UNIVERSITY OF CALIFORNIA SAN DIEGO

Using Emerging Next-Generation Sequencing Technologies to Enhance the Lifecycle of Biopharmaceuticals

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioinformatics & Systems Biology

by

Isaac Sam Shamie

Committee in charge:

Professor Nathan E. Lewis, Chair
Professor Chris Benner, Co-Chair
Professor Hannah Carter
Professor Ben Croker
Professor Niema Moshiri

2022

The Dissertation of Isaac Sam Shamie is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my parents, who have given me strength and unconditional love to pursue my whims
To my siblings Louis, Jack, Ezra, Lorraine, Michelle, Stella, & Grace,
who always make home feel home
To Eddie & Hoff, for collaborating in life, and Michael, Ikey & Bobby for the good times
To Ileena, Kevin, Curtis, Hratch, & all my other San Diego friends that kept me sane and happy
To all my teachers, PIs, and colleagues, for continued inspiration & curiosity of the world
To my cat Agnes, my amazing co-working partner for the last 3 years
To life itself

EPIGRAPH

And life flows on
within you
and without you

- *George Harrison*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I thank my advisor Professor Nathan Lewis for his support, guidance, and encouragement throughout my graduate career. Additionally, committee member Professor Ben Croker has given me great support and invaluable guidance across multiple projects.

I express my gratitude to my dissertation committee members Professor Chris Benner, Professor Hannah Carter, and Professor Niema Moshiri for their input and advice. I also thank my Lewis Lab colleagues, especially Curtis, Ben, Hratch, Chintan, Hooman, Julie, Matt, Austin, Erick, Jasmine & Phil, who have all fueled my love of science through endless discussions and collaborations. Most importantly, I thank my family for believing in me throughout my life.

Chapter 2, in full, is a reprint of the material as it appears in "Isaac Shamie, Sascha H Duttke, Karen J la Cour Karottki, Claudia Z Han, Anders H Hansen, Hooman Hefzi, Kai Xiong, Shangzhong Li, Samuel J Roth, Jenhan Tao, Gyun Min Lee, Christopher K Glass, Helene Faustrup Kildegaard, Christopher Benner, Nathan E Lewis, A Chinese hamster transcription start site atlas that enables targeted editing of CHO cells, *NAR Genomics and Bioinformatics*, Volume 3, Issue 3, September 2021, lqab061, https://doi.org/10.1093/nargab/lqab061." The dissertation author was the primary investigator and author.

Chapter 3, in full, is currently being prepared for submission for publication of the material in a manuscript titled "Comparative single-cell lineage bias in human and murine hematopoietic stem cells", Isaac Shamie, Meghan Bliss-Moreau, Jamie Casey Lee, Nathan E. Lewis, Yanfang Peipei Zhu, Ben A. Croker. The dissertation author was the primary researcher and author of this material.

VITA

2014 Bachelor of Science in Biology, Carnegie Mellon University

2022 Doctor of Philosophy in Bioinformatics & Systems Biology, University of California San Diego

PUBLICATIONS

1. Alasfour, A. et al. Spatiotemporal dynamics of human high gamma discriminate naturalistic behavioral states. PLoS Comput. Biol. 18, e1010401 (2022).

2. Armingol, E. et al. Inferring a spatial code of cell-cell interactions across a whole animal body. BioRxiv 2020–2011 (2022).

3. Joshi, C. J. et al. StanDep: Capturing transcriptomic variability improves context-specific metabolic models (vol 16, e1007764, 2020). PLoS Comput. Biol. 18, (2022).

4. Kellman, B. P. et al. Multiple freeze-thaw cycles lead to a loss of consistency in poly (A)-enriched RNA sequencing. BMC Genomics 22, 1–15 (2021).

5. Zhu, Y. P. et al. Immune response to intravenous immunoglobulin in patients with Kawasaki disease and MIS-C. J. Clin. Invest. 131, (2021).

6. Shamie, I. et al. A Chinese hamster transcription start site atlas that enables targeted editing of CHO cells. NAR genomics and bioinformatics 3, lqab061 (2021).

7. Karottki, K. J. la C. et al. Awakening dormant glycosyltransferases in CHO cells with CRISPRa. Biotechnol. Bioeng. 117, 593–598 (2020).

8. Speir, M. et al. Ptpn6 inhibits caspase-8-and Ripk3/Mlkl-dependent inflammation. Nat. Immunol. 21, 54–64 (2020).

9. Xiong, K. et al. Reduced apoptosis in Chinese hamster ovary cells via optimized CRISPR interference. Biotechnol. Bioeng. 116, 1813–1819 (2019).

10. Alasfour, A. et al. Coarse behavioral context decoding. J. Neural Eng. 16, 016021 (2019).

11. Kuo, C.-C. et al. The emerging role of systems biology for engineering protein production in CHO cells. Curr. Opin. Biotechnol. 51, 64–69 (2018).

12. Jiang, X. & Shamie, I. K Doyle W, Friedman D, Dugan P, Devinsky O, Eskandar E, Cash SS, Thesen T, Halgren E. 2017. Replay of large-scale spatio-temporal patterns from waking during subsequent NREM sleep in human cortex. Sci. Rep. 7, 17380.

13. Krishnan, G. P. et al. Cellular and neurochemical basis of sleep stages in the thalamocortical network. Elife 5, e18607 (2016).

ABSTRACT OF THE DISSERTATION

Using Emerging Next-Generation Sequencing Technologies to Enhance the Lifecycle of Biopharmaceuticals

by

Isaac Sam Shamie

Doctor of Philosophy in Bioinformatics & Systems Biology

University of California San Diego, 2022

Professor Nathan E. Lewis, Chair
Professor Chris Benner, Co-Chair

Biopharmaceuticals are emerging as a promising avenue for treating a range of diseases [1], and bringing down the costs of production, as well as monitoring the response to treatment, is of high importance. One technology to assist in these needs is next generation sequencing (NGS). In this dissertation, I use emergent NGS techniques to provide valuable resources in two important cell types, the

Chinese hamster ovary (CHO) cell-line in Chapter 2, and CD34+ haematopoietic stem and progenitor cells (HSPCs) in Chapter 3.

CHO cells are currently the most used cell line for producing recombinant monoclonal antibodies, and optimization of this cell line including media control and gene engineering has brought down production costs [2]. However, costs can still be prohibitive, and the genome resources required for novel gene engineering techniques are limited by the resolution of the genome annotation. In Chapter 2, I revise the TSS genome annotation using two different TSS sequencing techniques across multiple tissues in the Chinese hamster. TSSs were detected in 15308 protein-coding genes, and detected TSSs in 13037 of these genes had TSSs revised by at least 10 base pairs from the nearest NCBI RefSeq TSS. More promoter motif elements were detected at the revised TSSs than at the NCBI TSSs. To demonstrate the accuracy and functionality of our revised TSS annotation we activated the dormant *Mgat3* gene in CHO cells by CRISPR activation [3] using a novel identified TSS.

CD34+ HSPC cells are the target in cytokine therapies to recruit cells to differentiate into desired immune lineages. Studying the clonal lineage multipotent capacity is important for understanding cellular response to therapy. Recent studies have shown that there is heterogeneity in HSPC clones, which are cells from the same phylogenetic origin, in both their growth and hematopoietic multipotent capacity. However, tracking HPSC clones in humans is limited. In Chapter 3, clonal heterogeneity is tracked across multiple donors in steady-state and in response to *ex vivo* cytokine cocktail using mitochondrial single-cell ATAC-seq. Cells are grouped into clones using naturally occurring somatic mitochondrial mutations, and their lineages assessed using regulatory regions in the nuclear open-chromatin. Larger clones make up a large fraction of donor HSPC donor populations, with no clone preferentially responding to culture. Most clones show multipotent capacity, differentiating into multiple immune lineage progenitors.

Overall, this dissertation provides an improved genome annotation in CHO cells and an analysis on HSPC lineage bias in steady-state and in response to cytokine treatment.

CHAPTER 1: INTRODUCTION

1.1 Biotherapeutics as emerging therapies

In 1982 the FDA approved the first biotherapeutic drug, Humulin, a human insulin recombinant protein produced in bacterial cells [4,5]. Since then the number of biotherapeutics, described by the World Health Organization as "biological medicinal products using genetically engineered bacteria, yeast, fungi, cells or even whole animals and plants" [6], has widely expanded. This includes therapies such as recombinant protein (RP) Enbrel, a fusion antibody for treating autoimmune diseases; cytokine-based therapy such as interleukin-2 to treat cancers; and autologous gene-therapy to treat sickle-cell disease [7]. 316 biopharmaceuticals were on the market in 2018 [1], and as of November 2021 there are 621 FDA-licensed biologics (the biological therapeutic product) [8], including monoclonal antibodies (mABs), hormones, clotting factors, and engineered cell-based products. Discovery of effective small molecule therapies has been in decline, leading to more research and investment into biotherapies. There are pressing needs, however, to both bring down production costs of these therapies [9–12] and to understand the downstream physiological effects of these molecules [13–16].

1.2 NGS technologies for advancing biotherapeutic development and monitoring

Complementing the research into biotherapies over the last 15 years is the rise of diverse and cost-effective next-generation sequencing (NGS) technologies, which perform millions of reactions in parallel to provide a high-throughput and sensitive technique to measure multiple molecules at once [17]. NGS has revolutionized how we understand molecular biology [18,19], and the first human genome solely using this technique was published in 2008 [20] (the first human genome published in 2000). As the cost of sequencing decreases annually, genetic testing, and mapping model organisms' genomes were promised to revolutionize our approach to human medicine. It was realized early on, however, that the genome was just the start to creating breakthrough therapeutics at a rapid pace, and NGS technology was then further

developed to not just study ‘genomics’, which measures an organism's complete set of DNA, but to investigate other areas of “-omics”. This includes transcriptomics, which, depending on the technique, measures expression of different RNA molecules; epigenomics, which measures different epigenomic signatures depending on the technique; and more recently, single-cell -omics, which measures the molecule type of interest in multiple single cells at once. Additionally, multi-omics allows the simultaneous measurements of distinct types of macromolecules. While -omics has been informative in developing and measuring effects of biotherapies, there still appears to be gaps in incorporating new NGS techniques to develop novel therapies [21,22]. Further, when producing therapeutics, much of the research is based on private datasets and proprietary knowledge, and there is a lack of focus on creating useful shared resources, widening information gaps that when filled, can benefit the entire field. When developing and testing novel therapies such as cytokine therapy and stem-cell transplantation, we want to understand the effects on the body in high resolution, which NGS provides. In this work, I help create valuable resources using recently-developed NGS techniques to the general biopharmaceutical industry to assist in producing and monitoring effects of biotherapies with two examples.

1.3 Characterizing the transcriptional architecture in industrially-relevant CHO cells

The Chinese hamster ovary (CHO) cell-line is one of the most used mammalian cell-line for production of recombinant proteins such as Enbrel and mABs [1]. After their immortalization, [23], their popularity as biotherapeutic producers rose due to their high growth rate [24], ease of genetic manipulation, and ability to produce complex post-translationally modified proteins that are not immunogenic in humans [25]. Optimizing CHO cells to increase production quantity and improving quality has been a priority to reduce costs. Over the past few decades, these optimization efforts have progressed from engineering the media and bioreactors to transgene codon sequence and more recently, cell engineering and synthetic biology [26,27].

Genome sequencing efforts in both CHO cells and the Chinese hamster have helped our ability to engineer CHO cells [28–30]. To improve cell performance and product quality, NGS technologies have been used to develop metabolic models, find differentially expressed genes between high- and low-producing cell lines, and discover context-specific promoters [31–35]. Costs, however, are still prohibitive, and differential expression does not provide enough actionable direction, as often dozens to hundreds of genes are detected across conditions. This has led the field to consider genetic engineering the host CHO cells either rationally [36,37], or in a systematic and unbiased manner through large-scale genetic screens [2] to better characterize gene effects. CRISPR activation (CRISPRa [38]) and other genetic engineering methods could prove instrumental to improving therapeutic protein production. To engineer gene expression, however, knowledge of the underlying regulatory elements is critical. This is due in part to the fact that the Chinese hamster genome annotation, which maps the gene regions in the genome, remains far from complete, especially for the approximately 50% of genes that are silenced in CHO cells, including many needed for producing more human-like proteins [34]. Specifically, no method that maps transcription start sites (TSSs) has been used in mapping the hamster annotation. TSS mapping is vital for using genetic engineering techniques such as CRISPRa and CRISPR inhibition (CRISPRi) relying on the guide RNAs targeting within 150 basepairs of the gene start site [39,40]. In Chapter 2, to assist in production of RPs, I intend to refine and extend the transcript annotation of the Chinese hamster using sensitive transcription start-site sequencing (TSS-Seq) methods. I will then demonstrate how this refinement can aid further optimization of the cell-line to produce recombinant proteins through effective gene engineering using CRISPR activation (CRISPRa).

1.4 Tracking clonal lineage bias in clinically-relevant HSPCs

Biotherapies targeting the immune-system have achieved recent success with the development of cytokine-based therapies, stem-cell targeted gene therapies, and CAR-T therapy. In beta-globin gene therapy in sickle-cell disease patients, CD34+ haematopoietic stem and progenitor cells (HSPCs) cells are either taken from the patient (‘autologously’) or from a donor (‘allogenic’), and genetically modified to

carry the therapeutic gene, and are then placed into the patient. These cells can then differentiate into circulating blood cells with the corrected gene, effectively treating the disease. Issues remain, however, in generating sufficient cells as well as tracking the lineage fates of the genetically modified HSPC clones once transplanted in the donor. CD34+ HSPCs are also an important target for different cytokine therapies that aim to recruit specific immune cells to a disease [41,42] such as with interleukin-2 [43], to promote differentiation of T- cells (amongst others), and Flt3l to promote dendritic cell differentiation [44,45]. The long-term effect of these HSPC cells, including their growth and lineage potential, still remains an active area of research [46–48]. Early hematopoietic stem cells (HSCs) are single cells with the potential to reconstitute the blood cell population, and HSC clones are downstream cells coming from the same HSC. These clones grow and respond differently to treatment. Recent studies have also shown that genetic mutations in specific clones increase their growth rate and correlates with bone-marrow malignancies [49–51]. Detecting a clone's potential for differentiating into different progenitor lineages, or its multipotency, in these HSCs is hampered by our ability to track cell clonal relationships in a human along with their downstream immune lineages. Studying HSPC lineage fate requires delineation of cell history, which requires the use of 'barcode', which is usually a nucleotide sequence to identify the cell, that can be passed onto daughter cells. Due to the limitation of inserting exogenous barcodes in healthy humans, tracking clonal relationships in humans has been limited to gene-therapy transplant studies, and cancer cells using somatic mutations. Single-cell NGS techniques that capture the epigenome and transcriptome have been used to infer lineage and clone relationships, but these measurements do not directly capture these cell relationships [52].

In Chapter 3, we explore clonal size heterogeneity and lineage-bias in native hematopoiesis as well as in response to cytokines in humans and mice. This establishes a baseline to understand the therapeutically relevant BM CD34+ cells.

1.5 Overview

In Chapter 2, I revise the Chinese Hamster transcription start-site annotation using multiple complementary NGS techniques (GRO-seq [53], 5'GRO-seq [54], csRNA-seq [55], ribosomal RNA-depleted RNA-seq and ATAC-seq [56], by capturing and confidently mapping nascent TSSs in CHO-K1 cells, 10 CH-derived tissues and hamster bone marrow derived macrophages (BMDMs). TSSs detected are integrated across all tissues to revise gene TSSs annotated by NCBI, and discover additional unannotated TSSs. These TSSs improve the capture of initiator and TATA-box motif elements compared to the prior annotation. Additionally, TSSs specific to certain tissues that corroborate with prior known tissue specific genes, as well as shared and distinct promoter elements across samples. A new TSS in MGAT3 is then corrobarated using CRISPRa to produce a clinically-relevant glycan.

In chapter 3, lineage bias is assessed across CD34+ HSPC clones in both native steady-state hematopoiesis across multiple healthy donors, and in response to cytokine treatment. This is done using a recently developed NGS technique that tracks cell clonality using somatic mitochondrial (MT) variants as naturally occurring barcodes, as well as nuclear open-chromatin regions to inform cell lineage through their epigenome. The technique, mitochondrial single-cell assay for transposase-accessible chromatin sequencing (mt-scATAC-seq), takes advantage of the MT's small genome size, high-mutation rate, and high per-cell copy-numbers to find both persistent and novel variants in cells of known phylogenetic (clonal) relationship. In this study, I show donors can be multiplexed in a sample and separated using germline MT mutations, and define clones using somatic MT mutations. Clonal detection showed that a few larger clones make up a large fraction of the clone population. Cells were then clustered using their open-chromatin regions and assigned lineage types based on functional annotation of these regions, including proximity to lineage genes and exposed lineage motifs in a cell. Most clones carry multipotent capacity, producing multiple downstream lineages. This capacity was further measured using the entropy of a clone's lineage fate, and is found to be consistent across clones.

Together, this dissertation provides improved understanding of biotherapeutically relevant cell types.

CHAPTER 2

2.1 Abstract

Chinese hamster ovary (CHO) cells are widely used for producing biopharmaceuticals, and engineering gene expression in CHO is key to improving drug quality and affordability. However, engineering gene expression or activating silent genes requires accurate annotation of the underlying regulatory elements and transcription start sites (TSSs). Unfortunately, most TSSs in the published Chinese hamster genome sequence were computationally predicted and are frequently inaccurate. Here, we use nascent transcription start site sequencing methods to revise TSS annotations for 15,308 Chinese hamster genes and 3,034 non-coding RNAs based on experimental data from CHO-K1 cells and 10 hamster tissues. We further capture tens of thousands of putative transcribed enhancer regions with this method. Our revised TSSs improves upon the RefSeq annotation by revealing core sequence features of gene regulation such as the TATA box and the Initiator and, as exemplified by targeting the glycosyltransferase gene Mgat3, facilitate activating silent genes by CRISPRa. Together, we envision our revised annotation and data will provide a rich resource for the CHO community, improve genome engineering efforts and aid comparative and evolutionary studies.

2.2 Introduction

Chinese hamster ovary (CHO) cells are the predominant mammalian system for large-scale production of clinical therapeutic proteins [1]. They are valued for their high growth rate [24], ease of genetic manipulation and ability to properly fold, assemble and produce complex post-translationally modified proteins that are not immunogenic in humans [25]. As of 2018, 84% of FDA approved monoclonal antibodies were produced in CHO cells [1] and by sales in 2020, 5 out of the top 10 drugs are CHO-derived recombinant proteins [57]. Optimizing CHO cells to increase production quantity and quality has been a priority for efforts to reduce the costs of biopharmaceuticals. Over the past few decades, these optimization efforts have progressed from engineering the media and bioreactors to transgene codon sequence and more recently, cell engineering and synthetic biology [26,37].

Genome sequencing efforts for CHO cells and the Chinese hamster [28,29] have been fundamental for studying and engineering CHO cells. In particular, they enabled systematic identification of genes associated with improved cell performance and product quality [32,33,35,36,58–60]. Furthermore, the sequences enabled the implementation of CHO cell engineering using tools including transcription activator-like effector nucleases (TALENs, [61], RNA-directed DNA methylation (RdDM) [62], CRISPR-Cas9 [38]) and others for genetic screens and the targeted inhibition or activation of genes [37,63]. However, the Chinese hamster genome annotation remains far from complete, especially for the approximately 50% of genes that are silenced in CHO cells, including many needed for producing more human-like proteins [34]. CRISPR activation (CRISPRa [38] and other genetic engineering methods could be instrumental to improve therapeutic protein production. However, to engineer gene expression, knowledge of the underlying regulatory elements is critical.

Recruitment of the RNA Polymerase II pre-initiation complex (RNAPII) by CRISPRa or blocking of the RNAPII by CRISPR inhibition (CRISPRi) and promoter editing [39,64] require knowledge of the polymerase's native transcription start site (TSS). Unfortunately, the vast majority of TSSs in the Chinese hamster RefSeq annotation were predicted computationally (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/) and may not correspond to the actual start sites *in vivo*. While previous work annotated 6,547 TSSs using steady-state 5′RNA ends by cap analysis gene expression (CAGE) in CHO cells [65], the data and annotation are not publicly available. Consequently, current inaccuracies in the annotation of the Chinese hamster genome and its TSSs present a major hurdle for targeted engineering of gene expression in CHO cells.

To remedy this issue, we generated multiple complementary experimental data types to accurately capture nascent transcription start sites (TSSs) at single nucleotide resolution [5′GRO-seq [54], csRNA-seq [55], GRO-seq [53]], expressed genes (ribosomal RNA-depleted RNA-seq), small RNAs [sRNA-seq [55]] and open chromatin (ATAC-seq [56]). To more comprehensively define regulatory elements in CHO cells, including for silenced genes, we interrogated not only CHO K1 cells but also 10 tissues and bone marrow derived macrophages (BMDMs) from Chinese hamsters of the original colony where CHO cells were derived [23]. Through this work, we developed a comprehensive compendium of Chinese hamster gene expression, including genes, enhancers, unstable divergent transcripts, diverse non-coding RNAs and their respective TSSs. Given their importance in deploying CRISPRa, we further analyzed the TSSs of

protein-coding genes. These data enabled us to accurately annotate the TSS or TSSs of 15 308 protein-coding genes and 3,034 non-coding RNAs. Notably, for 13,037 (85% of observed genes) genes, all detected TSSs were revised by >10 base pairs (bp) from the nearest NCBI RefSeq TSS, and 2607 (17%) by >150 bp. To demonstrate the accuracy and functionality of our revised TSS annotation we activated the dormant *Mgat3* (β-1,4-mannosyl-glycoprotein 4-β-N-acetylglucosaminyltransferase) in CHO cells by CRISPRa [3] using a novel identified TSS. In addition to accurate TSSs, the data generated provide insights into the DNA motifs and transcriptional regulatory pathways underlying tissue specificity in hamsters. Together, we envision our data and revised TSS annotation for the Chinese hamster will provide a rich resource for the CHO community, facilitate integrating the Chinese hamster into comparative studies, and improve engineering and manipulation for optimizing the production of therapeutic recombinant proteins in CHO cells.

2.3 Materials and Methods

2.3.1 Sample preparation

Female Chinese hamsters (*Cricetulus griseus*) were generously provided by George Yerganian (Cytogen Research and Development, Inc.) and housed at the University of California San Diego animal facility on a 12h/12h light/dark cycle with free access to normal chow food and water. All animal procedures were approved by the University of California San Diego Institutional Animal Care and Use Committee in accordance with University of California San Diego research guidelines for the care and use of laboratory animals. None of the used hamsters were subject to any previous procedures and all were used naively, without any previous exposure to drugs. Euthanized hamsters were quickly chilled in a wet ice/ethanol mixture (~50/50), organs were isolated, placed into Trizol LS, flash frozen in liquid nitrogen and stored at -80°C for later use. CHO-K1 cells were grown in F-K12 medium (GIBCO-Invitrogen, Carlsbad, CA, USA) at 37°C with 5% CO_2.

2.3.2 Bone marrow-derived macrophage (BMDM) culture

Hamster bone marrow-derived macrophages (BMDMs) were generated as detailed previously in[66]. Femur, tibia and iliac bones were flushed with DMEM high glucose (Corning), red blood cells were lysed, and cells cultured in DMEM high glucose (50%), 30% L929-cell conditioned laboratory-made

media (as source of macrophage colony-stimulating factor (M-CSF)), 20% FBS (Omega Scientific), 100 U/ml penicillin/streptomycin+L-glutamine (Gibco) and 2.5 μg/ml Amphotericin B (HyClone). After 4 days of differentiation, 16.7 ng/ml mouse M-CSF (Shenandoah Biotechnology) was added. After an additional 2 days of culture, non-adherent cells were washed off with room temperature DMEM to obtain a homogeneous population of adherent macrophages which were seeded for experimentation in Nunc Cell Culture dishes (Thermo Scientific) overnight in DMEM containing 10% FBS, 100 U/ml penicillin/streptomycin+L-glutamine, 2.5 μg/ml Amphotericin B and 16.7 ng/ml M-CSF. For Kdo2-Lipid A (KLA) activation, macrophages were treated with 10 ng/ml KLA (Avanti Polar Lipids) for 1 hour.

2.3.3 RNA-seq

RNA was extracted from organs that were homogenized in Trizol LS using an Omni Tissue homogenizer. After incubation at RT for 5 min, samples were spun at 21,000 *g* for 3 min, supernatant transferred to a new tube and RNA extracted following manufacturer's instructions. Strand-specific total RNA-seq libraries from ribosomal RNA-depleted RNA were prepared using the TruSeq Stranded Total RNA Library kit (Illumina) according to the manufacturer-supplied protocol. Libraries were sequenced 100 bp paired-end to a depth of 29.1–48.4 million reads on an Illumina HiSeq2500 instrument.

2.3.4 csRNA-seq protocol

Capped small RNA-sequencing was performed identically as described in [55]. Briefly, total RNA was size selected on 15% acrylamide, 7 M UREA and 1× TBE gel (Invitrogen EC6885BOX), eluted and precipitated over night at -80°C. Given that the RIN of the tissue RNA was often as low as 2, essential input libraries were generated to facilitate accurate peak calling [55]. csRNA libraries were twice cap selected prior to decapping, adapter ligation and sequencing. Input libraries were decapped prior to adapter ligation and sequencing to represent the whole repertoire of small RNAs. Samples were quantified by Qbit (Invitrogen) and sequenced using the Illumina NextSeq 500 platform using 75 cycles single end.

2.3.4 Global run-on nuclear sequencing protocol

Nuclei from hamster tissues were isolated as described in [67]. Hamster BMDM and CHO nuclei were isolated using hypotonic lysis [10 mM Tris-HCl pH 7.5, 2 mM $MgCl_2$, 3 mM $CaCl_2$] with 0.1% and 0.5% IGEPAL, respectively. Nuclei were flash frozen and later 0.5–1 × 106 nuclei in 200 μl GRO-freezing buffer [50 mM Tris-HCl pH 7.8, 5 mM $MgCl_2$, 40% Glycerol] were used in reactions with

3x NRO buffer [15 mM Tris-Cl pH 8.0, 7.5 mM $MgCl_2$, 1.5 mM DTT, 450 mM KCl, 0.3 U/µl of SUPERase In, 1.5% Sarkosyl, 366 µM ATP, GTP (Roche) and Br-UTP (Sigma Aldrich) and 1.2 µM CTP (Roche, to limit run-on length to ~40 nt)] as described in [68]. Run-on reactions were stopped, purified and GRO-seq and 5′GRO-seq libraries generated exactly as described in [66]. BrU enrichment was performed using a BrdU Antibody (Sigma B8434-200 µl Mouse monoclonal BU-33) coupled to Protein G (Dynal 1004D) beads. For each sample, 3 × 20 µl of Protein G beads were washed twice in DPBS+0.05% Tween 20 (DPBS+T) and then the antibody coupled in a total volume of 1 ml DPBS+T under gentle rotation. About 1 µl of antibody was used per 8 µl of beads. Samples were amplified for 14 cycles, size selected for 160–250 bp and sequenced on an Illumina NextSeq 500 using 75 cycles single end.

2.3.5 Assay for transposase-accessible chromatin sequencing (ATAC-seq) protocol

Approximately 150 k nuclei in 22.5 µl GRO freezing buffer (isolated as described for GRO-seq above) were mixed with 25 µl 2× DMF buffer [66mM Tris-acetate (pH = 7.8), 132 K-Acetate, 20 mM Mg-Acetate, 32% DMF] and tagmented using 2.5 µl DNA Tn5 (Nextera DNA Library Preparation Kit, Illumina) added. The mixture was incubated at 37°C for 30 min and subsequently purified using the Zymogen ChIP DNA purification kit (D5205) as described by the manufacturer. DNA was amplified using the Nextera Primer Ad1 and unique Ad2.n barcoding primers using NEBNext High-Fidelity 2× PCR MM for 8 cycles. PCR reactions were purified using 1.5 volumes of SpeedBeads in 2.5 M NaCl, 20% PEG8000, size selected for 140–240 bp fragments and sequenced using the Illumina NextSeq 500 platform using 75 cycles single end. This size range was selected to enrich for nucleosome-free regions.

2.3.6 CRISPRa

CRISPRa was carried out as previously described in [37]. Briefly, guide RNAs (gRNAs) were designed in a region proximal to our new revised TSS for *Mgat3* (NCBI GeneID: 100689076) and prioritized based on off-targets/proximity to the TSS. Target sequences and gRNA oligos are listed in **Tables S2.2** and **S2.3**, respectively. gRNAs were transfected along with a dCas9 VPR fusion plasmid [VPR-dCas9 (addgene #134601)] into mutant CHO-S cells carrying knockouts of *Mgat4a*,*4b* and *5*, *St3gal3,4* and *6*, *B3gnt2*, *Sppl3* and *Fut8* in biological triplicates. Non-targeting gRNAs were transfected with (NT-gRNA) and without VPR-dCas9 (NT-Cas9) as controls. Two days after transfection, cells were harvested to assess activation via qRT-PCR (in technical triplicate) and N-glycan analysis. Transcript

levels were normalized to the mean of *Hprt* and *Gnb1* and relative expression levels were calculated using the 2–ΔΔCt method [69].

2.3.7 Glycan quantification

N-Glycans were fluorescently labeled and quantified via LC-MS as described previously in [37]. Briefly, the supernatant was concentrated using Amicon® Ultra-4 Centrifugal Filter Units. Secretome proteins were fluorescently labeled with GlycoWorks RapiFluor-MS N-Glycan Kit (Waters, Milford, MA). N-linked glycan analysis was performed by LC-MS using a Waters Acquity Glycan BEH Amide 130 Å, 2.1 mm × 150 mm, 1.7 µm column (Waters, Milford) with a Thermo Ultimate 3000 HPLC with the fluorescence detector coupled online to a Thermo Velos Pro Iontrap MS (run in positive mode) and a separation gradient of 30–43% buffer. The amount of N-glycan was measured by integrating the areas under the normalized fluorescence spectrum peaks with Thermo Xcalibur software (Thermo Fisher Scientific) giving normalized, relative glycan quantities.

