doi: 10.1002/bit.22916.

[40] Michael S. Weiss, Beatriz Pealver Bernab, Abigail D. Bellis, Linda J. Broadbelt, Jacqueline S. Jeruss, and Lonnie D. Shea. Dynamic, Large-Scale Profiling of Transcription Factor Activity from Live Cells in 3d Culture. *PLoS ONE*, 5(11):e14026, November 2010. doi: 10.1371/journal.pone.0014026. URL http://dx.doi.org/10.1371/journal.pone.0014026.

[41] Chaolin Zhang, Zhenyu Xuan, Stefanie Otto, John R. Hover, Sean R. McCorkle, Gail Mandel, and Michael Q. Zhang. A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acids Research*, 34(8):2238–2246, 2006. ISSN 0305-1048. doi: 10.1093/nar/gkl248. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1456330/.

[42] Gregory M. Findlay, Evan A. Boyle, Ronald J. Hause, Jason C. Klein, and Jay Shendure. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature*, 513(7516):120–123, September 2014. ISSN 1476-4687. doi: 10.1038/nature13695.

[43] J A Sutherland, A R Turner, P Mannoni, L E McGann, and J M Turc. Differentiation of K562 leukemia cells along erythroid, macrophage, and megakaryocyte lineages. *Journal of biological response modifiers*, 5(3):250–262, June 1986. ISSN 0732-6580.

[44] Peter G. Fuhrken, Pani A. Apostolidis, Stephan Lindsey, William M. Miller, and Eleftherios T. Papoutsakis. Tumor Suppressor Protein p53 Regulates Megakaryocytic Polyploidization and Apoptosis. *Journal of Biological Chemistry*, 283(23):15589–15600, June 2008. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M801923200. URL `http://www.jbc.org/content/283/23/15589`.

[45] Lisa M Giammona, Peter G Fuhrken, Eleftherios T Papoutsakis, and William M Miller. Nicotinamide (vitamin B3) increases the polyploidisation and proplatelet formation of cultured primary human megakaryocytes. *British journal of haematology*, 135(4):554–566, November 2006. ISSN 0007-1048. doi: 10.1111/j.1365-2141.2006.06341.x.

[46] Petra Isabel Lorenzo, Elen Margrethe Brendeford, Siv Gilfillan, Alexey A. Gavrilov, Marit Leedsak, Sergey V. Razin, Ragnhild Eskeland, Thomas Saether, and Odd Stokke Gabrielsen. Identification of c-Myb Target Genes in K562 Cells Reveals a Role for c-Myb as a Master Regulator. *Genes & Cancer*, 2(8):805–817, August 2011. ISSN 1947-6019. doi: 10.1177/1947601911428224. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3278898/`.

[47] Nicola K Wilson, Samuel D Foster, Xiaonan Wang, Kathy Knezevic, Judith Schtte, Polynikis Kaimakis, Paulina M Chilarska, Sarah Kinston, Willem H Ouwehand, Elaine Dzierzak, John E Pimanda, Marella F T R de Bruijn, and Berthold Gttgens. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell stem cell*, 7(4):532–544, October 2010. ISSN 1875-9777. doi: 10.1016/j.stem.2010.07.016.

[48] Peter G. Fuhrken, Chi Chen, William M. Miller, and Eleftherios T. Papoutsakis. Comparative, genome-scale transcriptional analysis of CHRF-288-11 and primary human megakaryocytic cell cultures provides novel insights into lineage-specific differentiation. *Experimental Hematology*, 35(3):476–489, March 2007. ISSN 0301-472X. doi: 10.1016/j.exphem.2006.10.017.

[49] Aurelie de Thonel, Julie Vandekerckhove, David Lanneau, Subramaniam Selvakumar, Genevive Courtois, Adonis Hazoume, Mathilde Brunet, Sebastien Maurel, Arlette Hammann, Jean Antoine Ribeil, Yael Zermati, Anne Sophie Gabet, Joan Boyes, Eric Solary, Olivier Hermine, and Carmen Garrido. HSP27 controls GATA-1 protein level during erythroid cell differentiation. *Blood*, 116(1):85–96, July 2010. ISSN 1528-0020. doi: 10.1182/blood-2009-09-241778.

[50] Mira T. Kassouf, Hedia Chagraoui, Paresh Vyas, and Catherine Porcher. Differential use of SCL/TAL-1 DNA-binding domain in developmental hematopoiesis. *Blood*, 112(4):1056–1067, August 2008. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2007-12-128900. URL `http://bloodjournal.hematologylibrary.org/content/112/4/1056`.

[51] C Labbaye, M Valtieri, T Barberi, E Meccia, B Masella, E Pelosi, G L Condorelli, U Testa, and C Peschle. Differential expression and functional role of GATA-2, NF-E2, and GATA-1 in normal adult hematopoiesis. *The Journal of clinical investigation*, 95(5):2346–2358, May 1995. ISSN 0021-9738. doi: 10.1172/JCI117927.

[52] M Athanasiou, G Mavrothalassitis, L Sun-Hoffman, and D G Blair. FLI-1 is a suppressor of erythroid differentiation in human hematopoietic cells. *Leukemia*, 14(3):439–445, March 2000. ISSN 0887-6924.

[53] V. Matys, E. Fricke, R. Geffers, E. Gling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Mnch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378, January 2003. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkg108. URL `http://nar.oxfordjournals.org/content/31/1/374`.

