THE UNIVERSITY OF CHICAGO

MECHANISMS OF INNATE IMMUNE MEMORY IN BASIC AND CLINICAL MODELS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

INTERDISCIPLINARY SCIENTIST TRAINING PROGRAM: IMMUNOLOGY

BY

SARAH JIE SUN

CHICAGO, ILLINOIS

DECEMBER 2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Perhaps the most difficult part of writing this thesis has been writing this section. I feel indebted to a seemingly endless number of friends, family members, and mentors without whom none of this would be possible. I have to start by thanking my PhD mentor Luis. Luis has been an incredible mentor throughout my PhD not only in times of success but in really low moments. A few years into my PhD I began to have recurring problems with critical experiments for one of the projects I was working on. I remember it as one of the most frustrating points, honestly, in my life and I wanted nothing more than to crumple up my manuscript and throw it away. At that time Luis pushed me to continue to persevere and to make the most out of the project. When I felt that the failure of my experiments made the work worthless, he pushed me think about the interesting and novel aspects of the project and the ways in which it was still valuable. Now, in the end, it is a project that I am proud of, in large part due to the positively with which Luis has brought to my research both in this specific instance, but also more generally throughout our time working together. I am incredibly grateful and proud to have been given the opportunity to conduct research with Luis whom I respect tremendously both as mentor and scientist.

I would also like to thank the remaining members of my thesis committee, Francois Spitz, Bana Jabri, and Nicolas Chevrier. They are all incredible scientists and I feel lucky to have had the opportunity to present my work to them at each committee meeting. Over the course of my PhD they have brought innovative and interesting ideas to the table and have pushed me to think more deeply about my science. Because of them, this work became more rigorous and complete. Bana who is one of my PhD co-mentors, has always offered her time to mentor me on the

experimental aspects of my research projects. I have presented at her lab meetings for multiple years and am still struck by the innovate ideas which she seems to always have.

I want to thank every member, past and present, from the Barreiro lab for these past 4 years. It's difficult to think of another group of people who are as intelligent and hardworking, yet as kind and supportive as my fellow lab members. I feel that we are not just colleagues, but great, long-lasting friends. I wish all of them the very, very best for the future. My MSTP classmates likewise have been, and are, an incredibly humble, fun, and interesting group of people with whom I feel so fortunate to call friends. I cannot help but reflect back over the past 5 years, from the moment we first stepped into the BSLC for our first anatomy class, until now, when each of us are getting our PhDs.

Everyone in the medical scientist training program deserves a huge thanks for the support and guidance I have received through my PhD. I particularly want to thank Kristin McCann for all of her support throughout this time and her willingness to take time out of her schedule to give advice, whether that was to respond to one of my many emails, or even to talk with me on the phone in crisis moments. I thank the University of Chicago flow cytometry core and the Human Disease and Immune Discovery core for their assistance and work. I want to thank Cezary Ciszewski for his work on the BCG project on which the rest of the project relied. Finally, I want to thank my mom and dad for their support of me from day one. Throughout my life they have made tremendous sacrifices so that I could have the best, both in life and in my career. Without their support none of this would be possible.

ABSTRACT

Innate immune memory is a burgeoning field in immunology, yet there still remain many unanswered questions regarding its inherent duration in mitotic and post-mitotic cells, and its applicability in clinical settings. Clinically, the BCG vaccine is hypothesized to induce long-lasting innate immune memory in humans indirectly by reprogramming hematopoietic stem and progenitor cells (HSPCs) in the bone marrow. Although this model is supported by mouse studies, very little is known about whether reprogramming of HSPCs in response to BCG occurs in humans, and whether this has implications for the function of mature innate immune cells entering the peripheral circulation. In this work we first used single cell sequencing to probe the gene expression and chromatin accessibility landscape of human bone marrow samples before and 90 days after BCG vaccination and integrated these data with secondary response data collected on donor paired PBMCs. We find that the most uncommitted stem cells exhibit persistent signatures of myeloid bias and upregulation of immune genes. On the epigenetic level, downstream progenitors contained thousands of sites of differential accessibility at sites which enriched for the motifs of KLF/SP and EGR transcription factors which were predominantly active within upstream HSCs, suggesting that long-lasting TF activity and differential gene expression at the level of HSCs may impact the chromatin accessibility landscape of downstream progenitors. Myeloid bias, HSC activation signatures, and progenitor chromatin accessibility levels were all found to correlate significantly with Il1β secretion of donor paired PBMCs in response to a C. albicans challenge, indicating that BCG vaccination re-wires transcription factor activity, gene expression, chromatin accessibility, and lineage bias in human bone marrow in a

way that is linked to responses of PBMCs to secondary immune challenge with non-mycobacterial pathogens.

In addition, we utilized an in vitro macrophage model to study basic mechanisms by which innate immune memory can be retained within dividing cell populations. The encoding of innate immune memory has often been linked to histone post-translational modifications (PTMs). Since there is no known mechanism whereby stimulus-induced histone PTMs can be directly copied from parent to daughter strand during DNA replication, it is expected that these signatures be rapidly lost in dividing cells. Yet, in vivo studies have demonstrated that a state of trained immunity can persist for months, which paradoxically suggests that histone PTMs may persist. Using time course RNA-seq, ChIP-seq, and functional assays, we find that dividing macrophages harbor H3K4me1 signatures at hundreds of sites for at least 14 cell divisions after stimulus washout, however these marks are dynamic, timepoint specific, and tightly coupled to the continued activity of transcription factor (TF) circuits. Our work emphasizes the central role of continued TF activation in driving the continued detection of H3K4me1 signatures across cell divisions and suggests that innate immune memory in dividing cells may be a phenomenon that is mechanistically separate from that observed in non-dividing cells.

CHAPTER I: INTRODUCTION[a]

Overview of adaptive immune memory and the response to infection

The immune system is a highly complex network of diverse cell types. Any one defect affecting a particular component of the immune system can seriously undermine one's ability to fight infection, underscoring the clinical and scientific importance of understanding how all aspects of the immune system function in concert.

Cells of the vertebrate immune system are classically divided into two branches. The *adaptive* immune system is comprised of cells such as T- and B-lymphocytes that can form antigen-specific immunological memory. *Innate* immune cells include monocytes, dendritic cells, and neutrophils, and have historically been viewed as general, "non-specific" responders to pathogen challenge. Upon infection, a complex interplay between cells of the innate and adaptive immune systems is initiated. Innate immune cells are found both in the peripheral circulation, as well as within tissues and barrier sites and express germline encoded pathogen recognition receptors (PRRs) that recognize pathogen-associated or danger-associated molecular patterns introduced during infections. While recognition is not antigen-specific *per se*, different receptors recognize different classes of pathogens (i.e., gram negative or positive bacteria, viruses, fungi), allowing a response that is tailored to some extent[1]. Innate immune cells typically act as first responders to infection by phagocytosing pathogen, producing pro-inflammatory cytokines, presenting antigen,

[a]Parts of this section are reproduced with modification from Sun, S., & Barreiro, L. B. (2020). The epigenetically-encoded memory of the innate immune system. Current opinion in immunology, 65, 7-13.

and trafficking to secondary lymphoid organs to present this antigen to naïve T-cells[1]. The presentation of antigen by APCs (most notably dendritic cells) to naïve T-lymphocytes then initiates a series of selective expansion and differentiation events that lead to a memory T-cell pool and the formation of high-affinity immunoglobulins. The exact models and steps by which memory T-cells are formed are still debated, but generally, a naïve T-cell specific to a particular antigen first undergoes a high level of clonal expansion (believed to be around 400,000-fold in many cases[2]). A portion of these clonally expanded cells develop a memory-precursor phenotype and over time, in the presence of cytokine exposure, develop into fully fledged memory T-cells[3]. The end result is that after an initial infection, there is a larger number of T-cells capable of recognizing the antigens of that infection. Not only are the T-cell clones expanded, but they are intrinsically altered, harboring permanently rewired transcriptional programs driven by sustained changes in the levels of transcription factors such as T-bet, EOMES, BLIMP, and BCL6[3,4]. Thus, adaptive immune memory is driven by antigen-specific selective clonal expansion, and a rewiring of transcriptional programs caused by changes to the balance of critical transcription factors that collectively make memory lymphocytes more capable of rapidly engaging effector functions and expanding upon re-exposure to the same antigen[4].

Innate immune memory

The type of immunological memory described above is an exclusive hallmark of the adaptive immune system. However, the rigid view that other cells are devoid of any memory-like properties, has changed over the past 10-15 years with research in both mice and humans demonstrating memory-like features within innate immune cells[5–7]. While innate immune cells lack antigen specificity and the diversity of receptors that bestow T-cells and B-cells with the

collective ability to recognize a virtually unlimited number of antigens, they can undergo fundamental changes in response to immune challenge, can maintain these changes on a variable scale of days, weeks, or months, and can mount qualitatively and quantitatively altered responses to a broad range of secondary pathogens. Broadly, this ability to mount enhanced secondary responses after an initial priming event has been referred to as *trained immunity*, or *innate immune memory*.

The concept of trained immunity was formalized in the 2010's in a series of *in vitro* studies performed on human monocytes. These studies demonstrated that human monocytes isolated and placed in cell culture could gain short-term enhanced recall responses after priming with *C. albicans* or the *C. albicans* cell wall component, β-glucan[8,9]. In these initial experiments, the purified monocytes were stimulated with *C. albicans* or β-glucan for 24 hours, washed, rested for 1 week, then challenged again. Primed monocytes, as compared to naïve monocytes, secreted significantly higher levels of the proinflammatory cytokines IL6 and TNFα, but not the anti-inflammatory cytokine IL-10, in response to a wide range of secondary stimuli including lipopolysaccharide (LPS), *C. albicans*, and *Mycobacterium tuberculosis* (Mtb)[8,9].

Subsequent studies *in vivo* have led to more complex readouts, broadening the concept of innate immune memory to accommodate these diverse memory-like phenotypes. *In vivo*, a wide variety of different pathogens have been shown to modulate the innate immune system of mice. One group reported that intraperitoneal injection of LPS led to changes in the basal expression levels of Stat1 genes in peritoneal macrophages for three weeks and decreased the bacterial burden of S. aureus following a subsequent infection[10]. BCG vaccination protected SCID mice against subsequent *C. albicans* infection two weeks later[11] even though these mice lack a fully functional

adaptive immune system, suggesting a critical role for innate immunity. Additionally, respiratory infections with adenovirus[12], influenza[13], S. aureus[14] and E.coli[14], have all been shown to modulate alveolar macrophages on the functional, protein, gene expression, or epigenetic level within the lungs of mice for one or more months.

The role of epigenetics in innate immune memory

Although we still lack a clear picture of the mechanistic underpinnings of the various forms of innate immune memory that have been described, innate immune cells developing altered phenotypes following a primary challenge, almost always harbor epigenetic alterations. Thus, a predominant hypothesis is that epigenetic reprogramming induced upon a primary stimulation serves as the "language" by which innate immune memory is encoded.

In the most fundamental sense, epigenetic changes refer to changes that impact cell biology without being genetic - meaning that they do not change the DNA coding sequence. The most common way by which this can occur is through alterations in the way DNA is packaged. DNA is naturally packaged together with proteins called histones into chromatin of which the basic unit is the nucleosome (which is comprised of the histone protein and the 147 base pairs of DNA wrapped around it). The spacing and density of nucleosomes is highly heterogenous. Regions referred to as heterochromatin contain densely packed nucleosomes while euchromatin refers to those portions of chromatin where nucleosomes are spaced farther apart making the DNA more accessible[15]. Not surprisingly, while only 2-3% of DNA is accessible, these minority regions make up more than 90% of binding sites for transcription factors[16] highlighting the fact that chromatin structure is tightly intertwined with control of gene expression.

It is now well established that chemical modifications, either to histone proteins, or to DNA bases can change chromatin packaging, thereby modulating the ability of transcription factors to bind. Since this ultimately impacts gene expression and cellular behavior these modifications are seen as prototypical epigenetic modifications (although by some definitions only modifications heritable though replication are truly epigenetic thus excluding some histone modifications which are nonetheless referred to as epigenetic in this document). The chemical modification 5mC (DNA methylation) has been directly linked to gene repression. Its importance to cell biology is clear by the number of critical process in which DNA methylation is involved, ranging from X-chromosome inactivation, genomic imprinting, and the silencing of retroviral elements to its consistent dysregulation in various disease contexts such as cancer, in which DNA methylation levels are often unusually high at tumor suppressor regions[15]. The enzymes DNMT3a/b and DNMT1 are enzymatic writers of DNA methylation that catalyze DNA methylation de novo and retain DNA methylation across DNA replication respectively[17]. The TET1-3 enzymes subsequently can remove DNA methylation, meaning that this modification is ultimately reversible[17]. A large percentage of the genome is methylated, except for at CpG islands[17] which are CpG-dense regions occurring near promoters that serve as attractive binding sites for proteins that inhibit the activity of DNMT enzymes, ultimately preventing DNA-methylation mediated silencing of gene promoters[15].

Among histone modifications, methylation, acetylation, and phosphorylation are the most studied and believed to be the most important and prevalent[15]. Histone acetylation has generally been linked to active gene expression going as far back as the 1960's when the first HAT (histone acetyltransferase) was purified and cloned and found to be an orthologue of a protein known to be involved in gene activation – providing a link between histone acetylation and gene

expression upregulation[15]. One of the most widespread and studied histone acetylation modifications H3K27Ac for example, is a particular histone modification widely associated with active transcription and can be found at both active enhancers and promoter regions. Histone methylation depending on the type, is variably linked to both gene activation and repression. ChIP-seq studies, for example, have demonstrated that H3K4me3 is located at active promoter regions while H3K27me3 is associated with repressed regions and is generally anticorrelated with H3K4me3. As the number of such genome-wide ChIP-seq studies have increased it has generally become clear that promoters and enhancers, the primary regulatory regions, are associated with distinct combinations of histone modifications and DNA methylation that correlate not only with the region identity (enhancer versus promoter) but also the activation level of that region (see Table 0.1). While the correlative versus causal relationship between histone modifications and regulatory region activity remains to be completely understood, these modifications are believed to causally influence regulatory region activity and gene expression through both cis and trans mechanisms[15]. In trans, unique combinations of histone modifications can attract binding by unique combinations of proteins complexes containing transcription factors that promote or repress specific gene expression programs. In addition, in cis, histone modifications can increase or decrease steric hindrance making physically it less or more likely for TFs to bind[16].

Genomic element	Epigenetic marks	Notes
Silenced promoter	High DNA methylation	Methylated promoters are usually irreversibly silent
Poised/inactive promoter	H3K4me3, hypomethylated	May be more susceptible to DNA methylation in disease contexts
Active promoter	H3K4me3^{hi}, H3K27Ac, hypomethylated, high RNA PolII, moderate p300	
Poised enhancer	H3K4me1, moderate p300	Poised enhancers are established and bound by pioneering transcription factors
Active enhancers	H3K4me1, H3K27Ac, hypomethylated under certain conditions, p300, H3K4me2	
Latent enhancer	unmarked	Defined by a lack of marks associated with poised or active enhancers, but gains these marks when a cell is
Active Gene body	H3K36me3, DNA methylation	Unlike at regulatory regions, DNA methylation within gene bodies is associated with increased transcription
Heterochromatin, repressed regions	H3K9me3, H3K27me3	

Table 0.1. Specific combinations of histone modifications and DNA methylation levels occur at different regulatory elements and can also be used to identify these elements de novo (reproduced with modification from *Sun, S., & Barreiro, L. B. (2020). The epigenetically-encoded memory of the innate immune system. Current opinion in immunology, 65, 7-13.*)

Histone post translational modifications (PTMs) are the modifications that appear to be the most responsive to immune challenges. Immune-induced changes to the histone PTM landscape often involves the deposition or the loss of histone PTMs at known enhancer or promoter sites that often already harbor some low levels of these same PTMs at baseline (H3K4me1 at enhancers, H3K4me3 at promoters, and H3K27Ac at active enhancers and promoters[15,18]). Immune stimulation can then modulate the levels of these histone PTMs at their respective locations. For example, mouse bone marrow derived macrophages (BMDMs) stimulated with LPS quickly gain higher levels of H3K27Ac at more than 5000 pre-defined enhancers[19]. Similarly, human

monocytes stimulated with certain β-glucans, BCG, or LPS, undergo genome wide changes in H3K27Ac levels, as well as more modest changes in H3K4me1 and H3K4me3 levels at enhancers and promoters, respectively[5,9,10,20,21]. In addition to the deposition of modified histones at preexisting promoters and enhancers, it is also possible for pro-inflammatory stimuli to induce the de novo formation of enhancer elements through the deposition of H3K4me1 at sites with previously undetectable levels of this mark. The infection of human dendritic cells (DC) with Mycobacterium tuberculosis (Mtb) leads to the emergence of hundreds of de novo enhancers[22]. A recent paper investigating the effects of IL-1β and IFNG on human pancreatic islet cells identified 3800 regulatory elements responsive to cytokine stimulation and attributed 45% of changes to the induction of new regulatory elements[23].

The degree to which other epigenetic features may encode innate immune memory phenotypes remains unclear. DNA methylation has been particularly understudied, due to the belief that methylation marks are highly stable, and unlikely to respond to environmental perturbations on a short time scale. Despite this belief, recent studies have suggested that DNA methylation may be more plastic than previously appreciated. For example, infection of post-mitotic human DCs or macrophages with live Mtb led to an active loss of DNA methylation at thousands of enhancers throughout the genome[22,24] suggesting a potential role for this epigenetic mark in innate immune memory.

These stimulus-induced changes to the epigenome are thought to have direct consequences for future cellular behavior and function, although most current data is based on correlational analyses. For example, stimulating macrophages with LPS led to widespread deposition of H3K27Ac and H3K4me3 onto gene promoters. Although only some of these marks were

maintained, they coincided with genes that were primed, while tolerized genes lost these marks, linking increased H3K4me3 and H3K27Ac levels at specific gene promoters with priming and the loss of these marks with tolerance[25]. More recent work in an epidermal stem cell system, has suggested that the selective maintenance of activating histone modifications such as H3K27Ac and H3K4me1 at some sites after priming may involve the selective continued binding of transcription factors[26]. In a stem cell model, it has been demonstrated that long lasting increases in chromatin accessibility and histone PTMs in response to inflammation occurs only at a subset of "memory" peaks identified as binding sites for Stat3, FOS and JUN. Sites retaining H3K4me1 and H3K27Ac also remained bound by Stat3 and FOS following the resolution of inflammation. These memory peaks were then selectively bound rapidly by JUN upon a secondary stimulus, suggesting a global mechanism whereby transcription factors co-bind to chromatin, maintain increased levels of histone modifications, and keep the chromatin open, allowing more rapid binding of partner TFs upon a second challenge. Other epigenetic marks besides H3K4me1, H3K4me3, and H3K27Ac could also serve a similar role in marking chromatin. IFNβ stimulation of MEFs induced H3.3 and H3K36me3 at selective sites residing close to genes that were primed upon a secondary IFNβ challenge[27] and DNA methylation appears to be involved in the encoding of memory within NK cells which upon HCMV infection are primed to secrete increased IFNγ in response to a second HCMV infection[28]. In summary, most work has pointed to the same general paradigm whereby an initial stimulation induces chromatin-level changes - most often consisting of altered H3K4me1/H3K4me3/H3K27Ac levels, but also potentially involving retained binding of transcription factors, other histone PTMs, or DNA methylation- to keep the chromatin "book-marked" for faster binding upon a secondary challenge.

The intersection of metabolism and epigenetics in trained immunity

In addition, the regulation of epigenetics has been shown to be tightly linked to cell metabolism[9,29–31]. Multiple studies have suggested that the induction of epigenetic changes associated with trained immunity relies on concomitant metabolic rewiring[6], although the direction of causality remains unclear. On one hand, both in vitro and in vivo studies have demonstrated that primed cells have altered histone PTM levels at the promoter or enhancer regions of metabolic genes, which could reflect the fact that a primary stimulation induces epigenetic changes that impact the expression levels of metabolic genes which in turn, lead to the detected metabolic changes (in other words, metabolic changes are just a result of epigenetic changes). For example, human monocytes stimulated with beta glucan in vitro have increased histone methylation and acetylation on the promoters of genes involved in glycolysis and the mTOR pathway and the cells themselves have increased aerobic glycolysis[9]. Likewise, BCG vaccination induced changes in H3K27Ac in human monocytes at genes involved in similar metabolic pathways[29].

On the other hand, the byproducts and metabolites produced by metabolic processes can modulate the activity of enzymes that deposit or remove histone PTMs. For example, the TCA cycle metabolite fumarate downregulates the activity of the histone demethylase KDM5[30]. Moreover α-ketoglutarate produced during glutamine metabolism induces Jmjd3 to catalyze demethylation of H3K27me3 at promoters of immune related genes[31], demonstrating the ability of metabolism to regulate epigenetics. Thus, changes to metabolism may be required for changes in epigenetics to occur in the first place. In support of this idea, blocking the AKT-mTOR-HIF1α pathway abrogated BG-induced trained immunity in human monocytes, demonstrating the requirement of metabolic rewiring for trained immunity in this system[9].

Critical questions in the field of trained immunity

Despite the impressive pace at which new discoveries about innate immune memory are made, several key questions remain in the field, spanning from very basic mechanistic questions about the heritability of histone PTM-based memory across cell divisions to more general questions of clinical applicability.

First, on a basic science level, what dictates the longevity of innate immune memory? Mechanistic work linking innate memory phenotypes to epigenetic reprogramming would suggest the loss of innate memory should coincide with the return to a 'baseline' or pre-stimulation epigenetic landscape, however the kinetics by which this occurs have not been fully characterized. This question is made more complicated due to the diversity of possible scenarios. For example, is the experimental setting *in vitro* or *in vivo*, are the cells mitotic, or post-mitotic, and how easily cleared is the priming stimulus? Some information is already known about the *post-mitotic, in vitro* setting. In these settings we know that not all epigenetic modifications are maintained equally well. For instance, H3K27Ac is deposited very quickly onto activated enhancers and promoters[32], but also lost very rapidly (by 4 hours post stimulation in one study) even if the stimulus is still present[19]. In contrast, H3K4me1 has been shown to be a much more stable mark, requiring between 6 to 24 hours to appear[32], but also remaining at high levels at 24 hours of stimulation (compared to 4 hours for H3K27Ac)[28]. Similarly, DNA methylation changes have a slow onset but a high level of stability. In human dendritic cells infected with Mtb, it takes as long as 18 hours before active enhancers first become DNA demethylated, although these enhancers remained demethylated over the course of the 72-hour study period[24]. These kinetics likely differ a lot when comparing cells in culture to those within a mouse. In culture,

cells are isolated from interactions with other cells, including other adaptive immune cells with which there could potentially be crosstalk in an in vivo setting. In vivo, cells are also placed within a more dynamic microenvironment, and primary stimulus removal is less easily controlled. It also remains to be established whether self-renewing cells can independently retain epigenetic changes induced by a primary challenge. During mitosis, histones are displaced from the DNA wrapped around them during passage of the replication fork[33]. As of now, there is no known mechanism by which histone modifications can be faithfully copied from mother cell to daughter cells independently of the presence of TFs. Work in the fission yeast S. pombe demonstrated that H3K9me3 (critical for heterochromatin formation) can be inherited though more than 50 cell divisions only if the demethylase Epe1 was deleted[34,35]. However, in the native context, inheritance of H3K9me3 was shown to require the DNA binding of CREB family TFs indicating a requirement for transcription factor binding even in this case[36]. Although many of the *in vivo* studies discussed in this thesis suggest that some self-renewing cell populations may have the capacity to maintain stimulus-induced epigenetic changes, it remains difficult to prove whether this is due to a cell intrinsic ability to copy histone PTMs from parent to daughter strands via an unknown mechanism, relies on other epigenetic changes such as DNA methylation entirely, or rather is dependent upon cells, cytokines, or low-level stimulus persistence in the microenvironment.

Second, from a clinical perspective, does trained immunity have relevance to humans and generally how is human innate immune system in its entirety modulated by immune challenges? Trained immunity is hypothesized to play a potentially important role in the context of BCG vaccination which has been associated with the induction of heterologous protection against a variety of non-mycobacterial pathogens in multiple clinical trials and observational studies

published between present-day and the early 2000's[29,37–41], suggesting a role for innate immune training. The durability of this protection, reportedly spanning from 6 to 18 months, suggests that if innate immunity is involved, then memory must be induced in cells with the ability to survive and self-renew for at least as long as the window of protection. The human clinical data thus underscores the idea that innate immunity only has long-term clinical relevance if it can be encoded centrally in some form of long-lived innate immune cell or precursor type. New models of trained immunity taking these clinical data into account have proposed that long-lived innate immune memory could be induced within long-lived immune stem cells (referred to as hematopoietic stem and progenitor cells, HSPCs) residing in the bone marrow[5].

Briefly, "HSPC" is an umbrella term referring to all the cells at various stages of hematopoietic differentiation (the process through which a hematopoietic stem cell gives rise to mature cells of the immune system). General models of hematopoiesis have shifted over time. Originally, hematopoiesis was thought to occur through a simple bifurcation model whereby pluripotent HSCs first make a myeloid (innate immune cells, megakaryocytes, and erythrocytes)[42] versus lymphoid (T-cells, B-cells, or NK cells)[42] fate decision by differentiation into either a CMP (common myeloid progenitor) or a CLP (common lymphoid progenitor)[43]. However, it is now widely believed that at the earliest point of lineage choice HSCs choose between the CMP fate or prior to becoming a CLP must transition through an LMPP/MLP state in which they gain bias towards the lymphoid lineage but retain the potential to give rise to myeloid cells. HSCs (the stem cells at the top of the hematopoietic hierarchy) are subcategorized into 3 main types – the LT-HSC, ST-HSC, and MPP, based on their ability to reconstitute the entire immune system of a lethally irradiated mouse long-term[43]. LT-HSCs are defined by their ability to do this for at least 12 weeks, while ST-HSCs and MPP only have temporary reconstitution ability. Although most

of these definitions are based on work in mice, xenograft models involving transplantation of human HSCs into mice have demonstrated similar levels of heterogeneity with rare populations of HSCs having the ability to reconstitute the immune system long-term. Theoretically any HSC can divide symmetrically or asymmetrically giving rise to 2 stem cells, 2 progenitors, or 1 stem and 1 progenitor cell. Differential distributions of TFs within daughter HSCs are believed to dictate their ages[43], which highlights the overall importance of TFs in both the maintenance of HSC stem-like fate and differentiation. Although most studies are based in mouse models, a few key TFs are believed to be critical to maintaining HSC function or in pushing HSCs towards either a myeloid or lymphoid lineage. Within HSCs, HOX TFs are believed to play a critical role in HSC self-renewal[43,44]. Mice deficient in Hoxa9, for example, have severe hematopoietic defects. Notch and Wnt signaling are also generally seen as critical TFs for proper HSC function[44]. Then, in order for differentiation to happen, other TFs must be upregulated. In general, it is believed that the myeloid lineage is the "default" and that in order for lymphoid specification to occur certain TFs must be activated which repress the myeloid program. Indicative of this default myeloid program within HSCs is the fact that HSCs "prime" or lowly express certain lineage-specific TFs and genes including MPO, CEBPα, and MCSFR[44]. During lymphoid development, Ikaros (IKZF1) is believed to play a driving role, both in repressing the activity of myeloid-specific TFs but also in driving lymphoid programs. LMPPs that lack Ikaros, for example, fail to upregulate Flt3, leading to a loss of lymphoid committed CLPs. Ikaros also regulates the activity of Gfi1 which represses PU.1 and drives B-cell development[44]. On the myeloid side, GATA1, PU.1, and the CEBP TFs are critical in determining cell fate. GATA1 is a key TF that drives differentiation in the erythrocyte/megakaryocyte fates. PU.1 when active represses GATA1, preventing MEP differentiation and driving GMP differentiation instead.

GMPs can differentiate into either monocytes/macrophages or into granulocytes and this fate choice depends largely in the relative balance between PU.1 and the CEBP TFs with higher CEBP TF levels driving a granulocytic fate[44].

Within the field of innate immune memory, it is hypothesized that "trained" stem cells continuously give rise to trained innate immune cells, which would explain the functional and epigenetic changes observed in monocytes and whole blood for at least 28 days after vaccination[29,41], while simultaneously accounting for their short lifespan. A plethora of studies in mice have demonstrated that HSPCs are responsive to immune stimuli, including work from our lab which recently showed that intravenous vaccination of mice with BCG led to substantial changes in gene expression within HSCs and MPP of the bone marrow, as well as changes in HSPC cell proportions 4 weeks after vaccination[45]. Almost all current work studying trained immunity in humans has relied on epigenetic and stimulation assays of PBMCs *ex vivo*. Currently, we lack a thorough investigation of the impacts of immune challenges on human HSPCs.

Areas addressed in this thesis

In the second chapter of this thesis, we investigate the impact of BCG vaccination on HSPCs isolated from human bone marrow, with the aim of better understanding how a vaccine can modulate the human innate immune system in a way that is long-lasting. Our experiments help to directly inform whether current models (that the BCG vaccine can induce lasting changes to HSPCs that are transmitted to their progeny) are valid directly in humans and could help re-shape the way we think about the systemic impacts of vaccines beyond the adaptive immune system. In the third chapter we tackle the basic question of whether histone PTMs induced by

immune stimulation, thought to be critical for innate immune memory, can be propagated through multiple cell divisions in macrophages. More generally we explore the dynamic nature of the histone PTM landscape, gene expression, and functional properties of these cells in a dense time course. Together, the work presented in this thesis aims to utilize advanced experimental techniques combined with computational analysis to shed new light on some of foremost questions in the field of trained immunity.

CHAPTER II: BCG VACCINATION IMPACTS THE EXPRESSION AND EPIGENETIC LANDSCAPE OF HSPCs IN HUMAN BONE MARROW

INTRODUCTION

Trained immunity is hypothesized to play an important role in the context of the BCG vaccine. Although BCG is administered for the purpose of protecting against *Mycobacterium tuberculosis* infection, it has also shown effectiveness in other areas, such as in the protection against bladder cancer[46]. Moreover, clinical data suggests that the BCG vaccine may also provide a degree of heterologous protection against non-mycobacterial infections[29,37–41]. Clinical studies conducted in Africa and Europe have reported that BCG vaccination could decrease the risk of death due to secondary non-mycobacterial infections in the range of 6-18 months[37–39]. For instance, studies conducted in both Guinea Bissau and Denmark have reported that children not receiving the BCG vaccine have higher mortality rates due to causes other than Tuberculosis[37,38] and a different study reported that elderly adults receiving the BCG vaccine were significantly less likely to experience a new infection within the next year (HR=0.21, p=0.013)[39]. Overall, the global use of BCG as the only protective measure against Tuberculosis, combined with its other potential clinical uses, highlight the need for a more complete and molecular understanding of how BCG interacts with the immune system in its entirety.

