Transposable elements and the regulatory logic of hematopoietic differentiation

by

Mohamad Ali Najia
B.S. Biomedical Engineering, Georgia Institute of Technology, 2014

SUBMITTED TO THE HARVARD-MIT PROGRAM IN HEALTH SCIENCES & TECHNOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN MEDICAL ENGINEERING AND MEDICAL PHYSICS
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2023

Signature of the Author……
Harvard-MIT Program in Health Sciences and Technology
November 10, 2022

Certified by……
George Q. Daley, M.D., Ph.D.
Dean of the Faculty of Medicine, Harvard Medical School
Thesis Supervisor

Certified by……
Paul C. Blainey, Ph.D.
Associate Professor of Biological Engineering
Thesis Supervisor

Accepted by……
Collin M. Stultz, M.D., Ph.D.
Director, Harvard-MIT Program in Health Sciences and Technology
Nina T. and Robert H. Rubin Professor in Medical Engineering and Science
Professor of Electrical Engineering and Computer Science

Thesis Committee Members

James J. Collins, Ph.D. (Chair)
Termeer Professor of Medical Engineering and Science
Massachusetts Institute of Technology

George Q. Daley, M.D., Ph.D. (Thesis Supervisor)
Dean of the Faculty of Medicine
Caroline Shields Walker Professor of Medicine
Harvard Medical School

Paul C. Blainey, Ph.D. (Thesis Supervisor)
Associate Professor of Biological Engineering
Massachusetts Institute of Technology

George Church, Ph.D. (Reader)
Robert Winthrop Professor of Genetics
Harvard Medical School

To my family – parents Donna and Toufic, and sister Laila – for their enduring mentorship, encouragement, support and inspiration

Transposable elements and the regulatory logic of hematopoietic differentiation

by

Mohamad Ali Najia

Submitted to the Department of Health Sciences & Technology
on November 10, 2022 in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy in
Medical Engineering and Medical Physics

Abstract

The temporal regulation of gene expression by transcription factors, chromatin modifiers and *cis*-regulatory elements is central to establish cellular identity and function. Understanding this regulatory logic is critical for deriving select cell types *in vitro* for translational applications. The human hematopoietic system has long been a model system and an important source for adoptive cell therapies, yet, our understanding of the regulatory mechanisms that elicit commitment toward distinct hematopoietic lineages is continuously evolving.

In this thesis, I describe several studies on transposable elements (TEs) as natural and engineered sources of regulatory innovation that contribute to, and aid in the investigation of, dynamic cellular processes. Toward this end, I built comprehensive genome-wide enhancer-gene maps spanning the human hematopoietic system and identified that TEs in the human genome contribute to the transcriptional networks regulating lymphoid cells. De-repression of TEs in hematopoietic stem cells, enacted via modulation of TE chromatin silencing machinery, facilitates the development of natural killer (NK) cells during lymphoid differentiation. Specifically, knockout of the H3K9 methyltransferase *EHMT1* or transcriptional co-repressor *TRIM28* induced NK-fated progenitors that ultimately generated NK cells with diverse effector properties. We further leveraged TEs by repurposing the packaging function of the MLV gag polyprotein to create a non-destructive reporter of the transcriptional states of living cells, enabling the measurement of dynamic transcriptional processes. Through engineering and scientific inquiry, I established the utility of TEs as synthetic biology tools, furthering our understanding of hematopoietic lineage decisions and highlighting that modulation of TEs can be enabling for hematopoietic cell engineering.

Thesis Supervisor: George Q. Daley, M.D., Ph.D.
Title: Dean of the Faculty of Medicine, Harvard Medical School

Thesis Supervisor: Paul C. Blainey, Ph.D.
Title: Associate Professor of Biological Engineering, Massachusetts Institute of Technology

Table of Contents

Acknowledgements

When I was an undergraduate student striving to get into graduate school, I perceived that earning a PhD was a symbol of high intellect, hardwork and individual achievement. Upon reflection of my time at MIT, I think a PhD more accurately reflects the manifestation of nurtured curiosity and perseverance enabled by a community of mentors, role models, friends and family. It is because of this community that graduate school has been such a rewarding, exciting and fun period of my life.

First and foremost, I'm incredibly grateful for the mentorship from both of my advisors, Paul and George. Both share a steadfast optimism, enthusiastic embrace for out-of-the-box thinking, and a strong commitment to core values like humility, rigor and kindness. Through their guidance and example, they nurtured my development as a scientist, inspired me to think ambitiously, and gave me the confidence to pursue my scientific, professional and personal goals. They also left their marks in unique ways. Even before entering MIT, George played an outsized role in inspiring my scientific interests. In high school, I still remember the day that one of my science teachers dedicated time during her lecture to discuss a groundbreaking discovery from a research lab in Boston—that lab was George's and the discovery was human induced pluripotent stem cells. The thought that one can convert human skin fibroblasts into embryonic-like stem cells had sparked a scientific curiosity on how cells make fate decisions that has sustained my scientific interests in stem cells to this day. Sometimes after long and grueling days in the lab, I pause to reflect on how fortunate I am to be able to work in the same research group that catalyzed my scientific passions. Even with all the demanding responsibilities that come with being Dean, George still remains dedicated to his trainees. I will always remember his homemade risotto that he traditionally makes during lab retreats, and summer BBQ feasts that he hosted at his home. Simple gestures like making a home cooked meal for his trainees shows his caring and supportive nature. I'm also incredibly grateful for Paul and his unyielding excitement for science. His deep technical expertise and boundless optimism gave me the courage to tackle challenging problems. I'm also so appreciative of his openness to new ideas and collaborations, which fostered a creative lab environment and ultimately made me a better scientist.

I'm also appreciative of the support and mentorship from various other faculty members over the years. To my thesis committee members, George Church and Jim Collins, thank you for sharing your time and expertise. I'm appreciative of the thought provoking questions and enthusiasm for my projects. I distinctly remember the several brainstorming sessions I had with George Church and always leaving those meetings feeling excited about science again. Thank you to Thorsten Schlaeger for the collaborative projects that we worked on over the years, your boundless scientific expertise and unwavering commitment to scientific rigor—I always learned something new with every conversation. Thank you to Rizwan Romee for being an exceptionally helpful collaborator on NK biology, and for not only supporting my science but also in providing guidance on career opportunities. Additionally, my graduate school experience would not be the same without Trista North. I could not be more thankful for all the time she dedicated over years to mentor me, for her optimism, and for an open-door to discuss ideas and data. Through her

infectious laugh, she is able to make any environment she enters feel welcoming. I valued being able to contribute to projects in her lab and for her lab's reciprocity with my projects.

As a co-advised student, I was fortunate to be a member of two complementary research environments. When I first joined the Daley lab, the volume of people and scientific directions made finding a scientific niche a little daunting. But, I'm incredibly grateful to both Deepak and Caroline who selflessly provided me with early guidance and gave me critical feedback on all my crazy, unrealistic ideas. Throughout my time in the lab they have been role models as creative, rigorous and hardworking scientists, and I'm privileged to have learnt so much directly from them. Their thoughtful and quantitative way of thinking, and collaborative nature has been instrumental in advancing my research. To Ran and Rubul, thank you both for your down-to-earth and humble demeanors, the opportunity to have contributed to your projects, and the ease that we can discuss data and ideas with constant interspersed bouts of humor. To former members of the lab, Linda, Grant, Pavlos, John, Dan, Kal, and Jihan, thank you for making the lab such an intellectually stimulating environment and a really fun place to work. I also had the opportunity to work with very talented technicians over the years, particularly, Yang, Trevor, Luca and Arianna. Thank you all for your hard work, willingness to learn, and pursuit of excellence. And a special thank you to Arianna who is such a driven and compassionate person — you have taught me just as much as I've taught you, and we shared many laughs along the way!

I'm equivalently grateful for members of the Blainey Lab. Thank you to Miguel, Russell, Luke, Josh E, Josh P, Avtar, and Becca, for your friendship and continuous support. Your abilities to tackle challenging scientific problems with steadfast resolve has been inspiring. The many fun experiences that we shared outside of the lab like game nights, Friendsgivings and hiking trips contributed to making graduate school so memorable. To Emily, thank you for literally everything: for being a great friend, and carrying the lab on your shoulders. Finally, a big thank you to Anna, an amazing friend and someone who I valued working with in the lab. We became fast friends during your rotation and I think graduate school would have been very different without you. Between our 3 hour long pandemic lockdown phone calls, to the projects we worked on together, you have shaped so many aspects of my graduate school experience. Together with Peter Chen and Emily, we have shared so many laughs and fun experiences that I'm so grateful.

There are many other friends that have been incredibly supportive throughout my time at MIT. Thank you to HST classmates, Shriya and Kriti for always keeping life exciting and high energy. Our friendship was first shaped during the HST interview weekend and only grew while in graduate school as a result of the late-night study sessions, dinners, science projects and travels that we shared together. Thank you to Anu who I also first met on the grad school interview circuit. I really enjoyed working with you to organize a global biotech conference at the Broad and I can always count on you for scientific advice and a good laugh. Thank you also to Petar Todorov, Chris Ghadban, Shivam Shodhan, Chara Koutsioumpa, Mildred Jimenez, Jenna Aronson, and Elise Wilcox for such an enjoyable time working together (every Sunday morning team meeting for over a year) to pull off GS19. To George Chao, thank you for the travel

memories across South Africa as part of the HMS Translational Medicine program and for your continued friendship throughout grad school. Our frequent lunches in NRB were a hallmark of the post-pandemic-lockdown, return-to-normalcy period of the grad school. To Maurizio, thank you my Italian friend and for deeply insightful scientific conversations while also keeping life lighthearted (especially over one of your exceptional home cooked Italian dinners). And thank you Eden: you are such a thoughtful friend that I can talk to about anything, and being in your company melts away the background stresses of graduate school.

Going further back, I owe a huge debt of gratitude to former mentors that helped me along my scientific trajectory. In particular, I'm thankful to Todd McDevitt from Georgia Tech for taking a chance on a first-year undergrad, which ultimately manifested into 4 years in the lab. The experience in his lab further fostered my desire to pursue a PhD. I learnt an invaluable amount on embryonic stem cells and biomanufacturing from direct mentors, Barbara Nsiah, Jenna Wilson and Melissa Kinney. They also committed vast amounts of time to train me in fundamental lab techniques for which I'm incredibly grateful (especially now in my position as a grad student who now knows how much time and patience it takes to train people). I'm also appreciative of Sasha Gimelbrant at DFCI and Michael Eisen at UC Berkeley for welcoming me into their labs for summer research experiences, their invaluable feedback on my early technical writing, and brainstorming sessions for my graduate fellowship proposals. Finally, I'm so thankful for the mentorship that I received from Karen Adams in the Georgia Tech Fellowships Office throughout my undergraduate studies. She single-handedly made me a better writer and has opened up many doors. I'm so appreciative that she still stays in touch and takes interest in my career.

While at Georgia Tech, I made some lifelong friends that have played a huge role in my life. I can't begin to express how grateful I am to Tyler, Meng, Andy, Carlos, Candice, Pranay, Anj and Sai. We all met on the first day of college and have been strong friends ever since. Nearly all of them are rocket scientists and have opened my eyes to new fields. They are responsible for sending rockets, satellites and rovers beyond Earth's grasp and sometimes to new worlds. Their passions and line of work inspires me and definitely puts a failed experiment into perspective. We have shared so many experiences and memories—from global travels to deep philosophical conversations, from weddings to graduations, and from bar hops to just picking up the phone to talk. This PhD would be dramatically different without them. Finally, thank you to Michael Chen and TJ Kaplan. From our time running the undergraduate research journal until now it has been a lot of fun, and I'm grateful for all the support over the years.

Thank you to my high school friends Anisha and Lauren. Although we have all taken different paths, when we reunite, we find ourselves sharing old high school traditions like dinners at Sawasdee or long walks around Danvers. The passion, kindness and generosity that you both exude is so inspiring and I'm grateful for our long-lasting friendship.

Finally, I cannot begin to convey the unconditional love, support and wisdom provided by my family. My parents have been an inspirational and guiding light in my life. Through their example, they instilled a set of values in my sister and me from a young age: work hard, strive

for excellence, approach life with kindness, humility and compassion, and ultimately work to make a difference in people's lives. Simply, I would not be where I am in life without them and I am thankful for their endless supply of thoughtful advice, humor and support. I am grateful for my younger sister, Dr. Laila Najia and the intellectual debates we have about which organ is better: the heart or the blood. The passion she brings to her job as a cardiac pharmacist and the compassion she has for her patients is infectious and the catalyst for me to overcome from setbacks at the bench. To all my aunts, uncles and cousins—I am so appreciative of all your love and genuine interest in my research. Finally, to my late grandparents—despite not being able to see me defend my PhD, your influence and teachings still manifest to this day. Simply, this thesis would not be possible without your sacrifices, compassion, and endless love.

1 Introduction

1.1 Thesis Overview

A remarkable property of hematopoietic stem cells (HSCs) is their ability to integrate complex molecular processes in order to dynamically choose amongst multiple possible fates and generate the diverse compendium of cell types found in human blood [1,3]. Multiple layers of gene regulation underlie hematopoietic specification and differentiation during embryonic development and throughout adulthood. The dynamic regulation of gene expression, elicited through *cis*-regulatory elements, transcription factors and chromatin modifiers is central to hematopoietic fate choice and differentiation. The ability to modulate and control these processes can facilitate translational goals of generating specific hematopoietic cells *in vitro* for adoptive cell therapies [5,7].

Induced pluripotent stem cells (iPSCs) offer a scalable and renewing source to generate somatic cells for adoptive cell therapies. The design of precisely staged protocols to derive hematopoietic cells from iPSCs has been guided largely by understanding hematopoietic development in model organisms. While this approach has been generally successful in recapitulating aspects of hematopoietic development within *in vitro* cultures, it falls short when we have limited knowledge of *in vivo* processes. In particular, we have a limited understanding of the transcriptional and chromatin processes that generate hematopoietic cells from endothelial precursors. This gap in knowledge is reflected in the generation of highly myeloid-biased hematopoietic progenitors with limited lymphoid capacity to generate T cells. Interrogating the regulatory networks that govern hematopoietic specification from iPSCs can be informative for nominating transcriptional regulators that enhance the generation of lymphoid fates.

Following their embryonic formation, HSCs are subject to additional regulatory pressures throughout adulthood to determine their lineage output. Transcription factors and chromatin modifiers orchestrate on *cis*-regulatory elements to drive changes in transcription. Mounting evidence from various cell types highlights that transposable elements (TEs) in the human genome have been co-opted for various regulatory functions, including as transcriptional enhancers. Immune cells appear particularly prone to TE-mediated gene regulation as genome-wide chromatin profiling studies have revealed a significant fraction of TEs contribute to putative enhancers that modulate essential immune-related functions. Furthermore, TEs are strictly regulated by chromatin-based machinery and can thus influence local chromatin environments. Our understanding of how TEs influence the dynamic differentiation processes that generate immune cells from HSCs is unclear, but we hypothesize that they may influence cell fate decisions since they are potent sources of transcription factor and chromatin regulation.

Our ability to understand and manipulate cell fate decisions is also driven by new technologies. High-throughput transcriptional measurements have been indispensable for the molecular characterization of cell states and functions across various fields. The advent of single cell transcriptional profiling technologies further opened new terrains in the study of hematopoietic differentiation and fate choice by aiding in the discovery of new regulators and producing a more nuanced understanding of heterogeneous stem cell populations with variations in differentiation potential. However, a key limitation of these assays is that the act of making the measurement destroys the cell, which hinders insights on the dynamic transcriptional processes that mediate cell fate potential.

This thesis is focused on the studying and manipulating the regulatory processes dictating hematopoietic specification and differentiation with the goal of facilitating the derivation of hematopoietic cells *in vitro* for translational applications. We accomplished this through the development of molecular tools and hypothesis-driven inquiry that spans multiple layers of gene regulation. In particular, we systematically investigated enhancer-gene regulation in the human

hematopoietic system, uncovered a role for TEs in shaping transcriptional networks, and demonstrated that the modulation of these elements through chromatin machinery can influence hematopoietic lineage decisions (**Chapter 2**). We created genetically-encodable technologies to enable non-destructive transcriptional measurements in living cells, which advances the prospect of tracking dynamic transcriptional processes (**Chapter 3**). Finally, we integrated both transcriptional and chromatin profiling to dissect the temporal transitions of hematopoietic specification from iPSCs and identify transcriptional regulators that enhance the production of lymphoid cells *in vitro* (**Chapter 4**). Collectively, these studies advance our understanding of the regulatory logic guiding hematopoietic cell fate decisions, and provide tractable approaches to enhance the *in vitro* production of hematopoietic cells for translational applications.

1.2 Induced pluripotent stem cells as a source for hematopoietic cell therapies

Human iPSCs represent an ideal source for the scalable manufacture of off-the-shelf products for cell therapy. Numerous iPSC differentiation protocols have been developed to generate hematopoietic cells, yet all invariably result in short-lived progenitors with limited lymphoid potential [5,9]. A guiding principle in the field has been to recapitulate embryonic development to direct iPSCs under defined conditions with specific signaling pathway agonists and antagonists to specific hematopoietic fates. In this approach, cytokines, morphogens and/or small molecules are used to modulate signaling pathways in iPSCs-derived embryoid bodies (EBs). A common pattern in these protocols is the use of Wnt agonists and BMP4 to induce mesoderm specification, followed by VEGF to promote angiogenesis, and subsequently hematopoietic cytokine cocktails to drive hematopoiesis [12,116].

Recapitulating embryonic hematopoietic development *in vitro* remains a hurdle, due in large part to our incomplete understanding of the stage-specific regulators of hematopoietic development. *In vivo* blood development occurs in at least two waves, first marked by a primitive wave that takes place in the extraembryonic yolk sac and generates mostly myeloid

cells and nucleated erythrocytes. A definitive wave of intraembryonic hematopoiesis then follows, which creates HSCs and progenitors that last into adulthood and produce cells with adult-like characteristics[14]. In both cases, however, hematopoietic cells are generated from endothelial precursors, termed hemogenic endothelium (HE). The existence of different HE populations highlights one of the challenges associated with specifying iPSC-derived hematopoietic programs *in vitro*, as it is currently impossible to distinguish these populations on the basis of surface marker expression alone. An alternative approach to distinguishing the stages of hematopoiesis, which can be exploited to guide *in vitro* hPSC differentiation towards specific lineages, is the identification of transcription factors and chromatin modifiers that specifically regulate the different hematopoietic programs. Therefore, the interrogation of the chromatin and transcriptional processes underlying hematopoietic specification could generate hypotheses on specific sets of transcriptional regulators to modulate during iPSC differentiation.

1.3 Transposable elements as contributors to gene regulatory networks

Nearly half of the human genome is composed of transposable elements (TEs), which are increasingly being recognized not just as parasitic DNA, but as an important source of gene regulation [16,18,20,22]. In particular, endogenous retroviruses (ERVs), which comprise about 8% of the human genome, are sequences derived from ancient retroviruses whose germ-line infections have persisted through millions of years of evolution [24,26]. ERVs, like all retroviruses, contain 5' and 3' long terminal repeats (LTRs) that flank open reading frames encoding retroviral proteins. Over time, these LTRs accumulate mutations and often undergo homologous recombination [28,30,32]. In their capacity as retroviral promoters, LTRs are enriched for transcription factor motifs and thus are a particularly suitable substrate for evolving new regulatory elements that can be utilized for host gene regulation. For example, in the mouse two-cell stage embryo, MERVL elements serve as alternative promoters for a subset of mouse genes[34], while LTRs of a human ERV, MER41, can function as interferon-inducible enhancers[36].

Epigenomic mapping studies detected cell type-selective active enhancer signatures at thousands of TEs, suggesting that acquisition of tissue-specific or inducible regulatory functions by these elements is a widespread phenomenon that may have profound effects on host gene regulatory networks [38,40,42,44,46,48,50,52,54] . Furthermore, emerging evidence suggests that a large proportion of primate-specific *cis*-regulatory elements, as well as those that changed their activity most recently, since the separation of humans from chimpanzees, originate from TEs [20,54,56], highlighting how TEs have contributed to primate lineage and reshaped the human transcriptional landscape.

1.4 Cellular mechanisms regulating transposable elements

While TEs are critical to mammalian physiology, unmoderated transposition can compromise cellular health [58,60]. Thus, various cellular mechanisms exist to safeguard cellular homeostasis[62]. Histone modifications directly impact chromatin structure and the access to DNA binding proteins, which in turn affects transcription. H3K9 is the most common repressive histone mark observed at TE elements [44,64]. Histone methyltransferases including SETDB1, SUV39H1, SUV39H2, EHMT1 and EHMT2 mediate retrotransposon repression through deposition of H3K9 methylation [66,68,70,72,74,76,78,80]. Upon binding to H3K9 methylated nucleosomes, cofactor HP1 undergoes a switch from an auto-inhibitory state to a spreading competent state enabling the radial spread of H3K9 methylation and compaction of chromatin [82,84].

Additionally, another well known TE silencing mechanism is the KRAB zinc finger (KZFP) silencing pathway [86,88,90]. To initiate ERV repression, KZFPs must first bind to their DNA targets through their arrays of C2H2-type zinc finger domains. Once bound, the KRAB domain of the KZFP recruits the transcriptional co-repressor, TRIM28, NuRD complexes and SETDB1, which methylates proximal H3K9 residues to silence the transcription of KZFP-bound TE loci [88,92,94]. Although KZFP/TRIM28 repression is highly conserved between species, TEs are often species-specific. This observation is in line with the finding that KZFP repertoires are species-

specific and many have undergone positive selection in parallel with the rise of new classes of ERVs in the genome [92,96,98,100,102,104] .

1.5 Immune sensing of transposable elements

While TEs are largely repressed by chromatin silencing machinery, their reactivation and generation of retroviral nucleic acids and proteins has been documented during normal physiological processes, such as embryonic development[106] and specification of HSCs[108]. Given the exaptation of ERVs over millions of years, one hypothesis is that host cells should have developed immune tolerance to ERV products[110]. However, ERV-induced innate immune activation in cancer cells as well as in immune cells suggests that these elements are still recognized as foreign entities through innate immune sensing pathways [112,114,117]. ERVs can trigger an immune response by producing nucleic acids that resemble pathogen-associated molecular patterns (PAMPs), thereby triggering activation of pattern recognition receptors (PRRs) and ultimately the production of pro-inflammatory cytokines, chemokines and type I interferons (IFN-α and IFN-β)[118].

Toll-like receptors (TLRs) are one such class of PRRs that are transmembrane proteins characterized by different subcellular localization and cognate ligands[120]. TLRs 3, 7, 8, and 9 are localized to the endosomal lumen and bind nucleic acid ligands. TLR3 senses dsRNA, TLR7 and 8 detect ssRNA, and TLR9 recognises unmethylated CpG DNA and RNA:DNA hybrids [122,124]. Upon ligand binding, signaling from the TLRs involves dimerization and the recruitment of MyD88[120], which then activates a signaling cascade resulting in the nuclear translocation of transcription factors such as NF-κB and IRF 3 and 7[126] and in turn, activates the expression of proinflammatory cytokines and type I IFNs.

In addition to the TLRs, cytosolic RNAs are sensed by a series of PRRs, including RIG-I and MDA5 [91,128,131]. These belong to the RIG-I-like receptor (RLR) family, a class of RNA helicases that recognise viral RNA in the cytoplasm and induce the production of type I IFNs[122].

RIG-I senses short dsRNA and ssRNA, while MDA5 detects long dsRNA molecules[133]. Following detection, RIG-I and MDA5 signal through interaction with MAVS on the mitochondrial outer membrane[122], leading to the phosphorylation and nuclear translocation of IRF 3 and 7 to induce IFN expression along with other antiviral factors[135]. Treatment of cancer cells with DNA methyltransferase inhibitor, 5-Aza revealed that the antitumor effect is mediated through viral mimicry and TLR3 and MDA5/MAVS sensing of dsRNA derived from ERVs [112,114,117]. Sensing of TE-derived RNA by MAVS has also been documented hematopoietic regeneration, enabling HSCs to mount an inflammatory response necessary for their exit from quiescence [131].

Cells have also evolved a collection of sensors for cytosolic DNA PAMPs. cGAS is the best characterized DNA sensor to date and it plays a major role in cytosolic DNA sensing, whereby DNA sensing by cGAS results in the production of cGAMP, a secondary messenger molecule, which then activates STING and IRF3 for IFN induction [137,139]. Importantly, cGAS has been demonstrated to detect DNA from retroviruses[141]. The role of ERV DNA in stimulating the innate immune response first came from a study in mice investigating the 3' repair exonuclease 1 (TREX1)[143]. TREX1 is upregulated by the expression of type I IFN and NF-κB and it plays a role in regulating endogenous DNA PAMPs and DAMPs[145]. TREX1 knockout mice induce an IRF3-dependent immune response to cytosolic DNA, which is thought to include ERV cDNA, in a cGAS- and STING- dependent manner[147]. These observations suggest that some ERVs have the capability to get reverse transcribed into cDNA and stimulate host innate immunity.

During inflammation, IFN secretion can create a positive-feedback loop whereby the transcription factors that get activated can bind to and activate the transcription of more ERV loci. Type I IFNs stimulate natural killer cells and clonally expand cytotoxic T cells, linking innate with adaptive immunity[148]. *In silico* studies of the HERV-K 5'LTR has led to identification of multiple binding sites for pro-inflammatory transcription factors such as NF-κB[149]. NF-κB is commonly activated during an inflammatory response; its binding to the HERV LTR can potentially upregulate HERV transcription. Indeed, this has been demonstrated using LPS and

TNF-α treatments, in which NF-κB induction results in the increased expression of HERV-K, HERV-H, and HERV-W mRNAS[150]. Such interaction between TEs and inflammatory transcription factors provides a mechanism through which TEs can initiate/sustain inflammation.

1.6 Retroviral elements as tools for synthetic biology

A sizable fraction of TEs in the human genome consist of ERVs that have integrated into mammalian genomes throughout evolution [26,151–154]. While most ERVs have lost their original functions over the course of evolution, some retroelements have been co-opted for diverse roles in normal mammalian physiology. The gene *Arc* encodes a homolog of a retroviral core structural gene, *gag,* which forms virus-like particles (VLPs) to package and transfer its own mRNA between neural synapses [155–158]. Similarly, PEG10, an ERV-derived protein can package its own mRNA in VLPs and is involved in the formation of the placenta [159–161]. VLPs structurally resemble retroviruses, but lack viral genetic material and are consequently replication incompetent. The domestication of these retroelements for normal physiological functions underscores the opportunity of these elements for nucleic acid transfer.

VLP formation is minimally dependent on the core structural polyprotein, Gag [162–164]. VLPs assemble by associating a few thousand Gag proteins, which accumulate at the cytoplasmic side of the cell membrane through myristoylation of the N-terminal matrix (MA) protein. The subunits capsid (CA) and nucleocapsid (NC), and in murine leukemia virus (MLV) also p12, complete the Gag polyprotein. Gag self-assembly in cells is at least in part due to interactions between CA domains, whereas the NC domain is responsible for binding the viral genome [165–167]. The multimerized Gag proteins then bud from the plasma membrane and are secreted from the cell [168].

The minimal genetic requirements for VLP formation have motivated various studies to utilize retroviral components for bioengineering applications. In particular, the Gag polyprotein is

amenable for C-terminal fusions and therefore conducive for the packaging of proteins of interest into VLPs. Several studies have leveraged this property to deliver protein products to cells [162,163]. Additionally, Gag fusions to CRISPR/Cas9 genome editing agents enable the delivery of ribonucleoprotein complexes to cells for installing desired genomic alterations [63,65,169]. Outside of delivery applications, gag fusion proteins have also been utilized to readout protein-protein interactions within living cells [67]. Finally, endogenous retroelements such as PEG10 have been genetically modified to package and deliver select mRNA transcripts to cells, complementing existing approaches to tailor retroelements for delivery applications [170]. These studies ultimately highlight the versatility of retroelements and retroviral machinery for synthetic biology applications.

2 Transposable elements direct hematopoietic lineage decisions

Authors: Mohamad Ali Najia*, Deepak Jha*, Alaa Ali, Cheng Zhang, Arianna Markel, Katie Frenis, Ran Jing, Luca Hensch, Hu Li, Grant Rowe, Rizwan Romee, Trista North, Paul C. Blainey, George Q. Daley

Status: This manuscript is in preparation for submission to a scientific journal

Contributions: M.N., D.J. and G.Q.D. designed and conceived the study; M.N. performed all experiments with assistance from A.A., A.M., K.F., R.J. and L.H.; M.N., D.J., and C.Z. performed computational analysis; M.N. wrote the manuscript with input from all authors. H.L., G.R., R.R., T.N., P.C.B, and G.Q.D. supervised the research and provided funding.

2.1 INTRODUCTION

The regulation of gene expression is fundamental for the establishment of cellular identity and function. The coordinated activities of transcription factors (TFs) and chromatin factors on a DNA template orchestrate transcriptional networks that drive cellular phenotypes and govern cell fate decisions. The hematopoietic system is an archetypical example of the necessity of these mechanisms to regulate the dynamic differentiation processes that generate diverse, mature blood cell types from multipotent hematopoietic stem and progenitor cells (HSPCs) [171,172]. Yet, our understanding of how TFs and chromatin factors elicit commitment toward distinct hematopoietic lineages is continuously evolving. Insights into the regulatory architecture governing hematopoietic differentiation can facilitate novel approaches to derive specific blood cells *in vitro* for therapeutic applications.

Transposable elements in the human genome have garnered increased attention due to mounting evidence that evolution has co-opted them as transcriptional regulators in specific cellular contexts [173]. These ancestral elements, comprising retrotransposons and DNA

transposons, can serve as binding sites for TFs [38,40,174] and participate as transcriptional enhancers [16,175]. While TE-derived regulatory activity can be observed across multiple human tissues, TEs comprise a higher proportion of enhancer states in the hematopoietic lineage [176], suggesting that TEs are particularly important for hematopoietic transcriptional regulatory networks.

The regulatory contributions of TEs are tightly controlled by chromatin machinery. TRIM28, a well-documented suppressor of TEs, is recruited to specific genomic sites via direct interactions with KRAB-zinc finger proteins (KZFPs) [104,177]. Together this complex achieves TE silencing through deposition of H3K9 methylation and removal of histone acetylation via the NuRD complex [68,86,90]. Formation of heterochromatin more generally via H3K9 methyltransferases also functions to silence TE families [66,74,94,178]. Importantly, these processes can contribute to the dynamic activity of TEs during cell state transitions, such as somatic cell reprogramming and early embryogenesis [106,179,180].

Increasing evidence suggests that certain cell types are prone to TE-mediated gene regulation, particularly immune cells [36,181,182]. Genome-wide chromatin profiling studies in immune cells have revealed a significant fraction of TEs contribute to putative enhancers that modulate essential immune-related functions, such as interferon-inducible inflammasome activation [36,46,181]. Furthermore, endogenous retrovirus (ERV) proteins have been co-opted in immune cells to serve a dominant negative function, interfering with viral infection [26,183,184]. Speculation that TEs prompted the evolution of inflammatory gene regulatory networks, led us to ask whether TEs are involved in the dynamic differentiation processes that generate immune cells from HSPCs. We lack an understanding of the potential role of TEs in hematopoietic differentiation and lineage determination.

In this study, we systematically dissected the contributions of TEs to human hematopoiesis. We built a comprehensive cell type-resolved atlas of enhancer-gene regulation, comprising every major cell type in the human hematopoietic system. We identified lineage and

cell type-specific enrichments of TE families in putative enhancers. TEs were particularly wired to the regulatory landscape of lymphoid cells, serving as docking sites for critical TFs and exhibiting dynamic expression during lymphoid differentiation. TE derepression, achieved by modulation of regulators of heterochromatin formation within HSPCs, resulted in the surprising acquisition of natural killer (NK) cell fates in T or B cell supportive differentiation conditions. Specifically, knockout of the transcriptional co-repressor *TRIM28* or the H3K9 methyltransferase *EHMT1* generated distinct lymphoid progenitor populations with enriched activity for NK-relevant TFs. Further, NK cells derived as a result of knockout of *TRIM28* or *EHMT1* exhibited distinct classes of derepressed TEs and downstream effector properties. These findings deepen our understanding of the essential role of TE regulation during hematopoietic differentiation, and enable novel approaches to derive diverse sets of NK cells *ex vivo* for potential therapeutic applications.

2.2 RESULTS

2.2.1 Genome-wide framework for enhancer-gene inference in the hematopoietic system

To systematically investigate the regulatory contributions of TEs to human hematopoietic differentiation (**Figure 2.1A**), we first sought to build robust reference maps of enhancer-gene regulation. Although hematopoiesis has been the subject of numerous profiling efforts over decades to map *cis*-regulatory elements, chromatin states and transcription, less attention has been paid to the enumeration of enhancers and the genes they regulate within primary cells spanning the entirety of the human hematopoietic system. The Activity-by-Contact (ABC) model [185,186], developed to infer enhancer-gene regulation, robustly predicts CRISPRi perturbation experiments, and thus, effectively identifies functional enhancers. The model posits that an enhancer's relative contribution to gene transcription is dependent on its activity (measured by chromatin accessibility and H3K27ac ChIP-Seq) and contact frequency with the gene's promoter (measured by HiC or a power-law genomic distance-power law function). We

extended the applicability of the ABC model by defining enhancer activity solely in terms of chromatin accessibility, which consequently enables regulatory predictions for low-abundance cell types samples, where H3K27ac ChIP-Seq is experimentally infeasible, such as hematopoietic stem cells (HSCs). When benchmarked against experimental CRISPRi-FlowFISH screens in various hematopoietic cell types [185,186], we found that this modified model resulted in near comparable performance in identifying enhancer-gene pairs, as assessed by precision recall statistics (**Supplementary Figure 2.1A-C**). Additionally, we compared enhancer predictions from each of these hematopoietic cell lines to chromHMM-defined regulatory regions and observed that predicted enhancers were strongly enriched within active chromatin (**Supplementary Figure 2.1D**). The modified model predicted experimentally validated enhancer-gene interactions within the relevant cell type, including intronic enhancers regulating *RUNX1* [187] and *GATA2* [188] within HSCs (**Supplementary Figure 2.1E-F**), enhancers regulating *BCL11A* [189] in erythroid precursors (**Supplementary Figure 2.1G**), and enhancers regulating *CD9* [190] in megakaryocytes (**Supplementary Figure 2.1H**). Finally, the ABC model accurately identified known examples of TE-derived enhancers, specifically the MER41G enhancer that regulates *APOL1* within innate immune cells upon interferon-gamma stimulation [36] (**Supplementary Figure 2.1I**). These results indicate that our modified ABC framework provides a tractable and scalable approach to identify putative enhancers in the human hematopoietic system and the potential contributions of TEs to cell-type specific gene regulation.

2.2.2 A comprehensive atlas of human hematopoietic enhancer-gene regulation

We utilized our modified ABC model to build a compendium of genome-wide enhancer-gene maps for all major cell types and states in the human hematopoietic system. We curated publicly available chromatin accessibility data on primary human samples, spanning HSCs and hematopoietic progenitors [191] to mature, differentiated progeny in the myeloid [192–194], erythroid [195] and lymphoid [196] lineages (**Table S1**). In total, our dataset amounted to 258 samples,

comprising 65 different hematopoietic cell types or states. Each hematopoietic cell type was represented by at least 3 different donors in nearly every instance. We uniformly processed all the chromatin accessibility data (see **Methods**) and all samples exhibited high enrichment of signal over background (**Supplementary Figure 2.2A**, **Table S2**).

We applied our modified ABC model to generate genome-wide enhancer-gene maps for each of the 65 hematopoietic cell types in our dataset (see **Methods**). We identified a total of 3,793,020 enhancer-gene links representing a comprehensive resource of regulatory logic governing the human hematopoietic system (**Table S3**). Across all cell types, a gene was predicted to be regulated by 3.22±0.14 ABC enhancers and an ABC enhancer was predicted to regulate 2.35±0.12 genes (**Table S4**). Within a given cell type, ABC enhancers constituted on average only 11.2% of accessible chromatin regions (**Supplementary Figure 2.2B**), underscoring the specific nature of these genome-wide predictions. Furthermore, this dataset amounted to 207,648 unique ABC enhancers across all hematopoietic cell types, which we refer to as a pan-hematopoiesis ABC enhancer peakset. To facilitate subsequent analysis, we built a matrix of the quantitative degree of chromatin accessibility over the pan-hematopoiesis ABC enhancer peakset for all 258 samples. Chromatin accessibility signal in ABC enhancers was

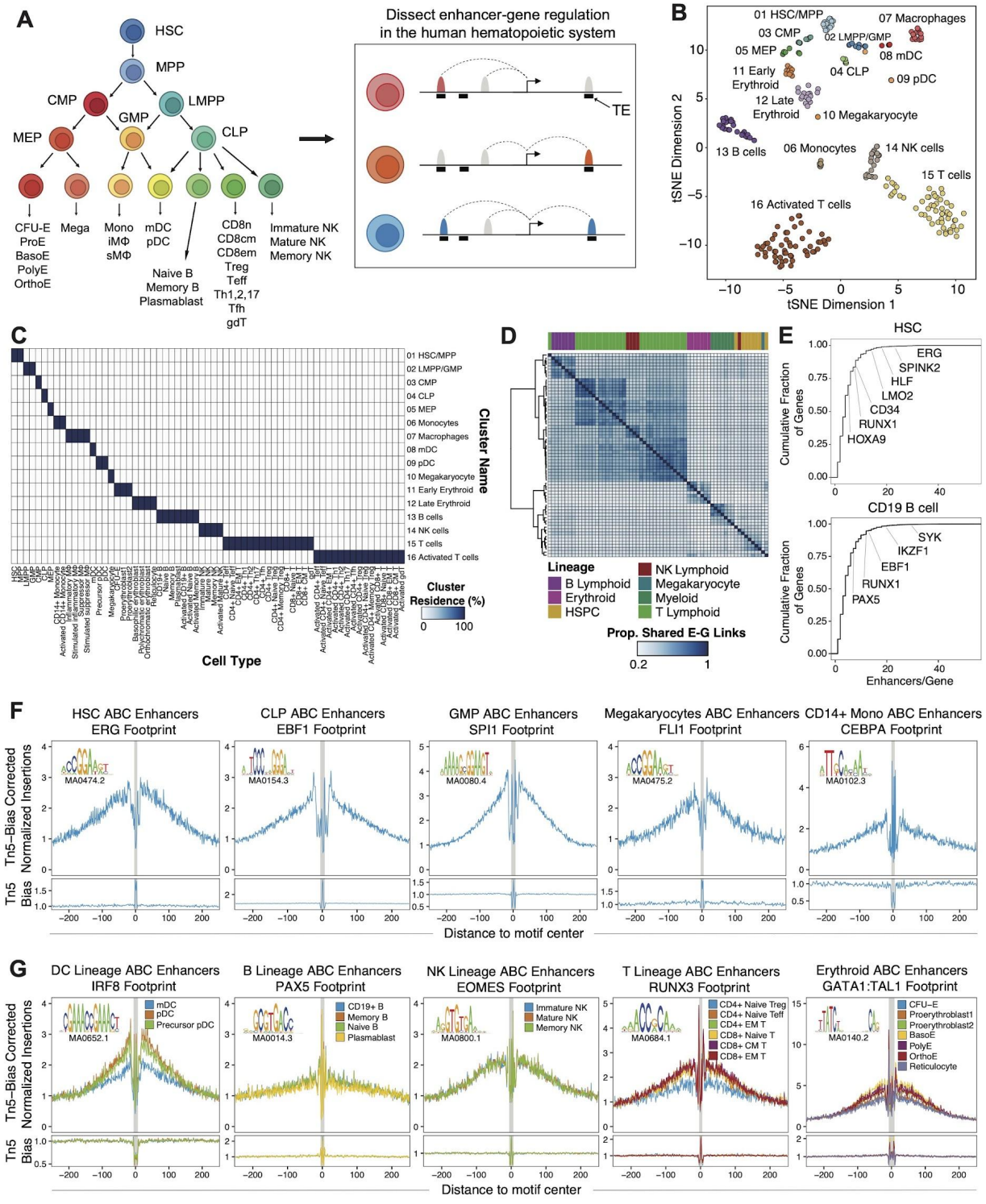


Figure 2.1. A comprehensive atlas of enhancer-gene regulation throughout the human hematopoietic system

(A) Schematic overview of the approach to generate enhancer-gene maps of the human hematopoietic system to dissect TE contributions to gene regulation. HSC = hematopoietic stem cell; MPP = multipotent progenitor; CMP = common myeloid progenitor; LMPP = lymphoid-primed multipotent progenitor; MEP = megakaryocyte-erythrocyte

progenitor; GMP = granulocyte-monocyte progenitor; CLP = common lymphoid progenitor; CFU-E = erythroid colony forming unit; ProE = proerythroblast; BasoE = basophilic erythroblast; PolyE = polychromatic erythroblast; OrthoE = orthochromatic erythroblast; Mega = megakaryocyte; Mono = monocyte; iMΦ = inflammatory macrophage; sMΦ = suppressor macrophage; mDC = myeloid dendritic cell; pDC = plasmacytoid dendritic cell; CD8n = naive CD8+ T cell; CD8cm = central memory CD8+ T cell; CD8em = effector memory CD8+ T cell; Treg = regulatory T cell; Teff = CD4+ T effector cell; Th1 = CD4+ T helper 1 cell; Th2 = CD4+ T helper 2 cell; Th17 = CD4+ T helper 17 cell; Tfh = CD4+ T follicular helper cell; gdT = γδT cell.
(B) Unsupervised t-SNE on the top 50 principal components for the 120,000 most variably accessible ABC enhancers across all hematopoietic cell types. Each dot represents a primary hematopoietic sample (see **Table S2**) and the colors represent clusters identified by density clustering.
(C) Cluster residence heatmap showing the percent of each FACS-identified hematopoietic cell type that resides within each of the 16 annotated clusters.
(D) Heatmap visualizing the proportion of enhancer-gene connections shared across all profiled hematopoietic cell types. An enhancer-gene link is considered to be shared between two cell types if the predicted gene is the same and the ABC enhancers overlap. The rows and columns are hierarchically ordered the same, and the lineage identity of the cell types is noted atop.
(E) The number of enhancer connections per gene for all genes with ABC predictions in HSCs (top) and CD19+ B cells (bottom). Cell-type specific regulators are noted within the plot.
(F-G) Transcription factor footprinting specifically within ABC enhancers of the noted cell types (F) or lineages (G). The sequence logo and JASPAR identifier of the transcription factor motif utilized for the footprinting analysis is noted as an inset on each plot. The Tn5 insertion bias tracks for each motif is shown below each footprint plot.

highly reproducible across technical and donor replicates for each hematopoietic cell type (**Supplementary Figure 2.2C**).

To visualize global patterns from our ABC predictions, we utilized *t*-SNE (**Figure 2.1B**) and density clustering to identify 16 distinct clusters based on chromatin accessibility within the pan-hematopoiesis ABC enhancer peakset (**Supplementary Figure 2.2D**). Accessibility of ABC enhancers alone was sufficient to delineate the broad set of hematopoietic cell types in our dataset, reflective of their cell-type specificity. We also observed concordance between the FACS-sorted, immunophenotypic identity of the samples and the unbiased determined clusters (**Figure 2.1C**). In further agreement with the cell type specificity of ABC enhancers, enhancer-gene links were distinct across cell types, yet share a higher fraction of links within a given lineage (**Figure 2.1D**). This observation suggests that lineage-level gene regulation is more finely tuned by ABC enhancer accessibility than by different sets of enhancer-gene connections. We further investigated the predicted enhancer-gene links and, notably, found that genes with complex enhancer landscapes were canonical regulators of cell identity, such as *ERG, LMO2, HOXA9* and *RUNX1* within HSCs and *SYK, IKZF1, EBF1,* and *PAX5* in CD19+ B cells (**Figure**

2.1E). More generally, ABC-linked genes were enriched for canonical biological processes for the given cell type (**Supplementary Figure 2.2E**). Finally, we used our ABC maps to infer potential activity of *trans*-factors in ABC enhancers. We performed TF footprinting specifically within a cell type's ABC enhancers using ATAC-Seq data and observed motif footprints for canonical TFs that regulate cellular identity (**Figure 2.1F**). TFs also displayed dynamic activity within ABC enhancers across cell types belonging to the same hematopoietic lineage (**Figure 2.1G**), further supporting that ABC predictions are reflective of transcriptional enhancers. Overall, our ABC maps provide a rich, informative and comprehensive resource to dissect enhancer-gene regulation in the human hematopoietic system.

2.2.3 TE families are enriched in ABC enhancers and encode for cell-type specific transcriptional regulators

Having demonstrated the utility of our ABC maps, we sought to systematically investigate how TEs contribute to hematopoietic gene regulation. We utilized the Repbase TE database and segregated the annotated TEs in the human genome by family-level classification. We then intersected the genomic coordinates of these TE families with ABC enhancers from each of the hematopoietic cell types in our dataset and identified 51 TE families that were significantly enriched in ABC enhancers (**Figure 2.2A**, **Table S5**, **Methods**). An overwhelming fraction of the significantly enriched TE families consisted of MER and LTR elements (43/51). Notably, TE families exhibited cell type and state-specific enrichments. In particular, LTR10A/F elements were selectively enriched within activated CD4 and CD8 T cell subsets. The specificity and degree of these enrichments prompted us to further investigate the basis of TE co-option in lymphoid cells. ATAC-Seq signal over LTR10A/F-containing ABC enhancers was pronounced only in activated states and generally low-to-inaccessible in resting states (**Figure 2.2B**), confirming our prior enrichment result and suggesting that these elements may be involved in

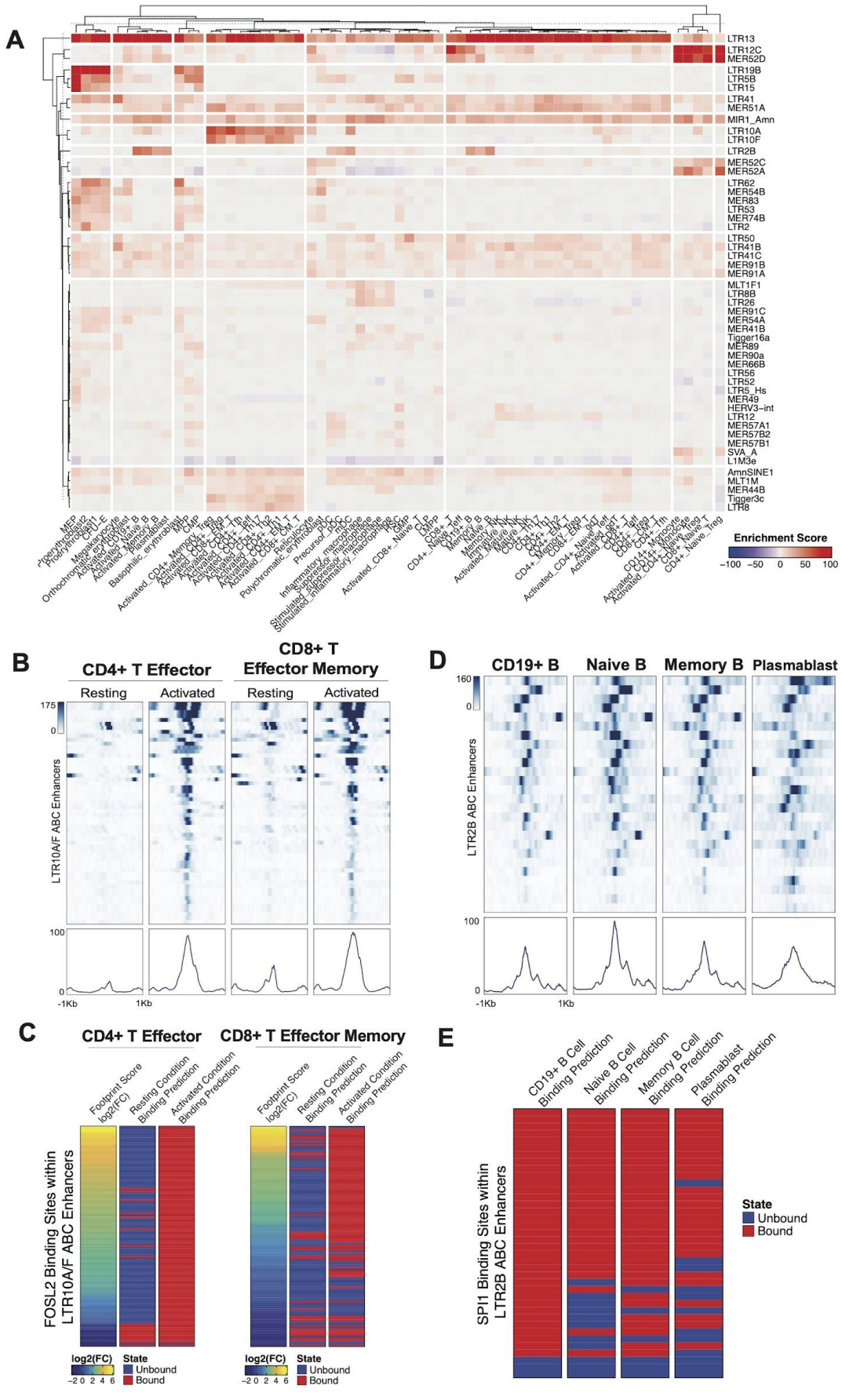

A
LTR13
LTR12C
MER52D
LTR19B
LTR5B
LTR15
LTR41
MER51A
MIR1_Amn
LTR10A
LTR10F
LTR2B
MER52C
MER52A
LTR62
MER54B
MER83
LTR53
MER74B
LTR2
LTR50
LTR41B
LTR41C
MER91B
MER91A
MLT1F1
LTR8B
LTR26
MER91C
MER54A
MER41B
Tigger16a
MER89
MER90a
MER66B
LTR56
LTR52
LTR5_Hs
MER49
HERV3-int
LTR12
MER57A1
MER57B2
MER57B1
SVA_A
L1M3e
AmnSINE1
MLT1M
MER44B
Tigger3c
LTR8
Enrichment Score
−100 −50 0 50 100
B
CD4+ T Effector
CD8+ T Effector Memory
Resting
Activated
Resting
Activated
LTR10A/F ABC Enhancers
175
0
100
0
-1Kb
1Kb
D
CD19+ B
Naive B
Memory B
Plasmablast
LTR2B ABC Enhancers
160
0
100
0
-1Kb
1Kb
C
CD4+ T Effector
CD8+ T Effector Memory
Footprint Score log2(FC)
Resting Condition Binding Prediction
Activated Condition Binding Prediction
FOSL2 Binding Sites within LTR10A/F ABC Enhancers
log2(FC)
−2 0 2 4 6
State
Unbound
Bound
E
CD19+ B Cell Binding Prediction
Naive B Cell Binding Prediction
Memory B Cell Binding Prediction
Plasmablast Binding Prediction
SPI1 Binding Sites within LTR2B ABC Enhancers
State
Unbound
Bound

Figure 2.2. TE families contribute to cell-type specific hematopoietic gene regulation
(A) Enrichment of transposable element families (rows) within ABC enhancers of human hematopoietic cell types (columns). Rows and columns are hierarchy clustered based on the enrichment score. Enrichment was determined using the GIGGLE framework [197] and a significantly enriched TE was defined if at least 20 elements of the TE family overlapped ABC enhancers of a cell type, had an odds ratio > 2.5 and a Fisher's two-tailed p-value < 0.01.
(B,D) Heatmap of ATAC-Seq signal over the set of LTR10A/F elements (n=56) (B) or LTR2B elements (n=26) (D) overlapping ABC enhancers. Each column is the aggregate ATAC-Seq signal across all donors and replicates. The average, normalized ATAC-Seq signal within a +/-1kb window of the elements is displayed on the bottom of each heatmap.
(C) Binding predictions for FOSL2 (JASPAR motif ID: MA0478.1) in LTR10A/F ABC enhancers using the TOBIAS transcription factor occupancy framework [198]. The footprint fold change column represents the matched change in FOSL2 footprint scores between the resting and activated T cell states. Log2(FC) is calculated as log2(activated/resting). The binding prediction columns depict whether individual FOSL2 binding sites were predicted by TOBIAS to be bound/unbound in resting and activated T cell states.
(E) Binding predictions for SPI1 (HOCOMOCOv11) in LTR2B ABC enhancers, analogous to (C).

regulating T cell activation. The LTR10A-G family has previously been documented to contain transcription factor motifs for the AP-1 family of TFs [199], which is consistent with the role of AP-1 in T cell activation. We therefore asked whether these LTR10A/F ABC enhancers may be regulated by AP-1 TFs. We leveraged the ATAC-Seq data on activated and resting T cells in our dataset and implemented the TOBIAS transcription factor occupancy framework [198] to infer TF occupancy in ABC enhancers. Indeed, nearly all AP-1 motifs within LTR10A/F ABC enhancers were predicted to be bound by AP-1 TFs in activated T cells, whereas binding was largely absent in the resting state (**Figure 2.2C**), further supporting the role of these elements as putative TE-derived enhancers. In addition, we also observed notable enrichment of LTR2B elements within ABC enhancers of B cells (**Figure 2.2B**), and the enhancer elements were highly accessible across the B cell lineage (**Figure 2.2D**). Furthermore, TOBIAS predicted binding of SPI1, an important TF regulating B cell identity [200,201] in LTR2B ABC enhancers (**Figure 2.2E**). While predictions at some SPI1 motifs were not consistently bound across the B cell lineage, this may reflect cell-type specific tuning of enhancer activity.