[54] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, November 2003. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.1239303. URL `http://genome.cshlp.org/content/13/11/2498`.

[55] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, January 1995. ISSN 0035-9246. URL `http://www.jstor.org/stable/2346101`.

[56] Gordon K Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25, January 2004. ISSN 1544-6115. doi: 10.2202/1544-6115.1027. URL `http://www.degruyter.com/view/j/sagmb.2004.3.issue-1/sagmb.2004.3.1.1027/sagmb.2004.3.1.1027.xml`.

[57] Erna Magnsdttir, Sabine Dietmann, Kazuhiro Murakami, Ufuk Gnesdogan, Fuchou Tang, Siqin Bao, Evangelia Diamanti, Kaiqin Lao, Berthold Gottgens, and M. Azim Surani. A tripartite transcription factor network regulates primordial germ cell specification in mice. *Nature Cell Biology*, 15(8):905–915, August 2013. ISSN

1465-7392. doi: 10.1038/ncb2798. URL http://www.nature.com/ncb/journal/v15/n8/abs/ncb2798.html.

[58] Eran Segal, Tali Raveh-Sadka, Mark Schroeder, Ulrich Unnerstall, and Ulrike Gaul. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, 451(7178):535–540, January 2008. ISSN 0028-0836. doi: 10.1038/nature06496. URL http://www.nature.com/nature/journal/v451/n7178/full/nature06496.html.

[59] Brian A. Aguado, Jia J. Wu, Samira M. Azarin, Dhaval Nanavati, Shreyas S. Rao, Grace G. Bushnell, Chaitanya B. Medicherla, and Lonnie D. Shea. Secretome identification of immune cell factors mediating metastatic cell homing. *Scientific Reports*, 5: 17566, 2015. ISSN 2045-2322. doi: 10.1038/srep17566.

[60] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[61] Aurélie C Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. Grouped graphical granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, 25(12):i110–i118, 2009.

[62] Pietro Zoppoli, Sandro Morganella, and Michele Ceccarelli. Timedelay-aracne: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC bioinformatics*, 11(1):154, 2010.

[63] Ali Shojaie and George Michailidis. Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010.

[64] Francesca Petralia, Pei Wang, Jialiang Yang, and Zhidong Tu. Integrative random forest for gene regulatory network inference. *Bioinformatics (Oxford, England)*, 31

(12):i197–205, June 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv268.

[65] Christopher J. Quinn, Todd P. Coleman, Negar Kiyavash, and Nicholas G. Hatsopoulos. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *Journal of Computational Neuroscience*, 30(1):17–44, February 2011. ISSN 1573-6873. doi: 10.1007/s10827-010-0247-2.

[66] Alexandre Irrthum, Louis Wehenkel, Pierre Geurts, et al. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

[67] Mark F Ciaccio, Vincent C Chen, Richard B Jones, and Neda Bagheri. The dionesus algorithm provides scalable and accurate reconstruction of dynamic phosphoproteomic networks to reveal new drug targets. *Integrative Biology*, 7(7):776–791, 2015.

[68] Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.

[69] Daniel Marbach, Robert J Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the national academy of sciences*, 107(14):6286–6291, 2010.

[70] Steven M. Boker, Minquan Xu, Jennifer L. Rotondo, and Kadijah King. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, 7(3):338–355, September 2002. ISSN 1082-989X.

[71] Tomáš Gedeon and Pavol Bokes. Delayed Protein Synthesis Reduces the Correlation between mRNA and Protein Fluctuations. *Biophysical Journal*, 103(3):377–385, August 2012. ISSN 0006-3495. doi: 10.1016/j.bpj.2012.06.025. URL `http:`

//www.ncbi.nlm.nih.gov/pmc/articles/PMC3414885/.

[72] Eric W. Weisstein. Bonferroni Correction, 2017. URL http://mathworld.wolfram.com/BonferroniCorrection.html.

[73] Socorro Gama-Castro, Heladia Salgado, Alberto Santos-Zavaleta, Daniela Ledezma-Tejeida, Luis Muñiz-Rascado, Jair Santiago García-Sotelo, Kevin Alquicira-Hernández, Irma Martínez-Flores, Lucia Pannier, Jaime Abraham Castro-Mondragón, Alejandra Medina-Rivera, Hilda Solano-Lira, César Bonavides-Martínez, Ernesto Pérez-Rueda, Shirley Alquicira-Hernández, Liliana Porrón-Sotelo, Alejandra López-Fuentes, Anastasia Hernández-Koutoucheva, Víctor Del Moral-Chávez, Fabio Rinaldi, and Julio Collado-Vides. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, 44(D1):D133–143, January 2016. ISSN 1362-4962. doi: 10.1093/nar/gkv1156.

[74] DV Klopfenstein, Haibao Tang, Fidel Ramirez, Olga Botvinnik, Brent Pedersen Bioinformatics, Patrick Flick, Kenta Sato ( ), Chris Mungall, Uwe Schmitt, Gregory Stupp, David DeTomaso, chri, Mark Fiers, Lucas van Dijk, Jeff Yunes, and Douglas Myers-Turnbull. Goatools v0.6.10. *Zenodo*, October 2016. doi: 10.5281/zenodo.159493. URL https://zenodo.org/record/159493.

[75] Interpolation (scipy.interpolate) — scipy v0.19.0 reference guide, 2017. URL https://docs.scipy.org/doc/scipy/reference/tutorial/interpolate.html.