One area of increasing focus has been on understanding how the BCG vaccine interacts with the innate branch of the immune system. Potentially long-lasting impacts of the BCG vaccine on innate immune cells has been hypothesized to underlie some of the vaccine's apparent heterologous protective effects due to the ability of innate cells to respond broadly to a range of

different pathogens. Innate immune cells can be "trained", at least in the short term, to mount stronger responses to heterologous secondary challenge, a phenomenon first shown to occur in human monocytes stimulated with beta glucan[8,9]. The hallmark features of beta glucan-induced trained immunity - reconfiguration to the histone modification and chromatin accessibility landscape and increased pro-inflammatory cytokine responses to heterologous secondary challenge – have also been observed following BCG vaccination. For example, one study found increased H3K27Ac levels in monocytes following BCG vaccination, which was accompanied by increased IL1β production by PBMCs secondarily challenged with Candida albicans[29]. Thus, it has been hypothesized that these BCG-induced changes in innate immune cells, may underlie its clinically observed heterologous protective effects.

One aspect that this model fails to explain is the durability of BCG-induced heterologous protection, reported to last for over a year in some studies[37–39]. In contrast, memory within peripheral blood innate cells would only be expected to last for a few days, given the short lifespan of these cells. To bridge this gap, newer models have proposed that BCG could impact the functional responses of innate immune cells through an indirect route, by inducing a memory-like state within long-lived hematopoietic stem and progenitor cells (HSPCs) that could be encoded in the immune cells they give rise to[5]. The idea that stem cells can harbor inflammatory memory is a paradigm that has been shown in other stem cell types such as within the skin and nasal airways. Epidermal stem cells of the skin primed with the inflammatory agent imiquimod (IMQ) gain long-lasting differences in chromatin accessibility for more than 180 days and are able to close a subsequent punch wound at a significantly faster rate compared to previously unexposed skin[47]. Epidermal stem cells can also migrate from an old niche to a new one following inflammation, yet still retain epigenetic signatures of the original niche from

which they came[48]. Another study found that chronic rhinosinusitis could imprint long lasting transcriptional and functional changes in basal stem cells. Even after 5 weeks in culture, basal stem cells from patients with polyps remained transcriptionally distinct from controls and exhibited tolerance like characteristics with a decreased capacity to respond to cytokine stimulation[49].

Whether BCG vaccination could trigger similar epigenetic memories within HSPCs remains an active area of investigation. Our lab recently showed that intravenous vaccination of mice with BCG led to substantial changes in gene expression within HSCs and MPP of the bone marrow, as well as changes in HSPC cell proportions 4 weeks after vaccination[45]. Moreover, even if the mice were treated with antibiotics following vaccination, bone marrow derived macrophages (BMDMs) from these mice harbored epigenetic, transcriptional, and functional differences for at least 5 months post vaccination, which is supportive of the hypothesis that innate immune training can be encoded at the level of stem cells and can be propagated to mature innate immune cells.

Nonetheless, very little is still known about the relevance of this model in humans, and about the overall impact of BCG vaccination on human bone marrow. Recent work by Cirovic et al. investigated a 20-person cohort from which bone marrow aspirates were collected before and 90 days after intradermal BCG vaccination, enabling a rare look into how BCG vaccination can impact human bone marrow[50]. Since this study was based on bulk RNA-seq performed on HSPCs sorted from these bone marrow samples, it has opened the door for many additional questions, such as which HSPC cell types are most impacted, and whether these expression changes are also coupled to epigenetic changes.

Here we combined single cell analyses, flow cytometry, functional analyses, and computational approaches performed on the samples from Cirovic et al. to address these questions. We performed droplet-based scRNA- and scATAC-sequencing on the human bone marrow aspirates from all 20 healthy individuals involved in the study, both before and 90 days after intradermal BCG vaccination or placebo. Even 90 days after a single intradermal vaccination, we found that the most uncommitted stem cells exhibited multiple hallmarks of granulocyte/neutrophil bias, which was accompanied by higher percentages of classically defined CMPs in BCG compared to placebo individuals. On the epigenetic level, we identified over 2000 sites of differential chromatin accessibility across multiple CD34 subpopulations. Specifically, within progenitor clusters, these peaks of differential chromatin accessibility were enriched for motifs of KLF/SP and EGR transcription factors which were predominantly active within upstream HSCs, suggesting that long-lasting TF activity and differential gene expression at the level of HSCs may impact the chromatin accessibility landscape of downstream progenitors. Within individual donors, the extent of myeloid bias and the expression levels of a core set of BCG-induced genes and transcription factors within HSCs were found to significantly correlate with increased Il1β secretion of donor paired PBMCs in response to a C. albicans challenge. BCG-induced changes in chromatin accessibility within downstream GMPs were also predictive of IL1B production capacity demonstrating that BCG vaccination induces both a protracted period of baseline activation within HSCs, myeloid skewing, and the accumulation of epigenetic memories in downstream progenitors, all directly correlated with changes in the production of cytokines by donor-matched PBMCs. These data indicate that BCG vaccination re-wires transcription factor activity, gene expression, chromatin accessibility, and lineage bias in human bone marrow in a

way that is linked to responses of PBMCs to secondary immune challenge with non-mycobacterial pathogens.

RESULTS

Single cell analysis of human bone marrow

Bone marrow aspirates and PBMCs were collected and cryopreserved from a cohort of 20 BCG-naïve volunteers prior to (D0) and 90 days following (D90) intradermal vaccination with BCG (n=15) or placebo (n=5) (Figure 1.1a). To specifically isolate hematopoietic stem and progenitor cells (HSPCs) from each bone marrow sample, we stained bone marrow aspirates with fluorescence-conjugated antibodies targeting CD34, a transmembrane phosphoglycoprotein specific to HSPCs[51]. We also stained all bone marrow aspirates with a panel of antibodies targeting canonical markers (CD3, CD56, CD14, etc. for mature immune cells and CD90, CD10, CD110, etc. to distinguish between CD34+ HSPC subtypes; 16 total markers, Table 1.1). We then used fluorescence activated cell sorting to sort out live, CD34+ cells for downstream droplet-based scRNA-seq and scATAC-seq processing while simultaneously collecting flow cytometry data (Figure 1.1b).

BCG vaccination leaves a lasting impact on gene expression within HSPCs

We started by first broadly asking whether BCG vaccination had a lasting impact on the gene expression landscape of each HSPC subtype. Our initial scRNA-sequencing data set contained 115,698 cells across all samples, which we filtered down to 92,014 high quality cells to retain for subsequent analyses. We clustered these cells into 23 initial groups, and then condensed these

groups into 13 non-overlapping clusters which we assigned to known HSPC cell subtypes (HSCs, CMPs, MLPs, GMPs, MEPs, and Pre-BNK cells) (Figure 1.1c) based on similarity to a pre-labelled reference. To verify our cluster assignments, we plotted the expression levels of known lineage-specific genes including GATA1 (erythroid), DNTT (lymphoid), MPO (myeloid/neutrophilic), and HOXA9 (stem) whose expression was restricted to specific clusters with matching cell-type labels (Figures 1.1d-g).

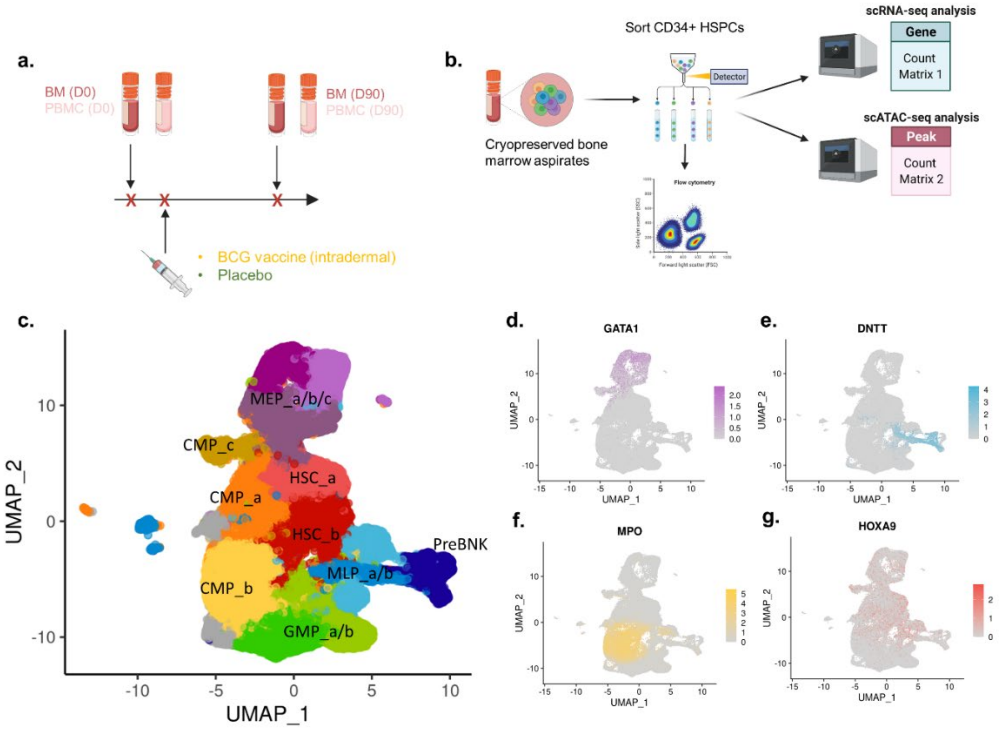

Figure 1.1. Bone marrow and PBMC sample processing

a. Overview schematic of experimental timeline and samples collected. Bone marrow aspirates and PBMCs were collected from 20 total donors on D0 prior to BCG (n=15) or placebo (n=5) and D90 (90 days after BCG or placebo). **b.** Schematic showing processing steps for all bone marrow samples. Cryopreserved bone marrow samples were stained with a cocktail of lineage and HSPC-specific antibodies to enable flow cytometric analysis of cellular composition as well

Figure 1.1, continued

as simultaneous sorting of all CD34+ cells. Sorted CD34+ cells were immediately processed for scRNA-seq and scATAC-seq according to the respective 10X genomics protocols. **c.** UMAP of the scRNA-seq data collected from bone marrow CD34+ HSPCs of BCG or placebo vaccinated individuals at D0 or D90. Clustering based on gene expression grouped cells into 13 non-overlapping clusters: HSC_a (n=9637), HSC_b (n=10953), CMP_a (n=9174), CMP_b (n=14918), CMP_c (n=1715), GMP_a (n=6631), GMP_b (n=6423), MEP_a (n=8871), MEP_b (n=5439), MEP_c (n=3811), MLP_a (n=5153), MLP_b (n=3837), PreBNK (n=3371). **d-g.** UMAP colored by expression levels of lineage defining genes (D. GATA1 – MEP; E. DNTT – lymphoid; F. MPO – myeloid/neutrophilic; HOXA9 – stem)

Antibody	Fluorochrome
L/D	L/D fixable blue
CD16	PE Alexa Fluor 700
CD14	Spark Nir 685
CD3	PerCP
CD56	PerCP eFluor 710
CD45RA	BUV496
CD38	BUV737
HLA-DR	V500
CD34	PE
CD33	PE-Vio770
CD45	BV650
CD4	BUV615
CD20	BV605
CD90	APC
CD10	FITC
CD110	BV421
CD8	BV711

Expected markers for each population	
CD4	L/D- CD45+ CD4+ CD8- CD20-
CD8	L/D- CD45+ CD4- CD8+ CD20-
B-cell	L/D- CD45+ CD4- CD8- CD20+
classical monocyte	L/D- CD45+ CD3- CD20- CD14+ CD16-
nc monocyte	L/D- CD45+ CD3- CD20- CD14+ CD16+
NK cell	L/D- CD45+ CD3- CD20- CD14- CD56+
DC	L/D- CD45+ CD3- CD20- CD14- HLA-DR+
HSC	L/D- CD34+ CD38- CD45RA- CD90+
MPP	L/D- CD34+ CD38- CD45RA- CD90-
CMP	L/D- CD34+ CD38+ CD45RA-
CLP	L/D- CD34+ CD38lo CD45RA+ CD90- CD10+
MLP	L/D- CD34+ CD38- CD45RA+ CD90-
GMP	L/D- CD34+ CD38+ CD45RA+ CD10-
MEP	L/D- CD34+ CD38lo CD45RA- CD110+ CD33-

Table 1.1. Antibody panel for bone marrow samples and cell type-defining surface markers

Cryopreserved bone marrow aspirates from each donor/timepoint were stained with the panel of fluorochrome conjugated antibodies shown on the left. Mature immune and HSPC cell types were defined based on the presence or absence of the cell surface proteins shown on the table to the right.

We then performed a differential gene expression analysis separately for each of the 13 unique clusters. For this analysis we collapsed single cell expression counts into pseudobulk expression (Figure 1.2a) to allow a bulk RNAseq-like approach. Since measured gene expression may naturally shift over time, we focused on genes whose across-time change in gene expression (D90 vs. D0) differed between the BCG vaccinated and placebo cohorts (Figure 1.2b). We refer to these genes as being *differentially regulated (DR)* across time due to BCG vaccination. We quantified the number of DR genes in the positive and negative directions independently for each cell-type (Figure 1.2c) which revealed that BCG effects on gene expression were generally upward biased and heterogenous, with a disproportionate percentage of DR genes residing within the most stem-like clusters, HSC_a and HSC_b (Figures 1.2c, d). Strikingly, in HSC_a and HSC_b, more than 200 and 150 genes respectively were significantly differentially regulated 90 days following BCG vaccination, even when using a stricter lfsr < 0.01 cutoff. Thus, these data demonstrate that a single intradermal BCG vaccination impacts the gene expression landscape of HSPCs in the bone marrow for at least 3 months.

To determine enriched pathways among DR genes we performed a gene set enrichment analysis (Figure 1.2e). We identified 27 Hallmark pathways that were enriched in at least one cluster ($p < 0.05$). Most enriched pathways were generally representative of either immune (blue), metabolism (red), or proliferation/apoptosis (green) pathways. The immune-related pathways, IL2/Stat5 signaling, complement, and inflammatory response, were significantly positively enriched (padj < 0.1) predominantly within HSC clusters, demonstrating a baseline activated immune state within stem cells. The 'TNFα via NFκB signaling pathway', however, was significantly enriched in HSC, CMP, GMP, MLP, and Pre-BNK clusters, indicating that immune gene expression related to NFκB signaling was impacted across multiple HSPC subtypes. The

metabolism-related pathway ‘oxidative phosphorylation’ also had strong, positive enrichments across almost all clusters, in line with previous observations that immune challenges can induce persisting metabolic changes[5,6,9]. We observed a dichotomy when comparing metabolic pathway enrichments in MEPs with that of HSCs and other non-MEP progenitor clusters. For example, HSCs and GMPs had no enrichments in glycolysis or MTOR signaling, but instead had significant (padj < 0.1) enrichments in stress-related pathways (hypoxia, reactive oxygen species, and apoptosis pathways). In direct contrast, MEPs trended towards pro-glycolytic and pro-proliferative enrichments (glycolysis and MTOR signaling and MYC signaling), but for the most part did not have significant enrichments of ROS, hypoxia, or apoptosis pathways, indicating that BCG vaccination heterogeneously rewires the metabolism of HSPCs. For a general comparison of pathway enrichments between HSPC subtypes, we performed a principal component analysis on each cluster using gsea pathway enrichment scores as input (Figure 1.2f). This revealed a tight and distinct clustering of HSC_a with HSC_b and of MEP_a with MEP_b and MEP_c, away from the zero-reference point. HSCs and MEPs, the cell types whose gene expression landscapes were the most impacted by BCG vaccination, clustered on opposite ends of PC1, further demonstrating the differential impact of BCG vaccination on these cell types. The other progenitor clusters, including CMPs, GMPs, MLPs, and PreBNK cells clustered together, closer to the zero-reference point, but in the negative direction of PC1, suggesting greater similarity to HSCs, compared to MEPs. These data collectively suggest that different HSPC cell types, namely HSCs and MEPs, may have a stronger propensity to maintain a lasting state of activation, compared to others. We note that of enriched pathways within the other myeloid and lymphoid progenitor clusters, very few were unique to only one cluster. Rather most were shared across all cell types. Thus, given that CMPs, GMPs, MLPs, and PreBNK cells appear to predominantly

harbor differences in gene expression that are shared with HSCs, we speculate that gene expression differences within these myeloid and lymphoid clusters may be retained as a form of transcriptional memory from activation within upstream HSCs. Overall, our data indicate that BCG vaccination variably impacts the expression of immune, metabolism, and proliferative genes across HSPC subtypes and has the strongest direct impact on HSCs and MEPs.

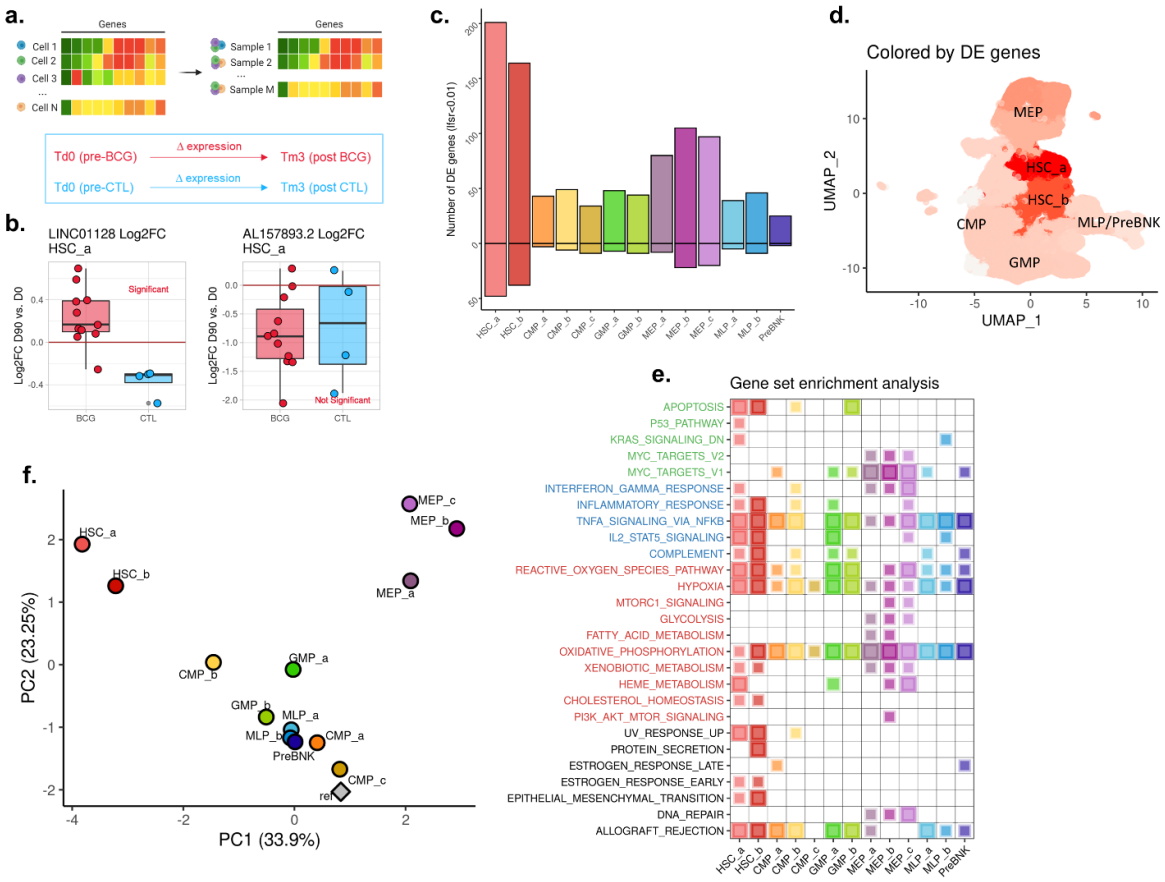

Figure 1.2. BCG vaccination has heterogenous impacts on gene expression after 90 days

a. Schematic showing the general scRNA-seq analysis approach. Raw 'CELL x GENE' UMI counts generated through the Seurat pipeline were transformed into 'SAMPLE x GENE' pseudobulk matrices for each cell-type/cluster. Pseudobulk expression was fit to a linear model that estimates and corrects for natural expression changes across time in placebo individuals and

Figure 1.2, continued

allows identification of BCG-specific effects on gene expression. **b.** Example boxplots showing a gene (LINC01128) for which BCG vaccination had a significant differential impact on expression compared to placebo and a non-significant gene (AL157893.2) that exhibited similar across-time changes in expression in both placebo and BCG vaccinated individuals. **c.** Bar graph summarizing the total number of significant genes (lfsr<0.01) in each cell type. Bars extending in the positive and negative direction indicate genes whose expression was impacted positively and negatively, respectively, by BCG vaccination compared to placebo. Bars are color-coded by cell type. **d.** UMAP colored to indicate the number of significant genes in each cluster as shown in c. Darker red colors indicate higher total numbers of significant genes. **e.** Summary plot of gene set enrichment analysis (GSEA) performed separately for each cell type. Genes were ordered by the rank statistic –log10(pval)*logFC and compared against Hallmark gene sets. Pathway names are manually color-coded according to category (green: apoptosis/proliferation, blue: immune, red: metabolic, black: other). Square size is scaled to –log10(pval). All shown squares are pathways with p<=0.05. All squares with border have padj<=0.1; squares are color-coded by cell type. **f.** Principal component analysis showing cell types clustered by Hallmark pathway enrichment (NES) scores as computed for GSEA in E. "Ref" point (gray diamond) is a vector of zeros, representing a baseline unaffected state with no enrichment of any pathway. NES values for pathways with p>0.05 were set to 0.

BCG vaccination impacts lineage bias of HSPCs

We next focused our analyses on the HSC clusters, HSC_a and HSC_b, which harbored the largest overall changes in gene expression. We asked whether BCG vaccination was associated with any systematic change in HSC lineage bias, given that DR genes within HSCs were related to NFκB, Stat5 activity, and ROS pathways (Figure 1.2e), which are known to be involved in myelopoiesis-like responses to bacterial infections[52–55]. Previous murine and human studies have reported that BCG vaccination can induce acute emergency myelopoiesis[56,57], although its persistence has not been thoroughly investigated. Thus, we asked whether long-lasting, inherent myeloid-leaning bias within HSCs was directly detectable 90 days after BCG vaccination. To computationally predict lineage bias we utilized CellRank[58], a similarity-based trajectory

inference method that utilizes RNA-velocity information (measurements of unspliced to spliced mRNA) to infer developmental directionality. Applying CellRank to HSCs in each sample, we computed the most likely terminal state (CMP, GMP, MLP, MEP, or Pre-BNK) for each cell (Figure 1.3a). Then for each individual we determined the percentage of HSCs predicted to differentiate into each terminal state and found the difference across time (Figure 1.3b). Even 90 days after a single intradermal vaccination, BCG vaccination had a significant ($p = 0.026$) and exclusive positive effect on differentiation biases towards the CMP_b terminal state (Figure 1.3c-h). This was also true when directly comparing the percentage of CMP_b biased HSCs in individuals of the BCG and placebo cohorts at day 90 (Figure 1.3i, $p = 0.018$). Although not reaching a p-value threshold of 0.05, individuals in the BCG-vaccinated cohort also had relative, positive shifts in overall CMP ($p = 0.104$; Figure 1.3f) bias and a compensatory decrease in erythroid bias ($p = 0.138$; Figure 1.3g).

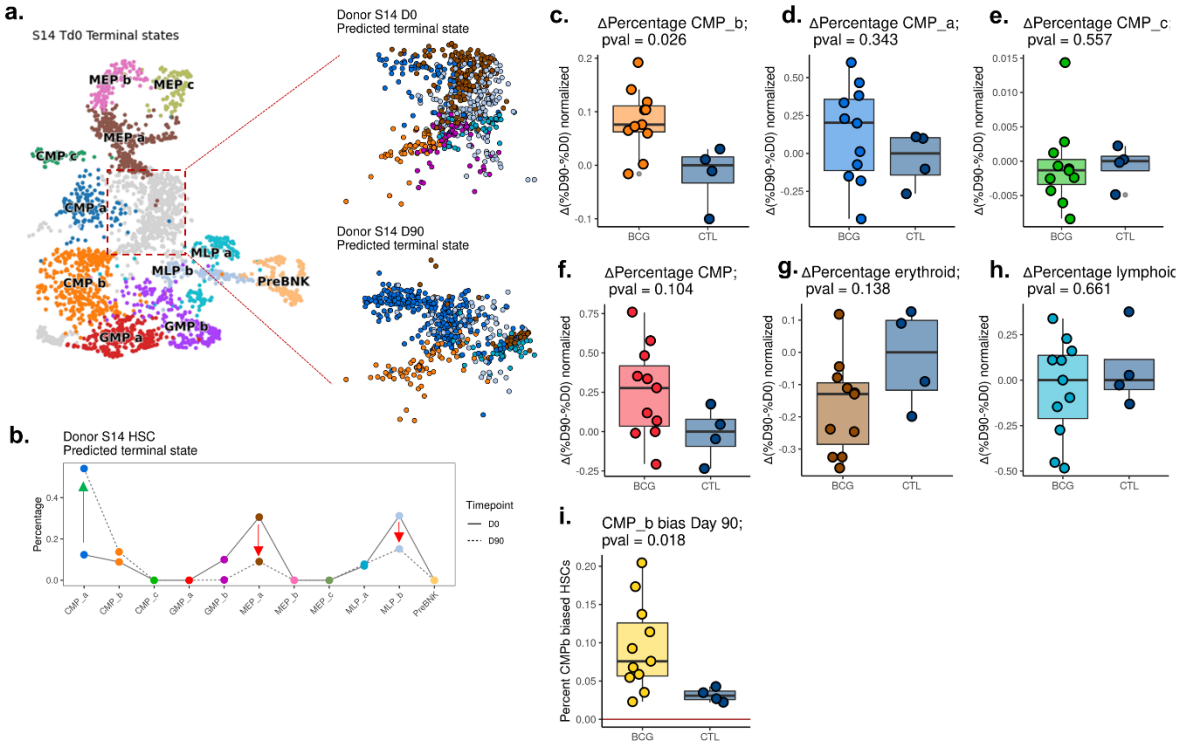

Figure 1.3. BCG vaccination increases granulocyte bias of HSPCs

a. Example CellRank analysis for donor S14. Terminal fates of individual HSCs (cells in HSC_a or HSC_b clusters; colored grey within the red dotted box) were predicted with CellRank at D0 and D90 for each donor. The righthand side of the figure shows HSCs color-coded by predicted terminal fate (orange: CMP_b, dark blue: CMP_a, brown: MEP_a, purple: GMP_b, light blue: MLP_b, turquoise: MLP_a) at D0 (top) and D90 (bottom). **b.** Line graph showing the percentage (y-axis) of HSCs from donor S14 predicted to differentiate into each possible fate (x-axis) at D0 (solid line) and D90 (dotted line). **c-h.** Comparison of the across-time change (%D90-%D0) in the percentage of biased-HSCs for different terminal fates. Figures D-F show data for CMP_b bias, CMP_a bias, and CMP_c bias respectively. Figures G-I show combined results for broad cell types: CMP = CMP_a + CMP_b + CMP_c; erythroid = MEP_a + MEP_b + MEP_c; lymphoid = MLP_a + MLP_b + PreBNK. All values are shown normalized to the median of the control group. **i.** Absolute percentages of CMP_b biased HSCs for each donor at D90 (p=0.018).

CMP_b, the myeloid cluster towards which HSCs were significantly biased, expressed clear neutrophilic/granulocytic signatures on the gene expression level, suggesting that long-lasting BCG-induced myelopoiesis was specifically oriented towards production of granulocytes and neutrophils (a long-lasting emergency granulopoiesis-like state). The granulocytic identity of CMP_b was exemplified by the exclusive expression of MPO (Figure 1.4d, e) and the strong expression of CSF3R (Figure 1.4b). Moreover, CMP_b displayed a CEBPA to CEBPB expression gradient (Figure 1.4a,c) in agreement with the known essential role in CEBPA in initial neutrophil differentiation and the later developmental roles of CEBPB, which largely replaces CEBPA activity[52,59]. We investigated whether there was evidence that HSCs may be responding to any bacterial remnants or PAMPs remaining at D90 by looking at the expression levels of cytokines IL6, IL3, IL1B, G-CSF, TNF, IFNG, and GM-CSF, which are produced by non-hematopoietic cells or by HSPCs themselves upon bacterial infection or LPS challenge[52,60,61]. However, in our data set we detected no differential gene expression of any of these cytokines in any HSPC cell type. Most were expressed at such low levels that they were excluded from the original gene expression matrix and not included in our linear model (Figure 1.4f). Cytokines IL3, IL6, IFNG, and G-CSF had negligible expression in all HSPC clusters and GM-CSF had appreciable expression only in the HSC_a and MEP_a clusters but with a padj value > 0.9 when comparing expression between the BCG and placebo cohorts. Likewise, several clusters expressed some IL1B and TNF, but we detected no differences in expression between cohorts, indicating that HSPCs themselves are likely not acutely responding to bacterial challenge at the 90-day timepoint.

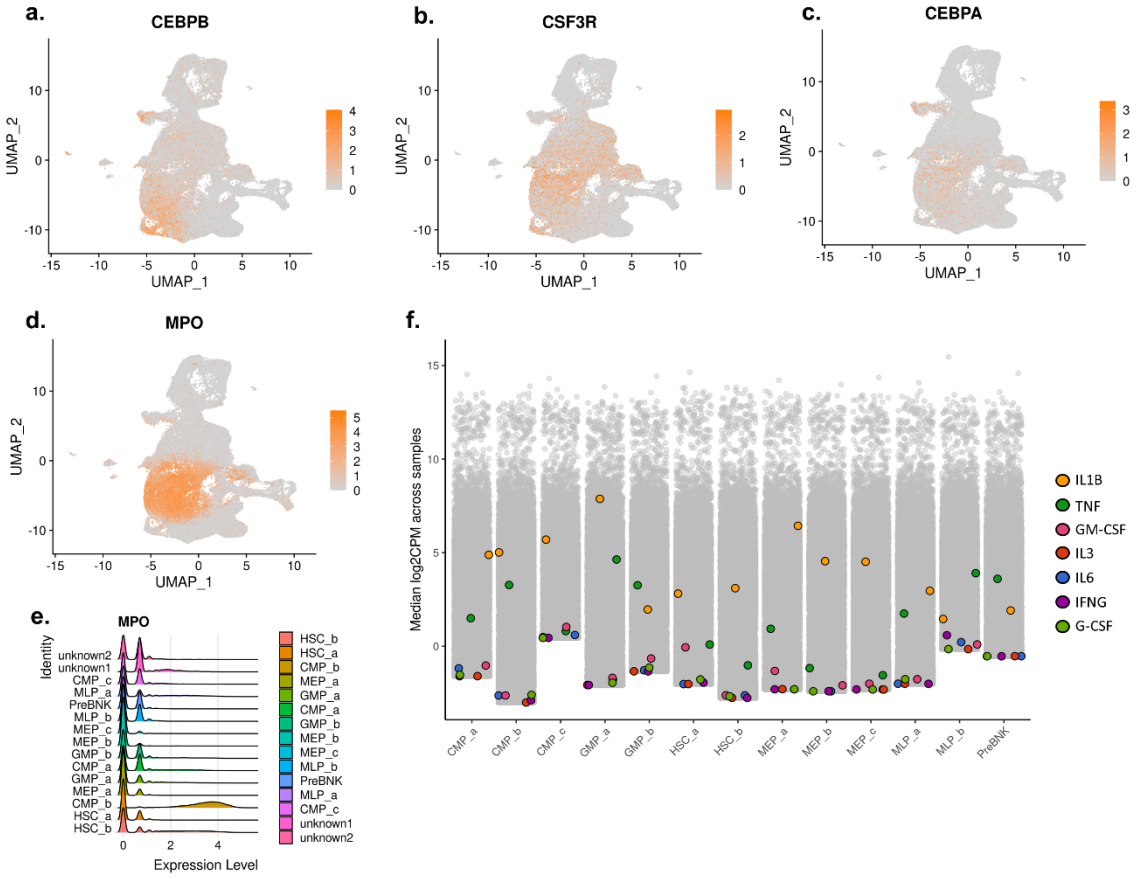

Figure 1.4. The CMP_b cluster is granulocytic

a-d. UMAPs colored by expression levels of granulocyte-specific genes (**a.** CEBPB; **b**. CSF3R; **c.** CEBPA; **d.** MPO) **e.** Comparison of MPO expression levels across clusters. **f.** Raw, normalized log2CPM expression levels of pro-inflammatory and myeloid-differentiation cytokines (IL1B, TNFA, GM-CSF, IL3, IL6, IFNG, G-CSF) in each cluster.