Overall, our analysis argues for a pervasive and cell type-specific contribution of TE families to hematopoietic ABC enhancers, and a particular role for TEs in transcriptional regulatory networks of lymphoid cells via TFs essential for cell identity and function.

2.2.4 TEs and their epigenetic regulatory machinery are dynamically expressed during hematopoietic differentiation

Hematopoiesis involves the dynamic control of transcriptional regulatory networks to generate a diverse repertoire of hematopoietic and immune cell types [202,203]. In this context, we asked whether TEs are dynamically regulated during hematopoietic differentiation, given their contribution to gene regulation within specific cell types. We quantified TE expression from RNA-Seq data (see **Methods**) on FACS sorted hematopoietic populations spanning the hematopoietic hierarchy, which identified 1,295 expressed TE families. As expected, hierarchical clustering of these hematopoietic populations based on expressed genes revealed lineage relationships that have been well documented in the field (**Figure 2.3A**). Notably, clustering based solely on expressed TEs also largely recapitulated these lineage relationships (**Figure 2.3B**), in concordance with the cell type specificity of TE expression within the hematopoietic system. This result is in line with observations of selective expression of particular TE families within pluripotent stem cells and mature somatic cells, as well as during dynamic processes such as embryonic development and cellular reprogramming [204]. To dissect the dynamics of TE expression during hematopoietic differentiation we identified differentially expressed TEs in hematopoietic populations relative to HSCs. Notably, expression of TEs increased during lymphoid differentiation, initiating at common lymphoid progenitors (CLP). Conversely, a larger number of TEs exhibited decreased expression in various myeloid and erythroid progenitors (**Figure 2.3C**).

Various cellular mechanisms exist to repress TEs within the human genome that largely depend on the formation of heterochromatin at TE loci. Thus, the dynamic expression of TEs

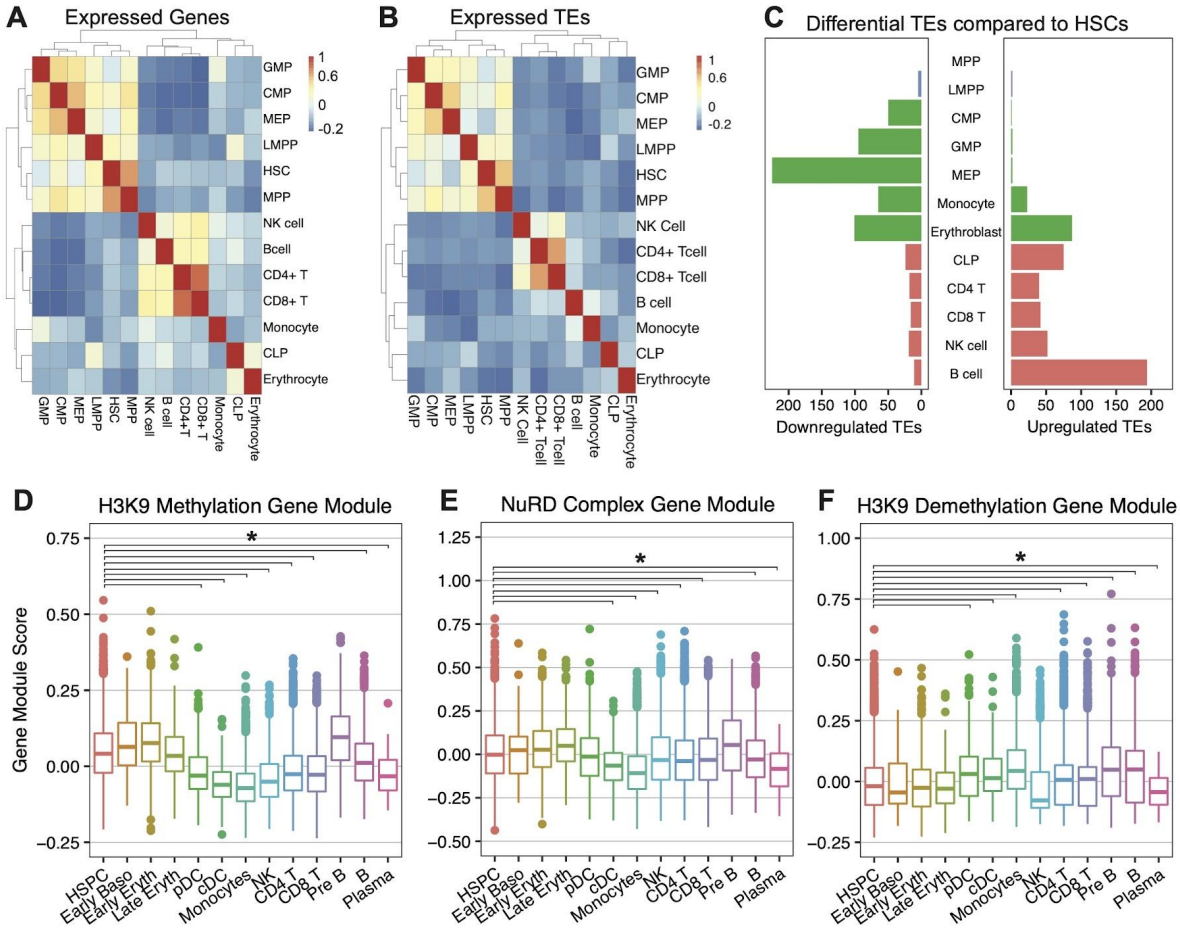


Figure 2.3. TEs and their regulatory machinery are dynamically expressed during hematopoietic differentiation

(A-B) Pearson hierarchical clustering of expressed genes (n=23,272) and expressed TEs (n=1,295) across hematopoietic populations profiled from Corces, *et al*. 2016. TEs were quantified using TEtranscript [205] from bulk RNA-Seq.

(C) Quantification of the number of differentially expressed TEs within each hematopoietic cell type compared to HSCs.

(D-F) Genetic regulators of heterochromatin formation are dynamically expressed across the human hematopoietic hierarchy. Single cell RNA-Seq data on healthy human donors from Granja, *et al.* 2019 were reanalyzed to calculate gene module scores for H3K9 methylation (GO:0051567), NuRD complex (GO:0016581), and H3K9 demethylation (GO:0032454) gene sets across all single cells. Gene module scores represent the average expression of genes in the set above background expression. Boxplots depict the median and IQR for each patient cohort, with whiskers extending to 1.5*IQR in either direction from the top or bottom quartile. Welch's *t*-test was used in (D-E) to test if the mean HSPC gene module score is greater than the means of other hematopoietic cell types, and in (F) to test if the mean HSPC gene module score is less than the means of other hematopoietic cell types. FDR corrected p-values are noted with (*) if statistically significant.

observed during hematopoietic differentiation prompted us to ask whether regulators of heterochromatin formation likewise exhibit dynamic expression during differentiation. We quantified the expression of gene modules involved in H3K9 methylation using a scRNA-Seq

atlas of the human hematopoiesis system. Mature cell types exhibited significantly reduced expression of H3K9 methylation genes compared to HSPCs (**Figure 2.3D**), consistent with the up-regulation of TE expression. The NuRD complex demonstrated similar expression dynamics as the H3K9 methylation gene program (**Figure 2.3E**), further supporting our observations that the expression of TE regulatory machinery is inversely correlated with TE expression during lymphoid differentiation. In further agreement with these observations, H3K9 demethylation modules were elevated in mature hematopoietic cell types compared to HSPCs (**Figure 2.3F**). The collective observations that TEs are upregulated during lymphoid differentiation with associated attenuation of TE regulatory machinery data suggest that TEs are dynamically regulated during hematopoietic differentiation.

2.2.5 Modulation of chromatin machinery regulating TEs directs lymphoid lineage decisions

We hypothesized that TEs may influence lymphoid lineage decisions, motivated by our data that TEs contribute to enhancer regulation in lymphoid cells and are dynamically expressed during lymphoid differentiation. Thus, we employed a gene-centric loss-of-function approach in HSPCs to systematically knockout regulators of heterochromatin formation as a means to derepress TEs and then study lineage transitions in T and B cell-supportive differentiation conditions (**Figure 2.4A**). H3K9 lysine methyltransferases, as well as the transcriptional co-repressor *TRIM28* exhibited dynamic expression across hematopoietic progenitors and mature, differentiated cell populations (**Figure 2.4B**), and were therefore prioritized for functional interrogation. We nucleofected CRISPR/Cas9 ribonucleoproteins (RNPs) into CD34+ human umbilical cord blood to generate replicate knockouts using two distinct gRNAs per gene, as well as a control targeting the *AAVS1* locus (**Table S6**). Frameshifting indel formation overall was

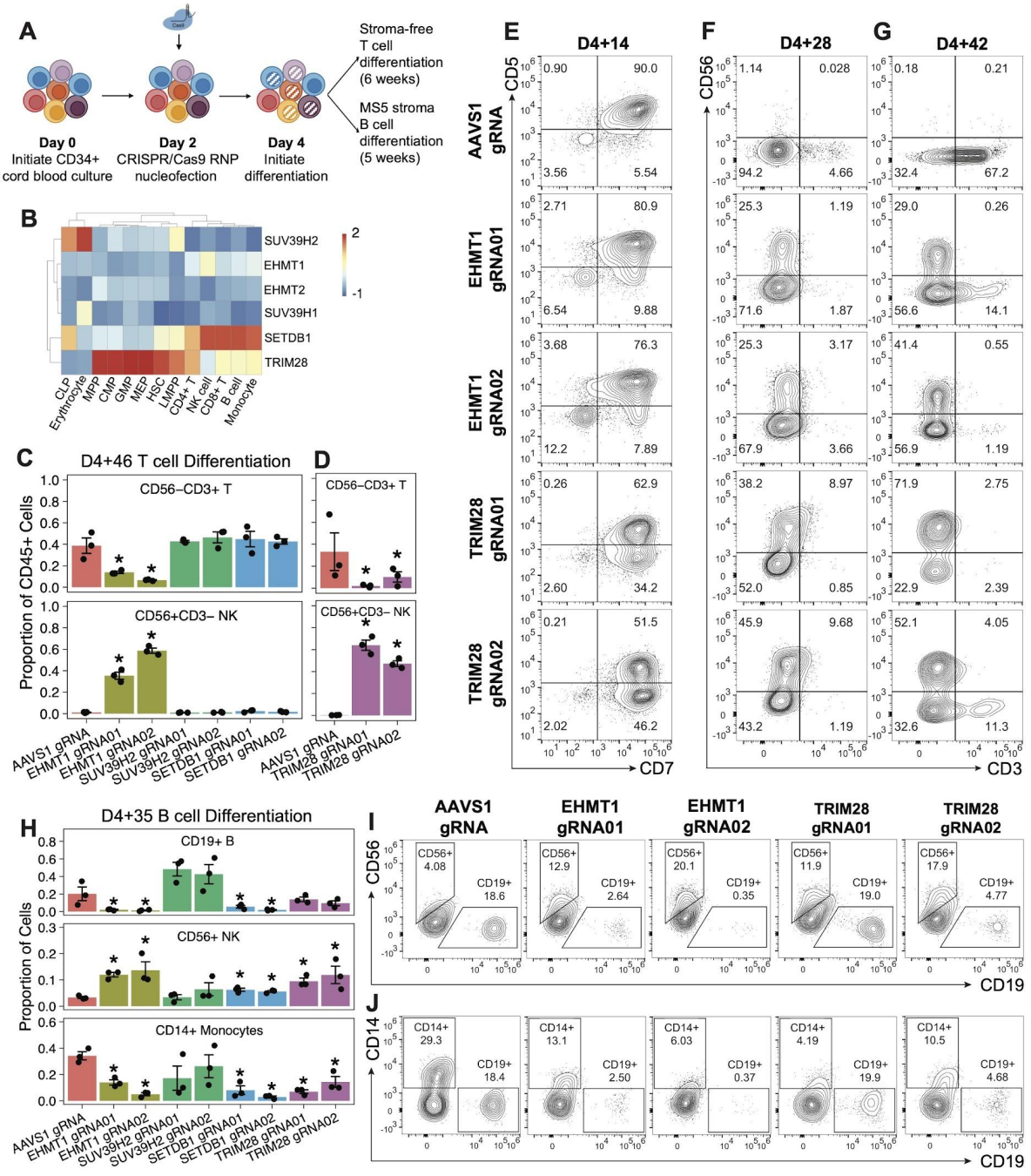


Figure 2.4. Modulation of TE regulatory machinery influences lineage output from hematopoietic stem and progenitor cells

(A) Experimental scheme to genetically knockout genes involved in TE regulation via heterochromatin formation within CD34+ umbilical cord blood and assess *in vitro* lymphoid differentiation.

(B) Expression heatmap of genes involved in the regulation of heterochromatin across major bone marrow hematopoietic cell types. The RNA-Seq data is reanalyzed from Corces *et al.* 2016. Heatmap color represents the Z-score of gene expression across all cell types.

(C-D) Fractional abundance of CD56-CD3+ T and CD56+CD3- NK cells within the CD45+DAPI- population quantified by flow cytometry following six weeks of *in vitro* T cell differentiation (n=3 independent differentiation

experiments per gRNA condition). * = p < 0.05 in comparison to the *AAVS1* gRNA control evaluated with a two-sided Wilcoxon rank-sum test.
(E) Flow cytometry profiling of CD5 and CD7 on lymphoid progenitors following 14 days of *in vitro* T cell differentiation. The plots are pregated on CD45+DAPI- cells.
(F-G) Flow cytometry profiling of CD56+CD3- NK and CD56-CD3+ T cell populations following 28 days (F) and 42 days (G) of *in vitro* T cell differentiation. The plots are pregated on CD45+DAPI- cells.
(H) Fractional abundance of CD19+ B cells, CD56+ NK cells and CD14+ monocytes within the CD45+DAPI- population quantified by flow cytometry following 35 days of differentiation in an MS5 co-culture assay (n=3 independent differentiation experiments per knockout condition). * = p < 0.05 in comparison to the *AAVS1* gRNA control evaluated with a two-sided Wilcoxon rank-sum test.
(I-J) Representative flow cytometry plots of CD19+ B cells, CD56+ NK cells and CD14+ monocytes across gRNA conditions targeting the *AAVS1* locus, *EHMT1* and *TRIM28*. The plots are pregated on CD45+DAPI- cells.

highly efficient for each gRNA across multiple donor pools, enabling experiments directly on the pool of nucleofected cells (**Supplementary Figure 2.3A-B**).

We first initiated T cell differentiation 48 hours post-delivery of RNPs using a stroma-free assay (**Methods**). After six weeks of *in vitro* differentiation (noted as D4+46), we observed that knockout of *EHMT1* resulted in a reduction of CD56-CD3+ T cells and an associated increase of CD56+CD3- natural killer (NK) cells compared to an *AAVS1* gRNA control (**Figure 2.4C**). Furthermore, knockout of *TRIM28*, a key regulator of TE silencing, phenocopied that of *EHMT1* (**Figure 2.4D**), highlighting a central role of TE regulatory machinery in the determination of NK cell fate. The proportional increase in CD56+CD3- NK cells at D4+46 was also accompanied by increased absolute numbers of CD56+CD3- cells (**Supplementary Figure 2.3C-D**). Sequencing of isolated NK cells confirmed the presence of frameshifting indels (**Supplementary Figure 2.3E-F**), further supporting the conclusion that the knockouts generate NK cells. To dissect the kinetics of this T-to-NK lineage bias, we performed time course profiling every two weeks over the six weeks of differentiation (**Figure 2.4E-G**). Notably, we observed an altered distribution of lymphoid progenitors at D4+14 characterized by an attenuated frequency of cells expressing the T cell-associated marker CD5 without affecting overall cell viability (**Figure 2.4E**, **Supplementary Figure 2.3G**). These data imply that NK lineage bias manifests in early stages of lymphoid differentiation. NK cells were evident an additional two weeks later at D4+28 and proportionally further increased by D4+46 (**Figure 2.4F-G**, **Supplementary Figure 2.3H**).

Finally, sustained treatment with a potent enzymatic inhibitor of GLP/G9a, UNC0642 [206] over the course of T cell differentiation resulted in a dose-dependent phenocopy of the *EHMT1* CRISPR knockout (**Supplementary Figure 2.3I-J**), leading to the acquisition of CD56+CD3- NK cells and general attenuation of CD56-CD3+ T cells. These data strongly suggest that modulation of H3K9 methyltransferase activity and regulation of heterochromatin facilitates NK lineage choice.

Our data in T cell-supportive differentiation conditions underscore a role for TE regulation in lymphoid fate decisions. We next asked whether similar mechanisms may be applicable during B-lymphoid differentiation. We replicated the CRISPR/Cas9 knockouts in CD34+ umbilical cord blood HSPCs and utilized an MS5-stroma co-culture assay, which is conducive for B cell development as well as myeloid differentiation. Consistent with our observations from the *in vitro* T cell differentiation assay, knockout of *EHMT1* resulted in NK lineage skewing at the expense of CD19+ B cells and CD14+ monocytes following five weeks of differentiation (**Figure 2.4H-J**). Knockout of *TRIM28* also enhanced the proportion of CD56+ NK cells at the expense of CD14+ monocytes, although we did not observe a statistically significant decrease in CD19+ B cells (**Figure 2.4H**). In further support of the modulation of TE regulatory machinery to facilitate NK fates, knockout of *SETDB1* also enhanced the generation of CD56+ NK cells (**Figure 2.4H**, **Supplementary Figure 2.3K**). As in T cell-supportive conditions, proportional increases in NK cells were accompanied by an increase in the absolute numbers of NK cells compared to the *AAVS1* gRNA control (**Supplementary Figure 2.3K-L**). In further accordance with our T cell differentiation experiment, lineage alterations may have manifested early during differentiation since knockout of *EHMT1* and *SETDB1* resulted in the precipitous decrease in CD19+ progenitors at D4+14 (**Supplementary Figure 2.3M**). Taken together, our data indicate that the modulation of *EHMT1* or *TRIM28* has a pronounced influence on NK cell development during lymphoid differentiation, further establishing that manipulation of chromatin machinery regulating TEs can direct hematopoietic lineage decisions.

2.2.6 Knockout of *EHMT1* or *TRIM28* generates distinct hematopoietic progenitors with early NK characteristics

We hypothesized that knockout of *EHMT1* or *TRIM28* influences NK lineage selection early during lymphoid differentiation. Motivated by our data demonstrating an altered distribution of progenitor populations within T cell-supportive conditions (**Figure 2.4E**), we utilized single cell RNA-Seq and ATAC-Seq technologies to dissect the molecular landscapes of hematopoietic progenitors across *EHMT1*, *TRIM28* and *AAVS1* gRNA conditions at D4+14. We measured chromatin accessibility across a total of 23,593 cells. All gRNA conditions exhibited a canonical fragment size distribution and high signal-to-noise (**Supplementary Figure 2.4A-D**). We first identified chromatin states by clustering scATAC-Seq profiles from all cells, revealing 8 distinct clusters (**Figure 2.5A**). We complemented these data by also capturing 25,288 single cell transcriptomes with scRNA-Seq across the same gRNA conditions. Clustering of all cells revealed 7 transcriptional states (**Figure 2.5B**). For both assays, we observed high concordance in clustering across different gRNAs targeting the same gene (**Supplementary Figure 2.4E-F**), establishing that the clustering results were driven by biological rather than technical effects. Furthermore, EHMT1 gRNA and TRIM28 gRNA conditions clustered distinctly from the AAVS1 gRNA control in both assays, indicating that knockout of *EHMT1* or *TRIM28* are driving distinct chromatin and transcriptional states within hematopoietic progenitors (**Figure 2.5C-D**). We first utilized the single cell transcriptional information to annotate the identity of scRNA-Seq clusters. Notably, *RAG1* expression was primarily restricted to cluster 0, which is enriched for AAVS1 gRNA cells and suggests a T cell-fated state (**Figure 2.5E**). In further support of this, cluster 0 also selectively expressed important genes required for T cell development, such as *LEF1, BCL11B, ZEB1, CD1B* and *RAG2* (**Figure 2.5F**). Interestingly, the gene encoding for CD56, *NCAM1*, was only expressed in clusters enriched for *EHMT1* and *TRIM28* knockout cells, and *IL2RB*, one of the earliest markers for NK-fated progenitors already was expressed in EHMT1 knockout clusters (cluster 1 and 5) (**Figure 2.5E**). EHMT1 knockout

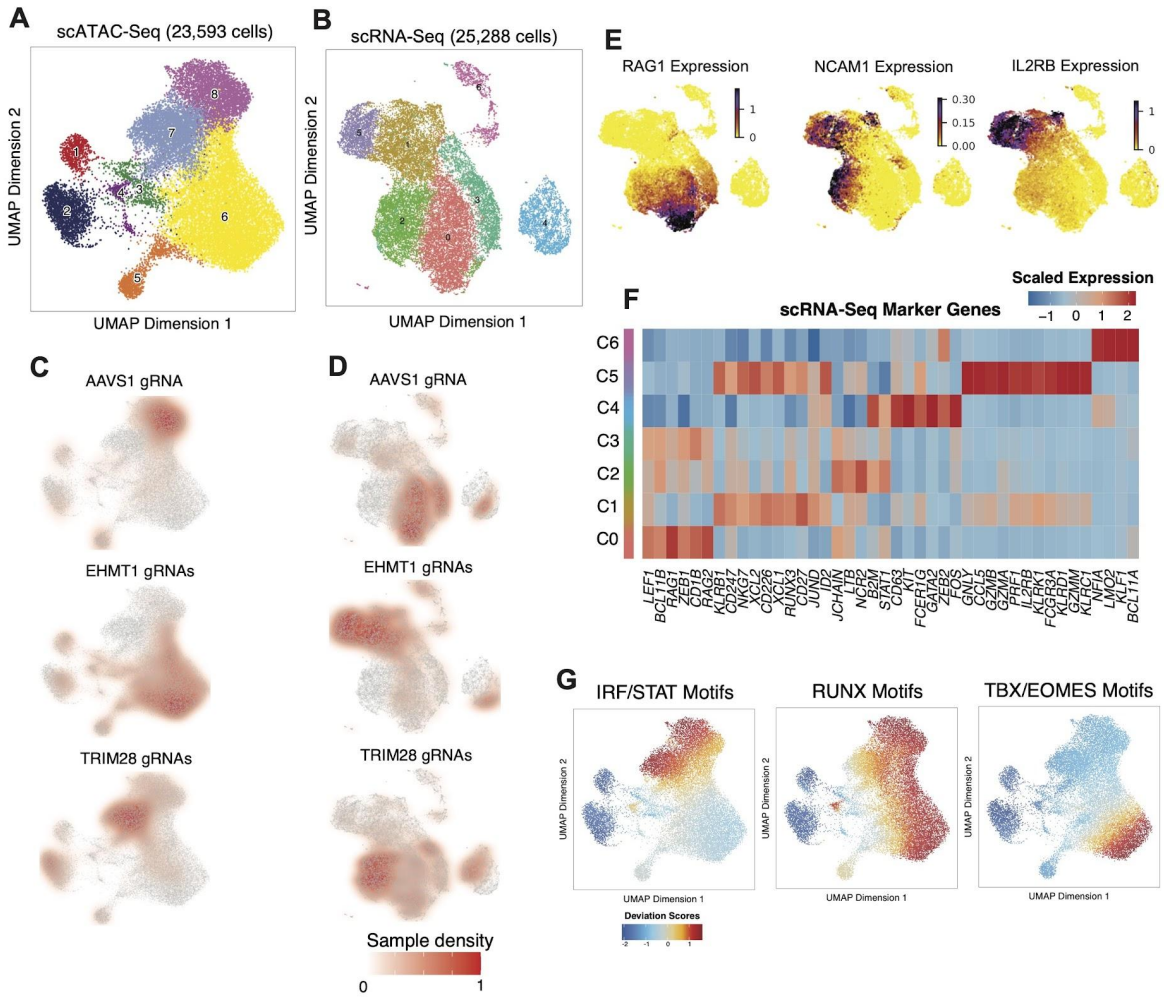


Figure 2.5. Knockout of *EHMT1* or *TRIM28* generate distinct hematopoietic progenitors with early NK characteristics

(A) scATAC-Seq latent semantic indexing UMAP projection and clustering of 23,593 cells encompassing *AAVS1, EHMT1* gRNA01, *EHMT1* gRNA02, *TRIM28* gRNA01 and *TRIM28* gRNA02 conditions at D4+14 of *in vitro* lymphoid differentiation.

(B) scRNA-Seq UMAP projection and clustering of 25,288 cells encompassing *AAVS1, EHMT1* gRNA01, *EHMT1* gRNA02, *TRIM28* gRNA01 and *TRIM28* gRNA02 conditions at D4+14 of *in vitro* lymphoid differentiation.

(C-D) scATAC-Seq UMAP plot (C) and scRNA-Seq UMAP plot (D) colored by the embedding density of cells from the indicated gRNA conditions.

(E) Visualization of the expression of T and NK cell associated genes overlaid on the scRNA-Seq UMAP.

(F) Select marker genes that are differentially expressed (FDR < 0.05, two-tailed Wilcoxon rank-sum test) in each transcriptional cluster from scRNA-Seq. Color scale corresponds to z-scored, log-transformed mean gene-expression counts for each cluster.

(G) UMAP projection of scATAC-Seq data colored by chromVAR TF motif bias-corrected deviations for the indicated factors. TF motifs represent a clustered archetype derived from [207].

clusters also positively enriched for NK genes, such as *KLRK1* (encoding for NKG2D), *FCGR3A* (encoding for CD16), granzymes and perforin (*GZMA*, *GZMB*, *GZMM*, *PRF1*), *CD247*, *CD226*

(encoding DNAM-1) and *ID2.* This expression is consistent with prior reports of transcriptional markers of NK progenitors within human fetal and adult tissues [208]. TRIM28 knockout cells within cluster 2 were characterized by high expression of *LTB, JCHAIN, NCR2,* and *B2M* with low levels of *RAG1* and *RAG2*, potentially suggesting an earlier stage of development compared to the EHMT1 knockout cells.

Collectively, the transcriptional data underscore the distinct, NK-biased states arising from EHMT1 and TRIM28 knockouts. We therefore next examined transcription factor activity exhibited in hematopoietic progenitors as a result of loss of *EHMT1* or *TRIM28*. We leveraged our scATAC-Seq data to infer TF activity with chromVAR [209] and identified key developmental TFs with cluster-specific activity (**Figure 2.5G**). Interestingly, IRF and STAT TFs exhibited prominent activity within TRIM28 knockout cells, which is consistent with the transcriptional upregulation of *STAT1* in TRIM28 knockout cells (scRNA-Seq cluster 2). We note that other studies have documented inflammatory signatures associated with derepression of TEs via loss of TRIM28, as well as the relevance of STAT1 in NK cell function. RUNX3 is expressed in developing NK cells and increases with NK maturation [210–212]. In line with this established biology, we observed RUNX TF activity within AAVS1 and EHMT1 gRNA scATAC-Seq clusters (clusters 6 and 8) and specific upregulation of *RUNX3* expression within EHMT1 gRNA scRNA-Seq clusters (clusters 1 and 5). Additionally, developmentally important TFs for NK maturation, TBX21/EOMES also exhibited localized activity within EHMT1 gRNA scATAC-Seq clusters. In further support of the distinct genetic regulation, RNA velocity (**Supplementary Figure 2.4G**) and scATAC-Seq trajectory analysis (**Supplementary Figure 2.4H**) are both reflective of cellular trajectories diverging from the *AAVS1* gRNA control population. Taken together, these data reflect the distinct transcriptional and chromatin states characterized by knockout of EHMT1 or TRIM28 and the induction of NK fates within hematopoietic progenitors.

2.2.7 *EHMT1* and *TRIM28* knockout NK cells exhibit distinct states and effector functions

To characterize the NK cells generated as a result of loss of EHMT1 or TRIM28, we differentiated lymphoid progenitors within NK supportive differentiation conditions for an additional two weeks then utilized the resulting NK cells for phenotypic and molecular assays (**Figure 2.6A**). NK cells isolated on D4+28 exhibited derepression of several hundred TE families, including LTR, MER, L1 and Alu elements (**Supplementary Figure 2.5A-B**). Interestingly, the classes of derepressed TEs were largely distinct between *TRIM28* and *EHMT1* knockouts. This indicates that EHMT1 and TRIM28 regulate separate classes of TEs within NK cells. In addition, 36 zinc finger proteins were transcriptionally upregulated in *TRIM28* knockout NK cells and 26 are annotated to contain a KRAB domain (**Supplementary Figure 2.5C**), consistent with the known mechanism-of-action of TRIM28-mediated TE silencing. This is in contrast to *EHMT1* knockout NK cells where only 3 KZFPs were transcriptionally upregulated and in line with the non-overlapping sets of TEs derepressed between *EHMT1* and *TRIM28* knockouts. In either case, derepressed TEs contained TF binding sites relevant for NK cell differentiation and function (**Supplementary Figure 2.5D**), supportive of a molecular model whereby TE derepression enables TFs to bind chromatin to drive selection of NK cell fate.

Based on these findings we further investigated the phenotypic properties of *EHMT1* and *TRIM28* knockout NK cells. NK cells were efficiently derived in all gRNA conditions and expressed canonical NK surface markers (**Supplementary Figure 2.6A**). RNA-Seq confirmed transcriptional downregulation of *TRIM28* and *EHMT1* within TRIM28 gRNA and EHMT1 gRNA conditions, respectively, and each condition expressed core NK signature genes [213] (**Supplementary Figure 2.6B-C**). All knockout NK cells displayed cytotoxicity against K562s in co-culture, further supporting their NK identity and that knockout of *EHMT1* or *TRIM28* does not inhibit the cytotoxic function of NK cells (**Supplementary Figure 2.6D**). Notably, nearly three-fold more NK cells were generated in *EHMT1* knockouts and for one gRNA targeting TRIM28 compared to the *AAVS1* gRNA control (**Figure 2.6B-C**). Knockout of *EHMT1* also

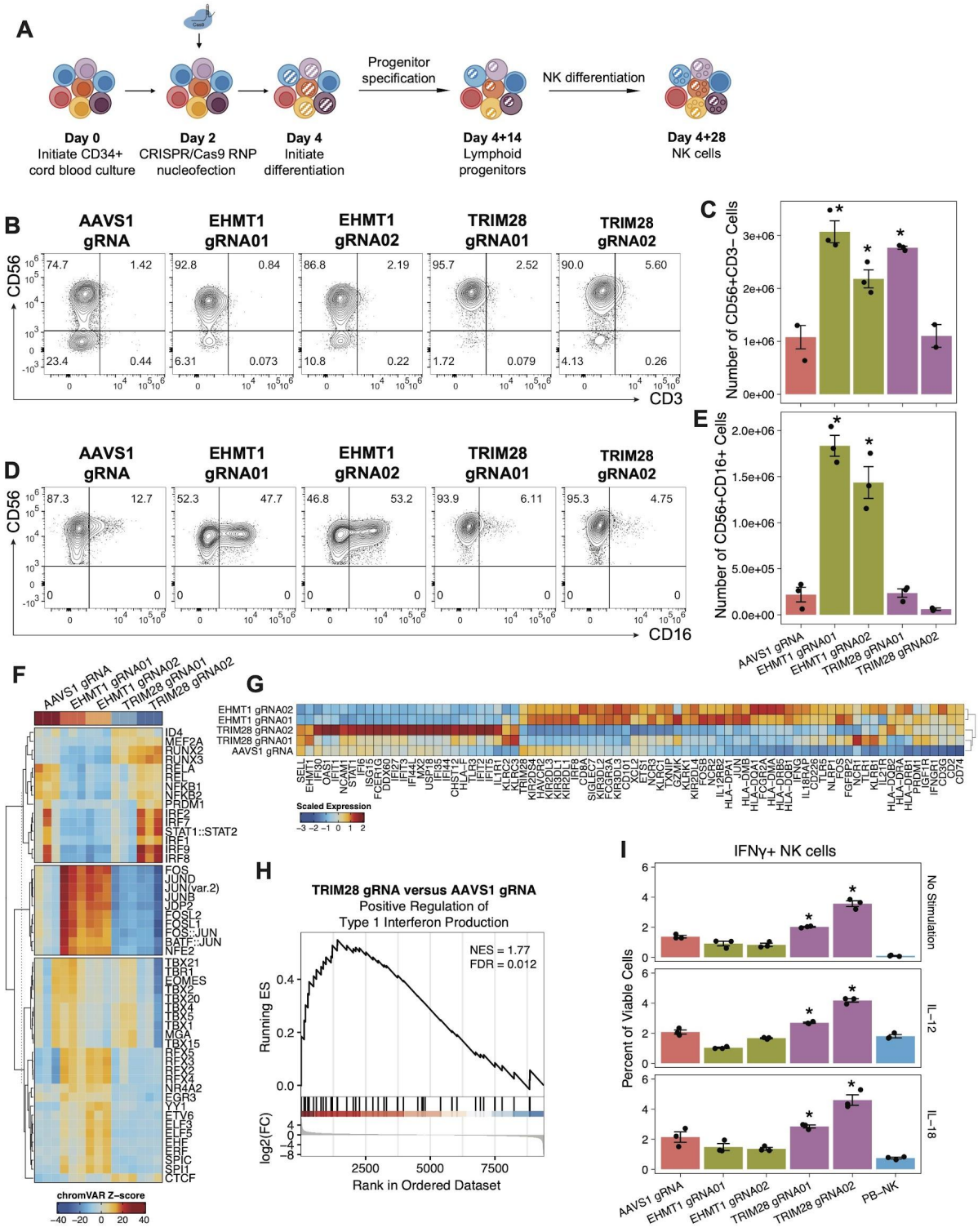


Figure 2.6. *EHMT1* and *TRIM28* knockout NK cells exhibit unique phenotypic and molecular states.
(A) Experimental scheme to genetically knockout EHMT1 or TRIM28 within CD34+ umbilical cord blood and perform NK-supportive differentiation.

(B) Expression of CD56 and CD3 assayed by flow cytometry on D4+28 of *in vitro* NK differentiation. Plots are pre-gated on viable CD45+ cells.
(C) Quantification of the number of CD56+CD3- NK cells on D4+28 of *in vitro* differentiation for all gRNA conditions. n=3 independent differentiation replicates per gRNA. * = $p < 0.05$ in comparison to the *AAVS1* gRNA control evaluated with a two-sided Wilcoxon rank-sum test.
(D) Expression of CD56 and CD16 assayed by flow cytometry on D4+28 of *in vitro* NK differentiation. Plots are pre-gated on viable CD45+CD56+ cells.
(E) Quantification of the number of CD56+CD16+ NK cells on D4+28 of *in vitro* differentiation for all gRNA conditions. n=3 independent differentiation replicates per gRNA. * = $p < 0.05$ in comparison to the *AAVS1* gRNA control evaluated with a two-sided Wilcoxon rank-sum test.
(F) Heatmap of changes in ATAC-Seq chromatin accessibility for the top TFs with the greatest accessibility variability between gRNA conditions. The color scale corresponds to chromVAR accessibility deviation z-scores of a TF across conditions. n=3 independent differentiation replicates.
(G) Expression heatmap of select genes exhibiting differential expression (log2(fold change) > 1 and FDR < 0.05, DESeq2) in at least one gRNA condition compared to the *AAVS1* gRNA control. The color scale corresponds to z-scored, log-transformed mean gene-expression counts.
(H) GSEA on RNA-Seq data comparing *TRIM28* gRNA versus *AAVS1* gRNA demonstrating over-representation of genes related to type I interferon production (GO:0032481) within *TRIM28* gRNA cells.
(I) Proportion of IFN-ɣ+ NK cells measured by intracellular flow cytometry across unstimulated and stimulated conditions. Each dot represents an independent differentiation replicate. * = $p < 0.05$ in comparison to the *AAVS1* gRNA control evaluated with a two-sided Wilcoxon rank-sum test.

resulted in a greater frequency of CD16+ NK cells (**Figure 2.6D-E**), suggestive of a more mature state. AP-1 transcription factors exhibited enhanced activity with EHMT1 knockout NK cells, whereas inflammatory TFs (IRFs, STATs and NFKB) were more active in TRIM28 knockouts, consistent with early chromatin signatures in lymphoid progenitors (**Figure 2.6F, Supplementary Figure 2.6E-F**). *TRIM28* knockout NK cells also showed transcriptional upregulation of various interferon related genes (**Figure 2.6G-H**), which equated to enhanced IFN-ɣ production at baseline and upon stimulation with IL-12 or IL-18 (**Figure 2.6I**). In further support of a more mature state, *EHMT1* knockout NK cells upregulated various KIR and HLA-II genes in comparison to *TRIM28* knockout and AAVS1 gRNA control populations. These data overall support the conclusion that knockout of *EHMT1* and *TRIM28* generate NK cells with distinct molecular and phenotypic properties.

2.3 DISCUSSION

While several lines of evidence suggest that TEs co-evolved with the emergence of the innate immune response, there has been little work to dissect the role of TEs during the dynamic

differentiation processes that generate functional immune cells from HSPCs. In this study, by combining analytical models, omics analysis and chemical/genetic perturbations, we reveal that the regulation of TEs provides instructive roles in the human hematopoietic system, particularly during lymphoid differentiation. We built genome-wide maps of enhancer-gene regulation on an individual cell type basis within the human hematopoietic system, and used this resource to understand the contribution of TEs to gene regulation. TEs exhibited dynamic activity during lymphoid differentiation, and derepression of TEs during lymphoid differentiation mediated by modulating regulators of heterochromatin formation resulted in the acquisition of NK cell fates, despite differentiation conditions designed to support T or B cells. In particular, knockout of *EHMT1* or *TRIM28* within HSPCs generated distinct lymphoid progenitor states that diverted from T fated progenitors, and resulted in derepression of different TE families within *in vitro* derived NK cells. We also demonstrate that knockout of *EHMT1* generated more CD16+ NK cells, whereas knockout of *TRIM28* resulted in elevated IFN-ɣ production, reflective of distinct NK states.

The availability of epigenomic data from various cell types has fueled scientific inquiry and hypothesis generation on the role of TEs to regulate transcriptional networks [46,176,214]. While abundant, these data often lack insights on putative functional enhancers [193,215,216], which hinders identifying true regulatory relationships [217]. To address this challenge, we employed the ABC model—which performs robustly to predict CRISPR experiments of endogenous enhancer activity—to generate enhancer-gene predictions within the hematopoietic system. The experimental difficulty of perturbing repetitive elements within primary cells often limits the degree to which individual elements can be functionally tested for regulatory activity. By utilizing state-of-the-art models to identify putative functional enhancers, the approach utilized in this study provides higher confidence of the regulatory potential of implicated TEs. Furthermore, in order to systematically study dynamic differentiation processes, we generated ABC predictions on a comprehensive set of hematopoietic cell types to capture stages of differentiation from

hematopoietic stem cells to erythroid, myeloid and lymphoid lineages, as well as stimulation states within immune populations. This resource can be utilized to understand gene regulation more generally due to the long standing application of hematopoiesis as a model of stem cell biology, and to investigate genetic bases of hematopoietic diseases.

A core finding from our study was that TEs are dynamically expressed during hematopoietic differentiation and progressively upregulated during lymphopoiesis with associated downregulation of regulators of heterochromatin formation. Indeed, this observation is consistent with reports of other dynamic processes, such as embryonic development, where widespread DNA-demethylation causes stage-specific TE expression [106]. We systematically perturbed a diverse set of regulators of heterochromatin within HSPCs to derepress TEs during hematopoietic differentiation. We discovered that knockout of H3K9 lysine methyltransferase, *EHMT1* or the transcriptional co-repressor *TRIM28* resulted in NK lineage skewing at the expense of B and T cell fates. We identified distinct TE families that were derepressed within *EHMT1* knockouts versus *TRIM28* knockouts, which is consistent with prior observations in murine embryonic stem cells demonstrating EHMT1/EHMT2 involvement in silencing MERVL elements with little involvement from TRIM28 [74]. These distinct sets of derepressed TEs may contribute to the different NK phenotypes observed between the two knockouts. Of note, although we knocked out SUV39H2 and we did not observe a phenotype; we believe this may be due to compensation by SUV39H1 as SUV39H1/2 double knockouts were needed to observe heterochromatin disruption in T cells [218].

Findings from this study lend compelling support to the long-standing hypotheses that the innate immune system evolved through co-option of endogenous viruses [26,184]. Intriguingly, our study demonstrates that TE derepression through modulation of TE regulatory machinery during lymphoid differentiation facilitates development of the NK cells—the very cells that are responsible for protection from viral pathogens. Seminal studies have implicated endogenous retroviruses in regulating innate immune pathways [16,36,182]. We observed upregulation of

interferon-related genes within *TRIM28* knockout NK cells, which is consistent with innate immune sensing of TE derived transcripts [118]. Furthermore, *TRIM28* knockout NK cells also experienced enhanced IFN-ɣ production, a critical inflammatory mediator that these cell secrete for antiviral innate immunity.

The long-standing interest in NK cells as adoptive immunotherapy for cancer treatment has garnered increased attention, reflective of clinical desires to complement or overcome limitations of T cell-based therapies [219,220]. Accumulating evidence over the past decade, however, has underscored the functional and molecular heterogeneity of human NK cells owing to transcription factor utilization, prior antigen exposure, developmental ontogeny, and tissue residence, among likely other factors [221–223]. These factors ultimately influence the anti-tumor activity of NK cells [224–227], although clinical studies have not fully resolved the specific NK subtypes that are most efficacious for cancer remission. Therefore, strategies to engineer and derive targeted populations of NK cells *in vitro* with distinct effector functions will be essential to develop effective therapies and better understand the cells responsible for cancer remission. Our study reveals the potential of leveraging TE regulatory machinery to modulate TEs during hematopoietic differentiation for the *in vitro* generation of NK cells with diverse properties for translational applications.

2.4 METHODS

2.4.1 CRISPR/Cas9 knockouts in CD34+ umbilical cord blood

Genetic knockouts of select genes were generated in primary hematopoietic stem and progenitor cells (HSPCs) via nucleofection of CRISPR/Cas9 ribonucleoproteins (RNPs). Two independent knockouts were generated for each gene using separate gRNAs and knockouts were replicated across two different donor pools of human umbilical cord blood. First, CD34+ umbilical cord blood cells (AllCells) were thawed from liquid nitrogen storage via dropwise addition of RPMI 1640 + 10% FBS. Cells were centrifuged at 200g for 8 minutes and subsequently washed with FACS buffer (PBS + 2% FBS). Cells were cultured for 48 hours prior to nucleofection at a density of 250,000 cells/mL in SFEM II media (StemCell Technologies, 09605) supplemented with 100 ng/mL SCF, TPO, FLT3L, and IL-6 (Peprotech) within a 5% O2, 5% CO2, 37C incubator. Next, the top two protospacer sequences for each desired gene were selected from the Brunello genome-wide knockout library (Broad Institute) based on the Rule Set 2 score. A control gRNA targeting the *AAVS1* locus was designed in Benchling. All gRNA sequences are available in **Table S6** and were synthesized by IDT as an Alt-R CRISPR/Cas9 sgRNA. RNPs were complexed by incubating 105 pmol of Cas9 protein (IDT, 1081058) with 125 pmol of sgRNA in a 5 uL total volume for 10 minutes at room temperature. CD34+ cord blood HSPCs were washed with FACS buffer and 125,000 cells were resuspended in 20 uL of P3 Primary Cell Nucleofection Solution (Lonza, V4XP-3032) per nucleofection. RNPs were nucleofected into HSPCs with 3.85 uM electroporation enhancer (IDT, 1075915) using a 16-reaction nucleovette and pulse code DZ-100 on a Lonza 4D-Nucleofector. Cells were cultured for 36-48 hours at 5% O2, 5% CO2 in 96-well U-bottom plates (Corning, 351177) then subsequently harvested for genomic DNA extraction and initiation of *in vitro* differentiation assays.

2.4.2 Quantification of indel frequencies

Genomic DNA was extracted from at least 50,000 cells using a custom extraction buffer (1 mM CaCl2, 3 mM MgCl2, 1 mM EDTA, 10 mM Tris-HCl, 1% Triton X-100, and 0.2 mg/ml Proteinase K) and then subjected to the following thermal program: 65C for 10 minutes then 95C for 15 minutes. gRNA-targeted genomic loci were PCR dialed-out from genomic DNA using NEBNext High Fidelity Master Mix (NEB, M0541S) and primers that flank ~200 bps of the expected Cas9 cut site (PCR primers are listed in **Table S6**). PCR amplicons were gel extracted from a 2% agarose gel and submitted for Sanger Sequencing using forward and reverse PCR primers. Indel frequencies for each gRNA were quantified using TIDE analysis (http://tide.nki.nl) where the reference nucleotide sequence was derived from a mock-nucleofected control. The proportion of frameshifting indels was then determined from the total TIDE indel frequency estimates.

2.4.3 *In vitro* T cell differentiation

Following 36-48 hours post-nucleofection of Cas9 RNPs, 5,000 cord blood HSPCs were differentiated to T cells using the StemSpan T cell Generation Kit (Stem Cell Technologies, 09940) per the manufacturer's instructions with n=3 replicate differentiations per gRNA in 24-well plates. After 14 days of differentiation, cells were collected and 50,000 cells/sample were replated into new 24-well plates. Similarly, after 28 days of differentiation, cells were collected and 250,000 cells were replated into new 24-well plates for the final two weeks of differentiation. Total viable cell counts were performed on Days 14, 28 and 42 of differentiation using a TrypanBlue exclusion dye and a hemocytometer. Aliquots of cells were collected for flow cytometry on days 14, 28 and 42 of differentiation to assess lymphoid progenitors (CD45, CD5 and CD7) and mature lymphoid populations (CD45, CD4, CD8, CD3, CD56, CD5 and CD7). Cells were stained with either DAPI or propidium iodide as a viability marker and the following antibodies for flow cytometry: CD45 APC-Cy7 (BD Biosciences, 557833), CD4 PE-Cy5

(Beckman Coulter Immunotech, IM2636U), CD8 BV421 (BD Horizon, RPA-T8), CD5 BV510 (BD Biosciences, UCHT2), CD3 PE–Cy7 (BD Pharmigen, UCHT1), CD7 PE (BD Pharmigen, 555361), CD56 FITC (BD Pharmigen, 362545).

2.4.4 *In vitro* B cell differentiation

MS5 stromal cells were seeded onto gelatin-coated 24-well plates at a density of 10,000/well. Cells were allowed to grow for 48 hours in Myelocult H5100 (Stem Cell Technologies, 05150) supplemented with 100 ng/mL SCF (R&D Systems), 50 ng/mL TPO (Peprotech), 10 ng/mL FLT3L (R&D Systems), and 25 ng/mL IL-7 (R&D Systems) in order to condition the media prior to seeding hematopoietic cells. Following 36-48 hours post-nucleofection of Cas9 RNPs, 2,500 CD34+ cord blood HSPCs were seeded onto MS5 stroma with at least n=3 replicate differentiations per gRNA per time-point. Fresh medium with cytokines was supplied weekly over five weeks of differentiation. Wells were harvested at regular intervals throughout the differentiation by gentle trituration to obtain total cell counts and for flow cytometric assessment of hematopoietic differentiation. Cells were stained with propidium iodide and either a hematopoietic lineage antibody panel, comprising CD45 APC-Cy7 (BD Biosciences, 557833), CD19 PE (BioLegend), CD56 FITC (BioLegend, 362545), CD33 APC (BioLegend), CD14 PerCP (BD Biosciences), or a B-cell specific antibody panel, comprising CD45 APC-Cy7 (BD Biosciences, 557833), CD19 PE (BioLegend), CD20 PE-Cy5 (BD Biosciences), IgM BV510 (BD Biosciences, 563113).

2.4.5 *In vitro* NK cell differentiation

CRISPR/Cas9 RNPs were delivered to CD34+ umbilical cord blood cells, as described above. NK differentiation was initiated 48 hours post-RNP nucleofection using a StemSpan NK cell Generation Kit (Stem Cell Technologies, 09960), per the manufacturer's instructions with n=3 replicate differentiations per gRNA in 24-well plates. Cells were replated into new 24-well plates

after 14 days of differentiation and harvested for flow cytometry profiling, functional assays, and molecular assays after 28 days of differentiation.

2.4.6 Flow cytometry

Cells were stained with antibodies and a viability dye (DAPI or propidium iodide) at room temperature for 20 minutes in the dark for flow cytometric profiling, followed by a wash step with FACS buffer (PBS + 2% FBS). Data was collected on a Sony MA900 and analyzed using FlowJo. Antibodies for flow cytometry were used at 1:100 dilution, unless otherwise stated.

2.4.7 UNC0642 treatment during T cell differentiation

CD34+ umbilical cord blood cells (AllCells) were thawed from liquid nitrogen storage and T cell differentiation was immediately initiated with a StemSpan T cell Generation Kit (Stem Cell Technologies, 09940), as described above. Cells were treated with UNC0642 (Cayman Chemical, 14604) over the entire duration of the differentiation and fresh compound was supplemented every 3-4 days to the media. A range of doses were tested (10 nM, 50 nM, 100 nM, and 500 nM) alongside a 0.5% DMSO vehicle control with n=3 replicate differentiations per condition. Cells were collected after 28 days of differentiation to assess NK and T cell populations by flow cytometry.

2.4.8 Single cell -omics assays on CRISPR-perturbed lymphoid progenitors

Umbilical cord blood HSPCs were genome edited with Cas9 RNPs targeting *EHMT1*, *TRIM28*, and the *AAVS1* locus, as described previously. The CRISPR-edited HSPCs were then differentiated to lymphoid progenitors over 14 days using the StemSpan T cell Generation Kit (Stem Cell Technologies cat. no. 09940) with n=3 replicate differentiations per gRNA. The cells were harvested, washed with FACS buffer (PBS + 2% FBS), stained with DAPI for 10 minutes at room temperature, and sorted with a Sony MA900 FACS to isolate viable, single cells. FACS

sorted cells were centrifuged at 300g for 5 minutes at 4C, resuspended in FACS buffer and the concentration was determined with a hemocytometer. scRNA-Seq and scATAC-Seq targeting 5,000 cells per gRNA was performed using the Chromium NextGEM Single Cell 3' Reagent Kit v3.1 (10X Genomics) and Chromium NextGEM Single Cell ATAC Reagent Kit v1.1 (10X Genomics), respectively. scRNA-Seq libraries were barcoded with dual Illumina indices, whereas scATAC-Seq libraries were barcoded with single indices. The concentration of each final library was quantified using a High Sensitivity DNA Bioanalyzer Assay (Agilent). scRNA-Seq libraries were equimolar pooled and shallow sequenced on an Illumina MiniSeq to determine the number of high-confidence cell barcodes per library. The scRNA-Seq libraries were then renormalized and deep sequenced on a NextSeq 550 in a 28-10-10-44 read configuration. scATAC-Seq libraries were equimolar pooled and sequenced on a NextSeq 550 in a 34-8-16-34 read configuration.

2.4.9 Bulk RNA-Sequencing

RNA from at least 5,000 cells was extracted using a 2X TCL lysis buffer (Qiagen cat. no. 1070498). At least two technical replicates were prepared per sample using the SMART-Seq2 protocol [69], with some modifications. Briefly, RNA was purified from cellular lysate using 2.2X RNA SPRI beads (Beckman Coulter cat. no. A63987) then immediately reverse transcribed in the presence of a template switching oligo (Exiqon) with Maxima RNase H-minus RT (Thermo Fisher Scientific cat. no. EP0751) using a polyT primer containing the ISPCR sequence. Whole transcriptome amplification proceeded with KAPA HiFi HotStart ReadyMix using an ISPCR primer according to the following thermal program: 98C for 3 minutes, 21 cycles of 98C for 15 seconds, 67C for 20 seconds, and 72C for 6 minutes, and a final extension step of 72C for 5 minutes. The amplified cDNA was cleaned with 0.8X DNA SPRI beads (Beckman Coulter cat. no. B23318). Ten nanograms of amplified cDNA was tagmented at 58C for 10 minutes in a 10 uL reaction containing 2 uL of 5X tagmentation buffer (50 mM Tris-HCl, 25 mM MgCl2 pH 8.0),

2 uL of Tris Buffer (10 mM Tris-HCl, 1% Tween-20 pH 8), and 4 uL Nextera Tn5 transposase (Illumina). The reaction was stopped with 1% SDS and incubated at 72C for 10 minutes, then 4C for 3 minutes. The tagmented library was cleaned with 1X DNA SPRI beads followed by an indexing PCR with NEBNext High Fidelity polymerase to incorporate sample index barcodes and Illumina flow cell handles. The index PCR proceeded with the following thermal program: 72C for 3 minutes, 98C for 30 seconds, 12 cycles of 98C for 10 seconds, 60C for 30 seconds, 72C for 30 seconds, and a final extension step of 72C for 5 minutes. The final libraries were pooled, diluted and sequenced on a MiniSeq with a 75-cycle High Output Kit with the following read configuration: 36-8-8-38.