2.3.8 RNA-seq processing

Sequence data for all RNA-seq (ribosomal-depleted RNA-seq, csRNA-seq, 5′GRO-seq, sRNA-seq, GRO-seq), data were quality controlled using FastQC (v0.11.6. Babraham Institute, 2010), and cutadapt v1.16 [70] was used to trim adapter sequences and low quality bases from the reads. Reads were aligned to the Chinese hamster genome assembly PICR and annotation GCF_003668045.1, part of the NCBI Annotation Release 103. Sequence alignment was accomplished using the STAR v2.5.3a aligner [71] with default parameters. Reads mapped to multiple locations were removed from analysis.

2.3.9 ATAC-seq processing

Sequence data for ATAC-seq was processed using the ENCODE ATAC-seq pipeline (https://github.com/kundajelab/atac_dnase_pipelines). The reads were trimmed using cutadapt v1.9.1. Reads were aligned using Bowtie2 v2.2.4 [72] to the same Chinese hamster genome. Peaks were called using MACS2 v2.1.0 [73] with a *P*-value of 0.01 and replicates were merged using irreproducible discovery rate (IDR) [74] of 0.1. The fold-change value is the number of normalized counts over the local background, taken as a 10,000 bp surrounding region.

2.3.10 Detecting TSSs

To call TSS peaks, the Homer [75] version 4.10 TSS pipeline was used with the command 'findPeaks -style tss' (http://homer.ucsd.edu/homer/ngs/tss/index.html). Briefly, fragment lengths are set to 1, and 150 bp regions significantly enriched with fragments above the local genomic background region, as well as 2-fold above the input data (GRO-seq and sRNA-seq). FDR correction of 0.01 across peaks in each sample was used. The samples are then merged together into our initial, putative experimental TSSs. Additionally, the total RNA-seq was used to call TSSs as stable if reads are identified between -100 and +500 bp upstream of the TSS.

Sample peaks were merged using the mergePeaks command in Homer. If samples have overlapping peaks, they are combined into one, where the start position is the minimum start position and the end is maximum end position. When merging the replicate peak expression in the same biological sample, the average counts per million (CPM) was used.

2.3.11 Revised promoter annotation

To annotate protein-coding TSSs, a distance threshold from the original annotations was enforced. Ultimately, we retained TSSs that are within -1000 bp and +1000 bp from the initial reported TSS. Additionally, TSSs found in introns, coding sequences, and opposite strand TSSs (divergent transcripts) found in the TSS region were removed (**Figure S2.1)**. There were two annotations used in this study to provide gene promoter landmarks, one from the NCBI RefSeq Annotation 103 release using the PICR genome, and the other with both NCBI's annotation and a proteogenomics annotation (doi:10.7303/syn17037372) that used RNA-seq, proteomics and Ribo-seq to refine gene mappings [76].

When samples are merged together, the TSSs that are merged may be offset by a few bp. Our revised annotation TSS location is assigned as the CHO TSS location if there is one present, or the location of the TSS in the sample which had the highest expression in CPM. An additional annotation integrating promoter TSSs found in either annotation is also reported.

The annotation provided (Supplementary Data 2-3 in [77]) includes the chromosome, start position (0-based index similar to bed format), strand, position, corresponding gene name, corresponding transcript, comma-separated list of biosamples that express the TSS, and a confidence score signifying the TSS having 2 CPM in at least 2 5′GRO-seq and/or csRNA-seq experiment.

2.3.12 Distal TSSs

Distal TSSs (dTSSs), or intergenic TSSs, were defined as being >1000 kilobase pairs (kbp) away from an annotated gene (ncRNA and protein-coding).

2.3.13 RNA-seq/TSS-seq comparison

To compare RNA-seq to TSS-seq, we used 1558 CHO samples of different lines that were a combination of in-house and public samples (see **Table S2.4** for accession IDs). These were quantified and converted into transcript per kilobase gene per million mapped reads (TPM) using Salmon with default parameters[78].

2.3.14 Read histograms

For **Figure 2.2A-B**, Homer annotatePeaks.pl with the -hist command was used to construct the histogram with a bin size of 1 bp, and the CPM per TSS was calculated. We restrict the maximum number of tags to count per nucleotide to 3 to prevent high-expressing TSSs from saturating the signal.

2.3.15 Motif analysis

Motif analysis of the core promoter elements the Initiator element and the TATA-Box seen in **Figure 2.2** were done using FIMO of the MEME Suite 5.0.2 package with default parameters [79], scanned across a 150 bp window centered on the TSS. Position weight matrix scores of the motifs are summed across all TSSs and converted into a $\log_2$-likelihood ratio score for each motif with respect to each sequence position and then converts these scores to *P*-values, with a cutoff of 0.0001.

For motif analysis in **Figure 2.3**, the promoter regions were -300 bp to +100 bp downstream of each TSS using Homer command 'findMotifsGenome.pl' with parameters '-size -300,100 -len 6,8,10'. For each sample, protein-coding TSSs with log2 CPM of 2 standard deviations above the mean were taken as enriched promoters. The background chosen was randomly selected GC-controlled regions. The negative $\log_e$*P*-value of the top 3 enriched motifs from each sample are taken and the TFs were clustered based on their enrichment *P*-values.

2.3.16 Tissue-specific gene enrichment analysis (TSEA)

TSEA was done using the webserver http://genetics.wustl.edu/jdlab/tsea/. This performs enrichment analysis using Fisher's Exact test, and the Benjamini-Hochberg corrected log *P*-values were

used for **Figure 2.3D**. Unique genes for each sample were defined as only one sample having an observed promoter in that gene. Homologous genes to the human set in TSEA were taken using gene names.

2.3.17 GlycoGene database

Human glycosylation genes and their associated enzyme classes were taken from the 'Enzymatic Activity' section of the GlycoGene Database [80]. Homologous genes were taken as described above.

2.4 Results

Nascent 5′ RNA sequencing across hamster tissues enables accurate reannotation of RNA start sites at single nucleotide resolution

Algorithms predicting gene annotations rely on highly conserved features such as protein domains (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/). Consequently, although gene exon regions are commonly assigned correctly, the annotations of their TSSs, and each associated promoter, are often inaccurate, as these features evolve rapidly and can relocate to non-homologous regions [81]. To correctly annotate the TSS of protein genes and non-coding RNAs (e.g. pri-miRNAs, lncRNAs and snoRNAs), it is necessary to experimentally determine these features. We therefore captured [5′GRO-seq [54], csRNA-seq [55]], active transcription [GRO-seq [53]], expressed genes (ribosomal RNA-depleted RNA-seq), small RNAs [sRNA-seq [55]], and open chromatin [ATAC-seq [56]] in CHO-K1 cells as well as ten tissues and bone marrow derived macrophages (BMDMs) in female hamsters from Dr George Yerganian, representing the original colony from which CHO cells were derived in 1957 [82] (**Figure 2.1A**; **Figure S2.1A** and **Table S2.1**). Unlike RNA-seq, 5′GRO-seq and csRNA-seq provide accurate TSSs of stable transcripts such as mRNAs (**Figure 2.1B**) or ncRNAs but also unstable RNAs such as enhancers RNAs (**Figure 2.1C**) [83,84] at single nucleotide resolution. Even for highly expressed genes, such as the Eukaryotic Translation Elongation Factor 1 Alpha (Eef1a1), RNA-seq and related methods that capture the complete transcriptome have limited information about the exact location where genes start and often fall short in the detection of the TSSs for less abundant transcripts (**Figure 2.1**). Capturing the TSSs of nascent transcripts further helps to avoid potential false-positive 5′ends caused by RNA processing or recapping of cytosolic (steady-state) mRNAs [85].

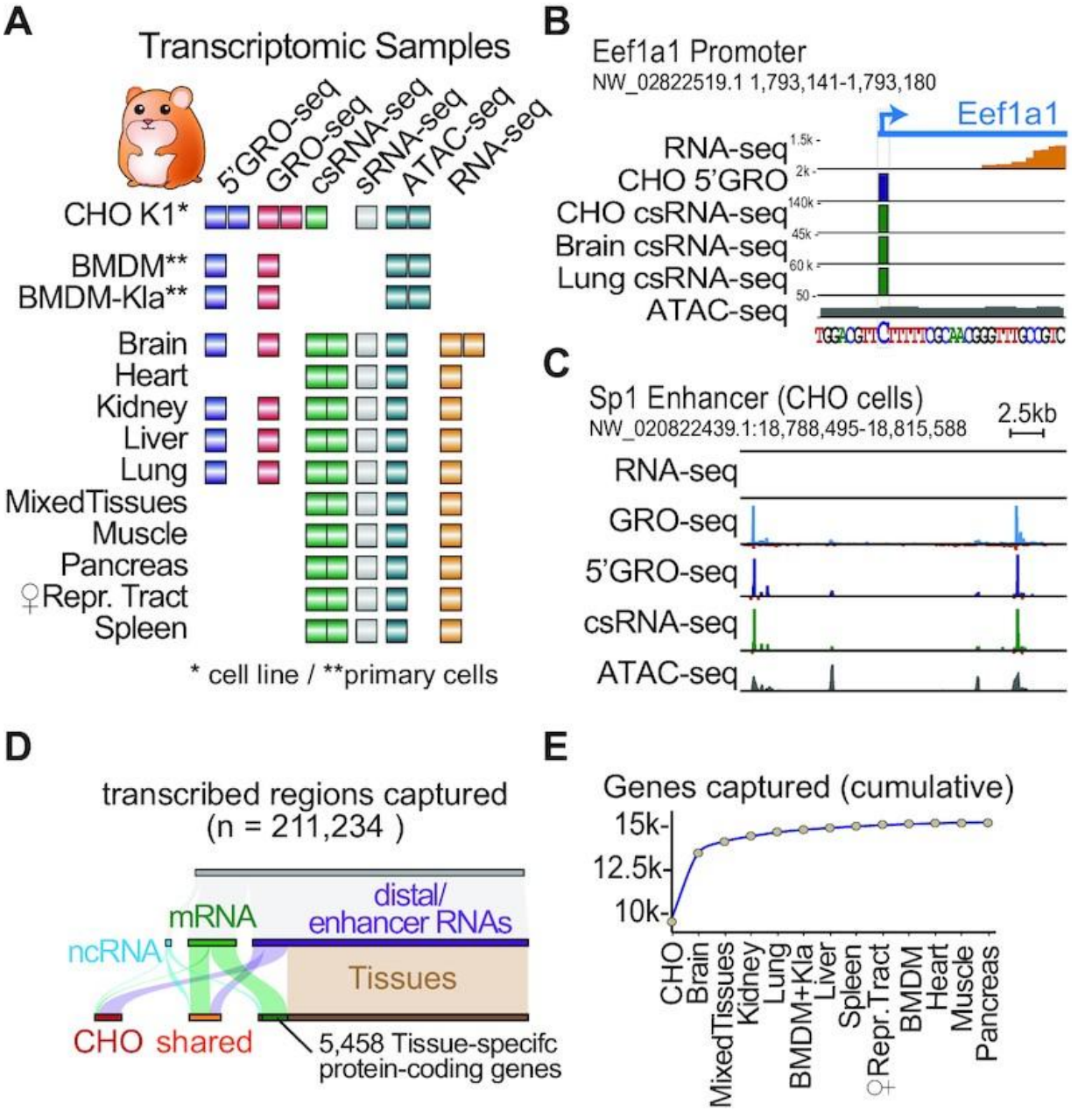


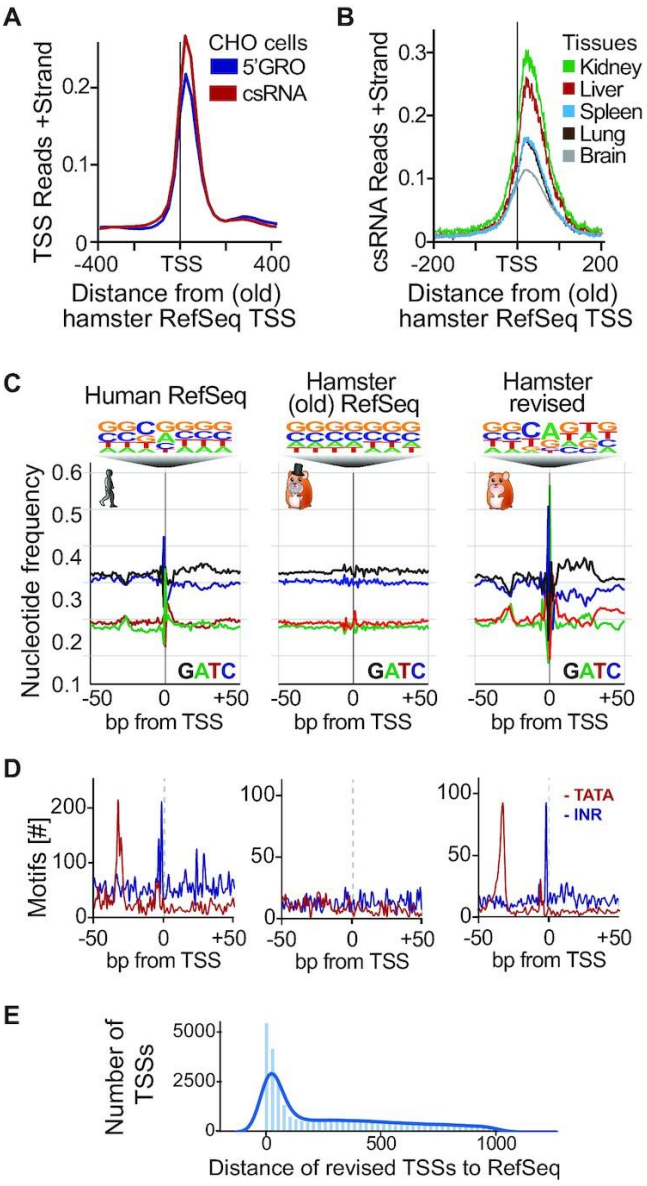
Figure 2.1 A Chinese hamster Transcriptome Atlas.
(**A**) Overview of datasets generated to identify transcription start sites. * Denote cell lines, ** denote primary cells. (**B** and **C**) IGV viewer of data. Units are in counts per million (CPM) (B) Example transcription start site at single-nucleotide resolution as defined by 5′GRO-seq and csRNA-seq (using GRO-seq and sRNA-seq as input, respectively) of the focused Eukaryotic Translation Elongation Factor 1 Alpha (Eef1A1) promoter in CHO cells and diverse tissues. Brain RNA-seq reads are shown in orange. (C) Example of unstable transcription start sites of enhancer RNAs that are poorly detected by conventional RNA-seq at the Sp1 'super enhancer' locus in CHO cells. Note: Raw IGV browser visualization data are provided in Figure S2.3. (**D**) Number of TSSs captured, grouped by TSS type and samples detected in (**E**). Cumulative plot across all samples of protein-coding genes with a TSS detected by csRNA-seq and/or 5′GRO-seq enrichment over GRO-seq and/or csRNA-seq. Sorted by taking CHO as the first sample, followed by hamster tissues.

As the primary goal of this study was the determination of confident TSSs, we employed two independent nascent TSS methods, csRNA-seq and 5′GRO-seq. However, while csRNA-seq accurately captures TSSs from total RNA, 5′GRO-seq requires several million purified nuclei, which was not feasible for some tissues. Using csRNA-seq allowed us to expand our analysis across more diverse hamster tissues. In addition, we employed GRO-seq and small RNA-seq (sRNA-seq) data as a background control (also known as input) to boost the confidence of TSS calls by 5′GRO-seq and csRNA-seq, respectively (**Figure S2.1B**). Next, we integrated ATAC-seq to filter TSSs that mapped outside of open chromatin regions (**Figure S2.4A**). Finally, as nascent TSS methods detect both stable and unstable TSSs, we used conventional ribosomal RNA-depleted RNA-seq to assign TSSs as stable if RNA-seq coverage was detected between -100 and +500 base pairs from the TSS [55]. Integrating these multiple independent data sets also enabled an intrinsic quality control metric and highlighted the confidence of captured TSSs. For example, the correlation among 5′GRO-seq replicates and between 5′GRO-seq and csRNA-seq were highly consistent in their position and expression strength (Pearson correlation of $r = 0.96$ and $r = 0.88$, respectively, **Figure S2.2**). A list of the 71 datasets generated in this study is provided in **Table S2.1**. These data capture over 210 000 transcribed regions at single-nucleotide resolution (**Figure 2.1D**, **Figure S2.3A–C**, Supplementary Data 1 in [77]) and provide a comprehensive view of the hamster transcriptome. The majority of these regions ($n = 154\,736$) mark putative distal regulatory elements (sometimes referred to as 'enhancers' for simplicity [84]) and unstable divergent transcripts, two common hallmarks of mammalian gene expression [53,86], as well as 3560 non-coding RNAs (**Figure 2.1D**). Importantly for protein engineering, we focus on the detected TSSs that mark the promoter or promoters of a cumulative 15 308 RefSeq protein-coding genes captured and their revised promoter TSSs (**Figure 2.1E**, Supplementary Data 2 in [77]). Functional gene groups that were less covered by our data include those associated with olfaction, taste, the male sex organ (testis), development and the adaptive immune system (**Figure S2.3D-E**). Together, our experimental data provide accurate TSSs for 72% of annotated hamster protein-coding genes and 3034 non-coding RNAs. We additionally leverage promoter TSSs predicted by a recent proteogenomics annotation [76] to detect and revise additional promoters (Supplementary Data 3 in [77]).

Realignment of NCBI Chinese hamster RefSeq TSSs exposes key features of transcription

Genome annotations are an essential part of many sequencing and bioinformatic analyses and TSSs provide the foundation for accurate annotation of 5′ ends. We therefore tested the rigor of our experimentally determined protein-coding TSSs and our revised annotation using a number of independent measures. First, we evaluated the relationship of our revised TSSs to the Chinese hamster RefSeq TSSs (GCF_003668045.1). TSSs called by either 5′GRO-seq or csRNA-seq displayed similar distributions (**Figure 2.2A)**. However, both experimentally determined TSSs displayed a clear offset from the RefSeq annotation. A comparable offset was also observed for protein-coding TSSs measured in diverse tissues (**Figure 2.2B**). To further explore these differences we next plotted the proximate DNA nucleotide frequency distributions for both RefSeq and our revised TSS. Basal transcription factors often bind core promoter elements to recruit and position the RNAP II transcription complex which preferentially initiates on purines [87–89]. These nucleotide preferences are clearly visible when analyzing the human RefSeq (GRCh38) annotation and in our revised hamster annotation, but not in the current Chinese hamster RefSeq annotation (**Figure 2.2C**). In addition to the increased information content in the TSS-proximate nucleotide frequencies, the TATA box and Initiator (Inr) core promoter elements [90–92], were found at the expected -30 and +1 bp positions respectively in the human RefSeq and in our revised hamster annotations, but not the old RefSeq annotation (**Figure 2.2D**).

Figure 2.2 An experimental realignment of TSS annotation for the Chinese hamster uncovers expected genomic elements
A comparison of our TSSs to Chinese hamster RefSeq annotation GCF_003668045.1 (**A** and**B**) Average normalized CPM around protein-coding reference TSSs. (A) Comparison of experimentally defined TSSs from CHO cells by 5′GRO-seq and csRNA-seq relative to the RefSeq annotation. (B) Comparison of experimentally defined TSSs from representative tissues relative to the RefSeq annotation. (**C**) Nucleotide frequency plots of TSSs and their relative information content in Human RefSeq, Chinese hamster RefSeq, and our revised Chinese hamster annotation. (**D**) Frequency of positional core promoter elements: the TATA box and the Initiator that are commonly found at -30 and +1, relative to the TSS. (**E**) Frequency of distance between revised TSSs observed and the nearest RefSeq TSS. (**F**) Summary of total protein-coding and non-distal ncRNA TSSs observed and their distances to RefSeq TSSs.



F

Promoters shifted from NCBI RefSeq annotation	Observed	Median distance of revision	Revised >10 bp	Revised >150 bp
Experimental protein-coding promoters	30760	158 bp	28419	15557
RefSeq protein-coding promoters (n=35679)	20627	54 bp	18286	5552
RefSeq protein-coding coding genes (n=21488)	15308	40 bp	13037	2607
Experimental ncRNA promoters	3560	136 bp	3073	1717
RefSeq ncRNA promoters (n=8144)	3034	83 bp	2547	1198

Next, we utilized published epigenetic chromatin states from CHO samples [93] which revealed a striking enrichment of our revised TSSs in the 'active promoter' category, highlighting that our experimental CHO dataset is consistent with prior published CHO chromatin states (**Figure S2.4B**). On the contrary, both our revised promoter TSSs and the NCBI Refseq TSSs fell into more quiescent states, suggesting these regions are near silenced CHO genes. Lastly, we integrated 1558 CHO RNA-seq samples [34,94–96] to assess potential false positive and false negative TSSs in our revised annotation. Genes where we failed to experimentally detect a TSS showed little to no expression across the CHO RNA-seq datasets while those where we captured a CHO TSS were consistently expressed (**Figure S2.5**), suggesting a low false discovery rate.

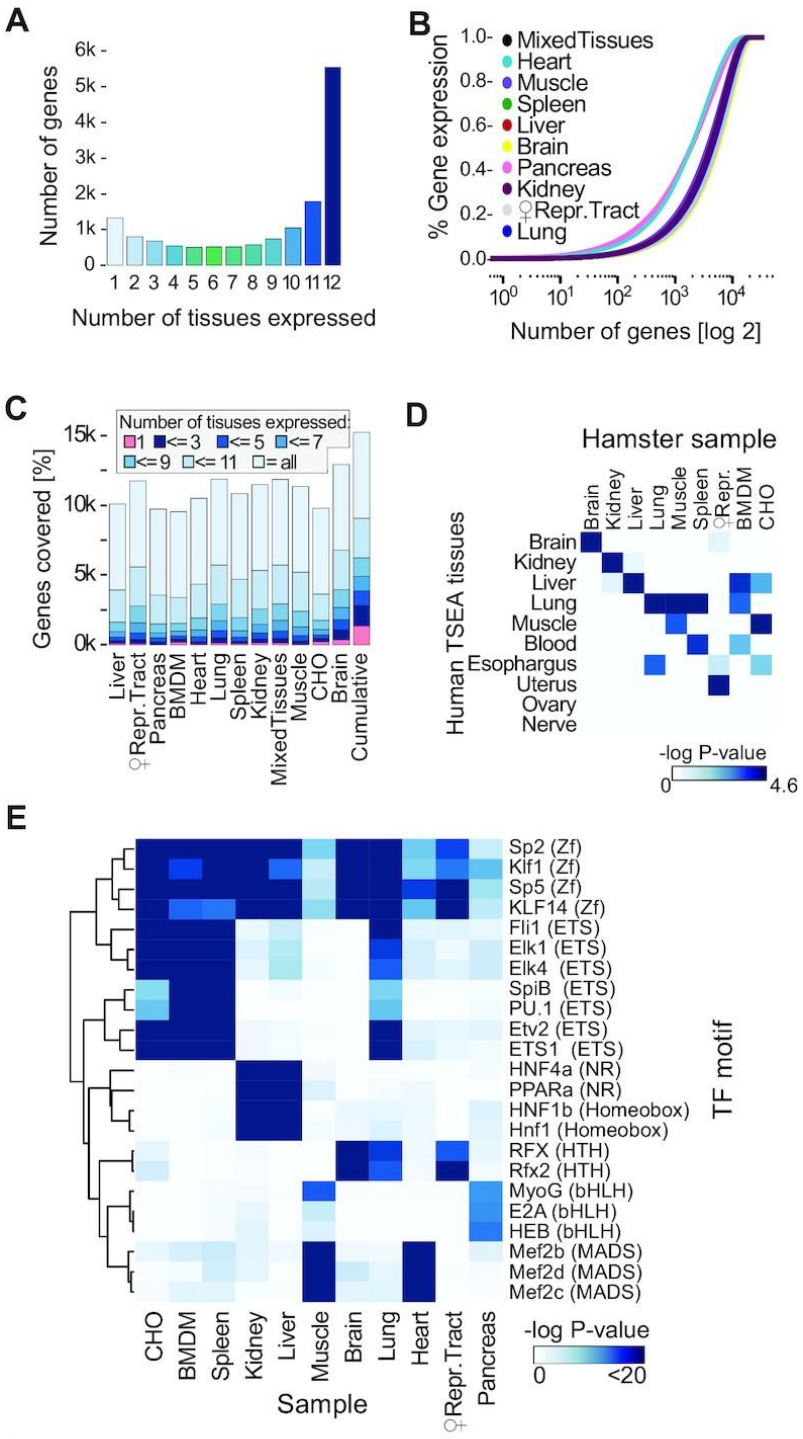
Overall, the distance of protein-coding TSSs to the nearest RefSeq TSS varied widely (**Figure 2.2E**), with a median distance of 158 bp (**Figure 2.2F**). Notably, RefSeq promoters (that represent different transcript isoforms) with a detectable TSS were revised by a median of 54 bp, and 5552 of the promoter TSSs were revised by >150 bp. When we look further at the smallest revision across the promoter TSS' of each gene, the median distance is 40 base pairs. Importantly, 13,037 were revised >10 bp, and 2607 >150 bp (**Figure 2.2E-F**). ncRNAs TSSs also varied, and had a median distance of 83 bp, and 1,198 revised by >150 bp (**Figure 2.2F**). In summary, these observations provide an independent validation for our revised annotation and stress the importance of experimental TSS data for accurate genome annotations.

Tissue-specific TSS and gene expression patterns in the Chinese hamster

Capturing the protein-coding TSSs across tissues and cell lines revealed that about 1/3 of annotated genes were ubiquitously expressed, while only a comparatively small number of genes were tissue-specific (**Figure 2.3A**). Using ribosomal-depleted RNA-seq to measure the steady-state transcriptome highlights the variation of gene expression across tissues (**Figure 2.3B**). The number of genes with detected mRNA were 9596 and 9850 in bone marrow-derived macrophages (BMDMs, pooled rested and stimulated conditions) and CHO cells, respectively. Meanwhile, the number spanned from 9782 genes with measured mRNA in the pancreas to 13,007 in the brain (**Figure 2.3B**). The number of tissue-specific genes is related to the tissues' degree of specialization and the number of different cell types found within the tissue, but also affected by high abundance transcripts that can hinder detection of less abundant ones [97]. In the pancreas, for example, much of transcription is directed towards expressing

secretory enzymes such as chymotrypsinogen or carboxypeptidase [98], while in the brain, a higher diversity of transcripts are expressed [99,100].

Figure 2.3 Composition of diverse tissue-specific Chinese hamster transcriptomes
(**A**) Experimentally detected genes and the number of tissues wherein they were confidently expressed, as defined by csRNA-seq and 5′GRO-seq. (**B**) Cumulative plot of the distribution of transcript abundance as defined by RNA-seq in various tissues. The transcriptome of highly specialized tissues such as the heart or the pancreas is more dominated by the high expression of a small set of specific RNAs than those of complex tissues such as the brain. (**C**) Comparison of gene expression distributions across tissues as defined by csRNA-seq and 5′GRO-seq. (**D**) Tissue-specific gene enrichment analysis (TSEA) comparing the gene expression patterns of our samples as defined by csRNA-seq and 5′GRO-seq to orthologous human pre-defined tissue-specific genes. -loge (*P*-value) values are shown. (**E** and**F**) Motif analysis with Homer. Significance of hypergeometric enrichment of the motifs shown as -loge (*P*-value). **(E)** Transcription factor motifs (top 3 per sample) enriched in TSSs for each tissue highlight conservation and factors involved in maintaining tissue-specific expression patterns. **(F)** Transcription factor motifs enriched in all protein-coding gene-associated TSSs in the revised TSS annotation.


A
Number of genes
6k
5k
4k
3k
2k
1k
0
1 2 3 4 5 6 7 8 9 10 11 12
Number of tissues expressed
B
% Gene expression
1.0
0.8
0.6
0.4
0.2
0.0
MixedTissues
Heart
Muscle
Spleen
Liver
Brain
Pancreas
Kidney
♀Repr.Tract
Lung
10^0 10^1 10^2 10^3 10^4
Number of genes [log 2]
C
Genes covered [%]
15k
10k
5k
0k
Number of tisuses expressed:
1 <=3 <=5 <=7
<=9 <=11 =all
Liver
♀Repr.Tract
Pancreas
BMDM
Heart
Lung
Spleen
Kidney
MixedTissues
Muscle
CHO
Brain
Cumulative
D
Hamster sample
Brain
Kidney
Liver
Lung
Muscle
Spleen
♀Repr.
BMDM
CHO
Human TSEA tissues
Brain
Kidney
Liver
Lung
Muscle
Blood
Esophargus
Uterus
Ovary
Nerve
-log P-value
0
4.6
E
Sp2 (Zf)
Klf1 (Zf)
Sp5 (Zf)
KLF14 (Zf)
Fli1 (ETS)
Elk1 (ETS)
Elk4 (ETS)
SpiB (ETS)
PU.1 (ETS)
Etv2 (ETS)
ETS1 (ETS)
HNF4a (NR)
PPARa (NR)
HNF1b (Homeobox)
Hnf1 (Homeobox)
RFX (HTH)
Rfx2 (HTH)
MyoG (bHLH)
E2A (bHLH)
HEB (bHLH)
Mef2b (MADS)
Mef2d (MADS)
Mef2c (MADS)
TF motif
CHO
BMDM
Spleen
Kidney
Liver
Muscle
Brain
Lung
Heart
♀Repr.Tract
Pancreas
Sample
-log P-value
0
<20
F
Motif
Match
-log P-value
Sp2/KLF motif
1.182e+03
ETS1/ ETS motif
7.654e+02
NFY / CCAATmotif
4.290e+02

Of the genes for which TSSs were confidently detected, 40% were expressed in all 12 tissues or cell types and another 19% were found in 11 samples. Approximately 8% of captured genes were unique to a tissue or cell type which increased to 18% and 25% for genes expressed in <3 or <5 samples, respectively (**Figure 2.3C**). The genes underlying these tissue-specific gene expression signatures in hamsters at large resembled those of analogous human tissues, as determined by Tissue-Specific Gene Enrichment Analysis (TSEA [101], **Figure 2.3D**). Capturing the TSS of a given gene across multiple tissues provided an additional control, in addition to our use of two distinct methods for TSS detection. Nevertheless, for many conserved genes (in which at least one TSS was detected in each sample), the respective promoters detected differed among tissues (**Figure S2.6D**). Additionally, within conserved promoters, there were small but slight shifts of the called tissue TSS from the revised TSS annotation (**Figure S2.6E**), together showing a remarkable diversity in 5′ ends. This finding highlights regulatory plasticity as a critical factor to maintain gene expression in distinct cell types.