Despite finding no changes in cytokine expression, several genes that are involved in promoting neutrophil development, or whose expression is known to change during emergency granulopoiesis were differentially expressed in HSCs 90 days post-BCG (Figure 1.5a-e). These included genes such as KLF6 (Figure 1.5c) and IRF1 (Figure 1.5d), which are transcription factors that promote neutrophil development and reach maximal expression in mature

neutrophils[62,63]. We detected decreased expression of MAPK14 (Figure 1.5b), a prototypical p38 MAPK, in line with reports that p38 MAPKs inhibit granulopoiesis, unlike their MAPK counterparts ERK and JNK[64]. Importantly we detected a near-significant increase in gene expression of CEBPB (Figure 1.5e, lfsr = 0.129), the transcription factor most critical for driving a granulopoiesis response[52,59], and a significant increase in suppressor of cytokine signaling 5 (SOCS5, Figure 1.5a) which is in the same family as SOCS3, also known to be upregulated during emergency granulopoiesis[65].

To further validate our computational findings of BCG-induced granulopoiesis, we looked within our previously collected flow cytometry data (Figure 1.1b, Table1.1) to see whether proportions of CD34 cell subtypes as defined by classical cell surface markers were also altered between BCG and placebo vaccinated groups. In agreement with the myeloid/CMP biased differentiation predicted by CellRank, we found a significant increase in the number of CMPs in the bone marrow of BCG vaccinated versus placebo volunteers at D90 (Figure 1.5f) and a non-significant trend towards increased GMPs (Figure 1.5i). Although we did not find a statistically significant difference in MEPs, we found significantly lower numbers of CLPs and a trend towards decreased MLPs in BCG vaccinated individuals, confirming that the granulopoiesis-like phenotype we detected computationally using CellRank also manifested in increased numbers of classically defined myeloid progenitors at the expense of other lineages (Figure 1.5g, h). Notably, although we combined LT-HSCs, ST-HSCs, and MPPs under the composite label 'HSC' in our single cell dataset, we were able to approximate changes in MPP percentages in our flow cytometry data, since our antibody cocktail contained MPP defining cell surface markers ($CD34^+$ $CD38^-$ $CD45RA^-$ $CD90^-$). In the flow cytometry data, percentages of MPPs were also higher ($p = 0.0945$) in the bone marrow of BCG vaccinated individuals, resembling the

inflammation induced MPP expansions reported in mice challenged with LPS, beta glucan, and BCG[38,61] (Figure 1.5l).

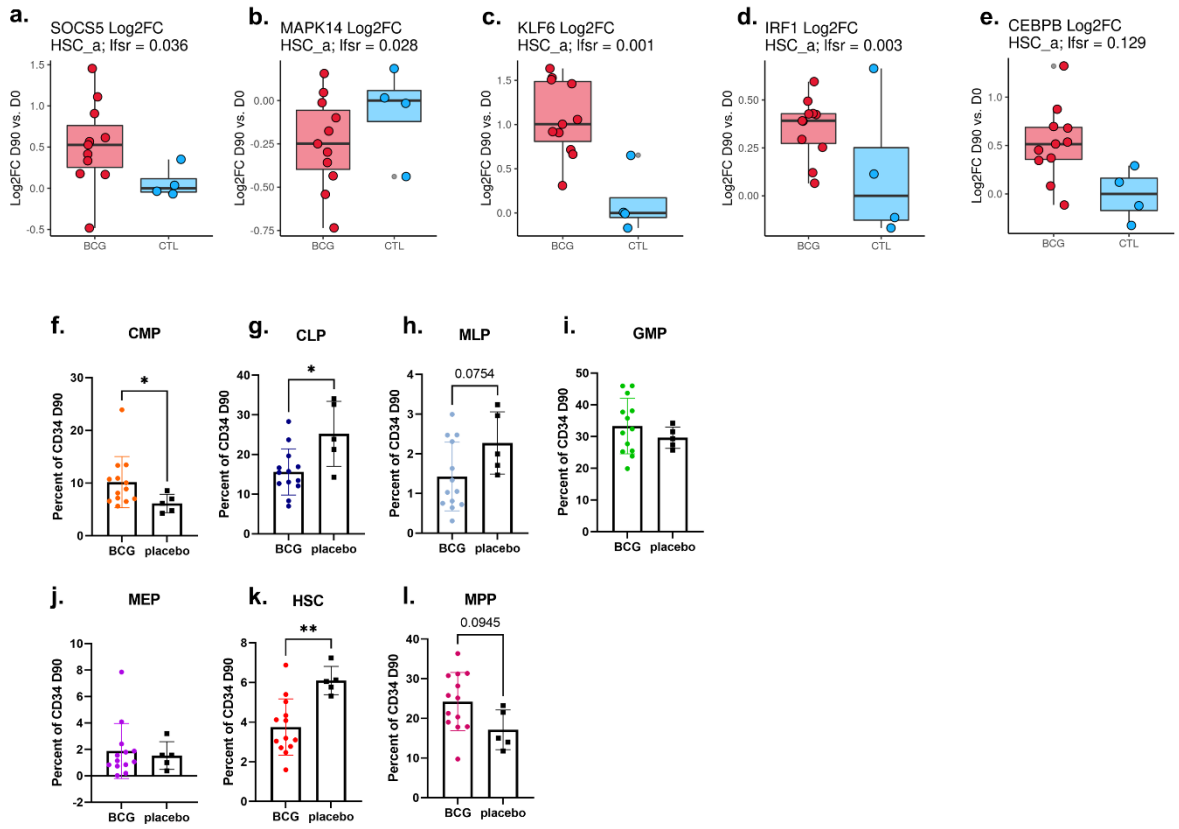

Figure 1.5. Genes involved in granulocyte development are differentially expressed in HSCs; Altered proportions of HSPC cell types determined by flow cytometry

a-e. D90 vs. D0 Log2FC expression of select genes with reported roles in neutrophil development or granulopoiesis for each donor. **f-i.** Bar graphs showing the percentage of each cell type among live CD34+ HSPCs at D90 as determined by flow cytometry analysis.

These lineage-bias analyses suggested that differentially expressed genes within HSCs as a whole could be explained by two possible scenarios (Figure 1.6a). First, BCG vaccination could induce shifts in HSC subtype proportions (here a subtype being defined by HSC lineage bias) without inducing novel changes in gene expression *within* any lineage-biased group. More specifically, overall changes in gene expression could simply be the result of increased proportions of CMP_b biased HSCs, even if gene expression programs *within* CMP_b biased HSCs was unaffected by vaccination. Alternatively, BCG vaccination could induce both a shift in subtype proportions, and within-group changes in gene expression.

To differentiate between these two models, we subset HSCs by their CellRank predicted terminal fates (Figure 1.6b). Then within each terminal fate group we performed a differential gene expression analysis as we had previously done for the broader HSC_a and HSC_b groups in Figure 1.2a and b. Although some of the terminal fate groups did not contain enough Td0 and Tm3 samples from both placebo and BCG vaccinated individuals to perform a reliable differential gene expression analysis, the biggest HSC subgroups – those with bias towards CMP_a, CMP_b, MLP_a, or MLP_b, all harbored significant differences in gene expression (Figure 1.6c). Apart from MLP_b, the other three clusters harbored more than 200 DE genes as a result of BCG vaccination, showing that differential gene expression detected within HSCs as a whole is reflective of both shifts in composition, and within subtype changes in gene expression.

More generally, the clear heterogeneous nature of the HSCs prompted us to perform further unbiased sub clustering of the HSCs. We subset all cells belonging to clusters HSC_a and HSC_b and then clustered the cells into subgroups based on the most variably expressed genes, forming 10 total HSC subtypes (Figure 1.6d, e). Several of the HSC subgroups defined by this

unbiased gene expression-based clustering approach contained a majority of cells with a single lineage bias (Figure 1.6f, Figure 1.7). For example, most cells within subcluster 8 were biased towards the MLP_b lineage, more than 75% of cells within subcluster 4 had CMP_a bias, and the majority of CMP_b biased cells overlapped with subcluster 2, demonstrating that many HSCs with different lineage biases have different baseline gene expression programs and therefore can be naturally grouped into different clusters. However other subclusters contained a mixture of cells with different lineage biases. For example, subclusters 0 and 9 contained almost evenly split proportions of MEP, MLP, and CMP biased cells (Figure 1.7), indicating that it is also possible for cells with very similar baseline gene expression programs to harbor potentially stochastic differentiation velocities towards different lineages.

Since there already exist known sub-classifications of HSCs (LT-HSC, ST-HSC, MPP) we also assessed whether our 10 HSC subclusters could be roughly identified as one of these three known groups. The standard markers used to differentiate between MPPs and LT-HSCs/ST-HSCs are a lack of the cell surface proteins CD90, CD49f, and CD45RA[51,66]. Although we lacked information on protein levels in our scRNA-seq dataset, we found that expression of CD90, CD49f, and CD45RA was variable across the different HSC subclusters (Figure 1.6g) and lowest within subclusters 2 and 8, indicating that these subclusters are most likely MPPs. Since the vast majority of CMP_b biased cells fell within subcluster 2 and BCG vaccination increases the proportion of these CMP_b biased cells, this indicated that BCG vaccination specifically promotes increased CMP_b biased *MPPs*, in agreement with our flow cytometry data in which we detected increased percentages of MPPs in BCG vaccinated compared to placebo individuals (Figure 1.5l), and with previous work in mouse models demonstrating that among "subtypes" of HSCs (LT-HSCs, ST-HSCs, or MPPs) inflammation and infection often induces the most clear

effects and expansions on MPPs[45,67]. Differential gene expression performed on each subcluster identified more than 50 DE genes, mostly in the upward direction, within each HSC subgroup. Notably, PCA to compare differentially expressed pathways between different subtypes demonstrated clear separation between the MPP-like cluster c2 (dominated by CMP_b biased cells) and the MPP-like cluster c8 (dominated by MLP_b biased cells), suggesting that the total differential gene expression program detected within the broad HSC_a and HSC_b clusters is actually reflective of heterogenous differences in expression across smaller HSC subclusters and that HSC subclusters dominated by CMP_b biased cells, compared to those dominated by MLP_b biased cells, are impacted differently by BCG vaccination. Collectively, these data indicate that BCG vaccination induces differential immune, stress, and metabolism gene expression most predominantly within HSCs and MEPs, as well as a granulopoiesis-like bias inherent within HSCs manifesting as increased proportions of CMP_b biased MPPs. This suggests that a single intradermal BCG vaccination induces a protracted period of increased immune activation leading to some changes in gene expression within all cell types but prominent activation signatures within HSCs in human bone marrow.

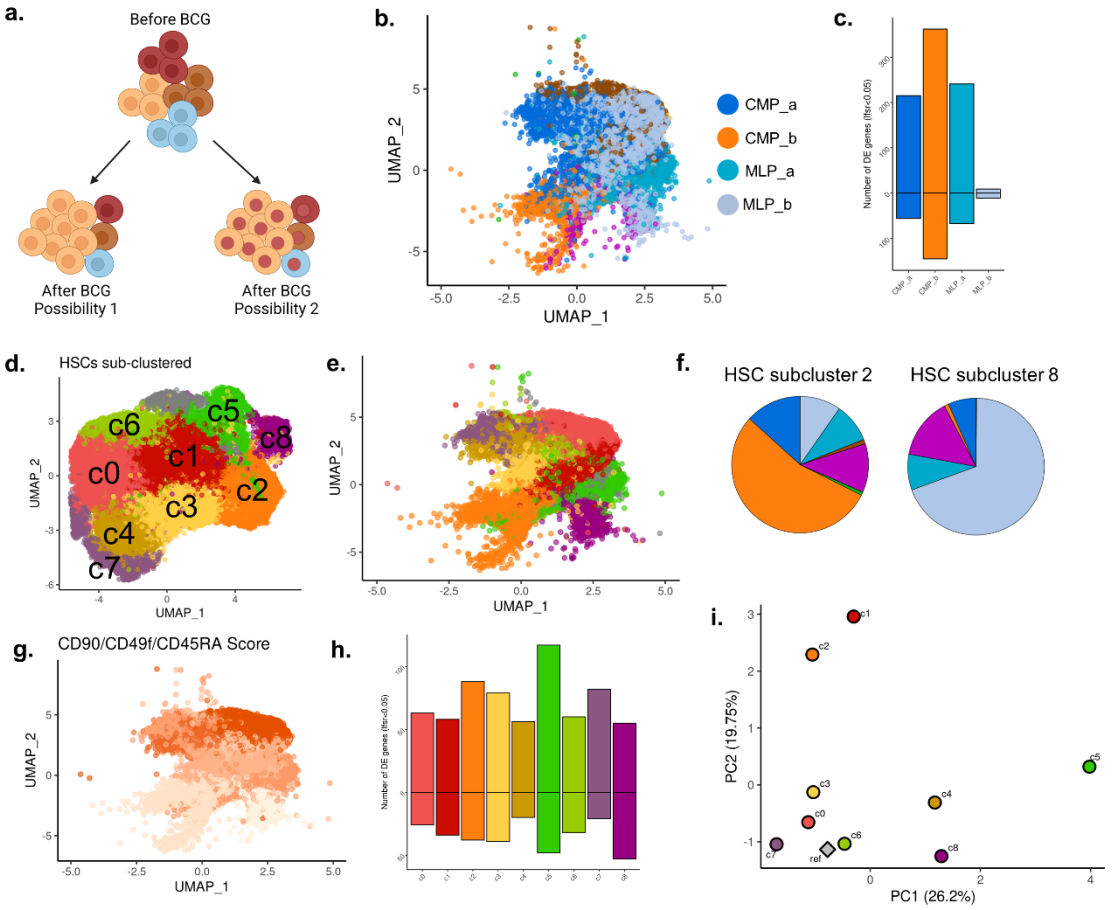

Figure 1.6. BCG induces differential expression within HSCs of the same lineage bias

a. Proposed models for how BCG vaccination impacts total gene expression within HSCs. In proposed model 1 (left, bottom), BCG vaccination alters the proportion of HSCs biased towards the CMP_b lineage but does not induce gene expression changes within any given lineage biased group. The net effect is still a global change in expression due to the compositional shift. In proposed model 2 (right bottom) BCG vaccination induces changes in lineage bias in addition to changes in gene expression within any lineage biased group, leading not only to increased numbers of CMP_b biased cells, but also changes in gene expression among CMP_b biased HSCs. **b.** UMAP of all cells classified as HSC_a or HSC_b. Cells were grouped and colored according to the predicted terminal fate assigned b CellRank in Figure 3a. **c.** The number of differentially expressed genes within each lineage-biased subgroup. Colors of the bars match the colors in b. Upward and downward bars indicate increased or decreased gene expression due to BCG vaccination respectively. **d.** UMAP of all HSCs (HSC_a and HSC_b) subclustered using the Seurat pipeline. **e.** Subclusters projected onto the original UMAP **f.** Pie charts indicating the proportion of cells within each subcluster with different predicted terminal fates. The pie charts show example subclusters which contained a majority of cells with a single predicted terminal fate. Subcluster 2 contained more than 50% of CMP_b biased HSCs and the majority of cells in subcluster 8 were biased towards the MLP_b lineage. **g.** Subclusters colored by CD90/CD49f/CD45RA (average z-score across the three genes). Darker shading indicates a

Figure 1.6, continued

higher score. Subclusters with lighter shading (2 and 8) are predicted to be more MPP-like since MPPs are defined by their lack of all three surface markers on the protein level. **h.** The number of differentially expressed genes within each subcluster. **i.** PCA of each subcluster based on Hallmark gene set enrichment scores. Subclusters are shown relative to a reference "zero enrichment" cluster (grey diamond).

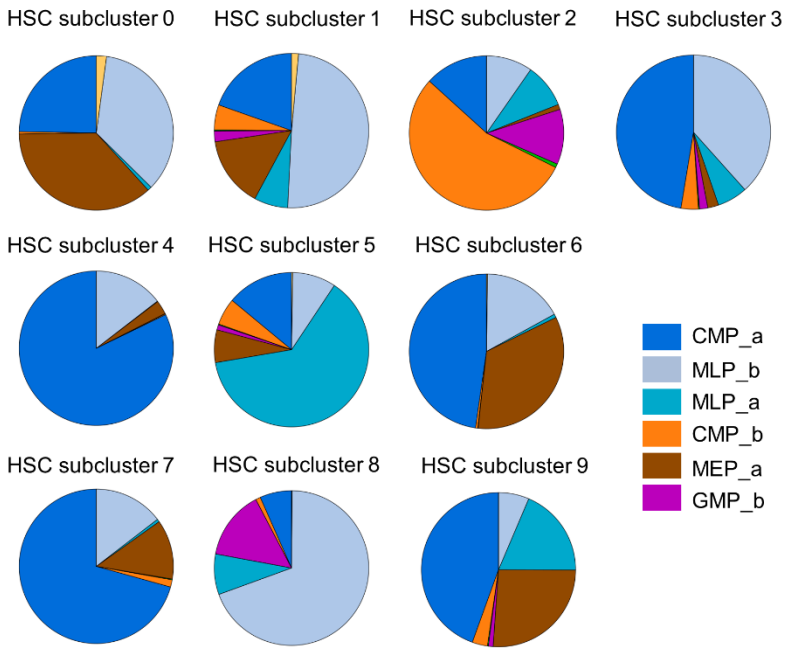

Figure 1.7. Lineage bias composition of each HSC subcluster

Pie charts showing the proportion of HSCs within each subcluster with Cellrank predicted lineage bias towards the CMP_a, CMP_b, GMP, MEP, MLP_a, or MLP_b fate.

BCG vaccination impacts the chromatin accessibility of immune progenitors

Since changes in gene expression are often coupled to changes in the epigenome, and because epigenetic alterations are believed to be central to innate immune memory[6], we next investigated whether BCG vaccination was associated with changes in the epigenetic landscape of HSPCs using the scATAC-sequencing data we collected on bone marrow before and after BCG vaccination, in parallel with our gene expression data (Figure 1.1b). We retained 58,988 total high-quality cells in this dataset, for an average of 1,552 cells per sample. We clustered these cells into 16 primary groups (Figure 1.8a) and labeled the clusters according to cell type using 'gene activity' scores calculated for each gene based on the accessibility of nearby peaks (Figure 1.8b).

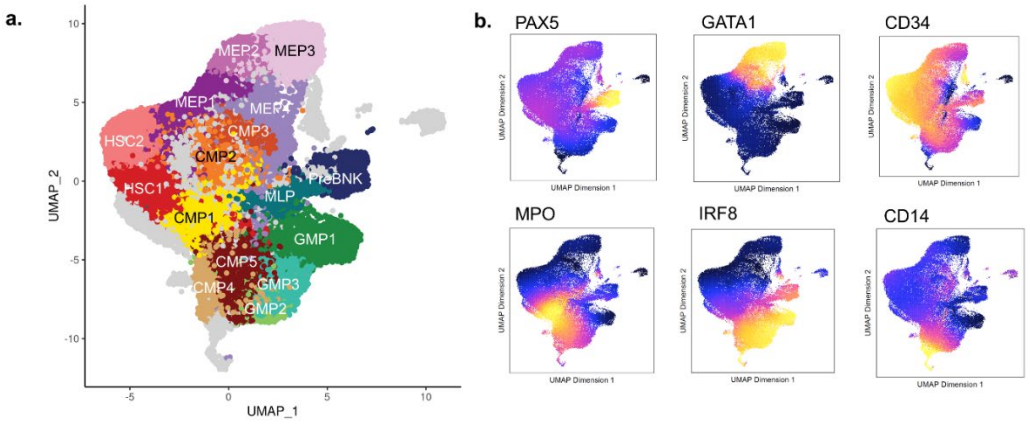

Figure 1.8. scATAC-sequencing profiles on human bone marrow

a. UMAP of the scATAC-seq data collected from bone marrow CD34+ HSPCs of BCG or placebo vaccinated individuals at D0 or D90. Clustering based on chromatin accessibility grouped cells into 16 clusters: CMP1 (n=5041), CMP2 (n=2942), HSC1 (n=5750), HSC2 (n=5052), MEP1 (n=4436), PreBNK (n=1876), GMP1 (n=3798), CLP (n=2146), MEP2 (n=2653), MEP3 (n=3886), CMP3 (n=1458), MEP4 (n=2230), CMP4 (n=1952), CMP5

Figure 1.8, continued

(n=5590), GMP2 (n=962), GMP3 (n=2806). **b.** UMAPs colored by gene activity scores of lineage defining genes (PAX5 – lymphoid; GATA1 – MEP; CD34 – stem; MPO – granulocytic/myeloid; IRF8 – DC; CD14 - myeloid). Gene scores are indicative of the degree of chromatin accessibility within a 100 kb window on either side of the gene body.

In the same way we had previously detected *differentially regulated* genes, we asked whether BCG vaccination led to changes in peak accessibility (*differentially regulated*, or DR, peaks). Across all clusters we identified more than 2000 DR peaks (Figure 1.9a-e), demonstrating that BCG vaccination not only has lasting impacts on gene expression, but also on the epigenetic landscape of HSPCs. While these data were in line with our expectation that changes in gene expression would be coupled to some changes in chromatin accessibility, we had initially expected DR peak counts to mirror DR gene counts (more DR peaks in HSCs and fewer DR peaks in downstream progenitor cell types). In contrast we found an unexpectedly large number of DR peaks within peripheral CMP, GMP, and MEP clusters, and lower counts within HSCs (Figure 1.9a). Clusters CMP3 and GMP2, which are small myeloid clusters, harbored the greatest changes in chromatin accessibility, although many DR peaks were also detected within MEPs, MLPs, and other CMP/GMP clusters.

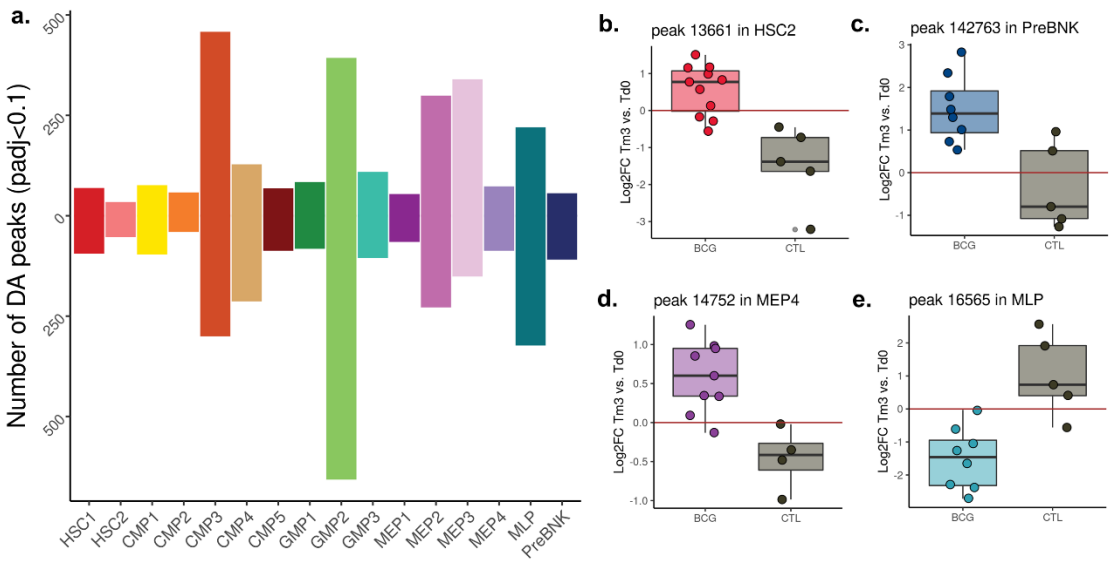

Figure 1.9. BCG vaccination impacts the chromatin accessibility landscape of HSPC progenitors after 90 days

a. The total number of significant peaks (FDR<0.1) for each cluster. Bars extending in the positive and negative direction indicate peaks whose accessibility was impacted positively and negatively, respectively, by BCG vaccination compared to placebo. Bars are color-coded by cluster. **b-e.** Log2FC accessibility (D90 vs. D0) for significant peaks found within different clusters (**b.** HSC_b, **c.** PreBNK, **d.** MEP_c, **e.** MLP_a).

To better understand the nature of BCG-induced differences in chromatin accessibility, we started by performing a transcription factor (TF) motif enrichment analysis at DR peaks. To simplify the analysis, we grouped DR peaks within similar clusters into broader groups (i.e., DR peaks within CMP1-5 were combined into a unique 'CMP' peak set; Figure 1.10a) and then used HINT[68] to find TF motifs enriched within DR peaks of each broad group (Figure 1.10b). Significantly enriched TF motifs (FDR< 0.001) belonged to a large number of zinc finger transcription factors, including 4 different KLF transcription factor classes, as well as EGR1.

ERG, FOS/JUN, and multiple ETS transcription factor groups were also enriched. The most significant enrichments resided largely within myeloid lineage clusters (CMPs and GMPs), and both the identity and degree of enrichment of significant TFs were highly similar when comparing CMPs with GMPs. Likely due to the overall lower numbers of DR peaks, enriched motifs within erythroid and lymphoid clusters were less significant although they involved similar motifs to those found within CMPs and GMPs.

Overall, these peak counts and motif enrichments suggested that on the one hand, BCG vaccination broadly impacts gene expression programs in HSCs, while having limited detectable effects on chromatin accessibility. On the other hand, only a limited set of DR genes were detected in the downstream progenitor clusters - CMP, GMP, MLP, and PreBNK, even though changes in chromatin accessibility were more pronounced. To more formally investigate whether there was any evidence of differential TF activity within downstream progenitors remaining at D90, we performed a regulon analysis using SCENIC[69,70], an approach that groups genes into transcription factor modules, each module containing a transcription factor and it's predicted gene targets. Gene expression data can then be used to infer whether there is differential activity of entire TF modules (regulons) based on correlated patterns of differential expression of many genes within the module. We define *transcription factor (TF) activity* from this point on as evidence for differential expression of TF target genes as indicated by significant regulon scores. Not unexpectedly, BCG vaccination had the largest impact on TF activity within HSCs and MEPs (Figure 1.10 c, d), which we predicted would be the case given that HSCs and MEPs had the largest total number of detected DR genes in our scRNA-sequencing dataset (Figure 1.2c,d). In comparison, we found fewer than 10 regulons with persisting differential activity within the other progenitor clusters (CMPs, GMPs, MLPs, PreBNK). The lack of differential activity was

most predominant within CMPs or GMPs (Figure 1.10c) which, when using a p<0.05 threshold, had only 6 and 2 TFs respectively with differential activity by day 90, supporting the general idea that progenitors, apart from MEPs, have a much more limited repertoire of differential TF activity and gene expression compared to HSCs.

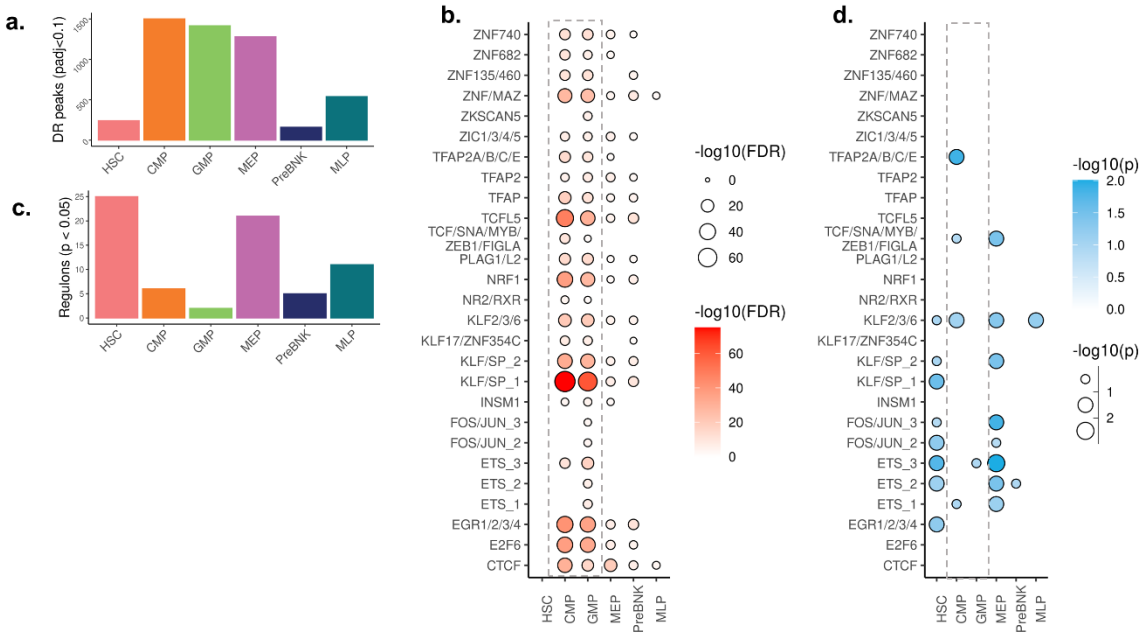

Figure 1.10. Changes in chromatin accessibility and TF activity are uncoupled

a. Number of unique differentially regulated peaks within broad cluster groups HSC (HSC1 and HSC2), CMP (CMP1-5), GMP (GMP1-3), MEP (MEP1-4), PreBNK, and MLPs. **b.** Transcription factor motif enrichments for broad cluster groups. Both circle size and color are scaled to –log10(FDR). TFs shown with a circle have FDR < 0.001 and are present in at least 15% of DR peaks in at least one cell type. Areas with no circle indicate an enrichment with FDR >= 0.001. **c.** Number of transcription factor regulons with differential activity (p < 0.05) in each broad cluster group. **d.** Bubble plot indicating differential regulon activity of transcription factors matching those in b. Circle size and color are scaled to –log10(p). Areas with no circle indicate p >= 0.15

DR peaks across different HSPC clusters are bound by a core set of shared TFs

To further investigate the role of transcription factors in the establishment and/or maintenance of DR peaks we next performed genome-wide transcription factor foot printing. This analysis goes beyond motif enrichments and can be used to determine whether there is evidence of physical transcription factor binding within a peak before and after BCG vaccination. In general, and in line with the motif enrichment analysis in Figure 1.10b, we observed low levels of TF footprint enrichments at DR peaks in HSCs. However, among TFs with enriched binding at DR peaks in HSCs, we found that most were only enriched ($p < 0.05$) either before vaccination or after, but not at both timepoints (Figure 1.11a). The transcription factor groups EGR1 and ETS_2, which had enriched levels of binding at DR peaks only after BCG vaccination, also had increased activity within HSCs at D90 as determined by regulon analysis. Other TF groups had differential activity at D90 (Figure 1.10d) but not differential binding enrichments at D90 compared to D0, suggesting that they continue to bind the same numbers of DR peaks, but still upregulate or downregulate their target genes perhaps through stronger or more frequent binding. In contrast, most of the top TFs with enriched binding within DR peaks of CMPs and GMPs were strongly enriched both before and after BCG vaccination (Figure 1.11b, c), suggestive of roughly equal binding at DR peaks at 90 days post-BCG vaccination compared to before vaccination. We found this to be generally true for all of the progenitor clusters, including MEPs, MLPs, and PreBNKs (Figure 1.11d-f), supporting the idea that while HSCs continue to have altered transcription factor activities and/or binding patterns 90 days following BCG vaccination, transcription factors within progenitor clusters return closer to pre-BCG activity levels and binding patterns at this time point.