2.4.10 Bulk ATAC-Sequencing

NK cells were FACS sorted following *in vitro* differentiation from CD34+ umbilical cord blood and at least 50,000 cells per sample were used as input for ATAC-Seq. Nuclei isolation, tagmentation and library construction were followed as described in the Omni-ATAC-Seq protocol [228]. The concentration of the final ATAC-Seq libraries was quantified using a High Sensitivity DNA Bioanalyzer Assay (Agilent) in the size range of 100-1000 bp. The libraries were then equimolar pooled and shallow sequenced on a MiniSeq to determine the quality of the libraries. The ATAC-Seq libraries were processed with the PEPATAC pipeline [229] to determine the number of deduplicated, aligned reads to the hg38 genome and pertinent QC metrics, such as TSS enrichment and fragment length distributions. The libraries were then renormalized, pooled and sequenced on a NextSeq 500 with a 150-cycle High Output Kit v2 with the following read configuration: 76-8-8-75.

2.4.11 NK functional assays

In vitro derived NK cells were collected on D4+28, washed with FACS buffer, and then cultured overnight in RP-10 medium (RPMI-1640 supplemented with 10% FBS, 1×

penicillin/streptomycin, 2 mM L-glutamine, and 7.5 mmol HEPES) with 1 ng/mL recombinant human IL-15 (Miltenyi). K562 cells were cultured in RP-10 medium and labeled with 5 μM of CellTrace Violet (Thermo Fisher Scientific) in PBS for 20 minutes at 37°C. K562 were washed twice with RP-10 and co-cultured at various effector:target ratios. To measure NK cell cytotoxicity, NK cells and target cells were co-cultured for 4 hours, then stained with 2 μL of PE-Annexin V (Biolegend) and 2 μL of 7-AAD (BD Biosciences) in 50 μL Annexin V binding buffer (Biolegend) for 15 minutes at room temperature. To measure intracellular IFNγ and degranulation, NK cells were stimulated with recombinant human IL-12 (R&D), recombinant human IL-18 (R&D), or were co-cultured target cells for 1 hour, followed by the addition of 0.2 μL BD GolgiPlug (BD Biosciences), 0.13 μL BD GolgiStop (BD Biosciences) and 1 μL of APC-CD107a (Biolegend). After an additional 5 hours of co-culture, cells were stained for intracellular IFNγ using BD Cytofix/Cytoperm (BD Biosciences). Cells were acquired using BD LSRFortessa and analyzed using FlowJo.

2.4.12 Comparison of ABC model modifications to genetic perturbation data

We investigated various modifications to the ABC model and evaluated the enhancer-gene predictions against the results of genetic perturbation experiments on the candidate enhancer elements. First, we utilized genetic perturbation data in K562s (Supplementary Table 6a from Fulco, et al. 2019) and defined true-positive enhancer-gene pairs as previously described [185,186]. Next, we utilized K562 epigenomic data (**Table S1**) and a 10-cell type averaged HiC dataset [185] to generate enhancer-gene predictions using the ABC model on tissue-specific, non-ubiquitously expressed genes. We also evaluated different definitions of enhancer "Activity" in the model by only using chromatin accessibility data, as well as approximating HiC data with a power-law estimate. We then intersected each experimentally tested enhancer-gene pair with the predictions from the modified ABC models and calculated precision-recall statistics using the `precrec` package in R [230]. Furthermore, we evaluated the predictive performance of a

chromatin accessibility-only ABC model against an expanded compendium of genetic perturbations in other hematopoietic cell types. We utilized ATAC-Seq data on GM12878, BJAB, Jurkat, and THP1 cell lines +/- stimulation (**Table S1**) to generate ABC maps and calculated precision-recall statistics using genetic perturbation data in the same cell lines from Nasser, et al. 2021 (Supplemental Table 5 from their manuscript).

2.4.13 Curation of public chromatin accessibility data for ABC model predictions

We curated published chromatin accessibility data on primary, FACS sorted cells to capture all major cell types and states within the human hematopoietic system. To be consistent across different studies, we primarily utilized ATAC-Seq to measure chromatin accessibility. We downloaded raw fastq files for HSPC populations from [191], erythroid subpopulations corresponding to temporal stages of differentiation from [195], myeloid and plasmacytoid dendritic cells from [194], precursor dendritic cells and megakaryocytes from [192], macrophages from ENCODE, and immune cells +/- stimulation from [196]. Each hematopoietic cell type, designated by a FACS immunophenotype (see **Table S2**) was represented by at least 3 different human donors within our final curated dataset. In the case of the Calderon, et al. dataset, we excluded ATAC-Seq samples with a reported TSS score <6 (9/175 samples). The comprehensive list of samples utilized for ABC predictions is outlined in **Table S2**.

2.4.14 Uniform processing of chromatin accessibility data

We implemented the PEPATAC pipeline [229] to uniformly process and align raw ATAC-Seq data to the hg38 genome. All fastq files were first trimmed to remove Illumina adapter sequences using Skewer with command line options “-f sanger -t 20 -m pe -x”. We then performed pre-alignments with Bowtie2 to remove reads mapping to the mitochondrial genome, alpha satellite repeats, Alu repeats, and ribosomal DNA repeats using “-k 1 -D 20 -R 3 -N 1 -L 20 -i S,1,0.50 -X 2000 -rg-id” options. Bowtie2 was used to align the remaining reads to the hg38 genome using

"--very-sensitive X 2000 --rg-id" options. We isolated uniquely mapped reads with samtools and options "-f 2 -q 10 -b -@ 20". Samblaster was used to mark duplicate reads, resulting in the final aligned, deduplicated BAM file that was used in all downstream analyses. For DNase-Seq data, we downloaded hg38-aligned BAM files from the ENCODE data portal.

2.4.15 Generation of enhancer-gene predictions using the ABC model

We generated enhancer-gene maps for each hematopoietic cell type in our dataset using the ABC model (https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction) as previously described by Nasser, et al. 2021 with some modifications. First, we called peaks with MACS2 using the ATAC-Seq sample with the highest TSS score amongst all donor samples of a hematopoietic cell type. To enable comparisons between ATAC-Seq and DNase-Seq data, we used nucleosome-free ATAC-Seq reads (outputted from the PEPATAC pipeline) for defining candidate enhancers and quantifying enhancer activity. Candidate enhancer elements overlapping the ENCODE hg38 blacklist (https://www.encodeproject.org/files/ENCFF356LFX) were excluded. We downloaded gene annotations from the UCSC genome browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/refGene.txt.gz) to define TSSs and gene bounds for the model. Next, we only used chromatin accessibility data to quantify enhancer "Activity" (supported by our previous analysis) and samples from different donors of the same cell type were included as replicate experiments in the model. Chromatin accessibility in each candidate enhancer element was quantile-normalized to the distribution observed in K562s. We also used a power-law estimate as the input for the "Contact" term in the model. ABC scores were then computed for each candidate enhancer-gene pair within a 5 Mb window of a gene's TSS and we retained enhancer-gene predictions with an ABC score > 0.02. We only made enhancer-gene predictions for non-ubiquitously expressed genes, as defined in [185]. All ABC predictions are reported in **Table S3**.

2.4.16 Overlap of ABC enhancers with chromHMM regulatory states

We verified the predictions from our modified ABC model by intersecting the predicted ABC enhancers from GM12878, BJAB, Jurkat, and THP1 cell lines +/- stimulation with matched ChIP-Seq defined chromHMM regulatory states from the Roadmap Epigenomics Project. We downloaded chromHMM 15-state models and determined the number of ABC enhancers that overlapped each chromHMM state for each cell type. We calculated the expected background as the coverage of a chromHMM state in the human genome and determined the statistical significance of ABC enhancers overlapping a state using a binomial test in R.

2.4.17 Construction of a pan-hematopoiesis ABC enhancer peakset and chromatin accessibility counts matrix

To facilitate global analyses, we determined the unique, non-redundant set of ABC enhancers across all ABC predictions in our dataset and constructed a chromatin accessibility counts matrix over these regions for all samples. To do this we first shrunk the ABC enhancer regions by 150 bps on either side using `bedtools slop` and then determined the unique set of ABC enhancers per hematopoietic cell type. We concatenated all unique ABC enhancers from each cell type into one bed file and used `bedtools merge` to create a non-overlapping list of ABC enhancer regions. This list represents the unique set of ABC enhancers across all cell types in our dataset, which we term a pan-hematopoiesis ABC enhancer peakset. Next, we used hg38-aligned, deduplicated BAM files and `bedtools multicov` to count chromatin accessibility reads within this peakset for all samples in our dataset. We then constructed an ABC enhancers x samples counts matrix within R.

2.4.18 Clustering and visualization of pan-hematopoiesis ABC enhancers

Prior to clustering, we CPM normalized the ABC enhancer counts matrix using the edgeR package: `cpm(matrix, log=TRUE, prior.count=5)`. To correct for technical covariates, we then regressed out differences in data quality by using the TSS score as a covariate with the `removeBatchEffect()` function in the limma package. Finally, we performed quantile normalization using `normalize.quantiles()` from the preprocessCore package. We used the top 120,000 ABC enhancer by row variance using `rowVars()` in the matrixStats package for clustering. PCA was then performed using `prcomp_irlba()` on the first 50 PCs without scaling. We then calculated the Euclidean distance with `dist()` across samples within the first 9 PCs (since the first 9 PCs explained approximately 80% of the variance). We then performed density clustering using `densityClust()` to get a decision plot (**Figure S2D**). We extracted cluster assignments for rho = 0.8 and delta = 0.08 with `findClusters()`. We visualized the cluster assignments using *t*-SNE: `Rtsne(perplexity=40, theta=0, max_iter=1000, pca=TRUE)`. Finally, we created a confusion matrix to understand which samples reside in which clusters and visualized the matrix with the pheatmap package in R.

2.4.19 Transcription factor footprinting within ABC enhancers

We used the ChrAccR R package (v.0.9.17) to perform transcription factor footprinting analysis using ATAC-Seq data. We created a DsATAC object using hg38-aligned, deduplicated BAM files from the PEPATAC pipeline and ABC enhancers as the input peakset. Biological replicates for each cell type (e.g. different donor samples) were merged with `mergeSamples(dsa, mergeGroups="bio_group", countAggrFun="sum")`. We then performed footprinting analysis specifically within ABC enhancer peaks using getMotifFootprints() and PWMs from the JASPAR motif database.

2.4.20 Visualization of ABC enhancer-gene predictions

We visualized ATAC-Seq sequencing tracks with pyGenomeTracks (https://github.com/deeptools/pyGenomeTracks). We utilized bigwig files, outputted from the PEPATAC pipeline, containing Tn5 offset-corrected insertion sites. We performed reads-in-peaks normalization as described in [231] when comparing multiple samples. ABC enhancer-gene links were visualized as arcs weighted by the ABC score and centered on the predicted gene's TSS and the midpoint of the ABC enhancer.

2.4.21 Processing of the RepeatMasker transposable element reference database

We downloaded RepeatMasker TE annotations within the hg38 human reference genome from the UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/rmsk.txt.gz). We only considered TEs belonging to DNA, LINE, LTR, RC, Retroposon, and SINE classes and residing within standard chromosomes. Furthermore, we filtered out ambiguous TE annotations and all tRNA related families. We constructed a GRanges object in R of genomic intervals for each TE family and used it for all downstream analyses.

2.4.22 Enrichment of transposable elements in ABC enhancers

We utilized the GIGGLE framework [197] to generate indices of genomic intervals for TE families. The non-redundant set of ABC enhancers for each cell type were queried against the TE family database using the GIGGLE search function with a genome size of 3099922541 bp. Significantly enriched TE families were defined as those TE families with at least 20 overlaps with ABC enhancer regions, an odds ratio > 2.5 and a -log10(p-value) < 0.01 in at least one cell type.

2.4.23 Transcription factor occupancy analysis within ABC enhancers

We predicted TF occupancy within TE-containing ABC enhancers using the TOBIAS transcription factor occupancy framework (https://github.com/loosolab/TOBIAS), as described in [198]. In brief, we merged ATAC-Seq BAM files from various donors belonging to the same cell type and performed bias correction of the ATAC-Seq signal with `ATACorrect` across all peaks in the pan-hematopoiesis ABC enhancer peakset. We then calculated footprint scores on the bias-corrected ATAC-Seq signal with `FootprintScores` and estimated transcription factor occupancy using `BINDetect` and the position frequency matrix of the TF motif. In the case of predicting AP-1 occupancy, we identified differential binding with `BINDetect` between resting and activated T cell states within TE-containing ABC enhancers.

2.4.24 TE quantification from RNA-Seq and differential expression analysis

RNA-Seq data on FACS-sorted hematopoietic populations from [191] was downloaded and reanalyzed. First, fastq files were mapped to the hg38 reference transcriptome with kallisto. The resulting BAM files were used with TEtranscripts [205] (https://github.com/mhammell-laboratory/TEtranscripts) to quantify the expression of TEs. Curated GTF files on TE annotations were downloaded from the Hammell Lab website for TEtranscript quantification. Differential expression was subsequently performed with DESeq2 1.10.1 [232].

2.4.25 Calculation of gene module scores for regulators of heterochromatin from single cell RNA-Seq data

We reanalyzed scRNA-Seq data on healthy human hematopoiesis previously reported in [233]. We downloaded the processed scRNA-Seq data from the publication's GitHub repository (https://github.com/GreenleafLab/MPAL-Single-Cell-2019) and imported it as a Seurat object in R. We scored gene sets to generate a new aggregate expression level corrected for the

background expression of each cell, as described previously [234]. Gene sets were downloaded from the Gene Ontology database and scored using the `AddModuleScore` function in Seurat. Gene module scores were visualized across the transcriptional clusters originally annotated by [233].

2.4.26 Processing scRNA-Seq data on *in vitro* derived lymphoid progenitors

scRNA-Seq libraries were demultiplexed and fastqs were generated with cellranger mkfastq (10x Genomics, version 7.0.1). Reads were aligned to the hg38 reference genome and quantified using cellranger count (10x Genomics, version 7.0.1). Downstream analysis was performed using Seurat (version 3.2.3) within R. Count matrices from the cellranger output were preprocessed by filtering for cells and genes (percent mitochondrial reads < 20%, at least 1000 genes detected/cell, and less than 20,000 UMIs/cell). Normalization and variance stabilization of the count data was performed using sctransform and regressing out the proportion of mitochondrial reads and cell cycle phase. We subsequently performed PCA and used the first 20 PCs to run UMAP analysis. We identified transcriptional clusters with FindClusters() at a resolution of 0.2. Differentially expressed genes were determined for each cluster by a Wilcoxon rank-sum test with an FDR cutoff of 0.01 and log2(fold-change) of at least 0.25.

2.4.27 Processing scATAC-Seq data on *in vitro* derived lymphoid progenitors

Sequencing libraries were demultiplexed and fastqs were generated with cellranger atac mkfastq (10x Genomics, version 1.2.0). Reads were aligned to the hg38 reference genome and quantified using cellranger count (10x Genomics, version 1.2.0). All downstream analysis was performed with ArchR (version 1.0.2) in R. We performed quality control filtering and excluded cells that had a TSS enrichment less than 5 and fewer than 1000 aligned fragments. We also excluded putative doublets with filterDoublets(). We reduced dimensionality with two rounds of latent semantic indexing (LSI) with a cluster resolution of 0.2. We clustered on the iterative LSI

dimensions using Seurat's SNN graph clustering function at a resolution of 0.3. We subsequently added a UMAP with addUMAP() and the default parameters. To help annotate scATAC-Seq clusters, marker genes were identified with Gene Scores using getMarkerFeatures() and accounting for biases in TSS enrichment and the number of aligned fragments. Pseudobulk replicates were created per gRNA condition per cluster followed by peak calling with MACS2. A union peakset was created using an iterative overlap removal method implemented in addReproduciblePeakSet(). TF activity was inferred with chromVAR using the clustered TF motif archetype collection from [207].

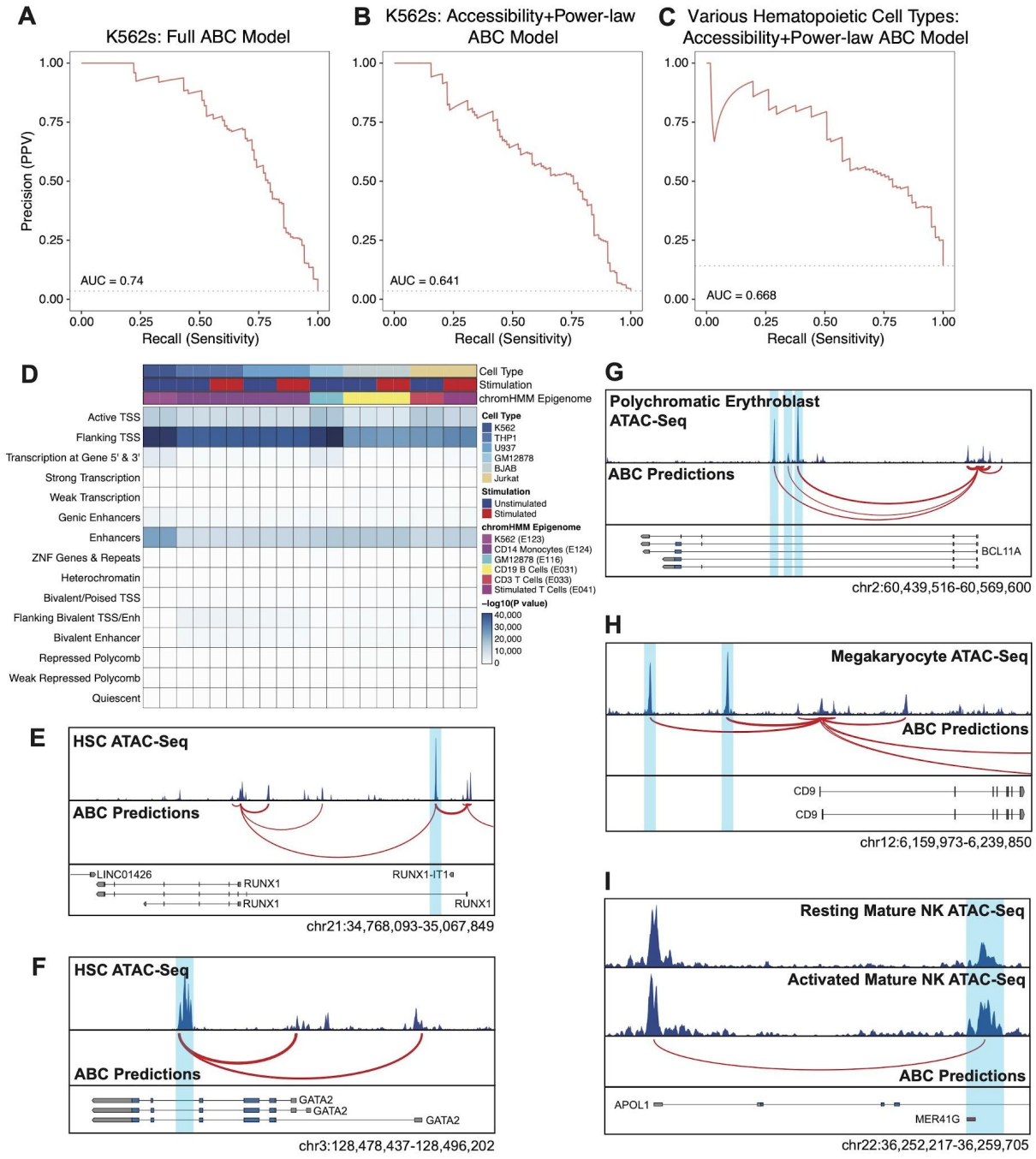
2.4.28 Processing and analysis of bulk ATAC-Seq on *in vitro* derived NK cells

ATAC-Seq libraries were demultiplexed with bcl2fastq and processed with the PEPATAC pipeline, as described above. A consensus peakset across all samples was generated using the PEPATACr package in R and a count matrix was constructed using bedtools multicov to count ATAC-Seq reads from hg38-aligned, deduplicated BAM files within the consensus peakset. Using the ATAC-Seq count matrix as an input, we inferred transcription factor activity with chromVAR and CIS-BP motif matches within these peaks from motifmatchr. GC bias-corrected deviations were computed using the chromVAR `deviations` function. A comparable analysis was performed using TE families rather than motifs to determine derepressed TEs within knockout conditions compared the AAVS1 gRNA control. Statistical significance of TE accessibility deviations was computed with the `differentialDeviations` function from chromVAR and FDR-corrected p-values < 0.01 were considered significant. Differentially accessible peaks were determined with DESeq2 using the raw ATAC-Seq count matrix and statistically significant peaks were identified at an independent hypothesis weighting (IHW) value of 0.01. Differentially accessible peaks were then clustered with k-means clustering and used as the input to GREAT for annotation of biological processes.

2.4.29 Processing and analysis of bulk RNA-Seq on *in vitro* derived NK cells

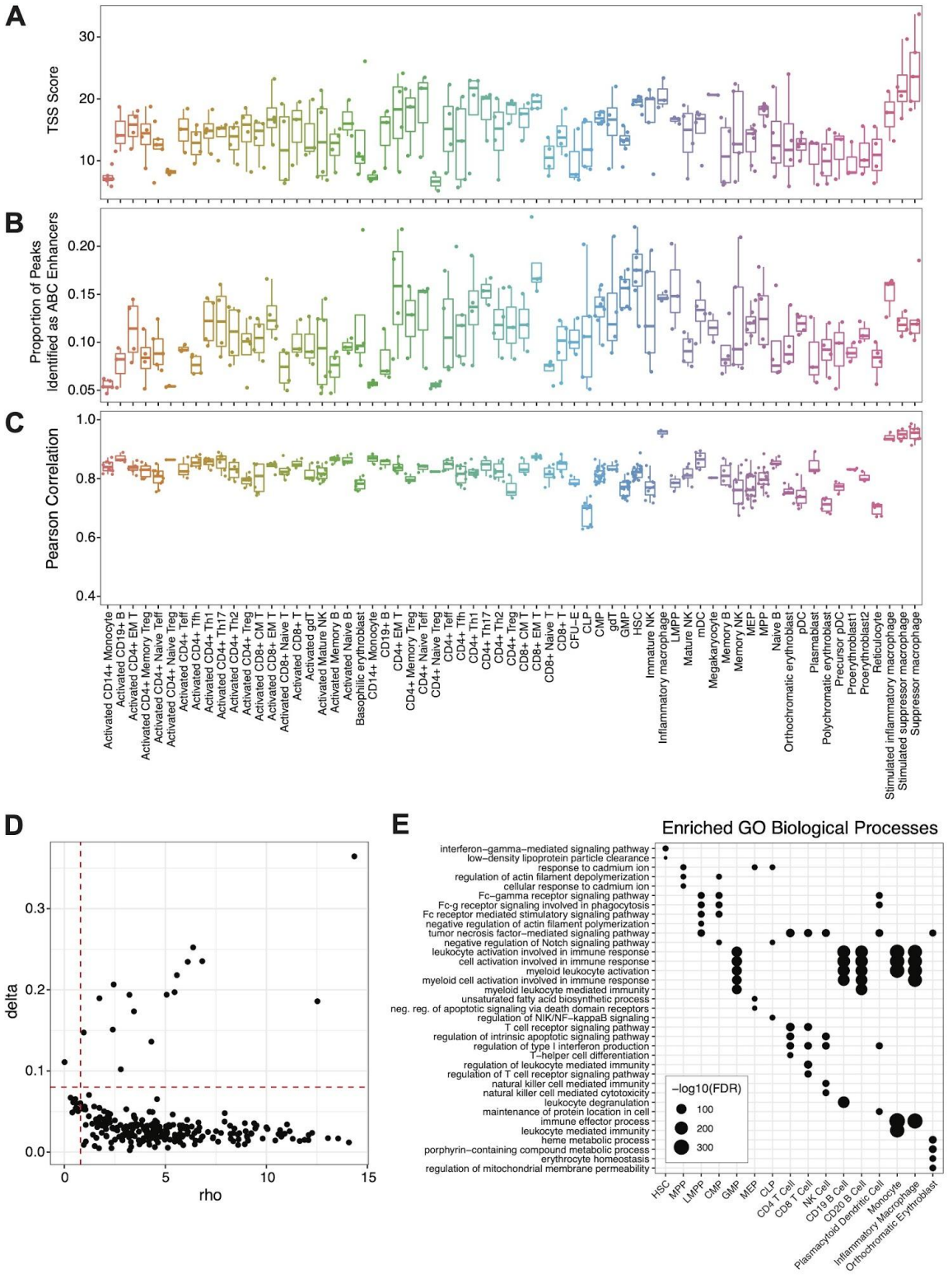
RNA-Seq paired end reads were pseudo-aligned with kallisto [85] to the hg19 reference transcriptome. Transcript-level abundance estimates were imported into R with the tximport package and constructed into a gene-summarized count and abundance matrices for all samples. Differential gene expression analysis was performed using DESeq2 [232] on the estimated count matrix. Statistically significant genes varying between a knockout population and the AAVS1 gRNA control population were identified at an FDR<0.05.

2.5 SUPPLEMENTAL FIGURES



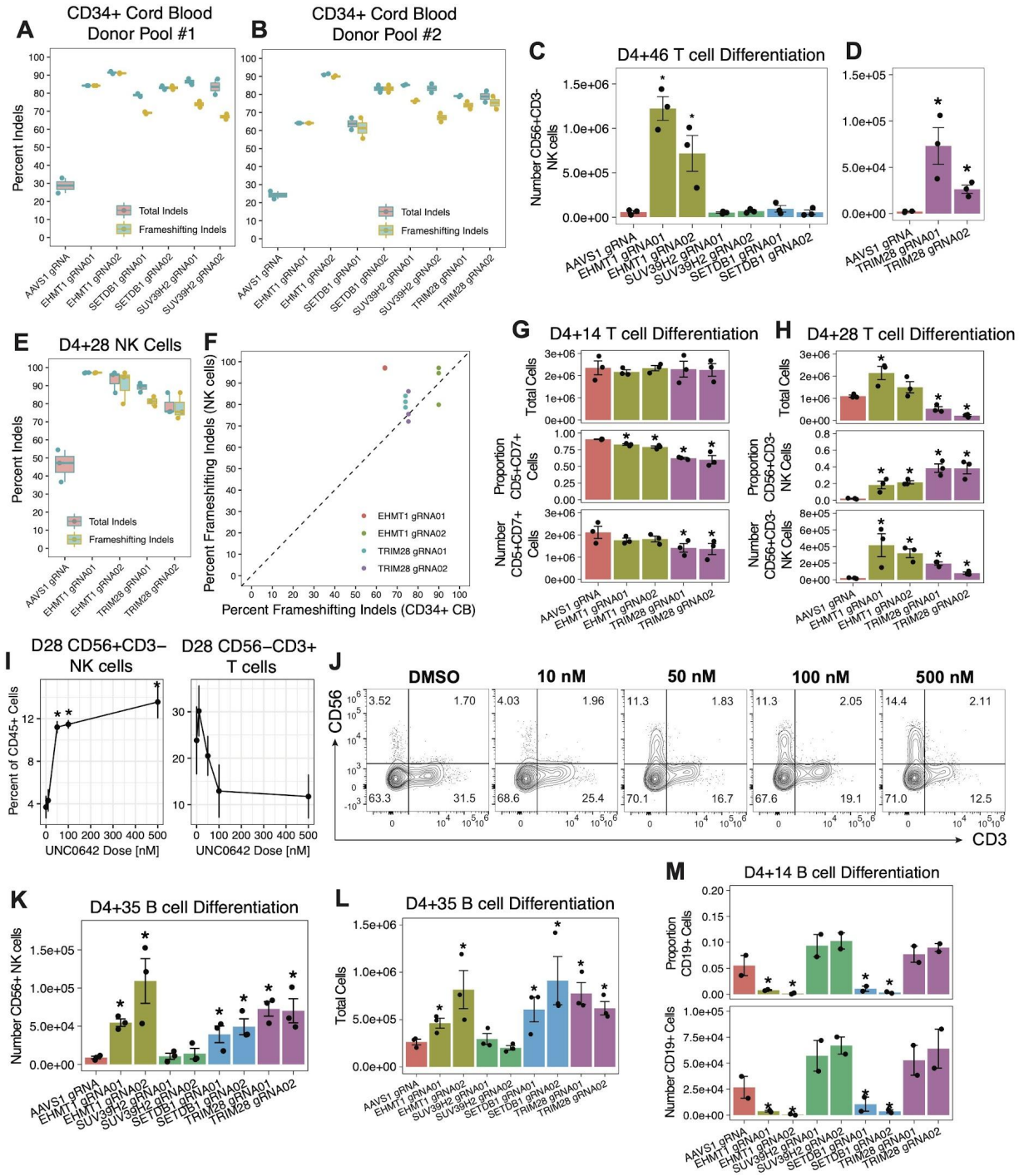
Supplementary Figure 2.1. related to Figure 2.1.
(A-C) Precision-recall curves comparing ABC enhancer-gene predictions to experimental CRISPR data in K562s (A-B) or an expanded compendium of hematopoietic cell lines: GM12878, THP1 +/- stimulation, Jurkat +/- stimulation, and BJAB +/- stimulation (C). CRISPRi-FlowFISH screen data was derived from [185] for analysis in (A-B) and from [186] for analysis in (C). The full ABC model in (A) refers to measurements of enhancer activity with H3K27ac ChIP-Seq and chromatin accessibility, and contact frequency measurements from a 10-cell type

averaged HiC dataset, as originally reported in [185]. The modified ABC model in (B-C) utilizes only chromatin accessibility for enhancer activity and a power-law of genomic distance to approximate HiC.
(D) Heatmap of the enrichment of predicted enhancers from the modified ABC model within ChIP-Seq-defined chromHMM states from the Roadmap Epigenomics Project. Enrichment was determined with a binomial test. The chromHMM epigenomes used for the enrichment analysis are noted along the top and the ATAC-Seq data used to make the ABC predictions was derived from [186].
(E-I) Normalized ATAC-Seq sequencing tracks depicting ABC enhancer-gene linkages across various hematopoietic cell types. Known and experimentally verified enhancers are highlighted with blue shading. The thickness of the enhancer-gene link is scaled by the ABC score. The genomic regions visualized are noted in hg38 coordinates below each sequencing track.



Supplementary Figure 2.2. related to Figure 2.1.

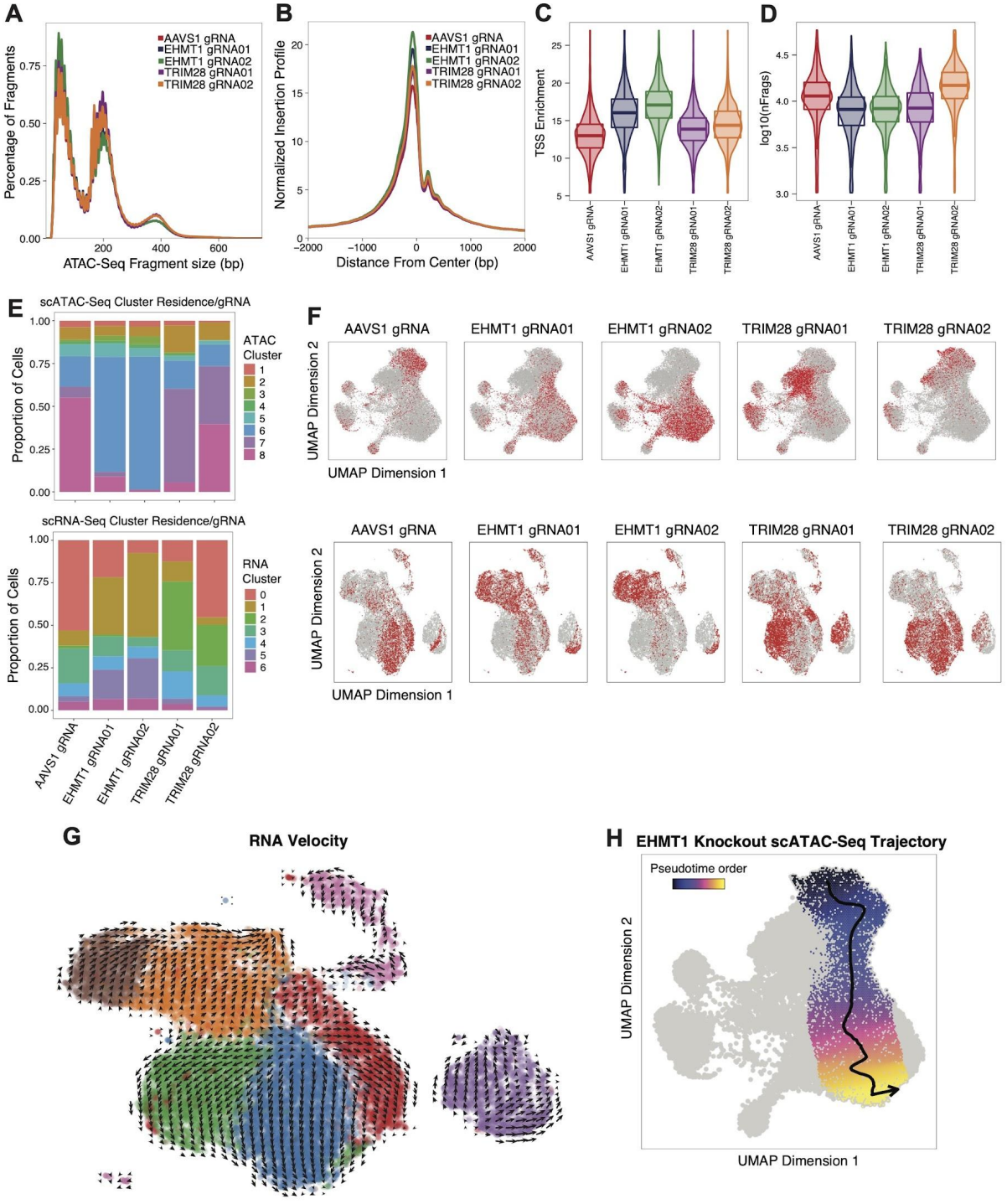
(A) Box and whisker plots of TSS enrichment scores for all samples within each hematopoietic cell type. Each dot represents an individual ATAC-Seq sample. The hinges represent the 25th to 75th percentile.
(B) Box and whisker plots of the fraction of ATAC-Seq peaks within a sample identified as ABC enhancers for each hematopoietic cell type. Each dot represents an individual ATAC-Seq sample. The hinges represent the 25th to 75th percentile.
(C) Box and whisker plots of the Pearson correlations of chromatin accessibility in the pan-hematopoiesis ABC enhancer peakset between all samples (technical replicates and different donors) in each hematopoietic cell type. The hinges represent the 25th to 75th percentile.
(D) The decision metrics used for the density clustering methodology, where rho is the number of points that are closer than the cutoff distance to a given point and delta is the distance between a given and any other point with higher density. Clusters were determined with rho = 0.8 and delta = 0.08.
(E) Dot plot of enrichment of GO Biological Processes within ABC-linked genes for each hematopoietic cell type noted. Enrichment p-values, which were determined by a binomial test were FDR corrected and only terms with an FDR < 0.01 are plotted.



Supplementary Figure 2.3. related to Figure 2.4.

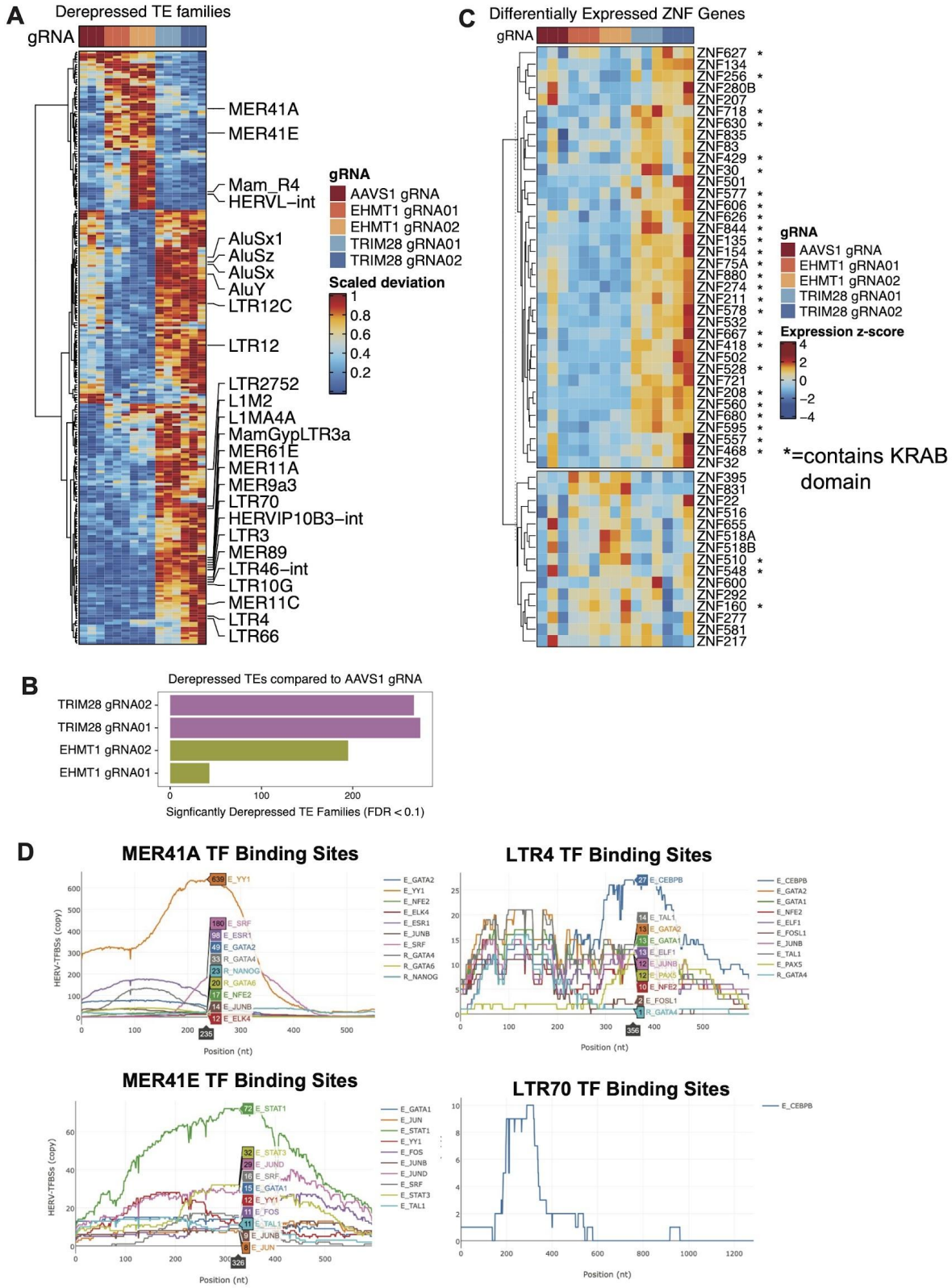
(A-B) Quantification of indels by TIDE analysis for each CRISPR/Cas9 RNP-mediated knockout in CD34+ umbilical cord blood HSPCs. An aliquot of the population of CD34+ cord blood cells were collected for gDNA extraction two days following nucleofection of CRISPR/Cas9 RNPs. The indel locus was PCR dialed-out from the genome (**Table S6**) and Sanger sequenced. Replicates represent TIDE quantification from forward and reverse Sanger sequencing traces of the indel amplicon.

(C-D) Quantification of the absolute numbers of CD56+CD3- NK cells on D4+46 of *in vitro* T cell differentiation across all gRNA conditions (n=3 replicate differentiations per gRNA). * = $p < 0.05$ in comparison to the *AAVS1* gRNA using a two-tailed Wilcoxon rank-sum test.
(E) Quantification of indels in an analogous manner as described in (A-B) on CD56+CD3- NK cells on D4+28 of *in vitro* T cell differentiation. Replicates represent n=3 independent differentiations per gRNA.
(F) Concordance of frameshifting indel frequencies quantified in the starting CD34+ HSPC population versus frameshifting indel frequencies within D4+28 NK cells.
(G-H) Quantification of total viable cells, the proportion of CD5+CD7+ lymphoid progenitors (G) or CD56+CD3- NK cells (H) as determined by flow cytometry, and the absolute cell counts of each population across all gRNA conditions from *in vitro* T cell differentiation. Replicates represent n=3 independent differentiations per gRNA. * = $p < 0.05$ in comparison to the *AAVS1* gRNA using a two-tailed Wilcoxon rank-sum test.
(I) The proportion of CD56+CD3- NK and CD56-CD3+ T cells within the CD45+DAPI- population, quantified on D28 of *in vitro* T cell differentiation starting from CD34+ umbilical cord blood HSPCs treated continuously with UNC0642. Error bars represent the standard deviation across n=3 independent differentiations per dose of UNC0642. * = $p < 0.05$ compared to the DMSO control using a two-tailed Wilcoxon sum-rank test.
(J) Representative flow cytometry plots of data summarized in (I). The plots are pre-gated on CD45+DAPI- cells.
(K-L) Quantification of the absolute numbers of CD56+CD3- NK cells (K) and total viable cells (L) on D4+35 of the MS5 co-culture assay across all gRNA conditions (n=3 replicate differentiations per gRNA). * = $p < 0.05$ in comparison to the *AAVS1* gRNA using a two-tailed Wilcoxon rank-sum test.
(M) Quantification of the proportion (top) and absolute cell counts (bottom) of CD19+ B cell progenitors on D4+35 of the MS5 co-culture assay. Replicates represent n=2 independent differentiations per gRNA. * = $p < 0.05$ in comparison to the *AAVS1* gRNA using a two-tailed Wilcoxon rank-sum test.



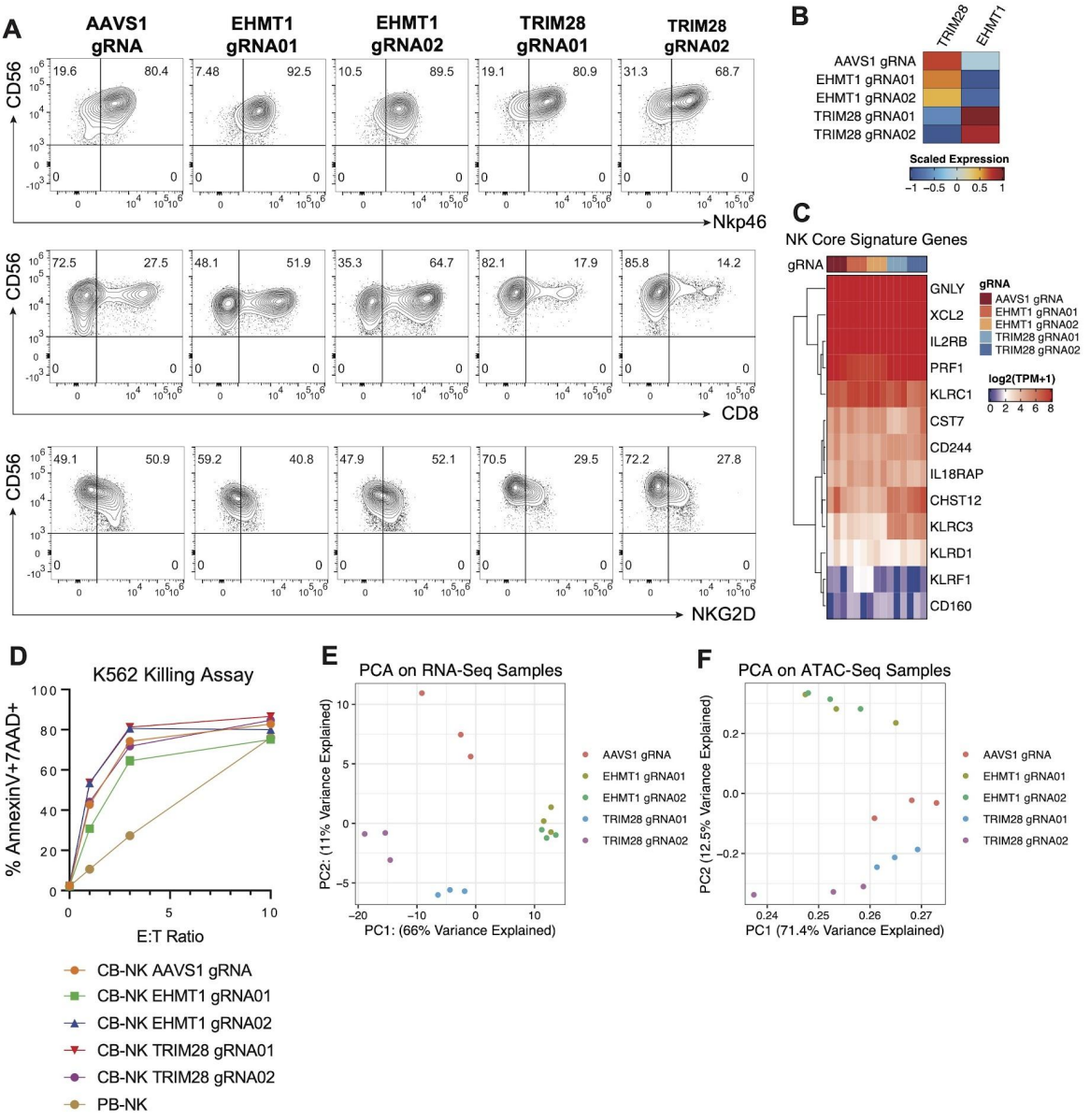
Supplementary Figure 2.4. related to Figure 2.5.
(A) Aggregated scATAC-Seq fragment size distributions for each gRNA library demonstrating sub-, mono- and multi nucleosome spanning fragments.
(B) Enrichment of ATAC-Seq accessibility +/-2kb of transcription start sites for each gRNA library.
(C) Violin and box-whisker plot of the normalized TSS enrichment for each single cell passing quality control filters per gRNA library.

(D) Violin and box-whisker plot of the number of total aligned fragments for each single cell passing quality control filters per gRNA library.
(E) Proportion of cells from a gRNA condition residing within scATAC-Seq (top) or scRNA-Seq (bottom) clusters.
(F) UMAP projections of scATAC-Seq (top) or scRNA-Seq (bottom) and cells belonging to each gRNA condition are highlighted.
(G) RNA Velocity analysis (steady-state model) projected onto the scRNA-Seq UMAP projection.
(H) Pseudotime trajectory representation of the divergence of EHMT1 gRNAs from AAVS1 gRNA control cells using scATAC-Seq data and overlaid on the UMAP projection of single cells.



Supplementary Figure 2.5. related to Figure 2.6.

(A) ATAC-Seq chromatin accessibility of TE families that are significantly derepressed in either EHMT1 or TRIM28 gRNA conditions compared to the AAVS1 gRNA control at D4+28 of NK differentiation. The color scale corresponds to a scaled chromVAR deviation.
(B) Quantification of the number of significantly derepressed TE families per gRNA in comparison to the AAVS1 gRNA control (FDR < 0.1) at D4+28 of NK differentiation. n=3 replicate differentiations per gRNA.
(C) Gene expression heatmap of zinc finger proteins that are differentially expressed in EHMT1 or TRIM28 gRNA conditions compared to the AAVS1 gRNA control. Differentially expressed genes were determined with DESeq2 and defined as log2(fold-change) > 1 and FDR < 0.05. The (*) designates whether the zinc finger is annotated to contain a KRAB domain in the Uniprot database. n=3 differentiation replicates per gRNA.
(D) Number of transcription factor binding sites mapped on each consensus position of the TE. The x-axis indicates nucleotide positions of the TE family consensus sequence, and the y-axis indicates number of TE copies harboring the transcription factor binding sites at each position.



Supplementary Figure 2.6. related to Figure 2.6.

(A) Expression of NK surface markers assayed by flow cytometry on D4+28 of *in vitro* NK differentiation. Plots are pre-gated on viable CD45+CD56+ cells.

(B) Scaled expression of EHMT1 and TRIM28 across all gRNA conditions demonstrating that the knockouts result in transcriptional attenuation of their respective target gene.

(C) Expression of a core NK signature gene set, derived from [213] across all gRNA conditions. n=3 replicate differentiations per gRNA.

(D) K562 killing assay using 0:1, 1:1, 3:1 and 10:1 mixtures of *in vitro* derived NK cells from all gRNA conditions to K562 cells. Co-cultures were incubated for 4 hours and then Annexin V+7AAD+ cells were assessed by flow cytometry.

(E-F) Principal-component analysis (PCA) plots of RNA-Seq gene expression (left) and ATAC-Seq chromatin accessibility profiles (right) for all gRNA conditions on D4+28 of *in vitro* NK differentiation and n=3 independent differentiation replicates per gRNA.

3 Live-cell transcriptomics via virus-like particles

Authors: Jacob Borrajo*, Mohamad Ali Najia*, Anna Le, FuNien Tsai, George Q. Daley, Paul C. Blainey

Status: This manuscript is in preparation to be submitted to a journal

Contributions: J.B. and P.C.B. conceived the original idea. J.B., M.N., A.L., and F.T. planned and performed experiments. J.D.B. and M.N. performed computational analysis. J.D. and M.N. prepared the manuscript with input from all authors. G.Q.D and P.C.B. supervised the research and provided funding.

3.0 ABSTRACT

Measuring the transcriptional states of living systems provides insight into their biological status and associated molecular mechanisms. However, non-destructive methods capable of retrieving and monitoring transcriptome-wide RNA information at multiple time points from the same living cells have yet to be developed. We overcame this limitation with a genetically encodable technology by repurposing the Gag polyprotein from murine leukemia virus (MLV) to create engineered virus-like particles (VLPs) that can export RNA molecules from cells for readout in a process we term cellular “self-reporting.” We rationally engineered Gag by adding poly(A)-binding domains to enhance RNA detection, and multiplexed cellular self-reporting from live co-cultures using VLPs bearing orthogonal epitopes for affinity purification. By using this simple and scalable approach, we collected quantitative transcriptome-wide RNA information with minimal perturbation from diverse human cell types, including epithelial, fibroblast, and induced pluripotent stem cells. Finally, we demonstrated that live-cell self-reported transcriptome measurements faithfully detect temporal changes in gene expression programs upon TNFα stimulation in HT1080 cells. Self-reporting of transcriptional states with engineered VLPs

enables facile live-cell transcriptome-wide measurements from the same biological samples over time. This simple molecular approach offers access to true time series data collection using established RNA-seq library construction, sequencing procedures, and analysis procedures.

3.1 INTRODUCTION

Biological research is rooted in time series analyses of living organisms and cells. The utility of light microscopy was dramatically enhanced by the introduction of genetically encodable non-toxic fluorescent proteins, which established an important interface between the dynamic analysis of live cell activity and the burgeoning field of molecular biology[2,4]. In today's genomic era, high-throughput RNA measurement using RNA sequencing (RNA-seq) is indispensable for the molecular characterization of cell states and functions across the life sciences[6,8,10]. However, RNA-seq has been applied almost exclusively as a destructive method in which biological samples are lysed for RNA extraction or transcripts are physically retrieved from a small number of cells via invasive instrumental procedures[11,13,15,17]. Destructive RNA-seq methods cannot produce true time series data from the same sample, precluding comparison of such measurements at different time points. These limitations hobble the ability of RNA-seq to directly characterize dynamic functional activity or state changes in cells, and prevent analysis of cells that are not physically accessible for destructive analysis. The popularity of computational methods that indirectly infer the time evolution of transcriptional dynamics from destructive "snapshot" data and recognition of the limits of such "pseudotime" approaches highlight the need for non-destructive RNA-seq methods[19,21,23,25,27,29,31]. To record RNA levels from the same cells over time, several groups have pioneered optical methods that use endogenous or exogenous probes to monitor a few native or tagged target transcripts at a time[33,35,37,39,41,43,45]. More recently, molecular recording methods have allowed information about the age of transcripts or prior transcriptional states to be encoded in the DNA or RNA of living systems for readout at a single end-point[47,49,51,53]. While promising, these methods require a prior transcriptional hypothesis, provide only summary-level information, have limited temporal dynamic ranges, and do not retain the living biological sample after the measurement.