[76] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science Engineering*, 13 (2):22–30, March 2011. ISSN 1521-9615. doi: 10.1109/MCSE.2011.37.

[77] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*,

pages 51 – 56, 2010.

[78] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, August 2008.

[79] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[80] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, 9(3):90–95, May 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.55.

[81] Harley H. McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3):814–819, February 1997. ISSN 0027-8424, 1091-6490. URL `http://www.pnas.org/content/94/3/814`.

[82] Michal Ronen, Revital Rosenberg, Boris I Shraiman, and Uri Alon. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proceedings of the national academy of sciences*, 99(16):10555–10560, 2002.

[83] Mukesh Bansal, Giusy Della Gatta, and Diego di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics (Oxford, England)*, 22(7):815–822, April 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl003.

[84] Nizamul Morshed, Madhu Chetty, and Nguyen Xuan Vinh. Simultaneous learning of instantaneous and time-delayed genetic interactions using novel information theoretic scoring technique. *BMC Systems Biology*, 6:62, 2012. ISSN 1752-0509. doi: 10.1186/1752-0509-6-62. URL `http://dx.doi.org/10.1186/1752-0509-6-62`.

[85] Szymon Jozefczuk, Sebastian Klie, Gareth Catchpole, Jedrzej Szymanski, Alvaro Cuadros-Inostroza, Dirk Steinhauser, Joachim Selbig, and Lothar Willmitzer. Metabolomic and transcriptomic stress response of Escherichia coli. *Molecular Systems Biology*, 6:364, May 2010. ISSN 1744-4292. doi: 10.1038/msb.2010.18.

[86] Jeremiah J. Faith, Boris Hayete, Joshua T. Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J. Collins, and Timothy S. Gardner. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8, January 2007. ISSN 1545-7885. doi: 10.1371/journal.pbio.0050008.

[87] Gary D. Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, 2003. ISSN 1471-2105. doi: 10.1186/1471-2105-4-2. URL `http://dx.doi.org/10.1186/1471-2105-4-2`.

[88] Ophir Shalem, Orna Dahan, Michal Levo, Maria Rodriguez Martinez, Itay Furman, Eran Segal, and Yitzhak Pilpel. Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Molecular Systems Biology*, 4:223, 2008. ISSN 1744-4292. doi: 10.1038/msb.2008.59.

[89] Caroline M. Li and Robert R. Klevecz. A rapid genome-scale response of the transcriptional oscillator to perturbation reveals a period-doubling path to phenotypic change. *Proceedings of the National Academy of Sciences of the United States of America*, 103 (44):16254–16259, October 2006. ISSN 0027-8424. doi: 10.1073/pnas.0604860103.

[90] David A. Orlando, Charles Y. Lin, Allister Bernard, Jean Y. Wang, Joshua E. S. Socolar, Edwin S. Iversen, Alexander J. Hartemink, and Steven B. Haase. Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature*,

453(7197):944–947, June 2008. ISSN 1476-4687. doi: 10.1038/nature06955.

[91] Y. L. Ganduri, S. R. Sadda, M. W. Datta, R. K. Jambukeswaran, and P. Datta. Tdca, a transcriptional activator of the tdcabc operon of escherichia coli, is a member of the lysr family of proteins. *Molecular & general genetics: MGG*, 240(3):395–402, September 1993. ISSN 0026-8925.

[92] Tomohiro Shimada, Nobuyuki Fujita, Kaneyoshi Yamamoto, and Akira Ishihama. Novel roles of camp receptor protein (crp) in regulation of transport and metabolism of carbon sources. *PLOS ONE*, 6(6):e20081, June 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0020081. URL `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0020081`.

[93] Jason Crack, Jeffrey Green, and Andrew J. Thomson. Mechanism of oxygen sensing by the bacterial transcription factor fumarate-nitrate reduction (fnr). *The Journal of Biological Chemistry*, 279(10):9278–9286, March 2004. ISSN 0021-9258. doi: 10.1074/jbc.M309878200.

[94] Katy C. Kao, Linh M. Tran, and James C. Liao. A Global Regulatory Role of Gluconeogenic Genes in Escherichia coli Revealed by Transcriptome Network Analysis. *Journal of Biological Chemistry*, 280(43):36079–36087, October 2005. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M508202200. URL `http://www.jbc.org/content/280/43/36079`.

[95] Junmei Zhang, Robert Sprung, Jimin Pei, Xiaohong Tan, Sungchan Kim, Heng Zhu, Chuan-Fa Liu, Nick V. Grishin, and Yingming Zhao. Lysine Acetylation Is a Highly Abundant and Evolutionarily Conserved Modification in Escherichia Coli. *Molecular & Cellular Proteomics : MCP*, 8(2):215–225, February 2009. ISSN 1535-9476. doi: 10.1074/mcp.M800187-MCP200. URL `https://www.ncbi.nlm.nih.gov/pmc/`

articles/PMC2634580/.

[96] Jakob Runge, Vladimir Petoukhov, Jonathan F. Donges, Jaroslav Hlinka, Nikola Jajcay, Martin Vejmelka, David Hartman, Norbert Marwan, Milan Paluš, and Jürgen Kurths. Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications*, 6:ncomms9502, October 2015. ISSN 2041-1723. doi: 10.1038/ncomms9502. URL https://www.nature.com/articles/ncomms9502.