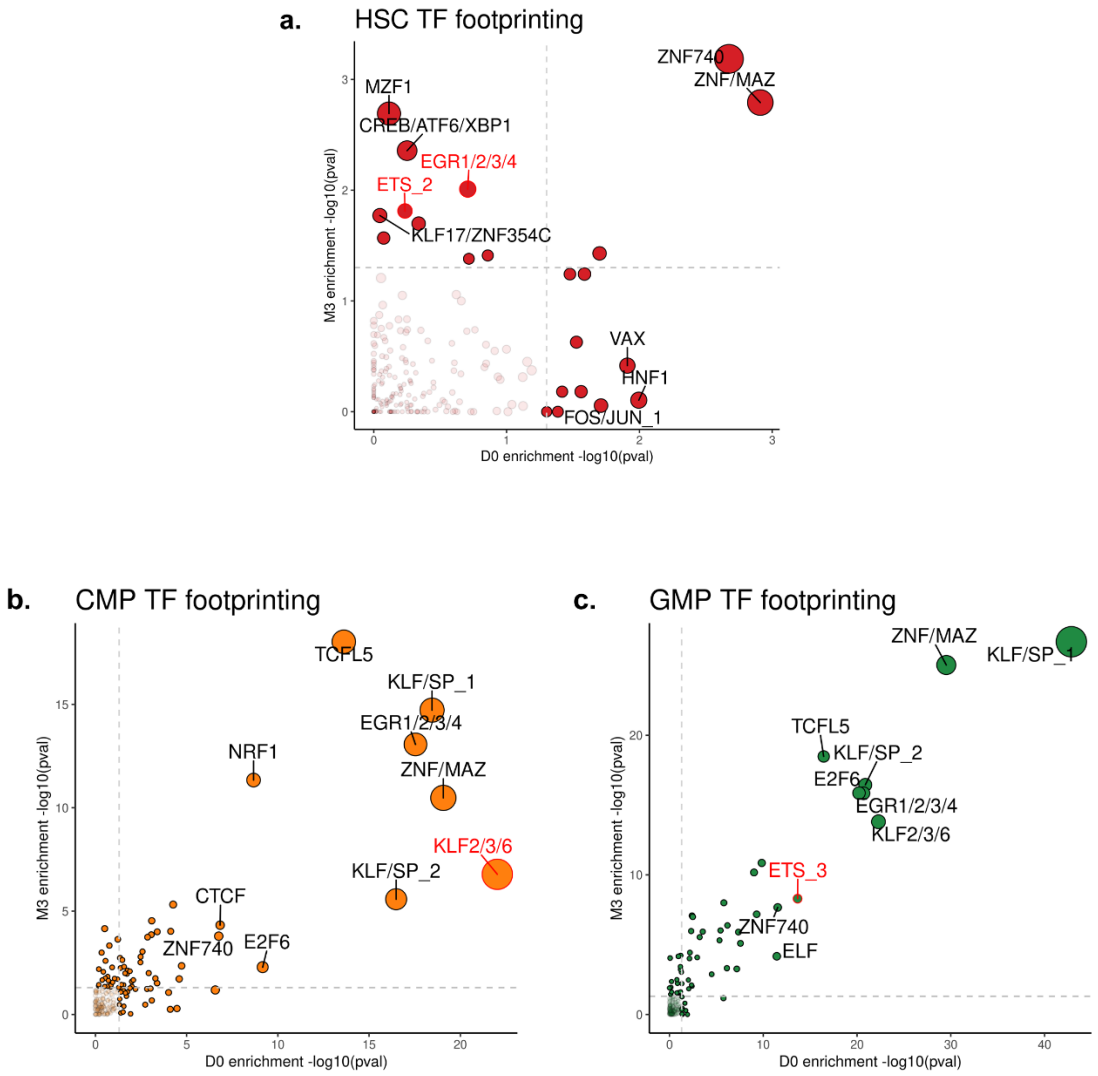

Figure 1.11. Transcription factor footprinting

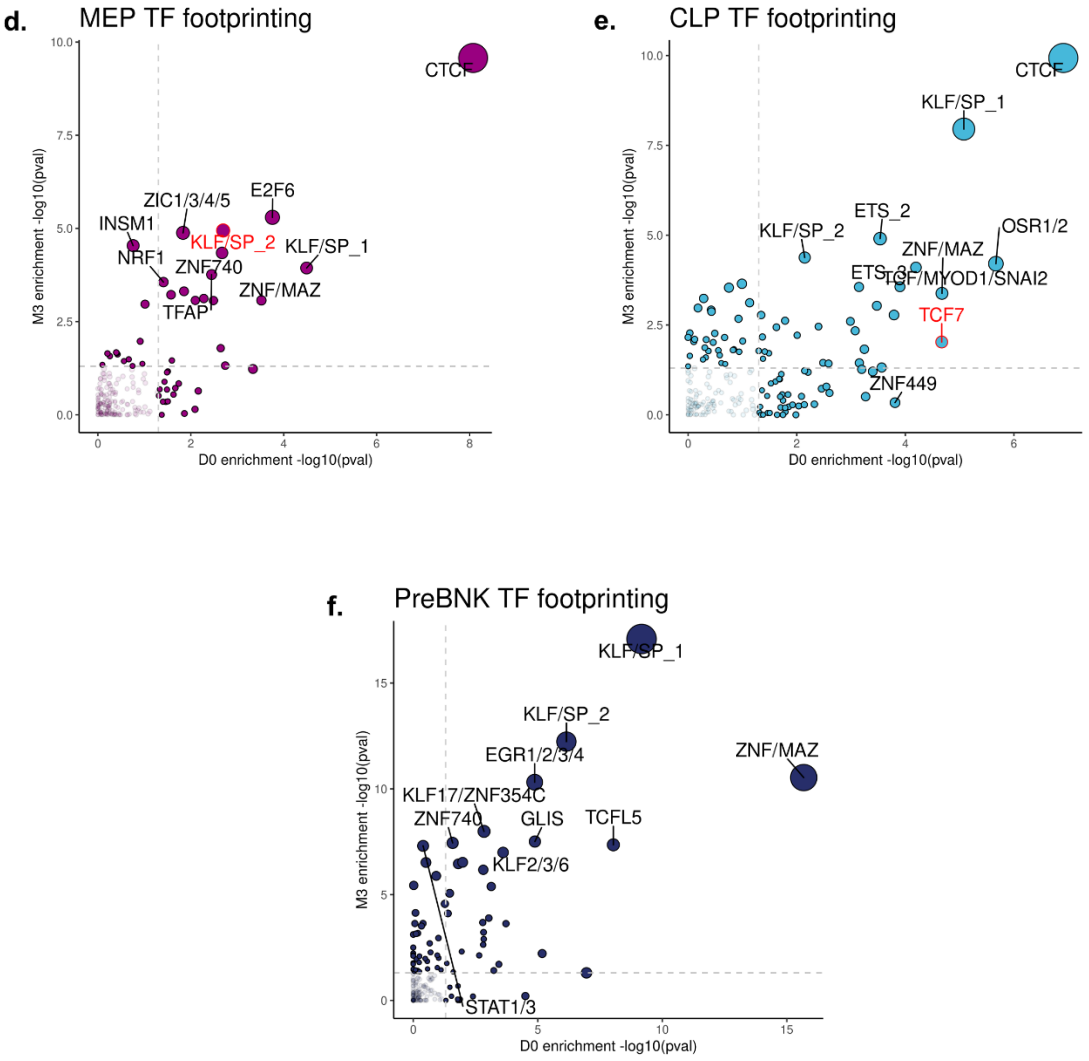

Figure 1.11, continued

a-f. Chi-squared transcription factor footprinting enrichments for each cluster. Footprints overlapping DR peaks were used as foreground and all footprints overlapping any peak was used as background. The x-axis shows the D0 (prior to BCG) enrichment and the y-axis shows the D90 (post vaccination) enrichments. The top 10 most enriched TFs (either at D0 or D90) for each cluster are labelled. Point size is scaled to –log10(p-value). Dotted vertical and horizontal lines indicate p = 0.05 on the x and y axes respectively. Highlighted transcription factors (red label and outline) are transcription factors with differential activity at D90 within the cluster.

From the foot printing enrichments performed for each progenitor cluster shown in Figure 1.11, we noticed that every cluster contained a standout set of prominently enriched TFs and that a high degree of overlap existed when comparing these top-enriched transcription factors between clusters. For example, the transcription factor group 'KLF/SP_1' was among the most enriched transcription factors within all five progenitor clusters and 'EGR1/2/3/4' was highly enriched within CMPs, GMPs, and PreBNKs, suggesting that DR peaks within progenitor clusters may have been established at the binding sites of a common repertoire of transcription factors. To more systematically investigate the degree to which the most enriched TFs were shared across clusters we made a list of the top 10 most enriched TFs at DR peaks of each cluster (Table 1.2) and counted the occurrence of each TF among this top 10 table. We found nine core transcription factors appearing among the top 10 within at least three clusters. The TF group ZNF/MAZ was ranked in the top 10 for all 6 clusters and ZNF740/VEZF1, KLF/SP, and KLF/SP were in the top 10 for 5 clusters. In total, 75% of all TFs listed in Table 1.2 were shared with at least one other cluster and 68% were shared with at least two other clusters, demonstrating that DR peaks across HSPC clusters are constitutive binding sites of a common set of transcription factors.

TF rank	HSC	CMP	GMP	MEP	MLP	PreBNK
1	ZNF740	KLF2/3/6	KLF/SP_1	CTCF	CTCF	KLF/SP_1
2	ZNF/MAZ	ZNF/MAZ	ZNF/MAZ	E2F6	KLF/SP_1	ZNF/MAZ
3	MZF1	KLF/SP_1	KLF2/3/6	KLF/SP_2	OSR1/2	KLF/SP_2
4	CREB/ATF6/XBP1	TCFL5	KLF/SP_2	ZIC1/3/4/5	ETS_2	EGR1/2/3/4
5	EGR1/2/3/4	EGR1/2/3/4	EGR1/2/3/4	INSM1	ZNF/MAZ	TCFL5
6	HNF1	KLF/SP_2	E2F6	KLF/SP_1	TCF7	KLF17/ZNF354C
7	VAX	NRF1	TCFL5	ZNF740	KLF/SP_2	GLIS
8	ETS_2	E2F6	ETS_3	TFAP	TCF/MYOD1/SNAI2	ZNF740
9	KLF17/ZNF354C	CTCF	ZNF740	NRF1	ETS_3	STAT1/3
10	FOS/JUN_1	ZNF740	ELF	ZNF/MAZ	ZNF449	KLF2/3/6

clusters in the top 10
6
5
4
3
2
1

Table 1.2. Top 10 transcription factor classes with enriched binding in each HSPC cell type

Table showing the top 10 most significantly enriched transcription factor classes (ranked by p-value) within each HSPC cluster. Transcription factors are colored according to the number of times they appear in the table, which corresponds to the numbers of clusters within which they are highly enriched.

Among the top nine shared transcription factors with highly enriched binding in at least three clusters, we noticed that several had continued differential activity within HSCs (Table 1.3). Four of the transcription factors among the top seven (~57%) had differential regulon activity ($p<0.1$) within HSCs, representing a significant enrichment (OR=8.05, p=0.012). We note that not all four of these differentially active TFs had clear differential binding at D0 compared to D90 at DR peaks (Figure 1.11a) within HSCs, suggesting that BCG vaccination does not always

change the identity or types of sites bound by these TFs, but rather the strength at which they up- or down-regulate their targets. Nonetheless, the data collectively suggest that establishment of differential chromatin accessibility within progenitor clusters likely utilized, in part, a common set of transcription factors, and that a significant proportion of these common transcription factors, although exhibiting no differential activity in progenitor clusters themselves, have continued differential activities within upstream HSCs. Our data support a model whereby this continued activation within HSCs actively shapes the chromatin accessibility landscape of cells downstream (Figure 1.12).

TF	Occurrence	HSC diff activity pval
ZNF/MAZ	6	0.23
KLF/SP_1	5	0.03
KLF/SP_2	5	0.1
ZNF740	5	No activity
EGR1/2/3/4	4	0.06
CTCF	3	0.23
KLF2/3/6	3	0.1
TCFL5	3	No activity
E2F6	3	0.28

Table 1.3. Top shared transcription factor classes

Table showing the top transcription factor classes shared across 3 or more clusters. The 'Occurrence' column indicates the number of clusters for which the TF class falls with the top 10 most enriched. Among TFs with the same total number of occurrences, order was determined by the sum of ranks across clusters. The 'HSC diff activity pval' corresponds to the p-value for differential regulon activity within HSCs as determined in Figure 10d.

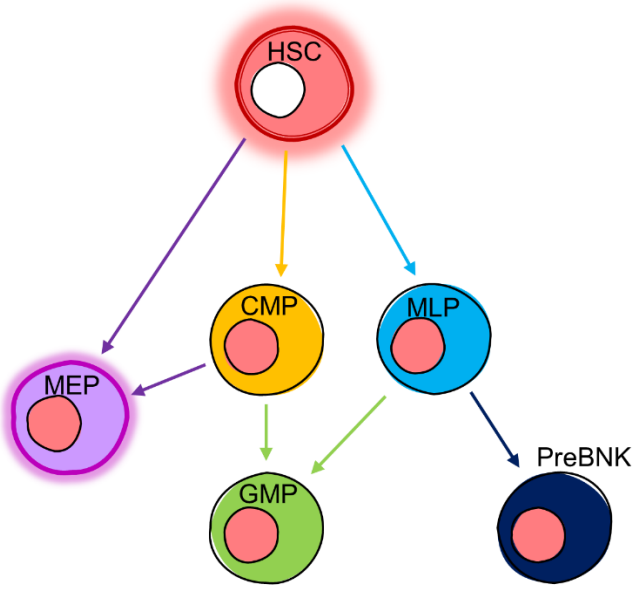

Figure 1.12. A preliminary model. BCG-induced differences in chromatin accessibility within progenitor clusters (red nucleus) occur at binding sites of TFs harboring continued differential activity within upstream HSCs (red cytoplasm). HSCs and MEPs exhibit the largest continued differences in gene expression and transcription factor activity while other clusters exhibit fewer signs of continued activation. Due to the hierarchical nature of hematopoietic differentiation (whereby HSCs differentiate into progenitor cell types), we hypothesize that activation of transcription factor circuits within upstream HSCs, directly impacts chromatin accessibility of downstream progenitors.

DR peaks are not directly transmitted across differentiation

Initially, we hypothesized that differentially active TFs within HSCs may help establish memory-like DR peaks within immediate downstream progenitors which could be maintained across further differentiation. To test this model more directly we looked for patterns of DR peak sharing across clusters (Figure 1.13). For example, if differential transcription factor activity in HSCs were to induce "memory" peaks transmitted to CMPs and then to GMPs, one would expect to find a high level of overlap between DR peaks within HSCs and CMPs and between CMPs and GMPs. However, when we looked at patterns of sharing for individual peaks in HSCs, CMPs, and GMPs, we found only 5 (0.2%) shared peaks between HSCs and CMPs, and only 16 (0.5%) DR peaks shared between CMPs and GMPs, indicating a lack of long-term propagation of the same DR peaks across differentiation. Similarly, only 0.06% of DR peaks within the HSC-MEP trajectory, and no peaks in the HSC-MLP-PreBNK trajectory were shared – collectively making it unlikely that a constant state of activation within HSCs leads to sites of differential accessibility which are precisely retained across hematopoietic differentiation.

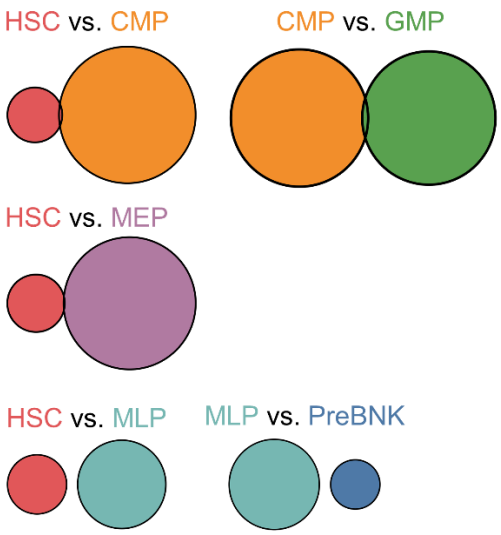

Figure 1.13. DR peaks do not overlap across clusters

Venn diagrams are proportional to the total number of DR peaks (FDR<0.1) within each group and show the overlap of peaks within each expected trajectory (top row: HSC-CMP and CMP-GMP; middle: HSC-MEP; bottom: HSC-MLP and MLP-PreBNK)

Thus, DR peaks within progenitor clusters enrich for a shared repertoire of TFs which remain activated within upstream HSCs, but for which we do not detect differential activity in downstream progenitors. Nonetheless, the locations of these DR peaks shift across differentiation, suggesting that there must be some active mechanism in place to modulate these across-differentiation changes. Taking all of these factors into account we hypothesized that activated TFs in HSCs (either due to increased protein levels of these TFs, or post-translational modifications such as phosphorylation) may remain upregulated or phosphorylated across differentiation, yet shift from binding a wide repertoire of open sites within HSCs and

upregulating their own gene targets, to binding preferentially to a more restricted subset of sites within downstream progenitors that are co-bound by cluster specific TFs, thus regulating a much more limited repertoire of cluster-specific gene targets. In other words, TFs maintain the potential for activated binding across differentiation, but bind to altered sites within downstream progenitor clusters in a manner that is strongly influenced by the cluster specific trans environment. This is supported by that fact that in addition to enriching for shared, HSC-active TFs such as KLF and EGR1, almost every cluster also enriched strongly for at least one cluster-specific TF. Examples include ZIC1/3/4/5 and INSM1 within MEPs, OSR1/2 and TCF7 within MLPs, and GLIS and STAT1/3 in PreBNKs (Table 1.2). Although CMPs and GMPs were bound by an extremely similar repertoire of TFs, there were also differences in the co-enriched factors within these two clusters. For example, ETS_3 and ELF transcription factors were much more enriched within GMPs. Moreover, although not within the top 10, FOS and JUN transcription factors were also enriched within GMPs but not at all within CMPs, further exemplifying differences in the trans environment even between closely related clusters of the same lineage.

To further investigate this hypothesis, we performed a focused analysis of the HSC – CMP – GMP differentiation trajectory (Figure 1.14). As shown previously, HSCs contain many transcription factors with sustained differential activity 90 days following BCG vaccination (Figure 1.10c,d; Figure 1.14a) – KLF transcription factors KLF/SP_1, KLF/SP_2, KLF2/3/6, and the EGR1/2/3/4 transcription factors being of particular interest due to the fact that these are also strongly enriched within DR peaks in downstream progenitor clusters of multiple lineages (Table 1.3) and are therefore particularly implicated as TFs that shape the downstream epigenetic landscape. Within CMPs, the immediate cluster downstream of HSCs, we no longer detected significant differential activity of KLF/SP_1, KLF/SP_2, or EGR1/2/3/4, however we continued

to detect some differential activity (i.e. increased expression of TF targets) of the KLF2/3/6 transcription factor group (Figure 1.11b, Figure 1.14b). In line with this observation, the KLF2/3/6 TF group was the most enriched TF group at DR peaks within CMPs (Table 1.2, Figure 1.14c), indicating that while most activated TFs in HSCs lose activity during the process of CMP differentiation, KLF2/3/6 continued to activate its downstream targets and plays a predominant role in shaping the DR peak landscape in CMPs. DR peaks in CMPs co-enriched for the HSC-active TFs KLF/SP_1, KLF/SP_2, and EGR1/2/3/4 (Figure 1.14c), which no longer actively induced their own targets (as indicated by a lack of differential activity) but likely shifted to a supportive role in driving the differential expression of KLF2/3/6 targets. Many HSC-specific TFs were not enriched at DR peaks within CMPs including MZF1, CREB/ATF6/XBP1, HNF1, and VAX. Instead, DR peaks in CMPs also co-enriched for a new set of TFs including TCFL5 and NRF1, exemplifying the idea that the trans environment is dynamic across differentiation (Figure 1.14c). In GMPs, immediately downstream of CMPs, we no longer detected differential activity of KLF2/3/6 (Figure 1.14d). Instead, KLF2/3/6 co-bound together with KLF/SP_1, KLF/SP_2, and EGR1/2/3/4, to a new set of DR peaks which co-enriched for ETS transcription factors (Figure 1.14e). The loss of KLF2/3/6 activity upon CMP to GMP differentiation suggests that KLF2/3/6 shifts from promoting its own target genes at CMP sites, to acting as a co-binding factor to promote increased ETS target expression at GMP DR peaks. This hypothesis was supported by the fact that DR peaks in GMPs uniquely co-enriched for ETS_3 binding and that ETS_3 exhibited differential activity within GMPs but not CMPs (Figure 1.14d, e).

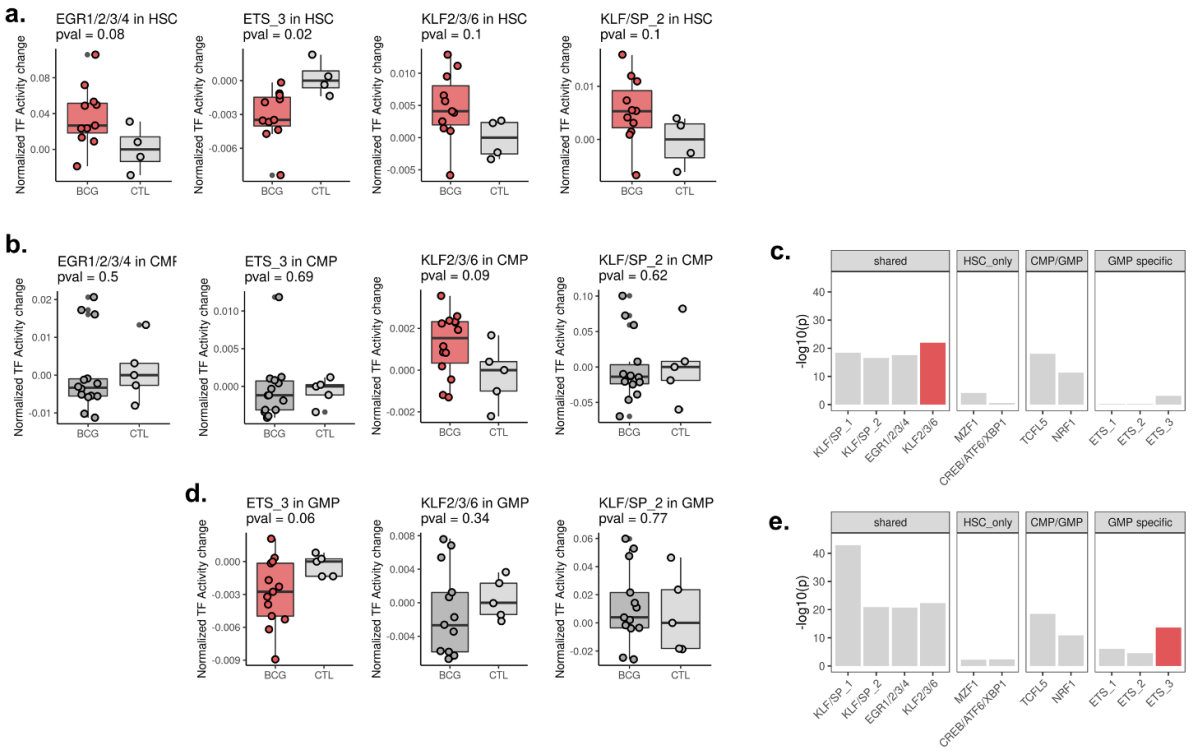

Figure 1.14. HSC-active TFs traverse unique trans-environments to induced cluster-specific DR peaks.

a. Transcription factors with differential activity within HSCs. **b.** Activity levels for the same transcription factor groups in **a** for CMPs. Red boxplots are for TFs with $p < 0.1$ differential activity. **c.** Barplots showing the enrichments of TFs in CMPs grouped by type: 1) shared and HSC-active, 2) co-enriched at DR peaks in HSCs only, 3) co-enriched in DR peaks in CMP/GMP clusters only, 4) and co-enriched only at DR peaks in GMPs. **d.** Activity levels of TFs in GMPs. EGR1/2/3/4 is not shown because this TF group had no detectable activity within GMPs. **e.** As in **c** for GMPs.

The overall model proposed for HSC to CMP and CMP to GMP differentiation is notably also consistent with additional transcription factor footprinting data in other progenitor clusters. For example, DR peaks within MLPs uniquely and strongly enrich for TCF7 binding (Table 1.2), and regulon analysis shows that TCF7 has significant differential activity only within MLPs (Figure 1.11c) suggesting that TCF7 cooperates with KLF and EGR1 TFs to module expression of its own targets in a similar manner to how ETS_3 cooperates with KLF and EGR1 TFs within GMPs to activate ETS_3 targets. Collectively these data support the hypothesis that HSC-active TFs traverse a unique trans-environment during differentiation, and shift from activating their own targets to co-binding together with cluster-specific TFs, selectively shaping the epigenetic landscape in a way that is tailored to each cell type (Figure 1.15).

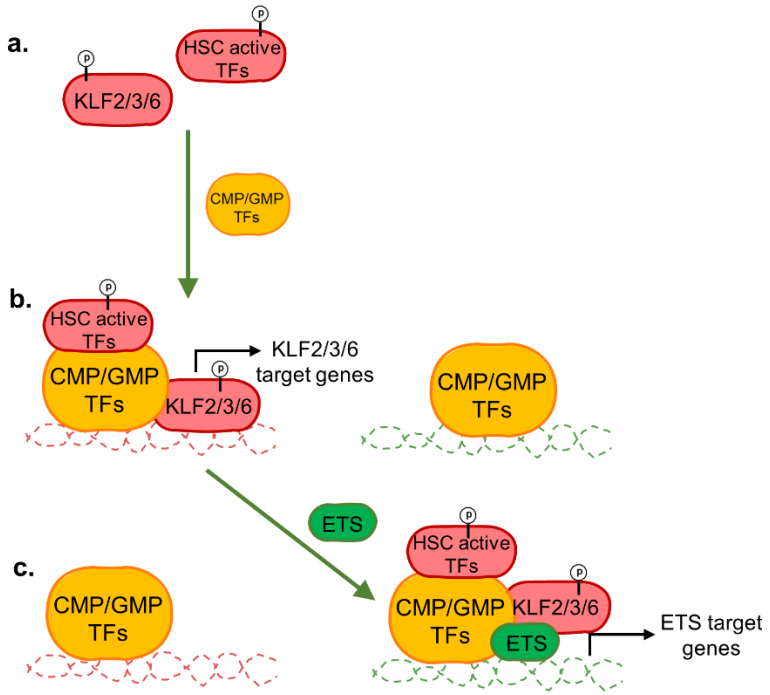

Figure 1.15. Proposed final model for how a shared set of BCG-responsive TFs modulate the downstream epigenetic landscape of progenitor clusters. **a.** Many TFs are activated in HSCs. **b.** Active TFs within HSCs co-bind with CMP/GMP specific transcription factors to target CMP specific peaks that regulate KLF2/3/6 targets. **c.** During differentiation from a CMP to a GMP, HSC-active TFs shift towards peaks co-bound by ETS_3 to differentially regulate ETS_3 targets. Evidence for a similar paradigm within MLPs is evidenced by the MLP-specific differential activity of TCF7 and co-enrichment of TCF7 at DR peaks in MLPs.

BCG-induced differential chromatin accessibility within GMPs predicts increased IL1B secretion by PBMCs

Given that BCG vaccination induced changes in chromatin accessibility in the bone marrow at day 90, we asked whether changes within the most differentiated clusters (for example, GMPs or

PreBNKs) could have functional implications for the mature immune cells that these progenitors give rise to. First, to get a general sense of the genes associated most closely to DR peaks within progenitor clusters, we performed an over-representation analysis by assigning each DR peak to its closest gene and comparing the resulting DR peak-associated genes to a background set of all peak associated genes. Overall, we found that the strongest immune-related enrichments were present in GMPs, whose DR peaks enriched for biological process pathways such as 'myeloid leukocyte activation' (Figure 1.16a). GMPs also notably had several significant enrichments in biological process pathways related to MAPK signaling and myeloid development (Figure 1.16a) and multiple reactome pathways related to TLR signaling, signaling by interleukins, and GM-CSF signaling (Figure 1.16b). Upstream CMPs had fewer total significant pathway enrichments but also enriched for MAPK and myeloid development pathways, while very few significant pathways were found in MEPs, PreBNKs, and MLPs.