Inspired by exosome and viral biology, we sought to overcome limitations of conventional RNA-seq by engineering a synthetic RNA export pathway that enables mammalian cells to "self-report" transcriptional states in real-time. Exosomes are physiological extracellular vesicles produced by cells that contain components of cell membranes and cytosol (including RNA) and can transport molecules between cells[55]. In contrast, retroviruses have evolved the ability to specifically package and export their genomic RNA in addition to delivering and propagating their RNA genome to complete their life cycle. Retroviral export pathways are an attractive substrate for engineering an RNA export system in particular, as a self-reporting technology based on select retroviral proteins could be orthogonal to host biology and activated in the time, place, and strength desired by researchers to achieve tailored reporting of transcriptional information. Specifically, we hypothesized that by repurposing a retroviral RNA export pathway, we could continuously sample RNA from living cells to enable transcriptome-wide non-destructive RNA-seq measurements.

3.2 RESULTS

3.2.1 MLV gag VLPs non-specifically package and export cellular RNAs

Virus-like particles (VLPs) produced by mammalian cells transfected with Gag, the core structural polyprotein for retroviruses, package cellular RNAs non-selectively in the absence of cis-acting viral packaging signals[57]. While VLPs already have a broad range of technological applications including vaccinology[59], the delivery of protein payloads to cells[61,63,65], and measurement of protein-protein interactions[67], we hypothesized that VLPs could also be leveraged to create an engineered RNA export pathway that would allow living cells to self-report transcriptional information to the extracellular environment (**Figure 3.1A-C**). We envision that live-cell transcriptional profiling will ultimately enable researchers to follow dynamic biological activities through time at the whole-transcriptome level from cells in a wide range of living biological systems and follow up with further studies of the intact cells or organism after

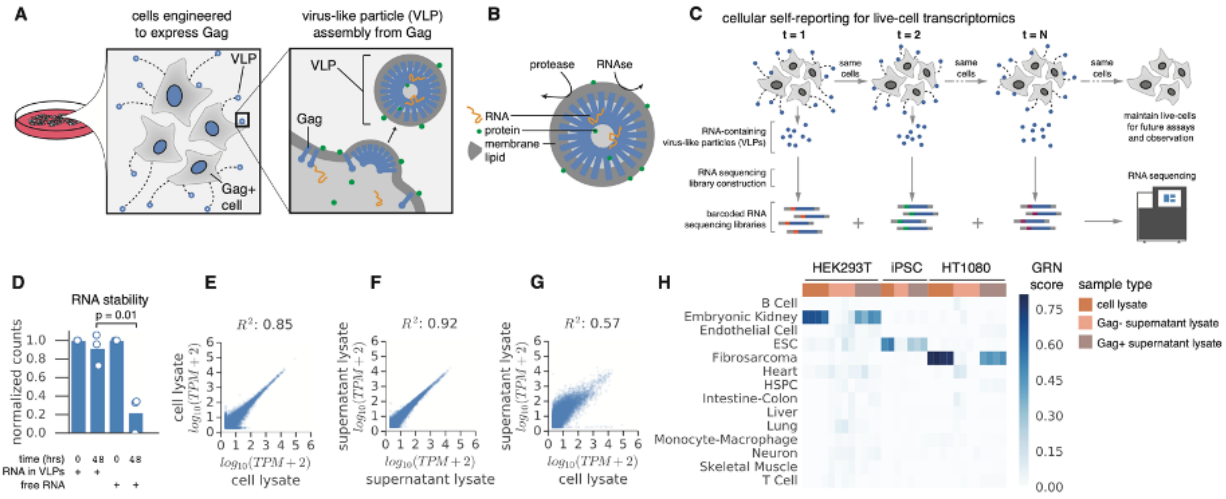


Figure 3.1. Self-reporting via VLPs for live-cell transcriptomics
(**A**) Self-reporting leverages virus-like particles (VLPs) to export RNA from living cells. Gag accumulates to assemble VLPs.
(**B**) VLPs can package several different types of cargos, including RNA, protein, and metabolites.
(**C**) Example of time-point collection using cellular self-reporting, where multiple time-points can be collected from the same biological samples.
(**D**) RNA stability in media at 37°C.
(**E**) RNA-seq replicate concordance of pGag+, pFLAG-VSV-G+ HEK293T cell lysates (biological replicates).
(**F**) RNA-seq replicate concordance of pGag+, pFLAG-VSV-G+ HEK293T supernatant lysates (biological replicates).
(**G**) RNA-seq sample representation showing pGag+ HEK293T lysate vs. pGag+ supernatant lysate.
(**H**) CellNet gene regulatory network (GRN) scores of RNA-seq from lysates and supernatants with and without Gag expression for HEK293T, iPSC, and HT1080 cells.

time series transcriptional analysis. In order to evaluate retroviral-based VLPs as the basis of such a technology platform, we transduced HEK293T with doxycycline (dox)-inducible murine leukemia virus (MLV) Gag fused to a P2A-linked GFP reporter (**Supplemental Figure 3.1A**) to validate full-length Gag expression and translation. After induction with doxycycline, we confirmed Gag expression via flow cytometry (**Supplemental Figure 3.1B**) and purified cell supernatants via ultracentrifugation. We then performed electron microscopy on purified supernatant from dox-induced cells to confirm the presence of VLPs (**Supplemental Figure 3.1C**). Further, we measured a 5.9 ± 1.2 fold increase (mean ± SD, n = 3, p = 0.0024) in exported GAPDH mRNA via quantitative reverse-transcription PCR (RT-qPCR) in dox-induced cell supernatants relative to baseline (**Supplemental Figure 3.1D**), confirming that Gag expression increases RNA export from living cells. Additionally, we characterized the stability of

VLPs and found that packaged RNA was stabilized for at least 48 hours in media at 37°C, whereas RNA not protected by VLPs degraded under the same conditions (**Figure 3.1D**). Importantly, HEK293T and iPS self-reporting cells displayed normal morphology (**Supplemental Figure 3.1E-F**). Finally, to further confirm that cellular RNAs were indeed packaged inside VLPs, we treated cellular media with benzonase nuclease and observed a 10-fold enrichment of RNA-seq reads that aligned to the human transcriptome from Gag-expressing cells compared to cells expressing a mutant Gag protein lacking the MA domain (GagΔMA) that is unable to form VLPs (**Supplemental Figure 3.2**). These results confirmed that the measured RNA was exported within VLPs induced by Gag expression in cells engineered to self-report their transcriptional state.

To characterize the RNA content of the VLPs, we performed RNA-seq[69] of VLPs harvested from HEK293T cells transfected with the self-reporting constructs. RNA-seq data showed that replicate concordance for VLP-packaged RNA was comparable to that of lysate-based RNA-seq (**Figure 3.1E-F**). Strikingly, we observed that the exported RNA was substantially representative of cellular lysate (**Figure 3.1G**) and that the relative abundance of transcripts measured in purified VLPs correlated with supernatant. To estimate the export rate of cellular mRNAs from individual cells, we used FACS to sort single HEK293T cells with stably-integrated dox-inducible Gag expression cassettes into independent wells. We harvested supernatant from the single-cell cultures and processed both the supernatant and cellular lysate for RNA-seq. The resulting read counts supported an estimate of the Gag-dependent export rate at >100 detected transcripts per cell per 24 hours for this construct (**Supplemental Figure 3.3**), which represents ~0.1% of mRNAs in the transcriptome with a typical mammalian cell possessing ~200,000 mRNA molecules[71]. We then tested whether Gag-based RNA self-reporting could be generalized beyond HEK293T cells by stably integrating Gag expression constructs into the genome of iPS and HT1080 cells. We assessed whether the RNAs

packaged and exported in VLPs in these cell lines were reflective of the cellular transcriptome and

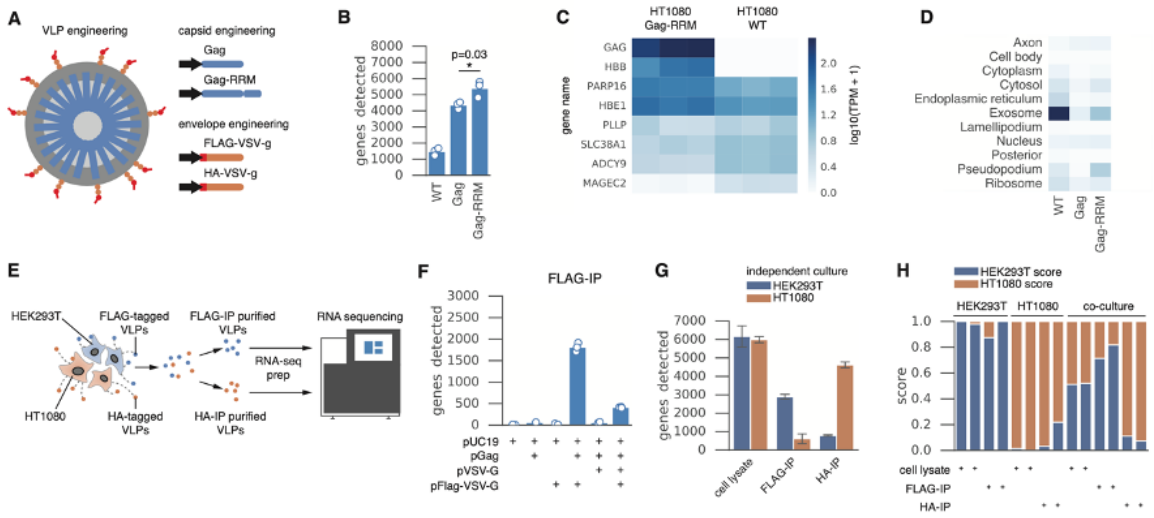


Figure 3.2. VLP engineering for multiplexed and selective RNA self-reporting
(**A**) Schematic of an engineered VLP.
(**B**) Gag–RRM leads to an increase in detected genes in HT1080 cells.
(**C**) Differentially expressed genes in HT1080 cells expressing Gag–RRM with columns as biological replicates (n=3).
(**D**) Importance of RNA localization/association annotation in predicting abundance in supernatant samples using a gradient boosted tree model.
(**E**) Schematic of co-culture experiment. HT1080 and HEK293T cells are independently transduced with VSV-G envelope proteins harboring different affinity tags. Cells are mixed and VLPs are then purified for downstream RNA-seq.
(**F**) RNA-seq of HEK293T supernatant purified via FLAG immunoprecipitation.
(**G**) RNA-seq results showing high quality libraries for matching affinity-based immunoprecipitations in independent culture.
(**H**) Cell-line specific gene expression detected for appropriate cell types in mono-culture and co-culture.

sufficiently informative to distinguish cell states. To do so, we sequenced VLP-derived RNAs produced from each cell type and applied CellNet[73] to ascertain the active regulatory networks and predict the biological identity of the sample. The VLP-derived transcriptomes supported specific classification to the expected identities, mirroring the classification performance of conventional lysate RNA-seq controls (**Figure 3.1H, Supplemental Figure 3.4**). These data highlight how the high-dimensional transcriptional readout afforded by self-reporting is amenable for gene regulatory network analysis and that VLP-derived RNAs are reflective of cellular states in the self-reporting cells.

3.2.2 Engineered gag proteins fused to RNA binding domains

We hypothesized that VLPs could be engineered to tailor RNA packaging via modification of the RNA binding domain in the Gag polyprotein, as well as display affinity-tagged envelope proteins to enable fascicle VLP purification (**Figure 3.2A**). Classically, the C-terminal nucleocapsid (NC) domain in the Gag polyprotein recognizes the cis-acting packaging signal in the viral RNA genome, but this basic domain also interacts nonspecifically with negative charged RNA molecules[57,75]. While a broad range of known RNA-binding domains could, in principle, be fused with Gag to optimize RNA capture for different export applications, we focused here on poly(A)-binding domains with the goal of enhancing mRNA export. Thus, we selected a tandem RNA recognition motif domain RRM1-2 from human PABPC4 to generate a Gag fusion protein, as RRM1-2 has been shown to interact with polyadenosine RNA with high affinity[77]. We transduced HEK293T and HT1080 cells with lentivirus containing the designed fusion construct to establish integrated cell lines with constitutive expression (**Supplemental Figure 3.5A**). After purifying VLPs from supernatants and preparing high-quality RNA-seq libraries (**Supplemental Figure 3.5B-E**), we measured the increase in the number of genes detected from Gag-expressing cells relative to the wild-type controls in HT1080 cells and found that fusing RRM1-2 to Gag resulted in the detection of more genes in HT1080 cells relative to wild-type Gag (**Figure 3.2B**). This demonstrates that engineered Gag fusion proteins support RNA export and that RNA export profiles can be tuned by modulating the RNA binding activity of engineered Gag fusion proteins. Interestingly, we found that expression of Gag–RRM and Gag were minimally perturbative of transcriptional profiles in HT1080 and HEK293T cells (**Figure 3.2C, Supplemental Figure 3.6A-B**), and that Gag expression led only to detectable upregulation of SERF1A expression in iPS cells (**Supplemental Figure 3.6C**). We observed minimal bias in the annotated localization of transcripts in self-reported RNA-seq data relative to cell lysate controls (**Figure 3.2D,**

Supplemental Figure 3.7). Indeed, exosome association was most predictive in the case of wild-type HEK293T and HT1080 control cells lacking an engineered export pathway, where a significant fraction of the RNA sequences may have in fact derived from naturally occurring exosomes. We also observed that VLPs exhibit a modest preference for packaging longer RNA molecules (**Supplemental Figure 3.8**).

3.2.3 VLP capsid engineering enables multiplexed transcriptional readouts

Given that mammalian cells naturally exist in complex environments composed of multiple interacting cell types, we sought to multiplex cellular self-reporting in a workflow where we simultaneously sample VLPs produced by multiple cell populations in co-culture (**Figure 3.2E**). We engineered the vesicular stomatitis virus G (VSV-G) envelope protein with an N-terminal FLAG tag and co-transfected HEK293Ts with MLV Gag to pseudotype VLPs. FLAG immunoprecipitation (IP) on cellular media revealed epitope-dependent specific isolation of VLPs by western blot (**Supplemental Figure 3.9A**). RNA-seq from the supernatant lysates led to detection of ~2000 genes from cells expressing Gag (**Supplemental Figure 3.9B**), while RNA-seq following FLAG IP led to high quality libraries only from Gag and FLAG-VSV-G expressing cells as expected (**Figure 3.2F**), and resulted in the detection of ~1700 genes. Again, we observed good replicate concordance and RNA representation in the supernatant IP RNA-seq data (**Supplemental Figure 3.9C-E**). Furthermore, we validated a similar IP approach to isolate HA-tagged VLPs from Gag and HA-VSV-G expressing cells and observed performance comparable to our FLAG-IP results (**Supplemental Figure 3.10A-B**). Additionally, we confirmed the epitope specificity of each IP protocol (**Supplemental Figure 3.10C-D**), which suggested the possibility of purifying orthogonal VLP populations from a mixture. To test multiplexing of RNA self-reporting, we independently transduced HEK293T and HT1080 cells with Gag, along with VSV-G envelope proteins tagged with FLAG and HA, respectively (**Supplemental Figure 3.11**). We were able to construct high-quality RNA-seq libraries from

purified VLPs of the independently cultured cell lines only when the IP antibody was cognate for each envelope tag (**Figure 3.2G**). The IP-based VLP purification performed well across a wide

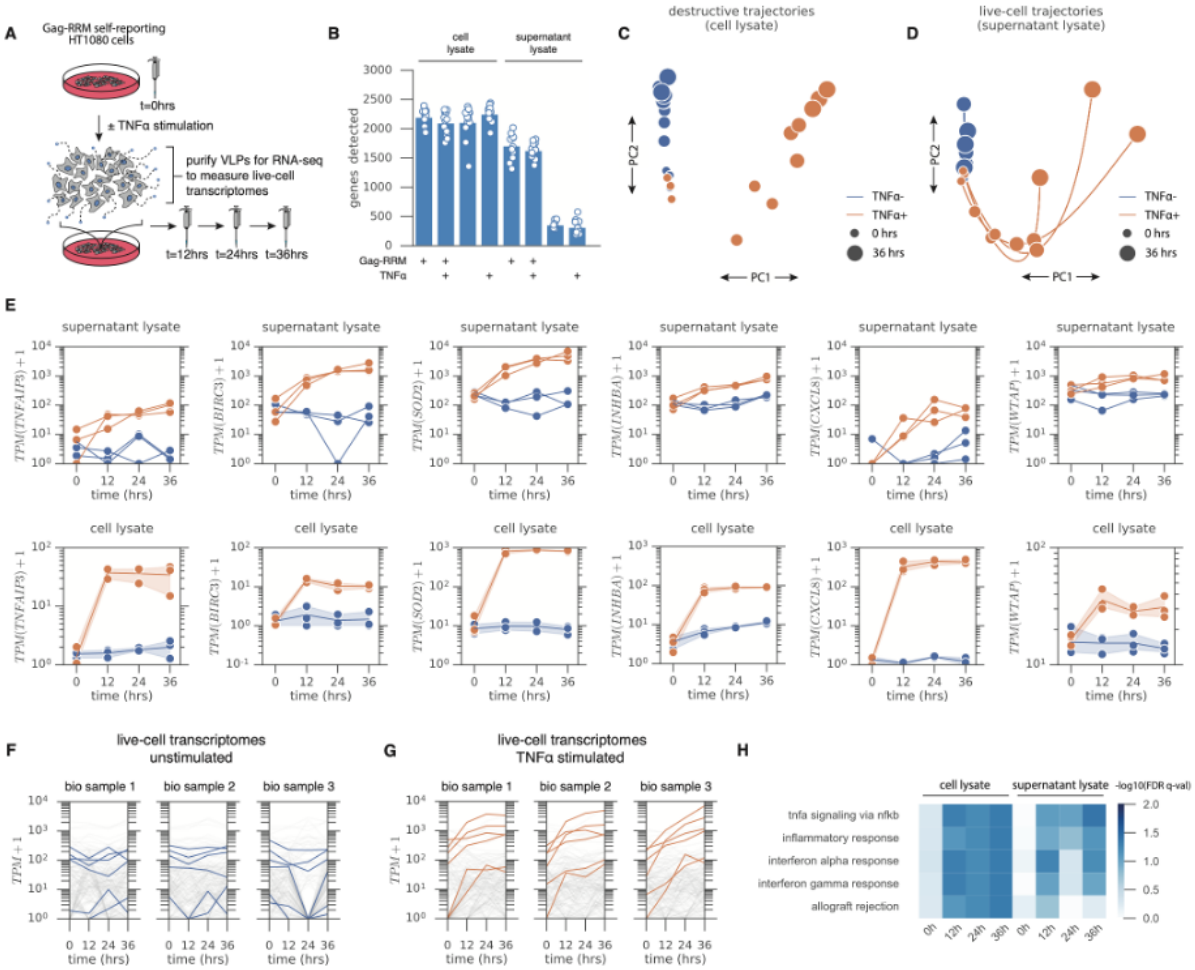


Figure 3.3. Self-reporting reveals TNFα response in longitudinally monitored HT1080 cells
(**A**) HT1080 cells with stable single-copy integrated Gag–RRM self-reporting constructs were longitudinally monitored with and without TNFα stimulation by harvesting supernatant with VLPs every 12 hours.
(**B**) Genes detected in RNA-seq libraries by sample type (± Gag–RRM expression, ± TNFα stimulation).
(**C**) PCA on differentially expressed genes from lysate controls showing destructive trajectories for cells in TNFα-stimulated (orange) and unstimulated (blue) conditions.
(**D**) PCA on differentially expressed genes from supernatants of self-reporting cells for live-cell transcriptional trajectories in TNFα-stimulated (orange) and unstimulated (blue) conditions.
(**E**) Comparison of differentially expressed genes between TNFα-stimulated (orange) and unstimulated (blue) conditions in supernatants of self-reporting cells. RNA-seq of VLPs from supernatant lysates (top) and RNA-seq of cell lysates (bottom).
(**F**) Live-cell transcriptomes for biological samples without TNFα stimulation highlighting differentially expressed genes (gray) and TNFAIP3, BIRC3, SOD2, INHBA, CXCL8, WTAP (blue).
(**G**) Live-cell transcriptomes for biological samples with TNFα stimulation highlighting differentially expressed genes (gray) and TNFAIP3, BIRC3, SOD2, INHBA, CXCL8, WTAP (orange).
(**H**) Gene set enrichment analysis (50 Hallmark gene sets) on lysates and supernatants from self-reporting cells (stimulated vs. unstimulated).

temporal range of media (cell supernatant containing VLPs) sampling durations, where we were able to detect genes significantly above background in as little as three hours of sampling (**Supplemental Figure 3.12A-B**). CellNet classification of transcripts detected by RNA-seq of IP-purified VLPs over various sampling windows further corroborated that the exported transcripts reflected the lysate transcriptome (**Supplemental Figure 3.12C-D**), underscoring the temporal resolution achievable with self-reporting. Finally, RNA-seq analysis of IP-purified, VLP-derived RNA from HEK293T / HT1080 co-cultures revealed cell line-specific gene expression (**Figure 3.2H, Supplemental Figure 3.13**), demonstrating the simultaneous specific analysis of independent cellular self-reporting information streams using the multiplexed IP approach with engineered VLPs.

3.2.4 Self-reporting captures transcriptional dynamics in living cells

Next, we sought to apply live-cell RNA self-reporting, wherein RNAs are continuously integrated in the culture media via VLPs to provide insights into temporal transcriptional programs in living cells. To do this, we sought to detect NF-κB signaling in stimulated HT1080 cells. We cultured Gag–RRM self-reporting HT1080 cells, and collected cell supernatant prior to TNFα stimulation and every 12 hours throughout the stimulation in order to capture transcriptional dynamics from the same cells over a 36-hour period (**Figure 3.3A**). To compare self-reporting RNA-seq to traditional lysate RNA-seq, we prepared separate cell populations for destructive lysis at each timepoint, and detected >1000 genes at each time point from VLP-containing cellular media and the lysate controls (**Figure 3.3B**). Principal components analysis (PCA) of expression data for genes with differential expression upon stimulation showed clear separation of stimulated cells at times post-stimulation in cell lysate samples as expected (**Figure 3.3C**). We observed similar separation when conducting PCA on genes differentially expressed in RNA-seq data from VLP-containing media samples at post-stimulation time points (**Figure 3.3D**). When conducting PCA with the same gene sets on supernatants from wild-type HT1080 cells lacking an engineered

RNA export pathway, TNFα-stimulated cells did not cleanly separate from unstimulated cells (**Supplemental Figure 3.14A-C**), indicating that background signals were insufficient to detect the transcriptional response of cells to TNFα stimulation.

Next, we compared the expression of genes identified as differentially regulated in standard cell lysate RNA-seq and live-cell RNA-seq of cell supernatants. Overall, live-cell RNA-seq data were concordant with the RNA-seq expression profiles of control cell lysate samples (**Figure 3.3E**). This indicated that VLP-derived RNA collected from the supernatants of self-reporting cells faithfully carried transient biological information (**Figure 3.3F-G**). We measured upregulation of genes known to be involved in TNFα signaling, notably TNFAIP3, BIRC3, SOD2, INHBA, and CXCL8[79,81]. We next performed gene set enrichment analysis (GSEA) to investigate whether cellular self-reporting accurately reported biological pathway activity[83]. We ran GSEA for the 50 Hallmark gene sets on RNA-seq data from the supernatant of TNFα-stimulated and unstimulated self-reporting cells at each time point. The top enriched gene set was TNFα signaling via NF-κB, confirming that dynamic activity of biological pathways can be non-destructively resolved in living cells by RNA-seq through the engineered self-reporting approach (**Figure 3.3H**). Altogether, these results demonstrate that VLP-based self-reporting enables live-cell transcriptomics, which can capture both transient transcriptional responses and meaningful biological pathway information from the same living cells over time.

3.3. DISCUSSION

Export of RNA via engineered VLPs makes live-cell transcriptomics possible through the cellular self-reporting paradigm. Live-cell RNA-seq enables new categories of experiments in which the transcriptome of a living cell sample is monitored over time without sampling the cells themselves. However, there are some limitations of our current protocol. Self-reporting requires collection of cell-free liquid samples at each time point of interest. This sets up a tradeoff

between time resolution/sensitivity and alteration of the conditions experienced by the cells that is set by the amount and frequency of supernatant sampling. Further, while fold-changes in gene expression detected in self-reported RNA were strongly correlated to fold-changes in cellular lysates, they do not correspond exactly, which complicates quantitative comparisons across live-cell and destructive RNA-seq studies (**Supplemental Figure 3.14D**). We note that the self-reporting data from VLPs presented here represent a smaller quantity of RNA per sample per time point than is available from lysing the self-reporting cells themselves. As a result, there is less statistical power for differential gene expression calls from a single time point relative to a lysate from a similar sample.

The genetically encoded format of the live-cell self-reporting system could ultimately enable targeted longitudinal transcriptomic measurements of cells, tissues, and organs in their physiological and spatial context. For example, the expression of self-reporting machinery could be driven by cell-type-specific promoters in transgenic animals to measure the transcriptomes of target cells in their native environment. The data shown here demonstrate that self-reported single-cell measurements are possible, and future refinements may lead to increased export rates, thereby enabling improvements to sensitivity and time resolution, as well as practical single-live-cell transcriptomics experiments with many time points. The self-reporting approach may also be adapted for non-destructive proteomic and metabolomic measurements. Altogether, the results of this study demonstrate that cellular self-reporting with engineered VLPs allows longitudinal analyte monitoring in the same biological samples, circumventing the need for physical access to samples and the lysis of living samples to obtain molecular analytes while preserving the living samples for further study after the longitudinal molecular assays are complete.

3.4 METHODS

3.4.1 Cell Culture

HEK293FT (Thermo Fisher Scientific Cat. No. R70007) and HT1080 (ATCC CCL-121) were cultured in a humidified incubator at 5% CO2 at 37C using high-glucose DMEM (Invitrogen) complemented with 10% FBS and 1% penicillin/streptomycin. Cell cultures were kept at low passage (<10) and regularly tested for mycoplasma contamination. iPS cells (1157-2 line, Boston Children's Hospital) were maintained in StemFlex media (Stem Cell Technologies), supplemented with 1% penicillin/streptomycin, and cultured on plasticware coated with growth factor-reduced matrigel (Corning) diluted in DMEM-F/12 according to manufacturer instructions. iPS cells were propagated as colonies using ReleSR (Stem Cell Technologies).

3.4.2 Molecular Cloning

All plasmid vectors were generated by Gibson Assembly using NEB's Gibson Assembly Master Mix, according to the manufacturer's protocol. The xenotropic MLV Gag ORF (UniProt Q27ID9) was synthesized by IDT as a gBlock and cloned in frame with a P2A-GFP sequence into the pcDNA3 backbone to generate the episomal expression vector, pxN01. As a negative control for VLP formation, we also created a truncation mutant with the MA domain of Gag (amino acids 2-130) removed (termed GagΔMA) to abolish VLP budding and maturation on lipid membranes. The pLenti backbone was used to generate the lentivirus vector pLV_CSR9 utilizing a CAG promoter to drive expression of an Gag-P2A-GFP-P2A-zeoR ORF. A dox inducible Gag vector (pLxN01) was generated by Gibson Assembly using the pCW57.1 vector. Vectors expressing Gag fusion proteins were generated by cloning the minimal RNA recognition motifs of RNA binding proteins (gBlocks synthesized by IDT) to the C-terminus of Gag with a 5X serine-glycine linker. Epitope-tagged VSV-G constructs were generated utilizing the VSV-G ORF from pMD2.G (a gift from D. Trono) to insert a FLAG (DYKDDDDK) or HA (YPYDVPDYA) sequence following the signal peptide at amino acid position 27. The epitope- tagged ORFs

were cloned into pcDNA3 or pLenti backbones with a 2A-mCherry-2A-puromycin cassette to derive episomal or lentivirus vectors, respectively. Finally, piggyBac vectors for expression in iPS cells were generated with a piggyBac transposon backbone (System Biosciences) containing a CBX3 ubiquitous chromatin opening element (UCOE) upstream of an EF1-a promoter driving expression of Gag or GagΔMA in frame with a P2A ribosomal skipping peptide and mNeonGreen reporter.

3.4.3 Lentivirus Production

HEK293FT cells were seeded at 1e6 cells/well in 6-well plates. The following day, cells were transfected when 90-95% confluent with pMD2.G (Addgene #12259), psPAX2 (Addgene #12260), and a lentiviral transfer plasmid (2:3:4 ratio by mass) using Lipofectamine 2000 (Thermo Fisher). Media was exchanged after 6 hours and viral supernatant was harvested 48 hours after transfection and filtered through 0.45 µm cellulose-acetate filters (VWR cat. no. 28145-481).

3.4.4 Generation of stable MLV Gag-expressing cells

Stably self-reporting HEK293FT and HT1080 cells were generated through sequential lentiviral transductions to first introduce a constitutive Gag expression vector followed by a constitutive epitope-tagged VSV-G expression vector. Cells were transduced by adding an appropriate amount of processed viral supernatant supplemented with polybrene (8 ug/mL) to the cellular media in order to achieve ~30% fluorescent reporter-positive cells for single-copy integrations. The cells were incubated in virus for 24 hours followed by antibiotic selection 48 hours post-transduction at the following concentrations: 1 µg/mL puromycin (Thermo Scientific A1113802), and 300 µg/mL zeocin (Thermo Scientific R25001). Stable Gag-expressing iPS cells were generated through piggyBac transposition to achieve long-term transgene expression and avoid lentiviral silencing. iPS cells were single cell dissociated with TrypLE reagent (Thermo

Scientific) and 800,000 cells were nucleofected with 2.5 ug of Super piggyBac Transposase (SBI cat. no. PB200A-1) and 10 ug of transposon plasmid using Lonza Cell Line Nucleofector Kit V (Lonza cat. no. VCA-1003) on an Amaxa IIb Nucleofector with program B-016. iPS cells post- nucleofection were cultured with 1 uM Y-27632 ROCK inhibitor (Stem Cell Technologies) for 24 hours and FACS sorted one week later to enrich for mNeon+ cells.

3.4.5 Transient transfection of MLV Gag

HEK293T were seeded into 6-well plates at a density of 100,000 cells/cm2. The next day, cells were transfected with 2000 ng of total DNA comprising a Gag expression plasmid, VSV-G expression plasmid and pUC19 plasmid (2:3:4 ratio by mass) using Lipofectamine 2000 (Thermo Fisher). The media was changed 6 hours post-transfection and was collected for downstream processing 48 hours post-transfection.

3.4.6 VLP purification from cellular media

VLP-containing cellular media from transient transfection experiments or stable Gag-expressing cells was first processed to remove cellular debris by centrifugation at 2,000 g for 10 minutes at 4C. Cleared cellular media from iPS cells was treated with 200 U/mL benzonase nuclease (Sigma cat. no. E1014) at 37C for 60 minutes with mixing every 15 minutes to degrade free ribosomal RNAs and background RNAs. Next, the media was further filtered through a 0.45 um cellulose acetate filter (VWR cat. no. 28145-481) and subsequently concentrated by centrifugation with a 100 kDa Amicon cutoff filter (Millipore Sigma cat. no. UFC5100) at 2500 g for 30 minutes at 4C. The retentate within the filter was either frozen at -80C for storage, used directly as a supernatant control for downstream assays, or used as an input for immunoprecipitation-based isolation of VLPs. For FLAG-based immunoprecipitation of VLPs, 20 uL of Anti-FLAG M2 Magnetic Beads (Sigma cat. no. M8823) were used per sample and washed 3x with TBS buffer. The beads were resuspended in 500 uL TBS + 1% Tween-20 and

incubated with the VLP retentate on a rotisserie at 4C overnight. The next day, the beads were washed 3x with TBS + 1% Tween-20 and eluted in 500 ng/uL 3x FLAG peptide (Sigma cat. no. F4799) at 1200 RPM shaking for 30 minutes at 4C. For HA-based immunoprecipitation of VLPs, 50 uL of His-Tag Dynabeads (Invitrogen cat. no. 10103D) were used per sample according to the manufacturer's protocol. HA-tagged VLPs were eluted in 50 uL of elution buffer (300 mM imidazole, 50 mM sodium phosphate, 300 mM NaCl and 0.01% Tween-20) at 1200 RPM shaking for 30 minutes at 4C. The eluted VLPs were used directly in downstream assays or stored at -80C.

3.4.7 RT-qPCR of VLPs

HEK293T cells stably transduced with lenti packaging inducible Gag (pLxN01) were cultured in T75 culture flasks for 72 hours ±Dox (10 ug/mL) before 10 mL of media was harvested from the flasks. The supernatant was stored at 4C for 24 hours before centrifugation at 600 rcf for 10 minutes at 4C to pellet any floating cells that were inadvertently collected with the cell media. The top 9 mL of media were then taken, and centrifuged at 2000 rcf for 10 minutes at 4C to pellet cellular debris. The top 8.5 mL of media were then collected and filtered through a 0.45 um cellulose-acetate syringe filter (VWR cat. no. 28145-481). 8 mL of filtrate was collected and stored at 4C overnight before ultracentrifugation. The filtrate was prepared for ultracentrifugation using 31.5 mL tubes (Beckman Coulter cat. no. 358126) with 1 mL of 70% sucrose cushion followed by 5 mL of 20% sucrose, and filtrate diluted in 1xPBS to bring to a total of 30 mL. The prepared tube was then ultracentrifuged (Thermo (Sorvall) WX80 Ultra-Centrifuge) with a swinging-bucket rotor (AH-629 (36 mL), Thermo cat. no. 54284) at 26,000 RPM for 2 hours at 4C. The interface between the 70% and 20% sucrose was collected with a syringe and then frozen at -80C. For RT-qPCR, the purified supernatant was thawed and 5 uL was taken per replicate and combined with 5 uL of 2xTCL (Qiagen cat. no. 1070498) for lysis. Lysed supernatant was then treated with 22 uL of RNAClean XP (Beckman Coulter cat. no.

A63987) and purified according to standard protocol, rinsing with 100 uL of 80% EtOH. The magnetic beads were eluted in qScript mix (Quanta Bio cat. no. 95047-025) and cDNA synthesis was performed according to the manufacturer's protocol. qPCR was performed on 1 uL of cDNA with JumpStart Taq ReadyMix (Millipore Sigma cat. no. P2893-100RXN) using Rox as a reference dye (Millipore Sigma cat. no. R4526-5ML), primers spanning GAPDH exon junctions (Fwd: `gaaggctggggctcatttgc`, Rev: `ggaggcattgctgatgatct`) and a custom TaqMan probe (Seq: `atctctgccccctctgctgatg`, ordered as a FAM TaqMan probe from IDT). qPCR was conducted on a Stratagene Mx3000P qPCR System real-time PCR.

3.4.8 RNA-Sequencing

RNA from cells, supernatant, or IP-purified VLPs was extracted using 2X TCL lysis buffer (Qiagen cat. no. 1070498). At least two technical replicates were prepared per sample using the SMART-Seq2 protocol as previously published[69] with some modifications. Briefly, the lysed samples were 2.2X RNA SPRI (Beckman Coulter cat. no. A63987) cleaned and reverse transcribed in the presence of a template switching oligo (Exiqon) with Maxima RNase H-minus RT (Thermo Fisher Scientific cat. no. EP0751) using a polyT primer containing the ISPCR sequence. Whole transcriptome amplification proceeded with KAPA HiFi HotStart ReadyMix using an ISPCR primer according to the following thermal program: 98C for 3 minutes, 27 cycles of 98C for 15 seconds, 67C for 20 seconds, and 72C for 6 minutes, and a final extension step of 72C for 5 minutes. The amplified cDNA was cleaned with 0.8X DNA SPRI beads (Beckman Coulter cat. no. B23318). Ten nanograms of DNA was tagmented at 58C for 10 minutes in a 10 uL reaction containing 2 uL of 5X tagmentation buffer (50 mM Tris-HCl, 25 mM MgCl2 pH 8.0), 2 uL of Tris Buffer (10 mM Tris-HCl, 1% Tween-20 pH 8), and 4 uL Nextera (Illumina). The reaction was stopped with 1% SDS and incubated at 72C for 10 minutes, then 4C for 3 minutes. The tagmented library was cleaned with 1X DNA SPRI beads followed by 12 cycles of PCR with NEBNext High Fidelity polymerase to incorporate sample index barcodes

and Illumina flow cell handles. The final libraries were pooled, diluted and sequenced on a NextSeq-500 (Illumina) in paired-end mode using a 75 cycle High Output Kit v2.

3.4.9 Western-bloting

Cells were washed with PBS and protein lysate was extracted using an extract solution (150 mM NaCl, 1% Triton X-100, 50 mM Tris-HCl pH 7.5) supplemented with a protease inhibitor (Sigma cat. no. 4693159001). The cellular lysate was incubated on a rotisserie for 30 minutes at 4C, followed by centrifugation at 10,000 g for 5 minutes at 4C. The protein lysate, IP-purified VLPs, or cleared cellular media was then used directly for Western-blotting. Samples were reduced using 2X Tris-Glycine SDS Sample Buffer (Life Technologies cat. no. LC2676) and 10X NuPAGE Reducing Agent (Life Technologies cat. no. NP0009) at 98C for 5 minutes. The reduced samples were then run on a Novex WedgeWell 8-16% Tris-Glycine Gel (Life Technologies cat. no. XP08165BOX) for 50 minutes at 225V. The gel was then transferred to a PVDF membrane (Life Technologiescat. no. IB24002) using an iBlot 2 device (Life Technologies). The PVDF membrane was blocked in 5% milk (Bio-Rad cat. no. 1706404XTU) for 1 hour at room temperature with gentle shaking, followed by an overnight incubation at 4C with gentle shaking in primary antibodies to Gag at 1/2000 dilution (Abcam cat. no. ab100970) and to actin at 1/5000 dilution (Abcam cat. no. ab179467). The next day the membrane was washed 3x with 5% milk and incubated for 4 hours at 4C in an anti-rabbit secondary antibody (Sigma cat. no. 41176). The membrane was imaged on a Azure Biosystems C600 Imaging System.

3.4.10 Transcriptional profiling of TNFa stimulation

25,000 HT1080 cells were plated in TC-treated 96-well plates (VWR cat. no. 62406-117) with 100 uL of media per well. After every 12 hours, media was changed with pre-warmed media ± TNFα at a final concentration of 30 ng/mL (Invivogen cat. no. rcyc-htnfa), introducing

TNF-α at each time point. Destructive lysates were collected by aspirating media and lysing with 50 uL of 1xTCL (Qiagen cat. no. 1031576) and stored at -20C until the final time point. Supernatants from self-reported cells were carefully collected (to not disturb adherent cells) and stored at 4C until the final time point. After the final time point, supernatants were centrifuged at 2000 rcf for 10 minutes at 4C and the top 20 uL of media was carefully collected (to not disturb adherent cells) and mixed with 20 uL of 2xTCL (Qiagen cat. no. 1070498) for lysis. Samples were collected every 12 hours throughout the time course. RNA-seq libraries were prepared from 20 uL of supernatant lysate and 10 uL of cell lysate using SMART-Seq2 with 27 WTA cycles for supernatant libraries and 21 WTA cycles for cell libraries.

3.4.11 Single cell export rate measurements

HEK293T cells were stably transduced with pLxN01 lentivirus to create a single copy integrated cell-line. Single HEK293T cells were sorted into a 384 well plate (Corning cat. no. 8794BC) with 10 uL of media using a Sony SH800Z cell sorter with a 100 um chip in plate mode at 200-300 events per second. Dox-induced conditions were induced 24 hrs before sorting, and were sorted into Dox+ media (10 ug/mL) from the top quartile of GFP+ cells. Immediately after sorting, the plate was centrifuged at 200 rcf for 1 minute at room temperature, and then immediately placed in a tissue culture incubator. The plate was cultured for 24 hours after which 10 uL of supernatant was harvested and directly lysed with 2xTCL (Qiagen cat. no. 1070498). Immediately after harvesting supernatant, corresponding cell lysates were prepared by using 10 uL of 1xTCL (Qiagen cat. no. 1031576). RNA-seq libraries were assembled using SMART-Seq2 with 27 WTA cycles for supernatant libraries and 21 WTA cycles for cell libraries. Libraries were sequences on the Illumina MiniSeq platform using a 150-cycle kit (8-75-75-8 read configuration). Paired-end reads were pseudo-aligned using Kallisto (version 0.43.1) (43) using a reference generated by concatenating MLV Gag (Uniprot: Q27ID9, GAG_XMRV6), the human cDNA reference (Homo_sapiens.GRCh38.cdna.all.fa.gz), and cow cDNA reference

(Bos_taurus.UMD3.1.cdna.all.fa.gz) with an index build from 31-mers. Supernatants were only considered for analysis if the corresponding lysates had at least 1e4 unique transcript isoforms detected, indicating the cell was present in the lysate sample (not the supernatant sample). For conservative estimation of export rates, we assumed that each unique gene detected was a unique RNA molecule measured (no read counts or UMI counts were used in the calculation).

3.4.12 RNA-Seq processing

RNA sequencing paired-end reads were pseudo-aligned using Kallisto (version 0.43.1)[85]. The hg19 cDNA fasta reference from UCSC was appended with the coding sequences of gag, gagΔMA, mNeon and VSV-G in order to generate a custom Kallisto index via the “kallisto index” command. Differential gene expression analysis was performed using DESeq2 (version 1.30.0) with R (version 4.0.3) on the estimated count matrix output from Kallisto. Gene set enrichment analysis was performed using GSEA (version 4.1.0) on TPM output from Kallisto. Downstream analysis was performed with custom python scripts (python versions 2.7 and 3). Sequencing reads were downsampled to match sequencing depth for all samples when producing plots showing the number of genes detected.

3.4.13 Co-culture analysis

RNA sequencing data was filtered for genes with TPM greater than 10, and then basis vectors were generated by finding genes that were exclusively detected in cellular lysates of HEK293T or HT1080 cells. Cell-type specific scores were calculated by taking the inner product of the binarized basis vectors, and the binarized RNA-seq expression vector for each sample. The relative score was then calculated by dividing by the sum of the inner products, to determine which inner product was stronger, thus inferring the sample of origin.

3.4.14 Predictive modeling and RNA localization analysis

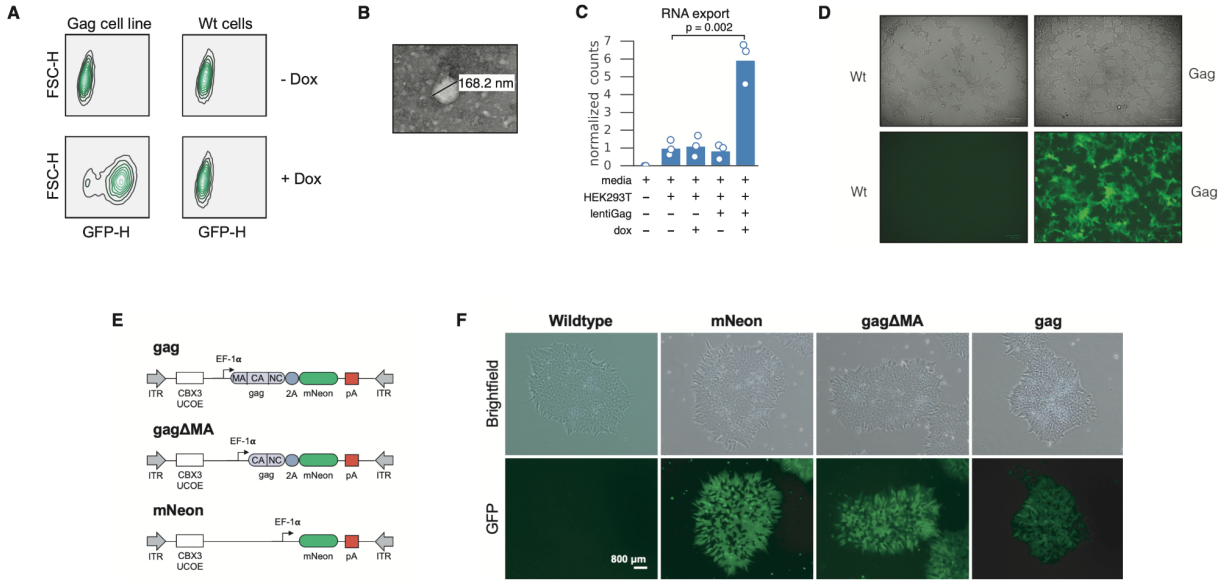
A matrix of supernatant to lysate TPM ratios was log transformed, and then several features were selected to build a model for supernatant to lysate TPM ratio prediction. RNA location was selected as a feature, and the RNALocate[87] database was used to annotate subcellular localization. Other features were transcript length (including UTRs and CDS), GC content (%). Further, we treated homology (7-mer sequences) between each transcript and the MLV genome as a feature, to account for any structural or sequence specificity in RNA export. Briefly, we counted the matches of 7-mers from the MLV genome for each transcript, and then binned the location of the 7-mer into 100 discrete bins along the positional axis of the MLV genome. Using the measured log-transformed supernatant to lysate RNA-seq measurements from HEK293T and HT1080 cell lines, along with the features above, we split the data into a training and test set and constructed a gradient boosting regressor {'n_estimators': 500, 'max_depth': 30, 'min_samples_split': 10, 'learning_rate': 0.01, 'loss': 'ls','subsample': 0.1,'verbose':1, 'criterion':'friedman_mse','min_samples_leaf':2,'min_weight_fraction_leaf':0.0}. We then looked at feature importance scores to better understand RNA properties that would influence export bias.

3.4.15 Inference of gene regulatory networks with CellNet

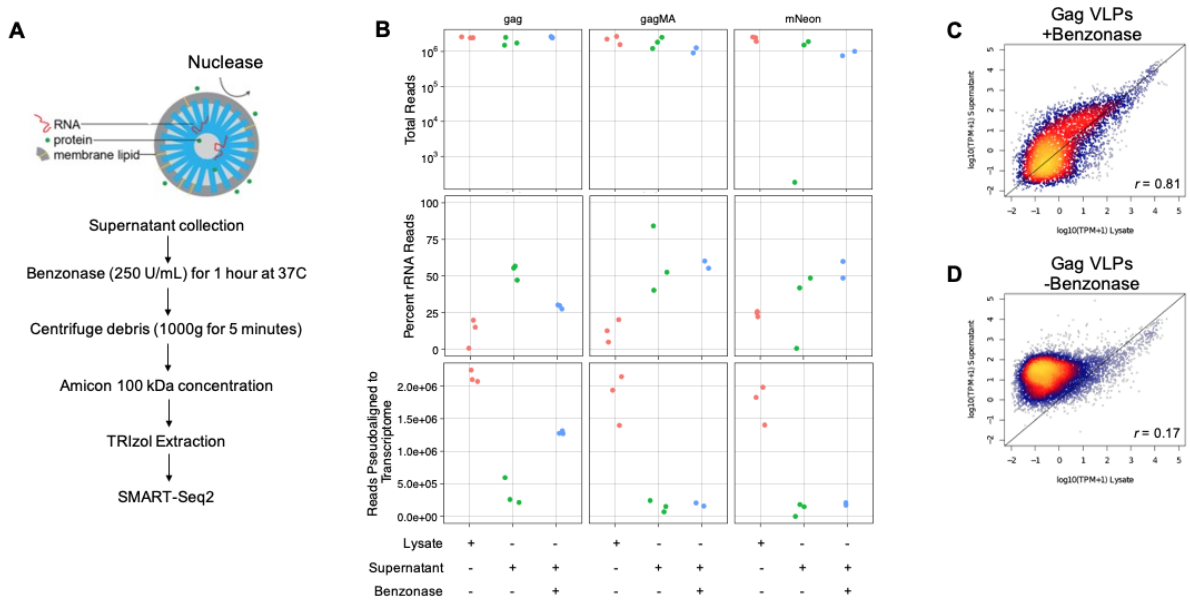
We utilized CellNet[73] to derive active gene regulatory networks (GRNs) from lysate and VLP RNA-seq samples and to assess the ability of VLP-derived RNAs to reflect the GRNs of the producing cell. We first retrained the CellNet classifier to include embryonic kidney and fibrosarcoma GRNs. We curated lysate RNA-seq samples from HEK293Ts and HT1080s and quantified transcript abundances with `cn_salmon()` using a pre-prepared Salmon transcript index, salmon.index.human.122116.tgz available from https://github.com/pcahan1/CellNet. We constructed new cell-type specific GRNs with `cn_make_grn()` using samples from the June 20, 2017 edition of the human CellNet Processor (https://github.com/pcahan1/CellNet) and our

HEK293T and HT1080 samples. We assessed the random forest classifier using `cn_splitMakeAssess()` and generated a new CellNet processor object with `cn_make_processor()`. We then applied CellNet with this retrained classifier using default settings to lysate and VLP RNA-Seq samples and plotted the sample classification scores as a heatmap in R.

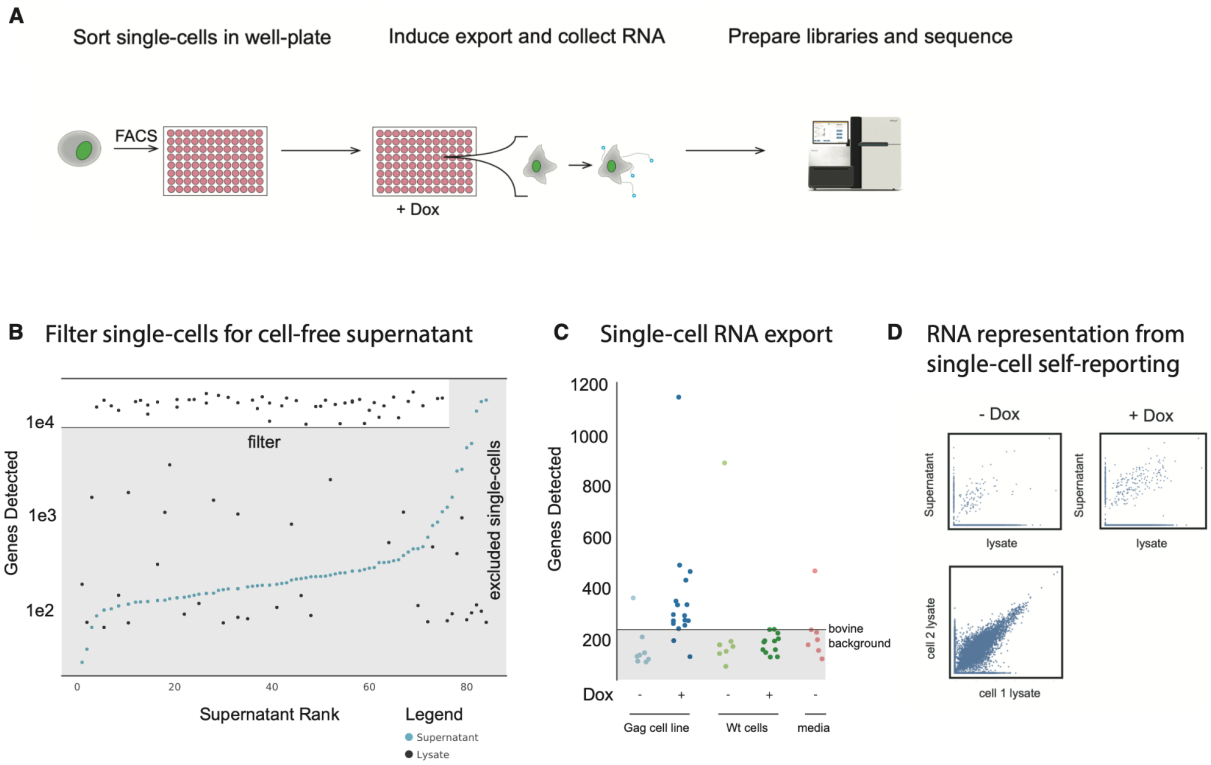
3.5 SUPPLEMENTAL FIGURES



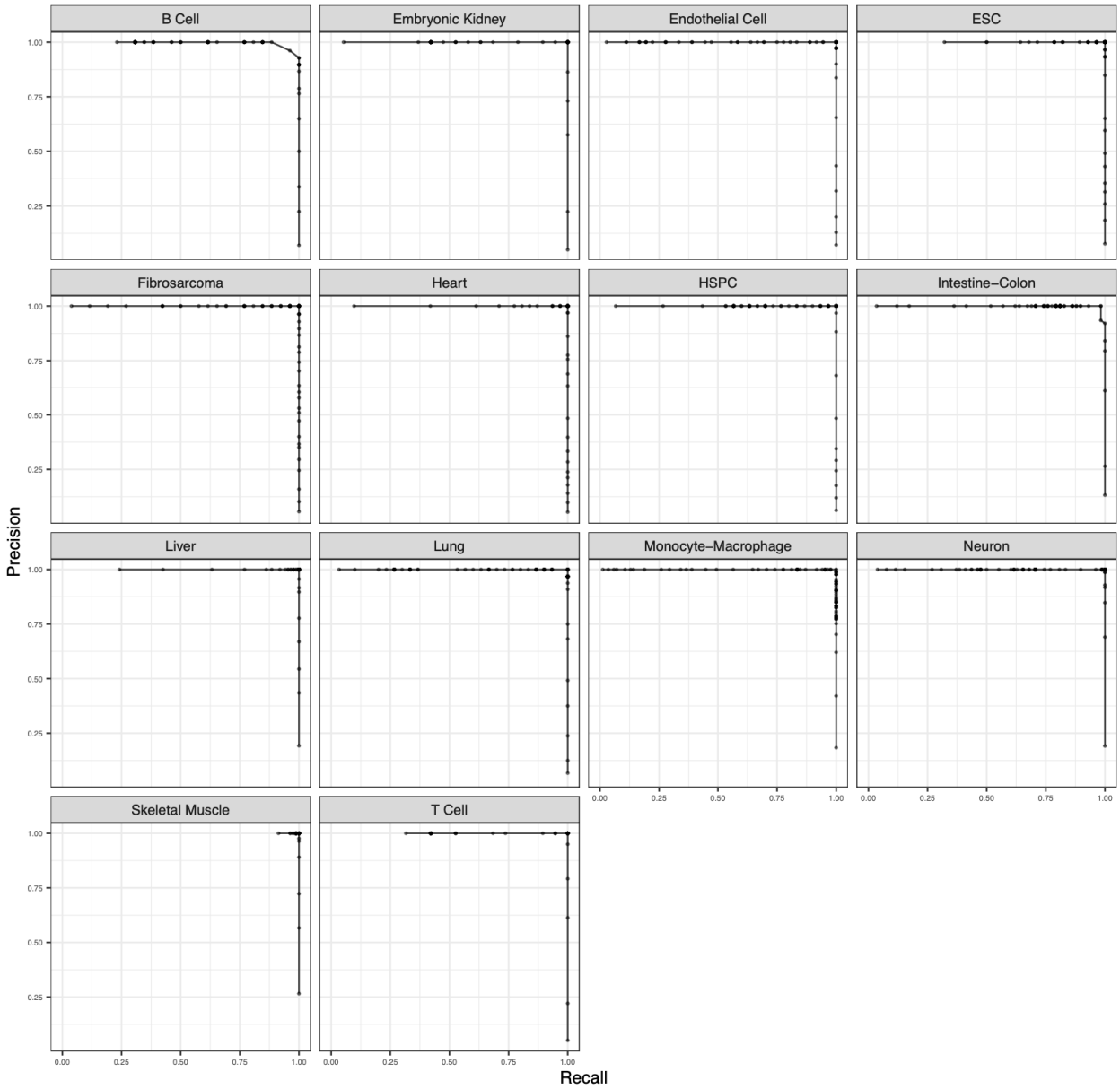
Supplemental Figure 3.1 Characterization of stably integrated Gag+ cell lines. (**A**) Flow cytometry on dox-inducible Gag+ HEK293T cell lines generated with lentiviral transduction. (**B**) Negative stain electron micrograph showing a VLP. (**C**) RT-qPCR results from supernatant purified from wild-type and Gag+ HEK293T cell lines ± doxycycline. GAPDH copy number was used as a proxy for exported RNA. Doxycycline induction led to VLP formation and RNA export. (**D**) HEK293T cell morphology with Gag expression. (**E**) PiggyBac expression vector diagram. (**F**) Brightfield and GFP images of iPS cells stably transposed with Gag, GagΔMA, and mNeon piggyBac constructs.