[97] Matthew E. Studham, Andreas Tjrnberg, Torbjrn E.M. Nordling, Sven Nelander, and Erik L. L. Sonnhammer. Functional association networks as priors for gene regulatory network inference. *Bioinformatics*, 30(12):i130–i138, June 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu285. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4058914/.

[98] Steven M. Hill, Laura M. Heiser, Thomas Cokelaer, Michael Unger, Nicole K. Nesser, Daniel E. Carlin, Yang Zhang, Artem Sokolov, Evan O. Paull, Chris K. Wong, Kiley Graim, Adrian Bivol, Haizhou Wang, Fan Zhu, Bahman Afsari, Ludmila V. Danilova, Alexander V. Favorov, Wai Shing Lee, Dane Taylor, Chenyue W. Hu, Byron L. Long, David P. Noren, Alexander J. Bisberg, HPN-DREAM Consortium, Gordon B. Mills, Joe W. Gray, Michael Kellen, Thea Norman, Stephen Friend, Amina A. Qutub, Elana J. Fertig, Yuanfang Guan, Mingzhou Song, Joshua M. Stuart, Paul T. Spellman, Heinz Koeppl, Gustavo Stolovitzky, Julio Saez-Rodriguez, and Sach Mukherjee. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature Methods*, 13(4):310–318, April 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3773.

[99] Matthew T. Rondina, Andrew S. Weyrich, and Guy A. Zimmerman. Platelets as Cellular Effectors of Inflammation in Vascular Diseases. *Circulation research*, 112(11):

1506–1519, May 2013. ISSN 0009-7330. doi: 10.1161/CIRCRESAHA.113.300512. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3738064/.

[100] Cara C. Bertozzi, Alec A. Schmaier, Patricia Mericko, Paul R. Hess, Zhiying Zou, Mei Chen, Chiu-Yu Chen, Bin Xu, Min-min Lu, Diane Zhou, Eric Sebzda, Matthew T. Santore, Demetri J. Merianos, Matthias Stadtfeld, Alan W. Flake, Thomas Graf, Radek Skoda, Jonathan S. Maltzman, Gary A. Koretzky, and Mark L. Kahn. Platelets regulate lymphatic vascular development through CLEC-2-SLP-76 signaling. *Blood*, 116 (4):661–670, July 2010. ISSN 1528-0020. doi: 10.1182/blood-2010-02-270876.

[101] Laurie J. Gay and Brunhilde Felding-Habermann. Contribution of platelets to tumour metastasis. *Nature Reviews. Cancer*, 11(2):123–134, February 2011. ISSN 1474-1768. doi: 10.1038/nrc3004.

[102] Lorna M. Williamson and Dana V. Devine. Challenges in the management of the blood supply. *Lancet (London, England)*, 381(9880):1866–1875, May 2013. ISSN 1474-547X. doi: 10.1016/S0140-6736(13)60631-5.

[103] Takuya Matsunaga, Ikuta Tanaka, Masayoshi Kobune, Yutaka Kawano, Maki Tanaka, Kageaki Kuribayashi, Satoshi Iyama, Tsutomu Sato, Yasushi Sato, Rishu Takimoto, Tetsuji Takayama, Junji Kato, Takafumi Ninomiya, Hirofumi Hamada, and Yoshiro Niitsu. Ex vivo large-scale generation of human platelets from cord blood CD34+ cells. *Stem Cells (Dayton, Ohio)*, 24(12):2877–2887, December 2006. ISSN 1066-5099. doi: 10.1634/stemcells.2006-0309.

[104] Shi-Jiang Lu, Feng Li, Hong Yin, Qiang Feng, Erin A. Kimbrel, Eunsil Hahm, Jonathan N. Thon, Wei Wang, Joseph E. Italiano, Jaehyung Cho, and Robert Lanza. Platelets generated from human embryonic stem cells are functional in vitro and in the microcirculation of living mice. *Cell Research*, 21(3):530–545, March 2011. ISSN

1748-7838. doi: 10.1038/cr.2011.8.

[105] Naoya Takayama and Koji Eto. Pluripotent stem cells reveal the developmental biology of human megakaryocytes and provide a source of platelets for clinical application. *Cellular and Molecular Life Sciences*, 69(20):3419–3428, October 2012. ISSN 1420-682X. doi: 10.1007/s00018-012-0995-4. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3445798/`.

[106] S. S. Tung, S. Parmar, S. N. Robinson, M. De Lima, and E. J. Shpall. Ex vivo expansion of umbilical cord blood for transplantation. *Best Practice & Research Clinical Haematology*, 23(2):245–257, June 2010. ISSN 1521-6926. doi: 10.1016/j.beha.2010.06.004. URL `http://www.sciencedirect.com/science/article/pii/S1521692610000344`.

[107] Nadim Mahmud, Benjamin Petro, Sudhakar Baluchamy, Xinmin Li, Simona Taioli, Donald Lavelle, John G. Quigley, Montha Suphangul, and Hiroto Araki. Differential effects of epigenetic modifiers on the expansion and maintenance of human cord blood stem/progenitor cells. *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation*, 20(4):480–489, April 2014. ISSN 1523-6536. doi: 10.1016/j.bbmt.2013.12.562.

[108] S. Querol, S. G. Gomez, A. Pagliuca, M. Torrabadella, and J. A. Madrigal. Quality rather than quantity: the cord blood bank dilemma. *Bone Marrow Transplantation*, 45(6):970–978, June 2010. ISSN 1476-5365. doi: 10.1038/bmt.2010.7.