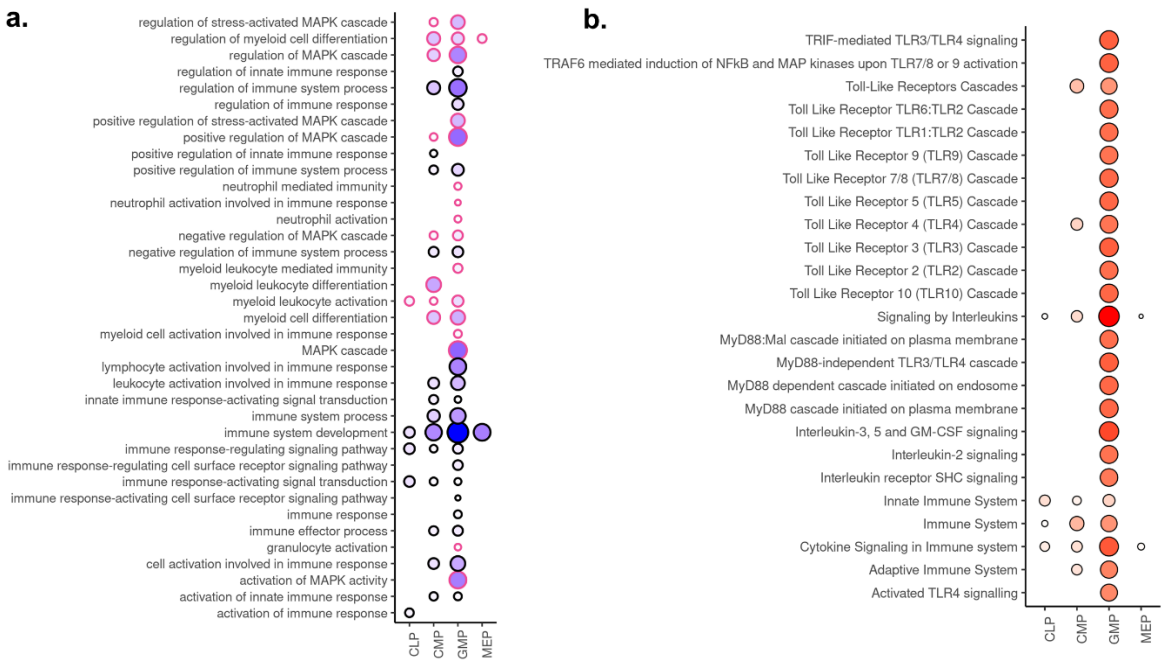

Figure 1.16. Gene ontology analysis of DR peak-associated genes

DR peaks within each cluster were assigned to the gene with the closest TSS. Gene ontology enrichment analysis was performed for biological process pathways (**a**) and reactome pathways (**b**) using all peak associated genes as background and genes associated with DR peaks as foreground. Plot circle size and shading darkness are both scaled to –log10(p-value) of enrichment. In **a,** pathways related to immunity, immune development, and MAPK signaling are outlined in pink.

These results suggested that DR peaks within GMPs in particular, may regulate immune pathways, and thus could impact immune functionality within myeloid cells entering the peripheral circulation. As a measure of peripheral blood cell immune functionality, we used

cytokine secretion data from donor-matched PBMCs[43] that had been collected in the original study from which our BM samples were derived (Figure 1.17a). This data set contained secreted cytokine concentrations for each donor in response to a 24-hour stimulation with heat killed *C. albicans* and demonstrated that BCG vaccination increases the ability to PBMC to secrete IL1B and IL6 in response to *C. albicans*.

Since GMPs were the most enriched for immune related pathways, we investigated whether BCG-induced changes in chromatin accessibility of GMPs was directly predictive of cytokine secretion by donor paired PBMCs. We leveraged the fact that even within BCG vaccinated individuals there was clear heterogeneity wherein some individuals had larger changes in peak accessibility compared to other BCG vaccinated individuals. In the PBMC cytokine data set, we also observed a high level of heterogeneity in cytokine secretion levels between individuals. Thus, if differential accessibility at DR peaks were truly related to functional alterations in PBMCs, one would expect to find individuals with the greatest magnitude of epigenetic rewiring to be the same individuals with the greatest increases in cytokine secretion. To formally determine whether this was true, we used elastic net regression to determine whether levels of differential accessibility at DR peaks (raw Tm3 vs. Td0 log2FC values) had power to predict cytokine responses (FC Tm3 vs. Td0) across individual donors. This analysis strikingly showed that fold change cytokine secretion of IL1B could be predicted by log2FC DR peak accessibility to a high level of accuracy (Figure 1.17b; $R = 0.72$, $p = 0.0037$). Notably this was only the case for IL1B, but not IL6, the other cytokine found to be significantly primed in response to BCG vaccination, demonstrating that altered chromatin accessibility could selectively predict one facet of the altered immune response by PBMCs. These data establish a link between BCG-induced

rewiring of chromatin accessibility in GMPs, and specific rewiring of *C.albicans*-induced IL1B responses by matched PBMCs.

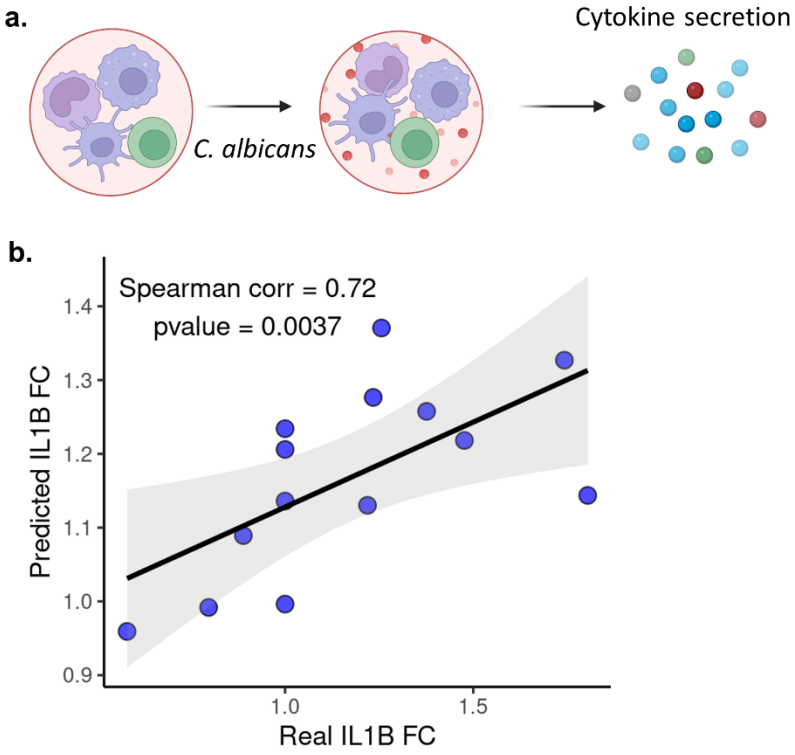

Figure 1.17. Chromatin accessibility changes in GMPs predict changes in IL1B secretion of PBMCs in response to C. albicans challenge

a. Schematic outlining PBMC experiment. PBMCs were collected from each donor at the same time collection of bone marrow samples (Td0 and Tm3). PBMCs were stimulated with heat-killed C. albicans for 24 hours at both time points and cytokine secretion of IL1B, IL6, TNF, IFNG, IFNA, IL1RA, and IL10 were measured using ELISA. **b.** Results of the elastic net regression. The scatterplot shows real IL1B FC (D90 vs. D0) for each donor on the x-axis and predicted values from the model on the y-axis

Finally, given that DR peaks within GMPs were found to enrich for binding of TFs which had persistent differential activity within HSCs, we hypothesized that the same individuals harboring the largest changes in GMP chromatin accessibility and increased cytokine secretion, may also be the same individuals harboring the greatest BCG-induced changes in transcription factor activity and differential gene expression in HSCs, which would further support the hypothesis that persistent gene expression and TF activity in HSCs is directly linked to differential chromatin accessibility in downstream progenitors, which is linked to IL1B secretion capacity of mature immune cells in the periphery. First, using the same approach as for DR peaks within GMPs, we used elastic net regression to determine whether levels of differential expression of DR genes within HSCs had power to predict IL1B responses (FC Tm3 vs. Td0) across individual donors (Figure 1.18a). Here we found differential gene expression within HSCs to have strong and significant predictive power ($R = 0.815$, $p=2x10^{-4}$), establishing that changes in IL1B secretion capacity are also tightly linked to day-90 differential gene expression within HSCs. Log2FC responses of hundreds of individuals genes within HSCs correlated significantly (padj < 0.1) with IL1B responses, further supporting the elastic net regression findings (Figure 1.18 b-d). Importantly, we noticed that the log2FC values of several transcription factors such as FOSB (Figure 1.18 b) and KLF6 (Figure 1.18 c; in the KLF2/3/6 family) were among significantly correlated genes with $R > 0.8$. Thus, we directly tested whether TF activity scores in HSCs were correlated with IL1B production. We found that activity scores of TFs in HSCs had remarkably strong correlations with IL1B production but not IL6 (Figure 1.18 e-g), thus formally demonstrating that BCG-induced differential TF activity and gene expression in HSCs, progenitor chromatin accessibility, and peripheral immune cell cytokine secretion are linked processes.

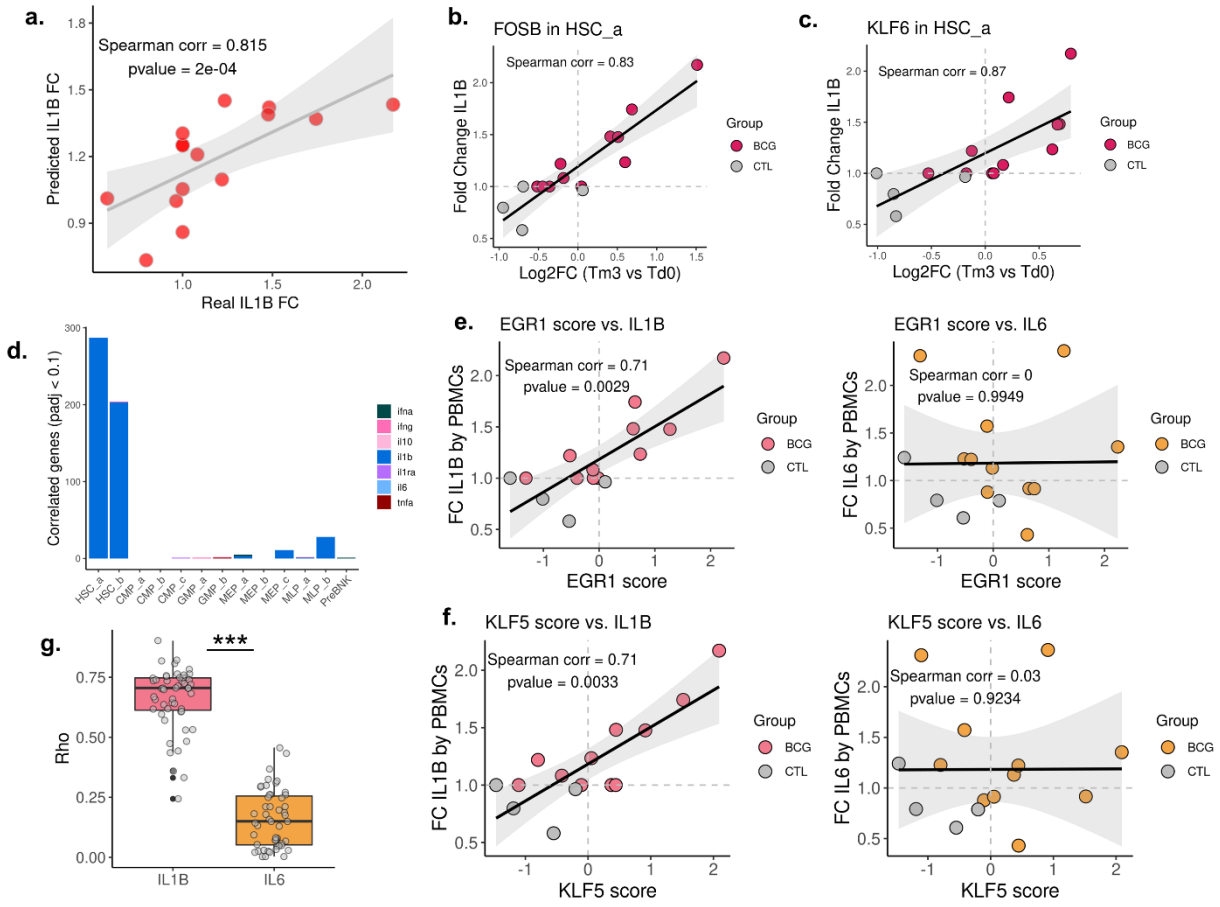

Figure 1.18. Gene expression and TF activity in HSCs correlates strongly with IL1B production by PBMCs

a. Results of the elastic net regression using gene expression data for HSCs. The scatterplot shows real IL1B FC (D90 vs. D0) for each donor on the x-axis and predicted values from the model on the y-axis. **b-c.** Spearman correlations between log2FC expression levels of the transcription factors FOSB and KLF6 and fold change IL1B secretion between D0 (before) and D90 (3 months-post BCG vaccination). **d.** quantification of the total number of DR genes in HSCs with significant spearman correlations with fold change IL1B secretion for each cytokine tested. **e-f.** Example scatter plots correlating TF activity (regulon) scores in HSCs from the regulon analysis described in Figure 10d, with fold change IL1B (left) and fold change IL6 (right) in PBMCs. **g.** Violin plot comparing spearman rho values for IL1B and IL6 ($p < 2.2e-16$)

DISCUSSION

Overview

Since clinical evidence and mouse models have suggested that the BCG vaccine may impact the immune system at the stem-cell level[9,45,47–49,67,71,72], we used single cell RNA and ATAC sequencing on HSPCs isolated from human bone marrow aspirates to investigate how BCG vaccination effects gene expression and chromatin accessibility after 3 months. Our data indicated that BCG vaccination impacted the bone marrow through multiple modalities. We detected differential expression of nearly 200 genes within the most primitive and stem-like HSCs, even 90 days following a single intradermal vaccination. HSCs, together with MEPs, harbored changes in the expression of genes related to immunity, metabolism, and apoptosis/proliferation while fewer changes in gene expression were detected in other progenitor clusters. By taking a focused look at the consequences of differential expression within HSCs, we used a computational approach to predict that the differential gene expression program biased them towards the granulocytic fate – a finding that was supported by flow cytometric data demonstrating that the bone marrow of BCG vaccinated individuals contained significantly higher percentages of CMPs relative to that of individuals in the placebo group. When integrated with single cell ATAC sequencing data collected to probe the epigenetic landscape of the same samples, we found evidence that the granulocytic-oriented differential gene expression program in HSCs was driven by the increased activity of several transcription factors, which were the same transcription factors enriched within differentially accessible peaks of downstream progenitors. These epigenetic changes were not directly propagated through differentiation - rather our data supported a model whereby activated transcription factors within HSCs traversed

a continuously changing trans environment, binding to unique sets of peaks within each progenitor cluster during the differentiation process. Changes in gene expression and transcription factor activity in HSCs, and differential accessibility of downstream progenitors, all correlated directly with IL1B secretion by donor paired PBMCs in response to a C. albicans challenge, supporting the hypothesis that long-lasting activation within HSC at the "top" of the differentiation hierarchy directly influenced the epigenetic landscape of downstream progenitors, which entered the circulation as functionally reprogrammed cells.

Lasting effects of BCG vaccination on gene expression are heterogeneous and centered on HSCs and MEPs

BCG induced lasting changes in the expression of immune, metabolism, and apoptosis/proliferation genes heterogeneously across cell types. HSCs and MEPs, compared to other cell types, had the largest number of differentially regulated genes at day 90. Pathways enriched among DR genes in HSCs were characteristic of a stress-like response and included apoptosis, ROS, and hypoxia pathways while DR genes in MEPs enriched for pro-proliferative and glycolytic pathways. Since HSCs can differentiate into MEPs, we considered the possibility that differential expression in MEPs could be retained from the differential expression occurring in upstream HSCs. However, the clear differences between activated pathways in HSCs and MEPs were more supportive of a scenario whereby lingering signals at D90 acted directly on HSCs and MEPs to induce different activation programs. Upon sub-clustering HSCs into different sub-groups, we had found that MEP-biased HSCs harbored fewer differences in gene expression compared to myeloid or lymphoid biased sub-groups, further supporting this idea. On the other hand, our data were more supportive of a model whereby myeloid and lymphoid

progenitors may have derived some of their limited differential expression programs from activation that was originally induced within upstream HSCs. Both our GSEA results, and HSC sub-clustering analyses supported this idea. Most pathways enriched in these progenitors were similar or shared with pathways enriched within HSCs, and sub-clusters of HSCs biased towards the myeloid or lymphoid lineages harbored hundreds of DR genes. As discussed later in this section, this model is also in line with our general findings that active TFs, which drove the differential expression of many gene networks in HSCs, switched to a supportive role by co-binding together with cluster-specific TFs at specific sites in downstream progenitors to maintain a limited number of differential gene expression programs, while shutting down most of the gene networks which were differentially expressed in HSCs.

In either case, these observations shed new light on the specific cell types that are most likely to be persistently affected by an immune stimulus and open up new avenues of investigation into better understanding, for example, whether there are particular shared receptors on HSCs and MEPs that allow them, but not other clusters, to respond to BCG-induced signals present in the bone marrow at day 90. HSCs are known to express both Toll-like receptors (TLRs) and pathogen recognition receptors (PRRs) through which they can directly sense pathogens and pro-inflammatory cytokines[73,74]. In comparison, much less is known about the repertoire of such receptors on MEPs. One could also speculate an alternative model in which an initial state of large-scale activation and differential gene expression was initially induced in all cell types directly, but in which only HSCs and MEPs continued to maintain a high level of activation over an extended period of time while others developed a more refractory state, shutting down many differential gene expression programs. We found that CMP and GMP progenitor clusters harbored many epigenetic changes, not only in the positive direction, but also in the negative

direction – implying that at several sites, chromatin closes following BCG vaccination. Closing of chromatin could prevent continued differential expression of select genes much in the same way that strong LPS exposure can lead to a refractory state to secondary LPS stimulations[25,75]. Although our work demonstrated that PBMCs in the periphery secreted higher levels of proinflammatory cytokines in response to secondary stimuli, it remains possible for progenitors in the bone marrow to selectively close and shut down certain gene expression programs while increasing chromatin accessibility to promote other expression programs, similarly to how antimicrobial and inflammatory pathways are primed and tolerized respectively within the same macrophages stimulated with LPS[25]. From an evolutionary perspective, this could be beneficial, preventing the release of overly activated myeloid cells into the periphery where they could induce tissue damage[76].

Another important question relates to the source of continued activation, for which two general models are believed to be possible[52,73]. In the indirect model, pro-inflammatory cytokines released by other cell types such as stromal cells or bone marrow resident immune cells could activate differential gene expression programs within HSCs and MEPs. Alternatively, pathogens could be directly detected through PRRs which are, at least on HSCs, known to be present and functional. We hypothesize that continued activation within HSCs and MEPs was due to indirect activation by proinflammatory cytokines due to the fact that cultures of these bone marrow samples were negative for BCG, directly suggesting a lack of direct PRR ligation[50]. Moreover, even if bacterial remnants remained in the bone marrow to stimulate TLRs and PRRs, this would likely be accompanied by some increase in the expression of pro-inflammatory cytokines such as IL6[60,61] which one would have expected to detect on the gene expression level, but which was not the case in our data. These data suggest that HSCs and MEPs were more likely responding to

proinflammatory cytokines produced by other cell types, than secreting these cytokines themselves in response to direct pathogen sensing. Interactions between HSPCs and other mature immune cells and stromal cells could be an important area of further investigation in this regard. Previous work has already demonstrated that different cell types within the hematopoietic niche can interact with and regulate HSCs. For example, it has been demonstrated that TLR signaling can suppress the differentiation of osteoclasts, which is in turn believed to promote HSC differentiation[77,78]. Other studies have reported roles for bone marrow resident macrophages in controlling the properties and functions of HSCs[79]. Thus, better understanding the niche within which our HSPCs resided could provide substantial new insight into continued sources of activation.

HSCs have granulocytic bias

Through a more focused analysis we found that HSCs were biased towards the myeloid, and specifically granulocytic fate at D90, thus exhibiting a low-level emergency granulopoiesis-like state. The finding that BCG vaccination could induce emergency granulopoiesis was not surprising on its own given that neonatal mice vaccinated with BCG-Denmark had increased numbers of neutrophils in the spleen 4-5 days following vaccination, and increased numbers of neutrophils were found in the peripheral blood of human newborns given the same vaccine[56]. However, our findings were novel because they suggested that this state could persist for at least 3 months. HSCs expressed many signatures of interleukin-1 signaling at D90, including increased transcription factor activity of CEBPB, IRF1, and EGR1 all involved in activated IL1β signaling[80]. IL-1 has been shown to play an important role in driving emergency granulopoiesis in mouse models of inflammation and humans[80–82]. A recent study demonstrated that the

induction of inflammation through ligature-induced periodontitis induced a strong upregulation of IL1β and G-CSF within the bone marrow extracellular fluid which was accompanied by multiple signatures of emergency granulopoiesis including increased MPP3 cells, GMPs, and peripheral granulocytes[80]. Moreover, the specific deletion of the IL-1 receptor within HSPCs abrogated the induction of emergency granulopoiesis, directly showing the necessity of IL1 signaling in mice. In this study, HSCs harbored increased accessibly at binding sites of EGR1, IRF1, and CEBPB (the same TFs activated in our data) following the resolution of inflammation, and HSCs within the bone marrow were primed to induce an even stronger granulopoiesis response to a second inflammatory event, suggesting that an initial bout of emergency granulopoiesis could increase the likelihood of a stronger second occurrence. Unlike in the situation in mice, BCG vaccination within humans appears to induce a longer period of continued activation, as opposed to transient activation followed by the induction of epigenetic memory within HSCs. However, it is possible that the emergency granulopoiesis response induced in response to BCG vaccination in humans, once resolved, could eventually prime stronger granulopoiesis responses to subsequent infections. Although the total number of differentially accessible sites within HSCs was not large at D90, EGR1 was among the transcription factors most enriched at the differentially accessible sites we did detect in HSCs, supporting the idea that IL1β-induced emergency granulopoiesis in humans can also induce similar epigenetic changes in HSCs compared to that observed in mice, which could potentially have similar priming effects.

Beyond the detection of emergency granulopoiesis, we found evidence that BCG vaccination could differentially impact different subclusters of HSCs. We determined that although the composition of lineage-biased HSCs changes and is marked by an increase in granulocyte-biased

HSCs, differences in gene expression in the broader HSC clusters are not only reflective of these changes in composition, but also reflective of true changes in gene expression occurring within each lineage-biased group of cells. In other words, even among all cells biased towards the granulocyte fate, we detected BCG-induced differences in expression. Differences in expression within HSCs did not appear to be completely homogenous. The MPP-like HSC cluster containing a majority of CMP_b (granulocyte) biased cells harbored BCG-induced differences in expression that differed from those found within the other MPP-like cluster containing a majority of MLP-biased cells, implying that any changes that HSCs propagate through differentiation may differ between different lineages.

Chromatin accessibility changes in progenitors are linked to continued differential transcription factor activity in HSCs

Single-cell ATAC sequencing performed in parallel with scRNA-seq revealed that BCG vaccination induces differential accessibility at thousands of total sites and impacts all HSPC clusters. Surprisingly, the largest number of significant sites were within progenitor clusters, for which we had detected relatively lower numbers of DR genes. The presence of differential chromatin accessibility within progenitors in the absence of equally large numbers of differentially expressed genes was suggestive of a, broadly defined, memory-like state wherein epigenetic changes were present despite a lack of evidence that progenitors are being directly activated by an external stimulus. This disconnect between chromatin accessibility and gene expression was most obvious within CMPs and GMPs and less apparent within MEPs for which we detected relatively high levels of differentially expressed genes *and* differentially accessible peaks within MEPs. Moreover, MLPs and PreBNK cells harbored fewer numbers both of

epigenetic and gene expression changes, making it difficult to conclusively determine to what extent epigenetic changes are reflective of current, versus past, activation. With that being said, we did not detect differential transcription factor activity for most of the TFs enriched at DR peaks in these progenitor clusters, suggesting that DR peaks in MLPs, MEPs, and PreBNKs may have been reflective both of past or upstream activation and of current differential gene expression programs.

Two questions stood out from these data. First, why did HSCs have such few changes in chromatin accessibility despite having clear signatures of continued activation, including hundreds of differentially expressed genes and active transcription factors? We speculate that because HSCs are pluripotent and can theoretically "choose" to differentiate down any lineage, most of the chromatin would generally be expected to be more open at baseline, allowing the stem cells to maintain the potential to make diverse fate decisions. We reason that on one hand, it could be challenging to detect the increased opening of something that is already very accessible and on the other hand, it would likely not, from an evolutionary perspective, be very beneficial for stem cells to dramatically close chromatin, as this could lead to a long-lasting impairment of pluripotency. Future investigations into the baseline levels of openness of peaks within HSCs, would directly indicate whether chromatin is, at baseline, truly more open in stem cells.

Second, what drove the chromatin accessibility changes seen in progenitor clusters, given the lack of evidence for large-scale activation? We performed transcription factor foot printing and found that CMPs and GMPs are highly enriched for binding of specific transcription factors families, including EGR1 and multiple KLF families. Enrichments of these TFs were shared across almost all other progenitor clusters as well, which suggested a potential shared set of TFs

responsible for inducing DR peaks. We identified a core set of 9 transcription factor classes with significantly enriched binding within at least 3 clusters. Strikingly, among the top 7 of these TFs, 4 (almost 60%) continued to have differential activity within upstream HSCs at day 90, nearly 4 times the proportion one would have expected by mere chance ($p=0.012$). These data suggested that current differential transcription factor activity within HSCs at D90 could have directly influenced downstream epigenic changes. However, we found a very low rate of physical overlap when comparing differentially accessible peaks between progressively differentiated clusters of the same lineage (for example, HSC vs. CMP, CMP vs. GMP), directly demonstrating that sites of differential accessibility within one cluster were not precisely maintained or copied during the differentiation process. The lack of peak sharing, despite the common set of enriched transcription factors suggested that in general, differential chromatin accessibility was established by a core set of required transcription factors interacting with a cluster-specific trans-environment. Indeed, when looking at the top-most enriched TFs within each cluster we found that DR peaks within most clusters enriched both for common TFs, and at least one other TF that was unique only to the cluster. For example, MLPs and GMPs both had strong enrichments for KLF/SP1 binding, but MLPs co-enriched for TCF7 binding, while GMPs uniquely co-enriched for ETS binding.

Based on these collective findings, we hypothesize a model whereby BCG vaccination induces continued activation of a core set of TFs within HSCs (including KLF/SP and EGR families), either though increases in TF abundance, which we found to be true for KLF6, or through post-translational modifications such as phosphorylation. Within HSCs these activated TFs bound to several sites to promote activation of their respective gene targets, leading to many detected DR genes in our scRNA-seq data and differential TF activity detected in the regulon analysis.

However, during differentiation (for example, HSC to CMP differentiation), binding of these TFs was restrained by their interactions with cluster specific TFs, leading to cluster specific DR peaks and a more limited differential gene expression program. In line with this model, DR peaks in MLPs enriched for common HSC-active TFs such as KLF/SP_1 but also strongly enriched for TCF7, a unique TF to MLPs which had differential activity only in MLPs, exemplifying the idea that a common set of transcription factors co-bind to DR peaks together with a cluster-specific TF to help promote its gene expression program. The same paradigm was found to be true within GMPs, in which ETS_3 binding was enriched at DR peaks and accompanied by significant ETS_3 differential activity (Figure 1.19).

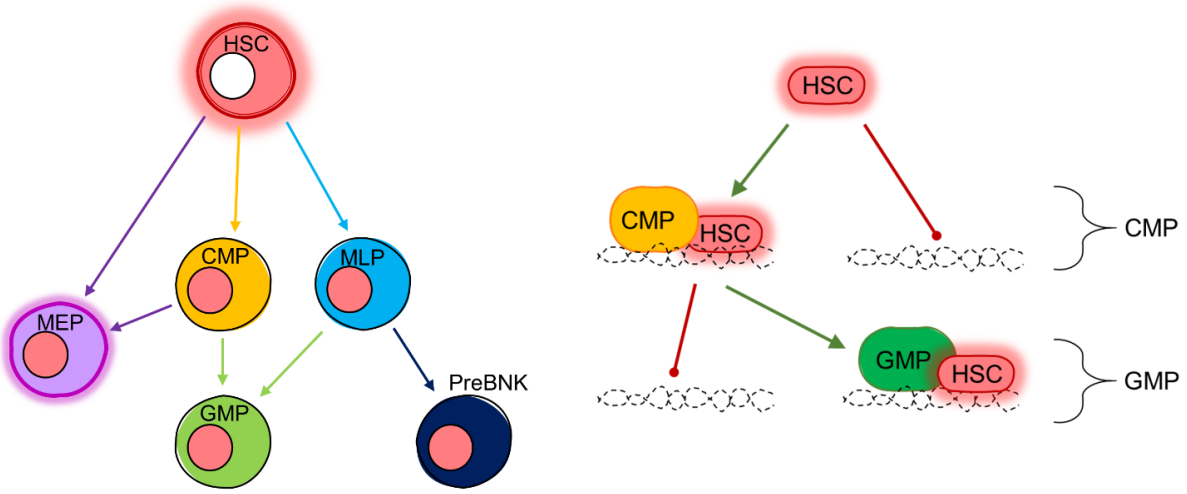

Figure 1.19. Model 1

BCG vaccination induces TF activation and differential gene expression within HSCs and MEPs (halo around HSC and MEPs, *left figure*). Changes in chromatin accessibility in downstream progenitors occurs at binding sites of these HSC-active TFs which co-bind with cluster specific TFs (HSC color-matched nucleus in progenitors, *left figure*). The figure on the right depicts an activated TF in HSCs that co-binds with cluster-specific TFs in CMPs and then GMPs, leading to cell-type specific peaks and gene expression programs.

This model would explain why differentially accessible peaks within every progenitor cluster enriched strongly for at least one transcription factor containing current differential activity within HSCs yet also each co-enriched for binding of a cluster-specific transcription factor. The fact that 3 out of the 4 core transcription factors enriched across many progenitor clusters and harboring differential activity within upstream HSCs were of the KLF or KLF/SP family was notable and supportive of this hypothesis given the known role of KLF/SP transcription factors as ubiquitous transcription factors that are still able to mediate tissue and context specific expression due in large part to co-binding[83]. For example, in one study corticosteroids were found to induce gene expression through the coordinated co-binding of KLF4 and the glucocorticoid receptor (which also acts as a transcription factor) to sites that harbor a CACCC box and GRE element[84]. Within macrophages, on the other hand, KLF4 cooperated with STAT6 to induce an M2 polarized state[85], demonstrating that the same KLF transcription factor can perform completely different functions when placed in different cellular contexts and forced to interact with different binding partners.

One added layer of complexity that also likely contributed to cluster specific DR peaks is the fact that different clusters falling along different pseudotimes, such as CMPs and GMPs, are derived from HSCs from different points in time. In other words, the CMPs profiled in our data were not the same exact cells which gave rise to the GMPs profiled in this same data set. Any epigenetic memory of upstream HSC activation harbored within CMPs is actually representative of HSCs at a more recent point in time compared to the memory encoded within GMPs, which is reflective of the state of HSCs from a time point farther in the past. Thus, this adds an additional level of complexity whereby, in addition to cluster specific interactions between shared HSC-active TFs and cluster-specific TFs, the exact inflammatory memory propagated from HSCs downward is

not constant over time, but changes depending on the dynamic nature of upstream HSC activation. When plotted onto a UMAP, our scATAC data indicated that within each lineage, the clusters harboring the greatest total number of differentially accessible peaks were all approximated to have the same pseudotime (Figure 1.20), suggesting that a previous stronger wave of inflammation once acted on upstream HSCs, encoding more extensive epigenetic changes now present within progenitor clusters at later pseudotimes, while clusters at earlier pseudotimes in our data, have fewer total numbers of DR peaks due to a dampened level of activation within upstream HSCs (Figure 1.20).