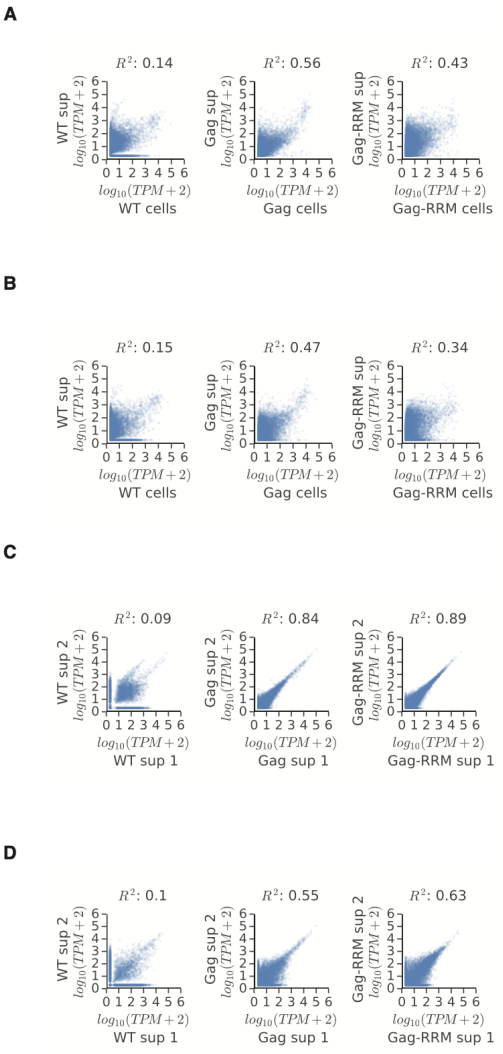
Supplemental Figure 3.2 Characterization of RNAs packaged in VLPs via nuclease digestion of cellular media. (A) VLP processing workflow for iPS cells. **(B)** Plot of RNA-Seq reads +/- benzonase treatment on media from Gag, GagΔMA, and mNeon expressing iPS cells to provide evidence that transcriptional signal is dependent both on Gag expression and the formation of exported VLPs able to protect RNA cargo from enzymatic degradation (reads pseudoaligned to the human transcriptome). **(C, D)** TPM concordance between supernatant and cell lysate +/- benzonase treatment from Gag expressing cells.



Supplemental Figure 3.3 Single-cell RNA self-reporting. (**A**) Single cell experiment overview. (**B**) Lysate signals and filter criteria. Samples were only included for supernatant analysis if cell lysates (black) had greater than 1e4 genes detected, to confidently call supernatant (cyan) signal as RNA from VLPs rather than from a dislodged or floating cell. (**C**) Supernatant signals showing RNA self-reporting for dox-induced Gag+ single cells. (**D**) Single-cell RNA-seq transcript abundance for supernatants vs. lysates (±dox), and lysate vs. lysate.

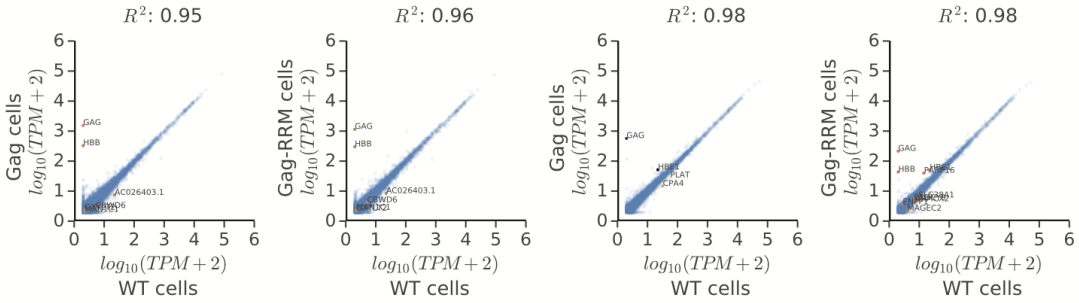


Supplemental Figure 3.4 CellNet training. Classification performance of a modified CellNet[73] random forest classifier trained on an expanded compendium of human cell types (including embryonic kidney and fibrosarcoma cell types, corresponding to HEK293T and HT1080 cells, respectively). Precision-recall curves show excellent performance for each cell-type classifier trained on human RNA-Seq data.



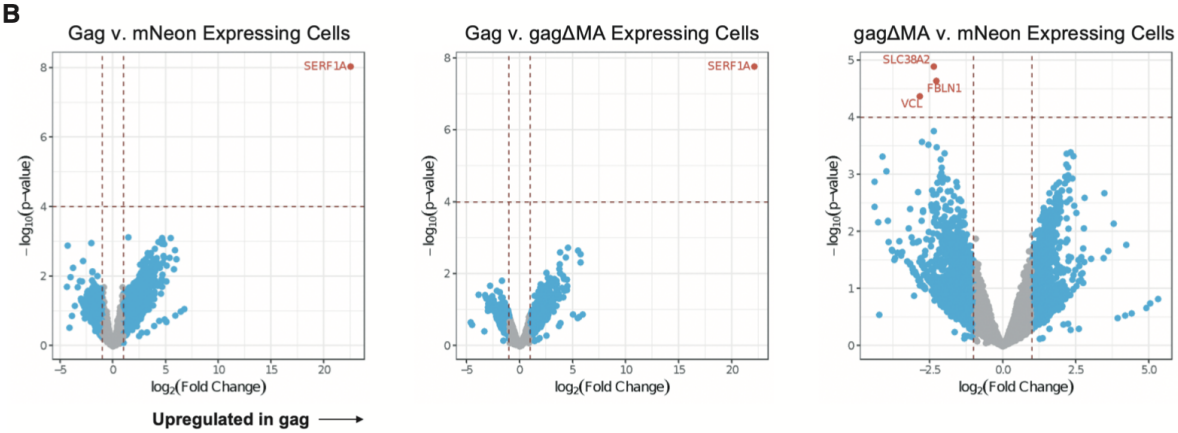
Supplemental Figure 3.5 Gag fusion characterization with stable, single-copy integrated cell lines. (**A**) RNA representation from wild-type HEK293T and HEK293T expressing different Gag constructs (**B**) RNA representation from wild-type HT1080 and HT1080 expressing different Gag constructs. (**C**) RNA-seq biological replicates of supernatant lysates from wild-type HEK293T and HEK293T expressing different Gag constructs. (**D**) RNA-seq biological replicates of supernatant lysates from wild-type HT1080 and HT1080 expressing different Gag constructs.

A

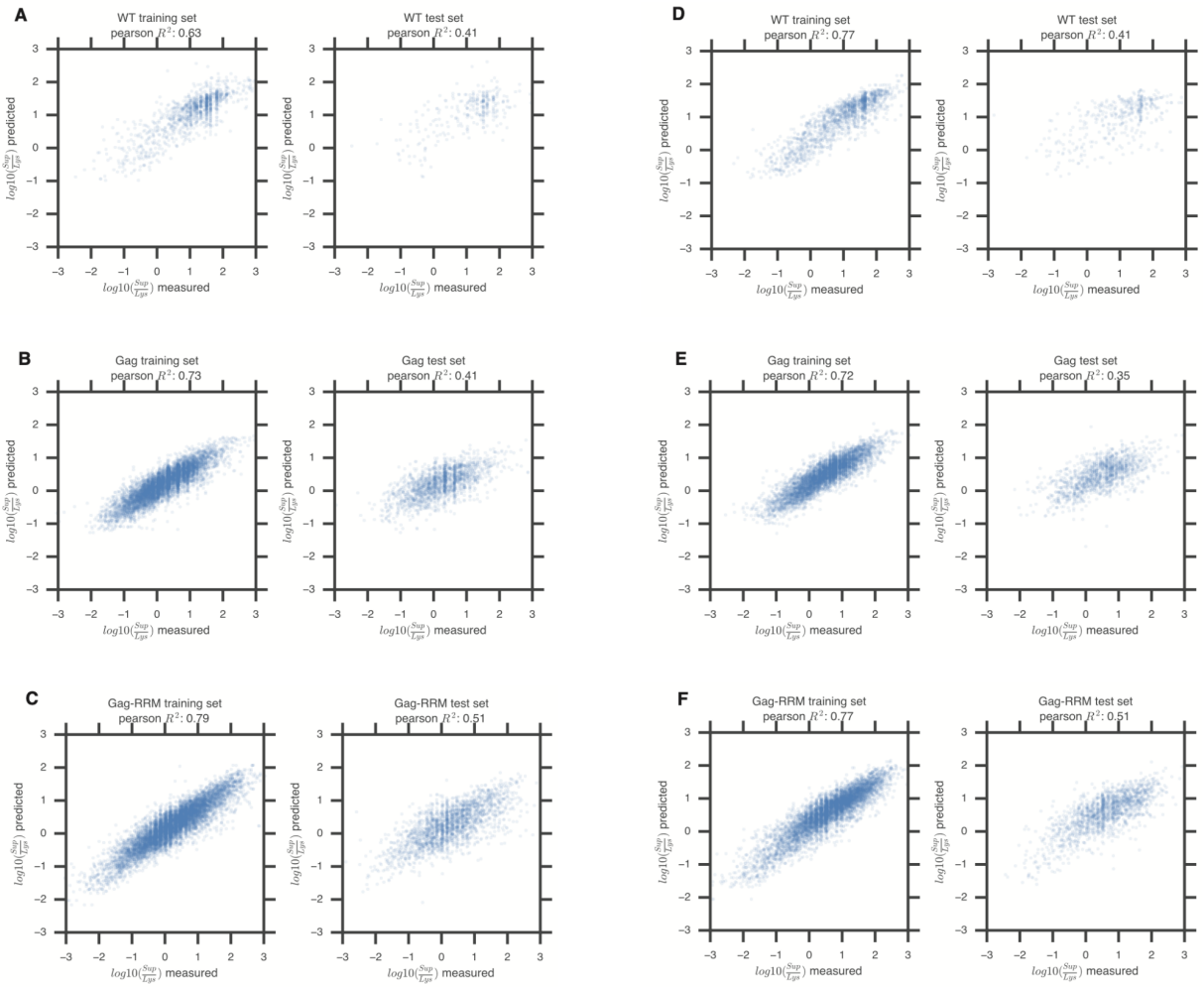


293T Gag vs wild-type	293T Gag-RRM vs wild-type	HT1080 Gag vs wild-type	HT1080 Gag-RRM vs wild-type
GAG	GAG	GAG	GAG
HBB	HBB	PLAT	HBB
AC026403.1	AC026403.1	CPA4	PARP16
MAN1C1	MAN1C1	HBE1	HBE1
CBWD6	CBWD6		MAGEC2
GXYLT2	GXYLT2		ADCY9
			SLC38A1
			PLLP
			ENPP1
			ARMCX2

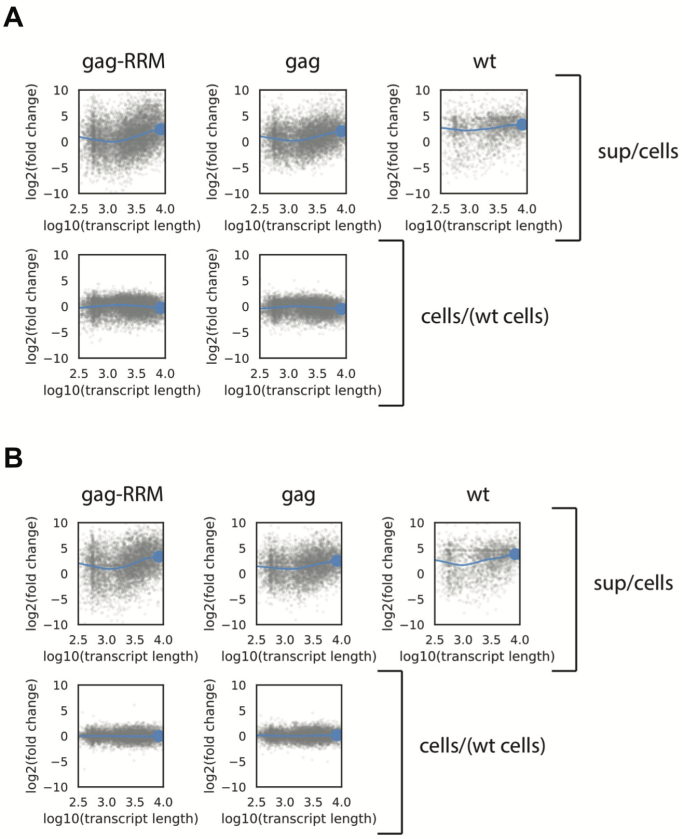
B



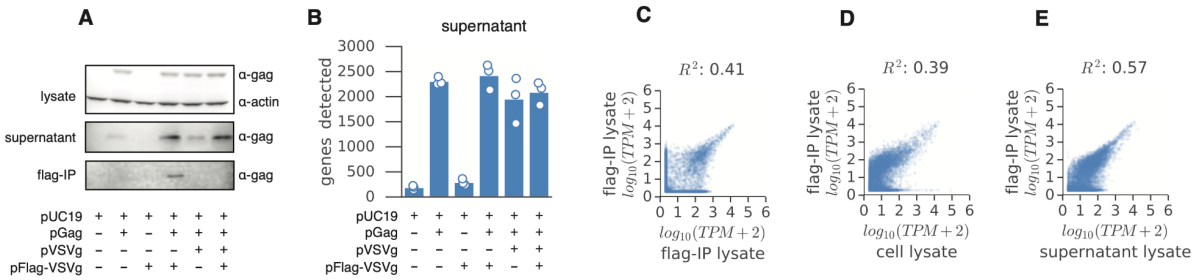
Supplemental Figure 3.6 Differentially expressed genes in self-reporting HEK293T and HT1080 cells. Using DESeq2, we identified differentially expressed genes (DEGs) in various self-reporting cell lines generated with stable single-copy integration of different self-reporting constructs. (**A**) We observed that self-reporting was minimally perturbative in Gag+ HEK293T, Gag–RRM+ HEK293T, Gag+ HT1080, and Gag–RRM+ HT1080 cells (left to right). (**B**) Volcano plots of DEGs detected in lysate from Gag and GagΔMA expressing cells in comparison to mNeon control cells.



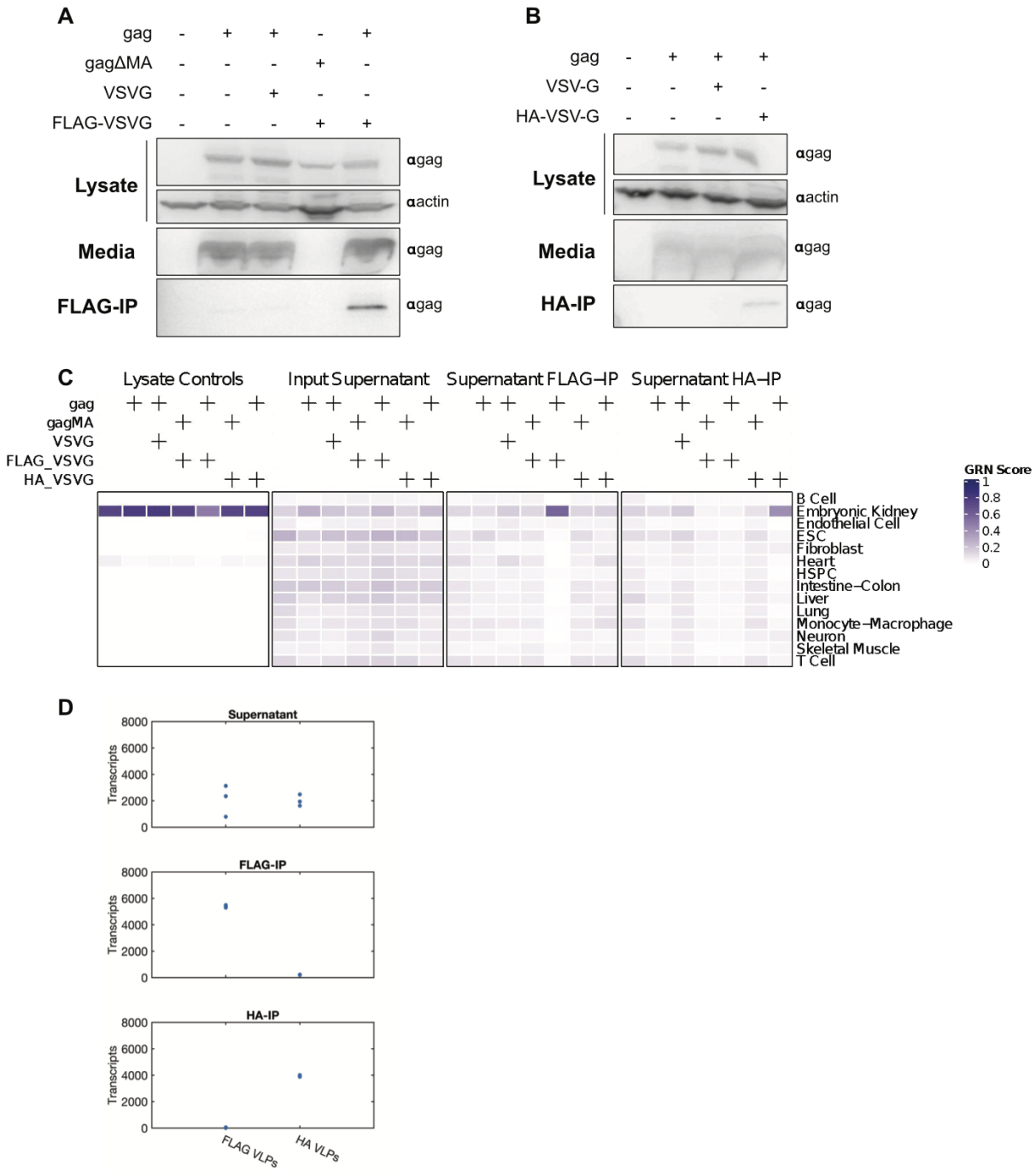
Supplemental Figure 3.7 Gradient-boosted tree regression performance for predicting ratios of self-reported RNA to lysate RNA. (**A**) Model performance for supernatant to cell lysate ratio predictions (training and test sets) for genes in wild-type HEK293T cells. (**B**) Model performance for supernatant to cell lysate ratio predictions (training and test sets) for genes in Gag+ HEK293T cells. (**C**) Model performance for supernatant to cell lysate ratio predictions (training and test sets) for genes in Gag–RRM+ HEK293T cells. (**D**) Model performance for supernatant to cell lysate ratio predictions (training and test sets) for genes in wild-type HT1080 cells. (**E**) Model performance for supernatant to cell lysate ratio predictions (training and test sets) for genes in Gag+ HT1080 cells. (**F**) Model performance for supernatant to cell lysate ratio predictions (training and test sets) for genes in Gag–RRM+ HT1080 cells.



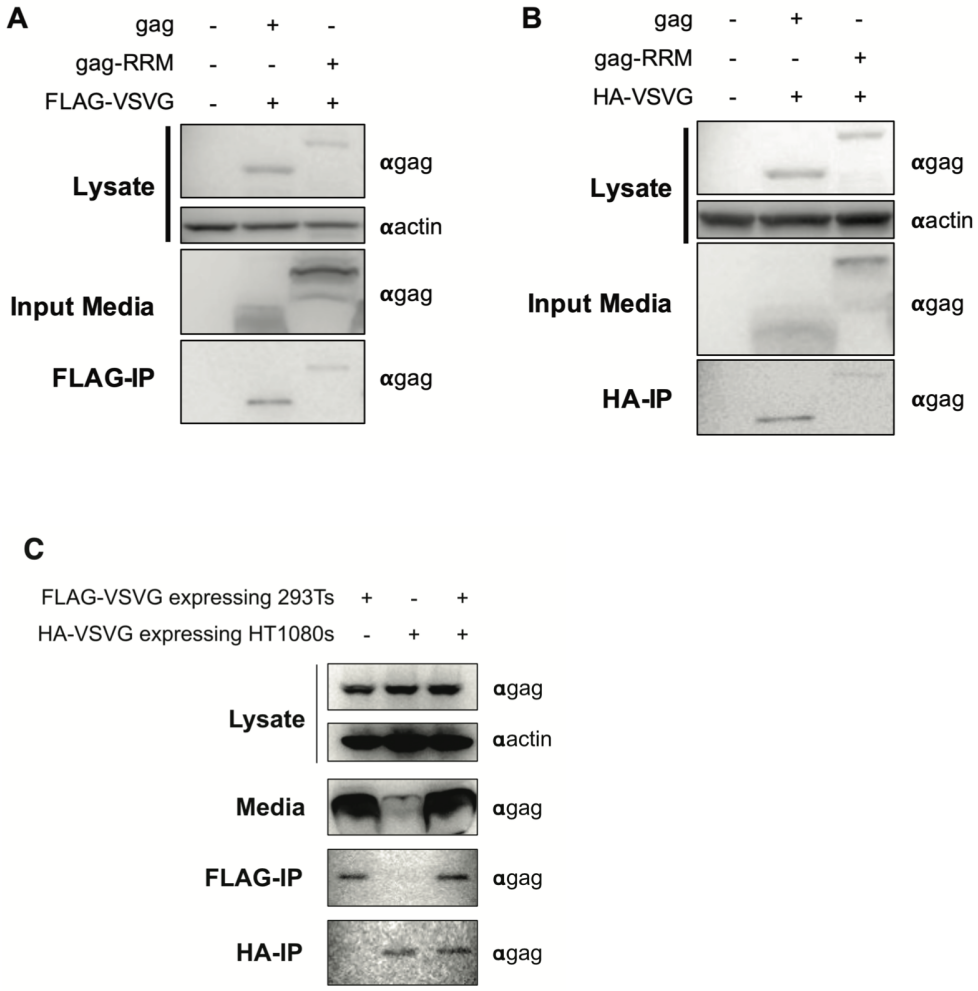
Supplemental Figure 3.8 Long RNA transcripts are preferentially packaged in VLPs. (**A**) Packaging size preference for HEK293T cells (lowess trendline in blue, depicting MLV genome size with a large blue dot). (**B**) Packaging size preference for HT1080 cells (lowess trendline in blue, depicting MLV genome size with a large blue dot).



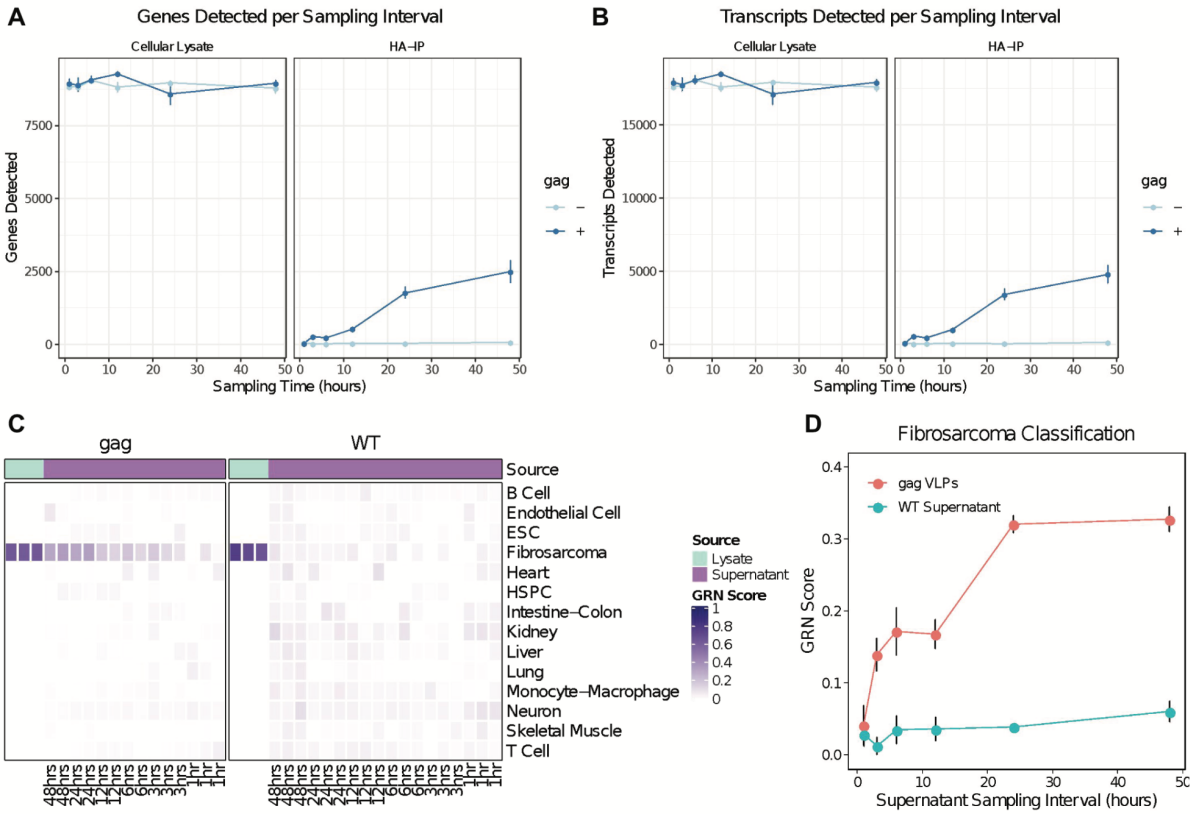
Supplemental Figure 3.9 Purification and characterization of VLPs with engineered envelopes. (**A**) Western blot on lysate, supernatant, and FLAG immunoprecipitation from HEK293T cell lines transfected with different constructs. (**B**) RNA-seq of supernatant purified for immunoprecipitation input. (**C**) RNA-seq replicate concordance of pGag+, pFLAG-VSV-G+ HEK293T supernatant after FLAG immunoprecipitation (FLAG-IP). (**D**) Transcript abundances for FLAG-IP on self-reporting supernatant vs. cell lysate for self-reporting cells. (**E**) Transcript abundances for FLAG-IP on self-reporting supernatant vs. input self-reporting supernatant.



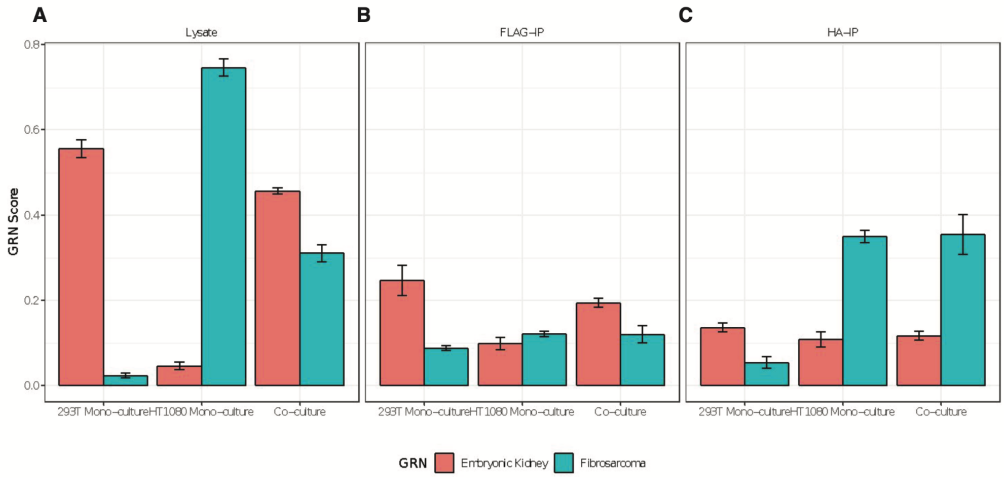
Supplemental Figure 3.10 Specificity of multiplexed immunoprecipitation-based isolation of epitope-tagged VLPs from transfected cells. (**A, B**) Western blots of cellular lysate, supernatant, and IP-purified media for both FLAG and HA-IP protocols. (**C**) CellNet classification of IP-purified VLPs from (A) and (B). (**D**) FLAG- or HA-tagged VLPs produced from HEK293T cells were processed with FLAG and HA-IP to assess the specificity of the purification to the target epitope. RNA-seq libraries were prepared from the input supernatant and all IP outputs.



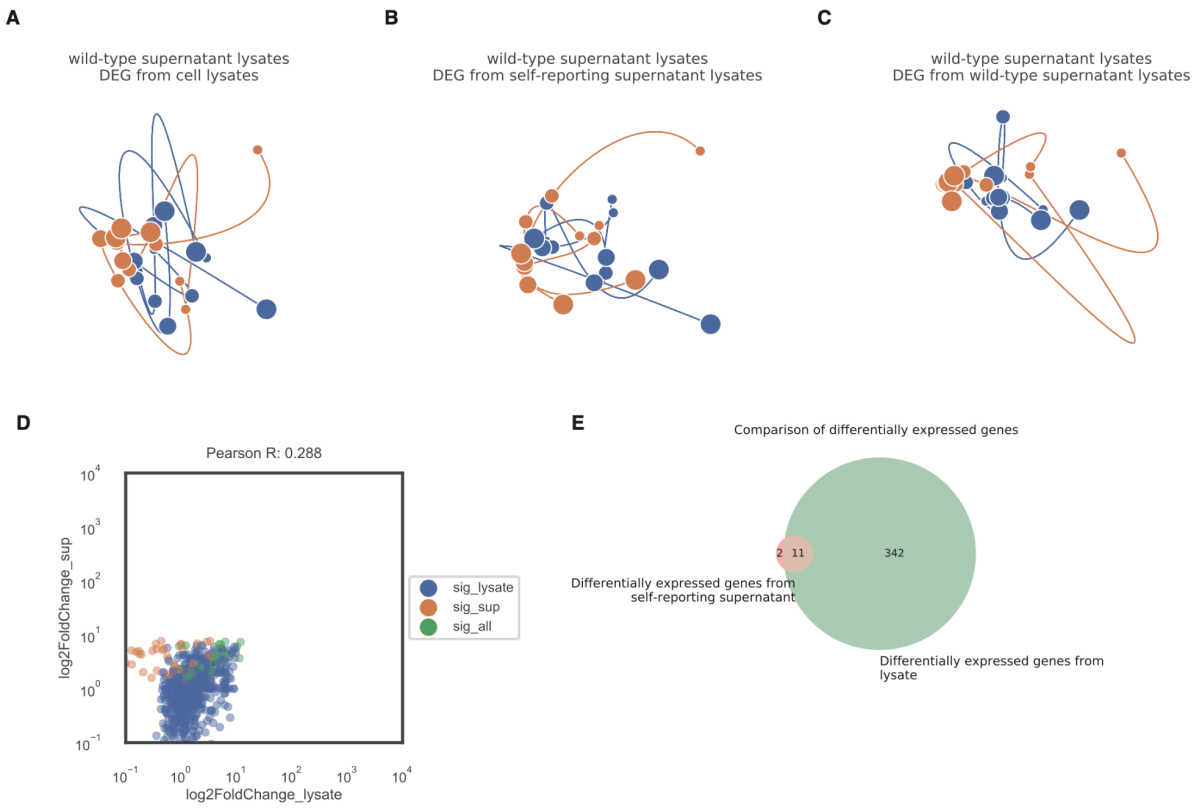
Supplemental Figure 3.11 Constitutive VLP production from cells stably expressing engineered Gag polyproteins and epitope-tagged VSV-G proteins. (**A**) Western blot from HEK293Ts demonstrates stable expression of Gag or Gag–RRM fusion proteins in addition to FLAG-VSV-G following lentiviral integration. Western blot for Gag on cellular media and FLAG-immunoprecipitated media demonstrates the constitutive production of FLAG-tagged VLPs. (**B**) Same as (A) for HT1080s expressing HA-VSV-G. (**C**) Co-culture experiment Western blot.



Supplemental Figure 3.12 Characterization of the VLP-derived RNAs as a function of sampling duration. (**A**) Genes detected as a function of length of sampling duration for HA-IP purified supernatants derived from Gag+ HT1080 cells. (**B**) Transcript isoforms detected as a function of sampling duration (HA-IP purified VLPs and cell lysate controls) in HT1080 cells. (**C**) CellNet classification per sampling interval length in HT1080 cells. (**D**) Fibrosarcoma GRN score as a function of sampling duration in HT1080 cells.



Supplemental Figure 3.13 CellNet classification of demultiplexed VLPs from HEK293T-HT1080 co-cultures. (**A**) GRN scores for lysates from HEK293T (FLAG-tagged VLPs) and HT1080 (HA-tagged VLPs) cells cultured independently and in co-culture. (**B**) GRN scores for FLAG-purified supernatants from HEK293T (FLAG-tagged VLPs) and HT1080 (HA-tagged VLPs) cells cultured independently and in co-culture. (**C**) GRN scores for HA-purified supernatants from HEK293T (FLAG-tagged VLPs) and HT1080 (HA-tagged VLPs) cells cultured independently and in co-culture.



Supplemental Figure 3.14 HT1080 TNFa stimulation. (**A**) PCA on wild-type HT1080 supernatant lysates using differentially expressed genes found in cell lysate RNA-seq (no clear ± TNFa separation observed). (**B**) PCA on wild-type HT1080 supernatant lysates using differentially expressed genes found in self-reporting supernatant lysate RNA-seq (no clear ± TNFa separation observed). (**C**) PCA on wild-type HT1080 supernatant lysates using differentially expressed genes found in wild-type HT1080 supernatant lysate RNA-seq (no clear ± TNFa separation observed). (**D**) Fold-change (TNFα vs. unstimulated) plot for supernatant lysate vs. cell lysate showing all differential expressed genes that are significant in either supernatant or lysate or both. (**E**) Venn diagram showing differentially expressed genes in supernatant and lysate upon TNF stimulation. Genes are filtered for significance if they appear significant in 2 or more time points.

4 Chromatin and transcriptional dynamics underlying human hematopoietic development identify regulators of lymphoid fate

Authors: Mohamad Ali Najia, Rebecca Soto, Caroline Kubaczka, Ran Jing, Trevor Bingham, Thorsten Schlaeger, Paul C. Blainey, Trista North, George Q. Daley

Status: This manuscript is in preparation for submission to a scientific journal

Contributions: M.N. designed and conceived the study; M.N. performed all experiments with assistance from C.K., R.J., and T.B.; R.S. performed all zebrafish experiments; M.N. performed computational analysis; M.N. wrote the manuscript with input from all authors. T.S., P.C.B., T.N., and G.Q.D. supervised the research and provided funding.

4.1 INTRODUCTION

The differentiation of human induced pluripotent stem cells (iPSCs) into hematopoietic lineages *in vitro* has broad applications in basic research and clinical medicine [5]. The design of experimental protocols to direct the differentiation of iPSCs to desired hematopoietic fates has largely been guided by understanding hematopoietic development in model organisms [12,235,236]. All hematopoietic cells emerge from an endothelial precursor, termed hemogenic endothelium, during embryonic development through a dynamic and complex process of endothelial-to-hematopoietic transition (EHT). Despite the general successes in recapitulating aspects of hematopoietic ontogeny in iPSCs protocols, our understanding of the molecular mechanisms driving hematopoietic emergence from hemogenic endothelium remain largely obscure.

The EHT process is regulated by an intricate balance of transcriptional regulators to ultimately repress endothelial identity and activate the hematopoietic program. Genetic and functional approaches provided important insights into transcription factors (TFs) regulating hematopoietic formation from hemogenic endothelium. For example, RUNX1 interacts in a heptad of TFs comprising FLI1, ERG, TAL1, LYL1, LMO2 and GATA2 to facilitate the EHT

process [237,238]. Downregulation of SOX7 within hemogenic endothelium is necessary for the subsequent emergence of hematopoietic progenitors [123,239]. The complex interplay between TFs and the need to reform gene networks during the transformation from endothelium to blood implies a significant degree of chromatin remodeling to facilitate the cell state transition. However, we lack a detailed characterization of the accessible chromatin landscape and transcription factor dynamics throughout the entire process of human EHT. Systematic efforts to characterize accessible chromatin landscapes during EHT from iPSC cultures would facilitate a more comprehensive understanding of the underlying gene regulatory dynamics, allow for the identification of molecular discordancies with *in vivo* hematopoietic population, and to nominate molecular interventions that promote the generation of desired cell types.

In this chapter, we characterized the dynamic transcriptional and chromatin landscapes during EHT in a cell type-resolved manner. We uncovered transcriptional regulators and *cis*-regulatory elements that drive the transition from endothelium to hematopoietic fates. Furthermore, using this rich molecular resource, we identified transcriptional regulators that are predicted to explain molecular discrepancies between in vitro derived hematopoietic progenitors and umbilical cord blood (UBC) derived progenitors. Functional follow up on one of those predictions, *KDM2B,* revealed a critical role in regulating lymphoid development within the context of human iPSC differentiation and zebrafish development.

4.2 RESULTS

4.2.1 A chromatin and transcriptional time course of human endothelial-to-hematopoietic transition

To comprehensively characterize the transcriptional and chromatin landscapes of human EHT, we differentiated iPSCs to hemogenic endothelium using an established step-wise serum-free protocol [119,146,236] (**Figure 4.1A**). CD34+ cells were isolated from EBs following 8 days of directed differentiation and cultured in supportive cytokines to facilitate EHT. Time course

profiling with broad lineage markers (CD34 and CD45) revealed the gradual acquisition of hematopoietic fates and attenuation of endothelial fates over 7 days in EHT-supportive conditions (**Figure 4.1B**). Furthermore, the frequency of CD34+CD45+ hematopoietic progenitors declined

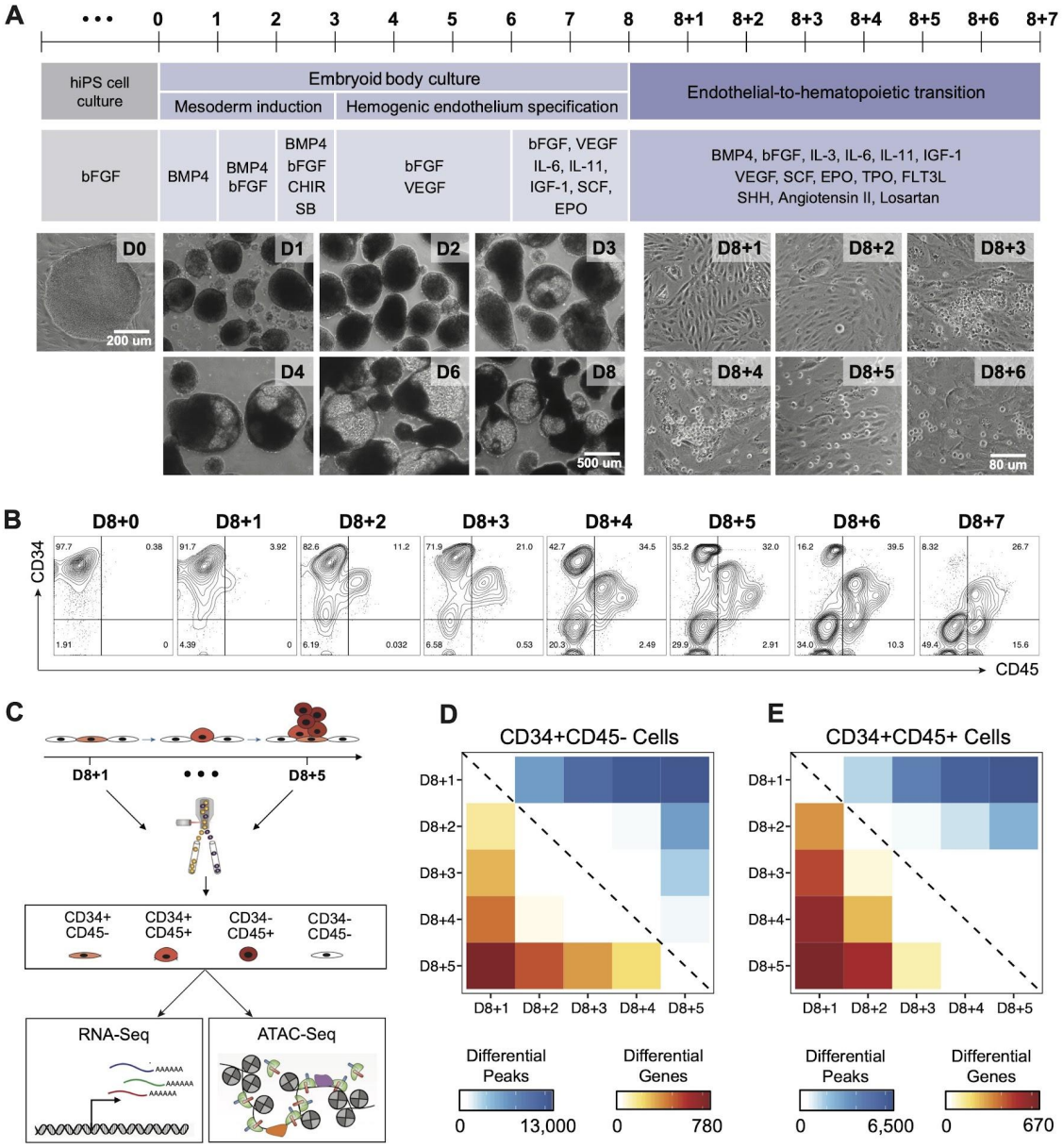


Figure 4.1. Temporal transcriptional and chromatin profiling of endothelial and hematopoietic populations throughout human EHT
(A) Schematic and representative images of the step-wise, defined protocol to generate CD34+ hemogenic endothelium from human iPSCs and endothelial-to-hematopoietic transition assay on the derived CD34s
(B) FACS gating scheme and time course profiling of CD34 and CD45 over EHT to endothelial and hematopoietic populations

(C) Schematic outline of generating matched transcriptome and open chromatin profiles in a cell type and time resolved manner during EHT
(D-E) Heatmap visualizing the number of differentially accessible peaks (top, blue shades) and differentially expressed genes (bottom, red shade) between time points for CD34+CD45- and CD34+CD45+ EHT populations.

with sustained culture in EHT-supportive conditions while a CD34-CD45+ population emerged, reflective of differentiation of hematopoietic progenitors. Thus, we FACS isolated cells based on CD34 and CD45 expression everyday over the course of EHT and performed RNA-Seq and ATAC-Seq to understand the dynamic transcriptional and chromatin processes underlying *in vitro* EHT (**Figure 4.1C**). We performed two independent iPSC differentiation experiments to obtain biological replicates for each population at each time point. Globally, we observed strong concordance in chromatin and transcriptional measurements within both CD34+CD45- endothelial (**Figure 4.1D**) and CD34+CD45+ hematopoietic populations (**Figure 4.1E**). The magnitude of differentially accessible loci and differentially expressed genes progressively changed over time, suggestive of dynamic cellular processes occurring within each population.

4.2.2 Global transcription factor activity governing EHT

Given the global and temporal changes in accessible chromatin and transcriptional landscapes during EHT, we first sought to determine transcriptional regulators that exhibited dynamic activity. We leveraged our ATAC-Seq data and utilized chromVAR [209] to identify transcription factor (TF) motifs associated with variably accessible loci across EHT populations and over time. Among the most variably accessible sequence motifs, we identified motifs from important hematopoietic TF such as *SPI1, BCL11A/B, GATA2,* and *GATA1* (**Figure 4.2A**), which suggests that these TFs may be driving different chromatin states during EHT. Additionally, motifs belonging to the AP-1 TF family and *TEAD4* motifs were among the most globally variable, which is consistent with the cooperative roles of AP-1 and TEAD4 in hemogenic endothelium for hematopoietic specification [240,241]. The motif for *SMARCC1*, a component of the SWI/SNF ATP-dependent chromatin remodeling complex, was also associated with variably accessible loci,

further suggesting chromatin remodeling processes during EHT. Next, we dissected differences in TF motifs across EHT populations and time points. Notably, the inferred activity of TFs was sufficient to delineate each FACS population (**Figure 4.2B**). Consistent with prior literature, *SOX, AP-1 and GFI1* motifs were selectively enriched in accessible chromatin of CD34+CD45- endothelial samples, whereas *RUNX, CBFB, SPI1, CEBP, and BCL11A/B* motifs

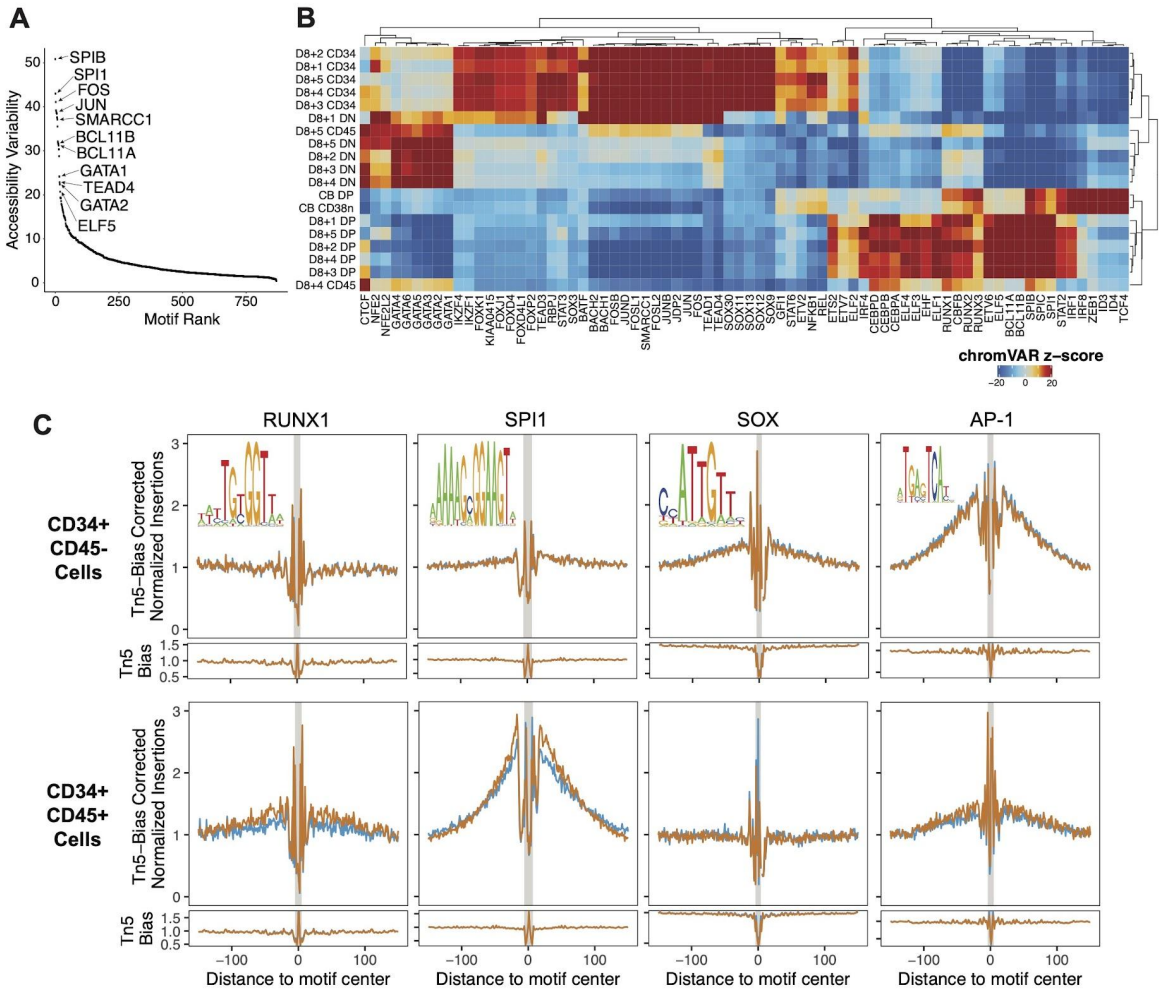


Figure 4.2. Activity of transcriptional regulators during EHT
(A) Transcription factor motifs with the greatest variability in chromatin accessibility across all ATAC-Seq samples.
(B) Heatmap showing changes in chromatin accessibility for the top TFs with greatest accessibility variability between all ATAC-Seq samples as measured by chromVAR. CD34 = CD34+CD45-, DP = CD34+CD45+, DN = CD34-CD45-, CD38n = CD34+CD45+CD38-
(C) ATAC-seq transcription factor footprinting plots visualizing the genome-wide Tn5 insertion density +/-150 bps of motif occurrences. Biological replicates are overlaid on each plot and the Tn5 bias plot for each motif is listed below each plot.

were selectively enriched within CD34+CD45+ hematopoietic samples. We also profiled hematopoietic progenitors from umbilical cord blood to understand how TF activity compares to the *in vitro* EHT-derived hematopoietic progenitors. While *RUNX, SPI1 and STAT2* motifs exhibited comparable activity, motifs belonging to *IRF8, ZEB1, ID3, ID4* and *TCF4* TFs exhibited reduced activity in EHT-derived CD34+CD45+ hematopoietic progenitors compared to umbilical cord blood progenitors (**Figure 4.2B**). Conversely, *CEPB* motifs were more accessible within EHT-derived hematopoietic progenitors compared to umbilical cord blood. These findings may suggest that differences in chromatin landscapes between umbilical cord blood and EHT-derived hematopoietic progenitors could be driven by these TFs. Collectively, this motif-centric analysis nominated several putative TFs with distinct activities across EHT and cord blood populations.