[109] Miguel A. Diaz, Marta Gonzalez-Vicent, Manuel Ramirez, Julian Sevilla, Alvaro Lassaletta, Antonio Perez, and Luis Madero. ALLOGENEIC CORD BLOOD TRANSPLANTATION IN CHILDREN WITH HEMATOLOGICAL MALIGNANCIES: A Long-Term Follow-Up Single-Center Study. *Pediatric Hematology and Oncology*, 26 (4):165–174, January 2009. ISSN 0888-0018. doi: 10.1080/08880010902773040. URL

https://doi.org/10.1080/08880010902773040.

[110] Vinod K. Prasad, Adam Mendizabal, Suhag H. Parikh, Paul Szabolcs, Timothy A. Driscoll, Kristin Page, Sonali Lakshminarayanan, June Allison, Susan Wood, Deborah Semmel, Maria L. Escolar, Paul L. Martin, Shelly Carter, and Joanne Kurtzberg. Unrelated donor umbilical cord blood transplantation for inherited metabolic disorders in 159 pediatric patients from a single center: influence of cellular composition of the graft on transplantation outcomes. *Blood*, 112(7):2979–2989, October 2008. ISSN 1528-0020. doi: 10.1182/blood-2008-03-140830.

[111] John E. Wagner, Juliet N. Barker, Todd E. DeFor, K. Scott Baker, Bruce R. Blazar, Cindy Eide, Anne Goldman, John Kersey, William Krivit, Margaret L. MacMillan, Paul J. Orchard, Charles Peters, Daniel J. Weisdorf, Norma K. C. Ramsay, and Stella M. Davies. Transplantation of unrelated donor umbilical cord blood in 102 patients with malignant and nonmalignant diseases: influence of CD34 cell dose and HLA disparity on treatment-related mortality and survival. *Blood*, 100(5):1611–1618, September 2002. ISSN 0006-4971. doi: 10.1182/blood-2002-01-0294.

[112] K. H. Yoo, S. H. Lee, H.-J. Kim, K. W. Sung, H. L. Jung, E. J. Cho, H. K. Park, H. A. Kim, and H. H. Koo. The impact of post-thaw colony-forming units-granulocyte/macrophage on engraftment following unrelated cord blood transplantation in pediatric recipients. *Bone Marrow Transplantation*, 39(9):515–521, May 2007. ISSN 0268-3369. doi: 10.1038/sj.bmt.1705629.

[113] Jo-Anna Reems, Nicolas Pineault, and Sijie Sun. In Vitro Megakaryocyte Production and Platelet Biogenesis: State of the Art. *Transfusion medicine reviews*, 24(1):33–43, January 2010. ISSN 0887-7963. doi: 10.1016/j.tmrv.2009.09.003. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2790431/.

[114] Valrie Cortin, Alain Garnier, Nicolas Pineault, Ral Lemieux, Lucie Boyer, and Chantal Proulx. Efficient in vitro megakaryocyte maturation using cytokine cocktails optimized by statistical experimental design. *Experimental Hematology*, 33(10):1182–1191, October 2005. ISSN 0301-472X. doi: 10.1016/j.exphem.2005.06.020.

[115] Javad Hatami, Pedro Z. Andrade, Denise Bacalhau, Fernando Cirurgio, Frederico Castelo Ferreira, Joaquim M. S. Cabral, and Cludia L. da Silva. Proliferation extent of CD34+ cells as a key parameter to maximize megakaryocytic differentiation of umbilical cord blood-derived hematopoietic stem/progenitor cells in a two-stage culture protocol. *Biotechnology Reports*, 4:50–55, December 2014. ISSN 2215-017X. doi: 10.1016/j.btre.2014.07.002. URL `http://www.sciencedirect.com/science/article/pii/S2215017X14000253`.

[116] Nikola Ivetic, Ishac Nazi, Nadia Karim, Rumi Clare, James W. Smith, Jane C. Moore, Kristin J. Hope, John G. Kelton, and Donald M. Arnold. Producing megakaryocytes from a human peripheral blood source. *Transfusion*, 56(5):1066–1074, May 2016. ISSN 1537-2995. doi: 10.1111/trf.13461. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/trf.13461`.

[117] Rika Miyazaki, Hajime Ogata, Tomoko Iguchi, Sinji Sogo, Taketoshi Kushida, Tomoki Ito, Muneo Inaba, Susumu Ikehara, and Yohnosuke Kobayashi. Comparative analyses of megakaryocytes derived from cord blood and bone marrow. *British Journal of Haematology*, 108(3):602–609, March 2000. ISSN 1365-2141. doi: 10.1046/j.1365-2141.2000.01854.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2141.2000.01854.x`.

[118] Qiang Feng, Namrata Shabrani, JonathanN. Thon, Hongguang Huo, Austin Thiel, KellieR. Machlus, Kyungho Kim, Julie Brooks, Feng Li, Chenmei Luo, ErinA. Kimbrel,

Jiwu Wang, Kwang-Soo Kim, Joseph Italiano, Jaehyung Cho, Shi-Jiang Lu, and Robert Lanza. Scalable Generation of Universal Platelets from Human Induced Pluripotent Stem Cells. *Stem Cell Reports*, 3(5):817–831, October 2014. ISSN 2213-6711. doi: 10.1016/j.stemcr.2014.09.010. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4235139/.