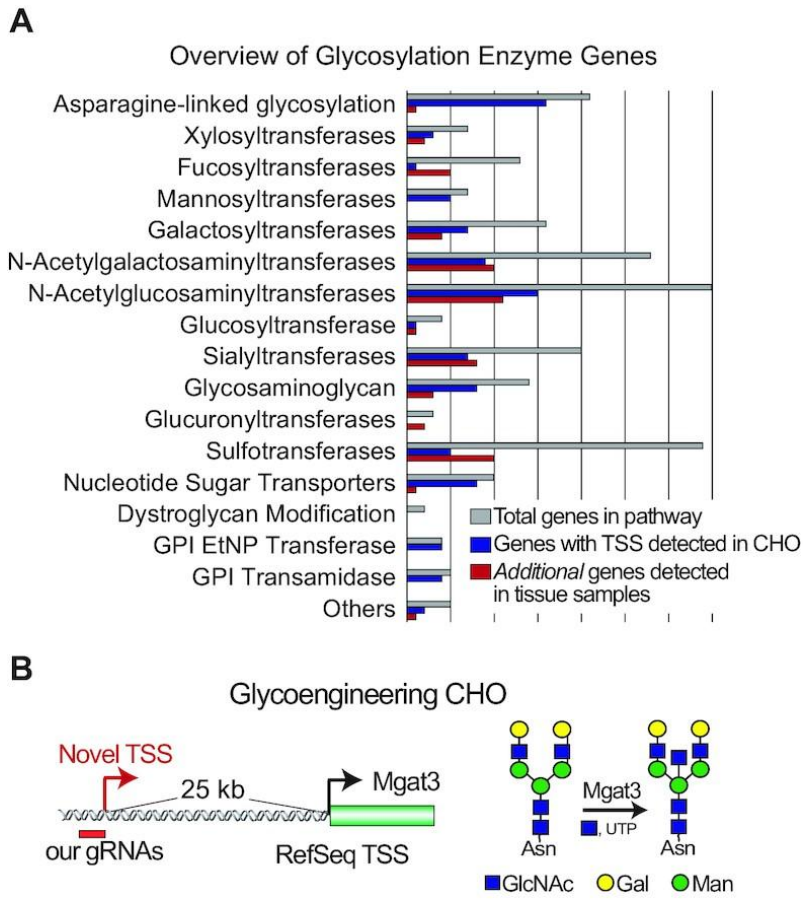
To gain insights into the underlying regulatory program, we next probed the promoters of tissue-specific genes' promoters for differentially enriched transcription factor binding motifs. To do this, we used Homer to scan for known motifs 300 bp upstream to 100 bp downstream of TSSs unique to a sample (see Materials and Methods section). The top 3 enriched motifs from each sample and their enrichment values are shown in **Figure 2.3E**. We found key regulators or lineage determining transcription factors with preferential expression and binding sites for each tissue such as RFX factors for the brain [102], HNF1 [103] and PPARα factors for the kidney and liver [104,105] or the MADS-box transcription factors Mef2b,c and d for the heart and muscle (**Figure 2.3E**) [106,107]. Closely related tissues, such as muscle and heart or liver and kidney, displayed a combination of shared and unique factors, which also became apparent for other tissues when more motifs were integrated into the analysis. This observation is in line with the hypothesis that tissue-specific regulatory pathways arise by tinkering with existing pathways, rather than complete innovation [108,109] of regulatory elements needed. On the other hand, ubiquitously expressed genes were enriched for the binding motifs of strong, ubiquitous activators such as SP2/KLF family members [110], ETS factors or NFY (**Figure 2.3F**). Together, these findings argue that a comparatively large fraction of genes, including ubiquitous transcription factors, ensure the cell's vital core programs, while a smaller number of genes effectively facilitates specialization. Moreover, our

identification of tissue-specific genes and transcription factors enriched in their promoters are consistent with other mammals, further validating our revised TSSs and RNA-seq data.

Profiling diverse hamster tissues identifies TSSs for important, but silenced genes in CHO cells

While CHO cells are exceptional protein production hosts, many genes that could improve product quality or quantity lay dormant. Indeed, about 50% of genes, including many that contribute to important human post-translational modifications, are silent [34]. We detected TSSs for only 46% of all protein-coding genes in CHO cells (**Figure 2.1E**). Integrating our TSSs from ten tissues and macrophages [111] confidently defined TSSs from an additional 5458 protein-coding genes. In addition, we identified alternative promoters responsible for transcript isoforms for 55% of the RefSeq annotated protein-coding promoters (**Figure S2.6C**). Our revised TSS annotation provides multiple promoters per gene along with additional promoters uncharacterized in RefSeq (**Figure S2.6A-B**). This isoform annotation is important as it facilitates the tailored expression of protein isoforms that can exhibit differential activity or distinct functions [112,113]. This characterization of >15 k protein-coding genes and >20 k annotated promoters provide the necessary foundation for ongoing efforts to optimize drug production in CHO cells through engineered activation of dormant genes. Given that most protein therapeutics are glycosylated, and the glycans can impact drug safety, efficacy and half-life [114], we next specifically investigated glycosylation-related genes in the context of our updated annotation (**Figure 2.4A**). When examining CHO homologues of curated human glycosylation enzymes, we detected dozens of TSSs across diverse classes of glycosylation enzyme genes (**Figure 2.4A**). Together, these new annotations should open up new possibilities for engineering gene expression programs, such as glycosylation in CHO cells.

Figure 2.4 Experimentally measured TSSs facilitates genome engineering to humanize glycosylation
(**A**) List of human glycosylation enzyme classes detected in our samples as defined by 5′GRO-seq/csRNA-seq in the Chinese hamster. The number of genes expressed in CHO cells (blue) and additional genes for which experimental TSSs were discovered in our tissue samples (red) are shown. (**B**) Overview of the RefSeq TSS targeted by guide RNAs with CRISPRa to induce *Mgat3* expression in CHO cells. The *Mgat3*-encoded glycosyltransferase catalyzes the addition of bisecting N-acetylglucosamine on glycoproteins, but is silenced in CHO cells. (**C**) Quantitative RT-PCR measurement of *Mgat3* expression in CHO cells and upon activation by the three designed gRNAs using our new TSSs. As a control, the cells were transfected with NT-gRNA (gRNA-Ctrl) or NT-gRNA and VPR-dCas9 (Cas9-Ctrl). (**E**) Comparison of the levels of bisecting N-acetylglucosamine in secretome following CRISPRa. As a control, the cells were transfected with NT-gRNA (gRNA-Ctrl). (E) Overview: Experimental TSS facilitates efficient engineering of *Mgat3* in an upstream promoter.


A
Overview of Glycosylation Enzyme Genes
Asparagine-linked glycosylation
Xylosyltransferases
Fucosyltransferases
Mannosyltransferases
Galactosyltransferases
N-Acetylgalactosaminyltransferases
N-Acetylglucosaminyltransferases
Glucosyltransferase
Sialyltransferases
Glycosaminoglycan
Glucuronyltransferases
Sulfotransferases
Nucleotide Sugar Transporters
Dystroglycan Modification
GPI EtNP Transferase
GPI Transamidase
Others
Total genes in pathway
Genes with TSS detected in CHO
Additional genes detected in tissue samples
B
Glycoengineering CHO
Novel TSS
25 kb
Mgat3
our gRNAs
RefSeq TSS
Mgat3
UTP
Asn
Asn
GlcNAc
Gal
Man
C
relative Mgat3 mRNA expression
n = 3
Fold Change
120
80
40
0
CHO
Cas9-Ctrl
gRNA1
gRNA2
gRNA3
gRNA 1-3
D
n = 3
% Total Glycan
1.2
0.8
0.4
0
ND
CHO
New TSS
F
CHO cells
Gene OFF
Experimental TSS
Gene ON
Cas9
VP 64
- high gene activation
- active protein
- isoform specificity

TSS detected in upstream promoter facilitates CRISPR activation of the dormant gene Mgat3 in CHO

To test the feasibility of genome engineering based on our revised annotations we aimed to activate *Mgat3* (**Figure 2.4B**), which is naturally dormant in CHO cells [115] using a novel identified alternative promoter stable TSS that is 25,481 bp upstream of the promoter previously used for *Mgat3* activation [37]. *Mgat3* is required for bisecting N-acetylglucosamines which play an important role in regulating complex glycosylation maturation and impact antibody effector function [116,117] and is hence well studied in both humans and CHO cells.

CRISPR/Cas9 enables rapid and cost-effective genome editing, gene inhibition (CRISPRi), and activation (CRISPRa) without altering the native DNA sequence [38,63,64]. However, the success of these and similar precise genome engineering approaches depends on accurate gene annotations [62,118]. Given that *Mgat3* has previously been targeted by CRISPRa [37], we used this experimental system to show that the gene can be activated by targeting an experimentally identified promoter, even when located >25 kb away from the RefSeq gene TSS. To activate *Mgat3* we designed three CRISPR guide RNAs (gRNAs) complementary to the DNA sequence near our alternative TSS (**Figure 2.4C**). CRISPRa resulted in a mean of 94-,109-, and 64-fold upregulation of *Mgat3* using the three different gRNAs individually (n = 3 samples each), and 73-fold for a mixture of the 3, as measured by qRT-PCR (**Figure 2.4D**). To test if the activation of *Mgat3* transcripts impacts glycan synthesis, we measured the relative abundance of glycans on the secretome. This analysis revealed that while undetectable in control cells, 1.08% of glycans were bisecting N-acetylglucosamines after *Mgat3* activation (**Figure 2.4E**). Together, these data show the use of our revised annotation for genome engineering. With our newly reported TSSs for 15,308 genes across >30 000 detected promoters, we anticipate further usage of these TSSs for cell line engineering.

2.5 Discussion

In this study we measured and analyzed the coding and non-coding RNA in the Chinese hamster genome using steady state and nascent RNA sequencing experiments for diverse hamster tissues and cell lines. Through this we were able to comprehensively map TSSs for >70% of annotated Chinese hamster

genes and non-coding RNAs, including many genes normally silenced in CHO cells. Importantly, these experimentally determined TSS enabled us to realign current RefSeq TSSs, which were predominantly computationally predicted and often inaccurate. Unlike the previous RefSeq TSS, our revised TSSs annotations display expected DNA nucleotide frequency features such as the Initiator motif or the TATA box in the core promoter. Furthermore, we demonstrated that accurate TSSs and knowledge of alternative promoters can be used to activate a silenced gene of interest using CRISPRa. Through this we present a resource to guide genome editing and genomic analysis of CHO cells.

Here, we captured 30 760 nascent protein-coding TSSs corresponding to 15 308 genes, along with 3560 ncRNAs (lncRNA, miRNA, snRNA, snoRNA and tRNA), and 176 914 distal peaks (enhancer RNAs etc.). This resource provides rich information for precise cell engineering. Furthermore, including diverse hamster tissues helps in efforts to fine tune existing CHO gene regulatory programs, as well as activate genes or pathways naturally encoded in the Chinese hamster genome but dormant in CHO cells. Our TSSs are a prerequisite for the design and testing of gRNAs and eventually, an effective gRNA library for the activation of diverse Chinese hamster genes by CRISPRa (**Figure 2.4F**). It can also complement existing data on epigenetic markers of CHO cells in efforts to find endogenous promoters that avoid silencing seen with common viral promoters or harness endogenous regulatory circuits involved in ER stress or cold shock [119].

Our transcriptomic datasets also provide a comprehensive resource for future research and discovery. In addition to our gene-centric atlas of Chinese hamster TSSs reported here, our data cover a plethora of transcriptomic features that remain to be explored including miRNAs, pri-miRNAs and well over a 100 k putative distal regulatory elements that are commonly referred to as enhancers (Supplementary Data 2 in [77]). Although beyond the focus of this manuscript, this extensive, transcript stability-independent resource of TSSs could also aid to improve our understanding of how gene expression is regulated in hamsters and how tissue-specific regulatory programs emerged. While a key advantage of CRISPRa is the ability to activate desired genes independent of tissue-specific transcription factors, future engineering efforts may be more tailored towards adjusting transcriptional programs, rather than one or a few specific genes. For example, our definition of transcription factors that were highly enriched in the promoters of tissue-specific genes provides a first step to advance our understanding of which and why specific genes or pathways are silent in CHO cells. Improved knowledge of how gene

regulatory networks function in hamsters may ultimately allow us to predict how activation of one gene impacts the hamster regulome and to eventually fine-tune desired regulatory programs, rather than individual genes [120]. Going beyond capturing TSSs, our data also contain maps of open chromatin, as defined by ATAC-seq, nascent transcription, as defined by GRO-seq, and mature RNAs, as defined by ribosomal RNA-depleted RNA-seq for CHO cells, hamster macrophages and diverse hamster tissues that were primarily used in this study as a critical input for the identification of high-confidence TSSs. Our data thus also provide a rich resource for future studies and enable the integration of the Chinese hamster into comparative or evolutionary studies, for example, as an outgroup to mice [66].

In summary, our data have enabled the development of a compendium of experimentally defined TSSs and transcriptomic features from multiple tissues and cell types from the same hamster colony from which CHO cells were generated. Our revised annotation shows considerable improvement over the current RefSeq by several measures including agreement with published RNA-seq datasets, TSS information content as well as core promoter motifs. More broadly, these findings emphasize the importance of refined TSS mapping methods such as 5′GRO-seq/GROcap or csRNA-seq for accurate annotation of a gene's 5′ end. The TSS is a landmark in gene regulation and its accuracy becomes imperative in an era of genetic engineering. We further envision that our data and annotation will provide a rich resource for the CHO community and beyond as the Chinese hamster is further included in comparative and evolutionary studies. At its core, the improved TSSs map will aid CHO gene engineering efforts aiming to improve the quality and quantity of desired recombinant proteins and ultimately reduce drug manufacturing costs.

2.6 Data Availability

All sequencing data are submitted to the Gene Expression Omnibus (GEO) with GEO ID GSE159044. The Supplementary Data provided is also uploaded to Synapse (synapse.org), with ID syn22969187. This includes our revised protein-coding promoter TSS annotation, in which each of TSS has an associated RefSeq transcript and gene association. This is done for both NCBI RefSeq (Supplementary Data 2 in [77]) and with RefSeq in conjunction with the proteogenomics annotation reported in [76] (Supplementary Data 3 in [77]). Open-chromatin regions merged across samples are provided on synapse as a bed file as well.

In addition, we release all our TSSs detected (Supplementary Data 1 in [77]), which include distal TSSs (putative enhancer regions, divergent transcripts), as well as non-coding RNA promoter TSSs and protein-coding TSSs, along with the CPM from each tissue per TSS and the respective TSS locations of the tissue if it expressed that TSS. This will allow researchers studying regulatory elements to have easy access to a comprehensive TSS dataset.

2.7 Supplementary Figures and Tables

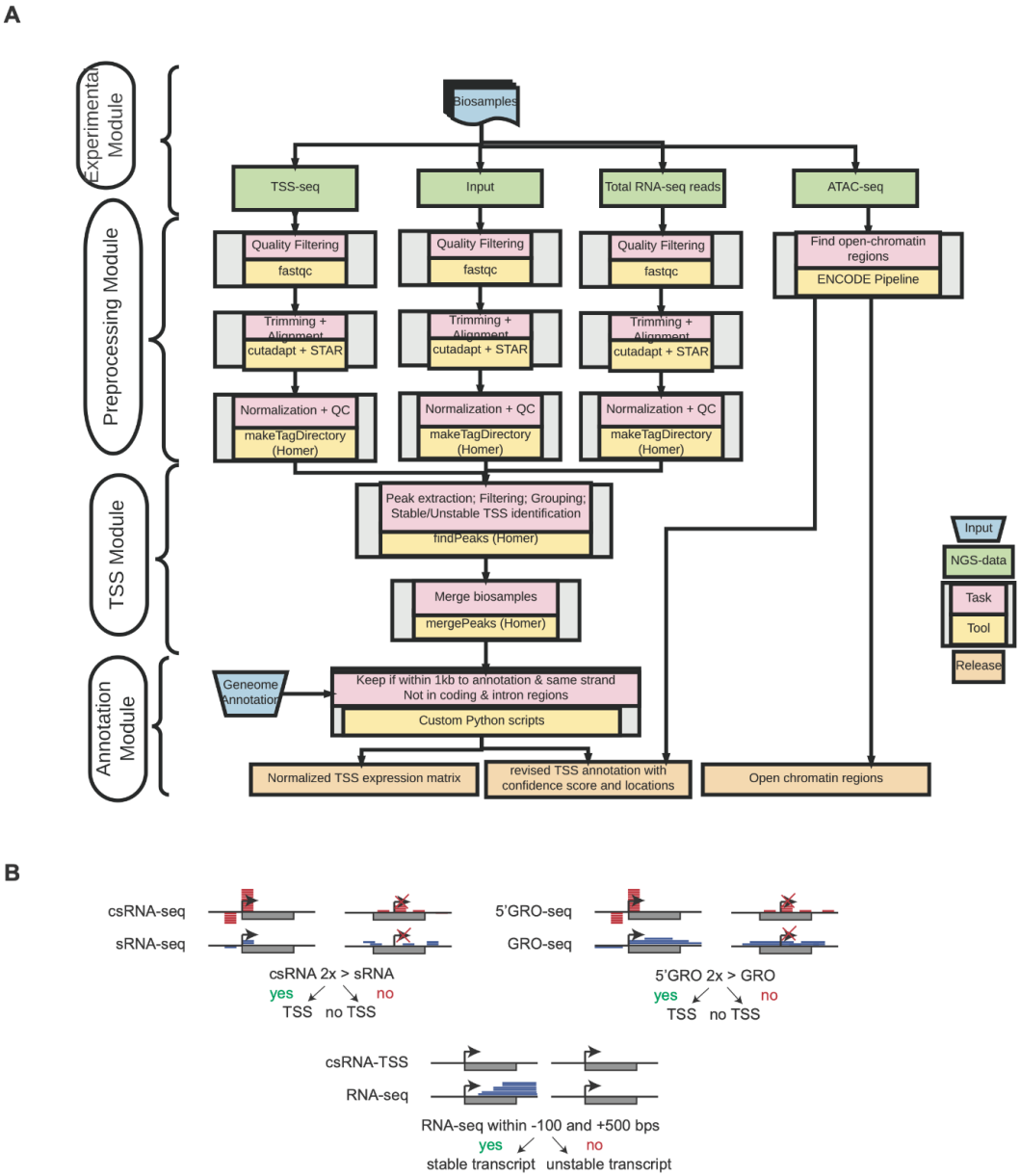


Figure S2.1 Computational workflow for TSS annotation
(A) Bioinformatics pipeline for promoter TSS annotation. This shows the steps required along with the commands used to run these steps. (B) Scheme of transcription start site identification for csRNA-seq, and how stable TSSs are called.

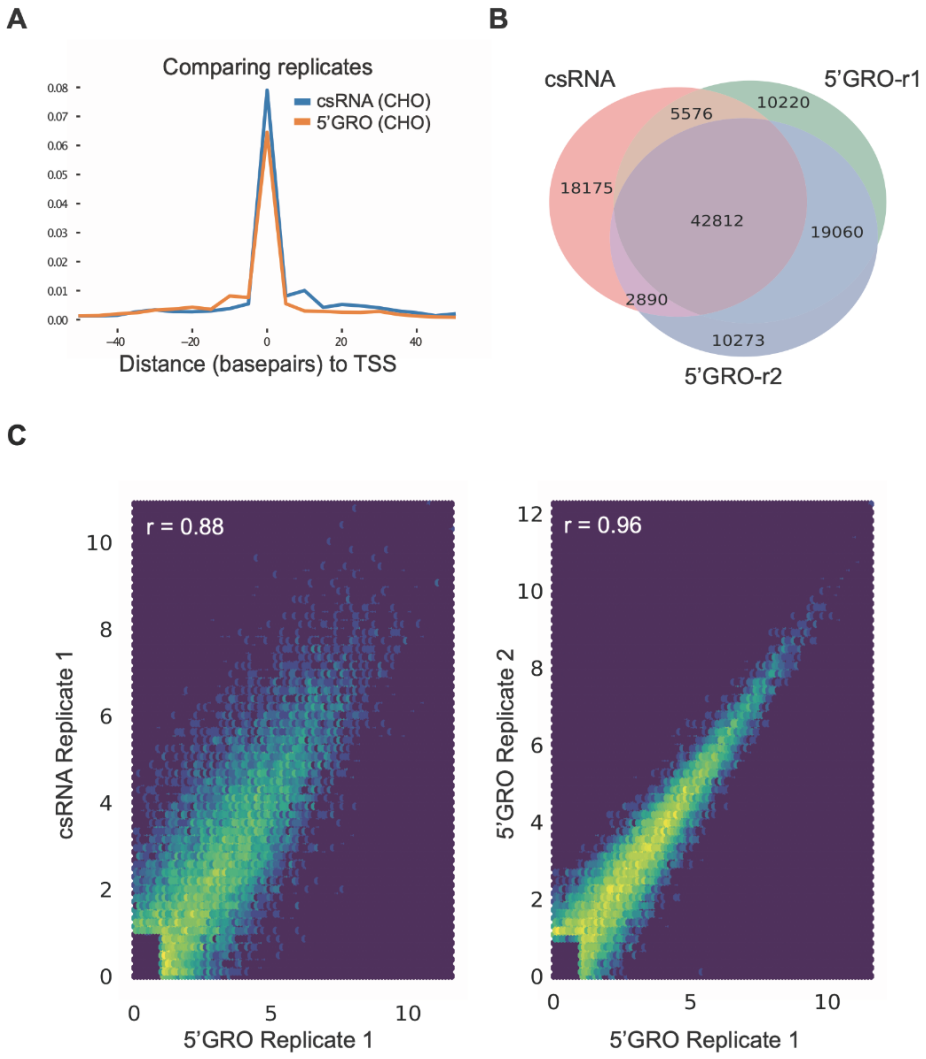
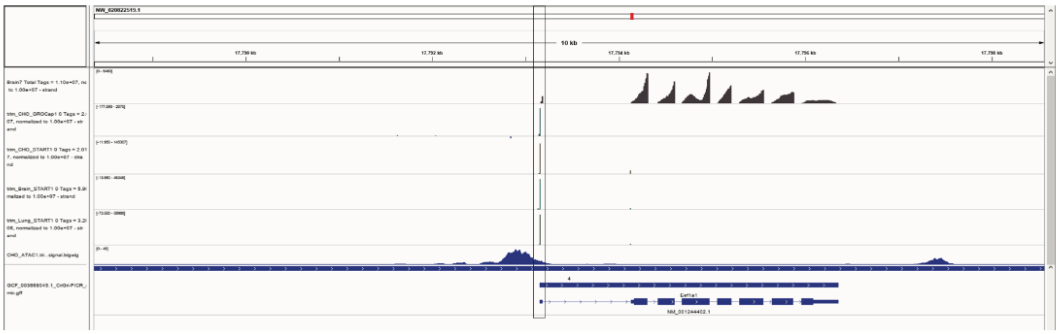
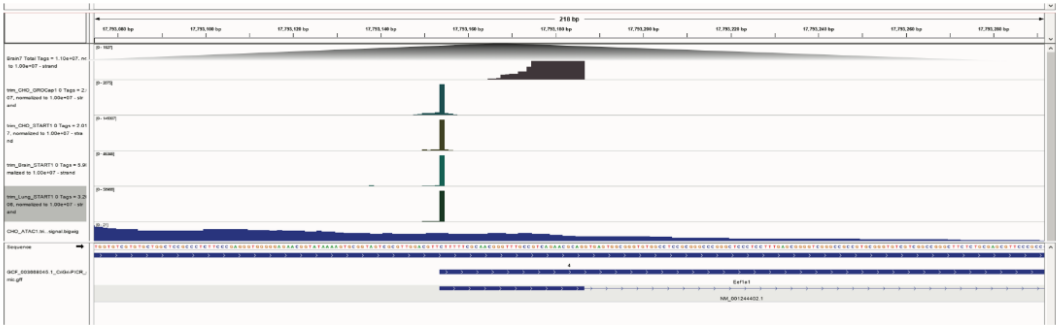


Figure S2.2 TSS similarity across experiments
(A) Density plot of the distance between peaks in different CHO replicates relative to a CHO GROcap sample B. CHO csRNA and GROcap sample A are shown nearby. **(B)** Number of overlapping total TSSs across all CHO replicates. **(C)** Density scatterplot of CHO replicates. Values are in log2 CPM. Left: CHO GRO-cap A vs csRNA-seq (pearson r=0.88, p-value < 0.001) Right: CHO GRO-cap replicates. (r=0.96, p-value < 0.001).

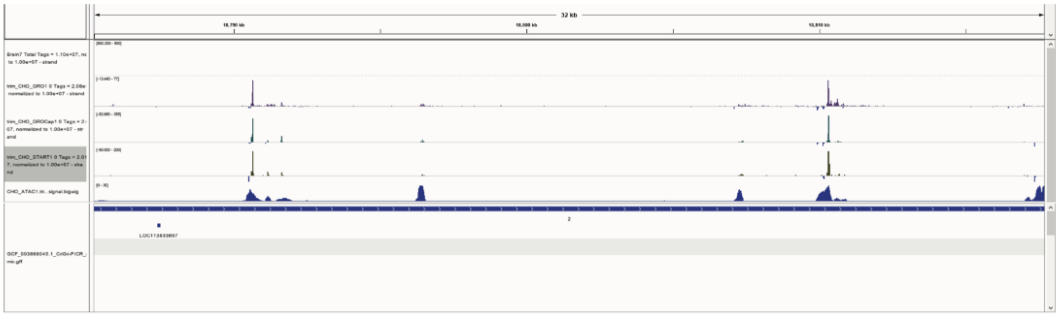
A



B



C



D

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Olfactory transduction	RT		1040	22.1	0.0E0	0.0E0
	KEGG_PATHWAY	Taste transduction	RT		31	0.7	2.3E-6	3.2E-4
	KEGG_PATHWAY	Autoimmune thyroid disease	RT		39	0.8	1.3E-3	1.1E-1
	KEGG_PATHWAY	Neuroactive ligand-receptor interaction	RT		103	2.2	1.5E-3	1.0E-1
	KEGG_PATHWAY	Jak-STAT signaling pathway	RT		53	1.1	3.7E-2	8.8E-1
	KEGG_PATHWAY	Cytokine-cytokine receptor interaction	RT		68	1.4	4.8E-2	9.0E-1
	KEGG_PATHWAY	Ribosome	RT		73	1.5	7.7E-2	9.6E-1

E

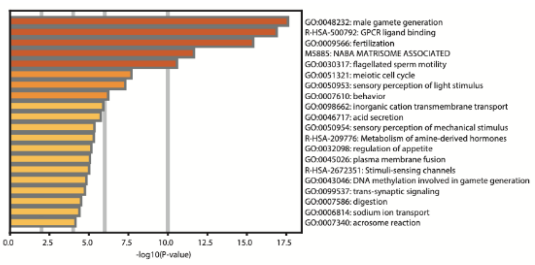


Figure S2.3. TSS sequencing highlights diversity of regions captured.
(A) IGV browser shots from Eukaryotic Translation Elongation Factor 1 Alpha 1Gene and (B) promoter. (C) IGV browser shot of the Sp1 "super enhancer". (D) Enriched pathways using DAVID and gene ontology using Metascape (E) of genes for which TSSs were not detected in our samples.