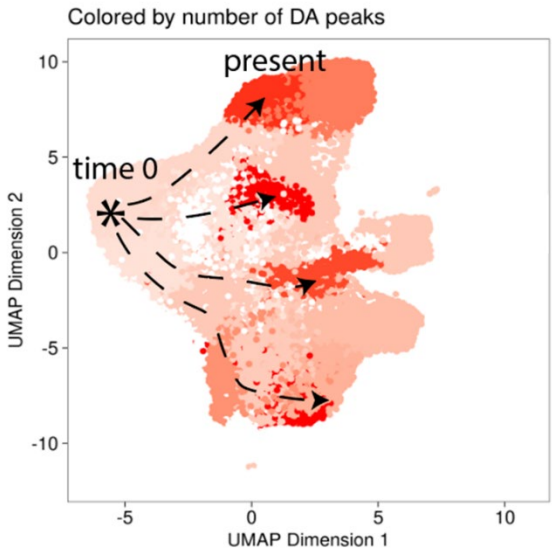

Figure 1.20. Model 1 part 2

Clusters at different pseudotimes harbor epigentic memory induced by HSCs at different timepoints. Clusters with the greatest number of differentially accessible peaks appear at similar pseudotimes suggestive of a past wave of heightened HSC activity.

Altogether, model 1 suggested that interactions between common TFs and cluster-specific TFs, combined with the dynamic nature of HSC activation across time, led to different sets of differentially accessible peaks within different lineages and different pseudotimes. Notably, these data do not completely rule out other possibilities. The second possible scenario is that the epigenetic changes within each cluster were encoded independently through direct immune activation acting on each progenitor, representing a fundamentally different model of continued direct activation as opposed to the downward propagation of HSC transcription factor-based memory proposed in model 1. In this case it would still not be surprising that differentially accessible peaks all enrich for a common set of transcription factors. These shared transcription factors could just reflect the fact that BCG vaccination induces a response within each cluster that involves the activation of a shared set of TFs in addition to cluster specific TFs (Figure 1.21).

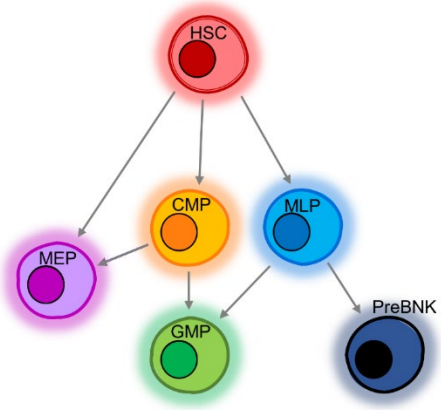

Figure 1.21. Model 2

BCG vaccination directly activates each HSPC cell type independently inducing differences in chromatin accessibility in each cluster.

Nonetheless, model 2 still leaves some observations unanswered. This model suggests that the nature by which HSCs and MEPs, compared to other progenitor clusters are activated is fundamentally different. Specifically, it suggests that HSCs and MEPs respond to external activation by engaging a strong differential expression program, whereas other progenitors only activate one or two transcription factors but gain many changes in chromatin accessibility. In that sense model 1 fit the data better since it modelled two different processes to account for the differences observed when comparing HSCs and MEPs with the other progenitor clusters: activation in HSCs and MEPs, and residual "memory" of activation within other downstream progenitors.

Many of the experiments that would be required to differentiate between the proposed ideas (Model 1 versus model 2) center around answering the key recurring question of whether BCG vaccination acts directly on each HSPC cluster individually, or whether only certain clusters such as HSCs can sense pathogens or pro-inflammatory cytokines and then propagate these gene expression and/or epigenetic features to downstream clusters. In theory, myeloid progenitors can sense and respond to certain cytokines. For example, G-CSF can be sensed directly by GMPs, causing them to expand[52], however it is not known exactly what happens in the context of BCG within humans. Experiments performed on individual HSPC clusters characterizing their capacity to respond directly to BCG or to pro-inflammatory cytokines would already provide great insight into determining whether model 2 is feasible.

In general, additional mechanistic-level experiments should center on understanding the consequences of knocking out some of the core transcription factors discovered in this data. Particularly interesting targets would be KLF6, KLF5, and EGR1 given that all three of these

TFs had increased activity within upstream HSCs at D90 and were significantly enriched among differentially accessible sites within multiple downstream progenitors. For example, one-by-one knockout of EGR1, KLF5, and KLF6, and triple knockout across all HSPCs followed by BCG vaccination would help determine the necessity and sufficiency of these transcription factors for the induction of differential chromatin accessibility. Another more challenging experiment to directly test the sufficiency of differential transcription factor activity within HSCs in encoding differentially accessibility within downstream progenitors would involve inducible knockout of EGR1 or KLF transcription factors within all non LSK (all progenitor) cells allowing these TFs only to be activated within HSCs. Subsequent time course scATAC-seq would determine if activation of these TFs within upstream HSCs is sufficient to induce the full repertoire of DR peaks and associated DR genes within downstream progenitors.

DR peaks and HSC differential expression predict cytokine secretion in PBMCs

Using cytokine secretion data from donor matched PBMCs, we found that log2FC chromatin accessibility values of GMPs had significant power to predict fold change IL1B secretion of PBMCs 3 months after BCG vaccination compared to before. Similarly, we found that log2FC expression values of more than 200 genes within HSCs, including several transcription factors, also had equally significant power to predict IL1B responses. In addition to gene expression, transcription factor activity scores as determined via regulon analysis correlated very strongly ($R > 0.7$) and significantly with IL1B. These results support two main ideas. First, the fact that peak accessibility within GMPs correlated with IL1B production suggests that GMPs harboring differences in peak accessibility continued to be epigenetically modified even as they differentiated into mature myeloid cells and entered the peripheral circulation, a model which is

in line with the idea that innate immune epigenetic memory encoded within stem cells of the bone marrow have the potential to produce epigenetically rewired innate immune cells[6]. Notably, chromatin accessibility differences within other downstream clusters such as PreBNK cells were not found to predict cytokine responses in the periphery, perhaps due to the fact that PreBNK cells had fewer total epigenetic changes and the fact that DR peaks did not enrich for nearby genes directly related to immunity and interleukin signaling as in GMPs.

Second, the fact that gene expression and transcription factor activity within HSCs was equally as predictive of IL1B responses directly supports the model that continued transcription factor activity (which manifests as differential gene expression) within HSCs played an instructive role in shaping the chromatin accessibility landscape of downstream progenitors. It implies that in general, those individuals harboring greater increases in transcription factor activity, matched the same individuals with greater log2FC increases in chromatin accessibility within progenitor clusters downstream, which were also the same individuals with the greatest BCG-induced increases in IL1B production in the periphery, a model that has been supported by many mouse studies[45,50,67,71,72,80] but for which mechanistic data is lacking.

Immediate future work will focus on investigating the properties of mature PBMCs in the periphery using 10x genomics multiome analysis through which chromatin accessibility and gene expression can be measured all within the same cell. These multiome analyses will enable direct investigations into the locations of differentially accessible peaks in different mature immune cell types and to characterize the relationship between epigenetic differences within the most differentiated progenitor clusters of the bone marrow, and their respective mature immune cell counter parts in the peripheral circulation. Given the complex binding patterns of

transcription factors during differentiation within the bone marrow, it is likely that mature immune cells may harbor differences in chromatin accessibility at sites that also enrich for KLF and EGR transcription factors, yet do not completely overlap the specific sites found within GMPs. Finally, the ability to characterize differences in gene expression within the same cell will provide clearer insight into the exact relationship between DR peaks and differential expression programs.

Limitations and primary future directions

Although I believe this work provides significant novel insight into human bone marrow in the context of BCG vaccination, there are many ways in which this work could be improved and extended.

First, as alluded to earlier, this work profiles HSPCs but does not include any of the surrounding stromal or mature immune cells that could play important roles in communicating with HSPCs. This has made it difficult to draw conclusions about the broader inflammatory state of each bone marrow sample, and therefore difficult to infer whether continued changes in gene expression, which we believe to be driving most of the overall changes seen across HSPCs, are driven by continued inflammation, or by intrinsic changes to the gene expression program of the HSCs. Although the exclusion of these broader bone marrow cell types in this work was largely cost driven, additional captures of whole bone marrow on a subset of samples as well as cytokine measurements of bone marrow comparing placebo and BCG vacationed donors would greatly inform the broader bone marrow environment within which the HSPCs are residing.

Second, in general, this work is limited primarily to genomic/computational analyses and therefore generates models that could benefit from additional experimental validation. This is

particularly the case for our integration of scRNA and scATAC-seq data in which we proposed that upstream TF activity within HSCs shapes the downstream chromatin accessibility landscape of progenitors. This hypothesis was drawn utilizing genomic and statistical evidence. Primarily, HSCs harbor numerous TFs with significant regulon activity, and these significantly overlap with TFs that are enriched within differentially accessible sites within downstream progenitors. Thus, we inferred a connection between continued TF activity within stem cells, and changes in chromatin accessibility downstream. Direct experiments to further investigate this model should generally involve, first, direct ChIP-seq or similar experimental approaches (CUT&RUN for low cell numbers) to directly probe the binding of the topmost enriched TFs within each celltype. The primary TFs targeted in these experiments should be KLF and EGR TFs because they were among the most highly enriched, commonly bound, and shared across at least 3 different progenitors. ChIP-seq binding for these transcription factors should be performed within each cell type and the binding sites should be compared to identified sites of differential accessibility within our scATAC-seq data to verify their binding to these sites. Additional ChIP-seq would ideally be performed for a series of cluster specifically enriched TFs (for example TCF7 enriched only in MLPs, or ETS3 enriched in GMPs). We would hypothesize that KLF and EGR binding sites within each cluster would also overlap with the binding sites of cluster specific TFs which in our model, are hypothesized to co-bind together with KLF and EGR TFs at cluster specific sites of differential accessibility that enriched for the shared set of KLFs and EGRs, but occur at different physical locations. Additional mouse experiments, although arguably riskier and further removed from the human setting, could also serve as experimental validation. For instance, one could compare whether downstream progenitors harbor differences in chromatin accessibility following BCG vaccination in WT mice versus mice treated with RNA-interference based

downregulation of KLF/EGR transcription factors, or heterozygotes with significantly decreased levels of these TFs. This would establish the necessity of these TFs within HSCs in establishing changes in chromatin accessibly within downstream progenitors. Oppositely, forced upregulation of these TFs within upstream HSCs would indicate whether increased levels of certain TFs are sufficient to induce downstream chromatin accessibility changes.

A third limitation of this work is the lack of coupled measurements in gene expression and chromatin accessibility within the same cells. Our current datasets were collected by splitting cells for either scATAC-seq or scRNA-seq ultimately leading to measurements on the same population of cells, although the exact cells profiled in the two datasets are not the same. Multiome analyses, at least on a few samples, would help to directly demonstrate whether or not changes in accessibility and gene expression are truly uncoupled, or whether power-differences partially drive these observations.

Finally, as the types of epigenetic experiments that can be performed at single cell resolution or small cell numbers expands, more detailed measurements of specific marks (either DNA methylation or histone modifications) would provide more specific insight into the molecular players that drive changes in chromatin accessibility in our data. Although changes in chromatin accessibility are often accompanied by other epigenetic changes or histone modifications, these cannot be directly probed with the current scATAC-seq data, limiting our interpretation of the molecular signatures that either drive, or are accompanied by these changes in accessibility.

MATERIALS AND METHODS

Bone marrow aspirate staining, sorting, and sample collection

Cryopreserved bone marrow aspirates were processed, following the steps detailed below, on 7 separate days/batches, each batch containing 1) males and females and 2) samples collected on both D0 and D90. Five out of the seven batches contained samples from both placebo and BCG vaccinated cohorts (two contained only BCG cohort samples when there were no remaining controls).

Initial thawing and incubation: Cryopreserved samples were thawed and cultured in RPMI 1640 (Fisher) supplemented with 10% fetal bovine serum (Corning), 2 mM L-glutamine (Fisher), 2% HEPES (Thermo Fisher Scientific), 1% non-essential amino acids (Thermo Fisher Scientific), 1% essential amino acids (Thermo Fisher Scientific), 0.14% 5N NaOH, 1mM sodium pyruvate (Thermo Fisher Scientific), 100U/ml penicillin (Thermo Fisher Scientific), and 100µg/ml streptomycin (Thermo Fisher Scientific) for 2 hours. After incubation, samples were washed with PBS, passed through a 100 µm filter, and counted.

Antibody staining: To prepare samples for flow cytometry analysis and sorting, cells were incubated with 1:50 Live/dead fixable blue (Invitrogen) at a final cell concentration of 1M cells/100 µL for 20 mins (on ice). Samples were washed with 1% BSA (Miltenyi Biotec) in PBS (used for all further washing and staining steps) and resuspended in F/C block solution (BD Biosciences) for 10 minutes. Cells were washed and resuspended in a cocktail of antibodies targeting mature and stem/progenitor cell surface markers (See Table 1 in Results-chapter 1) for a final cell concentration of 1M cells per 100 µL. After 30 minutes on ice, cells were washed,

resuspended, and passed through a 70 um Flowmi cell strainer (Fisher) immediately prior to sorting.

Sorting: Sorting was performed on a Symphony S6 cell sorter in the UChicago Human Disease and Immune Discovery core (HDID) using a 100 μm nozzle. For any given batch, samples collected from the same donor were sorted sequentially alternating between starting timepoints (for example, batch1: S1 D0, S1 D90, S2 D0, S2 D90; batch 2: S3 D90, S3 D0, S4 D90, S4 D0). Following sorting, CD34+ cells were washed in 1% BSA in PBS, counted, and then processed for single cell RNA and ATAC captures are described below:

Single cell RNA capture: Immediately prior to capture, samples were combined into two pools (2 or 3 samples per pool). Multiplexed cell pools were used as input for the single cell captures. For pools containing 2 or 3 samples, 6600 cells or 10,000 cells respectively were targeted for collection using the Chromium Single Cell 3' Reagent (v3.1 chemistry) kit (10X Genomics). Post Gel Bead-in-Emulsion (GEM) generation, the reverse transcription (RT) reaction was performed in a thermal cycler as described (53°C for 45 min, 85°C for 5 min), and post-RT products were stored at -20°C until downstream processing.

Single cell ATAC capture: Leftover cells in each pool not used for single cell RNA capture were lysed for 3 minutes to isolate nuclei, transposed, and used as input for the single cell ATAC captures. Variable numbers of nuclei (ranging from 2,026 to 9,085, depending on the number of leftover cells) were targeted for collection using the Chromium Next GEM Single Cell ATAC Reagent (v1.1 chemistry) kit (10X Genomics). Post Gel Bead-in-Emulsion (GEM) generation, the GEMs were incubated in a thermal cycler as described (72°C for 5 min, 98°C for 30 sec, 12

cycles of 98°C for 10 sec, 59°C for 30 sec and then 72°C for 1 min), and post-incubation products were stored at -20°C until downstream processing.

Bulk CD34- processing: Total RNA was extracted from the sorted CD34- cell fraction of each sample using the miRNeasy Micro kit (Qiagen) or miRNeasy Mini kit (Qiagen). RNA-sequencing libraries were prepared using the Illumina TruSeq protocol. Indexed cDNA libraries were pooled in equimolar amounts and sequenced single-end 100 bp reads on an Illumina NovaSeq.

Single cell library preparation and sequencing

Single cell RNA libraries: Post-RT reaction cleanup, cDNA amplification, and sequencing library preparation were performed as described in the Single Cell 3' Reagent Kits v3.1 User Guide (10X Genomics). Briefly, cDNA was cleaned with DynaBeads MyOne SILANE beads (ThermoFisher Scientific) and amplified in a thermal cycler using the following program: 98°C for 3 min, 11 cycles x 98°C for 15 s, 63°C for 20 s, 72°C for 1 min, and 72°C 1 min. After cleanup with the SPRIselect reagent kit (Beckman Coulter), the libraries were constructed by performing the following steps: fragmentation, end-repair, A-tailing, SPRIselect cleanup, adaptor ligation, SPRIselect cleanup, sample index PCR (98°C for 45 s, between 11 and 13 cycles x 98°C for 20 s, 54°C for 30 s, 72°C for 20 s, and 72°C 1 min), and SPRIselect size selection. Prior to sequencing, all multiplexed single-cell libraries were quantified using the KAPA Library Quantification Kit for Illumina Platforms (Roche) and pooled in an equimolar ratio. Libraries were sequenced 100 base pair (read1: 28, i7: 10, i5: 10, read2: 90) on an Illumina NovaSeq.

Single cell ATAC libraries: Post GEM incubation cleanup and sequencing library preparation were performed as described in the Single Cell ATAC Reagent Kits v1.1 User Guide (10X

Genomics). Briefly, post-incubation GEMs were cleaned up first with DynaBeads MyOne SILANE beads (ThermoFisher Scientific) and then with SPRIselect reagent (Beckman Coulter). Libraries were constructed by performing sample index PCR (98°C for 45 s, 9 or 10 cycles of 98°C for 20 s, 67°C for 30 s, 72°C for 20 s, and 72°C 1 min) followed by SPRIselect size selection. Prior to sequencing, all multiplexed single-cell libraries were quantified using the KAPA Library Quantification Kit for Illumina Platforms (Roche) and pooled in an equimolar ratio. Libraries were sequenced 100 base pair (read1: 50, i7: 8, i5: 16, read2: 50) on an Illumina NovaSeq.

Mapping, demultiplexing, and cell filtering

Single-cell RNA-seq data: FASTQ files from each multiplexed capture (n=14) were mapped to the GRCh38-2020-A-2.0.0 human reference genome using cellranger (v6.0.2) (10X Genomics). Demuxlet[86] was used to demultiplex each capture into its constituent samples based on genotypes in a common VCF file containing genotype (GT) and genotype likelihood (PL) for each individual. Demuxlet implements a statistical model to determine the likelihood of RNA-seq reads from any given single cell to map to a set of single nucleotide polymorphisms, therefore leveraging natural genetic variation to differentiate between samples from different individuals. Following demultiplexing, the Seurat (v3.2.3 Rv4.1.0) pipeline was used to retain only high-quality cells based on the following criteria: “singlet” as determined by Demuxlet (“doublets” and “ambiguous” cells removed), percent mitochondrial reads < 15%, and RNA read count (nCount_RNA) > 500. Out of the initial 115,698 cells captured across all batches, 92,014 were retained as high-quality singlets.

Single-cell ATAC-seq data: FASTQ files from each multiplexed capture (n=14) were mapped to the GRCh38-2020-A-2.0.0 human reference genome using cellranger-atac (v2.0.0) (10X Genomics). Demuxlet[86] was used to demultiplex each capture into its constituent samples as described above using the same common VCF file. Following demultiplexing, we used the ArchR (v1.0.1, ArchRGenome: hg38) pipeline to filter the data, retaining only high-quality cells. Cell filtering and the creation of ArrowFiles was performed in a single step using the createArrowFiles function on cells with "singlet" demuxlet status and using parameters minTSS = 4 and minFrags = 1000 to further retain only cells with a sufficient signal to background ratio (high accessibility at transcription start sites) and at least 1000 unique nuclear fragments. Across all batches, 58,988 cells were retained as high-quality singlets.

Clustering, cell type assignments, and UMAP analysis

scRNA-seq data: Following quality-control filtering, we split cells first by timepoint giving rise to two groups of cells: Td0 (from D0 samples, n=42,493) and Tm3 (from D90 samples, n=49,521). Since individuals received either the BCG vaccine or placebo, we further split Tm3 cells into two subgroups: BCG (n=37,999) or CTL (n=11,522) – leading to 3 final groups of cells: Td0, Tm3_BCG, and Tm3_CTL. We ran the function SCTransform separately for each group to normalize and scale UMI counts, to identify the most variable features, and to regress out variables corresponding to percent mitochondrial reads or capture. We then integrated the transformed data using the following Seurat functions: SelectIntegrationFeatures (nfeatures=3000), PrepSCTIntegration, FindintegrationAnchors, and IntegrateData. To perform dimensionality reduction downstream of integration we used the functions RunPCA (npcs=30),

RunUMAP (dims=1:30), FindNeighbors (dims=1:20), and FindClusters (resolution=0.5). This resulted in 23 preliminary clusters.

To annotate the clusters according to HSPC cell type, we used the FindTransferAnchors function (dims = 1:30, reference.reduction = "pca", reference.assay = "SCT", query.assay = "integrated") to map our integrated scRNA-seq data onto a pre-labelled human bone marrow reference dataset (thawed, stained, sorted, and processed for scRNA-seq as described above) we previously annotated using CellID[87].

scATAC-seq data: Following quality-control filtering and creation of arrow files for each sample, we combined all arrow files into a ArchRProject used in all downstream processing steps. Dimensionality reduction, batch effect correction, clustering, and UMAP visualization were performed using the following functions of the ArchR pipeline: addIterativeLSI (with parameters: iterations = 2, resolution = c(0,2), sampleCells = 10000, n.start = 10, varFeatures = 25000, dimsToUse = 1:30), addHarmony, addClusters (resolution=0.8, reducedDims=Harmony), and addUMAP (nNeighbors = 30, minDist = 0.5, metric = “cosine”). To annotate clusters according to HSPC cell type matching those in the scRNA-seq data, we first performed an unconstrained integration using the addGeneIntegrationMatrix function to broadly map each scATAC cluster to a cell type within our scRNA-seq data. Using this approach, we made the following preliminary assignments:

scATAC clusters "C5", "C6", "C19", "C8" → “HSC”

scATAC clusters "C22", "C17", "C2", "C4", "C3", "C1", "C21", "C16" → “CMP”

scATAC clusters "C10", "C23", "C24" → “GMP”

scATAC clusters "C7", "C14", "C15", "C18" → “MEP”

scATAC clusters "C12", "C13" → “MLP”

scATAC clusters "C11", "C9" → “PreBNK”

scATAC cluster “C20” → “unknown”

To generate more detailed cluster mappings (i.e., separating “HSCs” into “HSC_a” or “HSC_b”) we then performed a second-round constrained integration by rerunning addGeneIntegrationMatrix with the newly defined broad group labels, generating the following final cluster assignments: C1: CMP_a; C2: CMP_a; C3: CMP_b; C4: CMP_a; C5: HSC_b; C6: HSC_a; C7: MEP_a; C8: HSC_b; C9: PreBNK; C10: GMP_b; C11: PreBNK; C12: MLP_b; C13: MLP_b; C14: MEP_a; C15: MEP_c; C16: CMP_a; C17: CMP_c; C18: MEP_c; C19: HSC_b; C20: unknown1; C21: CMP_b; C22: CMP_b; C23: GMP_a; C24: GMP_a

scATAC-seq Peak calling

We called peaks using the ArchR function addReproduciblePeakSet which utilizes MACS2 to call cluster-specific peaks using pseudo-replicates, and then creates a merged peak set using iterative overlap peak merging. For peak calling we used the initial raw, unprocessed alignment data but with added cell type labels derived as described above.

Pseudobulk estimates

For downstream analyses of scRNA-seq data we summarized single cell expression into pseudobulk estimates for each sample (each unique donor-timepoint pair), allowing a bulk RNAseq-like approach to investigating effects of BCG vaccination on human bone marrow for each cell type. For each of the final 13 unique clusters with a defined cell type label (HSC_a,

HSC_b, CMP_a, CMP_b, CMP_c, GMP_a, GMP_b, MEP_a, MEP_b, MEP_c, MLP_a, MLP_b, and PreBNK) we summed raw UMI counts belonging to all cells from the same sample using the sparse_Sums function in textTinyR (v1.1.4). Thus, for each cluster we converted an initial cell by gene (n x m) matrix to a sample by gene (s x m) matrix.

For scATAC-seq data, we summarized single cell peak counts into pseudobulk estimates as described above, only using called peaks instead of genes. As described above, we summed raw peak counts belonging to all cells from the same sample using the sparse_Sums function in textTinyR separately for all clusters (n=24). Thus, for each cluster we converted an initial cell by peak (n x p) matrix to a sample by peak (s x p) matrix.

Modelling effect of BCG on gene expression and integration with mashr

Data filtering/normalization/transformation

Gene expression data: For each cell type, we analyzed pseudobulk gene expression as if it were bulk-RNA sequencing expression data. We first removed any samples for which there were fewer than 20 cells, and any samples for which there was not a matching Td0 or Tm3 timepoint (retaining only paired samples). Lowly expressed genes were filtered by removing all genes for which the median logCPM was below a cell-type specific threshold (thresholds: 1.5 for CMP_c and MLP_b; 2 for MEP_c and PreBNK, and 0.5 for all other clusters). Then, we normalized gene expression counts across all samples using the calcNormFactors function implemented in the edgeR R package (version 3.34.1) which utilizes the TMM algorithm (weighted trimmed mean of M-values) to compute normalization factors, and we log-transformed the data using the voom function from the limma package.

Peak accessibility data: Similarly for peak accessibility data, we analyzed each pseudobulk peak count matrix as if it were bulk-ATAC sequencing data. Data was filtered by removing samples with fewer than 20 cells and any samples for which there was not a matching Td0 or Tm3 timepoint. We filtered out low-count peaks for which the logCPM was below a cell-type specific threshold (0.75 for C4, C12, and C25, 1 for C18, 2 for C21, 2.5 for C9, C17, and C23, and 0.5 for all other clusters), then normalized peak counts across all samples using the calcNormFactors function in edgeR, and log-transformed the data using the voom function in limma.

Model fitting

We wanted to investigate the 90-day impact of BCG vaccination on gene expression and peak accessibility in human bone marrow by comparing expression/accessibility levels from vaccinated individuals at day 90 (after vaccination) and day 0 (prior to vaccination). However, expression and peak accessibility measurements can naturally change across time, independent of whether the individual received the BCG vaccine or only a placebo. Moreover, although individuals assigned to the placebo or BCG cohorts were matched for age, sex, and lack of previous BCG exposure there could be random preexisting baseline differences when comparing the cohorts. To correct for these effects, we independently fit scRNA and scATAC pseudobulk data to a mixed model to estimate the impact of time and cohort assignment on expression/accessibility while also giving an estimate of the independent contribution of BCG-vaccination to changes in expression/accessibility at day 90.

Separately, for each feature (genes or peaks) and each cell type, we fit the following model:

$$M_1: E(i,j) \sim \begin{cases} \beta_0(i) + Zu + \varepsilon(i,j) \; if \; Condition = placebo \; at \; D0 \\ \beta_0(i) + \beta_{BCG\_cohort}(i) + Zu + \varepsilon(i,j) \; if \; Condition = BCG \; at \; D0 \\ \beta_0(i) + \beta_{D90}(i) + Zu + \varepsilon(i,j) \; if \; Condition = placebo \; at \; D90 \\ \beta_0(i) + \beta_{D90}(i) + \beta_{BCG\_cohort}(i) + \beta_{vaccination}(i) + Zu + \varepsilon(i,j) \; if \; Condition = BCG \; at \; D90 \end{cases}$$

Where *E(i,j)* represents the estimate for each feature *i* and sample *j*. *E(i,j)* is modelled as a function of the fixed effects, β_0, β_{D90}, β_{BCG_cohort}, and $\beta_{vaccination}$, and the random effects Zu. β_0(i) represents the intercept for the feature i, β_{D90}(i) is the natural effect of time on feature *i*, β_{BCG_cohort}(i) represents pre-existing baseline differences in feature *i* between the control and BCG cohorts, and $\beta_{vaccination}$(i) represents the effect of BCG vaccination on feature *i* at D90. The vector *u* is an mx1 vector of random effects to control for individual donor differences where m is the number of unique donors (m=X; m=j/2). Z is an incidence matrix of 1's and 0's that maps each sample *j* to one of *m* individuals. The model was fit using the R package EMMREML**.**

Mashr

To increase our power to detect BCG-responsive genes shared or unique to each cell type, we applied Multivariate Adaptive Shrinkage in R (mashr version 0.2.57) to outputs from emmreml for scRNA-seq data. We did not apply mashr to scATAC data because peaks accessible enough to pass initial filtering steps are highly cell type specific, decreasing the utility of mashr in this context. For scRNA data, effect sizes were obtained by extracting the betas ($\beta_{vaccination}$) for each cell type and the standard error of the effect size for each gene was given by taking the square root of varbeta estimates from emmreml. Effect sizes and standard errors for each cell type were arranged into n x m matrices, n being the number of genes and m being the number of cell types. We then fit the mash model using canonical and data driven covariance matrices and then stringently defined significant genes as those with an lfsr < 0.01.

Gene set enrichment analysis

Gene set enrichment analyses (GSEA) were performed using the fgsea R package (version 1.18.0) with parameters: maxSize = 500, nperm=100000. To investigate biological pathway enrichments among BCG-responsive genes, we ordered genes by the rank statistic: -log10(lfsr)*PM where lfsr and PM (posterior mean) were output from running mashr as described above. The rank-ordered gene list was compared with the Hallmark gene sets from the MSigDB collections.

Velocyto, Cellrank, and terminal state prediction

We used velocyto followed by the CellRank[58] pipeline to determine single cell RNA-velocity measurements and to predict the terminal lineage fate of HSCs from each sample.

We first used the velocyto run10x command to quantify spliced and unspliced read counts (which are required downstream in the pipeline to estimate RNA velocities) for each gene within every cell of our scRNA-seq dataset. Then, separately for cells of each unique donor-timepoint sample, we ran the CellRank pipeline in python to predict terminal fates of individual HSCs within each sample. Briefly, for each sample, we first removed genes with very low spliced/unspliced mRNA counts, normalized and log-transformed the data, subset on only the top-most variable genes, and computed principal components and moments for velocity estimation using the following CellRank functions: scv.pp.filter_and_normalize (with parameters min_shared_counts=20 and n_top_genes=2000), sc.tl.pca, sc.pp.neighbors (with parameters n_pcs=30 and n_neighbors=30), and scv.pp.moments (with parameters n_pcs=None and n_neighbors=None). Next, we used dynamical modelling to estimate RNA velocities for each single cell using the function scv.tl.recover_dynamics and computed a velocity graph indicating

the likelihood that one cell will transition into another based on their RNA velocities and relative positions using scv.tl.velocity(mode="dynamical") and scv.tl.velocity_graph. Visualization of these velocity graphs was performed with the function scv.pl.velocity_embedding_stream.