To better resolve TF binding dynamics during EHT, we profiled TF occupancy at base-pair resolution through TF footprinting. TFs that are actively bound to accessible DNA can protect their motif from transposition by the Tn5 transposase used in the ATAC-Seq assay while the DNA adjacent to the bound TF can be accessible and a substrate for transposition. These phenomena are collectively referred to as a TF footprint. We performed genome-wide footprinting analysis using TF motifs that exhibited high variability between CD34+CD45- endothelium and CD34+CD45+ hematopoietic progenitors, and observed two broad classes of TF activity (**Figure 4.2C**). First, TFs such as RUNX1 and SPI1 created footprints within CD34+CD45+ hematopoietic progenitors with an associated increase in flanking accessibility, but not in CD34+CD45- endothelium. These findings support that these TFs are locally opening chromatin during the transition from endothelial to hematopoietic fates. Conversely, the activity of SOX and AP-1 TFs attenuated between endothelial and hematopoietic populations (**Figure 4.2C**), which is consistent with the role of AP-1 factors and SOX17 to regulate and specify hemogenic endothelium [123,242–244]. Overall, these results provide global insights into the cell-type and time dependent TF regulatory programs active during EHT from iPSCs.

4.2.3 Dynamics of the accessible chromatin landscape in hemogenic endothelium and hematopoietic progenitors

Insights from our analysis on transcriptional regulators suggested that the chromatin landscape is actively remodeled during EHT. We therefore asked how individual *cis*-regulatory elements change over time within each EHT population. Focusing on CD34+CD45- endothelial samples

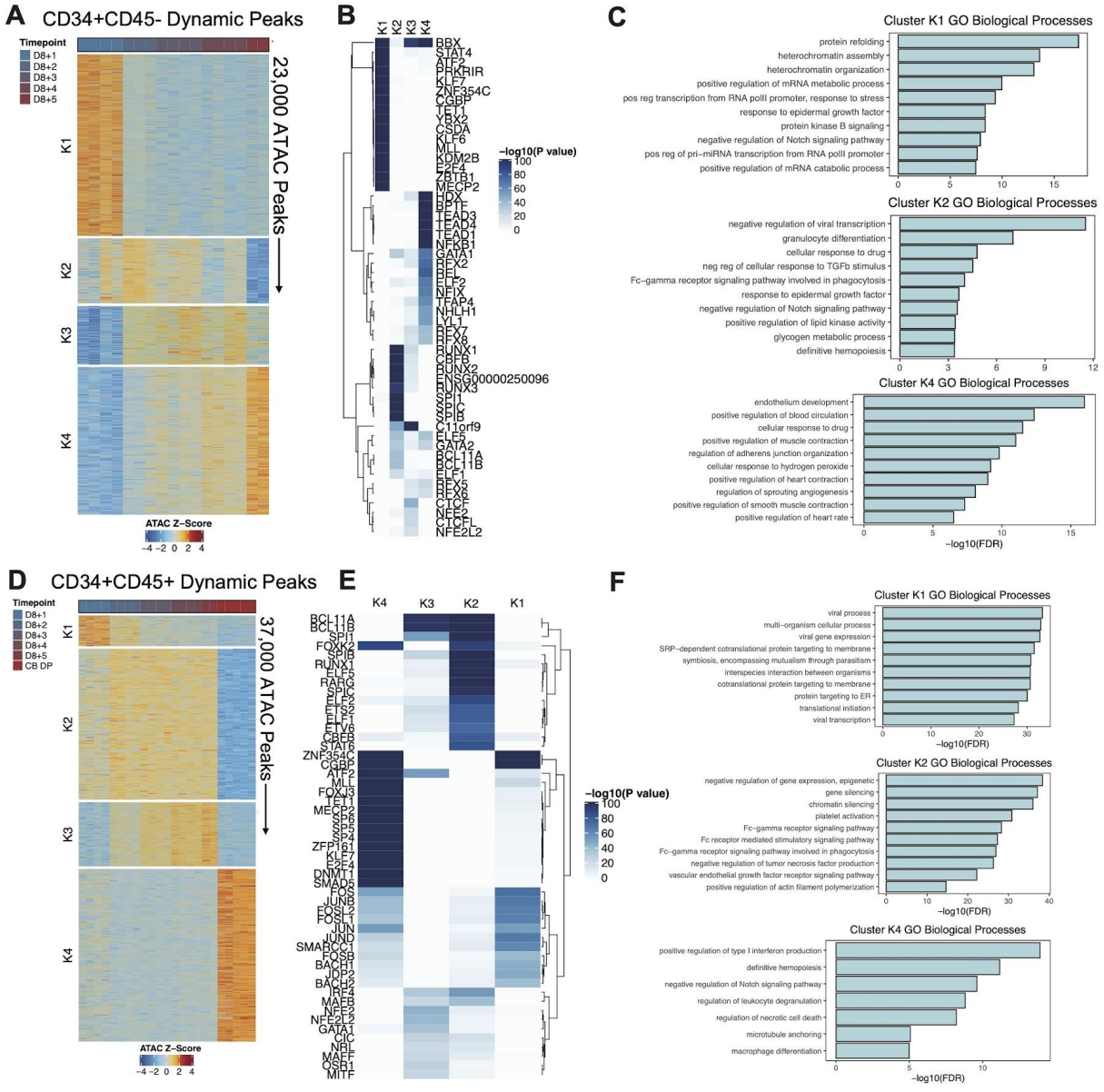


Figure 4.3. Cell-type and time-resolved dynamics of *cis*-regulatory elements
(A,D) Heatmap of clustered differentially accessible peaks using k-means clustering across EHT time points. The color bar corresponds to the accessibility Z-score of differentially accessible ATAC-Seq peaks.

(B,E) Enrichment of TF motifs in differentially accessible peak sets identified in (A,D). Enrichment is determined by a hypergeometric test –log10(P value) of the motif's representation within the cluster-specific peaks compared to all accessible loci.
(C,F) Top enriched Gene Ontology biological processes for clusters of peaks identified by GREAT analysis. The processes are ranked by the false discovery rate [FDR] of significance.

first, we identified approximately 23,000 ATAC-Seq peaks that displayed differential accessibility across time (**Figure 4.3A**) and k-means clustering segregated these dynamic peaks into four clusters. Cluster K1 contained regions of open chromatin showing the highest accessibility on D8+1 of EHT, which became less accessible with time. These peaks were particularly enriched for motifs of chromatin modifiers such as *TET1, MLL, KDM2B and MECP2,* (**Figure 4.3B**) and GREAT analysis [245] on these peaks further supported the involvement of chromatin-based biological processes (**Figure 4.3C**). Peaks within clusters K2 and K3 were primarily accessible between D8+2 and D8+4 of EHT and marked by motifs for hematopoietic TFs such as *RUNX1*, *CBFB*, *SPI1* and *BCL11A/B* (**Figure 4.3A-B**), suggesting that these *cis*-regulatory elements may be facilitating hematopoietic specification from hemogenic endothelium. In support of this hypothesis, GO Biological Processes such as "definitive hematopoiesis" and "granulocyte differentiation" were enriched within these peaks (**Figure 4.3C**). Finally, cluster K4 peaks gained accessibility incrementally as EHT progressed, with the greatest accessibility on D8+5 of EHT (**Figure 4.3A**). Since the CD34+ population isolated from EBs is a heterogeneous mixture of endothelial cell types with some containing hemogenic potential, the CD34+CD45- cells that remain at the end of EHT are most likely cells that do not have hemogenic potential, consistent with the enrichment of endothelium-related GO terms within these peaks (**Figure 4.3C**).

The temporally-specific set of *cis*-regulatory elements witin the CD34+CD45- endothelial population that enriched for hematopoietic-related TF motifs then prompted us to dissect the dynamic regulatory landscape within the CD34+CD45+ hematopoietic population (**Figure 4.3D-F**). We performed an analogous differential analysis of ATAC-Seq peaks, but also compared to CD34+CD45+ umbilical cord blood progenitors to understand the *cis*-regulatory elements that differ within *in vitro* EHT-derived hematopoietic progenitors. We identified four clusters of

dynamic peaks by k-means clustering (**Figure 4.3D**). Consistent with our analysis in CD34+CD45- endothelial populations, a set of accessible loci (cluster K2) were accessible throughout all time points and enriched for hematopoietic related TF motifs (**Figure 4.3E**). However, in contrast to the cluster K2 peaks in CD34+CD45- population, the cluster K2 peaks in the CD34+CD45+ hematopoietic population were enriched for GO Biological processes involved in chromatin and gene silencing (**Figure 4.3F**). These data would suggest that these *cis*-regulatory elements may contribute to the silencing of an endothelial transcriptional program. We also identified a significant number of peaks that were accessible within umbilical cord blood progenitors but largely inaccessible within the *in vitro* derived progenitors (cluster K4, **Figure 4.3D**). The regulation of inflammation, and in particular type I interferons (**Figure 4.3E-F**), may account for the discrepancy between chromatin landscapes. Exposure to interferons has been implicated in the developmental emergence and maturation of HSCs [89,99,246]. Overall, this analysis identified the set of dynamic *cis*-regulatory elements underlying EHT and provided insights into the transcriptional regulators and biological processes that differentiate EHT-derived hematopoietic progenitors from umbilical cord blood progenitors.

4.2.4 Systematic comparison of *in vitro* derived hematopoietic progenitors to *in vivo* progenitor populations

Thus far, our analysis of *cis*-regulatory elements and transcriptional regulators within EHT populations and across time demonstrated different chromatin landscapes between EHT-derived hematopoietic progenitors and immunophenotypically matched umbilical cord blood progenitors. We therefore sought to leverage these chromatin landscapes to determine the *in vivo* hematopoietic populations that most closely resemble the EHT-derived hematopoietic progenitors. To do this, we utilized ATAC-Seq data on FACS isolated hematopoietic stem and progenitor cells [191] as reference populations and applied CIBERSORT [247] to quantify the relative fraction of these reference populations within the EHT CD34+CD45+ populations. We identified

cis-regulatory elements that uniquely defined each HSPC population (**Figure 4.4A**), although, HSCs and MPPs shared a largely overlapping signature of *cis*-regulatory elements. The difficulty in resolving HSC and MPP populations is also reflected in single cell chromatin analysis of human HSPCs [248,249], so we therefore limit any conclusions distinguishing between HSCs and MPPs. Nonetheless, the application of this signature matrix via CIBERSORT

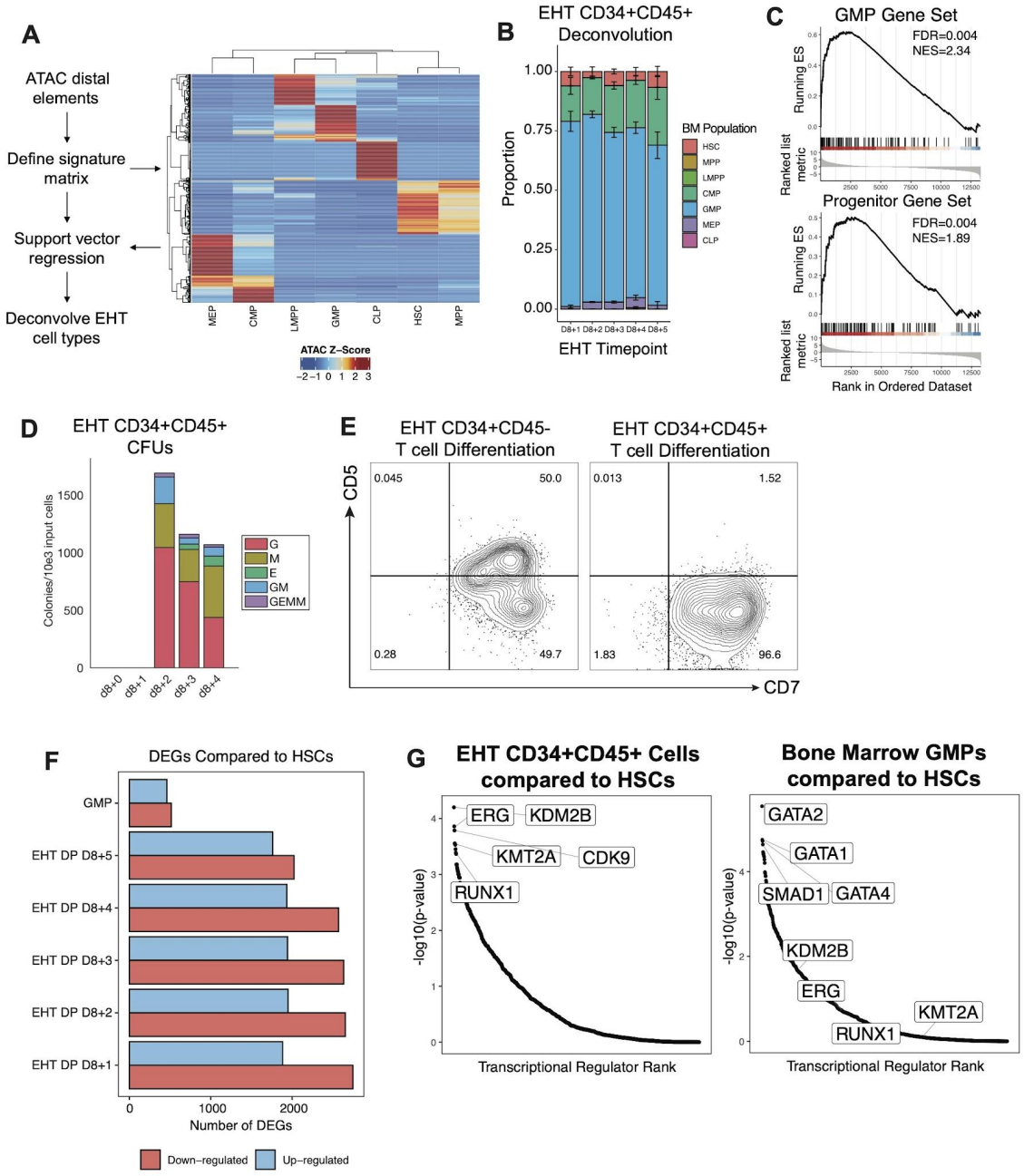


Figure 4.4. Systematic comparison of EHT-derived hematopoietic progenitors to in vivo progenitor populations
(A) CIBERSORT signature matrix visualizing sample-specific *cis*-regulatory elements on bone marrow HSPCs
(B) CIBERSORT deconvolution of the EHT CD34+CD45+ population using the signature matrix from (A).
(C) GSEA on GMP and Progenitor gene sets from [250] comparing differentially expressed genes between EHT CD34+CD45+ cells and bone marrow HSCs.
(D) Quantification of colony-forming potential from EHT CD34+CD45+ cells. E, erythroid; GM, granulocyte, monocyte; M, monocyte; G, granulocyte; GEMM, granulocyte, erythroid, monocyte, megakaryocyte.
(E) Flow cytometry contour plots of CD5 and CD7 following 14 days in T cell-supportive differentiation conditions to derive T cell lineages.
(F) Quantification of the number of differentially expressed genes for GMPs and EHT CD34+CD45+ populations in comparison to bone marrow HSCs.
(G) Ranked set of transcriptional regulators predicted by BART analysis to regulate down-regulated genes compared to bone marrow HSCs

revealed a substantial myeloid signature consisting of CMP and GMP states within our CD34+CD45+ EHT hematopoietic progenitors (**Figure 4.4B**). Differential gene expression analysis and GSEA also demonstrated significant enrichment of GMP and progenitor gene sets[250], further corroborating the chromatin similarity to *in vivo* myeloid progenitors (**Figure 4.4C**). CD34+CD45+ cells produced from *in vitro* EHT also maintained myeloid differentiation potential, as assayed by CFUs where the majority of colonies generated were granulocytes or macrophages (**Figure 4.4D**). Furthermore, while CD34+CD45- endothelium isolated from EBs harbored the capacity to generate CD5+CD7+ proT cells within T cell-supportive differentiation conditions, CD34+CD45+ cells produced from EHT only generated CD7+ progenitors, suggesting a deficiency in lymphoid differentiation (**Figure 4.4E**).

The myeloid bias exhibited by hematopoietic progenitors produced during EHT prompted us to determine the molecular underpinnings that explain why these cells lack features of multipotency. Several thousand genes were differentially expressed between CD34+CD45+ EHT progenitors and *in vivo* isolated bone marrow HSCs (**Figure 4.4F**). We performed BART analysis [251] to nominate transcriptional regulators that may account for the transcriptional discrepancy between HSCs and EHT progenitors. Interestingly, ERG, KMT2A, and RUNX1 were among the top predictions that regulate genes downregulated in EHT hematopoietic progenitors compared to HSCs (**Figure 4.4G**). These predictions are in strong alignment with a transcription factor cocktail that, when overexpressed during EHT, confers multilineage

hematopoiesis within engrafted NSG mice [146]. KMT2A is known to regulate *HOXA* genes, and HOXA5 and HOXA9 were both necessary along with ERG, RUNX1 and LCOR to produce engraftable hematopoietic cells from EHT. Notably, in addition to these predictions consistent with experimental findings, BART analysis prioritized chromatin modifier *KDM2B* as the top transcriptional regulator to account for transcriptional discrepancies between EHT progenitors and HSCs. We then replicated this analysis with GMPs and identified *GATA2* as the top prediction (**Figure 4.4G**), consistent with the critical role of GATA2 in maintaining HSCs [188,252–254]. These data suggest that the predicted transcriptional regulators for EHT hematopoietic progenitors are not simply a result of molecular differences due to a GMP-like state. Rather, the predictions suggest molecular differences as a result of *in vitro* specification. This multi-omic analysis enabled the molecular characterization of EHT-derived hematopoietic progenitors and predictions of transcriptional regulators that could promote increased molecular resemblance to *in vivo* hematopoietic populations.

4.2.5 KDM2B promotes lymphoid commitment from hematopoietic progenitors

We then asked whether modulation of *KDM2B* could enhance lymphoid commitment from EHT-derived hematopoietic progenitors. We focused on promoting lymphoid commitment because of the myeloid-biased state of EHT-derived hematopoietic progenitors. Since BART predictions were performed on downregulated genes, we hypothesized that overexpression of *KDM2B* could promote lymphoid fates. We designed a doxycycline (dox)-inducible expression vector to express the *KDM2B* ORF (**Figure 4.5A**). As a member of the PRC1.1 complex, KDM2B plays a central role in post-translationally modifying histones by demethylating H3K36 via a Jumonji domain. Therefore, we also designed a mutant ORF with the Jumonji domain omitted to determine if histone demethylation activity is necessary for potential phenotypes. We derived CD34+ hemogenic endothelium as previously described, infected the CD34+ population with dox-inducible vectors following one day in EHT culture conditions and administered dox 24

hours later to induce expression (**Figure 4.5B**). We observed robust induction 24 hours after dox administration (**Figure 4.5C**), however, induction did not significantly alter the proportion of CD34+CD45+ hematopoietic progenitors generated after seven days in EHT culture conditions (**Figure 4.5D**). We then FACS isolated CD34+CD45+ hematopoietic progenitors and initiated T cell differentiation. Notably, we observed a significant proportion of CD5+CD7+ proT cells in KDM2B overexpression conditions relative to no dox controls (**Figure 4.5E**). Furthermore,

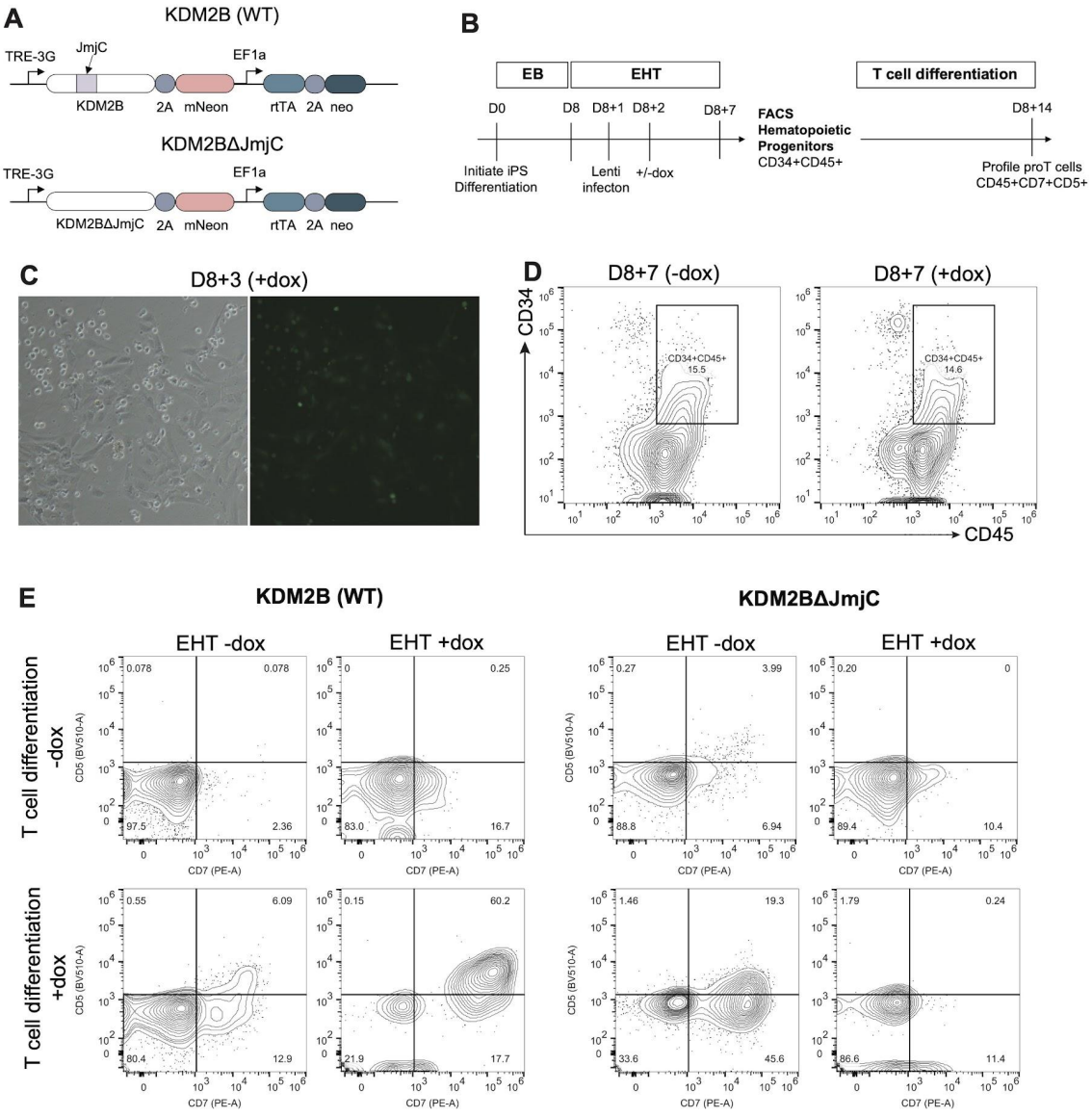


Figure 4.5. Overexpression of KDM2B enhances lymphoid development from iPS cells
(A) Schematic construct maps of dox-inducible overexpression vectors
(B) Experimental scheme to test KDM2B overexpression during EHT and T cell differentiation

(C) Brightfield (left) and GFP (right) microscopy visualizing gene induction following lentiviral infection of EHT cultures and dox administration.
(D) Flow cytometry profiling of CD34 and CD45 within EHT cultured infected with dox-inducible vectors but treated with or without dox.
(E) Flow cytometry contour plots of CD5 and CD7 following 14 days in T cell-supportive differentiation conditions to derive T cell lineages from D8+7 CD34+CD45+ EHT cells.

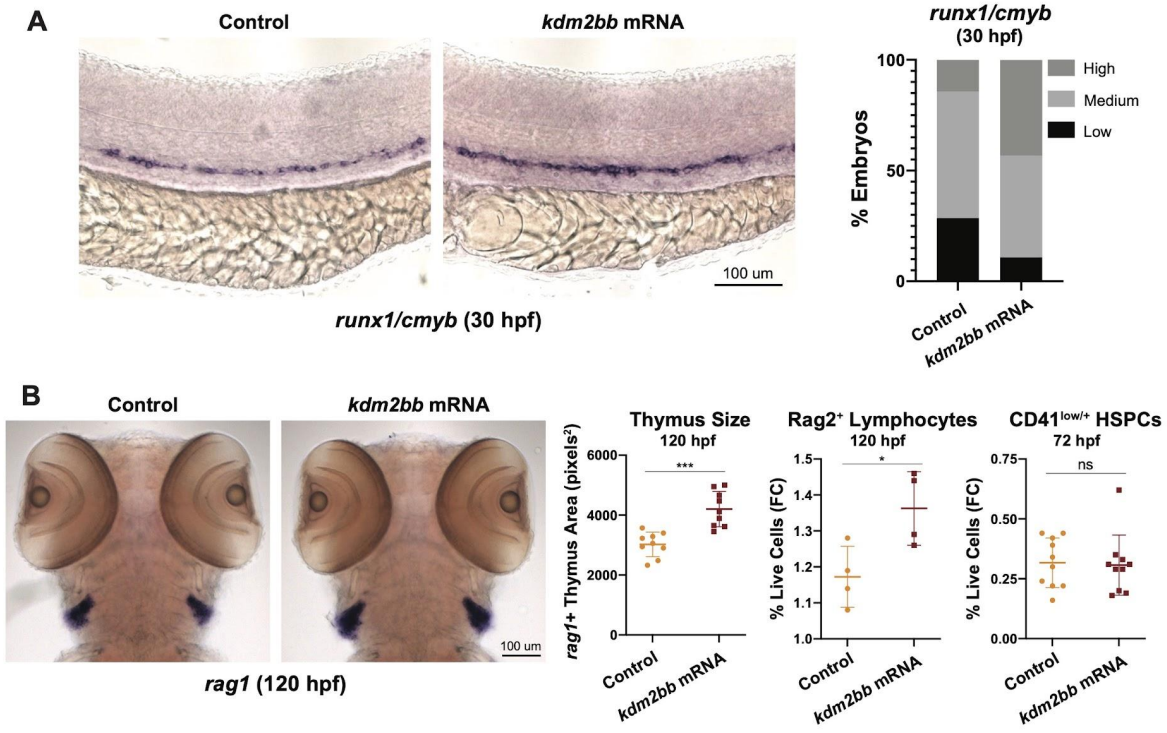


Figure 4.6. KDM2B regulates T cell lineages in zebrafish
(A) WISH for *runx1*/*cmyb* in control/mRNA-injected embryos at 30 hpf.
(B) WISH for *rag1* in the thymus of control and mRNA-injected embryos at 120 hpf (left). Quantification of thymic area from microscopy via Fiji. Quantification of Rag2:GFP+ lymphocytes at 120 hpf and CD41$^{low/+}$ HSPCs at 72 hpf by flow cytometry (FC) as a percentage of live cells in control and kdm2bb mRNA conditions (right; each dot represents a biological replicate; two-tailed unpaired Student's t-test, *p < 0.05. Error bars indicate SD).

sustained dox administration during EHT and T cell culture resulted in the highest frequency of CD5+CD7+ proT cells. The enhanced frequency of proT cells appeared to be Jumonji domain dependent as overexpression of a Jumonji domain lacking ORF did not phenocopy the wildtype protein.

We then asked whether there is a conserved role of *KDM2B* in lymphoid commitment. We utilized zebrafish as an experimentally tractable model that has proven invaluable for

identifying genes impacting hematopoietic development. We performed mRNA overexpression of the zebrafish ortholog, *kdm2bb* within single cell embryos and observed increased expression of conserved HSPC markers, *runx1* and *cmyb*, at 30 hours post-fertilization (hpf) (**Figure 4.6A**). Increased proportions of *rag1*+ and *rag2*+ lymphocytes were observed at 120 hpf (**Figure 4.6B**), in line with observations from human iPSC experiments. Notably, the frequency of CD41$^{low/+}$

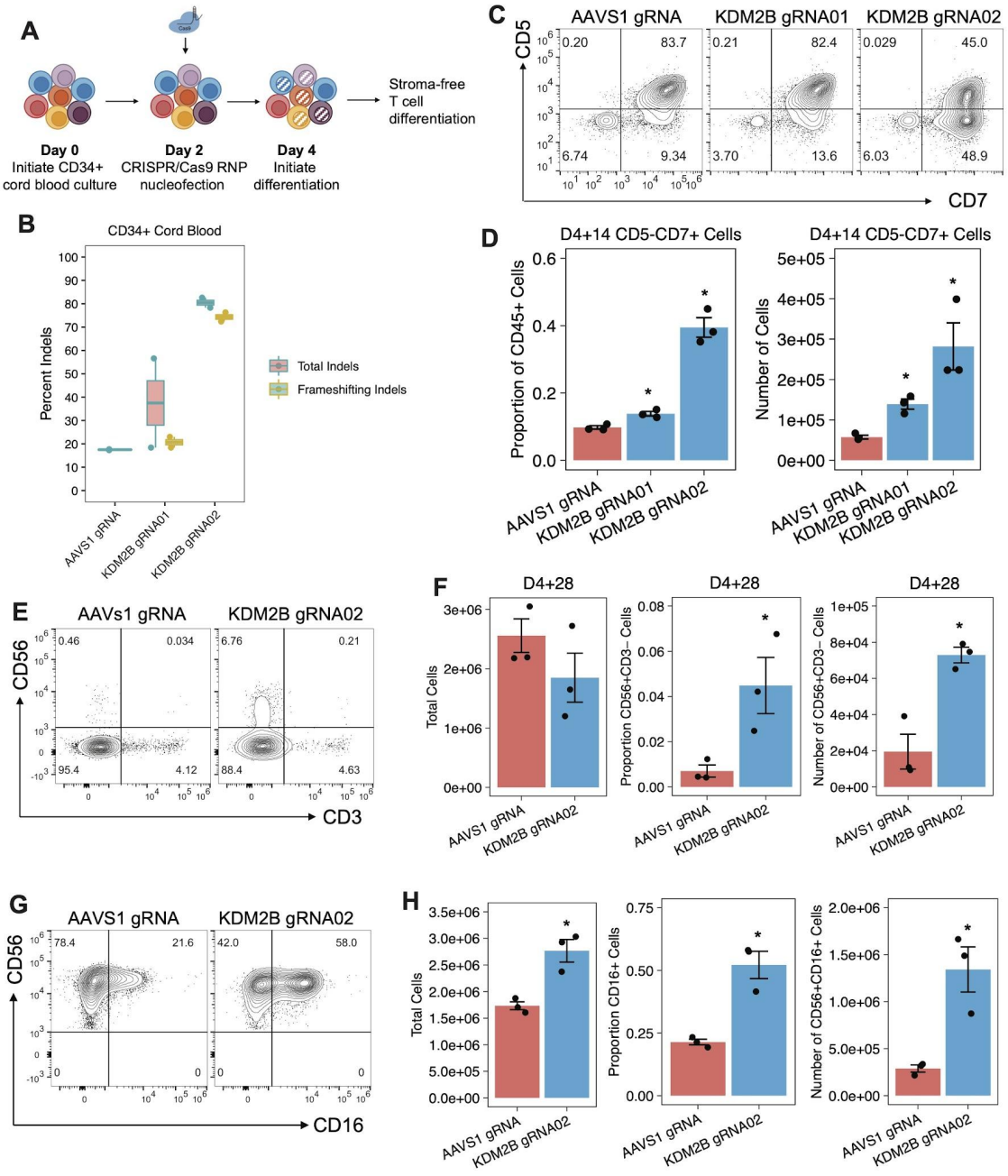


Figure 4.7. Knockout of KDM2B alters lymphoid progenitors and facilitates NK development
(A) Experimental scheme to knockout genes within CD34+ umbilical cord blood and assess T lymphoid differentiation
(B) Quantification of indels via TIDE analysis 48 hours following nucleofection of CRISPR/Cas9 RNPs into CD34+ HSPCs
(C) Flow cytometry contour plots of CD5 and CD7 following 14 days in T cell-supportive differentiation conditions to derive T cell lineages.
(D) Quantification of CD5, CD7 populations across n=3 replicate differentiation experiments.
(E) Flow cytometry contour plots of CD56 and CD3 following 28 days in T cell-supportive differentiation conditions to derive T cell lineages.
(F) Quantification of total viable cells and CD56, CD3 populations following 28 days in T cell supportive differentiation conditions across n=3 replicate experiments.
(G) Flow cytometry contour plots of CD56 and CD16 following 28 days in NK cell-supportive differentiation conditions to derive NK cell lineages.
(H) Quantification of total viable cells and CD56, CD16 populations following 28 days in NK cell supportive differentiation conditions across n=3 replicate experiments.

HSPCs at 72 hpf was consistent between control and *kdm2bb* overexpression conditions, suggesting that *kdm2bb* overexpression does not expand HSPCs generally, but promoted lymphoid commitment from an earlier stage of hematopoietic development.

Finally, to further understand the contributions to lymphoid commitment, we performed knockout experiments within umbilical cord blood hematopoietic stem and progenitor cells since they maintain lymphoid competency. We utilized CRISPR/Cas9 RNPs to knockout *KDM2B* within HSPCs (**Figure 4.7A**) and identified one gRNA that efficiently generated frameshifting indels (**Figure 4.7B**). Following knockout, we initiated T cell differentiation and observed a significant reduction in lymphoid progenitors expressing the T cell associated marker, CD5 (**Figure 4.7C-D**). These results are in line with the overexpression phenotype during EHT demonstrating an increased frequency of CD5+CD7+ cells. Continued differentiation within T cell supportive conditions interestingly resulted in the acquisition of CD56+CD3- natural killer (NK) cells (**Figure 4.7E-F**), potentially as a consequence of an altered distribution of lymphoid progenitors early in the differentiation. Finally to follow up on these findings, we replicated the knockouts in cord blood HSPCs and performed an NK-specific differentiation. Under these conditions, knockout of *KDM2B* resulted in the enhanced production of CD16+ NK cells following four weeks of differentiation (**Figure 4.7G-H**), further implicating a role for KDM2B within lymphoid lineage development.

4.3 DISCUSSION

In this chapter, we systematically dissected the chromatin and transcriptional landscapes of hematopoietic fate specification during EHT from iPSCs. We identified the temporal activity of transcription factors within hemogenic endothelium and hematopoietic progeny, which allowed for the identification of differences in molecular landscapes between iPSC-derived hematopoietic progenitors and umbilical cord blood progenitors. Furthermore, we sought to improve iPSC differentiation protocols by enhancing the production of lymphoid fates by predicting transcriptional regulators that may explain molecular discrepancies between iPSC-derived hematopoietic progenitors and umbilical cord blood progenitors. We experimentally validated the top prediction, KDM2B in iPSC differentiation and zebrafish contexts to enhance progenitor T cell commitment. Knockout of KDM2B in umbilical cord hematopoietic progenitors attenuated proT cell fates, yet enhanced NK cell development, suggesting a role in early lymphoid fate selection. Further investigation is needed to understand the mechanism through which KDM2B facilitates T cell fates. KDM2B, a component of the PRC1.1 complex, is a H3K36me2 demethylase. We demonstrated that the enhancement of progenitor T cells from iPSCis dependent on the Jumonji domain, suggesting a role for H3K36. Moreover, PRC2 complexes have also been shown to regulate early hematopoietic specification and T cell differentiation [255,256]. Thus, because H3K27 and H3K36 methylation are mutually exclusive histone marks, we hypothesize that overexpression of KDM2B results in increased removal of H3K36 methylation to enable a suitable substrate for EZH2 to deposit H3K27 methylation. In further support of this hypothesis, inhibition of EZH2 in human or mouse hematopoietic progenitors facilitates NK cell development, yet attenuates T cell maturation [256]. In conclusion, this study revealed new insights into molecular regulators of lymphoid fate and serves as a framework to refine differentiation protocols for the generation of lymphoid cells from iPSCs.

4.4 METHODS

4.4.1 Cell Culture

Human iPSCs (1157.2 line from Boston Children's Hospital) were cultured on hES-Qualified Matrigel Basement Membrane Matrix (Corning) with StemFlex media (Gibco) for gene editing, single cell cloning, and expansion culture. iPSCs were fed every other day and propagated with ReLeSR (Stem Cell Technologies) once the cultures reached 70% confluence. HEK293FT cells were obtained from ATCC and cultured in complete media, consisting of DMEM+GlutaMAX and 10% FBS. Cells were fed every other day and propagated upon reaching 70% confluency using TrypLE.

4.4.2 Differentiation of human iPSCs to hemogenic endothelium

Human iPSCs were cultured on irradiated mouse embryonic fibroblasts (Gibco) for one passage prior to the initiation of hematopoietic differentiation. A comprehensive protocol for the generation of CD34+ hemogenic endothelium is available as a supplemental protocol and adapted from Sugimura et al 2017, Ditadi et al 2015, Sturgeon et al 2014. Briefly, iPSC colonies were detached with Collagenase IV for 10 minutes at 37C and seeded into ultra-low attachment 10cm dishes (Corning) to form EBs in SFD media, supplemented with GlutaMAX (ThermoFisher), ascorbic acid (1 mM), α-monothioglycerol (400 μM), holo-transferrin (150 μg/mL), and BMP-4 (10 ng/uL). bFGF (5 ng/uL) was added to the EB cultures following 24 hours. The media was completely changed following an additional 18 hours to day 0 containing media, supplemented with bFGF (5 ng/mL), CHIR99021 (3 μM) and SB431542 (6 μM). Following 30 hours, the media was completely changed to StemPro-34 media (Thermo Fisher Scientific), supplemented with ascorbic acid (1 mM), holo-Transferrin (150 μg/mL), α-monothioglycerol (400 μM), bFGF (5 ng/uL), and VEGF (15 ng/mL). On day 6 of differentiation the media was changed to day 3 containing media supplemented additionally with IL-6 (10

ng/mL), IL-11 (5 ng/mL), IGF-1 (25 ng/mL), SCF (50 ng/mL) and EPO (2 U/mL) and then cultured until day 8. Cultures were maintained in a 5% CO2, 5% O2, 90% N2 environment.

4.4.3 Endothelial-to-hematopoietic transition assay

EBs were dissociated after 8 days of directed differentiation using the Embryoid Body Dissociation Kit (Miltenyi Biotec) according to the manufacturer's protocol. The resulting single cell suspension was stained through a 40 uM filter and then CD34+ cells were enriched with MACS beads (Miltenyi Biotec) according to the manufacturer's protocol. CD34+ cells were plated onto Matrigel-coated 24-well plates (100,000 cells/well) and cultured at 5% O2 in StemPro-34 based media, supplemented with BMP4 (10 ng/mL), bFGF (5 ng/mL), IL-3 (30 ng/mL), IL-6 (10 ng/mL), IL-11 (5 ng/mL), IGF-1 (25 ng/mL), VEGF (5 ng/mL), SCF (100 ng/mL), TPO (30 ng/mL), FLT-3L (10 ng/mL), EPO (2 U/mL) and SHH (20 ng/mL). Unless otherwise stated, CD34+ hemogenic endothelium underwent EHT over the course of 7 days with regular half media changes every other day. Floating and/or adherent cells were collected for downstream functional or molecular assays. For time course profiling, adherent cells were singularized with Accutase and pooled with the floating cell fraction. Cells were then stained with antibodies for 20 minutes at room temperature in the dark then FACS sorted for RNA-Seq and ATAC-Seq.

4.4.4 Bulk RNA-Sequencing and processing

RNA from at least 5,000 cells was extracted using a 2X TCL lysis buffer (Qiagen cat. no. 1070498). At least two technical replicates were prepared per sample using the SMART-Seq2 protocol [69], with some modifications. Briefly, RNA was purified from cellular lysate using 2.2X RNA SPRI beads (Beckman Coulter cat. no. A63987) then immediately reverse transcribed in the presence of a template switching oligo (Exiqon) with Maxima RNase H-minus RT (Thermo Fisher Scientific cat. no. EP0751) using a polyT primer containing the ISPCR sequence. Whole

transcriptome amplification proceeded with KAPA HiFi HotStart ReadyMix using an ISPCR primer according to the following thermal program: 98C for 3 minutes, 21 cycles of 98C for 15 seconds, 67C for 20 seconds, and 72C for 6 minutes, and a final extension step of 72C for 5 minutes. The amplified cDNA was cleaned with 0.8X DNA SPRI beads (Beckman Coulter cat. no. B23318). Ten nanograms of amplified cDNA was tagmented at 58C for 10 minutes in a 10 uL reaction containing 2 uL of 5X tagmentation buffer (50 mM Tris-HCl, 25 mM MgCl2 pH 8.0), 2 uL of Tris Buffer (10 mM Tris-HCl, 1% Tween-20 pH 8), and 4 uL Nextera Tn5 transposase (Illumina). The reaction was stopped with 1% SDS and incubated at 72C for 10 minutes, then 4C for 3 minutes. The tagmented library was cleaned with 1X DNA SPRI beads followed by an indexing PCR with NEBNext High Fidelity polymerase to incorporate sample index barcodes and Illumina flow cell handles. The index PCR proceeded with the following thermal program: 72C for 3 minutes, 98C for 30 seconds, 12 cycles of 98C for 10 seconds, 60C for 30 seconds, 72C for 30 seconds, and a final extension step of 72C for 5 minutes. The final libraries were pooled, diluted and sequenced on a NextSeq with a 75-cycle High Output Kit with the following read configuration: 36-8-8-38. Reads were then pseudoaligned with kallisto and a TPM matrix generated for all samples. Differentially expressed genes were identified using DESeq2 and defined as a |fold-change| > 2 and FDR < 0.01. Prediction of transcriptional regulators was performed using BART via the online website interface (https://zanglab.github.io/bart/).

4.4.5 Bulk ATAC-Sequencing

EHT cells following FACS isolation were used as input for ATAC-Seq. Nuclei isolation, tagmentation and library construction were followed as described in the Omni-ATAC-Seq protocol [228]. The concentration of the final ATAC-Seq libraries was quantified using a High Sensitivity DNA Bioanalyzer Assay (Agilent) in the size range of 100-1000 bp. The libraries were pooled and sequenced on a NovaSeq to a depth of at least 10 million aligned, deduplicated reads. The ATAC-Seq libraries were processed with the PEPATAC pipeline [229] to generate

hg38-aligned, deduplicated BAM files. A non-redundant, non-overlapping peaks set was generated as outlined in [231] and a count matrix constructed for all samples. The matrix was CPM normalized using the edgeR package in R followed by quantile normalization. Differentially accessible peaks were identified using the raw count matrix and DESeq2 in R. Differentially accessible peaks were then used as input to GREAT (http://bejerano.stanford.edu/great/public/html/). Motif enrichment within differentially accessible peaks was computed by using CIS-BP motif matches from motifmatchr and calculating the significance of the enrichments with a hypergeometic test in R. Transcription factor footprinting was performed via the ChrAccR package in R using genome-wide occurrences of motifs. Global, genome-wide activity was measured using the chromVAR package in R where the raw counts within ATAC-Seq peaks an CIS-BP motif matches within these peaks (determed with motifmatchr) were used as inputs. GC bias-corrected deviations were then computed using the chromVAR “deviations” function.

4.4.6 Hematopoietic colony assay

Analysis of hematopoietic colony potential was performed by plating 1,000 gene edited CD34+ cord blood cells into methylcellulose media (MethoCult SF H4636, Stem Cell Technologies) and cultured on 3.5cm dishes in a humidified chamber at 20% O2. Colonies consisting of erythroid lineages (BFU-E), myeloid lineages (granulocytes/macrophages) or mixed lineages were quantified blindly by an independent researcher after 14 days of culture.

4.4.7 CRISPR/Cas9 knockouts in CD34+ umbilical cord blood

Genetic knockouts of select genes were generated in primary hematopoietic stem and progenitor cells (HSPCs) via nucleofection of CRISPR/Cas9 ribonucleoproteins (RNPs). Two independent knockouts were generated for each gene using separate gRNAs and knockouts were replicated across two different donor pools of human umbilical cord blood. First, CD34+

umbilical cord blood cells (AllCells) were thawed from liquid nitrogen storage via dropwise addition of RPMI 1640 + 10% FBS. Cells were centrifuged at 200g for 8 minutes and subsequently washed with FACS buffer (PBS + 2% FBS). Cells were cultured for 48 hours prior to nucleofection at a density of 250,000 cells/mL in SFEM II media (StemCell Technologies, 09605) supplemented with 100 ng/mL SCF, TPO, FLT3L, and IL-6 (Peprotech) within a 5% O2, 5% CO2, 37C incubator. Next, the top two protospacer sequences for each desired gene were selected from the Brunello genome-wide knockout library (Broad Institute) based on the Rule Set 2 score. A control gRNA targeting the *AAVS1* locus was designed in Benchling. All gRNA sequences were synthesized by IDT as an Alt-R CRISPR/Cas9 sgRNA. RNPs were complexed by incubating 105 pmol of Cas9 protein (IDT, 1081058) with 125 pmol of sgRNA in a 5 uL total volume for 10 minutes at room temperature. CD34+ cord blood HSPCs were washed with FACS buffer and 125,000 cells were resuspended in 20 uL of P3 Primary Cell Nucleofection Solution (Lonza, V4XP-3032) per nucleofection. RNPs were nucleofected into HSPCs with 3.85 uM electroporation enhancer (IDT, 1075915) using a 16-reaction nucleovette and pulse code DZ-100 on a Lonza 4D-Nucleofector. Cells were cultured for 36-48 hours at 5% O2, 5% CO2 in 96-well U-bottom plates (Corning, 351177) then subsequently harvested for genomic DNA extraction and initiation of *in vitro* differentiation assays.

4.4.8 Quantification of indel frequencies

Genomic DNA was extracted from at least 50,000 cells using a custom extraction buffer (1 mM CaCl2, 3 mM MgCl2, 1 mM EDTA, 10 mM Tris-HCl, 1% Triton X-100, and 0.2 mg/ml Proteinase K) and then subjected to the following thermal program: 65C for 10 minutes then 95C for 15 minutes. gRNA-targeted genomic loci were PCR dialed-out from genomic DNA using NEBNext High Fidelity Master Mix (NEB, M0541S) and primers that flank ~200 bps of the expected Cas9 cut site. PCR amplicons were gel extracted from a 2% agarose gel and submitted for Sanger Sequencing using forward and reverse PCR primers. Indel frequencies for each gRNA were

quantified using TIDE analysis (http://tide.nki.nl) where the reference nucleotide sequence was derived from a mock-nucleofected control. The proportion of frameshifting indels was then determined from the total TIDE indel frequency estimates.

4.4.9 *In vitro* T cell differentiation

Umbilical cord blood HSPCs or iPSC-derived cells were differentiated to T cells using the StemSpan T cell Generation Kit (Stem Cell Technologies, 09940) per the manufacturer's instructions with n=3 replicate differentiations per gRNA in 24-well plates. After 14 days of differentiation, cells were collected and 50,000 cells/sample were replated into new 24-well plates for continued differentiation.

4.4.10 Molecular cloning

Plasmid vectors were designed in Benchling and generated via Gibson Assembly. The wildtype KDM2B ORF was amplified from a whole transcriptome amplification cDNA library using specific primers and cloned with a P2A-mNeon translation reporter into a custom dox-inducible vector. NEB Stable Competent *E. coli* (NEB cat #C3040H) were transformed with the Gibson Assembly reaction and grown overnight at 30C on agar plates with 50 mg/ml carbenicillin. Individual colonies were picked for liquid culture in LB media supplemented with ampicillin and plasmid DNA was subsequently isolated using a QIAGEN Plasmid Plus Midi Kit with endotoxin removal. Plasmid sequences were fully verified by Sanger sequencing (Genewiz, Inc).

4.4.11 Lentivirus production

HEK293FT cells were seeded into 15-cm plates or multi-well plates at a density of 100,000 cells/cm2. After 20 hours, cells were transfected with pMD2.G (Addgene #12259), psPAX2 (Addgene #12260), and a lentiviral transfer plasmid (2:3:4 ratio by mass) using Lipofectamine 3000 (Thermo Fisher L3000015). Media was exchanged after 4 hours and supplemented with 2

mM caffeine 20 hours post-transfection to increase viral titer. Viral supernatant was harvested 48 hours after transfection and filtered through 0.45 μm PVDF filters (Millipore SLHVR04NL).

4.4.12 Zebrafish whole mount *in situ* hybridization

Embryos were fixed in 4% paraformaldehyde overnight at 4°C, and processed with established protocols and published probes for *runx1*, *cmyb* and *rag1* (http://zfin.org/ZFIN/Methods/ThisseProtocol.html). Phenotypic variation (n ≥ 20 embryos/condition, n ≥ 3 replicate clutches) was qualitatively analyzed and graphically depicted as the percentage of total scored exhibiting high (increased)/ medium (normal)/ low (decreased) expression compared with the median of sibling controls. Images were acquired with a Zeiss Axio Imager A1/Axio Cam MRC using Axiovision LE software. Thymic area quantified using Fiji (NIH).

4.4.13 Zebrafish mRNA injections

5' capped *kdm2bb* mRNA was synthesized from commercially available plasmid constructs (Horizon Discovery) using the mMessage mMachine SP6 Kit (Invitrogen) and injected at the 1-cell embryo stage at a concentration of 300 ng/μl. Injected embryos were stage-matched to control siblings.

4.4.14 Zebrafish embryo dissociation and flow cytometry

Pools of embryos/larvae were dissociated by Liberase (Sigma) in a 34°C circulating water bath and resuspended in 1X PBS/1mM EDTA. FACS was performed on LSRFortessa (BD) with 5 fish per replicate. Prior to analysis, cells were labeled with 5nM SYTOX Red dead cell stain

(Thermo Fisher Scientific), as described (Cortes et al., 2016). Data analyzed using FlowJo X Software.

5 Conclusions & Future Directions

In this thesis, I describe novel molecular tools and biological insights that collectively advance our understanding of dynamic cellular processes. This body of work spans multiple layers of gene regulation that converge on transcription with the goal of studying, measuring and directing hematopoietic cell fate decisions. In particular, we investigate enhancers and chromatin modifiers contributing to the transcriptional regulatory networks governing hematopoietic differentiation and specification. We further extend our ability to understand dynamic cellular processes by engineering a synthetic RNA export system in mammalian cells to measure temporal transcriptional changes within living cells. Collectively, this thesis provides evidence that 1) the modulation of chromatin machinery instructs lymphoid fates during hematopoietic differentiation and development, 2) retroviral machinery can be repurposed to enable live-cell readouts of cellular transcriptomes, and 3) the derepression of TEs during lymphoid differentiation highlights their involvement within NK cell lineage choice. While these observations and novel technologies have the potential to influence current practice, there are several additional avenues for future interrogation and research that may further their biological and therapeutic reach.

5.1. Future Directions

5.1.1 iPSC derived NK cells for adoptive cell therapy

In **Chapter 2**, we arrived at the surprising finding that TE derepression enacted by knocking out regulators of heterochromatin formation lead to the acquisition of NK cells within B and T cell supportive differentiation conditions. We then further characterized the NK cells derived from *EHMT1* or *TRIM28* knockout HSPCs and identified distinct molecular and phenotypic properties despite sharing a core NK signature. Many efforts are emerging for developing and engineering

NK cell-based cancer immunotherapies. However, there are challenges to overcome, such as clinical-grade *ex vivo* expansion, limited *in vivo* persistence and sub-optimal infiltration to solid tumors. iPSCs represent an amenable platform to install genetic modification to endow NK cells with new functionality or augment their activity, as well provide scalable production of targeted cell populations[257–264]. Thus, the modulation of *EHMT1* or *TRIM28* during iPS differentiation could be a logical application of the insights derived in **Chapter 2** to endow iPS-derived NK cells with distinct effector functions. The impact of small molecule inhibitors of EHMT1/2 can be assessed during specific differentiation windows and/or *TRIM28* degron iPSCs can be engineered to tunably and selectively modulate TRIM28 during differentiation. Notably, knockout of *EHMT1* produced 3-fold more NK cells than the *AAVS1* gRNA control, which could aid in enhancing *in vitro* production of iPS-NK cells. Furthermore, a higher proportion of CD16+ NK cells were derived by knockout of *EHMT1*, which reflects a more mature, peripheral blood-like state. Conversely, *TRIM28* knockout NK cells were more potent IFN-γ producers, which can further augment the inherent ability of NK cells to produce interferons upon stimulation. This property may be of particular therapeutic interest in light of a recent study demonstrating loss of IFNγR signaling in solid tumors as a mechanism of resistance to CAR-based immunotherapies[265]. Finally, our observation that knockout of *TRIM28* results in elevated IFN-γ production raises interesting questions about the mechanism of epigenetic “priming” therapies to induce viral mimicry in cancers. Studies have demonstrated that TE derepression in solid or liquid tumors results in the production of ERV proteins that stimulate adaptive immune responses. It would be interesting to determine if the chromatin inhibitors used in these studies are also acting on NK cells to derepress TEs and enhance their inflammatory signaling to contribute to the anti-tumor effect.