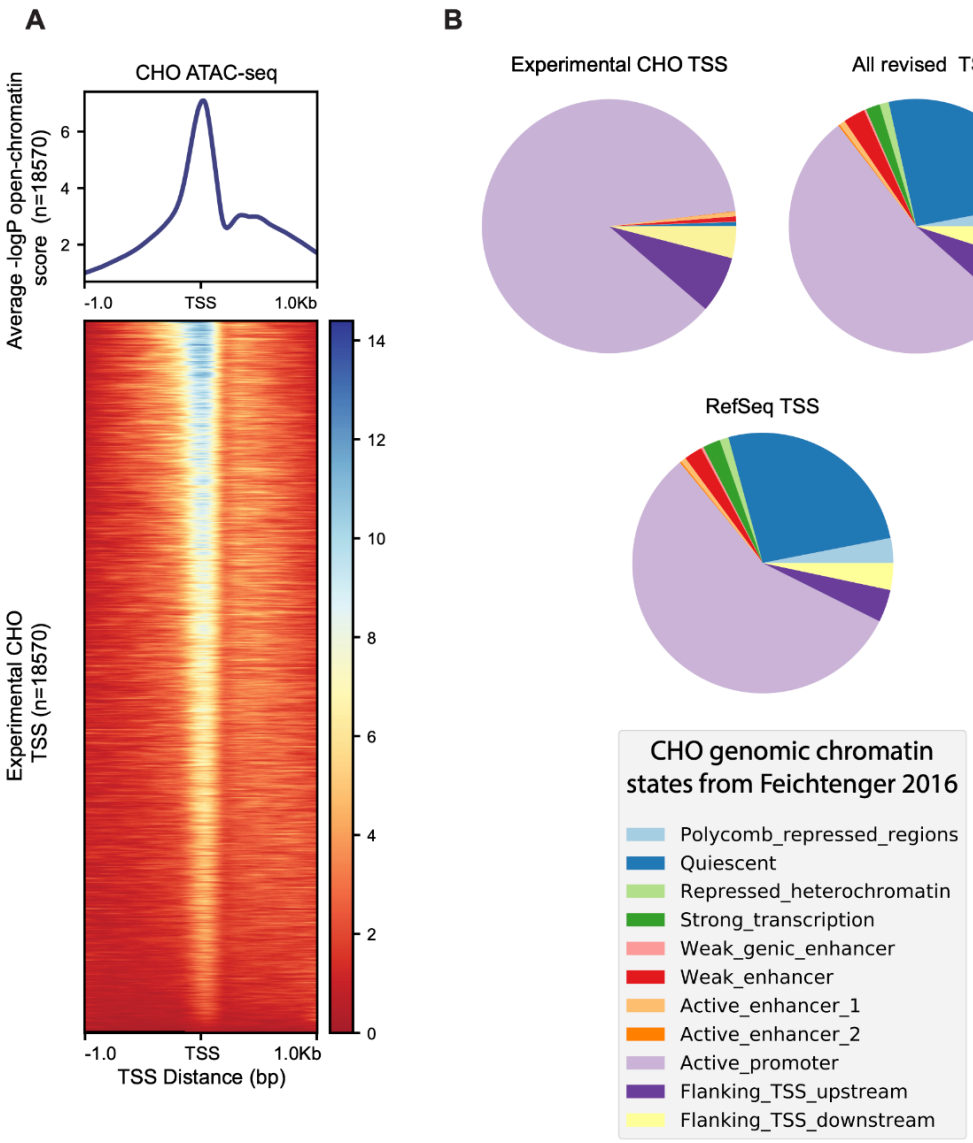


Figure S2.4. Epigenetic characterization of protein-coding experimental TSSs.
(A) ATAC-Seq pileup over TSS regions. Values are -log10 p-value of the base pair being in open-chromatin. Top: Histogram of a CHO ATAC-Seq sample using all revised protein-coding TSSs that were expressed in CHO. Bottom: The same regions, in heatmap form, where each row is a revised TSS. **(B)** Overlap of protein-coding TSS with chromatin modifications: our experi- mental CHO TSSs, our updated annotation (all protein-coding revised TSSs) and the RefSeq TSSs grouped by chromatin state made using chromHMM with histone marks from CHO in Feichtinger et al. 2016.

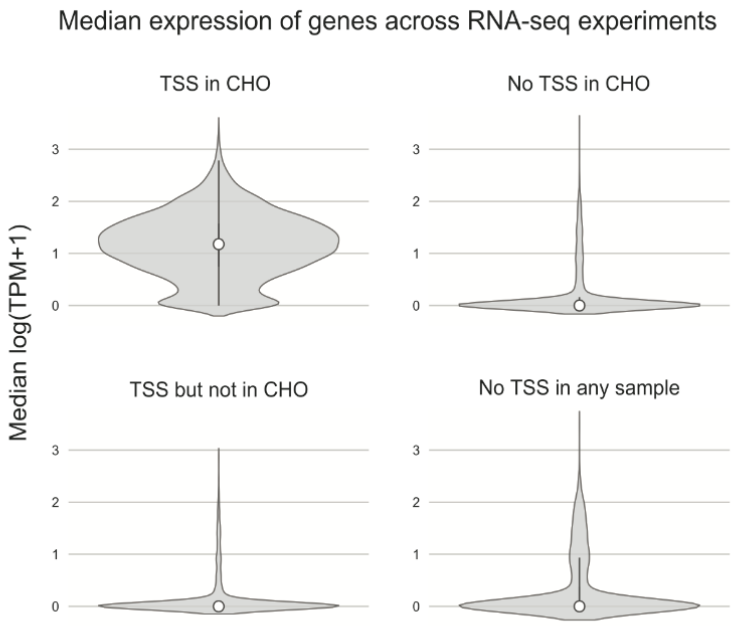


Figure S2.5. Detected TSSs found in expressed CHO genes. Violin plots of gene expression in CHO cells, grouped based on experimental TSS findings. 1,558 RNA-seq samples were used across different CHO cell lines and experiments. Expression is increased in genes where there are TSSs.

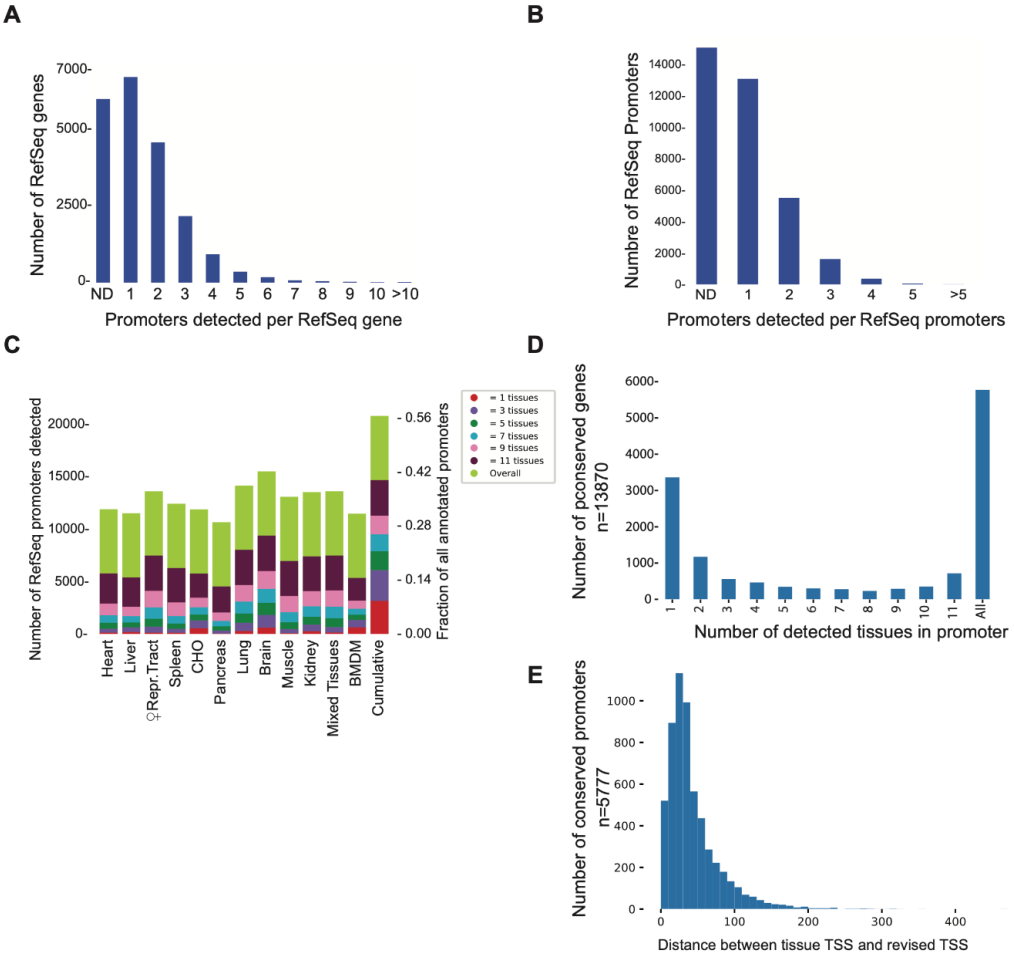


Figure S2.6. Diversity of promoter and TSS usage across annotation and tissues.
(A) Number of detected promoters per RefSeq gene (corresponding to different isoforms). **(B)** Number of revised promoters per RefSeq promoters. **(C)** Fraction of RefSeq promoters detected by each tissue and cell-type. **(D)** Bar plot of how many tissues detected in promoters of conserved genes (defined as having at least one promoter TSS detected in all samples). **(E)** Distribution of the maximum distance (bps) between the final revised TSS annotation based on integrating the tissue TSSs, and the initial tissue TSSs in the same promoter region in conserved promoters (each tissue having a detected TSS in the promoter).

Table S2.1 Samples and sequencing experiments in the Chinese hamster and CHO cells

Biological Source	5'GRO-seq	GRO-seq	csRNA-Seq	sRNA-seq	ATAC-seq	RNA-seq
CHO K1	2	2	1	1	2	0
Bone-marrow derived macrophages (BMDMs)	1	1	0	0	2	0
BMDMs with 1h Kdo2-Lipid A (KLA)	1	1	0	0	2	0
Brain	1	1	2	1	1	2
Heart	0	0	2	1	1	1
Kidney	1	1	2	1	1	1
Liver	1	1	2	1	1	1
Lung	1	1	2	1	1	1
Mixed Tissues	0	0	2	1	0	1
Muscle	0	0	2	1	0	1
Pancreas	0	0	2	1	0	1
Reproductive Tract	0	0	2	1	0	1
Spleen	0	0	2	1	1	1
Sum	8	8	21	11	12	11

Table S2.2 Mgat3 gRNA target sequences

	Target sequence
Mgat3	
gRNA_1	CTAGCTCTAGAAGCCGTCTTGG
gRNA_2	ATATCAAACTCCCACTAGCAGG
gRNA_3	TAAAGTCCAAGCATGCTAGAGG

Table S2.3 Mgat3 gRNA sequences

Name	oligo
CHOoptPICR_Mgat3_gRNA1_Fwd	GGAAAGGACGAAACACCGCTAGCTCTAGAAGCCGTCTGTTTTAGAGCTAGAAAT
CHOoptPICR_Mgat3_gRNA2_Fwd	GGAAAGGACGAAACACCGATATCAAACTCCCACTAGCGTTTTAGAGCTAGAAAT
CHOoptPICR_Mgat3_gRNA3_Fwd	GGAAAGGACGAAACACCGTAAAGTCCAAGCATGCTAGGTTTTAGAGCTAGAAAT
CHOoptPICR_Mgat3_gRNA1_Rev	CTAAAACAGACGGCTTCTAGAGCTAGCGGTGTTTCGTCCTTTCCACAAGATAT
CHOoptPICR_Mgat3_gRNA2_Rev	CTAAAACGCTAGTGGGAGTTTGATATCGGTGTTTCGTCCTTTCCACAAGATAT
CHOoptPICR_Mgat3_gRNA3_Rev	CTAAAACCTAGCATGCTTGGACTTTACGGTGTTTCGTCCTTTCCACAAGATAT

Table S2.4 CHO RNA-seq Accession IDs for data in Figure S2.5

Public RNA-seq CHO data that was used for Supplementary Figures, in addition to unpublished CHO datasets.
ERR359637
ERR359638
ERR366009
ERR366010
SRR035274
SRR035275
SRR035276
SRR035277
SRR035278
SRR035279
SRR035280
SRR035281
SRR035282
SRR035283
SRR035284
SRR035285
SRR950107
SRR950108
SRR950109
SRR1516214
SRR1516215
SRR1516216
SRR1516217
SRR2922597
SRR2922598
SRR2922599
SRR2922600
SRR2922601
SRR2922602
SRR2922603
SRR2922604
SRR2922605
SRR2922606
SRR3401745
SRR3401746
SRR3401747
SRR3401748
SRR3401749
SRR3401750
SRR3401751
SRR3401752

2.8 Acknowledgements

We like to thank Marten A. Hoeksema for culturing BMDMs.

Dedication: The authors dedicate this work to Dr. George Yerganian (1924–2019), who provided the hamsters for this study, and for the original CHO cells in 1957.

Chapter 2, in full, is a reprint of the material as it appears in "Isaac Shamie, Sascha H Duttke, Karen J la Cour Karottki, Claudia Z Han, Anders H Hansen, Hooman Hefzi, Kai Xiong, Shangzhong Li, Samuel J Roth, Jenhan Tao, Gyun Min Lee, Christopher K Glass, Helene Faustrup Kildegaard, Christopher Benner, Nathan E Lewis, A Chinese hamster transcription start site atlas that enables targeted editing of CHO cells, NAR Genomics and Bioinformatics, Volume 3, Issue 3, September 2021, lqab061, https://doi.org/10.1093/nargab/lqab061." The dissertation author was the primary investigator and author.

2.9 Funding

National Institutes of Health/National Institute of General Medical Sciences [K99GM135515 to S.H.D]; National Institutes of Health [AI135972, GM134366 to C.B.]; Novo Nordisk Foundation [NNF10CC1016517, NNF20SA0066621 to N.E.L., A.H.H.; NNF16OC0021638 to H.F.K.]; Cancer Research Institute Irvington Postdoctoral Fellowship Program (to C.Z.H.).

Conflict of interest statement. None declared.

CHAPTER 3

Comparative single-cell lineage bias in human and murine hematopoietic stem cells

Isaac Shamie[1,*], Meghan Bliss-Moreau*[2,3], Jamie Casey Lee[4,*], Nathan E. Lewis[1,4,Φ], Yanfang Peipei Zhu[4,Φ], Ben A. Croker[2,3,4,Φ]

[1]Department of Bioengineering, UC San Diego, La Jolla, CA, USA

[2]Division of Hematology/Oncology, Boston Children's Hospital, Boston MA, USA

[3]Department of Pediatrics, Harvard Medical School, Boston MA, USA.

[4]Division of Rheumatology, Allergy & Immunology, Department of Pediatrics, School of Medicine, UC San Diego, La Jolla, CA, USA.

Keywords: mitochondrial lineage tracing, scATAC-Seq, hematopoietic stem cells, CD34+, HSPCs, lineage-bias

Abstract word count:

Word count:

The authors declare no financial conflicts of interest.

*contributed equally

[Φ]contributed equally

Correspondence: Ben A. Croker, email: bcroker@health.ucsd.edu; Yanfang Peipei Zhu, email: peipeizhu@health.ucsd.edu; Nathan E. Lewis, email: nlewisres@ucsd.edu

3.1 Abstract

The biased commitment of hematopoietic stem cells (HSC) to myeloid, erythroid, and lymphoid lineages is influenced by cell-intrinsic and epigenetic differences in the HSC population. To investigate the nature of lineage commitment bias in human HSC, we use mitochondrial single cell (sc) RNA-Sequencing (mt-scATAC-Seq) which exploits somatic mutations in mitochondrial DNA, acting as natural barcodes, to track the ex vivo differentiation potential of HSC to myeloid and erythroid cells. Clonal lineages of human CD34+ cells and their mature progeny were normally distributed across the hematopoietic lineage tree without evidence of significant skewing. These data suggest that the variation in stem cell lineage commitment is restricted in normal bone-marrow hematopoiesis.

3.2 Introduction

Hematopoietic stem cells (HSCs) are classically considered to have the capacity for complete regeneration of the hematopoietic compartment. More recent analyses indicate additional complexity and heterogeneity in the HSC compartment, with lineage-restricted or lineage-biased HSCs considered a feature of mammalian hematopoiesis [121–134]. In humans, bone-marrow CD34+ cells were discovered to contain both stem and progenitor cells (HSPCs) with different lineage differentiation capabilities [135–137], including long-term HSCs (LT-HSCs), which can re-populate the required downstream lineages but who's heterogeneity in self-renewal and differentiation capabilities are still being uncovered. Understanding the hematopoietic capacity in HSC 'clones' (cells related to the same HSC) in the BM and other adult tissues requires precise delineation of differentiation trajectories from stem cells to mature cells at single-cell resolution. Clonal relationships between HSPC and mature hematopoietic lineages have also been explored using inducible DNA ''barcoding'' methodologies during embryogenesis or postnatal life, or upon the transplantation of virally-transduced barcoded single HSCs to lethally-irradiated or hemo-ablated adult murine or NHP recipients [138–144]. In humans, however, *in vivo* clonal-barcoding has been limited to xenograft transplantation, cancer clonality, and integration site tracking of gene-edited HSPCs in patients receiving gene-therapy [145–147]. The identification of cell surface markers that are enriched in specific subsets of HSPCs has enabled the prospective isolation of HSPC subpopulations for cellular and biochemical analysis *ex vivo*, and for precise cellular tracking of HSC and their progeny in transplanted mice. In mice and humans, next generation sequencing (NGS) and gene expression analyses have identified differentially expressed surface receptor genes modulated within the LT-HSC and downstream multipotent progenitor (MPP) populations, and enabled development of antibodies for purification and phenotypic analysis [148–151], although numerous distinct definitions still currently exist for them. More recently, HSPC subtypes have been assigned using single-cell NGS that measures a cell's transcriptome or epigenome. Coupled with computational lineage-tracing techniques [52,152–154], these methods have revealed overlapping subtypes as FACS-based methods, but with additional diversity and

characterization. However, these studies infer cell lineage trajectories using computational trajectory inference and dimensionality reduction methods, which assumes a direct coupling of the transcriptome or epigenome to cell clonality [52].

Tracking related cells in humans *in vivo* and *ex vivo* is limited due to the inability to add genetic barcodes or transplantation in healthy humans. Detecting clones in intact cells is sometimes done using their somatic mutation similarities, but this suffers from low mutation rate and coverage. Recently, Ludwig et al. [155] reported that mitochondrial somatic mutations can be used as natural barcodes to track single-cells, and a later protocol deemed deemed mitochondrial single-cell ATAC-seq (mt-scATAC-seq) was developed that simultaneously measures mitochondrial (MT) DNA and nuclear open-chromatin regions in single cells [156]. Using mitochondrial variants to track clones is advantageous due to the MT genome having a high per-cell copy-number, a smaller genome, a higher mutation rate than the nuclear genome, and factors such as random genetic drift and relaxed replication allow the genome to reach higher heteroplasmic proportion (fraction of MT copies with a variant) [155,157–159]. Additionally, lineage bias of HSPC progenitors can be assessed using the detected lineage markers in the open-chromatin regions of the nuclear genome.

Here we report on inherent HSC lineage bias potential in steady-state and cultured HSPC CD34+ cells' lineage potential using mt-scATAC-seq. We find heterogeneity in HSPC clonal sizes, with a few larger clones making up a large fraction of the population. Lineage markers, as determined using the open-chromatin epigenome, were assigned to cells, and limited differences in lineage bias across the HSPC clones. Overall, our data supports a model of hematopoiesis where HSC drives multi-lineage reconstitution of mammalian hematopoiesis without substantial lineage bias.

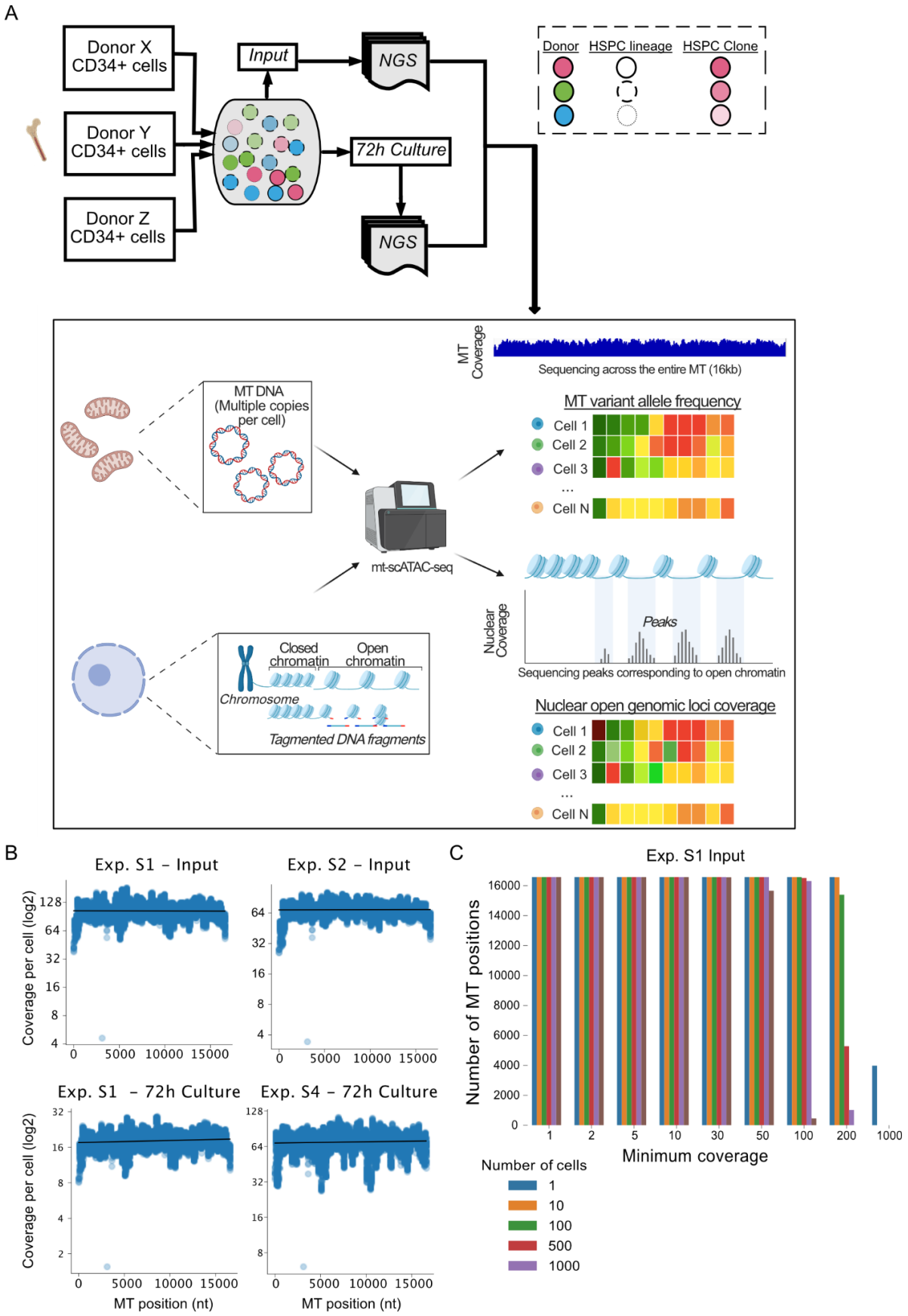
3.3 Results

3.3.1 mt-scATAC-seq defines clonal lineages in mobilized human CD34+ cells

To study differentiation of primary human CD34+ cells to committed myeloid and erythroid cells, we purified G-CSF-mobilized CD34+ cells from healthy donors, and cultured them for 72 hours in the presence of a cytokine cocktail made of SCF/IL-3/IL-6/Flt3L/G-CSF/GM-CSF. Donor CD34+ cells were multiplexed in three batch experiments for single cell capture and library preparation before or after culture (**Figure 3.1A**, **Table S3.1**, see Methods). Cells from eight healthy donors were analyzed in total, and Donors 1-4 contained verified mutations associated with clonal haematopoiesis of indeterminate potential [160]. We utilized mitochondrial DNA somatic mutations that were defined by scATAC-Seq to generate lineage-specific 'natural' barcodes, which were inherited by clones that originated from one single CD34+ cell to delineate clonality of these CD34+ stem cells and their progeny. In parallel, the lower-coverage reads in nuclear open-chromatin regions were used for assigning cell lineages by assessment of differential peaks in gene regulatory regions. These techniques allowed us to track the fate of these CD34+ HSPCs to assess their lineage potential at single cell level. This modified protocol exploits a fixation step and a modified permeabilization protocol to retain mitochondria in cells for subsequent capture and library preparation (see Methods, and [156]). Compared to scRNA-Seq with custom amplification of mitochondrial reads, this method achieves near complete coverage of the mitochondrial genome for optimal variant calling and cellular 'barcoding". Examination of single cell coverage across each mitochondrial DNA position in different donors yielded consistent high-level coverage of greater than 50x-100x in a minimum of 1000 cells (**Figure 3.1B-C** and **Figure S3.1A**). Although there was variable read coverage across sequencing runs, uniform coverage across the MT genome was observed (**Figure 3.1B**). Low-quality cells and MT alleles with low coverage and base-quality were filtered, and then additional variants were excluded using the Mitochondrial Genome Analysis Toolkit (MGATK) [156], which removes variants with a low correlation of allelic reads across strands and a low variance-mean ratio (**Figure 3.S1B**, see Methods). To de-multiplex each donor across conditions we used the Vireo

algorithm on germline MT variants (**Figure 3.S2**, see Methods). The donor predicted variant allele-frequency (VAF) revealed up to 28 (**Figure S3.2B**,**E**, **Table S3.2)** variants of high mean VAF (mean >0.7 in donor, <0.1 in others), highlighting many donor-specific variants, primarily transition mutations (**Figure S3.2C**). Performance was assessed by varying the number of donors and calculating the model loss, and the 'elbow rule' finds that the true number of donors is the inflection point in which performance gain is reduced with additional donors added to the model (**Figure S3.2D**). More donors did lead to lower recovery of donor specific variants (**Figure S3.2E**), which is possibly due to having less unique variants across more donors, but was still able to separate the donors confidently.

Figure 3.1 mt-scATAC-seq defines clonal lineages in mobilized human CD34+ cells
(A) Overview of mitochondrial mt-scATAC-seq workflow **(B)** Coverage across MT genome in single-cells across sequencing experiments. Black line is the mean at each position **(C)** Number of MT positions covered across a range of cells and coverage thresholds in the S1 input sample


A
Donor X CD34+ cells
Donor Y CD34+ cells
Donor Z CD34+ cells
Input
NGS
72h Culture
NGS
Donor
HSPC lineage
HSPC Clone
MT Coverage
Sequencing across the entire MT (16kb)
MT DNA (Multiple copies per cell)
MT variant allele frequency
Cell 1
Cell 2
Cell 3
...
Cell N
mt-scATAC-seq
Nuclear Coverage
Peaks
Sequencing peaks corresponding to open chromatin
Closed chromatin
Open chromatin
Chromosome
Tagmented DNA fragments
Nuclear open genomic loci coverage
Cell 1
Cell 2
Cell 3
...
Cell N
B
Exp. S1 – Input
Exp. S2 – Input
Exp. S1 – 72h Culture
Exp. S4 – 72h Culture
Coverage per cell (log2)
MT position (nt)
C
Exp. S1 Input
Number of MT positions
Minimum coverage
Number of cells
1
10
100
500
1000

3.3.2 Single CD34+ stem cells expand into clones of variable sizes identified by mitochondrial variants

Clone assignment is critical to track the fate of CD34+ stem cells at the single cell level. In addition to the separation of donors using donor-specific germline variants, we used somatic MT variants to facilitate clone assignment in human CD34+ cells pre- and post- cytokine-driven differentiation at 72 hours of *ex vivo* culture. To achieve this goal, we used a previously reported approach [156] and detected clones using a community-based k-nearest neighbors (KNN) clustering algorithm on VAFs across single-cells. After removing clones with fewer than 5 cells, we detected 26-50 clones per donor (**Table 3.1**). We find that a few clones make up majority of the population in donor 1 (**Figure 3.2A, Table 3.1**), with 25% of the clones making up 50% of the number of cells in the HSPC pool, followed by longer tail of small clones, with 77.7% making up 90%. Interestingly, no clone in Donor 1 appeared to be preferentially expanded in culture (**Figure 3.2B)**, suggesting that both large and small clones are actively contributing to hematopoiesis. Similar clone size heterogeneity was seen across donors **(Table 3.1, Figure 3.2E),** although the donors that were sequenced after culture appeared to have larger clones make up a larger fraction of their population (donors 5-8, **Figure 3.2E**). This could be in part due to some clones preferentially expanding in culture that was not observed in donor 1 and donor 2 due to lower number of cells detected (**Table S3.1)**. The clones detected had a range of distinct MT variants that separated those cells from the other clones (**Figure 3.2C, Figure 3.2D)**. However, as the MT genome is heteroplasmic, clone assignment was based on VAF rather than binary variant calls. Indeed, cells in clones were distinguished across a range of VAF (**Figure S3.3).** Interestingly, we noticed that some variants were shared across clones, seen in variant 5581G (i.e. allele G at position 5581 of the MT genome) (**Figure 3.2D)**. These variants shared across clones are predicted to have arisen from a common ancestral stem cell. We note that variant loss is also possible, as clones ‘D1.30’ (numbered by size) and ‘‘D1.32 in donor 1 share barcode 14233G but clone ''D1.32’ is missing 5581G. We also compared different variant-calling and clone-detection workflows (see Methods), and find a high concordance of cell pair clonal relationships observed across parameters (**Figure S3.4A-B**), but we did find lower performance when the

number of nearest neighbors in the KNN algorithm was low (resulting in larger, more sparse clusters), leading to lower consistency in clones detected when re-running the method in subsampled cells (**Figure S3.4C,** see Methods**).**