We then used a velocity Kernel to formally predict the terminal lineage fate of each HSC for each sample. We first used the commands VelocityKernel and vk.compute_transition_matrix on the single cell data, pre-processed as described above, to compute a cell-cell transition matrix based on RNA velocity. We combined this velocity kernel with a connectivity kernel to create a less noisy combined kernel (combined_kernel = 0.8 * vk + 0.2 * ck). Using a GPCCA (Generalized Perron Cluster Analysis) estimator, we computed a schur decomposition with g.compute_schur(n_components=20). Finally, we pre-defined all possible terminal states using g.set_terminal_states (with possible states: MLP_a, MLP_b, GMP_a, GMP_b, MEP_a, MEP_b, MEP_c, CMP_a, CMP_b, CMP_c, PreBNK) and then calculated the terminal state probabilities for each HSC using g.compute_absorption_probabilities(use_petsc=True, n_jobs=5, solver='gmres').

To compare terminal state differentiation probabilities across time for any given donor we labelled each single HSC with the terminal state towards which it had the greatest differentiation probability. For each donor (excluding donors with fewer than 20 total HSCs) we then computed the percentage of HSCs at day 0 and day 90 having maximal differential probability towards each terminal state and computed the difference across time (%day90 - %day0), leading to a "differentiation-shift" score for each possible terminal state, for each donor. Differentiation-shift scores were normalized by subtracting the median score among placebo vaccinated individuals

from every donor. For statistical comparisons of normalized differentiation-shift scores of BCG and placebo groups we used a Wilcoxon test.

MPP score

Calculation of the MPP score was based on known differences in the expression of CD90, CD49f, and CD45RA between MPPs and LT/ST-HSCs. To calculate the score, we first obtained mean expression values for each of the three genes across single cells for each HSC subcluster. The unprocessed mean values were centered and scaled using the *scale* function in R to normalize all values to a mean=0 and standard deviation of 1. Scaled scores for each individual gene were averaged to generate the final composite score.

scHINT

Preprocessing: Transcription factor motif enrichments and foot printing were performed using HINT-ATAC from the Regulatory Genomics Toolbox[68]. Raw bam files for each 10X capture were split by vaccination cohort, timepoint, and assigned cell type using *samtools* (v1.9) *view*. We focused on comparing BCG samples at D0 and D90, so only BCG samples (i.e., BCG_D0_HSC, BCG_D90_HSC, BCG_D0_CMP, BCG_D90_CMP, …) were processed further. Matching bam files from each capture were merged using *samtools merge* to generate BCG D0 and BCG D90 merged bam files for HSCs, CMPs, GMPs, MEPs, MLPs, and PreBNK (12 total files) for downstream foot printing and motif analyses.

Motif enrichment: To determine which motifs were present within DR peaks we performed *rgt-motifanalysis matching* on DR peaks using the JASPAR CORE Vertebrates set of curated position frequency matrices[88] to determine whether specific motifs were significantly enriched

we used the *rgt-motifanalysis enrichment* function with cluster-specific DR peaks as the foreground and the shared total peak set as the background for all clusters.

Foot printing: To predict the locations of transcription factor footprints we ran the *rgt-hint footprinting* function with parameters --atac-seq --paired-end --organism=hg38 on all peaks for each merged bam file generated in the preprocessing step. To predict which transcription factors were likely bound at each predicted footprint, we used the *rgt-motifanalysis matching* function to find motifs present within footprints.

Assigning DR peaks to genes and GO enrichment

To investigate which genes were located closest to differentially regulated peaks, we assigned each DR peak to the gene with the closest TSS using the Homer function *annotatePeaks* with default parameters. This peak-gene association was performed separately for DR peaks within CMPs, GMPs, HSCs, MEPs, MLPs, and PreBNK clusters. To determine whether specific pathways were enriched among genes closest to DR peaks we specified the parameter -GO when running the annotatePeaks function which outputs peak-gene assignments and gene ontology enrichments using DR peaks as foreground peaks and a total peak set (common to all clusters) as background. Output gene ontology enrichment p-values were corrected with the *p.adjust* function in R.

Regulon analysis

We used pySCENIC[70], the python implementation of the SCENIC pipeline[69], to predict transcription factor activity levels within each cluster. Briefly, we first created a loom file for each cluster for which the analysis was to be performed using the *build_loom* function implemented in the SCopeLoomR package. Then we used the *pyscenic grn* function on the loom

object to derive co-expression modules from the single cell expression data. Next, the *pyscenic ctx* function was run with default parameters to search for transcription factor motifs at promoter regions among members of each co-expression module and to trim targets lacking the target transcription factor motifs. Finally, *pyscenic aucell* was used to generate an activity score for each pruned co-expression module for every cell.

To compare transcription factor module activity scores across different conditions, we averaged activity scores for all cells belonging to the same sample to generate average TF activity scores per donor per timepoint. For each donor, we computed the Tm3/Td0 activity score ratio to compute the fold change in activity score across time. Then we compared Tm3/Td0 activity scores between donors of the placebo versus BCG cohorts and used the Wilcoxon rank sum test to derive a p-value.

Elastic net regression

We built an elastic net model using the *glmnet* R package[89] to determine whether the magnitude of BCG-induced differential accessibility of peaks within progenitors, or differential gene expression in HSCs, was predictive of the log2FC value of cytokine production of PBMCs after BCG vaccination. To choose the optimal value of alpha, we tested alphas ranging from 0 to 1 in increments of 0.1 and chose the alpha that maximized the R2 value between the elastic net predicted IL1B log2FC values, and their experimentally measured values. The regularization parameter lambda was chosen to minimize mean-squared error during n-fold internal cross-validation.

We used a leave-one-out cross-validation approach to generate predicted IL1B log2FC values for each donor. We first separated all samples (each sample corresponding to a donor) into training

and test samples and quantile normalized the raw log2FC values (BCG vs. placebo) for each differentially accessible peak, or differentially expressed gene, within each sample to a standard normal distribution. Then, we split the test sample from the training samples, and on the remaining training samples, quantile normalized across samples to a standard normal distribution. For each peak/gene within the test sample, we compared log2FC differential accessibility/expression to the empirical cumulative distribution function for the training samples. This allowed us to estimate the quantile into which the peak/gene fell and to assign this quantile value using the *qnorm* function in R.

Correlations

All correlations were performed with the *cor.test* function in R with parameter method= "spearman".

CHAPTER III: PERSISTANT EPIGENETIC SIGNATURES OF PREVIOUS ACTIVATION ARE COUPLED TO CONTINUED TRANSCRIPTON FACTOR ACTIVITY

INTRODUCTION

Although trained immunity has emerged as a focal point in immunology research there are still many unknowns regarding the mechanisms required for its development and maintenance. Historically, trained immunity has primarily been linked to histone post translational modifications (PTMs)[8,9,19,20,25,27], often co-occurring with differences in chromatin accessibility but not easily profiled at single cell resolution. Histone modifications are well known to be highly dynamic in response to stimulation[6,25,27,90], occur at well-defined regulatory regions such as enhancers and promoters[18,91,92], and can be studied in bulk populations using functional genomics techniques such as ChIP-seq. Early studies in vitro, demonstrated that monocyte-derived macrophages stimulated with β-glucan gain histone modifications H3K4me1, H3K4me3, and H3K27Ac at regulatory regions and that these changes can be detected throughout their in vitro lifespan[8,9,20]. These macrophages are reprogrammed ("trained") to secrete higher levels of proinflammatory cytokines IL-6 and TNFα upon a secondary challenge[8,9,20]. Due to the correlative nature of these studies, however, we do not know the extent to which changes in histone modifications are causally required for the induction of trained macrophages.

In dividing cells, such as tissue resident macrophages or pluripotent stem cells, the situation is further complicated. During DNA replication parental PTMs are diluted on sister chromatids. Thus, if the newly synthetized histones on daughter strands are not modified post-replication, the

information encoded in histone PTMs will be rapidly lost by dilution in successive rounds of cell division. Moreover, frequent histone exchange post-replication in active chromatin regions can further challenge the inheritance of histone PTMs[93], which is expected to primarily compromise the mitotic inheritance of active chromatin states. Supporting that view, a recent study found preservation across the cell cycle of biotinylated histones in repressed domains, but not among transcriptionally active sites[94,95]. This suggests that chromatin components that bookmark active regions – thought to be central for the trained immunity phenotype – should be short-lived, which seems hard to reconcile with the long-term effects – sometimes in the scale of several years post primary stimulus – described for trained immunity[12,14,29,38,39,45,47,67,71,96]. To date, most of the studies that have investigated long-term effects of trained immunity in self-renewing cell types have focused on tissue resident macrophages or hematopoietic stem and progenitor cells (HSPCs) using mouse models[97]. In chapter II we found that BCG vaccination also induced differential chromatin accessibility for at least 90 days in human bone marrow. While this work and other in vivo based studies may be the best way to study trained immunity in the biological context they pose barriers to answering basic questions related to the inherent ability of dividing cells to retain trained immunity signatures without input from other cell types in the microenvironment or from continued low-level stimulus-persistence and inflammation. While transplantation studies can partially solve this problem in mouse studies, it is often difficult to isolate enough cells to enable deep profiling over the course of many cell divisions.

Here we explored the mechanistic basis of innate immune memory in an isolated, dividing macrophage population through dense time course transcriptional, epigenetic, and functional profiling of macrophages following stimulation and washout of beta-glucan – a common trained immunity inducing stimulus. We find that trained macrophages are transcriptionally,

epigenetically, and functionally re-programmed for at least 14 cell divisions after stimulus washout. However, epigenetic signatures remaining through multiple rounds of cell division are always coupled with the continued activity of transcription factors – a finding consistent with our model that suggests that stimulus induced epigenetic signatures may not be self-sustained. In this macrophage model we find that many of the differences observed at late timepoints arise not from retention of beta-glucan induced signatures, but from new waves of coupled transcription factor activity and H3K4me1 deposition beginning days following stimulus washout. Thus, our data points to a dynamic process, as opposed to static retention and propagation of histone modifications, as underlying long-lasting trained immunity within isolated, dividing macrophages.

RESULTS

iBMDM$^{NF\kappa B\text{-}GFP}$ cells enable dense time course profiling of epigenetic and gene transcriptional dynamics after beta glucan stimulation

We sought to investigate how exposure to beta glucan impacts the transcriptional, epigenetic, and functional profile of dividing macrophages over the course of many cell divisions post stimulus-washout. Specifically, we wanted to choose an experimental setup that would allow 1) controlled stimulus application and removal, 2) synchronous cell divisions, and 3) a pure population of cells large enough to enable multimodal profiling at many timepoints (Figure 2.1a). Although we explored the possibility of using cultured primary bone marrow derived macrophages (BMDMs) as a pure population of dividing macrophages, we found it difficult to control their division rates and to passage these cells for an extended period of time. This

prompted us to explore the possibility of using immortalized BMDMs (iBMDMs) as a representative, and easily manipulatable, model of primary BMDMs.

To generate a reporter iBMDM line, we transduced iBMDMs (mouse BMDMs immortalized by infection with J2 retrovirus[97]) with lentiviral particles containing an NF-kB inducible GFP construct and then clonally selected successfully transduced iBMDMs to generate monoclonal reporter cells (hereafter referred to as iBMDM$^{NF\kappa B\text{-}GFP}$), which express GFP when NF-kB is active. A population of GFP+ cells emerged after stimulation of iBMDM$^{NF\kappa B\text{-}GFP}$ cells with β-glucan or Pam3CSK4 (a known NF-κB inducer), but not IFNα, confirming the specificity of the reporter and its ability to distinguish activated from non-activated cells within a population (Figure 2.1b). To characterize the division rates of our reporter line, we incubated iBMDM$^{NF\kappa B\text{-}GFP}$ cells with EdU (a thymidine analog that is incorporated into actively replicating DNA), harvested cells at multiple timepoints, and fluorescently labelled the EdU such that the extent of incorporation could be detected using flow cytometry. Using this method, we found that EdU positivity first reached a maximum in cells incubated for 12 hours, implying full replication of DNA within this time frame (Figure 2.1c). EdU positivity also peaked at 12 hours in the same time course incubation experiment performed on iBMDM$^{NF\kappa B\text{-}GFP}$ cells that had previously been stimulated with beta glucan, indicating that, on a population level, both naïve and post-stimulation iBMDM cells undergo one cycle of DNA replication within approximately 12 hours.

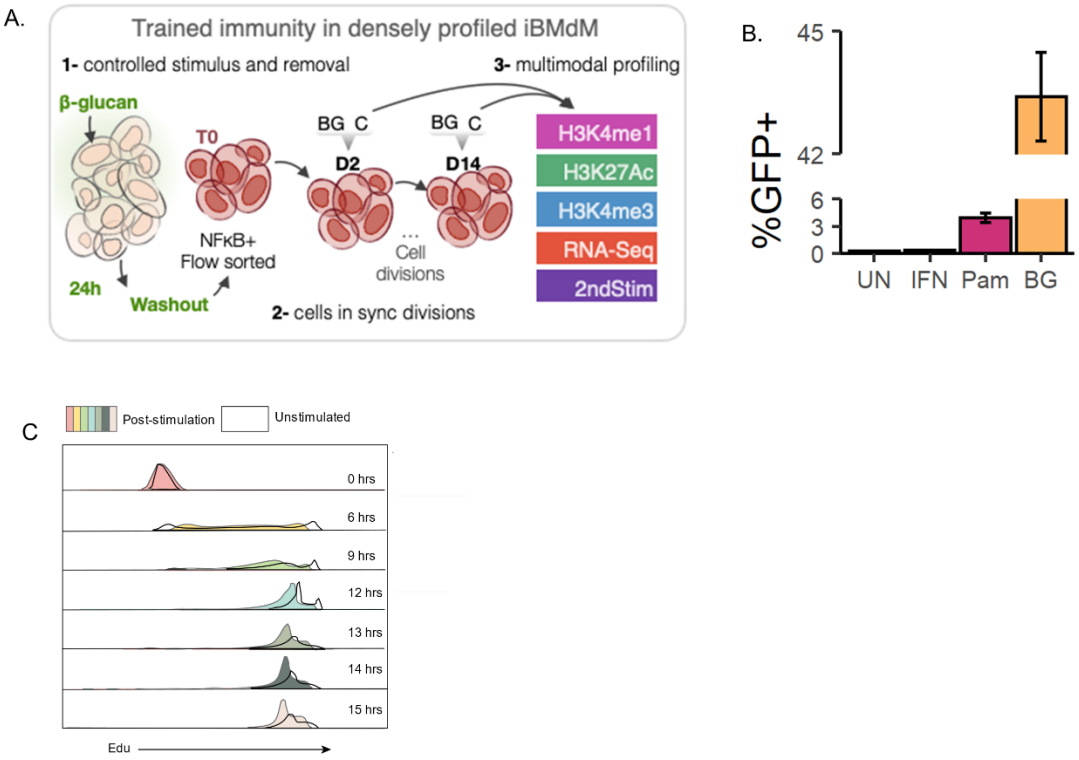

Figure 2.1. iBMDM reporters respond to beta glucan and divide rapidly

a. General experimental plan. **b.** Bar plot quantification (plotted as mean ± SD) of percent GFP+ cells among unstimulated cells, or cells stimulated with IFN (negative control), Pam3CSK4, or BG. Data is representative of 3 replicates per condition. **c.** Naïve (no stimulus) iBMDM$^{NF\kappa B\text{-}GFP}$ cells or iBMDM$^{NF\kappa B\text{-}GFP}$ cells previously stimulated with 30 μg/mL beta glucan were incubated with EdU for 6, 9, 12, 13, 14, or 15 hours. At each time point, paired unstimulated and post-stimulation cells were collected. EdU was fluorescently labelled (BV421) by Click-it reaction performed as described in the Click-it-EdU protocol. EdU incorporation was quantified by flow cytometry (gated on single cells)

Having established a reporter system with defined division rates, we characterized the initial gene expression response of iBMDM$^{NF\kappa B\text{-}GFP}$ cells to beta glucan stimulation. iBMDM$^{NF\kappa B\text{-}GFP}$ cells were stimulated with beta glucan and harvested for RNA-sequencing at multiple timepoints

to examine genome-wide dynamic changes in gene expression. iBMDM$^{NF\kappa B\text{-}GFP}$ cells rapidly upregulated hundreds of immune genes (Figure 2.2a,b). The total gene expression response peaked after 7 hours of beta glucan stimulation, after which the total number of differentially expressed genes declined (Figure 2.2b). The response of iBMDM$^{NF\kappa B\text{-}GFP}$ cells to beta glucan is comparable to the one engaged by primary BMDMs (Figure 2.2c; Pearson's r = 0.61, P< 1x10-16, 82% concordant in the direction of the effects), further supporting their validity as an experimental model to study gene regulatory responses to immune stimulation. To confirm that beta glucan could be fully removed after stimulation, we took advantage of the fact that iBMDM$^{NF\kappa B\text{-}GFP}$ cells act as a reporters of cell activation (NF-kB activity). We quantified GFP levels among sorted iBMDM$^{NF\kappa B\text{-}GFP}$ GFP+ cells compared to paired control cells at 12-hour intervals (every cell division) following BG washout (Figure 2.2d). GFP levels in the BG-stimulated cells were significantly higher compared to controls immediately following stimulation (T0) and for the next two cell divisions. By D3 population-level GFP differences became insignificant. Thus, we defined any timepoint beyond 2 cell divisions as representative of an inactivated state. Collectively these data demonstrated that iBMDM$^{NF\kappa B\text{-}GFP}$ cells divide within 12 hours, induce an immune response to BG resembling that of primary BMDMs, become strongly and selectively GFP+ in response to BG activation, and can return to an inactivated state within 3 cell divisions after beta glucan washout.

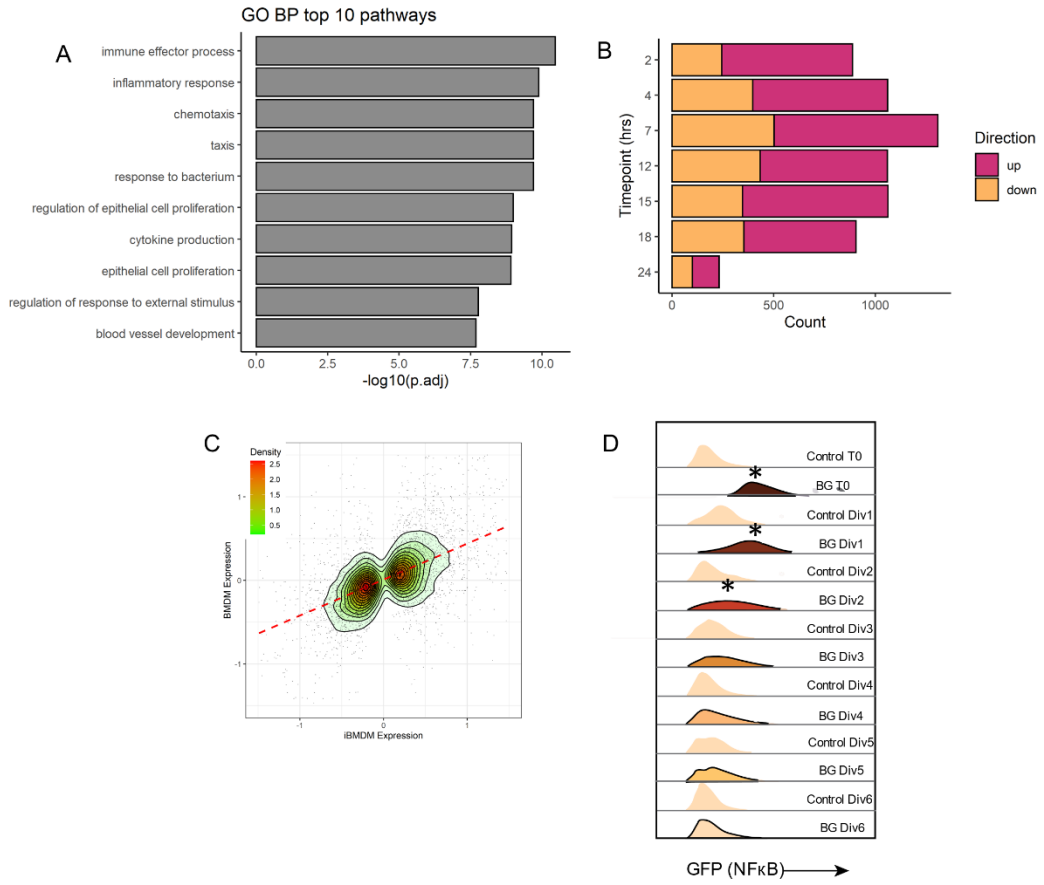

Figure 2.2. iBMDM reporters return to baseline within 3 divisions

a. Gene ontology analysis with biological process (BP) pathways performed on significant genes (FDR<0.05, abs(logFC)>1 after 7 hrs beta glucan stimulation). The top 10 pathways with the lowest p.adj values are shown. **b.** Quantification of the number of differentially expressed genes (Log2FC>1, FDR<0.05) in iBMDMs after various hours of BG stimulation. **c.** Correlation analysis of DE genes in primary BMDMs and iBMDMs at 7 hours of BG stimulation (Pearson's r = 0.61, P< 1x10-16, 82% concordant in the direction of the effects, red dotted line indicates line of best fit). **d**. Histograms from flow cytometry analysis of GFP levels (FITC channel) in BG^{exp} cells over time relative to paired controls.

Beta glucan experienced iBMDMs have long-lasting H3K4me1 signatures of previous beta glucan exposure

To assess the long-term impact of beta glucan on iBMDM$^{NF\kappa B\text{-}GFP}$ cells across multiple cell divisions, and to investigate whether stimulus-induced histone modifications could be propagated though cell divisions independently of continued gene activation, we designed a time course experiment to probe gene expression (via RNA-seq) and histone PTM levels (via ChIP-seq) of histone marks associated with promoters (H3K4 trimethylation, or H3K4me3), enhancers (H3K4 monomethylation, or H3K4me1), and their activation levels (H3K27 acetylation, or H3K27ac) every 2 cell divisions (up to 14 divisions) after washout of a 24-hour beta glucan stimulation. We stimulated iBMDM$^{NF\kappa B\text{-}GFP}$ cells with beta glucan for 24 hours and sorted GFP+ cells to enrich for cells that were responsive to the stimulation (Figure 2.3a). Following sorting, we returned control (C) and beta-glucan stimulated, GFP+ cells (BG^{GFP+}) to cell culture (hereafter referred to as BG-experienced/ BG^{exp}) and collected an aliquot of C and BG^{exp} cells at timepoints corresponding to 2, 4, 6, 8, 10, 12, and 14 cell divisions (referred to as D2, D4, etc.) post sorting for transcriptional and epigenetic profiling (Figure 2.3b). Time 0 (T0; immediately after 24 hours beta glucan) samples were collected in a separate set of experiments.

Using these time course data sets (3 independent replicates per time point), we first explored the impact of beta glucan on genome-wide dynamics of H3K4me3, H3K4me1, and H3K27Ac. We focused on these three histone modifications because of their hypothesized role in encoding innate immune memory[8,9,19,20,25,27,90]. For each time point, we first quantified the number of sites with significantly (false discovery rate (FDR) < 0.1) altered levels of each histone modification (Figure 2.3c,d). In line with fact that enhancer associated H3K4me1 and H3K27Ac are known to

be more highly responsive to immune-related stimuli and environmental perturbations compared to H3K4me3[19,90,91,98], we detected larger numbers of sites with altered levels of H3K4me1 and H3K27Ac in BG^{exp} cells collected at T0 (immediately after BG washout, Figure 2.3c,d). In total, we detected only 541 peaks (3% of total peaks tested) with significantly altered levels of H3K4me3 at T0. This number dropped more than 10-fold by D4 and to near-zero levels by D10 (Figure 2.3c,d). Likewise, while more than 7000 (~18%) peaks had altered levels of H3K27Ac at T0, this number dropped to only 70 peaks by D4 and to near-zero levels as early as D8. H3K4me1 was the most responsive histone PTM, with more than 23,000 significant sites (~33%) at T0 but also declined rapidly – to only ~6,500 sites (~5%) within 2 cell divisions.

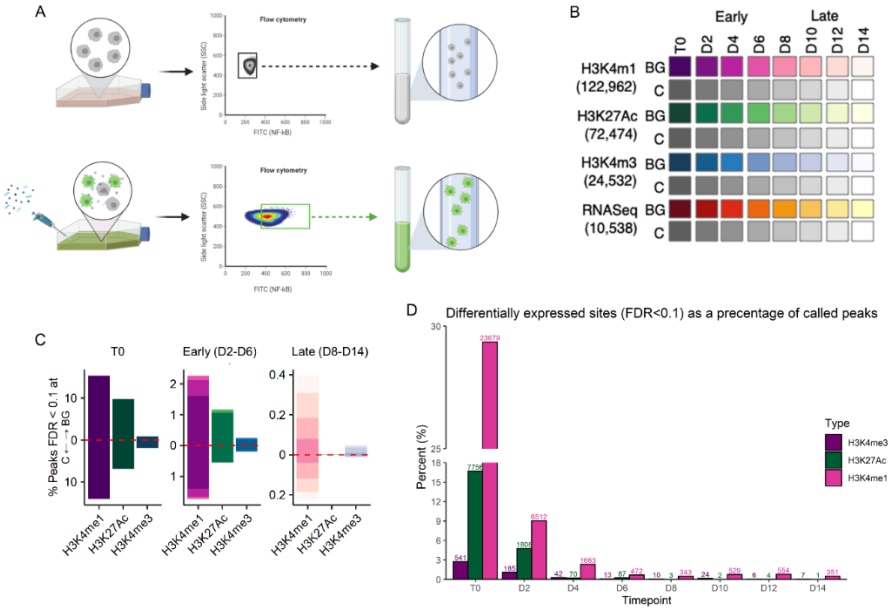

Figure 2.3. ChIP-seq and RNA-seq time course design

a. Cell stimulation and sorting workflow. iBMDM$^{NF\kappa B\text{-}GFP}$ cells are left unstimulated (C, top) or stimulated with 30 µg/mL β-glucan (BG, bottom) for 24 hours. After 24 hours C and BG cells

Figure 2.3, continued

were collected and sorted (C: gated on live, single cells; BG: gated on live, single, GFP+ cells). **b.** Sorted C and BG^{GFP+} cells were returned to cell culture. Aliquots of C and BG^{exp} cells were collected every 24 hours up to 168 hours (D2-D14) and used for downstream RNA-seq and ChIP-seq analysis, number of features that pass all quality control metrics are shown in parenthesis. **c.** Summary bar plots showing the proportion of differentially accessible H3K4me1, H3K4me3, and H3K27Ac peaks at T0, early (D2, D4 and D6) and late (D8 ,D10, D12, D14) timepoints. Red dotted line marks the direction of effect. **d.** Full bar graph showing the proportion of differentially accessible H3K4me1, H3K4me3, and H3K27Ac peaks at every time point

For all three histone PTMs, the initial kinetics of decay closely resembled a model of passive decay, suggesting a lack of mechanisms to preserve these histone PTMs through each round of cell division. This was apparent when we plotted the percentage of histone modifications significant at T0 remaining significant at each subsequent timepoint and compared this to a model of 50% loss with each cell division (Figure 2.4a; K-S test $P_{H3K4me1} = 0.9639$, $P_{H3K4me3} = 0.27$, $P_{H3K27Ac} = 0.6272$) – demonstrating that most beta-glucan induced histone modifications are readily lost with each round of DNA replication.

Despite this being the global pattern, we did detect significantly altered levels of H3K4me1 at more than 300 sites across the entire time course, even after 14 cell divisions following BG removal (Figure 2.3c late). Moreover, principal component analyses performed separately for each histone modification (Figure 2.4b-d) revealed H3K4me1 patterns to be distinct from that of H3K4me3 and H3K27Ac. In the H3K4me1 PCA we observed clear and significant separation along PC2 between control and BG^{exp} samples at every timepoint (Figure 2.4b). In the H3K27Ac PCA, BG^{exp} samples were generally distinguishable from control samples at most timepoints, likely reflecting the fact that although differences in H3K27Ac are no longer statistically

significant at later time points, some differences in the H3K27Ac landscape may remain. However, separation between BG^{exp} and control samples at late timepoints was not as clear as for H3K4me1 (Figure 2.4c). In the H3K4me3 PCA we observed no significant separation between control and BG^{exp} samples along either PC1 or PC2, consistent with the finding that H3K4me3 is the least responsive to BG stimulation and that few significant differences remain past D4 (Figure 2.4d).

We reasoned that the altered levels of H3K4me1 likely reflected a selective rewiring of the enhancer landscape, given that H3K4me1 occurs predominantly at enhancers[18,91,92] and can increase or decrease based on enhancer activity levels. To investigate whether BG stimulation induced long-lasting chromatin state transitions (e.g., a transition from an inactive enhancer to an active enhancer), we input H3K4me1, H3K27Ac, and H3K4me3 data sets collected at D14 into ChromHMM[99], which uses a multivariate hidden Markov model to define chromatin states along the genome. Using ChromHMM, we generated genome-wide state segmentations separately for control cells and BG^{exp} at D14 and then determined all regions across the genome for which state assignments differed between C and BG^{exp} (Figure 2.4e). While most of the genome was in the same state in both C and BG^{exp} samples by D14, we detected significant state transitions at genomic regions covering more than 44M total base pairs. All of the most prevalent state transitions occurred at enhancer regions and represented 3 main types of transitions: a complete loss of enhancer activity ('enhancer_inactive' to 'none'), a gain of new enhancer regions de novo ('none' to 'enhancer_inactive'), or a switch from an inactive to an active enhancer ('enhancer_inactive' to 'enhancer_active') (Figure 2.4e). These data show that beta glucan stimulation initially induced widespread changes in levels of H3K4me1 and H3K27Ac, all of which were initially lost with kinetics of passive decay. However, we found that differences

between histone PTMs emerged at late timepoints at which point BG^{exp} cells continued to harbor differences exclusively in H3K4me1. These continued differences were accompanied by both the loss of enhancer activity and the uncovering of novel latent enhancers[19].