[119] Sou Nakamura, Naoya Takayama, Shinji Hirata, Hideya Seo, Hiroshi Endo, Kiyosumi Ochi, Ken-ichi Fujita, Tomo Koike, Ken-ichi Harimoto, Takeaki Dohda, Akira Watanabe, Keisuke Okita, Nobuyasu Takahashi, Akira Sawaguchi, Shinya Yamanaka, Hiromitsu Nakauchi, Satoshi Nishimura, and Koji Eto. Expandable Megakaryocyte Cell Lines Enable Clinically Applicable Generation of Platelets from Human Induced Pluripotent Stem Cells. *Cell Stem Cell*, 14(4):535–548, April 2014. ISSN 1934-5909. doi: 10.1016/j.stem.2014.01.011. URL http://www.sciencedirect.com/science/article/pii/S1934590914000125.

[120] Rodolphe Besancenot, Ronan Chalign, Carole Tonetti, Florence Pasquier, Caroline Marty, Yann Lcluse, William Vainchenker, Stefan N. Constantinescu, and Stphane Giraudier. A Senescence-Like Cell-Cycle Arrest Occurs During Megakaryocytic Maturation: Implications for Physiological and Pathological Megakaryocytic Proliferation. *PLOS Biology*, 8(9):e1000476, September 2010. ISSN 1545-7885. doi: 10.1371/journal.pbio.1000476. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000476.

[121] G. H. Stein, L. F. Drullinger, A. Soulard, and V. Duli. Differential roles for cyclin-dependent kinase inhibitors p21 and p16 in the mechanisms of senescence and differentiation in human fibroblasts. *Molecular and Cellular Biology*, 19(3):2109–2117, March 1999. ISSN 0270-7306.

[122] David A. Alcorta, Yue Xiong, Dawn Phelps, Greg Hannon, David Beach, and J. Carl Barrett. Involvement of the cyclin-dependent kinase inhibitor p16 (INK4a) in replicative senescence of normal human fibroblasts. *Proceedings of the National Academy of Sciences*, 93(24):13742–13747, November 1996. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.93.24.13742. URL `http://www.pnas.org/content/93/24/13742`.

[123] Yingying Zhai, Xi Chen, Dehai Yu, Tao Li, Jiuwei Cui, Guanjun Wang, Ji-Fan Hu, and Wei Li. Histone deacetylase inhibitor valproic acid promotes the induction of pluripotency in mouse fibroblasts by suppressing reprogramming-induced senescence stress. *Experimental Cell Research*, 337(1):61–67, September 2015. ISSN 1090-2422. doi: 10.1016/j.yexcr.2015.06.003.

[124] Xi Chen, Yingying Zhai, Dehai Yu, Jiuwei Cui, Ji-Fan Hu, and Wei Li. Valproic Acid Enhances iPSC Induction From Human Bone Marrow-Derived Cells Through the Suppression of Reprogramming-Induced Senescence. *Journal of Cellular Physiology*, 231(8):1719–1727, August 2016. ISSN 1097-4652. doi: 10.1002/jcp.25270.

[125] Kohta Miyawaki, Hiromi Iwasaki, Takashi Jiromaru, Hirotake Kusumoto, Ayano Yurino, Takeshi Sugio, Yasufumi Uehara, Jun Odawara, Shinya Daitoku, Yuya Kunisaki, Yasuo Mori, Yojiro Arinobu, Hirofumi Tsuzuki, Yoshikane Kikushige, Tadafumi Iino, Koji Kato, Katsuto Takenaka, Toshihiro Miyamoto, Takahiro Maeda, and Koichi Akashi. Identification of unipotent megakaryocyte progenitors in human hematopoiesis. *Blood*, 129(25):3332–3343, June 2017. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2016-09-741611. URL `http://www.bloodjournal.org/content/129/25/3332`.

[126] Anthony E. Boitano, Jian Wang, Russell Romeo, Laure C. Bouchez, Albert E. Parker, Sue E. Sutton, John R. Walker, Colin A. Flaveny, Gary H. Perdew, Michael S. Denison, Peter G. Schultz, and Michael P. Cooke. Aryl hydrocarbon receptor antagonists

promote the expansion of human hematopoietic stem cells. *Science (New York, N.Y.)*, 329(5997):1345–1348, September 2010. ISSN 1095-9203. doi: 10.1126/science.1191536.

[127] Iman Fares, Jalila Chagraoui, Yves Gareau, Stphane Gingras, Rjean Ruel, Nadine Mayotte, Elizabeth Csaszar, David J. H. F. Knapp, Paul Miller, Mor Ngom, Suzan Imren, Denis-Claude Roy, Kori L. Watts, Hans-Peter Kiem, Robert Herrington, Norman N. Iscove, R. Keith Humphries, Connie J. Eaves, Sandra Cohen, Anne Marinier, Peter W. Zandstra, and Guy Sauvageau. Pyrimidoindole derivatives are agonists of human hematopoietic stem cell self-renewal. *Science*, 345(6203):1509–1512, September 2014. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1256337. URL `http://www.sciencemag.org/content/345/6203/1509`.