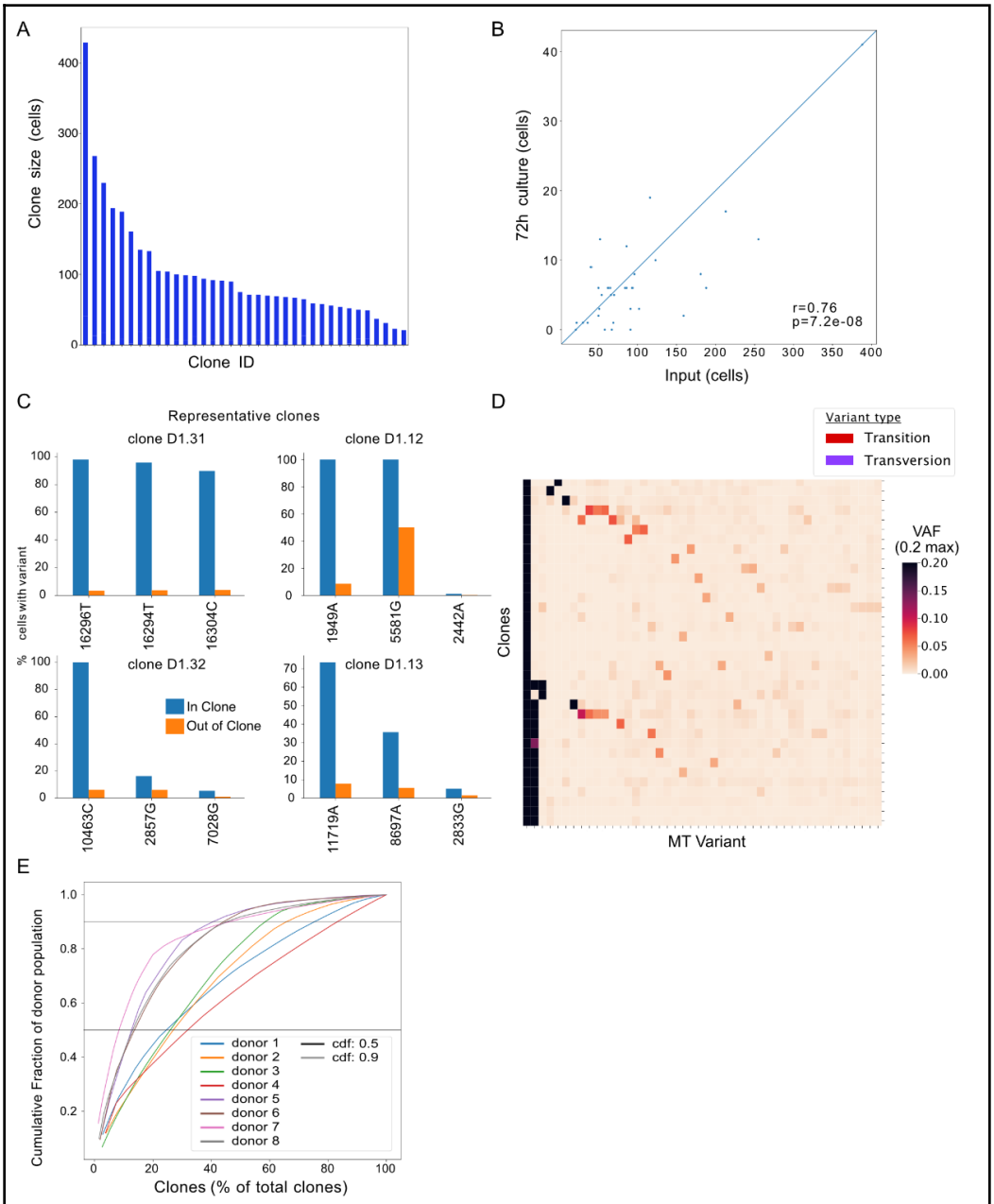


Figure 3.2 Single CD34+ stem cells expand into clones of variable sizes identified by mitochondrial variants

(A-D) Clones detected in donor 1 **(A)** Number of cells in each clone, colored by condition. **(B)** Scatterplot of number of cells across input and cultured cells. Pearson correlation r and p-value shown. **(C)** Barcodes detected in representative clones. In blue is the percent of cells with the barcode in the clone, and in orange is the percent in cells outside of the clone, shown for the top distinguishing barcodes for each clone. **(D)** Average VAF of each variant in each clone. Max cut-off at 0.2. Variant types shown for each variant and number of cells for each clone. **(E)** Cumulative distribution of the number of cells captured with increasing number of clones, sorted by largest to smallest, across all donors

Table 3.1 Clonal detection of human CD34+ BM cells using MT variants across 8 donors

		Number of cells in clone			Number of nuclear peaks			Number of cells in clone (fraction of donor)		
	Number of clones	mean +/- std	median	max	mean +/- std	median	max	mean +/- std	median	max
Donor										
Donor 1	36	101.61 +/- 79.23	73	429	5612.50 +/- 355.74	5611	6406	0.03 +/- 0.02	0.02	0.12
Donor 2	26	100.85 +/- 70.69	97.5	317	5405.59 +/- 531.32	5389	7457	0.04 +/- 0.03	0.04	0.12
Donor 3	34	38.62 +/- 24.36	41	91	3095.06 +/- 644.63	3020	5184	0.03 +/- 0.02	0.03	0.07
Donor 4	27	76.11 +/- 48.80	62	247	3216.47 +/- 366.80	3154	3892	0.04 +/- 0.02	0.03	0.12
Donor 5	33	63.39 +/- 74.85	26	264	3016.17 +/- 517.84	3042	3895	0.03 +/- 0.04	0.01	0.13
Donor 6	35	56.63 +/- 52.50	40	193	3324.59 +/- 492.69	3352	4498	0.03 +/- 0.03	0.02	0.1
Donor 7	41	30.68 +/- 40.60	10	213	3029.99 +/- 567.27	2948	4437	0.02 +/- 0.03	0.01	0.17
Donor 8	50	35.58 +/- 39.73	23	180	2401.89 +/- 573.19	2281	3902	0.02 +/- 0.02	0.01	0.1

3.3.3 mt-scATAC-seq identifies variable cell lineages in human CD34+ stem cell ex vivo culture

After we successfully identified the clonal lineage of CD34+ human HSPCs in each multiplexed donor using the mitochondrial open-chromatin captured by mt-scATAC-seq, we then used the nuclear genomic open-chromatin that was captured from the same experiments to identify active genomic loci to assign cell lineage therefore address the differentiation potential of each CD34+ cell. To achieve this goal, nuclear open-chromatin reads were processed using conventional scATAC-Seq tools, enabling dimension reduction and cell clustering. Quality control of experiments showed a comparable number of detected peaks across experiments (**Figure S3.5**). Focusing on the two donors that had both steady-state and cultured for 72 hours, we used the Signac protocol to integrate sample data, preprocess, and binarize data from cells (see Methods). These cells were embedded into a 2D map using uniform manifold approximation & projection (UMAP) and clustered. We identified 12 populations in the CD34+ cells before- and after- 72h cell culture based on analysis of the gene activity scores in exonic and promoter regions that were demonstrated to be critical for each cell lineage (**Figure 3.3, Table S3.3**). As peaks were also found in intergenic and other non-coding regions (**Figure S3.5D**), we confirmed markers using ChromVar transcription factor (TF) activity [161] as well, based on the TF motifs detected in a cell's open-chromatin peaks. We found that in different clusters gene activity scores are enriched at regions that

regulate the differentiation of HSPCs towards unique blood cell lineages, indicating the lineage-fate of each cluster of these *ex vivo* cultured CD34+ human HSPCs (**Figure 3.3C**). Among these clusters, 4 clusters display high KIT activity suggesting they are HSPC subsets. The trajectory of SPI1 activity from these clusters towards other clusters supports this annotation. The enrichment of CPOX/ANK1 identifies a megakaryocyte progenitor (MkP) cluster. Elevated activity at HBB/HBG2 suggests 1 cluster with erythrocyte potential. The combination of SPI1, FLT3, ATP8B4 suggests myeloid lineage potential in several clusters and these 5 clusters could then be further identified as monocytes, neutrophils, other granulocytes, and pDC based on analysis in gene regions shown in **Figure 3C**. Notably there is 1 cluster that displays pre-B phenotype, and was lost in post-culture samples (**Figure 3.3B**). Analysis of TF activities between clusters using chromVAR [161] supports the cell lineage assignment in each cluster (**Figure 3.3D**). We additionally ran the same pipeline by integrating cells across all 8 donors (**Figure S3.6A**), found similar UMAP lineages, and found cultured cells more represented in downstream progenitor lineages (**Figure S3.6B-D)**. The number of peaks detected across the UMAP was similar, although it was elevated in the detected pDC lineage (**Figure S3.5C**). Additionally, we found multiple clusters of similar type, such as two neutrophil clusters, that were not able to be resolved (**Figure 3.5A**).

To validate the results of mt-scATAC-seq we performed flow cytometry in parallel. Dimension reduction and automated clustering were performed on human CD34+ cells before- and after- 72h culture based on surface protein expression levels of HSPC markers CD34/CD117 (c-Kit), lymphoid lineage markers CD3/CD19/CD56, granulocyte lineage markers CD66b/FcεRIα/Siglec8, monocyte lineage markers CD14/CD16/CD86/CD11c, and additional developmental and maturation markers CD45/CD10/CD101/CD11b/HLA-DR. The result revealed high consistency with the mt-scATAC-seq that 13 clusters were identified including 4 HSPC subsets, 1 granulocyte-like cluster, 1 classical monocyte cluster, 1 non-classical monocyte cluster, 1 pre-B and 1 pDC cluster, 1 additional CD34+ cluster and 3 CD34- clusters (**Figure S3.7**). Similar to the mt-scATAC-seq results, the single pre-B and single pDC clusters disappeared after 72h cell culture, suggesting that these two populations may be a result of cross-contamination from the CD34+ cell sorting and undergo cell death in the *ex vivo* culture. On the

other hand, the frequency of these clusters identified by mt-scATAC-seq and flow cytometry are highly consistent (**Figure 3.3F**), suggesting the results from mt-scATAC-seq are valid. These results together, demonstrate that the CD34+ cells that were enriched from mobilized human blood consist of heterogeneous HSPC populations, and may differentiate into variable lineages of cells *ex vivo* in the presence of the cytokine cocktail made of SCF/IL-3/IL-6/Flt3L/G-CSF/GM-CSF. (**Figure 3.3B)**.

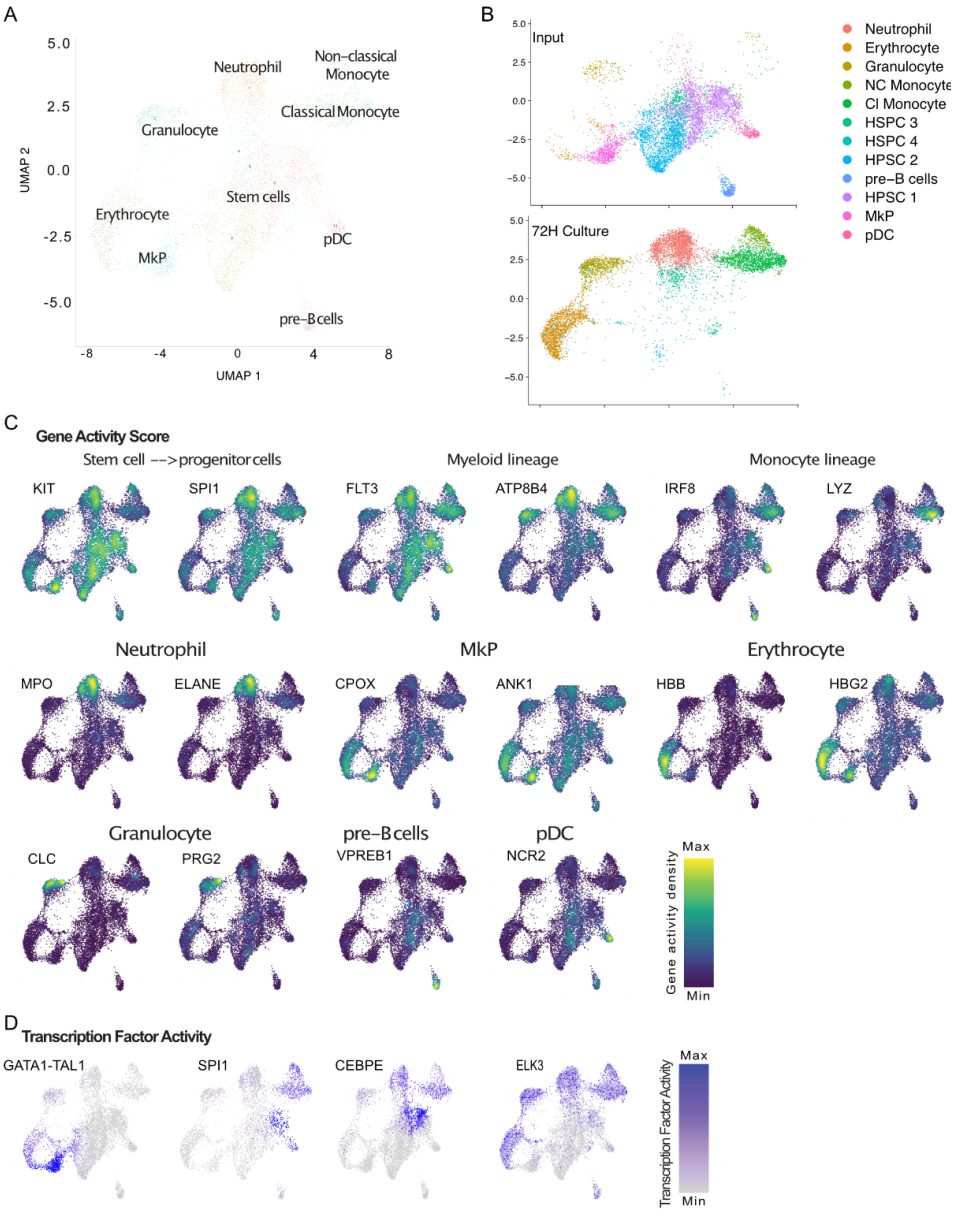


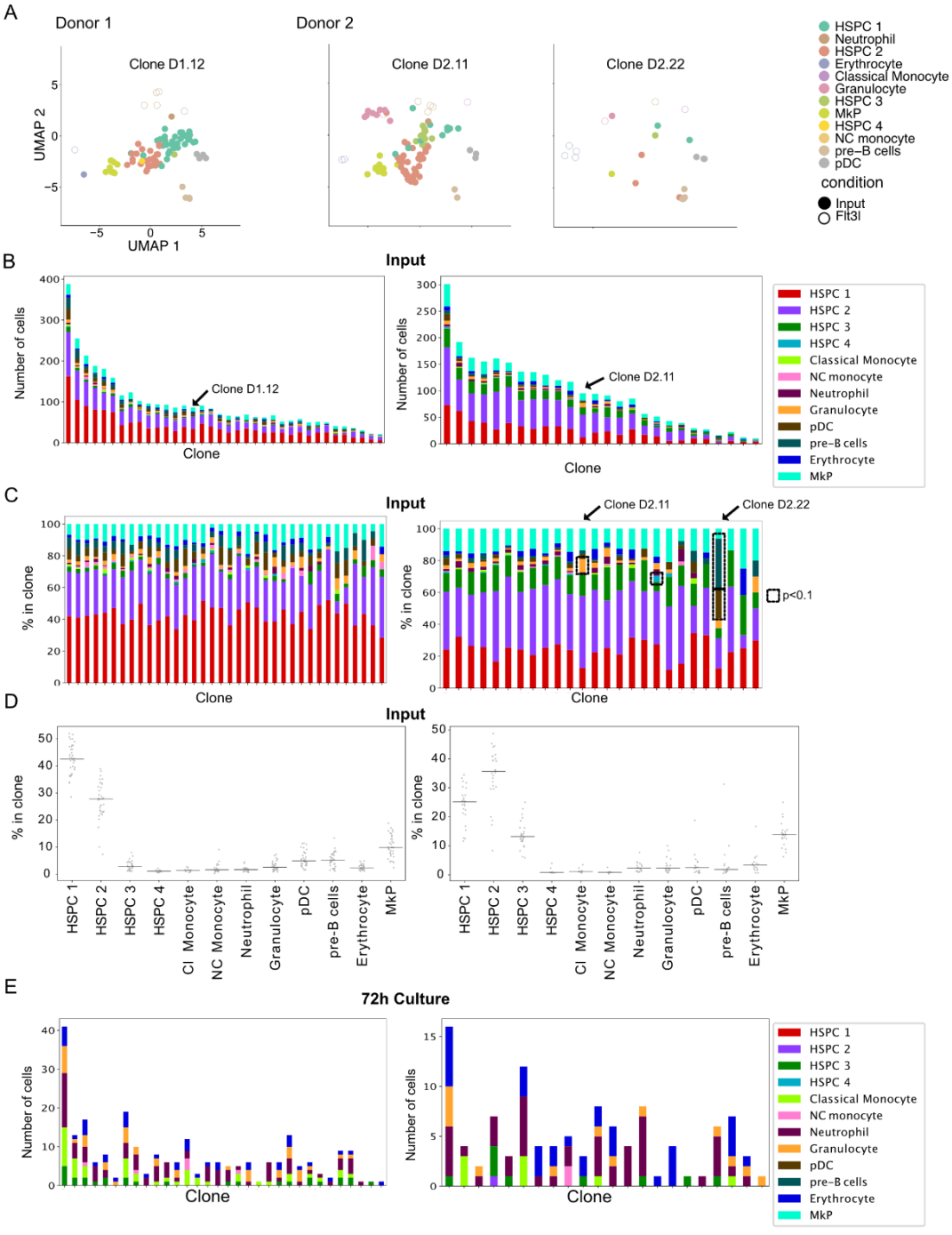
Figure 3.3 mt-scATAC-seq identifies variable cell lineages in human CD34+ stem cell steady-state and in *ex vivo* culture
2D UMAP embedding of single-cells in sequencing run S1 (2 donors) using nuclear ATAC-seq regions **(A)** UMAP of single-cells, with each cell colored by cluster labels, and manual HSPC type annotation overlaid on UMAP **(B)** UMAP split by conditions and colored by assigned cluster labels. (**C-D**) Lineage markers used to inform annotation **(C)** Gene activity scores for select markers overlaid on UMAP **(D)** Transcription factor activity scores for select markers

3.3.4 mt-scATAC-seq reveals minimum lineage-bias in steady-state and *ex vivo* differentiation of human CD34+ cells

We next looked to evaluate the multipotency of these clones and assess their lineage-bias by mt-scATAC-seq. When overlaying cells from clones on the UMAP using the donors in input and culture (**Figure 3.4A**), and analyzing their distribution (**Figure 3.4C)**, clones appeared randomly clustered with negligible evidence of lineage bias. Interestingly, while we detected variation in clone size within each cluster, most sizes are within a similar range. We tested if any clone was biased to certain HSPC clusters using a hypergeometric test on clone sizes in the input condition (see Methods). We found no biased clones in donor 1, but 4 clone-cluster pairs significant in donor 2. Clone 'D2.22' is significant in both pDC and pre-B annotated cells, clone 'D2.11' in granulocytes, and 'D2.17' in the HSPC 4 cluster (Benjamini-Hochberg adjusted p-value < 0.1). These were smaller clusters, but were not the only clones exclusively having these cells (**Figure 3.4B**). However, as mentioned earlier, the pDC and B-cells were not found in the *ex vivo* differentiated condition (**Figure 3.3C**). We also compared clone lineage bias from the clusters found across all donors (**Figure 3.5**). We looked to assess lineage bias in each clone using the entropy metric (see Methods), which measures how much information is conveyed in a distribution-if a clone is biased to one clone, it will have lower entropy. We see little variability across clone entropy, and the lower entropy clones were smaller (**Figure 3.5C**). We then examined the lineage proportions in each clone, and found that this varied widely across lineages (**Figure 3.5D**). Mainly, most clones were found in the first two HSPC clusters and a sizable fraction in the MkP population. However, the variation across clones within each lineage was smaller within each lineage compared to between them (**Figure 3.5D**). This was also measured across all donors using entropy in each lineage, converting the clonal counts into probability distributions for each lineage. (see Methods). Most donors shared similar entropy values, although Donors 5-8, which underwent *ex vivo* culture, had lower entropy in lymphoid based clusters (**Figure 3.5D**). While the majority of clones showed no skewing across clusters, we also determined if the resolution of clone detection influenced these data. To do this, we examined individual MT variant barcodes across the HSPC clusters and were unable to find significant biases across the clusters in donor 1

and donor 2 (**Figure S3.8**). Together, our data suggests that detectable HSPC clones have multi-potent capacity contributing to hematopoiesis in humans.

Figure 3.4. mt-scATAC-seq reveals minimum lineage-bias in the *ex vivo* differentiation of human CD34+ cells
Left: donor Donor 1, Right: Donor 2 **(A)** Cells in representative clone in each donor embedded in UMAP (B-D) Lineage fates of cells in input **(B)** Raw cell counts in each clone, colored by lineage cluster **(C)** Percent of immune lineage clusters across each clone for donors. Boxed values are significant ($p<0.1$) according to hypergeometric and non-parametric tests. **(D)** Percent of lineage in a clone, across all clones. (E) Same as (B), but for cells in 72 hour culture. Legend is the same as (B).


A
Donor 1
Donor 2
Clone D1.12
Clone D2.11
Clone D2.22
UMAP 2
UMAP 1
HSPC 1
Neutrophil
HSPC 2
Erythrocyte
Classical Monocyte
Granulocyte
HSPC 3
MkP
HSPC 4
NC monocyte
pre−B cells
pDC
condition
Input
Flt3l
B
Input
Number of cells
Clone
Clone D1.12
Clone D2.11
HSPC 1
HSPC 2
HSPC 3
HSPC 4
Classical Monocyte
NC monocyte
Neutrophil
Granulocyte
pDC
pre−B cells
Erythrocyte
MkP
C
Input
% in clone
Clone
Clone D2.11
Clone D2.22
p<0.1
D
Input
% in clone
HSPC 1
HSPC 2
HSPC 3
HSPC 4
Cl Monocyte
NC Monocyte
Neutrophil
Granulocyte
pDC
pre−B cells
Erythrocyte
MkP
E
72h Culture
Number of cells
Clone

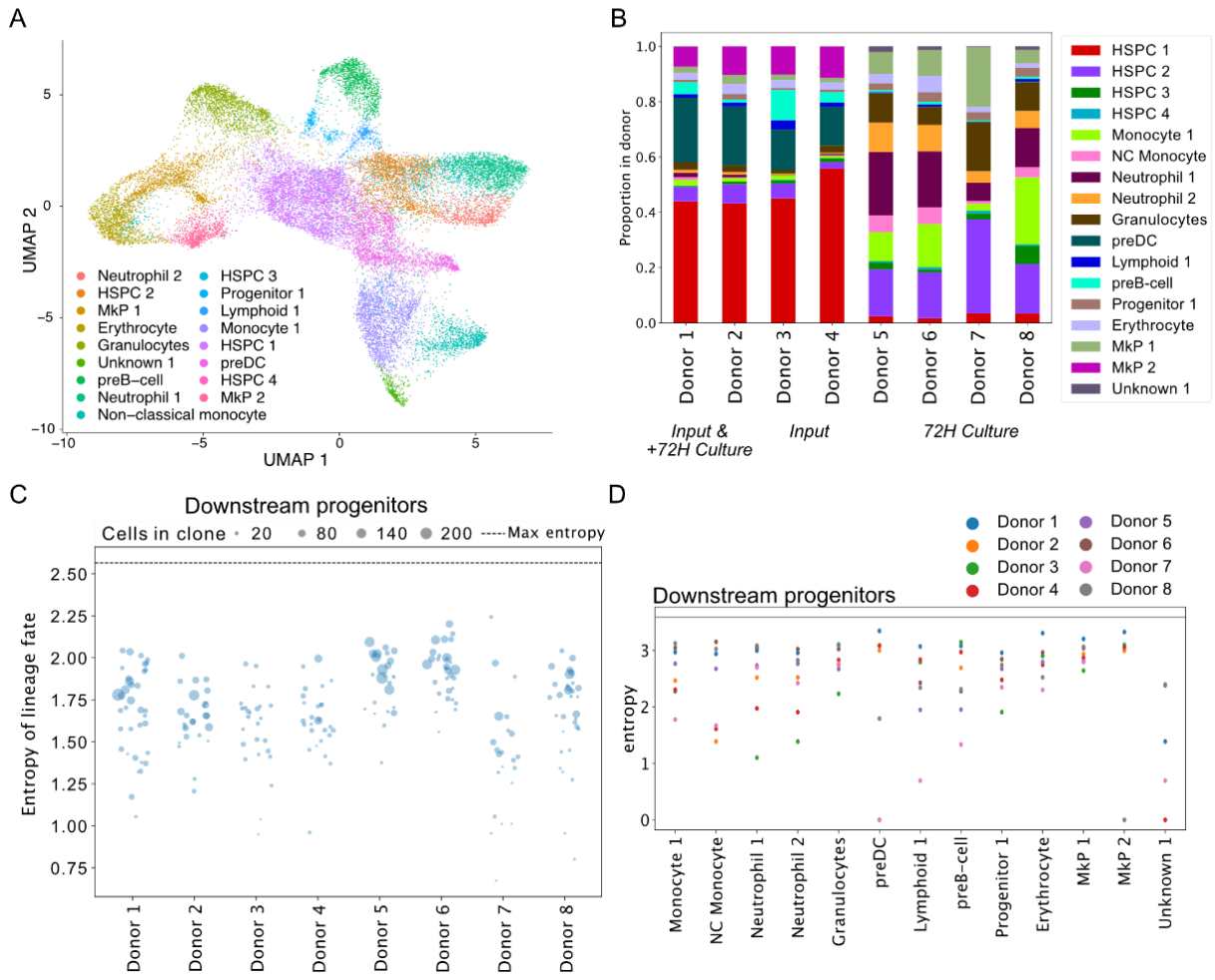


Figure 3.5. Minimum lineage-bias in human CD34+ HSPC clones across all donors

(A) Distribution of cells on UMAP, colored by annotated cluster labels **(B)** Proportion of cells across HSPC clusters in each donor **(C)** Entropy of lineage fate in each clone, sorted by rank within each donor. Same color legend as in (b) **(D)** Entropy of clonal bias in each lineage for each donor. Same color legend as in (B). Black line is maximum entropy possible given the maximum number of clones observe across donors.

3.4 Discussion

Here we report on tracking clonal HSPCs and their lineage commitment potential in humans and mice using cutting-edge NGS tools and innovative approaches. We provide evidence that CD34+ clones can proliferate and differentiate towards myeloid, lymphoid, and erythroid lineages without substantial variation across HSC clones. In humans, mt-scATAC-seq can be used to simultaneously track MT variants and open-chromatin regions in single-cells. Clones in the input CD34+ fraction are distributed diversely by ATAC–Seq signature, and differentiate to diverse lineages upon culture, suggesting that the MT variants found in clonal lineages are being inherited from a LT-HSC, and not from lineage-restricted progenitor cells. Longer term culture may enable dominant outgrowth of specific lineages and mask the true potential of CD34+ clonal lineages to generate diverse hematopoietic cell types [156]. These data are consistent with flow cytometric analysis of cultured cells over 72h. In total, we find little evidence of bias in lineage commitment in mobilized primary human peripheral blood CD34+ cells and in 72-hour cytokine culture *ex vivo*.

The methodology in this report also provides an opportunity to enhance the study of human HSPC *ex vivo* and *in vivo*. The study of genetic and biochemical regulators of HSC proliferation and differentiation can be explored using diverse methodology including clonal culture assays, single cell transcriptomics and epigenomics, and in vivo differentiation and function [162]. The effects of hematopoietic growth factors, inflammatory cytokines and pathogen-derived molecules on gene expression, cell function, and engraftment, can also be explored with higher resolution and confidence in the cell type in question. Barcoding and transplantation techniques provide excellent clone-detection methods, however the cellular and immune response to these editing techniques can introduce confounding effects that require consideration in experimental and therapeutic design, a factor that is avoided using naturally occurring mitochondrial DNA barcodes [155,156,163,164]. mt-scATAC-seq allows for clonal detection of intact

cells using MT barcodes as somatic mutations. We were able to multiplex donors by adopting a method to computationally separate donors using MT somatic mutations. As with all barcoding strategies, the limitations of our study include the sensitivity of the number of clones detected to coverage. Smaller clones may go undetected if they are present in low frequencies, and those may exhibit bias. However, prior studies have suggested those are not significant contributors to haematopoiesis at the time of sampling. Our ability to observe multipotent clones across conditions in high and low cell numbers suggest this impact is limited with this method. Single-cell NGS has allowed for rapidly advancing characterization of lineage subtypes [124,139,165–168]. This experiment also enables the assignment of lineage subtypes using nuclear chromatin accessibility at gene promoters, and is corroborated by flow cytometric analysis. These findings argue against substantial lineage bias in hematopoietic stem cells in humans.

3.5 Methods

Human primary cells samples

Cryopreserved CD34 + hematopoietic stem and progenitor cells were obtained from an industrial partner (Donors 1-4, Donors A-D for FACS sorting) or StemCell Technologies (Donors 5-8). The industrial partner samples contained healthy donors aged 51-64 with verified mutations associated with clonal hematopoiesis of indeterminate potential (CHIP) [160]. Samples, where applicable, were cultured for 72 hours in a cytokine culture consisting of SCF/IL-3/IL-6/Flt3L/G-CSF/GM-CSF. The CD34 + samples were de-identified and processed in both the mtscATAC-seq library preparation and FACS sorting.

mtscATAC-seq library preparation

mt-scATAC-seq libraries were generated by adapting the 10X Genomics protocol for single cell ATAC seq, according to modifications made by Lareau et al. [156]. This modified protocol exploits a fixation step and a modified permeabilization protocol to retain mitochondria in cells for subsequent capture and library preparation (**Figure 3.1A**). Whole cells were retained following an adapted protocol of 10X

Genomics Single-cell ATAC-seq [156]. Digitonin and Tween 20 were omitted in the lysis and wash buffers to generate a higher retention of mitochondrial DNA fragments per single cell. Cells were also fixed in 1% Formaldehyde to decrease the chance of mitochondrial DNA fragments from cross contaminating upon lysis. Compared to scRNA-Seq with custom amplification of mitochondrial reads, this method achieves near complete coverage of the mitochondrial genome for optimal variant calling and cellular 'barcoding".