5.1.2. Single cell resolution and multimodal readouts of self-reporting

A fundamental goal in molecular and stem cell biology is to understand how a cell's starting state (genetic, phenotypic and/or environmental) influences its phenotypic response to stimuli. Cellular heterogeneity in various contexts has been well documented, where seemingly similar cells can behave differently over time [266–270]. A challenge with trying to understand molecular determinants of cellular behavior is that most genome-wide profiling methods destroy the cell, which precludes follow-up experiments on the same cells. Recently, attempts to non-destructively measure transcriptomes from individual, living cells over time employ specialized hardware to extract cytoplasmic aliquots of mRNAs[271]. While this approach represents a technological advance, a limitation is the low throughput of cells that can be profiled, which hinders its utility depending on the rarity of the phenomena amongst single cells that one is intending to measure. A core advantage of self-reporting is that the technology is genetically encodable, and thus, allows for scalable throughput across a population of cells (**Chapter 3**). However, further technology development will be necessary to achieve single cell resolution. In this thesis, we demonstrate that the incorporation of different epitope tags into the VSV-G envelope protein enables multiplexing of different conditions within bulk cultures. An extension of this approach would be to utilize nucleic acid barcodes that are expressed within cells and get packaged in VLPs along with mRNAs. Optimization and refinement of droplet-based mRNA sequencing assays may allow the isolation of VLPs and profiling of the VLP-derived mRNAs to obtain single cell (or lineage) resolution. Repeated sampling on the same population of self-reporting cells may enable the reconstruction of lineage-level transcriptional trajectories.

5.1.3. Modulation of TEs during in vitro hematopoietic development

The temporal dissection of chromatin and transcriptional landscapes during endothelial-to-hematopoietic transition from iPSCs (**Chapter 4**) revealed several additional findings that could warrant further investigation. In particular, inflammatory signatures, such as type I interferons,

were transcriptionally attenuated and inaccessible by ATAC-Seq in iPSC-derived hematopoietic progenitors compared to umbilical cord blood hematopoietic progenitors. The up-regulation of interferon genes have also been observed during HSC maturation *in vivo* following emergence from hemogenic endothelium[89]. Furthermore, several studies have highlighted the importance of basal inflammatory levels in regulating hematopoietic development, particularly interferon signaling across zebrafish and mouse models [91,93,95,97,99]. Recently, TE-derived RNAs were discovered to be activators of RIG-I-like receptors during embryonic hematopoiesis, which ultimately induce inflammatory signals necessary for hematopoietic stem and progenitor cell generation. In this thesis, we provide evidence for upregulation of type I interferon genes and corresponding enhancement of IFN-γ production within *TRIM28* knockout NK cells (**Chapter 2**). Thus, similar chemical/genetic approaches which we employed to modulate regulators of heterochromatin could be applied transiently during hematopoietic specification from iPS cells to determine if 1) modulation of heterochromatin results in TE derepression and corresponding upregulation of inflammatory signaling, and 2) whether such modulations generate cells that are molecularly and phenotypically more similar to umbilical cord blood hematopoietic progenitors.

5.2. Conclusions

This dissertation spans multiple levels of gene regulation to collectively advance our understanding of dynamic cellular processes. We interrogated the role of TEs within human hematopoiesis identifying unique contributions to the regulatory wiring of lymphoid cells and their dynamic contributions to NK cell development. Furthermore, we repurposed retroviral machinery to create a live-cell reporter of transcriptional states and enable the measurement of transcriptional trajectories over time. Finally, we identify novel regulators of lymphoid fate by dissecting the molecular transitions underlying hematopoietic specification from iPSCs. The application and further investigation of insights derived from this thesis may enable new therapeutic opportunities for the treatment of hematological disease and cancer.

Appendix A: Inducible CRISPR-interference to study coding and non-coding elements during pluripotent stem cell differentiation

Authors: Mohamad Ali Najia, Trevor Bingham, Ran Jing, Thorsten Schlaeger, George Q. Daley

Contributions: M.N. designed and conceived the study; M.N. performed all experiments with assistance from T.B. and R.J; M.N. wrote the manuscript with input from all authors. T.S. and G.Q.D. supervised the research and provided funding.

A.1 INTRODUCTION

Induced pluripotent stem (iPS) cells are ideal sources to derive somatic cells for allogeneic cell therapies, in addition to serving as convenient *in vitro* models to study disease, development and differentiation. The differentiation of iPS cells consists of multiple highly orchestrated steps, temporally controlled by networks of transcription factors (TFs). TFs bind to *cis*-regulatory elements and recruit transcriptional machinery to modulate gene expression and establish lineage- and cell type-specific gene regulatory networks. In a developmental context, knockout-based approaches to perturb genes or *cis*-regulatory elements can often cloud scientific interpertations of a factor's role during differentiation due to pleiotropic or embryonic lethal phenotypes. Thus, the precise temporal control of gene expression is necessary to guide differentiation to desired cell types, and genetic tools to selectively manipulate regulatory elements are needed to understand the molecular circuitry that governs differentiation.

CRISPR-interference (CRISPRi), a catalytically dead Cas9 fused to a transcriptional repressor domain, has emerged as a facile and programmable molecular tool to selectively repress genomic loci[101]. When complexed with a gRNA targeting a desired locus, CRISPRi leads to the recruitment of repressive chromatin modifying machinery that results in transcriptional knockdown at promoters[103,105] or altered chromatin state at *cis*-regulatory

elements[107,109]. CRISPRi avoids the introduction of double-stranded DNA breaks, which enables reversible repression and temporal, titratable control over gene expression or *cis*-regulatory elements[111,113]. Furthermore, in contrast to a diversity of editing outcomes from catalytically active Cas9, CRISPRi generally confers a more homogeneous loss-of-function across a population of cells[115]. These distinct advantages enable CRISPRi to be an ideal approach to dissect the regulatory logic governing the differentiation of iPS cells to desired somatic cells.

In this work, we engineered iPS cells with CRISPRi machinery under the inducible control of a doxycycline-responsive promoter in order to study the role of individual *cis*-regulatory elements and genes during hematopoietic development. We demonstrate robust CRISPRi-mediated repression within iPS cells, as well as during embryoid body (EB) based differentiation. We further demonstrate that CRISPRi can be utilized during specific developmental windows, such as endothelial-to-hematopoietic transition (EHT) to dissect the temporal requirements and cell type-specific manifestations of individual enhancers to mediate the generation of hematopoietic progenitors. The inducible CRISPRi iPS cells generated in this study serve as a broadly enabling platform for basic science discovery and translational endeavors to generate desired somatic cell types.

A.2 RESULTS

To enable modular CRISPRi applications in iPS cells, we first generated clonal iPS cells with robust, inducible expression of the CRISPRi machinery. gRNA expression vectors can then be introduced into this established cell line to enable programmable repression. We constructed a piggyBac transposon expression vector to drive expression of a dCas9-KRAB transgene under the control of a dox-inducible promoter (**Figure A.1A**). This approach will enable long-term expression within iPS cells and limit transgene silencing that can result from lentiviral vectors. We utilized an mCherry translation reporter in order to determine cells by flow cytometry or microscopy that induce the CRISPRi machinery upon dox treatment. Following piggyBac

transposition and single cell cloning, we identified a karyotypically normal iPS clone (**Figure A.1B**) that resulted in rapid and homogeneous induction of dCas9-KRAB within 24 hours of dox administration (**Figure A.1C**) and negligible transcriptional leakiness in the absence of dox

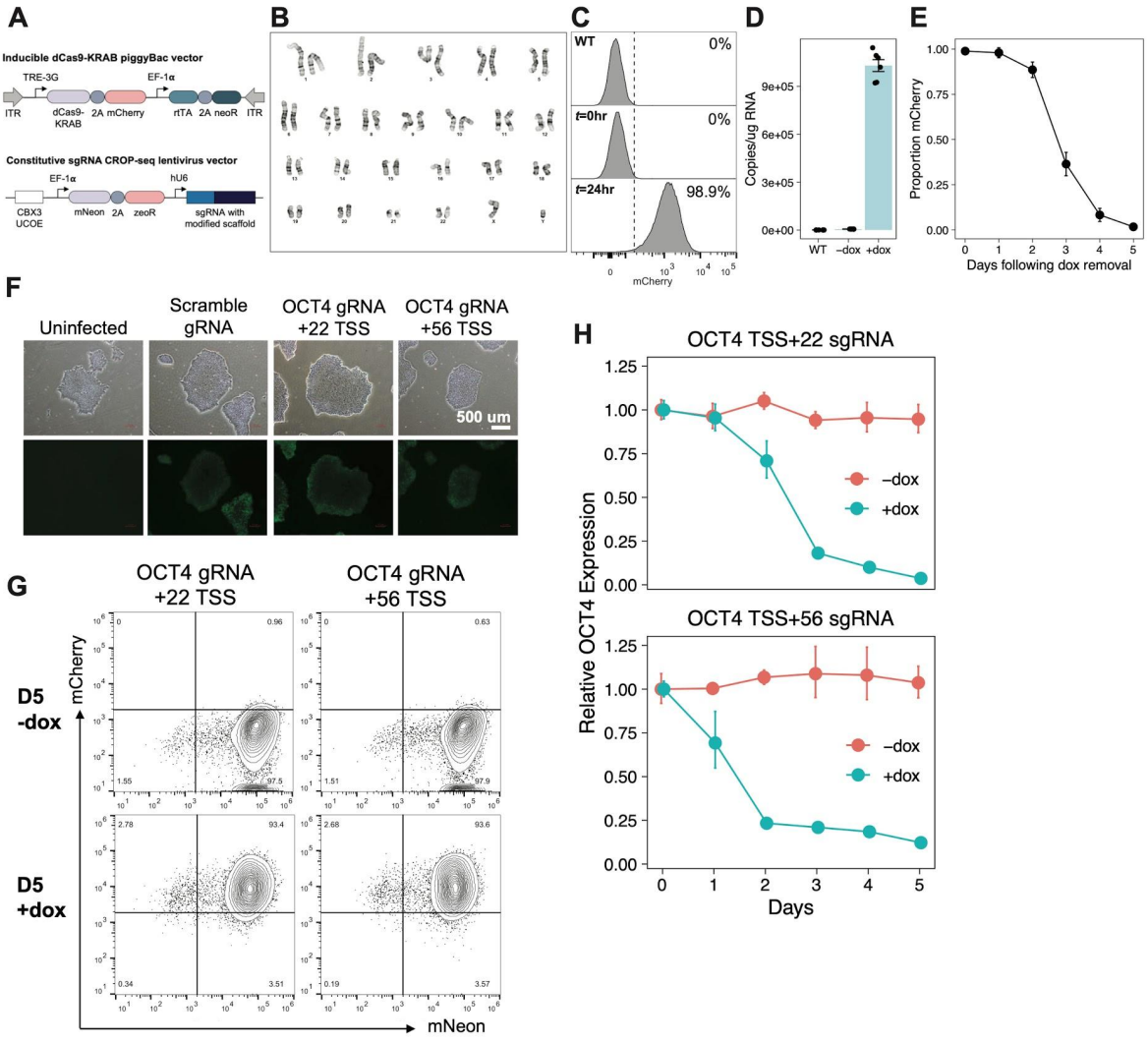


Figure A.1. Inducible CRISPR-interference in iPS cells
(A) Construct maps of the dox-inducible dCas9-KRAB piggyBac expression vector (top) and gRNA lentivirus expression vector (bottom).
(B) Karyotype of the clonal iPS line containing an integrated dCas9-KRAB piggyBac vector from (A).
(C) mCherry flow cytometry and **(D)** qPCR on the dCas9-KRAB transgene following 24 hours of dox treatment (1 ug/mL). Error bars represent standard deviation across n=3 replicates.
(E) Relaxation kinetics of dCas9-KRAB expression following withdrawal of dox assayed by flow cytometry for mCherry. Error bars represent standard deviation across n=3 replicates.
(F) Brightfield and GFP microscopy of inducible dCas9-KRAB iPS cells infected with CROPseq vectors.
(G) Flow cytometry for mNeon and mCherry following 5 days of dox treatment on CRISPRi iPS cells expressing OCT4 gRNAs.

(H) Validation of CRISPRi-mediated knockdown in iPS cells following lentiviral infection of gRNA vectors targeting the TSS of OCT4 and five days of dox treatment (1 ug/mL). Relative expression of OCT4 was normalized against ACTB using the 2^(–ΔΔCT) method.

(**Figure A.1D**). The expression of the dCas9-KRAB transgene returned to baseline within five days following dox removal (**Figure A.1E**), demonstrating tunable induction and relaxation of

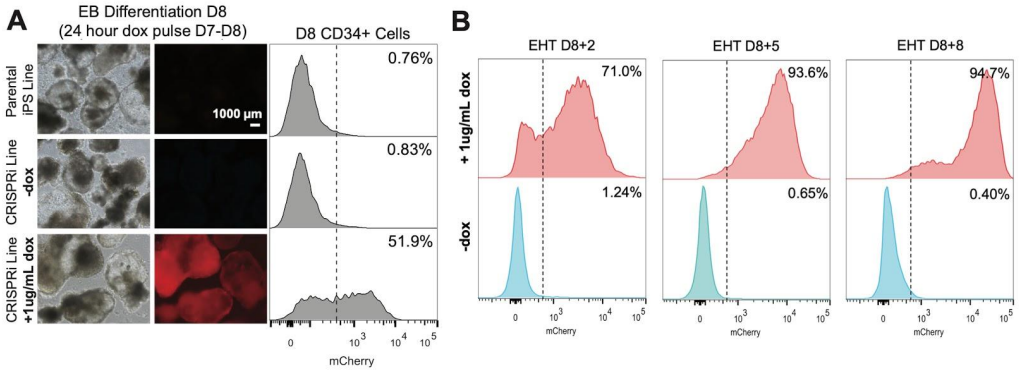


Figure A.2. Activity of CRISPRi machinery during iPS cell differentiation
(A) CRISPRi machinery is inducible during embryoid body differentiation evidenced by mCherry microscopy (left) and specifically within iPS-derived CD34+ cells by flow cytometry (right) following a 24-hour dox pulse.
(B) Flow cytometry time-course of mCherry expression during EHT demonstrating CRISPRi iPS cells retain hematopoietic differentiation capacity and can robustly induce expression of dCas9-KRAB during EHT.

the CRISPRi machinery. To validate CRISPRi-mediated repression, we first designed gRNAs targeting the transcriptional start site of pluripotency transcription factor, OCT4 and cloned them into a CROPseq gRNA expression vector (**Table S1, Figure A.1A**). We then infected the inducible dCas9-KRAB iPS line with the CROPseq gRNA vectors (**Figure A.1F**) and administration of dox resulted in robust transcriptional knockdown of OCT4 (**Figure A.1-H**).

Next, we characterized CRISPRi repression during specific time windows of iPS cell differentiation. We performed EB-based differentiation to derive hemogenic endothelium [116,119]. Administration of dox to EBs in a 24 hour pulse was sufficient to induce a majority of cells within the aggregates (**Figure A.2A**). We then isolated the CD34+ fraction of cells after 8 days of differentiation and plated them in a monolayer culture to undergo EHT. Sustained administration of dox during EHT resulted in ~95% of cells to induce dCas9-KRAB over 8 days (**Figure A.2B**).

These results are supportive of the long-term expression that can be achieved with our piggyBac expression vector and indicate induction of CRISPRi machinery is robust within iPS-derived cells following several weeks of differentiation. We then examined the application of inducible CRISPRi to interrogate the function of *cis*-regulatory factors during specific temporal

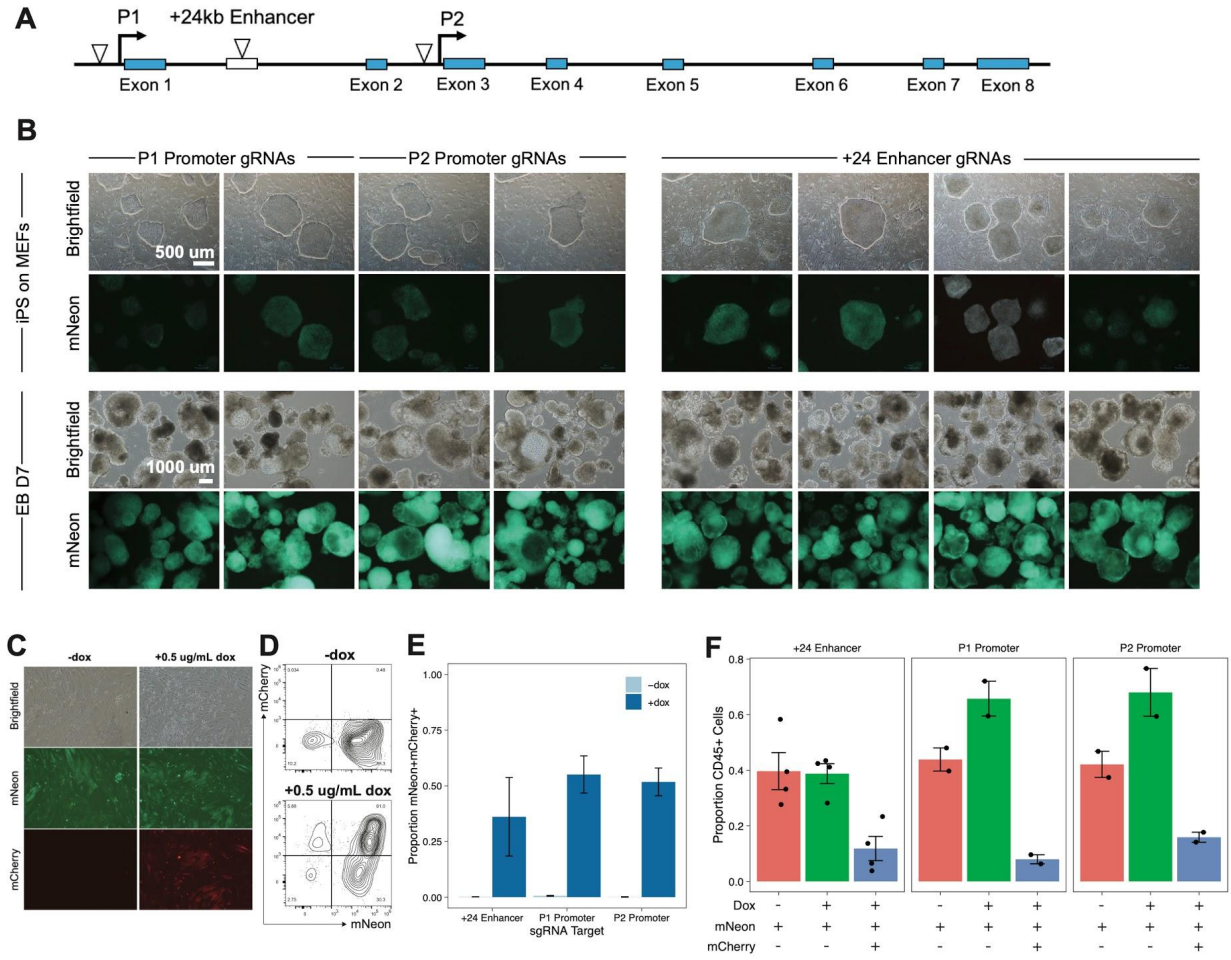


Figure A.3. CRISPRi interrogates cis-regulatory elements during endothelial-to-hematopoietic transition
(A) Schematic of the human RUNX1 locus. Triangles note gRNAs targeting the P1 and P2 promoters as well as the +24kb enhancer in order to validate CRISPR-interference at cis-regulatory elements and TSSs during EHT.
(B) Microscopy of inducible CRISPRi iPS cells expressing RUNX1-targeting gRNAs during EB-based hematopoietic differentiation.
(C) Representative microscopy and **(D)** flow cytometry of mNeon and mCherry expression from iPS-derived CD34s on d8+3 of EHT targeting RUNX1 regulatory elements. Endothelial and hematopoietic cells efficiently express mNeon (+90%) and at 0.5 ug/mL of dox treatment approximately half the cells induce expression of dCas9-KRAB (mCherry).
(E) Quantification of the proportion of mCherry+mNeon+ cells on d8+3 of EHT expressing +24kb enhancer gRNAs (n=4), P1 promoter gRNAs (n=2) and P2 promoter gRNAs (n=2).
(F) Proportion of CD45+ cells on d8+8 of EHT measured by flow cytometry for various reporter subpopulations. Each dot represents a distinct gRNA and error bars reflect one standard deviation. n=3 differentiation replicates.

windows. As a positive control, we focused on *cis*-regulatory elements known to regulate the expression *RUNX1*, a critical transcription factor required for hematopoietic commitment from hemogenic endothelium [121,123,125]. The human *RUNX1* gene is expressed from two distinct promoters and is regulated by an intronic enhancer located 24 kb from the P1 promoter [127]. We designed gRNAs targeting each of these RUNX1 regulatory elements (**Figure A.3A, Table S1**) and expressed them within our inducible CRISPRi iPS cells (**Figure A.3B**). We then generated EBs for each of the gRNA lines and differentiated them to hemogenic endothelium (**Figure A.3B**). We then isolated CD34+ hemogenic endothelium and plated the cells as a monolayer to undergo EHT. We administered a suboptimal dose of dox such that half the cells induced the CRISPRi machinery (**Figure A.3C-E**). This allowed us to compare induced and uninduced fractions within a gRNA condition and to control for variability observed between differentiation conditions. By flow cytometry, we could identify the mCherry+mNeon+ fraction of cells, indicative of CRISPRi induction and the mCherry-mNeon+ fraction which is uninduced. Consistent with the critical role of RUNX1 during EHT, we observed that gRNAs targeting RUNX1 regulatory elements resulted in a significant attenuation in the emergence of CD45+ hematopoietic cells (**Figure A.3F**). These data overall demonstrate the applicability of CRISPRi to selectively and tunably modulate *cis*-regulatory elements and genes during desired temporal windows of iPS cell differentiation.

A.3 DISCUSSION

In this work, we establish an inducible CRISPRi iPS line and demonstrate its utility to modulate *cis*-regulatory elements and genes during hematopoietic development. This *in vitro* platform can be useful to understand the genetic circuitry that governs differentiation to cell types of interest and to facilitate the production of therapeutically relevant cell types from iPS cells. Indeed, this same CRISPRi iPS line has recently been employed to tunably knockdown *EZH1* to enhance

the generation of adult-like T cells from iPS cells [129], as well as to understand the function of various transcriptional regulators during hematopoietic development [130]. Given the performance of this inducible CRISPRi system, we believe that it could be a powerful platform to perform large-scale pooled genetic screens during hematopoietic development. Current experimental approaches to understand hematopoietic development rely on mouse or zebrafish models where it is experimentally tractable to modulate only a handful of variables at a time. However, the versatility and scale of iPS cells coupled with the power of CRISPRi approaches theoretically enables the systematic interrogation of thousands of genes or elements within a single experiment. Furthermore, a significant translational goal in the hematopoiesis field is to generate engraftable hematopoietic stem cells (HSCs) from iPS cells. Large-scale knockdown screens with CRISPRi and coupled transcriptional readouts (e.g. Perturb-Seq [132,134,136]) could be informative to identify genes that induce an HSC-like transcriptional state from iPS cells and/or result in engraftment within NSG mice. Further refinements of this inducible CRISPRi iPS system could implement improved repressor domains, such as a ZIM3-derived KRAB domain that was recently identified [138]. Finally, we can build on the current power of CRISPRi for developmental screens by combining orthogonal dCas9-effectors to modulate combinations of genes simultaneously and for synergistic knockdown and activation of genes to direct cell fate toward desired lineages.

A.4 METHODS

A.4.1 Molecular cloning of CRISPR vectors

All CRISPR plasmids were designed in Benchling and cloned via Gibson Assembly. In brief, dox-inducible Cas9 or dCas9-KRAB vectors were designed by modifying the pB-rtTA vector (Addgene #126034). The UbC promoter in pB-rtTA was swapped for an EF1a promoter derived from CROPseq-Guide-Puro (Addgene #86708). A dCas9-KRAB-2A-mCherry transgene was derived from pHR-SFFV-KRAB-dCas9-P2A-mCherry (Addgene #60954) [103] and inserted upstream of the EF1a promoter with a TRE promoter. The resulting plasmid was termed PB03-NDi-dCas9-KRAB. A CROPseq vector (Addgene #86708) [140] for gRNA expression was modified by swapping the puromycin gene for a polycistronic mNeon-P2A-zeoR gene. The gRNA scaffold sequence was also swapped for an optimized sequence reported by [142] and the new plasmid was termed CROPseq-Zeo-GFP. NEB Stable Competent E. coli (NEB cat #C3040H) were transformed with the Gibson Assembly reaction and grown overnight at 30C on agar plates with 50 mg/mL carbenicillin. Individual colonies were picked for liquid culture in LB media supplemented with ampicillin and plasmid DNA was subsequently isolated using a QIAGEN Plasmid Plus Midi Kit with endotoxin removal. Plasmid sequences were fully verified by Sanger sequencing (Genewiz, Inc).

A.4.2 Maintenance of human pluripotent stem cells

Human iPS cells (1175.2 line, Boston Children's Hospital) were propagated on tissue culture-treated plates coated with Matrigel (Corning) and cultured in StemFlex medium (ThermoFisher). iPS cells were propagated upon 70% confluency using ReleSR (Stem Cell Technologies), unless otherwise stated.

A.4.3 Construction of inducible CRISPR iPS cells

iPS cells were engineered to express dox-inducible CRISPR vectors through piggyBac transposition. First, 1157.2 iPS cells were nucleofected (800,000 cells/reaction) with 2.5 µg of piggyBac transposase plasmid (SBI cat. no. PB200A-1) and 10 µg of transposon plasmid using Lonza Cell Line Nucleofector Kit V (cat. no. VCA-1003) with program B-016 on an Amaxa IIb Nucleofector. Nucleofected cells were cultured on matrigel-coated 6-well plates in StemFlex media supplemented with 1 µM Y-27632 (Stem Cell Technologies). Following five days of culture, iPS cells were exposed to 1 µg/mL of doxycycline hyclate (Sigma cat. no. D9891-1G) for 48 hours. mCherry+ iPS cells were single cell sorted using a Sony MA900 FACS into matrigel-coated 96-well plates containing conditioned media. Clones were prioritized for continued expansion and characterization based on stable growth rates, lack of mCherry expression following dox removal, and rapid, robust mCherry expression following reinduction with dox. Prioritized clones were submitted for karyotypic analyses (Cell Line Genetics), where 20 metaphases were analyzed using G-band karyotyping. Karyotypically normal clones were additionally characterized with knockdown/knockout assays, and verified to undergo hematopoietic differentiation and EHT. Validated iPS cell clones expressing either the dCas9-KRAB or Cas9 cassettes were scaled-up, banked, and used for all subsequent knockdown/knockout experiments.

A.4.4 Cloning and infection of gRNA expression vectors

gRNA sequences targeting the transcriptional start site of desired genes for CRISPRi-mediated knockdown were pulled from the Dolcetto Set A library (Broad Institute, Addgene #92385). gRNAs targeting non-coding elements, such as transcriptional enhancers and *cis*-regulatory elements, were designed in Benchling. gRNA target sequences and their reverse complement sequence were ordered as oligos (Integrated DNA Technologies), phosphorylated and annealed. The annealed oligos were cloned into the CROPseq-Zeo-GFP vector via Golden

Gate Assembly with Esp3I (Fisher Scientific) and T7 ligase (Epizyme). Lentivirus was produced for each gRNA expression vector by transfection of pMD2.G (Addgene #12259), psPAX2 (Addgene #12260), and the transfer plasmid (2:3:4 ratio by mass, 3 µg total) into HEK293FT cells using Lipofectamine 3000 (Thermo Fisher L3000015). Viral supernatants were harvested 48 hours after plasmid transfection and filtered through 0.45 µm PVDF filters (Millipore SLHVR04NL). Inducible dCas9-KRAB iPS cells were signularized with TrypLE (Gibco), plated on a matrigel-coated 6-well plate (160,000 cells/well), and then infected with individual CROPseq lentiviruses (MOI=0.25) with 8 µg/mL polybrene and 1 µM Y-27632 (Stem Cell Technologies). The following day, iPS cells were selected with 200 µg/mL zeocin (ThermoFisher R25001) for 24 hours. All iPS cells were mNeon+ following zeocin selection. iPS cells were then expanded as colonies, banked and then used for downstream differentiation experiments.

A.4.5 Flow cytometry

Cells were stained with antibodies and a viability dye (DAPI or propidium iodide) at room temperature for 20 minutes in the dark for flow cytometric profiling, followed by a wash step with FACS buffer (PBS + 2% FBS). Data was collected on a Sony MA900 and analyzed using FlowJo. Antibodies for flow cytometry were used at 1:100 dilution, unless otherwise stated.

A.4.6 CRISPR-interference assay

Inducible CRISPRi iPS cells expressing gRNAs were singularized with TrypLE and 50,000 cells were plated per well in matrigel-coated 24-well plates with 1 µM Y-27632. The following day, the StemFlex media was replaced to remove Y-27632 and dCas9-KRAB expression was induced with 1 ug/mL doxycycline supplemented to the media. Cells were harvested daily over the next five days through single cell dissociation. Total RNA was extracted from cells using an RNeasy Kit (Qiagen, 74106) per the manufacturer's instructions. cDNA was synthesized using 500 ng of purified RNA with a iScript cDNA Synthesis kit (Bio-Rad, 1708891). RT-qPCR was performed to

determine transcriptional knockdown of target genes using Kapa High Fidelity 2X ReadyMix, SYBR Green and Rox as a reference dye. RT-qPCR was also performed for *ACTB* as a housekeeping control gene. Relative expression of the gene of interest was normalized against *ACTB* using the difference in threshold-cycle (CT) values between the gene of interest and *ACTB* control by the $2^{-\Delta\Delta CT}$ method [144].

A.4.7 Differentiation of human iPS to hemogenic endothelium

Human iPS cells were cultured on irradiated mouse embryonic fibroblasts (Gibco) for one passage prior to the initiation of embryoid body differentiation. The CD34+ hemogenic endothelium was adapted from [116,119,146]. Briefly, iPS colonies were detached with Collagenase IV (Gibco) for 10 minutes at 37C and seeded into ultra-low attachment 10cm dishes (Corning) to form EBs in SFD media, supplemented with GlutaMAX (ThermoFisher), ascorbic acid (1 mM), α-monothioglycerol (400 μM), holo-transferrin (150 μg/mL), and BMP-4 (10 ng/uL). bFGF (5 ng/uL) was added to the EB cultures following 24 hours. The media was completely changed following an additional 18 hours to day 0 containing media, supplemented with bFGF (5 ng/mL), CHIR99021 (3 μM) and SB431542 (6 μM). Following 30 hours, the media was completely changed to StemPro-34 media (Thermo Fisher Scientific), supplemented with ascorbic acid (1 mM), holo-Transferrin (150 μg/mL), α-monothioglycerol (400 μM), bFGF (5 ng/uL), and VEGF (15 ng/mL). On day 6 of differentiation the media was changed to day 3 containing media supplemented additionally with IL-6 (10 ng/mL), IL-11 (5 ng/mL), IGF-1 (25 ng/mL), SCF (50 ng/mL) and EPO (2 U/mL) and then cultured until day 8. Cultures were maintained in a 5% CO2, 5% O2, 90% N2 environment.

A.4.8 Endothelial-to-hematopoietic transition assay

EBs were dissociated after 8 days of directed differentiation using the Embryoid Body Dissociation Kit (Miltenyi Biotec) according to the manufacturer's protocol. The resulting single

cell suspension was stained through a 40 uM filter and then CD34+ cells were enriched with MACS beads (Miltenyi Biotec) according to the manufacturer's protocol. CD34+ cells were plated onto Matrigel-coated 24-well plates (100,000 cells/well) and cultured at 5% O2 in StemPro-34 based media, supplemented with BMP4 (10 ng/mL), bFGF (5 ng/mL), IL-3 (30 ng/mL), IL-6 (10 ng/mL), IL-11 (5 ng/mL), IGF-1 (25 ng/mL), VEGF (5 ng/mL), SCF (100 ng/mL), TPO (30 ng/mL), FLT-3L (10 ng/mL), EPO (2 U/mL) and SHH (20 ng/mL). Unless otherwise stated, CD34+ hemogenic endothelium underwent EHT over the course of 7 days with regular half media changes every other day. Floating and/or adherent cells were collected for downstream functional or molecular assays.

A.5 SUPPLEMENTAL TABLES

Table S1: gRNA sequences

sgrna_id	Target Gene	Target Locus	Target Sequence
NTC	NA	NA	CCCGATGGACTATACCGAAC
grna_oct4_tss_a	OCT4	TSS	GGTGAAATGAGGGCTTGCGA
grna_oct4_tss_b	OCT4	TSS	**G**TTCACCAGGCCCCCGGCTTG
grna_oct4_tss_c	OCT4	TSS	GGGGCGCCAGTTGTGTCTCC
grna_oct4_pe_b	OCT4	Proximal enhancer	**G**CAGACATCTAATACCACGGT
grna_oct4_pe_c	OCT4	Proximal enhancer	**G**AGGGAGAACGGGGCCTACCG
grna_oct4_pe_f	OCT4	Proximal enhancer	**G**CCCTGGGTGGGGAAAACCAG
grna_oct4_de_b	OCT4	Distal enhancer	GTGCCGTGATGGTTCTGTCC
grna_oct4_de_c	OCT4	Distal enhancer	GGTCTGCCGGAAGGTCTACA
grna_oct4_de_e	OCT4	Distal enhancer	GCATGACAAAGGTGCCGTGA
grna_runx1_p2tss_a	RUNX1	P2 TSS	**G**CGGCGCAGGGCCGGGCAGCG
grna_runx1_p2tss_b	RUNX1	P2 TSS	**G**CGGGCGGGACGGGCGCCCCG
grna_runx1_p2tss_c	RUNX1	P2 TSS	GGCAGCGTGGTGCCCTGGCT
grna_runx1_p1tss_a	RUNX1	P1 TSS	**G**AGCCGAGTAGACTTTGCAAG
grna_runx1_p1tss_b	RUNX1	P1 TSS	**G**TTGAGATGGGCTGTGGAAAG
grna_runx1_p1tss_c	RUNX1	P1 TSS	**G**CTGCCAGGCTACCCAACTTG
grna_runx1_+24e_a	RUNX1	+24 Enhancer	GTAGAGCGGCCCCACCCTAG
grna_runx1_+24e_b	RUNX1	+24 Enhancer	**G**ATAAACCGGCAGTTAAAGCA

grna_runx1_+24e_c	RUNX1	+24 Enhancer	**G**TTGCAGAAGTTCCGTCGCTG
grna_runx1_+24e_d	RUNX1	+24 Enhancer	**G**AGCAACAGCCAGAAACGGCG

Table S2: RT-qPCR primers

Primer Name	Primer Sequence
ACTB Forward	GAGGCACTCTTCCAGCCTT
ACTB Reverse	AAGGTAGTTTCGTGGATGCC
huOCT4 Forward	CGAAAGAGAAAGCGAACCAGTATCGAGAAC
huOCT4 Reverse	GCGATCAAGCAGCGACTATGCACAACG

References

1. Antoniani, C., Romano, O., and Miccio, A. (2017). Concise Review: Epigenetic Regulation of Hematopoiesis: Biological Insights and Therapeutic Applications. STEM CELLS Transl. Med. *6*, 2106–2114. 10.1002/sctm.17-019
2. Shimomura, O., Johnson, F.H., and Saiga, Y. (1962). Extraction, Purification and Properties of Aequorin, a Bioluminescent Protein from the Luminous Hydromedusan, Aequorea. J. Cell. Comp. Physiol. *59*, 223–239. 10.1002/jcp.1030590302.
3. Kyba, M., and Daley, G.Q. (2003). Hematopoiesis from embryonic stem cells: lessons from and for ontogeny. Exp. Hematol. *31*, 994–1006. 10.1016/j.exphem.2003.08.002.
4. Tsien, R.Y. (1998). The green fluorescent protein. Annu. Rev. Biochem. *67*, 509–544. 10.1146/annurev.biochem.67.1.509.
5. Ditadi, A., Sturgeon, C.M., and Keller, G. (2017). A view of human haematopoietic development from the Petri dish. Nat. Rev. Mol. Cell Biol. *18*, 56–67. 10.1038/nrm.2016.127.
6. Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. Nat. Rev. Genet. *20*, 631–656. 10.1038/s41576-019-0150-2.
7. Irion, S., Nostro, M.C., Kattman, S.J., and Keller, G.M. (2008). Directed differentiation of pluripotent stem cells: from developmental biology to therapeutic applications. Cold Spring Harb. Symp. Quant. Biol. *73*, 101–110. 10.1101/sqb.2008.73.065.
8. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628. 10.1038/nmeth.1226.
9. Vo, L.T., and Daley, G.Q. (2015). De novo generation of HSCs from somatic and pluripotent stem cell sources. Blood *125*, 2641–2648. 10.1182/blood-2014-10-570234.
10. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M.

(2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349. 10.1126/science.1158441.

11. Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M., and Coleman, P. (1992). Analysis of gene expression in single live neurons. Proc. Natl. Acad. Sci. *89*, 3010–3014. 10.1073/pnas.89.7.3010.
12. Kennedy, M., Awong, G., Sturgeon, C.M., Ditadi, A., LaMotte-Mohs, R., Zúñiga-Pflücker, J.C., and Keller, G. (2012). T Lymphocyte Potential Marks the Emergence of Definitive Hematopoietic Progenitors in Human Pluripotent Stem Cell Differentiation Cultures. Cell Rep. *2*, 1722–1735. 10.1016/j.celrep.2012.11.003.
13. Sul, J.-Y., Wu, C.K., Zeng, F., Jochems, J., Lee, M.T., Kim, T.K., Peritz, T., Buckley, P., Cappelleri, D.J., Maronski, M., et al. (2009). Transcriptome transfer produces a predictable cellular phenotype. Proc. Natl. Acad. Sci. *106*, 7624–7629. 10.1073/pnas.0902161106.
14. Orkin, S.H., and Zon, L.I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. Cell *132*, 631–644. 10.1016/j.cell.2008.01.025.
15. Guillaume-Gentil, O., Grindberg, R.V., Kooger, R., Dorwling-Carter, L., Martinez, V., Ossola, D., Pilhofer, M., Zambelli, T., and Vorholt, J.A. (2016). Tunable Single-Cell Extraction for Molecular Analyses. Cell *166*, 506–516. 10.1016/j.cell.2016.06.025.
16. Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. *18*, 71–86. 10.1038/nrg.2016.139.
17. Cao, Y., Hjort, M., Chen, H., Birey, F., Leal-Ortiz, S.A., Han, C.M., Santiago, J.G., Paşca, S.P., Wu, J.C., and Melosh, N.A. (2017). Nondestructive nanostraw intracellular sampling for longitudinal cell monitoring. Proc. Natl. Acad. Sci. *114*, E1866–E1874. 10.1073/pnas.1615375114.
18. Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. Nat. Rev. Genet. *9*, 397–405. 10.1038/nrg2337.
19. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J.,

Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386. 10.1038/nbt.2859.

20. Rayan, N.A., del Rosario, R.C.H., and Prabhakar, S. (2016). Massive contribution of transposable elements to mammalian regulatory sequences. Semin. Cell Dev. Biol. *57*, 51–56. 10.1016/j.semcdb.2016.05.004.
21. Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods *advance online publication*. 10.1038/nmeth.3971.
22. Thompson, P.J., Macfarlan, T.S., and Lorincz, M.C. (2016). Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire. Mol. Cell *62*, 766–776. 10.1016/j.molcel.2016.03.029.
23. Liu, Z., Lou, H., Xie, K., Wang, H., Chen, N., Aparicio, O.M., Zhang, M.Q., Jiang, R., and Chen, T. (2017). Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. Nat. Commun. *8*, 22. 10.1038/s41467-017-00039-z.
24. Johnson, W.E. (2015). Endogenous Retroviruses in the Genomics Era. Annu. Rev. Virol. *2*, 135–159. 10.1146/annurev-virology-100114-054945.
25. Manno, G.L., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498. 10.1038/s41586-018-0414-6.
26. Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. *13*, 283–296. 10.1038/nrg3199.
27. Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. Cell *176*, 928-943.e22. 10.1016/j.cell.2019.01.006.

28. Greenwood, A.D., Ishida, Y., O'Brien, S.P., Roca, A.L., and Eiden, M.V. (2017). Transmission, Evolution, and Endogenization: Lessons Learned from Recent Retroviral Invasions. Microbiol. Mol. Biol. Rev. *82*, e00044-17. 10.1128/MMBR.00044-17.

29. Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554. 10.1038/s41587-019-0071-9.

30. Kassiotis, G., and Stoye, J.P. (2016). Immune responses to endogenous retroelements: taking the bad with the good. Nat. Rev. Immunol. *16*, 207–219. 10.1038/nri.2016.27.

31. Weinreb, C., Wolock, S., Tusi, B.K., Socolovsky, M., and Klein, A.M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. Proc. Natl. Acad. Sci. *115*, E2467–E2476. 10.1073/pnas.1714723115.

32. Stoye, J.P. (2012). Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat. Rev. Microbiol. *10*, 395–406. 10.1038/nrmicro2783.

33. Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S.M., Singer, R.H., and Long, R.M. (1998). Localization of ASH1 mRNA particles in living yeast. Mol. Cell *2*, 437–445. 10.1016/s1097-2765(00)80143-4.

34. Macfarlan, T.S., Gifford, W.D., Driscoll, S., Lettieri, K., Rowe, H.M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S.L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature *487*, 57–63. 10.1038/nature11244.

35. Wang, Z., Zhang, K., Wooley, K.L., and Taylor, J.-S. (2012). Imaging mRNA Expression in Live Cells via PNA·DNA Strand Displacement-Activated Probes. J. Nucleic Acids *2012*, 962652. 10.1155/2012/962652.

36. Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083–1087. 10.1126/science.aad5497.

37. Tutucci, E., Vera, M., Biswas, J., Garcia, J., Parker, R., and Singer, R.H. (2018). An improved MS2 system for accurate reporting of the mRNA life cycle. Nat. Methods *15*, 81–

89. 10.1038/nmeth.4502.

38. Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet. *42*, 631–634. 10.1038/ng.600.

39. Lovatt, D., Ruble, B.K., Lee, J., Dueck, H., Kim, T.K., Fisher, S., Francis, C., Spaethling, J.M., Wolf, J.A., Grady, M.S., et al. (2014). Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. Nat. Methods *11*, 190–196. 10.1038/nmeth.2804.

40. Chuong, E.B., Rumi, M.A.K., Soares, M.J., and Baker, J.C. (2013). Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat. Genet. *45*, 325–329. 10.1038/ng.2553.

41. Atmanli, A., Hu, D., Deiman, F.E., van de Vrugt, A.M., Cherbonneau, F., Black, L.D., III, and Domian, I.J. (2019). Multiplex live single-cell transcriptional analysis demarcates cellular functional heterogeneity. eLife *8*, e49599. 10.7554/eLife.49599.

42. Bourque, G., Leong, B., Vega, V.B., Chen, X., Lee, Y.L., Srinivasan, K.G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H.H., et al. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. *18*, 1752–1762. 10.1101/gr.080663.108.

43. Yang, L.-Z., Wang, Y., Li, S.-Q., Yao, R.-W., Luan, P.-F., Wu, H., Carmichael, G.G., and Chen, L.-L. (2019). Dynamic Imaging of RNA in Living Cells by CRISPR-Cas13 Systems. Mol. Cell *76*, 981-997.e7. 10.1016/j.molcel.2019.10.024.

44. Martens, J.H.A., O'Sullivan, R.J., Braunschweig, U., Opravil, S., Radolf, M., Steinlein, P., and Jenuwein, T. (2005). The profile of repeat-associated histone lysine methylation states in the mouse epigenome. EMBO J. *24*, 800–812. 10.1038/sj.emboj.7600545.

45. Nelles, D.A., Fang, M.Y., O'Connell, M.R., Xu, J.L., Markmiller, S.J., Doudna, J.A., and Yeo, G.W. (2016). Programmable RNA Tracking in Live Cells with CRISPR/Cas9. Cell

165, 488–496. 10.1016/j.cell.2016.02.054.

46. Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M.P., and Wang, T. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. *24*, 1963–1976. 10.1101/gr.168872.113.

47. Shipman, S.L., Nivala, J., Macklis, J.D., and Church, G.M. (2016). Molecular recordings by directed CRISPR spacer acquisition. Science *353*, aaf1175. 10.1126/science.aaf1175.

48. Trizzino, M., Kapusta, A., and Brown, C.D. (2018). Transposable elements generate regulatory novelty in a tissue-specific fashion. BMC Genomics *19*, 468. 10.1186/s12864-018-4850-3.

49. Tang, W., and Liu, D.R. (2018). Rewritable multi-event analog recording in bacterial and mammalian cells. Science, eaap8992. 10.1126/science.aap8992.

50. Cao, Y., Chen, G., Wu, G., Zhang, X., McDermott, J., Chen, X., Xu, C., Jiang, Q., Chen, Z., Zeng, Y., et al. (2019). Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. Genome Res. *29*, 40–52. 10.1101/gr.235747.118.

51. Schmidt, F., Cherepkova, M.Y., and Platt, R.J. (2018). Transcriptional recording by CRISPR spacer acquisition from RNA. Nature, 1. 10.1038/s41586-018-0569-1.

52. Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. *22*, 1798–1812. 10.1101/gr.139105.112.

53. Rodriques, S.G., Chen, L.M., Liu, S., Zhong, E.D., Scherrer, J.R., Boyden, E.S., and Chen, F. (2020). RNA timestamps identify the age of single molecules in RNA sequencing. Nat. Biotechnol., 1–6. 10.1038/s41587-020-0704-z.

54. Trizzino, M., Park, Y., Holsbach-Beltrame, M., Aracena, K., Mika, K., Caliskan, M., Perry, G.H., Lynch, V.J., and Brown, C.D. (2017). Transposable elements are the primary source

of novelty in primate gene regulation. Genome Res. *27*, 1623–1633. 10.1101/gr.218149.116.

55. Bang, C., and Thum, T. (2012). Exosomes: new players in cell-cell communication. Int. J. Biochem. Cell Biol. *44*, 2060–2064. 10.1016/j.biocel.2012.08.007.

56. Jacques, P.-É., Jeyakani, J., and Bourque, G. (2013). The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements. PLOS Genet. *9*, e1003504. 10.1371/journal.pgen.1003504.

57. Rulli, S.J., Hibbert, C.S., Mirro, J., Pederson, T., Biswal, S., and Rein, A. (2007). Selective and Nonselective Packaging of Cellular RNAs in Retrovirus Particles. J. Virol. *81*, 6623–6631. 10.1128/JVI.02833-06.

58. Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., et al. (2018). Ten things you should know about transposable elements. Genome Biol. *19*, 199. 10.1186/s13059-018-1577-z.

59. Donaldson, B., Lateef, Z., Walker, G.F., Young, S.L., and Ward, V.K. (2018). Virus-like particle vaccines: immunology and formulation for clinical translation. Expert Rev. Vaccines *17*, 833–849. 10.1080/14760584.2018.1516552.

60. Ryan, F.P. (2004). Human endogenous retroviruses in health and disease: a symbiotic perspective. J. R. Soc. Med. *97*, 560–565. 10.1177/014107680409701202.

61. Kaczmarczyk, S.J., Sitaraman, K., Young, H.A., Hughes, S.H., and Chatterjee, D.K. (2011). Protein delivery using engineered virus-like particles. Proc. Natl. Acad. Sci. *108*, 16998–17003. 10.1073/pnas.1101874108.

62. Dopkins, N., O'Mara, M.M., Lawrence, E., Fei, T., Sandoval-Motta, S., Nixon, D.F., and Bendall, M.L. (2022). A field guide to endogenous retrovirus regulatory networks. Mol. Cell *82*, 3763–3768. 10.1016/j.molcel.2022.09.011.

63. Mangeot, P.E., Risson, V., Fusil, F., Marnef, A., Laurent, E., Blin, J., Mournetas, V., Massouridès, E., Sohier, T.J.M., Corbin, A., et al. (2019). Genome editing in primary cells

and in vivo using viral-derived Nanoblades loaded with Cas9-sgRNA ribonucleoproteins. Nat. Commun. *10*, 45. 10.1038/s41467-018-07845-z.

64. Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560. 10.1038/nature06008.

65. Banskota, S., Raguram, A., Suh, S., Du, S.W., Davis, J.R., Choi, E.H., Wang, X., Nielsen, S.C., Newby, G.A., Randolph, P.B., et al. (2022). Engineered virus-like particles for efficient in vivo delivery of therapeutic proteins. Cell *0*. 10.1016/j.cell.2021.12.021.

66. Bulut-Karslioglu, A., De La Rosa-Velázquez, I.A., Ramirez, F., Barenboim, M., Onishi-Seebacher, M., Arand, J., Galán, C., Winter, G.E., Engist, B., Gerle, B., et al. (2014). Suv39h-dependent H3K9me3 marks intact retrotransposons and silences LINE elements in mouse embryonic stem cells. Mol. Cell *55*, 277–290. 10.1016/j.molcel.2014.05.029.

67. Eyckerman, S., Titeca, K., Quickelberghe, E.V., Cloots, E., Verhee, A., Samyn, N., Ceuninck, L.D., Timmerman, E., Sutter, D.D., Lievens, S., et al. (2016). Trapping mammalian protein complexes in viral particles. Nat. Commun. *7*, 11416. 10.1038/ncomms11416.

68. Matsui, T., Leung, D., Miyashita, H., Maksakova, I.A., Miyachi, H., Kimura, H., Tachibana, M., Lorincz, M.C., and Shinkai, Y. (2010). Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature *464*, 927–931. 10.1038/nature08858.

69. Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181. 10.1038/nprot.2014.006.

70. Leung, D.C., Dong, K.B., Maksakova, I.A., Goyal, P., Appanah, R., Lee, S., Tachibana, M., Shinkai, Y., Lehnertz, B., Mager, D.L., et al. (2011). Lysine methyltransferase G9a is required for de novo DNA methylation and the establishment, but not the maintenance, of

proviral silencing. Proc. Natl. Acad. Sci. U. S. A. *108*, 5718–5723. 10.1073/pnas.1014660108.

71. Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat. Rev. Genet. *14*, 618–630. 10.1038/nrg3542.

72. Garcia-Perez, J.L., Morell, M., Scheys, J.O., Kulpa, D.A., Morell, S., Carter, C.C., Hammer, G.D., Collins, K.L., O'Shea, K.S., Menendez, P., et al. (2010). Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. Nature *466*, 769–773. 10.1038/nature09209.

73. Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: Network Biology Applied to Stem Cell Engineering. Cell *158*, 903–915. 10.1016/j.cell.2014.07.020.

74. Maksakova, I.A., Thompson, P.J., Goyal, P., Jones, S.J., Singh, P.B., Karimi, M.M., and Lorincz, M.C. (2013). Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. Epigenetics Chromatin *6*, 15. 10.1186/1756-8935-6-15.

75. Rein, A., Datta, S.A.K., Jones, C.P., and Musier-Forsyth, K. (2011). Diverse interactions of retroviral Gag proteins with RNAs. Trends Biochem. Sci. *36*, 373–380. 10.1016/j.tibs.2011.04.001.

76. Yang, B.X., El Farran, C.A., Guo, H.C., Yu, T., Fang, H.T., Wang, H.F., Schlesinger, S., Seah, Y.F.S., Goh, G.Y.L., Neo, S.P., et al. (2015). Systematic identification of factors for provirus silencing in embryonic stem cells. Cell *163*, 230–245. 10.1016/j.cell.2015.08.037.

77. Safaee, N., Kozlov, G., Noronha, A.M., Xie, J., Wilds, C.J., and Gehring, K. (2012). Interdomain Allostery Promotes Assembly of the Poly(A) mRNA Complex with PABP and eIF4G. Mol. Cell *48*, 375–386. 10.1016/j.molcel.2012.09.001.

78. Rowe, H.M., Friedli, M., Offner, S., Verp, S., Mesnard, D., Marquis, J., Aktas, T., and

Trono, D. (2013). De novo DNA methylation of endogenous retroviruses is shaped by KRAB-ZFPs/KAP1 and ESET. Dev. Camb. Engl. *140*, 519–529. 10.1242/dev.087585.