[128] Andres F. Martinez, Richard D. McMahon, Marc Horner, and William M. Miller. A uniform-shear rate microfluidic bioreactor for real-time study of proplatelet formation and rapidly-released platelets. *Biotechnology Progress*, pages n/a–n/a. ISSN 1520-6033. doi: 10.1002/btpr.2563. URL `http://onlinelibrary.wiley.com/doi/10.1002/btpr.2563/abstract`.

[129] Antoine Blin, Anne Le Goff, Aurlie Magniez, Sonia Poirault-Chassac, Bruno Teste, Graldine Sicot, Kim Anh Nguyen, Feriel S. Hamdi, Mathilde Reyssat, and Dominique Baruch. Microfluidic model of the platelet-generating organ: beyond bone marrow biomimetics. *Scientific Reports*, 6:21700, February 2016. ISSN 2045-2322. doi: 10.1038/srep21700. URL `https://www.nature.com/articles/srep21700`.

[130] Jonathan N. Thon, Brad J. Dykstra, and Lea M. Beaulieu. Platelet bioreactor: accelerated evolution of design and manufacture. *Platelets*, 28(5):472–477, July 2017. ISSN 0953-7104. doi: 10.1080/09537104.2016.1265922. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5507711/`.

[131] Sean C. Bendall, Erin F. Simonds, Peng Qiu, El-ad D. Amir, Peter O. Krutzik, Rachel Finck, Robert V. Bruggner, Rachel Melamed, Angelica Trejo, Olga I. Ornatsky, Robert S. Balderas, Sylvia K. Plevritis, Karen Sachs, Dana Peer, Scott D. Tanner, and Garry P. Nolan. Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science*, 332(6030): 687–696, May 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1198704. URL `http://science.sciencemag.org/content/332/6030/687`.

[132] Xiaoyi Chen, Milena Hasan, Valentina Libri, Alejandra Urrutia, Benot Beitz, Vincent Rouilly, Darragh Duffy, tienne Patin, Bernard Chalmond, Lars Rogge, Lluis Quintana-Murci, Matthew L. Albert, Benno Schwikowski, and Milieu Intrieur Consortium. Automated flow cytometric analysis across large numbers of samples and cell types. *Clinical Immunology (Orlando, Fla.)*, 157(2):249–260, April 2015. ISSN 1521-7035. doi: 10.1016/j.clim.2014.12.009.

[133] Robert V. Bruggner, Bernd Bodenmiller, David L. Dill, Robert J. Tibshirani, and Garry P. Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, 111(26):E2770–E2777, July 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1408792111. URL `http://www.pnas.org/content/111/26/E2770`.

[134] Lin Lin, Greg Finak, Kevin Ushey, Chetan Seshadri, Thomas R. Hawn, Nicole Frahm, Thomas J. Scriba, Hassan Mahomed, Willem Hanekom, Pierre-Alexandre Bart, Giuseppe Pantaleo, Georgia D. Tomaras, Supachai Rerks-Ngarm, Jaranit Kaewkungwal, Sorachai Nitayaphan, Punnee Pitisuttithum, Nelson L. Michael, Jerome H. Kim, Merlin L. Robb, Robert J. OConnell, Nicos Karasavvas, Peter Gilbert, Stephen DeRosa, M. Juliana McElrath, and Raphael Gottardo. COMPASS identifies T-cell

subsets correlated with clinical outcomes. *Nature biotechnology*, 33(6):610–616, June 2015. ISSN 1087-0156. doi: 10.1038/nbt.3187. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4569006/.

[135] Holden T. Maecker, J. Philip McCoy, and Robert Nussenblatt. Standardizing immunophenotyping for the Human Immunology Project. *Nature Reviews. Immunology*, 12(3):191–200, February 2012. ISSN 1474-1741. doi: 10.1038/nri3158.

[136] Gisela Pachn, Isabel Caragol, and Jordi Petriz. Subjectivity and flow cytometric variability. *Nature Reviews Immunology*, 12(5):396, May 2012. ISSN 1474-1741. doi: 10.1038/nri3158-c1. URL https://www.nature.com/articles/nri3158-c1.

[137] Iftekhar Naim, Suprakash Datta, Jonathan Rebhahn, James S. Cavenaugh, Tim R. Mosmann, and Gaurav Sharma. SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 1: algorithm design. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology*, 85(5):408–421, May 2014. ISSN 1552-4930. doi: 10.1002/cyto.a.22446.

[138] Peng Qiu, Erin F. Simonds, Sean C. Bendall, Kenneth D. Gibbs, Robert V. Bruggner, Michael D. Linderman, Karen Sachs, Garry P. Nolan, and Sylvia K. Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature Biotechnology*, 29(10):886–891, October 2011. ISSN 1546-1696. doi: 10.1038/nbt.1991.

[139] Cariad Chester and Holden T. Maecker. Algorithmic Tools for Mining High-Dimensional Cytometry Data. *The Journal of Immunology*, 195(3):773–779, August 2015. ISSN 0022-1767, 1550-6606. doi: 10.4049/jimmunol.1500633. URL http://www.jimmunol.org/content/195/3/773.

[140] J. Paul Robinson and Mario Roederer. HISTORY OF SCIENCE. Flow cytometry strikes gold. *Science (New York, N.Y.)*, 350(6262):739–740, November 2015. ISSN

1095-9203. doi: 10.1126/science.aad6770.