Each patient sample was counted using a hemocytometer as well as a (insert Tali counter name). Whole cells were obtained by following the demonstrated protocol Nuclei Isolation for Single Cell ATAC sequencing (CG000169) with modifications made by Lareau et al. and as stated here. Once samples were thawed from liquid nitrogen according to common practice. Patient samples were fixed for 10 min at room temperature in 1% Formaldehyde or 1% Paraformaldehyde in PBS, subsequently glycine was added at a concentration of 0.125 M to quench the reaction. Cells were then washed twice in PBS in a low binding Eppendorf tube and centrifuged for 5 min at 400g in 4C then recounted utilizing the same method as before. Samples were then multiplexed by taking the same amount of cells per patient sample to get a total number of cells above 100,000 to account for any cell loss. If the cell count total for all patient samples was going to be less than 100,000 after pooling the cells, we followed the "Low Cell Input Nuclei Isolation" in the appendix. Whole cells were then retained by removing Tween 20 and Digitonin from the wash and lysis buffer as done by Lareau et al. Note that the samples were incubated in lysis buffer for exactly 3 min on ice prior to washing. After centrifugation of sample at 500 rcf for 5 min, the cell supernatant was discarded and the cell pellet was resuspended in a calculated volume of 1X Diluted Nuclei buffer–The calculated volume corresponding to the previous cell concentration count obtained. Cells were counted a final time and then processed according to the Chromium Single Cell ATAC Solution user guide using the Chromium Next GEM Single Cell ATAC Library Kit, Chromium Next GEM Single Cell ATAC Gel Bead Kit, Chromium Next GEM Chip H Single Cell Kit, and Single Index

Kit N Set A. Lastly, quality control (QC) tests were run on each library prep using Agilent TapeStation High Sensitivity D1000 (Agilent) and Qubit dsDNA HS Assay kit (Invitrogen) prior to sequencing.

Processing of mt-scATAC-seq sequencing fragments

Processing of mt-scATAC-seq reads was performed similarly to [156]. Briefly, fastq files were aligned using cellranger-atac v6.1.1 to the hg38 genome**.** We used the blacklisted genome from [156] by hard-masking nuclear regions that align to the MT with single bp errors. Cell barcodes and open-chromatin peaks were called and filtered using cellranger.

Variant calling in the MT genome

We filtered cells with less than 200 bps in the MT genome and removed fragment duplicates. The coverage for those cells are shown in **Figure 3.1B.** After, we filtered for cells and positions with high quality. Specifically, we removed positions with less than ten cells with at least 50x coverage, and with less than 10 cells having 5x coverage of a putative variant at that position. Additionally, the cells required an average Phred base quality score (BQ) of over 20 at the putative variant. After this, we use MGATK filters, which remove variants with low strand concordance and low variance-mean ratio for each variant across all cells in a sample. The thresholds used were the same as in the original paper, with concordance of 0.65 and log 10 variance-mean ratio of -2.

Separating multiplexed donor cells

To separate donors from the same sequencing run, we use the algorithm Vireo [169], which is a variational bayesian inference algorithm that reconstructs each donor's allele frequency profile (the donor’s mean AF is the latent variable) and assigns a probability of each cell to that donor. Any cell with less than 0.9 probability to be assigned to a clone is removed. The algorithm also assigns a ‘doublet’ probability for each cell, which is the likelihood of the cell being part of multiple donors versus one. Cells

with more than 0.1 probability of a doublet were also removed. To ensure the donors we called were correct, we varied the number of donors in Vireo +/- 2 from the true number of donors. The model's reconstruction error is saved for each, and the 'elbow rule' is used, which finds the error's inflection point upon increasing the number of donors. Donor specific homozygous variants were calculated as having a mean AF greater than 0.9 In all our cases, the true number of donors is where the elbow occurs.

Clonal detection using MT barcodes

After computationally separating the donors, we imputed single-cell variant AF for high-coverage positions to reduce spurious clone-calling, and then re-ran MGATK, which gave a new set of called variants for each donor. To detect cells of the same clone, we used the k-nearest-neighbors leiden-based community detection algorithm, similar to [156]. The resolution parameter was set to 30, as after varying that number we find it is robust from 30-50, and the cosine distance cutoff of the algorithm was set to 3.5.

We used three different variant processing methods to merge variants across conditions, and examined internal consistency across the different methods (**Figure S3.4)**. Method 'intersect' took variants that only overlapped in both conditions (if there were any) after de-multiplexing, while method 'union' took all variants found just before running MGATK. We ultimately used union-mgatk'', which takes all variants before running MGATK, imputes high-coverage positions, and reruns MGATK on each donor. Additionally, k=30 was used in the KNN algorithm after running for k as 3, 30 and 50.

To measure consistency across workflows, for each pair of workflows, we looked at each cell pair and determined if they were either a) assigned the same clone in both methods (Positive [P]-Positive) b) assigned different clones in both methods (Negative [N]-Negative) c) assigned the same clone in one method but not the other (P-N & N-P). We find that there is concordance across the methods (**Figure S3.4A-B)**. Although the number of clones detected did change, the variant barcodes were able to be distinguished across the methods..

To calculate the percentage of cells with the barcode in a clone and outside a clone in **Figure 3.2C**, we binarized variants with a minimum of 2 reads and an AF frequency of 0.001. The top 3 variants with the highest positive difference in percentage between clones and non-clones were chosen. For **Figure 3.2D**, complete-linkage using cosine similarity was used, setting AF of >0.2 to 0.2 to improve visibility. Barcodes with an average of less than 0.01 in each clone was removed. For **SF4**, the distribution of each barcode was plotted across cells in each clone. We used a boxenplot with default parameters in seaborn v0.11.2 , which is a modified form of a boxplot that better represents the distribution for large data [170].

Processing single-cell nuclear open-chromatin regions

We next examined the peaks detected using the nuclear open-chromatin reads in each cell. Briefly, we followed the Signac (V1.4) protocol to integrate the different conditions, preprocess and binarize the cells, run latent-semantic indexing (LSI), followed by UMAP dimensionality reduction, and KNN Louvain clustering to assign cluster labels. In **Figure 3.3** and **Figure 3.4** the integration was done across the Input and 72h culture for the sequencing batch S1, while **Figure S3.5**, **Figure S3.6**, and **Figure 3.5** was done by integrating across all sequencing runs.

For open-chromatin regions, when aggregating across experimental runs, we take the detected peaks and merge them by expanding the peaks when there is overlap across runs. Peaks less than 20 bp and > 10,000 bp were removed and fragment counts re-computed. We create a Signac model and remove regions with less than 10 cells, and cells with less than 200 features. We additionally filter by keeping peaks with: a) at least 10 and less than 15,000 fragments b) with at least 15% of the nucleotides in reads found in the peak is actually covered in the peak (since a read can span the peak region and outside the region). We also keep cells a) with a nucleosome signal of at least 4 (‘i.e. the ratio of mononucleosomal to nucleosome-free fragments per cell’), b) with a TSS enrichment of at least 0.2 (as defined here:

https://www.encodeproject.org/data-standards/terms/), and c) with a ratio of reads aligned to blacklist regions over reads aligned to peaks less than 0.05.

After this, we binarize the peaks and run term frequency–inverse document frequency (TF-IDF) followed by SVD, which combined is the latent-semantic indexing method. UMAP is then run on dimensions 2-50, as the first factor correlates with depth. After this, we integrate across runs using *FindIntegrationAnchors* of the Seurat package [171] using the lsi transformed data. After integration, we run UMAP on dimensions 2 to 30 of the integrated lsi components, then cluster using *FindNeighbors* and *FindClusters* with the SLM algorithm.

Annotating cell clusters using lineage markers

Cells were annotated by taking known lineage markers of both gene activity and TF activity and overlaying the density of the feature across the UMAP embedding. Gene activity scores for each gene was calculated by summing the number of peaks found in a gene and 2 kb upstream. Feature counts for each cell are divided by the total counts for that cell, multiplied by the median gene activity in that cell, and then natural-log transformed to get the activity score. TF activity was calculated using the chromVAR [161] extension in Signac, which estimates activity based on the number of TF motifs detected in a cell's open-chromatin peaks. Manual annotation was performed on the clusters using both the gene and TF activity in known markers.

Hypergeometric test to measure lineage bias in clones

To detect clonal bias towards a specific lineage, we used the hypergeometric cumulative distribution test for each clone-cluster pair, and p-values were adjusted using Benjamini-Hochberg to control the false discovery rate. A significance threshold of 0.1 was used, but to account for clone and cluster sizes affecting the test, we created a non-parametric null distribution in which the cluster labels for each cell were shuffled 1000 times and the p-values for each clone-cluster pair computed. The p-values in

each simulation were used as a background distribution, and empirical p-values were calculated for each clone-cluster pair, a significance of p=0.1 was used in reporting significance values in **Figure 3.4**.

Clone and lineage entropy measures

In order to measure the lineage-bias a clone has, we used the entropy metric (**Figure S3.5B**). To do this, we first removed the 'HSPC' lineage clusters and calculated the frequency in each clone, which was used as the probability distribution. We then calculated the standard entropy measure using entropy from the SciPy v1.7.3 stats package [172]. To calculate the maximum entropy possible, we took the largest number of clones across donors, and calculated the entropy over a uniform distribution across those clones (so for 50 clones, the probability for each clone will be 1/50). Entropy was also calculated in each lineage, which is measuring if there are clones over- or under-represented across lineages (**Figure S3.5C**). For a single lineage, we took all the clone proportions in that lineage, and then converted them into a probability distribution by dividing by their sum.

Flow-cytometry

Flow-cytometry was done for four healthy CD34+ donors, and culturing was done as mentioned above. The markers used for the UMAPs in **Figure S3.7** were HLA-DR, CD117, CD11c, CD11b, CD34, CD10, CD45, CD86, FceRIa, CD16, CD14, CD66b, CD101, Siglec8, CD3, CD19, CD56.

3.6 Supplementary Figures and Tables

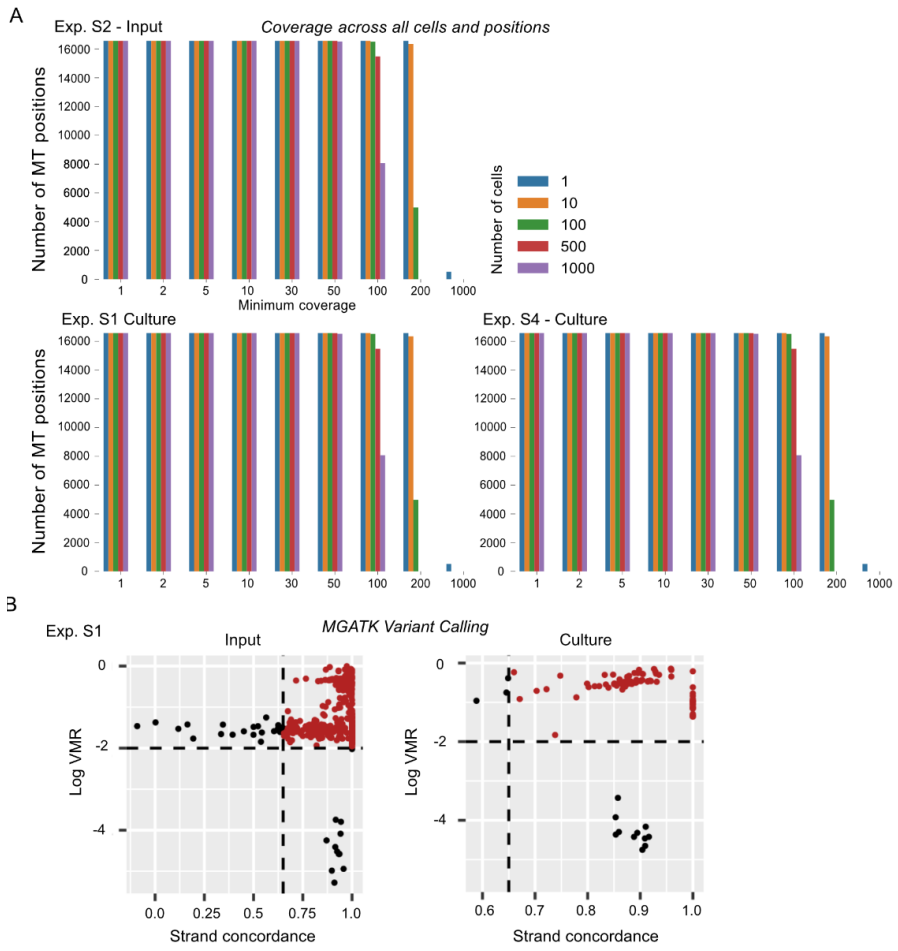


Figure S3.1 Coverage and variants called in mt-scATAC-seq experiments
(A) Detecting the number of cells with a certain number of fragments (coverage) across each position in the MT genome **(B)** MGATK algorithm used to call variants in the MT genome. Each point is a variant, and variants colored red pass the variance-mean ratio (VMR) and strand concordance thresholds.

Table S3.1 mt-scATAC-seq sequencing results

Experiment	Condition	Number of donors in group	Cells that pass nuclear open-chromatin QC	Cells that pass MT coverage QC and open-chromatin QC	Variants
S1	Cultured	2	7651	6907	75
S1	Input	2	6848	6500	197
S2	Input	2	4769	4214	107
S3	Cultured	4	12009	11203	312

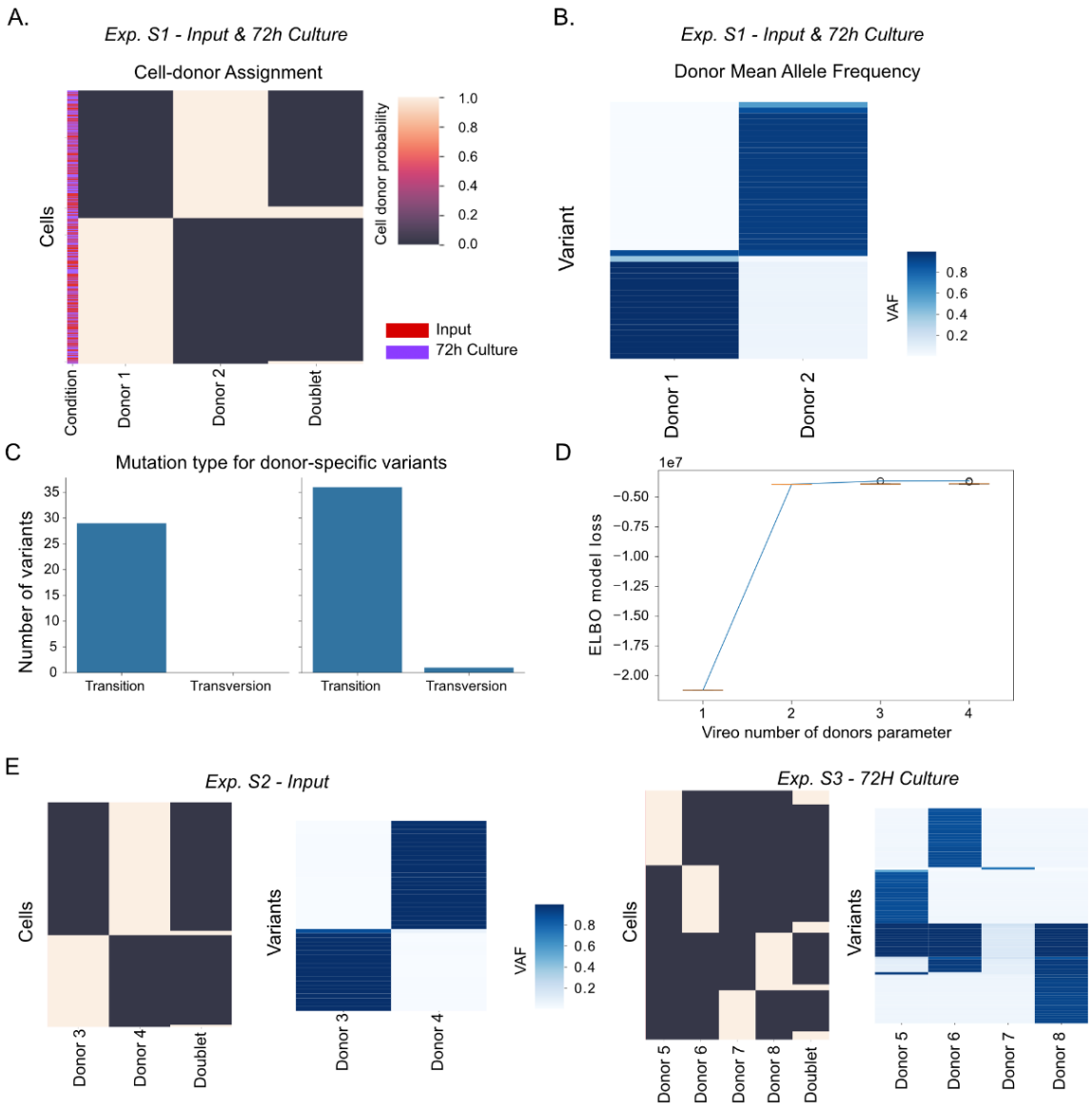


Figure S3.2. Cells are confidently assigned to a donor in a multiplexed run using germ-line MT variants
Experiment batch S1 containing 2 donors across two conditions was de-multiplexed. **(A)** The Vireo algorithm assigns a probability for each cell to a donor. This is done across all conditions for those multiplexed donors. Each row is a cell and each column is a donor. (Maximum of 1000 cells randomly chosen for visualization). For the conditions, red=input, blue=culture **(B)** The mean allele frequency for cells of the same predicted donor. The top variants in each donor is shown. **(C)** Transitions and transversions detected in each donor. Donor-specific variants defined as having greater than 0.8 VAF in over 90% of donor-assigned cells. **(D)** An elbow plot that plots the model loss ('ELBO') over varying the number of donors d. The point of decreasing model performance gains is the ideal parameter, which is 2 here **(E)** Same as (A) and (B), but for experiments batch S2 and batch S3

Table S3.2 Summary of donor

Donor variants defined as mean VAF>0.7 in donor, <0.1 in others

Donor	Experiment Batch ID	Number of cells after demultiplexing	Number of unique donor variants in batch	Verified CHIP mutations
Donor 1	S1	7002	17	Y
Donor 2	S1	5691	25	Y
Donor 3	S2	1730	19	Y
Donor 4	S2	2367	25	Y
Donor 5	S4	2630	20	N
Donor 6	S4	2605	23	N
Donor 7	S4	1675	1	N
Donor 8	S4	2414	19	N

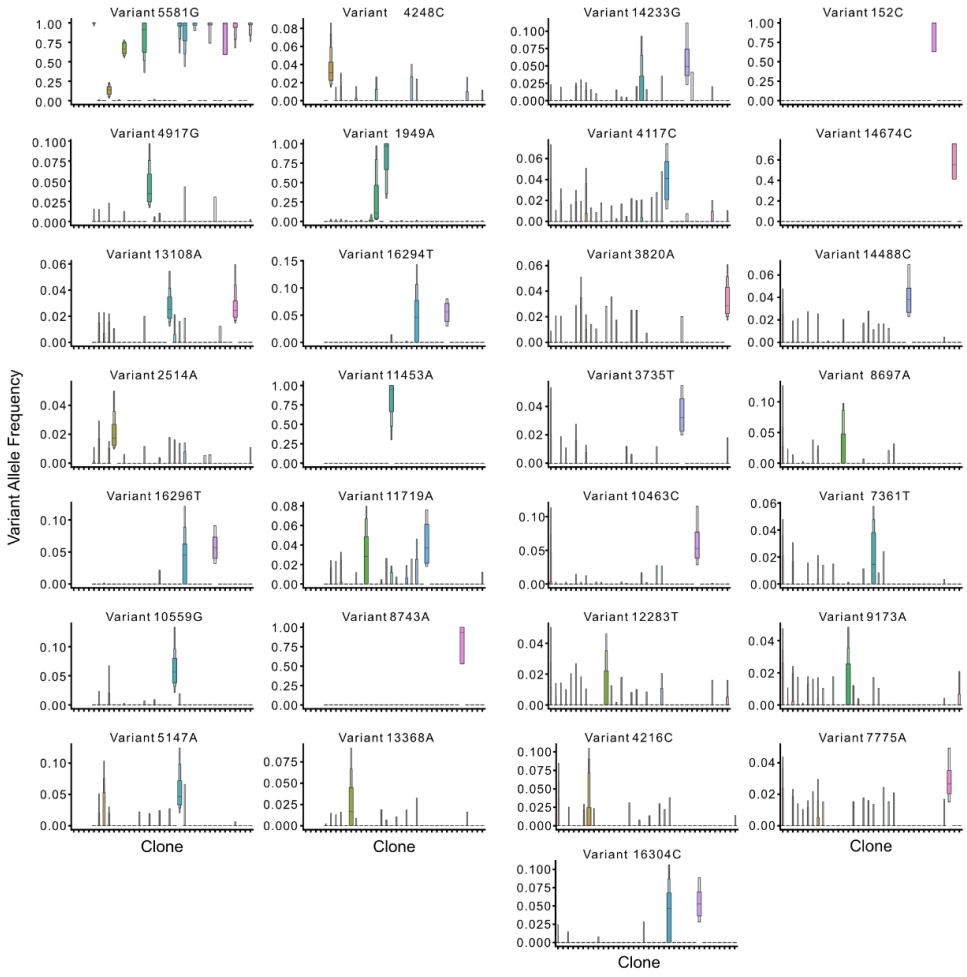
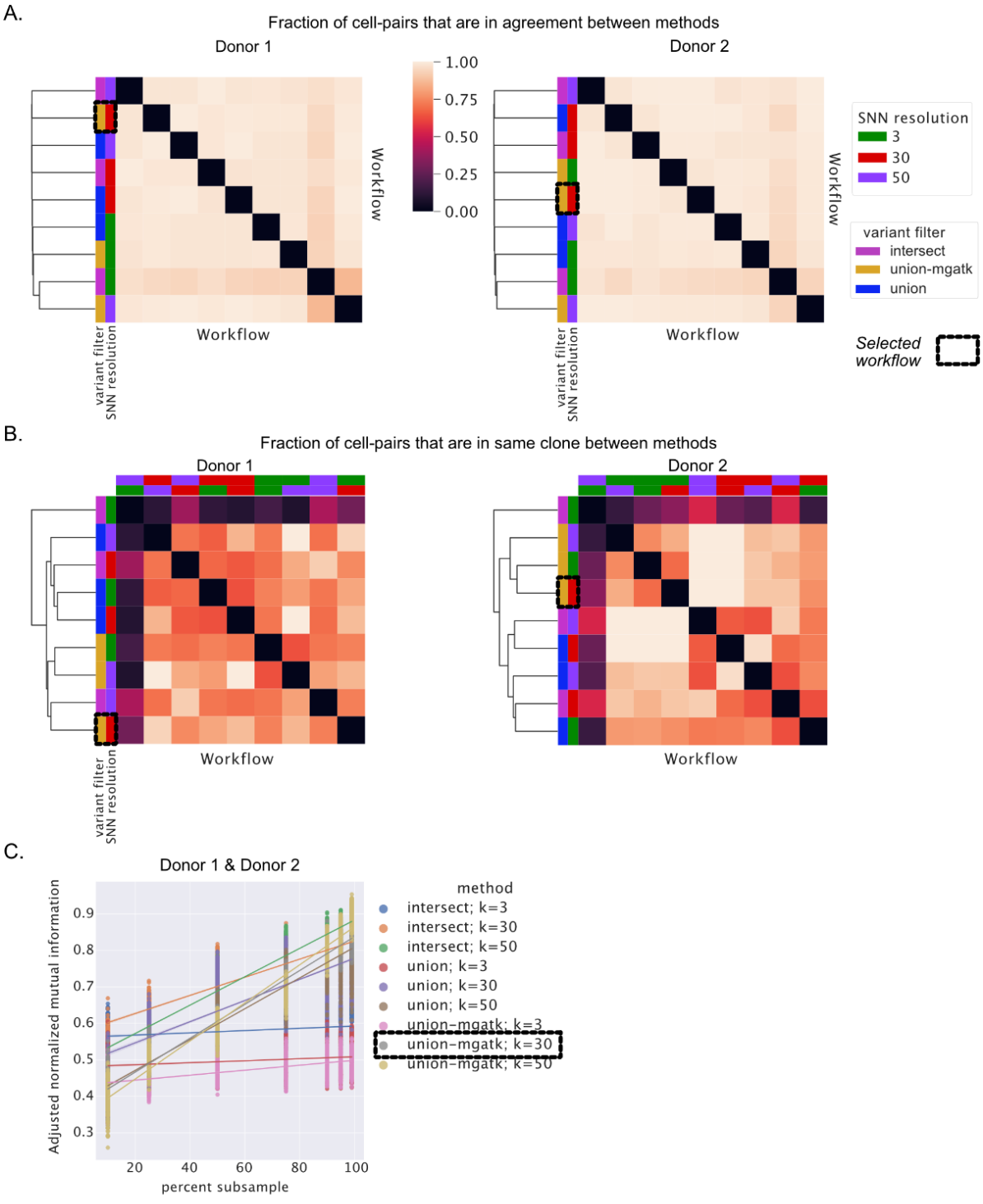


Figure S3.3. Variant allele frequency distribution in cells across clones VAF distribution in cells in each clone across variants used to detect clones in Donor 1

Figure S3.4. Comparing clone-calling workflows

Methods 'intersect', 'union', 'union-mgatk', (see Methods) and k nearest neighbor of 3, 30, and 50 were compared **(A)** Comparing each method by finding the number of cell pairs that are either in the same clone or in different clones in both methods. The score is then normalized to the total number of cell pairs. **(B)** Same as a, but only looking at scores in which both methods assign the pair of cells to the same clone. **(C)** Subsampling cells from 10-99% 1000 times, calculating adjusted normalized mutual information between clones in sub-sampled run and the full population.


A.
Fraction of cell-pairs that are in agreement between methods
Donor 1
Donor 2
1.00
0.75
0.50
0.25
0.00
Workflow
Workflow
variant filter
SNN resolution
SNN resolution
3
30
50
variant filter
intersect
union-mgatk
union
Selected workflow
B.
Fraction of cell-pairs that are in same clone between methods
Donor 1
Donor 2
Workflow
Workflow
variant filter
SNN resolution
C.
Donor 1 & Donor 2
Adjusted normalized mutual information
0.9
0.8
0.7
0.6
0.5
0.4
0.3
20
40
60
80
100
percent subsample
method
intersect; k=3
intersect; k=30
intersect; k=50
union; k=3
union; k=30
union; k=50
union-mgatk; k=3
union-mgatk; k=30
union-mgatk; k=50

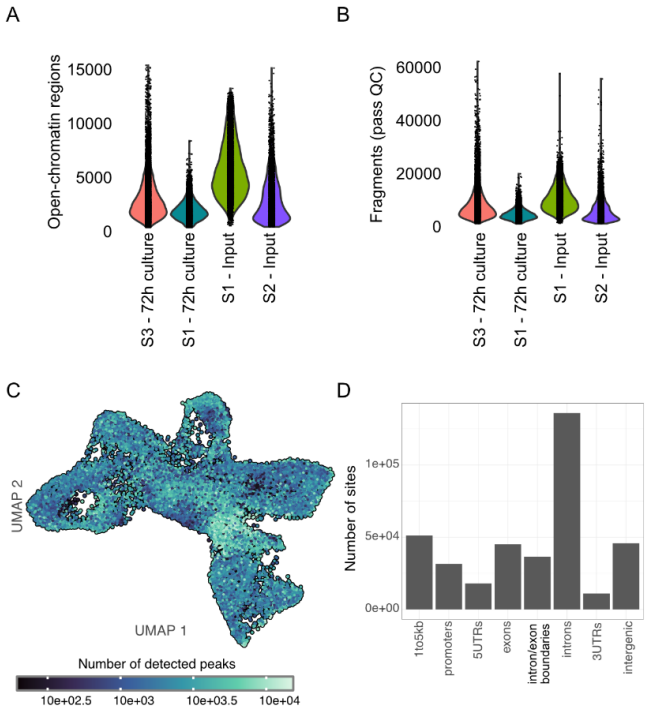


Figure S3.5. Detecting nuclear open-chromatin peaks

(A) Number of detected open-chromatin peak regions per cell (B) Number of fragments that are non-duplicated and pass QC filters (see Methods) (C) Number of peaks detected for each cell overlaid on UMAP (D) Annotating peaks based on location relative to a gene.