79. Martin, J.C., Chang, C., Boschetti, G., Ungaro, R., Giri, M., Grout, J.A., Gettler, K., Chuang, L.-S., Nayar, S., Greenstein, A.J., et al. (2019). Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. Cell *178*, 1493-1508.e20. 10.1016/j.cell.2019.08.008.
80. Rowe, H.M., Kapopoulou, A., Corsinotti, A., Fasching, L., Macfarlan, T.S., Tarabay, Y., Viville, S., Jakobsson, J., Pfaff, S.L., and Trono, D. (2013). TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. Genome Res. *23*, 452–461. 10.1101/gr.147678.112.
81. Sharif, O., Bolshakov, V.N., Raines, S., Newham, P., and Perkins, N.D. (2007). Transcriptional profiling of the LPS induced NF-κB response in macrophages. BMC Immunol. *8*, 1. 10.1186/1471-2172-8-1.
82. Canzio, D., Liao, M., Naber, N., Pate, E., Larson, A., Wu, S., Marina, D.B., Garcia, J.F., Madhani, H.D., Cooke, R., et al. (2013). A conformational switch in HP1 releases auto-inhibition to drive heterochromatin assembly. Nature *496*, 377–381. 10.1038/nature12032.
83. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550. 10.1073/pnas.0506580102.
84. Wolf, D., Cammas, F., Losson, R., and Goff, S.P. (2008). Primer binding site-dependent restriction of murine leukemia virus requires HP1 binding by TRIM28. J. Virol. *82*, 4675–4679. 10.1128/JVI.02445-07.
85. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527. 10.1038/nbt.3519.
86. Rowe, H.M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., Aktas, T., Maillard,

P.V., Layard-Liesching, H., Verp, S., Marquis, J., et al. (2010). KAP1 controls endogenous retroviruses in embryonic stem cells. Nature *463*, 237–240. 10.1038/nature08674.

87. Zhang, T., Tan, P., Wang, L., Jin, N., Li, Y., Zhang, L., Yang, H., Hu, Z., Zhang, L., Hu, C., et al. (2017). RNALocate: a resource for RNA subcellular localizations. Nucleic Acids Res. *45*, D135–D138. 10.1093/nar/gkw728.

88. Turelli, P., Castro-Diaz, N., Marzetta, F., Kapopoulou, A., Raclot, C., Duc, J., Tieng, V., Quenneville, S., and Trono, D. (2014). Interplay of TRIM28 and DNA methylation in controlling human endogenous retroelements. Genome Res. *24*, 1260–1270. 10.1101/gr.172833.114.

89. Chen, M.J., Lummertz da Rocha, E., Cahan, P., Kubaczka, C., Hunter, P., Sousa, P., Mullin, N.K., Fujiwara, Y., Nguyen, M., Tan, Y., et al. (2020). Transcriptome Dynamics of Hematopoietic Stem Cell Formation Revealed Using a Combinatorial Runx1 and Ly6a Reporter System. Stem Cell Rep. *14*, 956–971. 10.1016/j.stemcr.2020.03.020.

90. Ecco, G., Imbeault, M., and Trono, D. (2017). KRAB zinc finger proteins. Dev. Camb. Engl. *144*, 2719–2729. 10.1242/dev.132605.

91. Clapes, T., Lefkopoulos, S., and Trompouki, E. (2016). Stress and Non-Stress Roles of Inflammatory Signals during HSC Emergence and Maintenance. Front. Immunol. *7*. 10.3389/fimmu.2016.00487.

92. Najafabadi, H.S., Mnaimneh, S., Schmitges, F.W., Garton, M., Lam, K.N., Yang, A., Albu, M., Weirauch, M.T., Radovani, E., Kim, P.M., et al. (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat. Biotechnol. *33*, 555–562. 10.1038/nbt.3128.

93. Espín-Palazón, R., Stachura, D.L., Campbell, C.A., García-Moreno, D., Del Cid, N., Kim, A.D., Candel, S., Meseguer, J., Mulero, V., and Traver, D. (2014). Proinflammatory signaling regulates hematopoietic stem cell emergence. Cell *159*, 1070–1085. 10.1016/j.cell.2014.10.031.

94. Fukuda, K., and Shinkai, Y. (2020). SETDB1-Mediated Silencing of Retroelements. Viruses *12*, 596. 10.3390/v12060596.

95. Sawamiphak, S., Kontarakis, Z., and Stainier, D.Y.R. (2014). Interferon gamma signaling positively regulates hematopoietic stem cell emergence. Dev. Cell *31*, 640–653. 10.1016/j.devcel.2014.11.007.

96. Thomas, J.H., and Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger genes. Genome Res. *21*, 1800–1812. 10.1101/gr.121749.111.

97. He, Q., Zhang, C., Wang, L., Zhang, P., Ma, D., Lv, J., and Liu, F. (2015). Inflammatory signaling regulates hematopoietic stem and progenitor cell emergence in vertebrates. Blood *125*, 1098–1106. 10.1182/blood-2014-09-601542.

98. Lukic, S., Nicolas, J.-C., and Levine, A.J. (2014). The diversity of zinc-finger genes on human chromosome 19 provides an evolutionary mechanism for defense against inherited endogenous retroviruses. Cell Death Differ. *21*, 381–387. 10.1038/cdd.2013.150.

99. Li, Y., Esain, V., Teng, L., Xu, J., Kwan, W., Frost, I.M., Yzaguirre, A.D., Cai, X., Cortes, M., Maijenburg, M.W., et al. (2014). Inflammatory signaling regulates embryonic hematopoietic stem and progenitor cell production. Genes Dev. *28*, 2597–2612. 10.1101/gad.253302.114.

100. Corsinotti, A., Kapopoulou, A., Gubelmann, C., Imbeault, M., Santoni de Sio, F.R., Rowe, H.M., Mouscaz, Y., Deplancke, B., and Trono, D. (2013). Global and stage specific patterns of Krüppel-associated-box zinc finger protein gene expression in murine early embryonic cells. PloS One *8*, e56721. 10.1371/journal.pone.0056721.

101. Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. Cell *154*, 442–451. 10.1016/j.cell.2013.06.044.

102. Emerson, R.O., and Thomas, J.H. (2009). Adaptive Evolution in Zinc Finger Transcription

Factors. PLOS Genet. *5*, e1000325. 10.1371/journal.pgen.1000325.

103. Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell *159*, 647–661. 10.1016/j.cell.2014.09.029.

104. Imbeault, M., Helleboid, P.-Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature *543*, 550–554. 10.1038/nature21683.

105. Sanson, K.R., Hanna, R.E., Hegde, M., Donovan, K.F., Strand, C., Sullender, M.E., Vaimberg, E.W., Goodale, A., Root, D.E., Piccioni, F., et al. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nat. Commun. *9*, 1–15. 10.1038/s41467-018-07901-8.

106. Grow, E.J., Flynn, R.A., Chavez, S.L., Bayless, N.L., Wossidlo, M., Wesche, D.J., Martin, L., Ware, C.B., Blish, C.A., Chang, H.Y., et al. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature *522*, 221–225. 10.1038/nature14308.

107. Fulco, C.P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S.R., Perez, E.M., Kane, M., Cleary, B., Lander, E.S., and Engreitz, J.M. (2016). Systematic mapping of functional enhancer–promoter connections with CRISPR interference. Science *354*, 769–773. 10.1126/science.aag2445.

108. Lefkopoulos, S., Polyzou, A., Derecka, M., Bergo, V., Clapes, T., Cauchy, P., Jerez-Longres, C., Onishi-Seebacher, M., Yin, N., Martagon-Calderón, N.-A., et al. (2020). Repetitive Elements Trigger RIG-I-like Receptor Signaling that Regulates the Emergence of Hematopoietic Stem and Progenitor Cells. Immunity *53*, 934-951.e9. 10.1016/j.immuni.2020.10.007.

109. Klann, T.S., Black, J.B., Chellappan, M., Safi, A., Song, L., Hilton, I.B., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2017). CRISPR–Cas9 epigenome editing enables high-

throughput screening for functional regulatory elements in the human genome. Nat. Biotechnol. *35*, 561–568. 10.1038/nbt.3853.

110. Tugnet, N., Rylance, P., Roden, D., Trela, M., and Nelson, P. (2013). Human Endogenous Retroviruses (HERVs) and Autoimmune Rheumatic Disease: Is There a Link? Open Rheumatol. J. *7*, 13–21. 10.2174/1874312901307010013.
111. Mandegar, M.A., Huebsch, N., Frolov, E.B., Shin, E., Truong, A., Olvera, M.P., Chan, A.H., Miyaoka, Y., Holmes, K., Spencer, C.I., et al. (2016). CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. Cell Stem Cell *18*, 541–553. 10.1016/j.stem.2016.01.022.
112. Chiappinelli, K.B., Strissel, P.L., Desrichard, A., Li, H., Henke, C., Akman, B., Hein, A., Rote, N.S., Cope, L.M., Snyder, A., et al. (2015). Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. Cell *162*, 974–986. 10.1016/j.cell.2015.07.011.
113. Jost, M., Santos, D.A., Saunders, R.A., Horlbeck, M.A., Hawkins, J.S., Scaria, S.M., Norman, T.M., Hussmann, J.A., Liem, C.R., Gross, C.A., et al. (2020). Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs. Nat. Biotechnol. *38*, 355–364. 10.1038/s41587-019-0387-5.
114. Roulois, D., Loo Yau, H., Singhania, R., Wang, Y., Danesh, A., Shen, S.Y., Han, H., Liang, G., Jones, P.A., Pugh, T.J., et al. (2015). DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. Cell *162*, 961–973. 10.1016/j.cell.2015.07.056.
115. Smits, A.H., Ziebell, F., Joberty, G., Zinn, N., Mueller, W.F., Clauder-Münster, S., Eberhard, D., Fälth Savitski, M., Grandi, P., Jakob, P., et al. (2019). Biological plasticity rescues target activity in CRISPR knock outs. Nat. Methods *16*, 1087–1093. 10.1038/s41592-019-0614-5.
116. Sturgeon, C.M., Ditadi, A., Awong, G., Kennedy, M., and Keller, G. (2014). Wnt signaling

controls the specification of definitive and primitive hematopoiesis from human pluripotent stem cells. Nat. Biotechnol. *32*, 554–561. 10.1038/nbt.2915.

117. Goel, S., DeCristo, M.J., Watt, A.C., BrinJones, H., Sceneay, J., Li, B.B., Khan, N., Ubellacker, J.M., Xie, S., Metzger-Filho, O., et al. (2017). CDK4/6 inhibition triggers anti-tumour immunity. Nature *548*, 471–475. 10.1038/nature23465.

118. Gazquez-Gutierrez, A., Witteveldt, J., R Heras, S., and Macias, S. (2021). Sensing of transposable elements by the antiviral innate immune system. RNA N. Y. N, rna.078721.121. 10.1261/rna.078721.121.

119. Ditadi, A., Sturgeon, C.M., Tober, J., Awong, G., Kennedy, M., Yzaguirre, A.D., Azzola, L., Ng, E.S., Stanley, E.G., French, D.L., et al. (2015). Human definitive haemogenic endothelium and arterial vascular endothelium represent distinct lineages. Nat. Cell Biol. *17*, 580–591. 10.1038/ncb3161.

120. Lester, S.N., and Li, K. (2014). Toll-like receptors in antiviral innate immunity. J. Mol. Biol. *426*, 1246–1264. 10.1016/j.jmb.2013.11.024.

121. North, T.E., de Bruijn, M.F.T.R., Stacy, T., Talebian, L., Lind, E., Robin, C., Binder, M., Dzierzak, E., and Speck, N.A. (2002). Runx1 expression marks long-term repopulating hematopoietic stem cells in the midgestation mouse embryo. Immunity *16*, 661–672. 10.1016/s1074-7613(02)00296-0.

122. Gürtler, C., and Bowie, A.G. (2013). Innate immune detection of microbial nucleic acids. Trends Microbiol. *21*, 413–420. 10.1016/j.tim.2013.04.004.

123. Lilly, A.J., Costa, G., Largeot, A., Fadlullah, M.Z.H., Lie-A-Ling, M., Lacaud, G., and Kouskoff, V. (2016). Interplay between SOX7 and RUNX1 regulates hemogenic endothelial fate in the yolk sac. Development *143*, 4341–4351. 10.1242/dev.140970.

124. Rigby, R.E., Webb, L.M., Mackenzie, K.J., Li, Y., Leitch, A., Reijns, M.A.M., Lundie, R.J., Revuelta, A., Davidson, D.J., Diebold, S., et al. (2014). RNA:DNA hybrids are a novel molecular pattern sensed by TLR9. EMBO J. *33*, 542–558. 10.1002/embj.201386117.

125. Lie-A-Ling, M., Marinopoulou, E., Lilly, A.J., Challinor, M., Patel, R., Lancrin, C., Kouskoff, V., and Lacaud, G. (2018). Regulation of RUNX1 dosage is crucial for efficient blood formation from hemogenic endothelium. Development *145*. 10.1242/dev.149419.

126. O'Neill, L.A.J., Golenbock, D., and Bowie, A.G. (2013). The history of Toll-like receptors — redefining innate immunity. Nat. Rev. Immunol. *13*, 453–460. 10.1038/nri3446.

127. Ferrell, P.I., Xi, J., Ma, C., Adlakha, M., and Kaufman, D.S. (2015). The RUNX1 +24 enhancer and P1 promoter identify a unique subpopulation of hematopoietic progenitor cells derived from human pluripotent stem cells. Stem Cells Dayt. Ohio *33*, 1130–1141. 10.1002/stem.1940.

128. Clapes, T., and Trompouki, E. (2020). Hematopoietic regeneration under the spell of epigenetic-epitranscriptomic factors and transposable elements. Curr. Opin. Hematol. *27*, 264–272. 10.1097/MOH.0000000000000585.

129. Jing, R., Scarfo, I., Najia, M.A., Lummertz da Rocha, E., Han, A., Sanborn, M., Bingham, T., Kubaczka, C., Jha, D.K., Falchetti, M., et al. (2022). EZH1 repression generates mature iPSC-derived CAR T cells with enhanced antitumor activity. Cell Stem Cell *29*, 1181-1196.e6. 10.1016/j.stem.2022.06.014.

130. Lummertz da Rocha, E., Kubaczka, C., Sugden, W.W., Najia, M.A., Jing, R., Markel, A., LeBlanc, Z.C., dos Santos Peixoto, R., Falchetti, M., Collins, J.J., et al. (2022). CellComm infers cellular crosstalk that drives haematopoietic stem and progenitor cell development. Nat. Cell Biol. *24*, 579–589. 10.1038/s41556-022-00884-1.

131. Clapes, T., Polyzou, A., Prater, P., Sagar, Morales-Hernández, A., Ferrarini, M.G., Kehrer, N., Lefkopoulos, S., Bergo, V., Hummel, B., et al. (2021). Chemotherapy-induced transposable elements activate MDA5 to enhance haematopoietic regeneration. Nat. Cell Biol., 1–14. 10.1038/s41556-021-00707-9.

132. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-Seq: Dissecting Molecular

Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell *167*, 1853-1866.e17. 10.1016/j.cell.2016.11.038.

133. Kato, H., Takeuchi, O., Mikamo-Satoh, E., Hirai, R., Kawai, T., Matsushita, K., Hiiragi, A., Dermody, T.S., Fujita, T., and Akira, S. (2008). Length-dependent recognition of double-stranded ribonucleic acids by retinoic acid-inducible gene-I and melanoma differentiation-associated gene 5. J. Exp. Med. *205*, 1601–1610. 10.1084/jem.20080091.

134. Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell *167*, 1867-1882.e21. 10.1016/j.cell.2016.11.048.

135. Hiscott, J. (2007). Triggering the innate antiviral response through IRF-3 activation. J. Biol. Chem. *282*, 15325–15329. 10.1074/jbc.R700002200.

136. Replogle, J.M., Saunders, R.A., Pogson, A.N., Hussmann, J.A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E.J., Adelman, K., Lithwick-Yanai, G., et al. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell *0*. 10.1016/j.cell.2022.05.013.

137. Schoggins, J.W., MacDuff, D.A., Imanaka, N., Gainey, M.D., Shrestha, B., Eitson, J.L., Mar, K.B., Richardson, R.B., Ratushny, A.V., Litvak, V., et al. (2015). Corrigendum: Pan-viral specificity of IFN-induced genes reveals new roles for cGAS in innate immunity. Nature *525*, 144. 10.1038/nature14555.

138. Alerasool, N., Segal, D., Lee, H., and Taipale, M. (2020). An efficient KRAB domain for CRISPRi applications in human cells. Nat. Methods, 1–4. 10.1038/s41592-020-0966-x.

139. Sun, L., Wu, J., Du, F., Chen, X., and Chen, Z.J. (2013). Cyclic GMP-AMP Synthase Is a Cytosolic DNA Sensor That Activates the Type I Interferon Pathway. Science *339*, 786–791. 10.1126/science.1232458.

140. Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J.,

Schuster, L.C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. Nat. Methods *14*, 297–301. 10.1038/nmeth.4177.

141. Gao, D., Wu, J., Wu, Y.-T., Du, F., Aroh, C., Yan, N., Sun, L., and Chen, Z.J. (2013). Cyclic GMP-AMP synthase is an innate immune sensor of HIV and other retroviruses. Science *341*, 903–906. 10.1126/science.1240933.
142. Dang, Y., Jia, G., Choi, J., Ma, H., Anaya, E., Ye, C., Shankar, P., and Wu, H. (2015). Optimizing sgRNA structure to improve CRISPR-Cas9 knockout efficiency. Genome Biol. *16*, 280. 10.1186/s13059-015-0846-3.
143. Paludan, S.R., and Bowie, A.G. (2013). Immune sensing of DNA. Immunity *38*, 870–880. 10.1016/j.immuni.2013.05.004.
144. Schmittgen, T.D., and Livak, K.J. (2008). Analyzing real-time PCR data by the comparative C T method. Nat. Protoc. *3*, 1101–1108. 10.1038/nprot.2008.73.
145. Xu, J., Zoltick, P.W., Gamero, A.M., and Gallucci, S. (2014). TLR ligands up-regulate Trex1 expression in murine conventional dendritic cells through type I Interferon and NF-κB-dependent signaling pathways. J. Leukoc. Biol. *96*, 93–103. 10.1189/jlb.2A0713-393RR.
146. Sugimura, R., Jha, D.K., Han, A., Soria-Valles, C., da Rocha, E.L., Lu, Y.-F., Goettel, J.A., Serrao, E., Rowe, R.G., Malleshaiah, M., et al. (2017). Haematopoietic stem and progenitor cells from human pluripotent stem cells. Nature *advance online publication*. 10.1038/nature22370.
147. Ablasser, A., Hemmerling, I., Schmid-Burgk, J.L., Behrendt, R., Roers, A., and Hornung, V. (2014). TREX1 deficiency triggers cell-autonomous immunity in a cGAS-dependent manner. J. Immunol. Baltim. Md 1950 *192*, 5993–5997. 10.4049/jimmunol.1400737.
148. Zuniga, E.I., Hahm, B., and Oldstone, M.B.A. (2007). Type I interferon during viral infections: multiple triggers for a multifunctional mediator. Curr. Top. Microbiol. Immunol. *316*, 337–357. 10.1007/978-3-540-71329-6_16.

149. Manghera, M., and Douville, R.N. (2013). Endogenous retrovirus-K promoter: a landing strip for inflammatory transcription factors? Retrovirology *10*, 16. 10.1186/1742-4690-10-16.

150. Johnston, J.B., Silva, C., Holden, J., Warren, K.G., Clark, A.W., and Power, C. (2001). Monocyte activation and differentiation augment human endogenous retrovirus expression: implications for inflammatory brain diseases. Ann. Neurol. *50*, 434–442. 10.1002/ana.1131.

151. Guio, L., and González, J. (2019). New Insights on the Evolution of Genome Content: Population Dynamics of Transposable Elements in Flies and Humans. Methods Mol. Biol. Clifton NJ *1910*, 505–530. 10.1007/978-1-4939-9074-0_16.

152. Patel, M.R., Emerman, M., and Malik, H.S. (2011). Paleovirology - ghosts and gifts of viruses past. Curr. Opin. Virol. *1*, 304–309. 10.1016/j.coviro.2011.06.007.

153. Smit, A.F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. *9*, 657–663. 10.1016/s0959-437x(99)00031-3.

154. Goodier, J.L., and Kazazian, H.H. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell *135*, 23–35. 10.1016/j.cell.2008.09.022.

155. Myrum, C., Baumann, A., Bustad, H.J., Flydal, M.I., Mariaule, V., Alvira, S., Cuéllar, J., Haavik, J., Soulé, J., Valpuesta, J.M., et al. (2015). Arc is a flexible modular protein capable of reversible self-oligomerization. Biochem. J. *468*, 145–158. 10.1042/BJ20141446.

156. Ashley, J., Cordy, B., Lucia, D., Fradkin, L.G., Budnik, V., and Thomson, T. (2018). Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. Cell *172*, 262-274.e11. 10.1016/j.cell.2017.12.022.

157. Pastuzyn, E.D., Day, C.E., Kearns, R.B., Kyrke-Smith, M., Taibi, A.V., McCormick, J., Yoder, N., Belnap, D.M., Erlendsson, S., Morado, D.R., et al. (2018). The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA

Transfer. Cell *172*, 275-288.e18. 10.1016/j.cell.2017.12.024.

158. Korb, E., and Finkbeiner, S. (2011). Arc in synaptic plasticity: from gene to behavior. Trends Neurosci. *34*, 591–598. 10.1016/j.tins.2011.08.007.

159. Abed, M., Verschueren, E., Budayeva, H., Liu, P., Kirkpatrick, D.S., Reja, R., Kummerfeld, S.K., Webster, J.D., Gierke, S., Reichelt, M., et al. (2019). The Gag protein PEG10 binds to RNA and regulates trophoblast stem cell lineage specification. PloS One *14*, e0214110. 10.1371/journal.pone.0214110.

160. Ono, R., Nakamura, K., Inoue, K., Naruse, M., Usami, T., Wakisaka-Saito, N., Hino, T., Suzuki-Migishima, R., Ogonuki, N., Miki, H., et al. (2006). Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Nat. Genet. *38*, 101–106. 10.1038/ng1699.

161. Henke, C., Strissel, P.L., Schubert, M.-T., Mitchell, M., Stolt, C.C., Faschingbauer, F., Beckmann, M.W., and Strick, R. (2015). Selective expression of sense and antisense transcripts of the sushi-ichi-related retrotransposon--derived family during mouse placentogenesis. Retrovirology *12*, 9. 10.1186/s12977-015-0138-8.

162. Voelkel, C., Galla, M., Maetzig, T., Warlich, E., Kuehle, J., Zychlinski, D., Bode, J., Cantz, T., Schambach, A., and Baum, C. (2010). Protein transduction from retroviral Gag precursors. Proc. Natl. Acad. Sci. U. S. A. *107*, 7805–7810. 10.1073/pnas.0914517107.

163. Kaczmarczyk, S.J., Sitaraman, K., Young, H.A., Hughes, S.H., and Chatterjee, D.K. (2011). Protein delivery using engineered virus-like particles. Proc. Natl. Acad. Sci. *108*, 16998–17003. 10.1073/pnas.1101874108.

164. Shields, A., Witte, W.N., Rothenberg, E., and Baltimore, D. (1978). High frequency of aberrant expression of Moloney murine leukemia virus in clonal infections. Cell *14*, 601–609. 10.1016/0092-8674(78)90245-3.

165. Jones, T.A., Blaug, G., Hansen, M., and Barklis, E. (1990). Assembly of gag-beta-galactosidase proteins into retrovirus particles. J. Virol. *64*, 2265–2279.

10.1128/JVI.64.5.2265-2279.1990.

166. Hansen, M.S., and Barklis, E. (1995). Structural interactions between retroviral Gag proteins examined by cysteine cross-linking. J. Virol. *69*, 1150–1159. 10.1128/JVI.69.2.1150-1159.1995.

167. Hansen, M., Jelinek, L., Whiting, S., and Barklis, E. (1990). Transport and assembly of gag proteins into Moloney murine leukemia virus. J. Virol. *64*, 5306–5316. 10.1128/JVI.64.11.5306-5316.1990.

168. Andrawiss, M., Takeuchi, Y., Hewlett, L., and Collins, M. (2003). Murine Leukemia Virus Particle Assembly Quantitated by Fluorescence Microscopy: Role of Gag-Gag Interactions and Membrane Association. J. Virol. *77*, 11651–11660. 10.1128/JVI.77.21.11651-11660.2003.

169. Choi, J.G., Dang, Y., Abraham, S., Ma, H., Zhang, J., Guo, H., Cai, Y., Mikkelsen, J.G., Wu, H., Shankar, P., et al. (2016). Lentivirus pre-packed with Cas9 protein for safer gene editing. Gene Ther. *23*, 627–633. 10.1038/gt.2016.27.

170. Segel, M., Lash, B., Song, J., Ladha, A., Liu, C.C., Jin, X., Mekhedov, S.L., Macrae, R.K., Koonin, E.V., and Zhang, F. (2021). Mammalian retrovirus-like protein PEG10 packages its own mRNA and can be pseudotyped for mRNA delivery. Science *373*, 882–889. 10.1126/science.abg6155.

171. Huang, J., Liu, X., Li, D., Shao, Z., Cao, H., Zhang, Y., Trompouki, E., Bowman, T.V., Zon, L.I., Yuan, G.-C., et al. (2016). Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis. Dev. Cell *36*, 9–23. 10.1016/j.devcel.2015.12.014.

172. Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., et al. (2014). Chromatin state dynamics during blood formation. Science *345*, 943–949. 10.1126/science.1256271.

173. Fueyo, R., Judd, J., Feschotte, C., and Wysocka, J. (2022). Roles of transposable

elements in the regulation of mammalian transcription. Nat. Rev. Mol. Cell Biol., 1–17. 10.1038/s41580-022-00457-y.

174. Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Gonçalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P., and Odom, D.T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell *148*, 335–348. 10.1016/j.cell.2011.11.058.

175. Fuentes, D.R., Swigut, T., and Wysocka, J. (2018). Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. eLife *7*, e35989. 10.7554/eLife.35989.

176. Pehrsson, E.C., Choudhary, M.N.K., Sundaram, V., and Wang, T. (2019). The epigenomic landscape of transposable elements across normal human development and anatomy. Nat. Commun. *10*, 5640. 10.1038/s41467-019-13555-x.

177. Friedman, J.R., Fredericks, W.J., Jensen, D.E., Speicher, D.W., Huang, X.P., Neilson, E.G., and Rauscher, F.J. (1996). KAP-1, a novel corepressor for the highly conserved KRAB repression domain. Genes Dev. *10*, 2067–2078. 10.1101/gad.10.16.2067.

178. Karimi, M.M., Goyal, P., Maksakova, I.A., Bilenky, M., Leung, D., Tang, J.X., Shinkai, Y., Mager, D.L., Jones, S., Hirst, M., et al. (2011). DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements and chimaeric transcripts in mouse ES cells. Cell Stem Cell *8*, 676–687. 10.1016/j.stem.2011.04.004.

179. Ohnuki, M., Tanabe, K., Sutou, K., Teramoto, I., Sawamura, Y., Narita, M., Nakamura, M., Tokunaga, Y., Nakamura, M., Watanabe, A., et al. (2014). Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. Proc. Natl. Acad. Sci. U. S. A. *111*, 12426–12431. 10.1073/pnas.1413299111.

180. Friedli, M., Turelli, P., Kapopoulou, A., Rauwel, B., Castro-Díaz, N., Rowe, H.M., Ecco, G., Unzu, C., Planet, E., Lombardo, A., et al. (2014). Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency. Genome Res. *24*, 1251–

1259. 10.1101/gr.172809.114.

181. Ye, M., Goudot, C., Hoyler, T., Lemoine, B., Amigorena, S., and Zueva, E. (2020). Specific subfamilies of transposable elements contribute to different domains of T lymphocyte enhancers. Proc. Natl. Acad. Sci. *117*, 7905–7916. 10.1073/pnas.1912008117.

182. Adoue, V., Binet, B., Malbec, A., Fourquet, J., Romagnoli, P., van Meerwijk, J.P.M., Amigorena, S., and Joffre, O.P. (2019). The Histone Methyltransferase SETDB1 Controls T Helper Cell Lineage Integrity by Repressing Endogenous Retroviruses. Immunity *50*, 629-644.e8. 10.1016/j.immuni.2019.01.003.

183. Best, S., Tissier, P.L., Towers, G., and Stoye, J.P. (1996). Positional cloning of the mouse retrovirus restriction gene Fvl. Nature *382*, 826–829. 10.1038/382826a0.

184. Srinivasachar Badarinarayan, S., and Sauter, D. (2021). Switching Sides: How Endogenous Retroviruses Protect Us from Viral Infections. J. Virol. *95*, e02299-20. 10.1128/JVI.02299-20.

185. Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A., et al. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat. Genet. *51*, 1664–1669. 10.1038/s41588-019-0538-0.

186. Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F., et al. (2021). Genome-wide enhancer maps link risk variants to disease genes. Nature, 1–6. 10.1038/s41586-021-03446-x.

187. Nottingham, W.T., Jarratt, A., Burgess, M., Speck, C.L., Cheng, J.-F., Prabhakar, S., Rubin, E.M., Li, P.-S., Sloane-Stanley, J., Kong-a-San, J., et al. (2007). Runx1-mediated hematopoietic stem-cell emergence is controlled by a Gata/Ets/SCL-regulated enhancer. Blood *110*, 4188–4197. 10.1182/blood-2007-07-100883.

188. Gao, X., Johnson, K.D., Chang, Y.-I., Boyer, M.E., Dewey, C.N., Zhang, J., and Bresnick,

E.H. (2013). Gata2 cis-element is required for hematopoietic stem cell generation in the mammalian embryo. J. Exp. Med. *210*, 2833–2842. 10.1084/jem.20130733.

189. Bauer, D.E., Kamran, S.C., Lessard, S., Xu, J., Fujiwara, Y., Lin, C., Shao, Z., Canver, M.C., Smith, E.C., Pinello, L., et al. (2013). An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science *342*, 253–257. 10.1126/science.1242088.

190. Petersen, R., Lambourne, J.J., Javierre, B.M., Grassi, L., Kreuzhuber, R., Ruklisa, D., Rosa, I.M., Tomé, A.R., Elding, H., van Geffen, J.P., et al. (2017). Platelet function is modified by common sequence variation in megakaryocyte super enhancers. Nat. Commun. *8*, 16058. 10.1038/ncomms16058.

191. Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J., et al. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat. Genet. *advance online publication*. 10.1038/ng.3646.

192. Buenrostro, J.D., Corces, M.R., Lareau, C.A., Wu, B., Schep, A.N., Aryee, M.J., Majeti, R., Chang, H.Y., and Greenleaf, W.J. (2018). Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. Cell *0*. 10.1016/j.cell.2018.03.074.

193. Consortium, T.E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. 10.1038/nature11247.

194. Ulirsch, J.C., Lareau, C.A., Bao, E.L., Ludwig, L.S., Guo, M.H., Benner, C., Satpathy, A.T., Kartha, V.K., Salem, R.M., Hirschhorn, J.N., et al. (2019). Interrogation of human hematopoiesis at single-cell and single-variant resolution. Nat. Genet., 1. 10.1038/s41588-019-0362-6.

195. Ludwig, L.S., Lareau, C.A., Bao, E.L., Nandakumar, S.K., Muus, C., Ulirsch, J.C., Chowdhary, K., Buenrostro, J.D., Mohandas, N., An, X., et al. (2019). Transcriptional

States and Chromatin Accessibility Underlying Human Erythropoiesis. Cell Rep. *27*, 3228-3240.e7. 10.1016/j.celrep.2019.05.046.

196. Calderon, D., Nguyen, M.L.T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J.V., et al. (2019). Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat. Genet. *51*, 1494–1505. 10.1038/s41588-019-0505-9.
197. Layer, R.M., Pedersen, B.S., DiSera, T., Marth, G.T., Gertz, J., and Quinlan, A.R. (2018). GIGGLE: a search engine for large-scale integrated genome analysis. Nat. Methods *15*, 123–126. 10.1038/nmeth.4556.
198. Bentsen, M., Goymann, P., Schultheis, H., Klee, K., Petrova, A., Wiegandt, R., Fust, A., Preussner, J., Kuenne, C., Braun, T., et al. (2020). ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nat. Commun. *11*, 4267. 10.1038/s41467-020-18035-1.
199. Ivancevic, A., Simpson, D.M., and Chuong, E.B. (2021). Endogenous retroviruses mediate transcriptional rewiring in response to oncogenic signaling in colorectal cancer. 2021.10.28.466196. 10.1101/2021.10.28.466196.
200. Le Coz, C., Nguyen, D.N., Su, C., Nolan, B.E., Albrecht, A.V., Xhani, S., Sun, D., Demaree, B., Pillarisetti, P., Khanna, C., et al. (2021). Constrained chromatin accessibility in PU.1-mutated agammaglobulinemia patients. J. Exp. Med. *218*, e20201750. 10.1084/jem.20201750.
201. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589. 10.1016/j.molcel.2010.05.004.
202. Laurenti, E., and Göttgens, B. (2018). From haematopoietic stem cells to complex differentiation landscapes. Nature *553*, 418–426. 10.1038/nature25022.

203. Yamamoto, R., Wilkinson, A.C., and Nakauchi, H. (2018). Changing concepts in hematopoietic stem cells. Science *362*, 895–896. 10.1126/science.aat7873.

204. He, J., Babarinde, I.A., Sun, L., Xu, S., Chen, R., Shi, J., Wei, Y., Li, Y., Ma, G., Zhuang, Q., et al. (2021). Identifying transposable element expression dynamics and heterogeneity during development at the single-cell level with a processing pipeline scTE. Nat. Commun. *12*, 1456. 10.1038/s41467-021-21808-x.

205. Jin, Y., Tam, O.H., Paniagua, E., and Hammell, M. (2015). TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinforma. Oxf. Engl. *31*, 3593–3599. 10.1093/bioinformatics/btv422.

206. Liu, F., Barsyte-Lovejoy, D., Li, F., Xiong, Y., Korboukh, V., Huang, X.-P., Allali-Hassani, A., Janzen, W.P., Roth, B.L., Frye, S.V., et al. (2013). Discovery of an in vivo chemical probe of the lysine methyltransferases G9a and GLP. J. Med. Chem. *56*, 8931–8942. 10.1021/jm401480r.

207. Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., et al. (2020). Global reference mapping of human transcription factor footprints. Nature *583*, 729–736. 10.1038/s41586-020-2528-x.

208. Renoux, V.M., Zriwil, A., Peitzsch, C., Michaëlsson, J., Friberg, D., Soneji, S., and Sitnicka, E. (2015). Identification of a Human Natural Killer Cell Lineage-Restricted Progenitor in Fetal and Adult Tissues. Immunity *43*, 394–407. 10.1016/j.immuni.2015.07.011.

209. Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods *14*, 975–978. 10.1038/nmeth.4401.

210. Ebihara, T., Song, C., Ryu, S.H., Plougastel-Douglas, B., Yang, L., Levanon, D., Groner, Y., Bern, M.D., Stappenbeck, T.S., Colonna, M., et al. (2015). Runx3 specifies lineage commitment of innate lymphoid cells. Nat. Immunol. *16*, 1124–1133. 10.1038/ni.3272.

211. Guo, Y., Maillard, I., Chakraborti, S., Rothenberg, E.V., and Speck, N.A. (2008). Core

binding factors are necessary for natural killer cell development and cooperate with Notch signaling during T-cell specification. Blood *112*, 480–492. 10.1182/blood-2007-10-120261.

212. Levanon, D., Negreanu, V., Lotem, J., Bone, K.R., Brenner, O., Leshkowitz, D., and Groner, Y. (2014). Transcription factor Runx3 regulates interleukin-15-dependent natural killer cell activation. Mol. Cell. Biol. *34*, 1158–1169. 10.1128/MCB.01202-13.

213. Crinier, A., Milpied, P., Escalière, B., Piperoglou, C., Galluso, J., Balsamo, A., Spinelli, L., Cervera-Marzal, I., Ebbo, M., Girard-Madoux, M., et al. (2018). High-Dimensional Single-Cell Analysis Identifies Organ-Specific Signatures and Conserved NK Cell Subsets in Humans and Mice. Immunity *49*, 971-986.e5. 10.1016/j.immuni.2018.09.009.

214. Wang, J., Vicente-García, C., Seruggia, D., Moltó, E., Fernandez-Miñán, A., Neto, A., Lee, E., Gómez-Skarmeta, J.L., Montoliu, L., Lunyak, V.V., et al. (2015). MIR retrotransposon sequences provide insulators to the human genome. Proc. Natl. Acad. Sci. U. S. A. *112*, E4428-4437. 10.1073/pnas.1507253112.

215. Stunnenberg, H.G., International Human Epigenome Consortium, and Hirst, M. (2016). The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell *167*, 1145–1149. 10.1016/j.cell.2016.11.007.

216. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330. 10.1038/nature14248.

217. Gasperini, M., Tome, J.M., and Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat. Rev. Genet. *21*, 292–310. 10.1038/s41576-019-0209-0.

218. Keenan, C.R., Iannarella, N., Naselli, G., Bediaga, N.G., Johanson, T.M., Harrison, L.C., and Allan, R.S. (2020). Extreme disruption of heterochromatin is required for accelerated hematopoietic aging. Blood *135*, 2049–2058. 10.1182/blood.2019002990.

219. Liu, E., Marin, D., Banerjee, P., Macapinlac, H.A., Thompson, P., Basar, R., Nassif Kerbauy, L., Overman, B., Thall, P., Kaplan, M., et al. (2020). Use of CAR-Transduced Natural Killer Cells in CD19-Positive Lymphoid Tumors. N. Engl. J. Med. *382*, 545–553. 10.1056/NEJMoa1910607.

220. Miller, J.S., Soignier, Y., Panoskaltsis-Mortari, A., McNearney, S.A., Yun, G.H., Fautsch, S.K., McKenna, D., Le, C., Defor, T.E., Burns, L.J., et al. (2005). Successful adoptive transfer and in vivo expansion of human haploidentical NK cells in patients with cancer. Blood *105*, 3051–3057. 10.1182/blood-2004-07-2974.

221. Freud, A.G., Mundy-Bosse, B.L., Yu, J., and Caligiuri, M.A. (2017). The Broad Spectrum of Human Natural Killer Cell Diversity. Immunity *47*, 820–833. 10.1016/j.immuni.2017.10.008.

222. Rawat, P., and Das, A. (2022). Differential expression of disparate transcription factor regime holds the key for NK cell development and function modulation. Life Sci. *297*, 120471. 10.1016/j.lfs.2022.120471.

223. Cichocki, F., Grzywacz, B., and Miller, J.S. (2019). Human NK Cell Development: One Road or Many? Front. Immunol. *10*, 2078. 10.3389/fimmu.2019.02078.

224. Dege, C., Fegan, K.H., Creamer, J.P., Berrien-Elliott, M.M., Luff, S.A., Kim, D., Wagner, J.A., Kingsley, P.D., McGrath, K.E., Fehniger, T.A., et al. (2020). Potently Cytotoxic Natural Killer Cells Initially Emerge from Erythro-Myeloid Progenitors during Mammalian Development. Dev. Cell *0*. 10.1016/j.devcel.2020.02.016.

225. Wagner, J.A., Rosario, M., Romee, R., Berrien-Elliott, M.M., Schneider, S.E., Leong, J.W., Sullivan, R.P., Jewell, B.A., Becker-Hapak, M., Schappe, T., et al. (2017). CD56bright NK cells exhibit potent antitumor responses following IL-15 priming. J. Clin. Invest. *127*, 4042–4058. 10.1172/JCI90387.

226. Romee, R., Rosario, M., Berrien-Elliott, M.M., Wagner, J.A., Jewell, B.A., Schappe, T., Leong, J.W., Abdel-Latif, S., Schneider, S.E., Willey, S., et al. (2016). Cytokine-induced memory-like natural killer cells exhibit enhanced responses against myeloid leukemia. Sci.

Transl. Med. *8*, 357ra123. 10.1126/scitranslmed.aaf2341.

227. Romee, R., Schneider, S.E., Leong, J.W., Chase, J.M., Keppel, C.R., Sullivan, R.P., Cooper, M.A., and Fehniger, T.A. (2012). Cytokine activation induces human memory-like NK cells. Blood *120*, 4751–4760. 10.1182/blood-2012-04-419283.

228. Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Methods *14*, 959–962. 10.1038/nmeth.4396.

229. Smith, J.P., Corces, M.R., Xu, J., Reuter, V.P., Chang, H.Y., and Sheffield, N.C. (2021). PEPATAC: an optimized pipeline for ATAC-seq data analysis with serial alignments. NAR Genomics Bioinforma. *3*, lqab101. 10.1093/nargab/lqab101.

230. Saito, T., and Rehmsmeier, M. (2017). Precrec: fast and accurate precision–recall and ROC curve calculations in R. Bioinformatics *33*, 145–147. 10.1093/bioinformatics/btw570.

231. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*, eaav1898. 10.1126/science.aav1898.

232. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. 10.1186/s13059-014-0550-8.

233. Granja, J.M., Klemm, S., McGinnis, L.M., Kathiria, A.S., Mezger, A., Corces, M.R., Parks, B., Gars, E., Liedtke, M., Zheng, G.X.Y., et al. (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat. Biotechnol., 1–8. 10.1038/s41587-019-0332-7.

234. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196.

10.1126/science.aad0501.

235. Dou, D.R., Calvanese, V., Sierra, M.I., Nguyen, A.T., Minasian, A., Saarikoski, P., Sasidharan, R., Ramirez, C.M., Zack, J.A., Crooks, G.M., et al. (2016). Medial HOXA genes demarcate haematopoietic stem cell fate during human development. Nat. Cell Biol. *18*, 595–606. 10.1038/ncb3354.

236. Ng, E.S., Azzola, L., Bruveris, F.F., Calvanese, V., Phipson, B., Vlahos, K., Hirst, C., Jokubaitis, V.J., Yu, Q.C., Maksimovic, J., et al. (2016). Differentiation of human embryonic stem cells to HOXA+ hemogenic vasculature that resembles the aorta-gonad-mesonephros. Nat. Biotechnol. *advance online publication*. 10.1038/nbt.3702.

237. Wilson, N.K., Foster, S.D., Wang, X., Knezevic, K., Schütte, J., Kaimakis, P., Chilarska, P.M., Kinston, S., Ouwehand, W.H., Dzierzak, E., et al. (2010). Combinatorial Transcriptional Control In Blood Stem/Progenitor Cells: Genome-wide Analysis of Ten Major Transcriptional Regulators. Cell Stem Cell *7*, 532–544. 10.1016/j.stem.2010.07.016.

238. Pimanda, J.E., Ottersbach, K., Knezevic, K., Kinston, S., Chan, W.Y.I., Wilson, N.K., Landry, J.-R., Wood, A.D., Kolb-Kokocinski, A., Green, A.R., et al. (2007). Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development. Proc. Natl. Acad. Sci. *104*, 17692–17697. 10.1073/pnas.0707045104.

239. Costa, G., Mazan, A., Gandillet, A., Pearson, S., Lacaud, G., and Kouskoff, V. (2012). SOX7 regulates the expression of VE-cadherin in the haemogenic endothelium at the onset of haematopoietic development. Development *139*, 1587–1598. 10.1242/dev.071282.

240. Goode, D.K., Obier, N., Vijayabaskar, M.S., Lie-A-Ling, M., Lilly, A.J., Hannah, R., Lichtinger, M., Batta, K., Florkowska, M., Patel, R., et al. (2016). Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. Dev. Cell *36*, 572–587. 10.1016/j.devcel.2016.01.024.

241. Obier, N., Cauchy, P., Assi, S.A., Gilmour, J., Lie-A-Ling, M., Lichtinger, M., Hoogenkamp,

M., Noailles, L., Cockerill, P.N., Lacaud, G., et al. (2016). Cooperative binding of AP-1 and TEAD4 modulates the balance between vascular smooth muscle and hemogenic cell fate. Dev. Camb. Engl. *143*, 4324–4340. 10.1242/dev.139857.

242. Clarke, R.L., Yzaguirre, A.D., Yashiro-Ohtani, Y., Bondue, A., Blanpain, C., Pear, W.S., Speck, N.A., and Keller, G. (2013). The expression of Sox17 identifies and regulates haemogenic endothelium. Nat. Cell Biol. *15*, 502–510. 10.1038/ncb2724.

243. Jung, H.S., Uenishi, G., Park, M.A., Liu, P., Suknuntha, K., Raymond, M., Choi, Y.J., Thomson, J.A., Ong, I.M., and Slukvin, I.I. (2021). SOX17 integrates HOXA and arterial programs in hemogenic endothelium to drive definitive lympho-myeloid hematopoiesis. Cell Rep. *34*, 108758. 10.1016/j.celrep.2021.108758.

244. Kim, I., Saunders, T.L., and Morrison, S.J. (2007). Sox17 Dependence Distinguishes the Transcriptional Regulation of Fetal from Adult Hematopoietic Stem Cells. Cell *130*, 470–483. 10.1016/j.cell.2007.06.011.

245. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501. 10.1038/nbt.1630.

246. Kim, P.G., Canver, M.C., Rhee, C., Ross, S.J., Harriss, J.V., Tu, H.-C., Orkin, S.H., Tucker, H.O., and Daley, G.Q. (2016). Interferon-α signaling promotes embryonic HSC maturation. Blood *128*, 204–216. 10.1182/blood-2016-01-689281.

247. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457. 10.1038/nmeth.3337.

248. Ranzoni, A.M., Tangherloni, A., Berest, I., Riva, S.G., Myers, B., Strzelecka, P.M., Xu, J., Panada, E., Mohorianu, I., Zaugg, J.B., et al. (2020). Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. Cell Stem Cell. 10.1016/j.stem.2020.11.015.

249. Takayama, N., Murison, A., Takayanagi, S., Arlidge, C., Zhou, S., Garcia-Prat, L., Chan-Seng-Yue, M., Zandi, S., Gan, O.I., Boutzen, H., et al. (2020). The Transition from Quiescent to Activated States in Human Hematopoietic Stem Cells Is Governed by Dynamic 3D Genome Reorganization. Cell Stem Cell. 10.1016/j.stem.2020.11.001.

250. Laurenti, E., Doulatov, S., Zandi, S., Plumb, I., Chen, J., April, C., Fan, J.-B., and Dick, J.E. (2013). The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. Nat. Immunol. *14*, 756–763. 10.1038/ni.2615.

251. Wang, Z., Civelek, M., Miller, C.L., Sheffield, N.C., Guertin, M.J., and Zang, C. (2018). BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. Bioinformatics *34*, 2867–2869. 10.1093/bioinformatics/bty194.

252. Lim, K.-C., Hosoya, T., Brandt, W., Ku, C.-J., Hosoya-Ohmura, S., Camper, S.A., Yamamoto, M., and Engel, J.D. (2012). Conditional *Gata2* inactivation results in HSC loss and lymphatic mispatterning. J. Clin. Invest. *122*, 3705–3717. 10.1172/JCI61619.

253. Menendez-Gonzalez, J.B., Vukovic, M., Abdelfattah, A., Saleh, L., Almotiri, A., Thomas, L., Agirre-Lizaso, A., Azevedo, A., Menezes, A.C., Tornillo, G., et al. (2019). Gata2 as a Crucial Regulator of Stem Cells in Adult Hematopoiesis and Acute Myeloid Leukemia. Stem Cell Rep. *13*, 291–306. 10.1016/j.stemcr.2019.07.005.

254. Pater, E. de, Kaimakis, P., Vink, C.S., Yokomizo, T., Yamada-Inagawa, T., Linden, R. van der, Kartalaei, P.S., Camper, S.A., Speck, N., and Dzierzak, E. (2013). Gata2 is required for HSC generation and survival. J. Exp. Med. *210*, 2843–2850. 10.1084/jem.20130751.

255. Soto, R.A., Najia, M.A.T., Hachimi, M., Frame, J.M., Yette, G.A., Lummertz da Rocha, E., Stankunas, K., Daley, G.Q., and North, T.E. (2021). Sequential regulation of hemogenic fate and hematopoietic stem and progenitor cell formation from arterial endothelium by Ezh1/2. Stem Cell Rep. 10.1016/j.stemcr.2021.05.014.

256. Yin, J., Leavenworth, J.W., Li, Y., Luo, Q., Xie, H., Liu, X., Huang, S., Yan, H., Fu, Z., Zhang, L.Y., et al. (2015). Ezh2 regulates differentiation and function of natural killer cells

through histone methyltransferase activity. Proc. Natl. Acad. Sci. U. S. A. *112*, 15988–15993. 10.1073/pnas.1521740112.

257. Li, Y., Hermanson, D.L., Moriarity, B.S., and Kaufman, D.S. (2018). Human iPSC-Derived Natural Killer Cells Engineered with Chimeric Antigen Receptors Enhance Anti-tumor Activity. Cell Stem Cell *23*, 181-192.e5. 10.1016/j.stem.2018.06.002.

258. Cichocki, F., Bjordahl, R., Gaidarova, S., Mahmood, S., Abujarour, R., Wang, H., Tuininga, K., Felices, M., Davis, Z.B., Bendzick, L., et al. (2020). iPSC-derived NK cells maintain high cytotoxicity and enhance in vivo tumor control in concert with T cells and anti–PD-1 therapy. Sci. Transl. Med. *12*, eaaz5618. 10.1126/scitranslmed.aaz5618.

259. Knorr, D.A., Ni, Z., Hermanson, D., Hexum, M.K., Bendzick, L., Cooper, L.J.N., Lee, D.A., and Kaufman, D.S. (2013). Clinical-Scale Derivation of Natural Killer Cells From Human Pluripotent Stem Cells for Cancer Therapy. Stem Cells Transl. Med. *2*, 274–283. 10.5966/sctm.2012-0084.

260. Zhu, H., Blum, R.H., Bernareggi, D., Ask, E.H., Wu, Z., Hoel, H.J., Meng, Z., Wu, C., Guan, K.-L., Malmberg, K.-J., et al. (2020). Metabolic Reprograming via Deletion of CISH in Human iPSC-Derived NK Cells Promotes In Vivo Persistence and Enhances Anti-tumor Activity. Cell Stem Cell *27*, 224-237.e6. 10.1016/j.stem.2020.05.008.

261. Zhu, H., Blum, R.H., Bjordahl, R., Gaidarova, S., Rogers, P., Lee, T.T., Abujarour, R., Bonello, G.B., Wu, J., Tsai, P.-F., et al. (2020). Pluripotent stem cell–derived NK cells with high-affinity noncleavable CD16a mediate improved antitumor activity. Blood *135*, 399–410. 10.1182/blood.2019000621.

262. Liu, E., Tong, Y., Dotti, G., Shaim, H., Savoldo, B., Mukherjee, M., Orange, J., Wan, X., Lu, X., Reynolds, A., et al. (2018). Cord blood NK cells engineered to express IL-15 and a CD19-targeted CAR show long-term persistence and potent antitumor activity. Leukemia *32*, 520–531. 10.1038/leu.2017.226.

263. Imai, C., Iwamoto, S., and Campana, D. (2005). Genetic modification of primary natural

killer cells overcomes inhibitory signals and induces specific killing of leukemic cells. Blood *106*, 376–383. 10.1182/blood-2004-12-4797.

264. Rezvani, K., Rouce, R., Liu, E., and Shpall, E. (2017). Engineering Natural Killer Cells for Cancer Immunotherapy. Mol. Ther. J. Am. Soc. Gene Ther. *25*, 1769–1781. 10.1016/j.ymthe.2017.06.012.

265. Larson, R.C., Kann, M.C., Bailey, S.R., Haradhvala, N.J., Llopis, P.M., Bouffard, A.A., Scarfó, I., Leick, M.B., Grauwet, K., Berger, T.R., et al. (2022). CAR T cell killing requires the IFNγR pathway in solid but not liquid tumours. Nature *604*, 563–570. 10.1038/s41586-022-04585-5.

266. Emert, B.L., Cote, C.J., Torre, E.A., Dardani, I.P., Jiang, C.L., Jain, N., Shaffer, S.M., and Raj, A. (2021). Variability within rare cell states enables multiple paths toward drug resistance. Nat. Biotechnol. *39*, 865–876. 10.1038/s41587-021-00837-3.

267. Hoppe, P.S., Schwarzfischer, M., Loeffler, D., Kokkaliaris, K.D., Hilsenbeck, O., Moritz, N., Endele, M., Filipczyk, A., Gambardella, A., Ahmed, N., et al. (2016). Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. Nature *535*, 299–302. 10.1038/nature18320.

268. Shaffer, S.M., Dunagin, M.C., Torborg, S.R., Torre, E.A., Emert, B., Krepler, C., Beqiri, M., Sproesser, K., Brafford, P.A., Xiao, M., et al. (2017). Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature *546*, 431–435. 10.1038/nature22794.

269. Eling, N., Morgan, M.D., and Marioni, J.C. (2019). Challenges in measuring and understanding biological noise. Nat. Rev. Genet. *20*, 536–548. 10.1038/s41576-019-0130-6.

270. Gupta, P.B., Fillmore, C.M., Jiang, G., Shapira, S.D., Tao, K., Kuperwasser, C., and Lander, E.S. (2011). Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells. Cell *146*, 633–644. 10.1016/j.cell.2011.07.026.

271. Chen, W., Guillaume-Gentil, O., Rainer, P.Y., Gäbelein, C.G., Saelens, W., Gardeux, V., Klaeger, A., Dainese, R., Zachara, M., Zambelli, T., et al. (2022). Live-seq enables temporal transcriptomic recording of single cells. Nature *608*, 733–740. 10.1038/s41586-022-05046-9.