[141] Habil Zare, Ali Bashashati, Robert Kridel, Nima Aghaeepour, Gholamreza Haffari, Joseph M. Connors, Randy D. Gascoyne, Arvind Gupta, Ryan R. Brinkman, and Andrew P. Weng. Automated analysis of multidimensional flow cytometry data improves diagnostic accuracy between mantle cell lymphoma and small lymphocytic lymphoma. *American Journal of Clinical Pathology*, 137(1):75–85, January 2012. ISSN 1943-7722. doi: 10.1309/AJCPMMLQ67YOMGEW.

[142] C. B. Bagwell and E. G. Adams. Fluorescence spectral overlap compensation for any number of flow cytometry parameters. *Annals of the New York Academy of Sciences*, 677:167–184, March 1993. ISSN 0077-8923.

[143] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 0035-9246. URL `https://www.jstor.org/stable/2984875`.

[144] Greg Finak, Marc Langweiler, Maria Jaimes, Mehrnoush Malek, Jafar Taghiyar, Yael Korin, Khadir Raddassi, Lesley Devine, Gerlinde Obermoser, Marcin L. Pekalski, Nikolas Pontikos, Alain Diaz, Susanne Heck, Federica Villanova, Nadia Terrazzini, Florian Kern, Yu Qian, Rick Stanton, Kui Wang, Aaron Brandes, John Ramey, Nima Aghaeepour, Tim Mosmann, Richard H. Scheuermann, Elaine Reed, Karolina Palucka, Virginia Pascual, Bonnie B. Blomberg, Frank Nestle, Robert B. Nussenblatt, Ryan Remy Brinkman, Raphael Gottardo, Holden Maecker, and J. Philip McCoy. Standardizing Flow Cytometry Immunophenotyping Analysis from the Human ImmunoPhenotyping Consortium. *Scientific Reports*, 6:20686, February 2016. ISSN 2045-2322. doi: 10.1038/srep20686. URL `https://www.nature.com/articles/srep20686`.

[145] Mark F. Pittenger, Alastair M. Mackay, Stephen C. Beck, Rama K. Jaiswal, Robin Douglas, Joseph D. Mosca, Mark A. Moorman, Donald W. Simonetti, Stewart Craig, and Daniel R. Marshak. Multilineage Potential of Adult Human Mesenchymal Stem Cells. *Science*, 284(5411):143–147, April 1999. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.284.5411.143. URL http://science.sciencemag.org/content/284/5411/143.

[146] Sean J Morrison, Patricia M White, Christiane Zock, and David J Anderson. Prospective Identification, Isolation by Flow Cytometry, and In Vivo Self-Renewal of Multipotent Mammalian Neural Crest Stem Cells. *Cell*, 96(5):737–749, March 1999. ISSN 0092-8674. doi: 10.1016/S0092-8674(00)80583-8. URL http://www.sciencedirect.com/science/article/pii/S0092867400805838.

[147] P. Lcio, A. Parreira, MWM van den Beemd, EG van Lochem, ER van Wering, E. Baars, A. Porwit-MacDonald, E. Bjorklund, G. Gaipa, A. Biondi, A. Orfao, G. Janossy, JJM van Dongen, and JF San Miguel. Flow cytometric analysis of normal B cell differentiation: a frame of reference for the detection of minimal residual disease in precursor-B-ALL. *Leukemia*, 13(3):419–427, March 1999. ISSN 1476-5551. doi: 10.1038/sj.leu.2401279. URL https://www.nature.com/articles/2401279.

[148] Andrew Cron, Ccile Gouttefangeas, Jacob Frelinger, Lin Lin, Satwinder K. Singh, Cedrik M. Britten, Marij J. P. Welters, Sjoerd H. van der Burg, Mike West, and Cliburn Chan. Hierarchical Modeling for Rare Event Detection and Cell Subset Alignment across Flow Cytometry Samples. *PLOS Computational Biology*, 9(7):e1003130, July 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003130. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003130.

[149] Kerstin Johnsson, Jonas Wallin, and Magnus Fontes. BayesFlow: latent modeling of flow cytometry cell populations. *BMC Bioinformatics*, 17(1):25, January 2016. ISSN 1471-2105. doi: 10.1186/s12859-015-0862-z. URL `https://doi.org/10.1186/s12859-015-0862-z`.

[150] Saumyadipta Pyne, Xinli Hu, Kui Wang, Elizabeth Rossin, Tsung-I. Lin, Lisa M. Maier, Clare Baecher-Allan, Geoffrey J. McLachlan, Pablo Tamayo, David A. Hafler, Philip L. De Jager, and Jill P. Mesirov. Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences*, 106(21):8519–8524, May 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0903028106. URL `https://www.pnas.org/content/106/21/8519`.

[151] Chiaowen Hsiao, Mengya Liu, Rick Stanton, Monnie McGee, Yu Qian, and Richard H. Scheuermann. Mapping cell populations in flow cytometry data for cross-sample comparison using the Friedman-Rafsky test statistic as a distance measure. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology*, 89(1):71–88, January 2016. ISSN 1552-4930. doi: 10.1002/cyto.a.22735.

[152] Elizabeth Csaszar, Daniel C. Kirouac, Mei Yu, WeiJia Wang, Wenlian Qiao, Michael P. Cooke, Anthony E. Boitano, Caryn Ito, and Peter W. Zandstra. Rapid expansion of human hematopoietic stem cells by automated control of inhibitory feedback signaling. *Cell Stem Cell*, 10(2):218–229, February 2012. ISSN 1875-9777. doi: 10.1016/j.stem.2012.01.003.