Table S3.3 Lineage markers used to inform cell lineage cluster assignment

Lineage Type	Markers		
Stem—>progenitor cells	KIT	SPI1	
Myeloid lineage	FLT3	ATP8B4	
Erythroid lineage	CPOX	ANK1	
B cell lineage	HBB	HBG2	
pre-Bcells	VPREB1	CD81	CD79B
Monocyte lineage	IRF8	CTSZ	LYZ
pDC	KCNK1	PROC	NCR2
Basophil (granulocytes)	IL18R1	IL1RL1	IL5RA
Eosinophil (granulocytes)	CLC	PRG2	
Neutrophil	MPO	ELANE	PRTN3

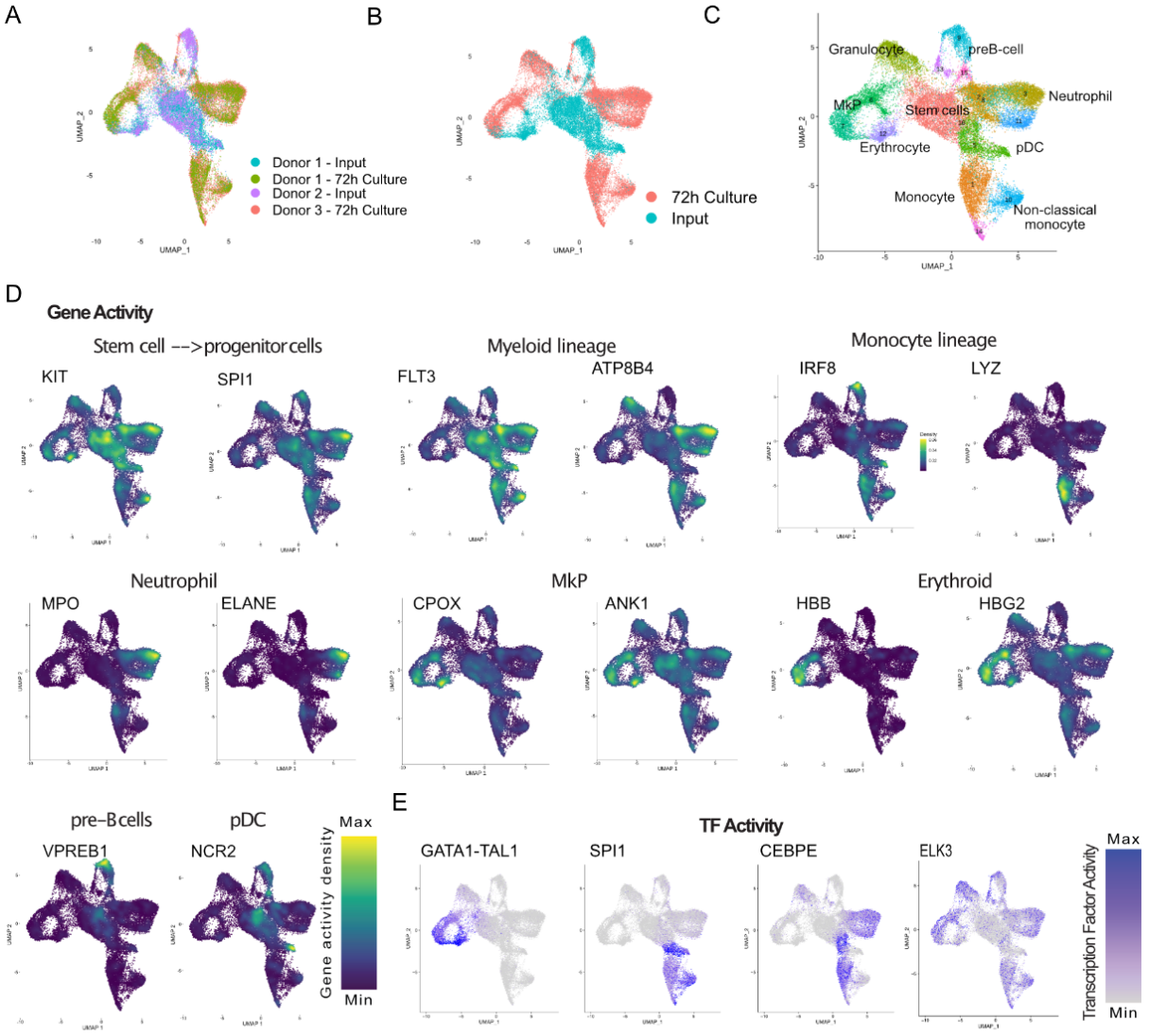


Figure S3.6. Characterizing cells by lineage markers in nuclear open-chromatin peaks across all donors
(A) Each batch sequencing run overlaid on UMAP (B) UMAP colored by condition. (C) Cells colored by clusters detected by Seurat's SNN method, and UMAP annotation labels overlaid on UMAP (D) Gene activity scores for select markers overlaid on UMAP (E) Transcription factor activity scores for select markers (F) Pseudotime values and trajectory using Monocle 3

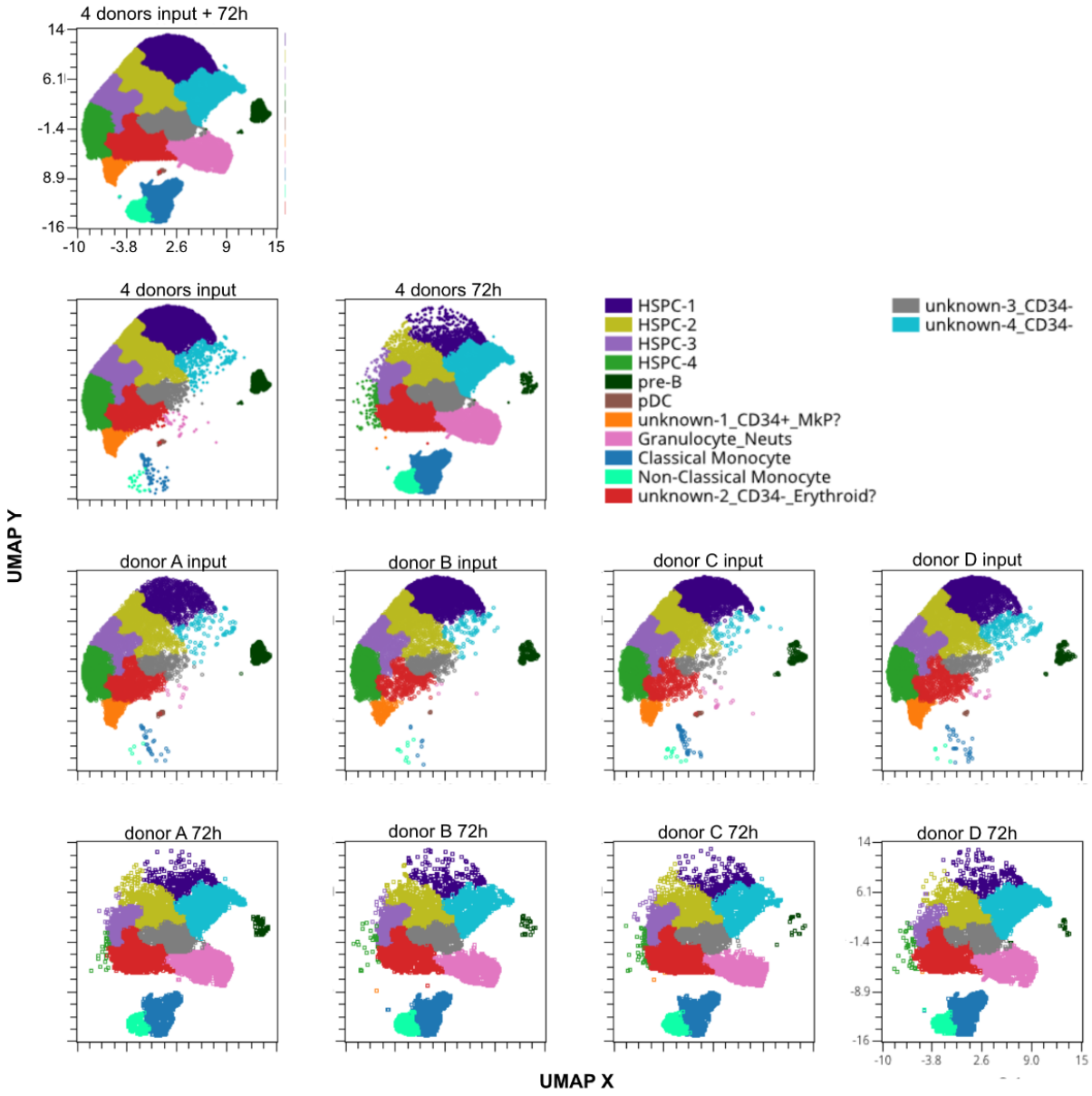
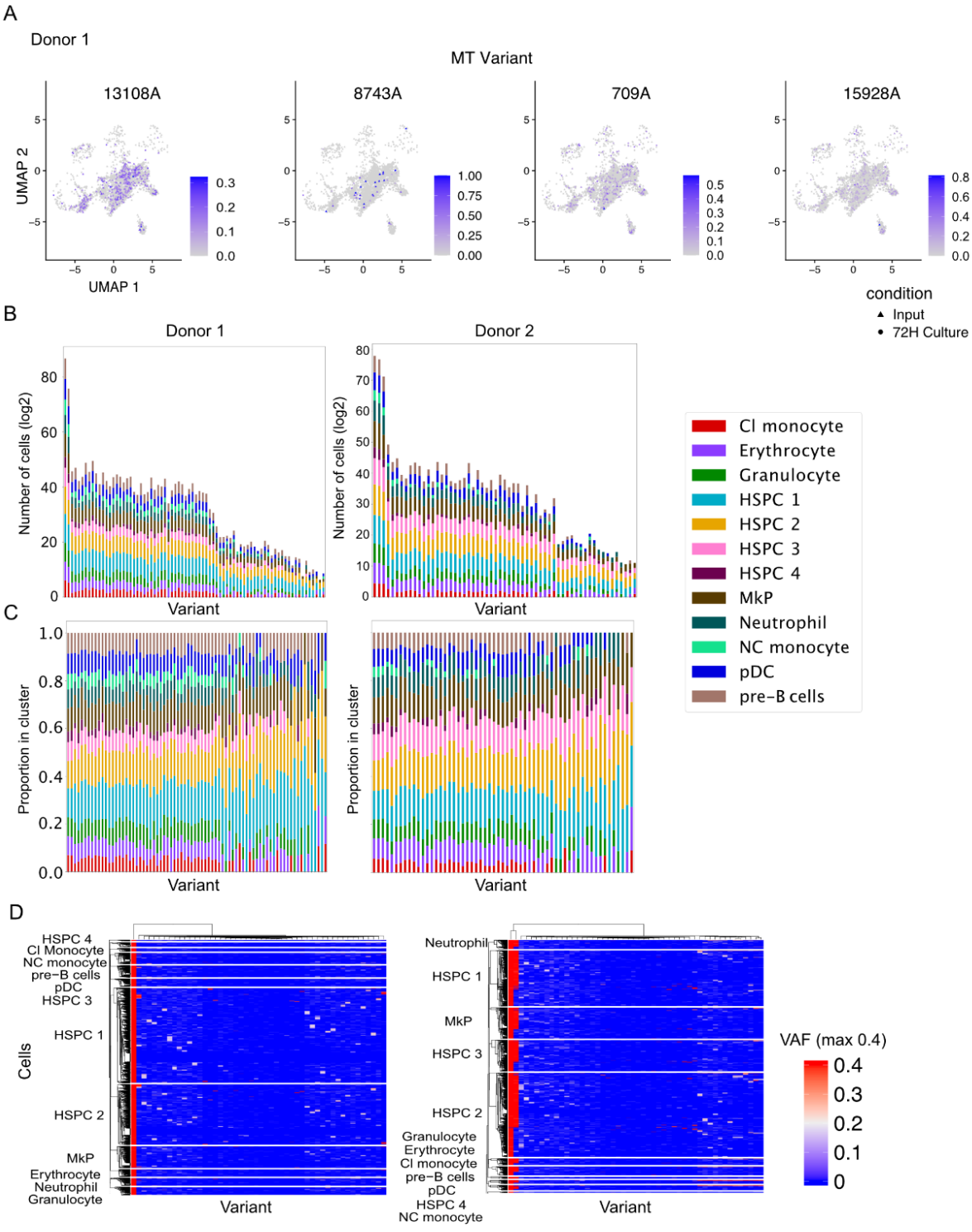


Figure S3.7 FACS sorting highlights differentiated lineages in CD34+ HSPCs after cytokine culture
UMAP of fluorescent markers enables lineage cluster detection in 4 donors in both input and cytokine culture. Top 2 rows: UMAP for all donors combined; bottom 2 rows: each donor separate

Figure S3.8. MT barcodes across lineage clusters
(A) MT variants across UMAP for selected variants in donor 1 **(B-D)** Donor 1 and Donor 2 shown **(B)** Total cell counts (log2) for barcodes (af>0.01, coverage>10) across hematopoietic clusters. **(C)** Similar to B, but normalized within each variant. **(D)** Cell-by-variant VAF heatmap for top differentiating variants in donors Donor 1 and Donor 2 ordered by single-linkage hierarchical clustering within each HSPC type followed by clustering on HSPC types.


A
Donor 1
MT Variant
13108A
8743A
709A
15928A
UMAP 2
UMAP 1
condition
Input
72H Culture
B
Donor 1
Donor 2
Number of cells (log2)
Variant
Cl monocyte
Erythrocyte
Granulocyte
HSPC 1
HSPC 2
HSPC 3
HSPC 4
MkP
Neutrophil
NC monocyte
pDC
pre-B cells
C
Proportion in cluster
Variant
D
Cells
HSPC 4
Cl Monocyte
NC monocyte
pre-B cells
pDC
HSPC 3
HSPC 1
HSPC 2
MkP
Erythrocyte
Neutrophil
Granulocyte
Variant
Neutrophil
HSPC 1
MkP
HSPC 3
HSPC 2
Granulocyte
Erythrocyte
Cl monocyte
pre-B cells
pDC
HSPC 4
NC monocyte
Variant
VAF (max 0.4)

3.7 Data Availability

Raw sequencing files will be submitted to NCBI's SRA upon acceptance of the manuscript for publication.

3.8 Acknowledgements

Chapter 3, in full, is materials in preparation for submission of the manuscript "Comparative single-cell lineage bias in human and murine hematopoietic stem cells," Isaac Shamie*, Meghan Bliss-Moreau*, Jamie Casey Lee*, Nathan E. Lewis, Yanfang Peipei Zhu, Ben A. Croker. The dissertation author was the primary investigator and author of this paper.

This work was supported by NIH Grant 5RO1HL124209, R35GM119850, the V Foundation for Cancer Research, Gilead, an American Heart Association CDA (to YPZ) and a T32 (5T32HL007574-36 to MBM).

Disclosure of Conflicts of Interest

The authors declare no competing financial or non-financial conflicts of interest.

CONCLUSION

In this dissertation, I argue for using more novel NGS techniques to help push our understanding of the production of biotherapeutics as well as tracking cellular response. This is done through establishing resources for CHO cells and CD34+ HSPCs in healthy donors. I show that transcription start site sequencing methods used in CHO cells and the Chinese hamster can be used to improve the genome annotation, and that these can be used for gene activation in CHO cells. Additionally, examining CD34+ cells in steady-state and *ex vivo* culture, we find that clones exhibit multipotency, suggesting the detected clones come from early HSCs, and treatment should lead to balanced response across different clones.

REFERENCES

1. Walsh, G. Biopharmaceutical benchmarks 2018. *Nature Publishing Group* **36,** 1136–1145 (2018).

2. Kuo, C.-C., Chiang, A. W. T., Shamie, I., Samoudi, M., Gutierrez, J. M. & Lewis, N. E. The emerging role of systems biology for engineering protein production in CHO cells. *Curr. Opin. Biotechnol.* **51,** 64–69 (2018).

3. Durocher, Y. & Butler, M. Expression systems for therapeutic glycoprotein production. *Curr. Opin. Biotechnol.* **20,** 700–707 (2009).

4. Altman, L. K. A NEW INSULIN GIVEN APPROVAL FOR USE IN U.S. *The New York Times* (1982). at <https://www.nytimes.com/1982/10/30/us/a-new-insulin-given-approval-for-use-in-us.html>

5. Landgraf, W., Medical Affairs Diabetes Division, Frankfurt, S.-A., Germany, Sandow, J., Centre of Pharmacology, Johann-Wolfgang-Goethe University, Frankfurt/Main & Germany. Recombinant Human Insulins – Clinical Efficacy and Safety in Diabetes Therapy. *European Endocrinology* **12,** 12 Preprint at https://doi.org/10.17925/ee.2016.12.01.12 (2016)

6. Biotherapeutic products. at <https://www.who.int/teams/health-product-policy-and-standards/standards-and-specifications/biotherapeutic-products>

7. Kanter, J. & Falcon, C. Gene therapy for sickle cell disease: where we are now? *Hematology Am. Soc. Hematol. Educ. Program* **2021,** 174–180 (2021).

8. Fda, U. S. Fact sheet: FDA at a glance. Preprint at https://www.fda.gov/about-fda/fda-basics/fact-sheet-fda-glance (2019)

9. Kay, E., Cuccui, J. & Wren, B. W. Recent advances in the production of recombinant glycoconjugate vaccines. *NPJ Vaccines* **4,** 16 (2019).

10. Beck, A., Goetsch, L., Dumontet, C. & Corvaïa, N. Strategies and challenges for the next generation of antibody–drug conjugates. *Nat. Rev. Drug Discov.* **16,** 315–337 (2017).

11. Barone, P. W., Wiebe, M. E., Leung, J. C., Hussein, I. T. M., Keumurian, F. J., Bouressa, J.,

Brussel, A., Chen, D., Chong, M., Dehghani, H., Gerentes, L., Gilbert, J., Gold, D., Kiss, R., Kreil, T. R., Labatut, R., Li, Y., Müllberg, J., Mallet, L., Menzel, C., Moody, M., Monpoeho, S., Murphy, M., Plavsic, M., Roth, N. J., Roush, D., Ruffing, M., Schicho, R., Snyder, R., Stark, D., Zhang, C., Wolfrum, J., Sinskey, A. J. & Springs, S. L. Viral contamination in biologic manufacture and implications for emerging therapies. *Nat. Biotechnol.* **38,** 563–572 (2020).

12. Silverman, A. D., Karim, A. S. & Jewett, M. C. Cell-free gene expression: an expanded repertoire of applications. *Nat. Rev. Genet.* **21,** 151–170 (2020).

13. Ozdemir, T., Fedorec, A. J. H., Danino, T. & Barnes, C. P. Synthetic Biology and Engineered Live Biotherapeutics: Toward Increasing System Complexity. *Cell Systems* **7,** 5–16 Preprint at https://doi.org/10.1016/j.cels.2018.06.008 (2018)

14. Gorovits, B. & Krinos-Fiorotti, C. Proposed mechanism of off-target toxicity for antibody--drug conjugates driven by mannose receptor uptake. *Cancer Immunol. Immunother.* **62,** 217–223 (2013).

15. Riley, R. S., June, C. H., Langer, R. & Mitchell, M. J. Delivery technologies for cancer immunotherapy. *Nat. Rev. Drug Discov.* **18,** 175–196 (2019).

16. Wculek, S. K., Cueto, F. J., Mujal, A. M., Melero, I., Krummel, M. F. & Sancho, D. Dendritic cells in cancer immunology and immunotherapy. *Nat. Rev. Immunol.* **20,** 7–24 (2020).

17. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30,** 418–426 (2014).

18. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nat. Methods* **5,** 16–18 (2008).

19. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17,** 333–351 (2016).

20. Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X.-Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A. & Rothberg, J. M. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).

21. Tripathi, P., Singh, J., Lal, J. A. & Tripathi, V. Next-Generation Sequencing: An Emerging Tool for Drug Designing. *Curr. Pharm. Des.* **25,** 3350–3357 (2019).

22. Lauschke, V. M. & Ingelman-Sundberg, M. Prediction of drug response and adverse drug

reactions: From twin studies to Next Generation Sequencing. *Eur. J. Pharm. Sci.* **130,** 65–77 (2019).

23. Puck, T. T., Cieciura, S. J. & Robinson, A. GENETICS OF SOMATIC MAMMALIAN CELLS. *The Journal of Experimental Medicine* **108,** 945–956 Preprint at https://doi.org/10.1084/jem.108.6.945 (1958)

24. Golabgir, A., Gutierrez, J. M., Hefzi, H., Li, S., Palsson, B. O., Herwig, C. & Lewis, N. E. Quantitative feature extraction from the Chinese hamster ovary bioprocess bibliome using a novel meta-analysis workflow. *Biotechnol. Adv.* **34,** 621–633 (2015).

25. Jenkins, N., Parekh, R. B. & James, D. C. Getting the glycosylation right: implications for the biotechnology industry. *Nat. Biotechnol.* **14,** 975–981 (1996).

26. Popp, O., Moser, S., Zielonka, J., Rüger, P., Hansen, S. & Plöttner, O. Development of a pre-glycoengineered CHO-K1 host cell line for the expression of antibodies with enhanced Fc mediated effector function. *MAbs* **10,** 290–303 (2018).

27. Karottki, K. J. la C., la Cour Karottki, K. J., Hefzi, H., Xiong, K., Shamie, I., Hansen, A. H., Li, S., Pedersen, L. E., Li, S., Lee, J. S., Lee, G. M., Kildegaard, H. F. & Lewis, N. E. Awakening dormant glycosyltransferases in CHO cells with CRISPRa. *Biotechnology and Bioengineering* **117,** 593–598 Preprint at https://doi.org/10.1002/bit.27199 (2020)

28. Xu, X., Nagarajan, H., Lewis, N. E., Pan, S., Cai, Z., Liu, X., Chen, W., Xie, M., Wang, W., Hammond, S., Andersen, M. R., Neff, N., Passarelli, B., Koh, W., Fan, H. C., Wang, J., Gui, Y., Lee, K. H., Betenbaugh, M. J., Quake, S. R., Famili, I., Palsson, B. O. & Wang, J. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat. Biotechnol.* **29,** 735–741 (2011).

29. Brinkrolf, K., Rupp, O., Laux, H., Kollin, F., Ernst, W., Linke, B., Kofler, R., Romand, S., Hesse, F., Budach, W. E., Galosy, S., Müller, D., Noll, T., Wienberg, J., Jostock, T., Leonard, M., Grillari, J., Tauch, A., Goesmann, A., Helk, B., Mott, J. E., Pühler, A. & Borth, N. Chinese hamster genome sequenced from sorted chromosomes. *Nat. Biotechnol.* **31,** 694–695 (2013).

30. Rupp, O., MacDonald, M. L., Li, S., Dhiman, H., Polson, S., Griep, S., Heffner, K., Hernandez, I., Brinkrolf, K., Jadhav, V., Samoudi, M., Hao, H., Kingham, B., Goesmann, A., Betenbaugh, M. J., Lewis, N. E., Borth, N. & Lee, K. H. A reference genome of the Chinese hamster based on a hybrid assembly strategy. *Biotechnol. Bioeng.* (2018). doi:10.1002/bit.26722

31. Nguyen, L. N., Novak, N., Baumann, M., Koehn, J. & Borth, N. Bioinformatic identification of Chinese hamster ovary (CHO) cold-shock genes and biological evidence of their cold-inducible promoters. *Biotechnol. J.* **15,** e1900359 (2020).

32. Ritter, A., Rauschert, T., Oertli, M., Piehlmaier, D., Mantas, P., Kuntzelmann, G., Lageyre, N., Brannetti, B., Voedisch, B., Geisse, S., Jostock, T. & Laux, H. Disruption of the gene C12orf35 leads to increased productivities in recombinant CHO cell lines. *Biotechnol. Bioeng.* **113,** 2433–2442 (2016).

33. Laux, H., Romand, S., Nuciforo, S., Farady, C. J., Tapparel, J., Buechmann-Moeller, S., Sommer, B., Oakeley, E. J. & Bodendorf, U. Degradation of recombinant proteins by Chinese hamster ovary host cell proteases is prevented by matriptase‐1 knockout. *Biotechnol. Bioeng.* **115,** 2530–2540 (2018).

34. Hefzi, H., Ang, K. S., Hanscho, M., Bordbar, A., Ruckerbauer, D., Lakshmanan, M., Orellana, C. A., Baycin-Hizal, D., Huang, Y., Ley, D., Martinez, V. S., Kyriakopoulos, S., Jiménez, N. E., Zielinski, D. C., Quek, L. E., Wulff, T., Arnsdorf, J., Li, S., Lee, J. S., Paglia, G., Loira, N., Spahn, P. N., Pedersen, L. E., Gutierrez, J. M., King, Z. A., Lund, A. M., Nagarajan, H., Thomas, A., Abdel-Haleem, A. M., Zanghellini, J., Kildegaard, H. F., Voldborg, B. G., Gerdtzen, Z. P., Betenbaugh, M. J., Palsson, B. O., Andersen, M. R., Nielsen, L. K., Borth, N., Lee, D. Y. & Lewis, N. E. A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. *Cell Systems* **3,** 434–443.e8 (2016).

35. Tang, D., Subramanian, J., Haley, B., Baker, J., Luo, L., Hsu, W., Liu, P., Sandoval, W., Laird, M. W., Snedecor, B., Shiratori, M. & Misaghi, S. Pyruvate Kinase Muscle‐1 Expression Appears to Drive Lactogenic Behavior in CHO Cell Lines, Triggering Lower Viability and Productivity: A Case Study. *Biotechnology Journal* **14,** 1800332 Preprint at https://doi.org/10.1002/biot.201800332 (2019)

36. Xiong, K., Marquart, K. F., la Cour Karottki, K. J., Li, S., Shamie, I., Lee, J. S., Gerling, S., Yeo, N. C., Chavez, A., Lee, G. M., Lewis, N. E. & Kildegaard, H. F. Reduced apoptosis in Chinese hamster ovary cells via optimized CRISPR interference. *Biotechnol. Bioeng.* **116,** 1813–1819 (2019).

37. Karottki, K. J. la C., Hefzi, H., Xiong, K., Shamie, I., Hansen, A. H., Li, S., Pedersen, L. E., Li, S., Lee, J. S., Lee, G. M. & Others. Awakening dormant glycosyltransferases in CHO cells with CRISPRa. *Biotechnol. Bioeng.* **117,** 593–598 (2020).

38. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346,** 1258096 (2014).

39. Horlbeck, M. A., Gilbert, L. A., Villalta, J. E., Adamson, B., Pak, R. A., Chen, Y., Fields, A. P., Park, C. Y., Corn, J. E., Kampmann, M. & Weissman, J. S. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife* **5,** (2016).

40. 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., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A. & Weissman, J. S. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167,** 1867–1882.e21 (2016).

41. Waldmann, T. A. Cytokines in Cancer Immunotherapy. *Cold Spring Harb. Perspect. Biol.* **10,** (2018).

42. Sim, G. C. & Radvanyi, L. The IL-2 cytokine family in cancer immunotherapy. *Cytokine Growth Factor Rev.* **25,** 377–390 (2014).

43. Jiang, T., Zhou, C. & Ren, S. Role of IL-2 in cancer immunotherapy. *Oncoimmunology* **5,** e1163462 (2016).

44. Cueto, F. J. & Sancho, D. The Flt3L/Flt3 Axis in Dendritic Cell Biology and Cancer Immunotherapy. *Cancers* **13,** (2021).

45. Tsapogas, P., Mooney, C. J., Brown, G. & Rolink, A. The Cytokine Flt3-Ligand in Normal and Malignant Hematopoiesis. *Int. J. Mol. Sci.* **18,** (2017).

46. Wang, W., Zhang, Y., Dettinger, P., Reimann, A., Kull, T., Loeffler, D., Manz, M. G., Lengerke, C. & Schroeder, T. Cytokine combinations for human blood stem cell expansion induce cell-type– and cytokine-specific signaling dynamics. *Blood* **138,** 847–857 Preprint at https://doi.org/10.1182/blood.2020008386 (2021)

47. Siena, S., Schiavo, R., Pedrazzoli, P. & Carlo-Stella, C. Therapeutic relevance of CD34 cell dose in blood cell transplantation for cancer therapy. *J. Clin. Oncol.* **18,** 1360–1377 (2000).

48. Tanhehco, Y. C. & Bhatia, M. Hematopoietic stem cell transplantation and cellular therapy in sickle cell disease: where are we now? *Curr. Opin. Hematol.* **26,** 448–452 (2019).

49. Park, D. S., Akuffo, A. A., Muench, D. E., Grimes, H. L., Epling-Burnette, P. K., Maini, P. K., Anderson, A. R. A. & Bonsall, M. B. Clonal hematopoiesis of indeterminate potential and its impact on patient trajectories after stem cell transplantation. *PLoS Comput. Biol.* **15,** e1006913 (2019).

50. Steensma, D. P. Clinical consequences of clonal hematopoiesis of indeterminate potential. *Hematology Am. Soc. Hematol. Educ. Program* **2018,** 264–269 (2018).

51. Valent, P., Kern, W., Hoermann, G., Milosevic Feenstra, J. D., Sotlar, K., Pfeilstöcker, M., Germing, U., Sperr, W. R., Reiter, A., Wolf, D., Arock, M., Haferlach, T. & Horny, H.-P. Clonal Hematopoiesis with Oncogenic Potential (CHOP): Separation from CHIP and Roads to AML. *Int. J. Mol. Sci.* **20,** (2019).

52. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21,** 410–427 (2020).

53. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322,** 1845–1848 Preprint at https://doi.org/10.1126/science.1162228 (2008)

54. Lam, M. T. Y., Cho, H., Lesch, H. P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M. U., Kim, A. S., Kosaka, M., Lee, C. Y., Watt, A., Grossman, T. R., Rosenfeld, M. G., Evans, R. M. & Glass, C. K. Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* **498,** 511–515 (2013).

55. Duttke, S. H., Chang, M. W., Heinz, S. & Benner, C. Identification and dynamic quantification of regulatory elements using total RNA. *Genome Res.* **29,** 1836–1846 (2019).

56. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10,** 1213–1218 (2013).

57. Urquhart, L. Top companies and drugs by sales in 2020. *Nat. Rev. Drug Discov.* **20,** 253 (2021).

58. Kim, C. L. & Lee, G. M. Improving recombinant bone morphogenetic protein-4 (BMP-4) production by autoregulatory feedback loop removal using BMP receptor-knockout CHO cell lines. *Metab. Eng.* **52,** 57–67 (2019).

59. Kol, S., Ley, D., Wulff, T., Decker, M., Arnsdorf, J., Schoffelen, S., Hansen, A. H., Jensen, T. L., Gutierrez, J. M., Chiang, A. W. T., Masson, H. O., Palsson, B. O., Voldborg, B. G., Pedersen, L. E., Kildegaard, H. F., Lee, G. M. & Lewis, N. E. Multiplex secretome engineering enhances recombinant protein production and purity. *Nat. Commun.* **11,** 1908 (2020).

60. Karottki, K. J. la C., la Cour Karottki, K. J., Hefzi, H., Li, S., Pedersen, L. E., Spahn, P., Joshi, C., Ruckerbauer, D., Bort, J. H., Thomas, A., Lee, J. S., Borth, N., Lee, G. M., Kildegaard, H. F. & Lewis, N. E. A metabolic CRISPR-Cas9 screen in Chinese hamster ovary cells identifies glutamine-sensitive genes. Preprint at https://doi.org/10.1101/2020.05.07.081604

61. Christian, M., Cermak, T., Doyle, E. L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A. J. & Voytas, D. F. Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186,** 757–761 (2010).

62. Gallego-Bartolomé, J., Liu, W., Kuo, P. H., Feng, S., Ghoshal, B., Gardiner, J., Zhao, J. M.-C., Park, S. Y., Chory, J. & Jacobsen, S. E. Co-targeting RNA Polymerases IV and V

Promotes Efficient De Novo DNA Methylation in Arabidopsis. *Cell* **176,** 1068–1082.e19 (2019).

63. 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., Lim, W. A., Weissman, J. S. & Qi, L. S. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154,** 442–451 (2013).

64. Chavez, A., Tuttle, M., Pruitt, B. W., Ewen-Campen, B., Chari, R., Ter-Ovanesyan, D., Haque, S. J., Cecchi, R. J., Kowal, E. J. K., Buchthal, J., Housden, B. E., Perrimon, N., Collins, J. J. & Church, G. Comparison of Cas9 activators in multiple species. *Nat. Methods* **13,** 563–567 (2016).

65. Jakobi, T., Brinkrolf, K., Tauch, A., Noll, T., Stoye, J., Pühler, A. & Goesmann, A. Discovery of transcription start sites in the Chinese hamster genome by next-generation RNA sequencing. *J. Biotechnol.* **190,** 64–75 (2014).

66. Link, V. M., Duttke, S. H., Chun, H. B., Holtman, I. R., Westin, E., Hoeksema, M. A., Abe, Y., Skola, D., Romanoski, C. E., Tao, J., Fonseca, G. J., Troutman, T. D., Spann, N. J., Strid, T., Sakai, M., Yu, M., Hu, R., Fang, R., Metzler, D., Ren, B. & Glass, C. K. Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function. *Cell* **173,** 1796–1809.e17 (2018).

67. Hetzel, J., Duttke, S. H., Benner, C. & Chory, J. Nascent RNA sequencing reveals distinct features in plant transcription. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 12316–12321 (2016).

68. Duttke, S. H. C., Lacadie, S. A., Ibrahim, M. M., Glass, C. K., Corcoran, D. L., Benner, C., Heinz, S., Kadonaga, J. T. & Ohler, U. Human promoters are intrinsically directional. *Mol. Cell* **57,** 674–684 (2015).

69. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2 (-Delta Delta C (T)) method. methods. 25: 402–408. 2001. *View Article: Google Scholar: PubMed/NCBI*

70. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).

71. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

72. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

73. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum,

C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).

74. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *aoas* **5,** 1752–1779 (2011).

75. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (2010).

76. Li, S., Cha, S. W., Heffner, K., Hizal, D. B., Bowen, M. A., Chaerkady, R., Cole, R. N., Tejwani, V., Kaushik, P., Henry, M., Meleady, P., Sharfstein, S. T., Betenbaugh, M. J., Bafna, V. & Lewis, N. E. Proteogenomic Annotation of Chinese Hamsters Reveals Extensive Novel Translation Events and Endogenous Retroviral Elements. *J. Proteome Res.* **18,** 2433–2445 (2019).

77. Shamie, I., Duttke, S. H., Karottki, K. J. la C., Han, C. Z., Hansen, A. H., Hefzi, H., Xiong, K., Li, S., Roth, S. J., Tao, J. & Others. A Chinese hamster transcription start site atlas that enables targeted editing of CHO cells. *NAR genomics and bioinformatics* **3,** lqab061 (2021).

78. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14,** 417–419 (2017).

79. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res.* **43,** W39–49 (2015).

80. Togayachi, A., Dae, K.-Y., Shikanai, T. & Narimatsu, H. in *Experimental Glycoscience: Glycobiology* (eds. Taniguchi, N., Suzuki, A., Ito, Y., Narimatsu, H., Kawasaki, T. & Hase, S.) 423–425 (Springer Japan, 2008).

81. Young, R. S., Hayashizaki, Y., Andersson, R., Sandelin, A., Kawaji, H., Itoh, M., Lassmann, T., Carninci, P., Bickmore, W. A., Forrest, A. R. & Others. The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Res.* **25,** 1546–1557 (2015).

82. Wurm, F. M. CHO Quasispecies—Implications for Manufacturing Processes. *Processes* **1,** 296–311 (2013).

83. Field, A. & Adelman, K. Evaluating Enhancer Function and Transcription. *Annu. Rev. Biochem.* **89,** 213–234 (2020).

84. Halfon, M. S. Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. *Trends Genet.* **35,** 93–103 (2019).

85. Fejes-Toth, K., Sotirova, V., Sachidanandam, R., Assaf, G., Hannon, G. J., Kapranov, P., Foissac, S., Willingham, A. T., Duttagupta, R., Dumais, E. & Gingeras, T. R. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457,** 1028–1032 (2009).

86. Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., Young, R. A. & Sharp, P. A. Divergent transcription from active promoters. *Science* **322,** 1849–1851 (2008).

87. Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72,** 449–479 (2003).

88. Danino, Y. M., Even, D., Ideses, D. & Juven-Gershon, T. The core promoter: At the heart of gene expression. *Biochim. Biophys. Acta* **1849,** 1116–1131 (2015).

89. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* **19,** 621–637 (2018).

90. Grosveld, G. C., Shewmaker, C. K., Jat, P. & Flavell, R. A. Localization of DNA sequences necessary for transcription of the rabbit beta-globin gene in vitro. *Cell* **25,** 215–226 (1981).

91. Huang, C. Y., Duttke, S. H. C., Kadonaga, J. T. & Others. The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes Dev.* **31,** 6–11 (2017).

92. Smale, S. T. & Baltimore, D. The 'initiator' as a transcription control element. *Cell* **57,** 103–113 (1989).

93. Feichtinger, J., Hern??ndez, I., Fischer, C., Hanscho, M., Auer, N., Hackl, M., Jadhav, V., Baumann, M., Krempl, P. M., Schmidl, C., Farlik, M., Schuster, M., Merkel, A., Sommer, A., Heath, S., Rico, D., Bock, C., Thallinger, G. G. & Borth, N. Comprehensive genome and epigenome characterization of CHO cells in response to evolutionary pressures and over time. *Biotechnol. Bioeng.* **113,** 2241–2253 (2016).

94. van Wijk, X. M., Döhrmann, S., Hallström, B. M., Li, S., Voldborg, B. G., Meng, B. X., McKee, K. K., van Kuppevelt, T. H., Yurchenco, P. D., Palsson, B. O., Lewis, N. E., Nizet, V. & Esko, J. D. Whole-Genome Sequencing of Invasion-Resistant Cells Identifies Laminin α2 as a Host Factor for Bacterial Invasion. *MBio* **8,** (2017).

95. Chiang, A. W. T., Li, S., Kellman, B. P., Chattopadhyay, G., Zhang, Y., Kuo, C.-C., Gutierrez, J. M., Ghazi, F., Schmeisser, H., Ménard, P., Bjørn, S. P., Voldborg, B. G., Rosenberg, A. S., Puig, M. & Lewis, N. E. Combating viral contaminants in CHO cells by engineering innate immunity. *Scientific Reports* **9,** Preprint at https://doi.org/10.1038/s41598-019-45126-x (2019)

96. Singh, A., Kildegaard, H. F. & Andersen, M. R. An Online Compendium of CHO RNA-Seq Data Allows Identification of CHO Cell Line-Specific Transcriptomic Signatures. *Biotechnol. J.* **13,** e1800070 (2018).

97. Yu, N. Y. L., Hallström, B. M., Fagerberg, L., Ponten, F., Kawaji, H., Carninci, P., Forrest, A. R. R., Hayashizaki, Y., Uhlén, M. & Daub, C. O. Complementing tissue characterization by integrating transcriptome profiling from the human protein atlas and from the FANTOM5 consortium. *Nucleic Acids Res.* **43,** 6787–6798 (2015).

98. Danielsson, A., Pontén, F., Fagerberg, L., Hallström, B. M., Schwenk, J. M., Uhlén, M., Korsgren, O. & Lindskog, C. The Human Pancreas Proteome Defined by Transcriptomics and Antibody-Based Profiling. *PLoS ONE* **9,** e115421 Preprint at https://doi.org/10.1371/journal.pone.0115421 (2014)

99. Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., van de Lagemaat, L. N., Smith, K. A., Ebbert, A., Riley, Z. L., Abajian, C., Beckmann, C. F., Bernard, A., Bertagnolli, D., Boe, A. F., Cartagena, P. M., Chakravarty, M. M., Chapin, M., Chong, J., Dalley, R. A., David Daly, B., Dang, C., Datta, S., Dee, N., Dolbeare, T. A., Faber, V., Feng, D., Fowler, D. R., Goldy, J., Gregor, B. W., Haradon, Z., Haynor, D. R., Hohmann, J. G., Horvath, S., Howard, R. E., Jeromin, A., Jochim, J. M., Kinnunen, M., Lau, C., Lazarz, E. T., Lee, C., Lemon, T. A., Li, L., Li, Y., Morris, J. A., Overly, C. C., Parker, P. D., Parry, S. E., Reding, M., Royall, J. J., Schulkin, J., Sequeira, P. A., Slaughterbeck, C. R., Smith, S. C., Sodt, A. J., Sunkin, S. M., Swanson, B. E., Vawter, M. P., Williams, D., Wohnoutka, P., Zielke, H. R., Geschwind, D. H., Hof, P. R., Smith, S. M., Koch, C., Grant, S. G. N. & Jones, A. R. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489,** 391–399 (2012).

100. Cardoso-Moreira, M., Halbert, J., Valloton, D., Velten, B., Chen, C., Shao, Y., Liechti, A., Ascenção, K., Rummel, C., Ovchinnikova, S., Mazin, P. V., Xenarios, I., Harshman, K., Mort, M., Cooper, D. N., Sandi, C., Soares, M. J., Ferreira, P. G., Afonso, S., Carneiro, M., Turner, J. M. A., VandeBerg, J. L., Fallahshahroudi, A., Jensen, P., Behr, R., Lisgo, S., Lindsay, S., Khaitovich, P., Huber, W., Baker, J., Anders, S., Zhang, Y. E. & Kaessmann, H. Gene expression across mammalian organ development. *Nature* **571,** 505–509 (2019).

101. Dougherty, J. D., Schmidt, E. F., Nakajima, M. & Heintz, N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res.* **38,** 4218–4230 (2010).

102. Sugiaman-Trapman, D., Vitezic, M., Jouhilahti, E.-M., Mathelier, A., Lauter, G., Misra, S., Daub, C. O., Kere, J. & Swoboda, P. Characterization of the human RFX transcription factor family by regulatory and target gene analysis. *BMC Genomics* **19,** 181 (2018).

103. Rey-Campos, J., Chouard, T., Yaniv, M. & Cereghini, S. vHNF1 is a homeoprotein that

activates transcription and forms heterodimers with HNF1. *EMBO J.* **10,** 1445–1457 (1991).

104. Tremblay, M., Sanchez-Ferras, O. & Bouchard, M. GATA transcription factors in development and disease. *Development* **145,** Preprint at https://doi.org/10.1242/dev.164384 (2018)

105. Berger, J. & Moller, D. E. The mechanisms of action of PPARs. *Annu. Rev. Med.* **53,** 409–435 (2002).

106. Lin, Q., Schwarz, J., Bucana, C. & Olson, E. N. Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C. *Science* **276,** 1404–1407 (1997).

107. Black, B. L. & Olson, E. N. Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins. *Annu. Rev. Cell Dev. Biol.* **14,** 167–196 (1998).

108. Sanetra, M., Begemann, G., Becker, M.-B. & Meyer, A. Conservation and co-option in developmental programmes: the importance of homology relationships. *Front. Zool.* **2,** 15 (2005).

109. Gregory, T. R. The Evolution of Complex Organs. *Evolution: Education and Outreach* **1,** 358–389 (2008).

110. Kadonaga, J. T., Courey, A. J., Ladika, J. & Tjian, R. Distinct regions of Sp1 modulate DNA binding and transcriptional activation. *Science* **242,** 1566–1570 (1988).

111. Raetz, C. R. H., Garrett, T. A., Reynolds, C. M., Shaw, W. A., Moore, J. D., Smith, D. C., Jr, Ribeiro, A. A., Murphy, R. C., Ulevitch, R. J., Fearns, C., Reichart, D., Glass, C. K., Benner, C., Subramaniam, S., Harkewicz, R., Bowers-Gentry, R. C., Buczynski, M. W., Cooper, J. A., Deems, R. A. & Dennis, E. A. Kdo2-Lipid A of Escherichia coli, a defined endotoxin that activates macrophages via TLR-4. *J. Lipid Res.* **47,** 1097–1111 (2006).

112. Spain, M. M., Caruso, J. A., Swaminathan, A. & Pile, L. A. Drosophila SIN3 isoforms interact with distinct proteins and have unique biological functions. *J. Biol. Chem.* **285,** 27457–27467 (2010).

113. Reyes, A. & Huber, W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* **46,** 582–592 (2018).

114. Solá, R. J. & Griebenow, K. Glycosylation of therapeutic proteins: an effective strategy to optimize efficacy. BioDrugs [Internet]. 2010 Feb 1 [cited 2017 Jan 27]; 24 (1): 9--21.

115. Stanley, P., Sundaram, S., Tang, J. & Shi, S. Molecular analysis of three gain-of-function CHO mutants that add the bisecting GlcNAc to N-glycans. *Glycobiology* **15,** 43–53 (2005).

116. Schachter, H. Biosynthetic controls that determine the branching and microheterogeneity of

protein-bound oligosaccharides. *Biochem. Cell Biol.* **64,** 163–181 (1986).

117. Umaña, P., Jean-Mairet, J., Moudry, R., Amstutz, H. & Bailey, J. E. Engineered glycoforms of an antineuroblastoma IgG1 with optimized antibody-dependent cellular cytotoxic activity. *Nat. Biotechnol.* **17,** 176–180 (1999).

118. Marx, N., Grünwald-Gruber, C., Bydlinski, N., Dhiman, H., Ngoc Nguyen, L., Klanert, G. & Borth, N. CRISPR-based targeted epigenetic editing enables gene expression modulation of the silenced beta-galactoside alpha-2,6-sialyltransferase 1 in CHO cells. *Biotechnol. J.* **13,** e1700217 (2018).

119. Nguyen, L. N., Baumann, M., Dhiman, H., Marx, N., Schmieder, V., Hussein, M., Eisenhut, P., Hernandez, I., Koehn, J. & Borth, N. Novel Promoters Derived from Chinese Hamster Ovary Cells via In Silico and In Vitro Analysis. *Biotechnol. J.* **14,** e1900125 (2019).

120. Edros, R., McDonnell, S. & Al-Rubeai, M. The relationship between mTOR signalling pathway and recombinant antibody productivity in CHO cell lines. *BMC Biotechnol.* **14,** 15 (2014).

121. Pang, W. W., Price, E. A., Sahoo, D., Beerman, I., Maloney, W. J., Rossi, D. J., Schrier, S. L. & Weissman, I. L. Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 20012–20017 (2011).

122. Sanjuan-Pla, A., Macaulay, I. C., Jensen, C. T., Woll, P. S., Luis, T. C., Mead, A., Moore, S., Carella, C., Matsuoka, S., Bouriez Jones, T., Chowdhury, O., Stenson, L., Lutteropp, M., Green, J. C. A., Facchini, R., Boukarabila, H., Grover, A., Gambardella, A., Thongjuea, S., Carrelha, J., Tarrant, P., Atkinson, D., Clark, S.-A., Nerlov, C. & Jacobsen, S. E. W. Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. *Nature* **502,** 232–236 (2013).

123. Carrelha, J., Meng, Y., Kettyle, L. M., Luis, T. C., Norfo, R., Alcolea, V., Boukarabila, H., Grasso, F., Gambardella, A., Grover, A., Högstrand, K., Lord, A. M., Sanjuan-Pla, A., Woll, P. S., Nerlov, C. & Jacobsen, S. E. W. Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells. *Nature* **554,** 106–111 (2018).

124. Grover, A., Sanjuan-Pla, A., Thongjuea, S., Carrelha, J., Giustacchini, A., Gambardella, A., Macaulay, I., Mancini, E., Luis, T. C., Mead, A., Jacobsen, S. E. W. & Nerlov, C. Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. *Nat. Commun.* **7,** 11075 (2016).

125. Pinho, S., Marchand, T., Yang, E., Wei, Q., Nerlov, C. & Frenette, P. S. Lineage-Biased Hematopoietic Stem Cells Are Regulated by Distinct Niches. *Dev. Cell* **44,** 634–641.e4 (2018).

126. Lu, R., Czechowicz, A., Seita, J., Jiang, D. & Weissman, I. L. Clonal-level lineage commitment pathways of hematopoietic stem cells in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 1447–1456 (2019).

127. Busch, K., Klapproth, K., Barile, M., Flossdorf, M., Holland-Letz, T., Schlenner, S. M., Reth, M., Höfer, T. & Rodewald, H.-R. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature* **518,** 542–546 (2015).

128. Park, S.-M., Deering, R. P., Lu, Y., Tivnan, P., Lianoglou, S., Al-Shahrour, F., Ebert, B. L., Hacohen, N., Leslie, C., Daley, G. Q., Lengner, C. J. & Kharas, M. G. Musashi-2 controls cell fate, lineage bias, and TGF-β signaling in HSCs. *J. Exp. Med.* **211,** 71–87 (2014).

129. Sarrazin, S., Mossadegh-Keller, N., Fukao, T., Aziz, A., Mourcin, F., Vanhille, L., Kelly Modis, L., Kastner, P., Chan, S., Duprez, E., Otto, C. & Sieweke, M. H. MafB restricts M-CSF-dependent myeloid commitment divisions of hematopoietic stem cells. *Cell* **138,** 300–313 (2009).

130. Muller-Sieburg, C. E., Cho, R. H., Karlsson, L., Huang, J.-F. & Sieburg, H. B. Myeloid-biased hematopoietic stem cells have extensive self-renewal capacity but generate diminished lymphoid progeny with impaired IL-7 responsiveness. *Blood* **103,** 4111–4118 (2004).

131. Kirschner, K., Chandra, T., Kiselev, V., Flores-Santa Cruz, D., Macaulay, I. C., Park, H. J., Li, J., Kent, D. G., Kumar, R., Pask, D. C., Hamilton, T. L., Hemberg, M., Reik, W. & Green, A. R. Proliferation Drives Aging-Related Functional Decline in a Subpopulation of the Hematopoietic Stem Cell Compartment. *Cell Rep.* **19,** 1503–1511 (2017).

132. Hoggatt, J., Mohammad, K. S., Singh, P. & Pelus, L. M. Prostaglandin E2 enhances long-term repopulation but does not permanently alter inherent stem cell competitiveness. *Blood* **122,** 2997–3000 (2013).

133. Rundberg Nilsson, A., Soneji, S., Adolfsson, S., Bryder, D. & Pronk, C. J. Human and Murine Hematopoietic Stem Cell Aging Is Associated with Functional Impairments and Intrinsic Megakaryocytic/Erythroid Bias. *PLoS One* **11,** e0158369 (2016).

134. Frisch, B. J., Hoffman, C. M., Latchney, S. E., LaMere, M. W., Myers, J., Ashton, J., Li, A. J., Saunders, J., 2nd, Palis, J., Perkins, A. S., McCabe, A., Smith, J. N., McGrath, K. E., Rivera-Escalera, F., McDavid, A., Liesveld, J. L., Korshunov, V. A., Elliott, M. R., MacNamara, K. C., Becker, M. W. & Calvi, L. M. Aged marrow macrophages expand platelet-biased hematopoietic stem cells via Interleukin1B. *JCI Insight* **5,** (2019).

135. Civin, C. I., Strauss, L. C., Brovall, C., Fackler, M. J., Schwartz, J. F. & Shaper, J. H. Antigenic analysis of hematopoiesis. III. A hematopoietic progenitor cell surface antigen defined by a monoclonal antibody raised against KG-1a cells. *J. Immunol.* **133,** 157–165

(1984).

136. DiGiusto, D., Chen, S., Combs, J., Webb, S., Namikawa, R., Tsukamoto, A., Chen, B. P. & Galy, A. H. Human fetal bone marrow early progenitors for T, B, and myeloid cells are found exclusively in the population expressing high levels of CD34. *Blood* **84,** 421–432 (1994).

137. Seita, J. & Weissman, I. L. Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2,** 640–653 (2010).

138. Naik, S. H., Perié, L., Swart, E., Gerlach, C., van Rooij, N., de Boer, R. J. & Schumacher, T. N. Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* **496,** 229–232 (2013).

139. Lin, D. S., Tian, L., Tomei, S., Amann-Zalcenstein, D., Baldwin, T. M., Weber, T. S., Schreuder, J., Stonehouse, O. J., Rautela, J., Huntington, N. D., Taoudi, S., Ritchie, M. E., Hodgkin, P. D., Ng, A. P., Nutt, S. L. & Naik, S. H. Single-cell analyses reveal the clonal and molecular aetiology of Flt3L-induced emergency dendritic cell development. *Nat. Cell Biol.* **23,** 219–231 (2021).

140. Sun, J., Ramos, A., Chapman, B., Johnnidis, J. B., Le, L., Ho, Y.-J., Klein, A., Hofmann, O. & Camargo, F. D. Clonal dynamics of native haematopoiesis. *Nature* **514,** 322–327 (2014).

141. Rodriguez-Fraticelli, A. E., Wolock, S. L., Weinreb, C. S., Panero, R., Patel, S. H., Jankovic, M., Sun, J., Calogero, R. A., Klein, A. M. & Camargo, F. D. Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553,** 212–216 (2018).

142. Spencer Chapman, M., Ranzoni, A. M., Myers, B., Williams, N., Coorens, T. H. H., Mitchell, E., Butler, T., Dawson, K. J., Hooks, Y., Moore, L., Nangalia, J., Robinson, P. S., Yoshida, K., Hook, E., Campbell, P. J. & Cvejic, A. Lineage tracing of human development through somatic mutations. *Nature* **595,** 85–90 (2021).

143. Kim, S., Kim, N., Presson, A. P., Metzger, M. E., Bonifacino, A. C., Sehl, M., Chow, S. A., Crooks, G. M., Dunbar, C. E., An, D. S., Donahue, R. E. & Chen, I. S. Y. Dynamics of HSPC repopulation in nonhuman primates revealed by a decade-long clonal-tracking study. *Cell Stem Cell* **14,** 473–485 (2014).

144. Koelle, S. J., Espinoza, D. A., Wu, C., Xu, J., Lu, R., Li, B., Donahue, R. E. & Dunbar, C. E. Quantitative stability of hematopoietic stem and progenitor cell clonal output in rhesus macaques receiving transplants. *Blood* **129,** 1448–1457 (2017).

145. Biasco, L., Pellin, D., Scala, S., Dionisio, F., Basso-Ricci, L., Leonardelli, L., Scaramuzza, S., Baricordi, C., Ferrua, F., Cicalese, M. P., Giannelli, S., Neduva, V., Dow, D. J., Schmidt, M., Von Kalle, C., Roncarolo, M. G., Ciceri, F., Vicard, P., Wit, E., Di Serio, C., Naldini, L.

& Aiuti, A. In Vivo Tracking of Human Hematopoiesis Reveals Patterns of Clonal Dynamics during Early and Steady-State Reconstitution Phases. *Cell Stem Cell* **19,** 107–119 (2016).

146. Six, E., Guilloux, A., Denis, A., Lecoules, A., Magnani, A., Vilette, R., Male, F., Cagnard, N., Delville, M., Magrin, E., Caccavelli, L., Roudaut, C., Plantier, C., Sobrino, S., Gregg, J., Nobles, C. L., Everett, J. K., Hacein-Bey-Abina, S., Galy, A., Fischer, A., Thrasher, A. J., André, I., Cavazzana, M. & Bushman, F. D. Clonal tracking in gene therapy patients reveals a diversity of human hematopoietic differentiation programs. *Blood* **135,** 1219–1231 (2020).

147. Scala, S., Basso-Ricci, L., Dionisio, F., Pellin, D., Giannelli, S., Salerio, F. A., Leonardelli, L., Cicalese, M. P., Ferrua, F., Aiuti, A. & Biasco, L. Dynamics of genetically engineered hematopoietic stem and progenitor cells after autologous transplantation in humans. *Nat. Med.* **24,** 1683–1690 (2018).

148. Kiel, M. J., Yilmaz, O. H., Iwashita, T., Yilmaz, O. H., Terhorst, C. & Morrison, S. J. SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121,** 1109–1121 (2005).

149. Rector, K., Liu, Y. & Van Zant, G. Comprehensive hematopoietic stem cell isolation methods. *Methods Mol. Biol.* **976,** 1–15 (2013).

150. Balazs, A. B., Fabian, A. J., Esmon, C. T. & Mulligan, R. C. Endothelial protein C receptor (CD201) explicitly identifies hematopoietic stem cells in murine bone marrow. *Blood* **107,** 2317–2321 (2006).

151. Christensen, J. L. & Weissman, I. L. Flk-2 is a marker in hematopoietic stem cell differentiation: a simple method to isolate long-term stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 14541–14546 (2001).

152. Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A. & Trapnell, C. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14,** 979–982 (2017).

153. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S. & Kharchenko, P. V. RNA velocity of single cells. *Nature* **560,** 494–498 (2018).

154. Li, C., Virgilio, M. C., Collins, K. L. & Welch, J. D. Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. *Nat. Biotechnol.* 1–12 (2022).

155. Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., Pelka, K., Ge, W., Oren, Y., Brack, A., Law, T., Rodman, C., Chen, J. H., Boland, G. M., Hacohen, N., Rozenblatt-Rosen, O., Aryee, M. J., Buenrostro, J. D., Regev, A. & Sankaran, V. G. Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176,** 1325–1339.e22 (2019).

156. Lareau, C. A., Ludwig, L. S., Muus, C., Gohil, S. H., Zhao, T., Chiang, Z., Pelka, K., Verboon, J. M., Luo, W., Christian, E., Rosebrock, D., Getz, G., Boland, G. M., Chen, F., Buenrostro, J. D., Hacohen, N., Wu, C. J., Aryee, M. J., Regev, A. & Sankaran, V. G. Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nat. Biotechnol.* **39,** 451–461 (2021).

157. Elson, J. L., Samuels, D. C., Turnbull, D. M. & Chinnery, P. F. Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age. *Am. J. Hum. Genet.* **68,** 802–806 (2001).

158. Stewart, J. B. & Chinnery, P. F. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.* **16,** 530–542 (2015).

159. Wallace, D. C. & Chalkia, D. Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harb. Perspect. Biol.* **5,** a021220 (2013).

160. Jaiswal, S. & Ebert, B. L. Clonal hematopoiesis in human aging and disease. *Science* **366,** (2019).

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

162. Chapple, R. H., Tseng, Y.-J., Hu, T., Kitano, A., Takeichi, M., Hoegenauer, K. A. & Nakada, D. Lineage tracing of murine adult hematopoietic stem cells reveals active contribution to steady-state hematopoiesis. *Blood Adv* **2,** 1220–1228 (2018).

163. Cromer, M. K., Vaidyanathan, S., Ryan, D. E., Curry, B., Lucas, A. B., Camarena, J., Kaushik, M., Hay, S. R., Martin, R. M., Steinfeld, I., Bak, R. O., Dever, D. P., Hendel, A., Bruhn, L. & Porteus, M. H. Global Transcriptional Response to CRISPR/Cas9-AAV6-Based Genome Editing in CD34+ Hematopoietic Stem and Progenitor Cells. *Mol. Ther.* **26,** 2431–2442 (2018).

164. Miller, T. E., Lareau, C. A., Verga, J. A., DePasquale, E. A. K., Liu, V., Ssozi, D., Sandor, K., Yin, Y., Ludwig, L. S., El Farran, C. A., Morgan, D. M., Satpathy, A. T., Griffin, G. K., Lane, A. A., Love, J. C., Bernstein, B. E., Sankaran, V. G. & van Galen, P. Mitochondrial variant enrichment from high-throughput single-cell RNA sequencing resolves clonal populations. *Nat. Biotechnol.* **40,** 1030–1034 (2022).

165. Jacobsen, S. E. W. & Nerlov, C. Haematopoiesis in the era of advanced single-cell technologies. *Nat. Cell Biol.* **21,** 2–8 (2019).

166. Feng, J., Pucella, J. N., Jang, G., Alcántara-Hernández, M., Upadhaya, S., Adams, N. M., Khodadadi-Jamayran, A., Lau, C. M., Stoeckius, M., Hao, S., Smibert, P., Tsirigos, A., Idoyaga, J. & Reizis, B. Clonal lineage tracing reveals shared origin of conventional and plasmacytoid dendritic cells. *Immunity* **55,** 405–422.e11 (2022).

167. Giladi, A., Paul, F., Herzog, Y., Lubling, Y., Weiner, A., Yofe, I., Jaitin, D., Cabezas-Wallscheid, N., Dress, R., Ginhoux, F., Trumpp, A., Tanay, A. & Amit, I. Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat. Cell Biol.* **20,** 836–846 (2018).

168. Lehnertz, B., Chagraoui, J., MacRae, T., Tomellini, E., Corneau, S., Mayotte, N., Boivin, I., Durand, A., Gracias, D. & Sauvageau, G. HLF expression defines the human hematopoietic stem cell state. *Blood* **138,** 2642–2654 (2021).

169. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* **20,** 1–12 (2019).

170. Heike, Wickham & Kafadar. Letter-value plots: Boxplots for large data. *J. Comput. Graph. Stat.*

171. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., 3rd, Hao, Y., Stoeckius, M., Smibert, P. & Satija, R. Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902.e21 (2019).

172. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17,** 261–272 (2020).