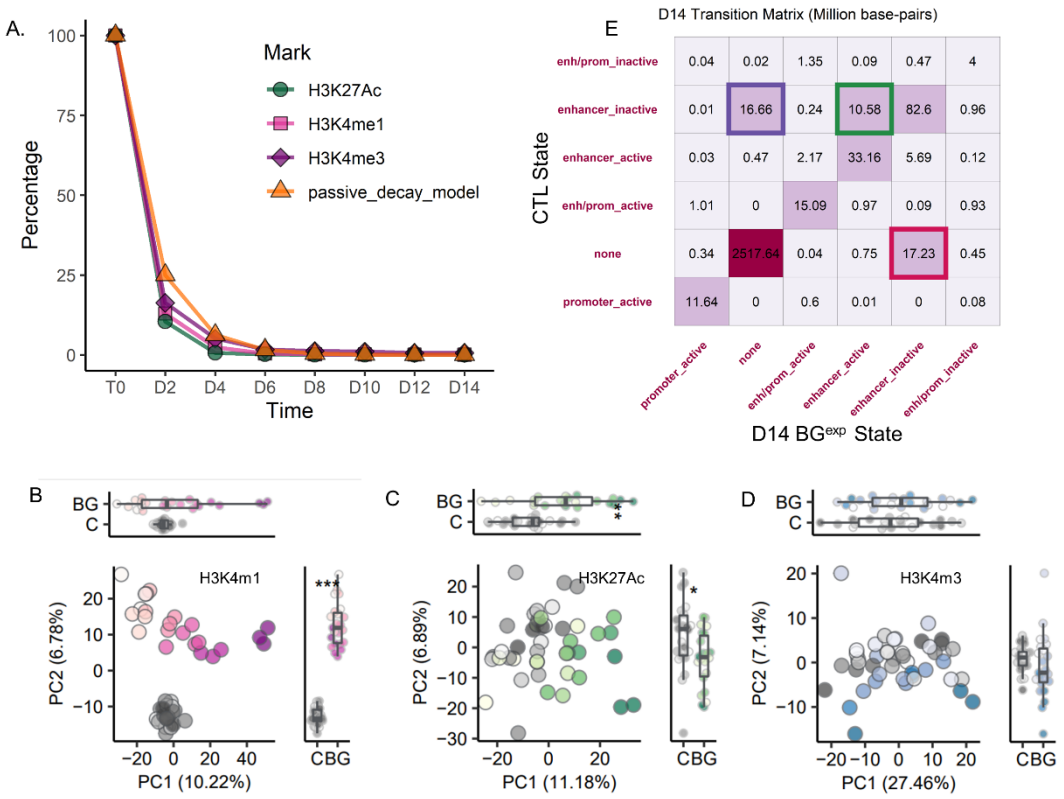

Figure 2.4. BG^{exp} dividing iBMDMs have long-lasting H3K4me1 signatures of previous beta glucan exposure

a. Line plot showing the percentage of significant peaks at T0 (FDR<0.05) remaining significant at each subsequent timepoint. Orange line models expected percentages assuming 50% loss with each cell division. **b.** principal component analysis (PCA) of H3K4me1 levels in BG^{exp} cells collected between D2 and D14 and time-paired controls across (across 3 experimental replicates). PCs were calculated using scaled and centered log2(counts per million) reads for

Figure 2.4, continued

each sample across all H3K4me1 peaks. **c** and **d**, represent same analysis as in b but for H3K27Ac and H3K4me3. **e.** Transition matrix displaying the proportion of base pairs in each transition state at D14. ChromHMM was used to segment the genome into 6 states using H3K4me1, H3K4me3, and H3K27Ac ChIP-seq data. Separate segmentations were performed with control and BG D14 ChIP-seq profiles. State assignments between control and BG D14 segmentation outputs were compared across the entire genome using 200 base pairs as the minimal unit.

Residual H3K4me1 differences are accompanied by changes in gene expression

At first glance, H3K4me1 profiles suggested that BG^{exp} cells may intrinsically retain select stimulus induced H3K4me1 despite dividing multiple times after loss of NFκB activity after only 3 cell divisions (Figure 2.2d). To better understand whether long-lasting differences in H3K4me1 were occurring in the presence or absence of any concomitant differential gene expression at all, we used our time series RNA-sequencing data to quantify the number of significantly differentially expressed (DE) genes between control and BG^{exp} macrophages (FDR<0.1) at each time point (Figure 2.5a,b). Not surprisingly, we detected the greatest number of DE genes at T0 and the first BG-washout timepoint, D2. This number dropped precipitously between divisions 2 and 4 from 1578 to 198 DE genes. However, we continued to detect similar numbers of DE genes (>100) at almost all subsequent time points (Figure 2.5b). Indeed, principal component analysis revealed that BG^{exp} samples remained clearly distinguishable from all control cells regardless of the collection time point (Figure 2.5c). Unsupervised hierarchical clustering confirmed both the long-lasting separability from control samples, and time point specific clustering of BG^{exp} samples (Figure 2.5e) based on gene expression, suggesting that exposure to beta glucan may permanently alter the baseline transcriptional state of macrophages.

Interestingly, we found that pc1*pc2 values in the gene expression PCA (Figure 2.5c) correlated strongly with pc1*pc2 values of the H3K4me1 PCA (Figure 2.4b), suggesting a relationship between patterns of continued differential gene expression and H3K4me1 levels (Figure 2.5g; correlation coef = -0.626). To better understand the nature of DE genes, we performed a gene set enrichment analysis (GSEA) to look for enriched hallmark pathways at each time point (D2-D14) as well as at multiple time points during BG stimulation (using data shown in Figure 2.2b). As expected, immune-related pathways such as 'Inflammatory response' and 'Tnfa signaling via nfkb', were the predominantly enriched pathways in direct response to BG stimulation (2-7 hours of stimulation) (Figure 2.5f). Surprisingly, we noticed that most of these beta-glucan responsive pathways were no longer enriched early on during the washout period (D2). However, multiple pathways, such as the complement pathway and the interferon alpha/gamma response pathways, reappeared at later timepoints (D10-D14) during the washout period. Other pathways such as Coagulation became enriched only post-BG while never being enriched in direct response to beta glucan. Moreover, when looking at the overlap of DE genes at the first and last timepoints, we found that more than 50% of DE genes at D14 were not significant at D2 (Figure 2.5d), suggesting that continued differential gene expression following beta glucan washout may not be indicative of retained immune activation *per se*, but rather indicative of new waves of expression emerging within the washout period.

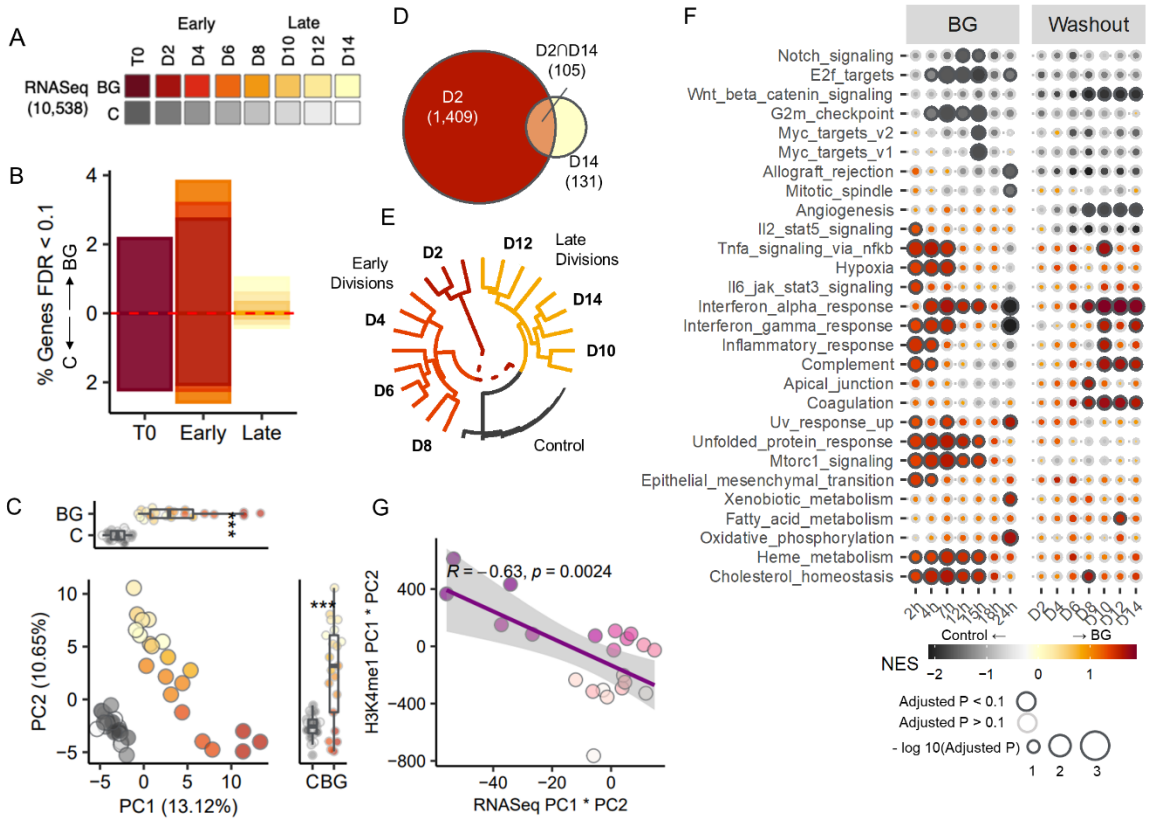

Figure 2.5. Transcriptional changes at late divisions correlate with H3K4me1 changes.

a. RNA-seq sample overview. **b.** bar plots showing the proportion of differentially expressed genes (DEG) at T0, early (D2, D4 and D6) and later (D8 ,D10, D12, D14). Red dotted line marks the direction of effect. **c.** principal component analysis (PCA) of gene expression for BG^{exp} cells collected between D2 and D14 and time-paired controls across (across 3 experimental replicates). PCs were calculated using scaled and centered log2(counts per million) reads for each sample across. **d.** Euler diagram displaying the overlapped DEG between the earliest division (D2) and latest one (D14). **e.** Unsupervised hierarchical clustering analysis of samples based on transformed log2(counts per million) values of the set of differentially expressed genes (FDR<0.05) detected across all time points. Log (CPM) values of control replicates within each time point were averaged and displayed as a single point. **f.** Gene set enrichment analysis of Hallmark pathways performed using the fgsea R package. A separate gene set enrichment analysis was performed at each time point. Genes were ordered in descending order according to -log10(pvalue)*log2FC values (NES is Normalized Enrichment Score). Colored boxes indicate significant (pvalue<0.05) enrichment and color shading corresponds to NES score as indicated in the legend. Only pathways with significant enrichments for at least one time point are displayed.

Figure 2.5, continued

g. Correlation analysis of PC1*PC2 values in PCA using gene expression data (x-axis) compared to PC1*PC2 values in PCA using H3K4me1 peaks values.

Post-BG changes in the H3K4me1 landscape are accompanied by signatures of altered transcription factor activity

Our gene expression and ChIP data collectively suggested that BG^{exp} cells co-regulate their enhancer and gene expression landscapes. We hypothesized that although we could detect differences in H3K4me1 as far out as 14 cell divisions after BG washout, this may be reliant on differential transcription factor activity, rather than self-sustained, as suggested by the fact that overall trends of differential gene expression matched H3K4me1 patterns. Since we had generally observed that individual genes and H3K4me1 peaks followed variable dynamics after BG-washout, we asked whether there existed a clear relationship between genes and peaks following similar trajectories. We reasoned that if this were the case, it would likely indicate that epigenetic remodeling is intertwined with remodeling of the gene expression program. First, for H3K4me1 peaks we focused on those peaks which were significantly different (FDR<0.05) at either only the beginning (DE_{D2}, N=4080), the end (DE_{D14}, N=124), or both timepoints of the time course ($DE_{D2, D14}$, N=51), representing non-persistent, induced, or retained peaks, respectively (Figure 2.6a). For each group, we plotted the average absolute log2FC of peaks in the group at each timepoint (Figure 2.6b). The log2FC values of DE_{D2} peaks declined steadily throughout the washout period while values for induced peaks rose beginning as early as D4. Log2FC values for the small set of retained peaks dipped between D2 and D4 but remained at

around 0.3 for most of the washout period before rising back to levels similar to D2 at the end. Next, we performed the same analysis on differentially expressed genes and found the same 3 patterns with markedly similar trajectories (Figure 2.6c,d; N=1081 DE_{D2}, N=100 DE_{D14}, N=70 $DE_{D2, D14}$). On a global scale and in line with our overall quantifications of DE genes and peaks in Figures 2.3c and 2.5b, we found that more than 85% of both genes and peaks were in the "non-persistent" group (Figure 2.6a,c) further confirming the overall finding that the vast majority of H3K4me1 and gene expression signatures initially induced by beta glucan are rapidly lost.

Among the small percentage of retained and induced features, we explored to what extent peaks and genes following the same trajectories (i.e., retained peaks and retained genes; induced peaks and induced genes) were related to each other. As a representative example, we focused on genes and H3K4me1 peaks both following an "induced" pattern. We performed a gene ontology analysis on genes for which gene expression dynamics belonged to the "induced group" (Figure 2.6e) and compared it to enriched transcription factor motifs within induced H3K4me1 peaks - all induced peaks in Figure 2.6a,b and a broader set of peaks whose patterns followed a significant upward linear trajectory (Figure 2.6f). Among induced genes we found a strong enrichment of viral response pathways (Figure 2.6e) and individually, we observed clear upward gene expression trajectories of transcription factors including Irf7 and Stat1 and downstream targets of these TFs such as Oas1a (Figure 2.6g-i). We found that 42% of the genes within the induced trajectory were direct targets of either the transcription factor Irf7 or Stat1 – these TFs themselves in the induced trajectory, suggesting that that the induced wave of novel gene expression as a whole may largely be driven by these TFs. Among enriched motifs at increasing H3K4me1 peaks we found strong, significant enrichment of IRF/ISRE motifs (Figure 2.6j). These motifs were enriched selectively among increasing H3K4me1 peaks, indicating the

specificity of IRF motifs enrichments to the induced trajectory and demonstrating consistency with the finding that viral response pathways were only enriched among induced genes. When we assigned each peak to its closest gene, we found that induced or non-persistent peaks enriched significantly for induced or non-persistent genes, respectively (P_{inc} = 8.55x10-5, P_{dec} = 1.07x10-42; Figure 2.6k). Constant peaks enriched for constant genes with a higher but near-significant p-value (P_{cons} = 0.195) likely due to the low power arising from the small total number of genes and peaks within this trajectory.

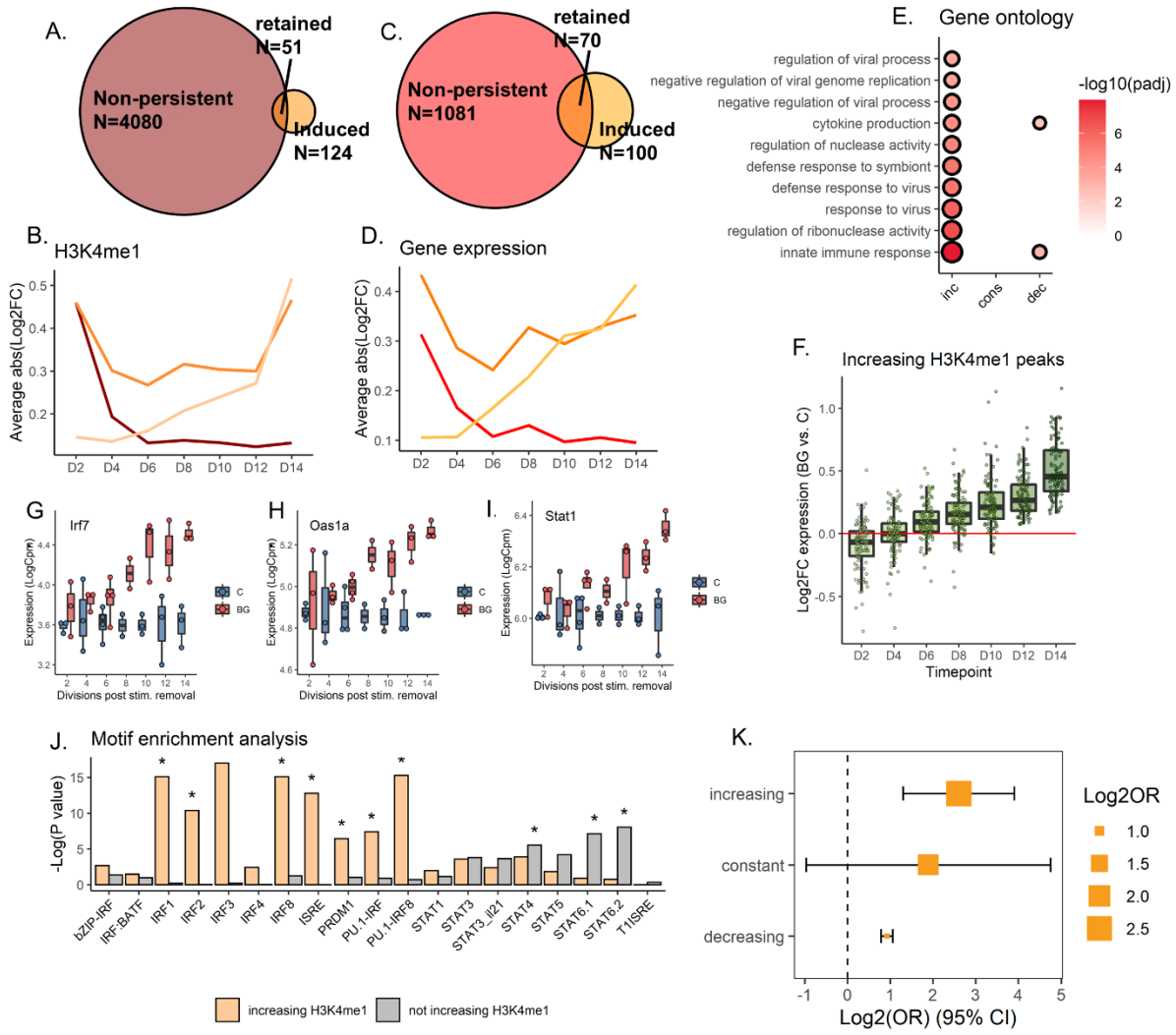

Figure 2.6. Post-BG changes in the H3K4me1 landscape are accompanied by signatures of altered transcription factor activity

a. Venn diagram of showing how H3K4me1 peaks were grouped into 3 trajectories based on differential levels at D2 (non-persistent), D14 (induced), or both timepoints (retained). **b.** Trajectories of significant (fdr<0.05) H3K4me1 peaks categorized as non-persistent, induced, or retained. Y-axis is the mean absolute value of log2FC differential abundance across all peaks in the group. **c.** Venn diagram of showing how genes were grouped into 3 trajectories based on differential expression at D2 (non-persistent), D14 (induced), or both timepoints (retained). **d.** Trajectories of significant (fdr<0.05) genes categorized as non-persistent, induced, or retained. Y-axis is the mean absolute value of log2FC differential expression across all genes in the group. **e.** Gene ontology analysis (Biological process pathways) on genes within each trajectory. Shown pathways are the top 10 most significantly enriched (fdr<0.05) pathways in the "increasing" group. Circle size and color are scaled to –log10(fdr). **f.** H3K4me1 peaks in the increasing

Figure 2.6, continued

trajectory. **g-i**. Examples of induced genes (green trajectory), Irf7, Oas1a, and Stat1. **j.** Summary of homer motif enrichment results for interferon motifs performed on increasing and non-increasing H3K4me1 peaks (* denotes fdr < 0.05). **k.** Log2 odds ratio (x axis) enrichment of genes in the 'increasing', 'constant', or 'decreasing' trajectories, among genes annotated to peaks in the 'increasing', 'constant', or 'decreasing' trajectories respectively

Collectively our data demonstrates that H3K4me1 peaks that are significant at D14 represent peaks that followed either a retained or induced trajectory following beta glucan washout and are significantly associated with transcription factor networks following the same expression patterns. We find little evidence that stimulus-induced histone modifications can be retained as an "epigenetic scar' independently of matched changes in transcription factor activities and gene expression in the context of an isolated and dividing cell.

Changes in H3K4me1 and gene expression are associated with evolving functional responses

It has been hypothesized that the presence of altered levels of H3K4me1 at enhancer regions may enable transcription factors to rapidly upregulate gene expression upon a secondary immune challenge, a phenomenon that is often referred to as "priming"[5,6]. A prototypical example of this was demonstrated in human monocytes stimulated with beta glucan, which subsequently secreted higher levels of proinflammatory cytokines following an LPS or Pam3CSK secondary challenge 6 days later and had differences in histone PTMs at that time point[8,9,20]. Since the small numbers of retained or induced H3K4me1 signatures we detected in this data varied across time, we hypothesized that primed genes may be quite different depending on the timepoint interrogated. To assess responses of BG^{exp} cells to secondary challenges, we stimulated a subset of control and

BG^{exp} cells with Pam3CSK4 for 5 hours at each time point and compared their ability to rapidly upregulate genes in response to the secondary stimulation. We detected small numbers of primed genes whose log2FC response to secondary stimulation in BG^{exp} cells was significantly different compared to controls. (Figure 2.7c, c-e, local false sign rate (lfsr) < 0.01). The largest total number of primed genes was detected at D2, also the time point with the greatest number of differences in all histone modifications profiled. We found various levels of overlap between primed genes at different time points (Figure 2.7b). The largest percentage of primed genes were specific to the first time point, D2, while the second most common pattern was shared priming across all time points. We also observed genes which were primed only at later timepoints (D12, D14, or both) and genes whose direction of priming reversed, consistent with the idea that the set of primed genes identified at one time point could differ significantly from the primed genes at another time point arguably because of the lack of overlap between significant H3K4me1 peaks at D2 compared to D14. These data suggest that as for H3K4me1, most priming is centered at early timepoints, and most of these primed genes are not maintained. Rather the repertoire of primed genes changes over time much in line with the overall dynamic nature of gene expression and H3K4me1 patterns we detected. Interestingly, GSEA on primed genes revealed that the strongest priming occurred at interferon alpha and gamma pathways early on – also the pathways whose baseline expression was induced most strongly at later time points (Figure 2.7f). We speculate that the reinduction of interferon and inflammatory pathways may be related to their priming at early timepoints. Given that priming at D2 and D4 still intersects with a state of continued NFκB activity and differential expression of other genes, it is not unfathomable that cells could self-engage their own primed pathways post-washout, thus driving new differential gene expression waves as seen here. This would explain why the pathways most strongly

responsive to BG stimulation, the most strongly primed pathways, and the pathways re-engaged post-washout are so strongly overlapping (Figure 2.5f, 2.7f).

Finally, we tested whether timepoint specific differential gene expression, epigenetics, and/or priming, may impact functional responses of iBMDMs to bacterial infection, mimicking a more biologically relevant situation. We infected control and BG^{exp} iBMDMs with *S. Typhimurium* at D2, D6, D10, D14, and D20 after BG washout and quantified the ability of iBMDMs to control bacterial growth over a 4-hour period using a CFU assay (Figure 2.7g). In these experiments BG^{exp} cells gained an approximately 2-fold increase in the ability to control bacterial growth, however this "trained" phenotype emerged only at later timepoints (D10-D20; Figure 2.7h) and followed kinetics highly correlated with the induced wave of baseline interferon expression we previously observed (Figure 2.7i). Thus, functional protection was much more closely correlated to the newly emerging gene expression program, H3K4me1 marks, and primed genes found at later timepoints, compared to the retained signatures of activation still present at D2. These data point to a role for newly induced changes post-washout, rather than the immediate post-BG priming, in shaping functional outcomes in response to secondary pathogen encounters.

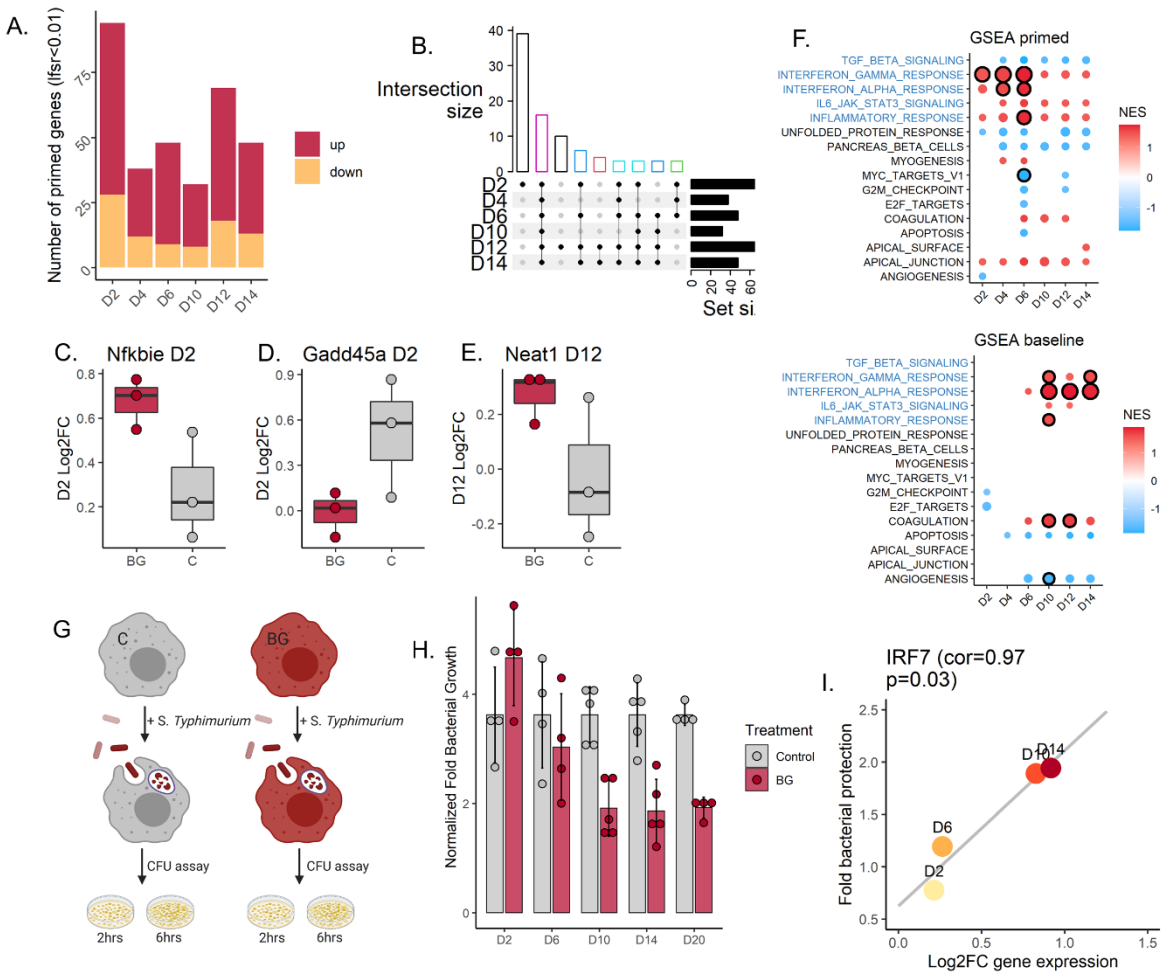

Figure 2.7. Changes in H3K4me1 and gene expression are associated with evolving functional responses

a. Bar plot quantification of the number of primed genes per time point (lfsr < 0.01). **b.** Upset plot showing overlap of primed genes across timepoints (showing only combinations with n>=3 genes). Bar colored by number of shared time points (black=1, red=2, green=3, dark blue=4, teal = 5, magenta=6). **c-e.** Individual examples of primed genes (Log2FC gene expression after 5 hours Pam stimulation of C or BG cells). **f.** Significantly enriched (circle with black outline, FDR<0.1) hallmark immune pathways among primed genes (top panel) and differentially expressed genes at baseline (bottom panel). Pathways with a circle but no outline have a p-value < 0.05. Circle size is scaled to –log10(padj). Color is scaled to normalized enrichment score (NES). **g.** Depiction of S. Typhimurium infection performed at D2, D6, D10, D14, and D20. At each time point 300,000 C and BGexp cells were infected at a MOI of 10 for 2 and 6 hours. Quantification of bacterial load during the 4-hour period was performed by CFU assay. **h.**

Figure 2.7, continued

Relative bacterial growth in BG^{exp} relative to timepoint and experiment paired control. The y-axis represents fold change bacterial growth over a 4-hour period in BGexp samples normalized to experiment-paired control. **i.** Scatterplot showing log2FC expression of IRF7 at D2, D6, D10, and D14 on the x-axis and bacterial protection on the y-axis, with line of best fit (p=0.03, pearson correlation = 0.967).

DISCUSSION

In this study, we performed time course ChIP- and RNA-sequencing to track the gene expression and epigenetic dynamics of dividing macrophages following a 24-hour stimulation with beta glucan. Our results demonstrate that a completely isolated, dividing macrophage population can harbor long-term changes in levels of H3K4me1 for at least 14 cell divisions following beta glucan stimulation. Notably, these differences made up a very small proportion of total stimulus responsive H3K4me1 peaks, of which over 85% were lost with kinetics matching a model of passive decay, suggesting that cells likely do not actively copy histone modifications during DNA replication. Nonetheless, the residual H3K4me1 "memory" was present even after all NF-kB activity had returned to baseline levels, demonstrating that H3K4me1 signatures can be detected independently of acute cell activation. In contrast, the other two histone modifications profiled – H3K4me3 and H3K27Ac returned to baseline levels after a few cell divisions post-BG stimulation, demonstrating a potentially unique role for H3K4me1 as a marker of previous beta glucan exposure. These data are consistent with previous findings that trained immunity can induce epigenetic reprogramming of long-lived cell populations, particularly at enhancer regions - which are the primary sites of H3K4me1[5,9,10,19–21]. Our experimental system is unique in that it demonstrates the presence of these signatures in a macrophage population undergoing rapid cell

divisions in culture, whereby cells are isolated from interactions with other cells and continually passaged and washed ensuring more complete removal of the stimulus. The idea that H3K4me1 signatures can remain in our experimental system seems to contradict current evidence that dividing cells lack the ability to actively copy PTMs during DNA replication[93–95]. Our results demonstrate how differences in the H3K4me1 landscape can be present after many cell divisions even though histone PTMs cannot be actively copied. We found that BG^{exp} cells evolve new sets of DE genes and associated H3K4me1 peaks during the washout phase, which largely contributed to why we were able to detect significant H3K4me1 peaks, even at D14. Primarily, we detected a novel set of interferon genes and transcription factors that were induced during the washout period and that were accompanied by increased H3K4me1 levels at peaks containing IRF/ISRE motifs. We found overall patterns of differential gene expression and H3K4me1 to be strongly correlated, supporting our hypothesis that differential transcription factor activity may be driving, and required for, the continued presence of H3K4me1 differences.

These results suggest that "memory" may be a misleading term, at least in our system. Rather, the gene expression and epigenetic program continues to be shaped, even after the stimulus has been removed, and the cells are no longer activated. Performing secondary stimulations of BG^{exp} cells with Pam3CSK, we assessed whether and how BGexp cells were primed at multiple timepoints and found that, although some genes were primed at multiple timepoints, many were primed in a time point specific manner. Priming was strongest at D2 – a time point during which the cells still had active NFκB activity and thousands of differential H3K4me1 peaks. We note that the specific pathways which were the most strongly primed at these early timepoints, including the interferon alpha and gamma pathways, were among the pathways most strongly upregulated during BG stimulation, confirming the basic principle that BG stimulation has a

short-term priming effect specifically on those pathways engaged during the initial immune response. Despite some degree of sharing across timepoints, the primed genes at D2 were largely unique, suggesting that initial priming likely decays in the same way as H3K4me1 marks decay. Functional protection, remarkably, followed the opposite pattern, and was much more closely correlated to the newly emerging gene expression program, H3K4me1 marks, and primed genes found at later timepoints. These data point to a role for newly induced changes post-washout, rather than the immediate post-BG priming, in shaping functional outcomes in response to secondary pathogen encounters. The fact that pathways that were initially responsive to BG stimulation, pathways exhibiting priming early on, and induced pathways at late timepoints were largely shared, suggest that newly emerged phenotypes are closely linked and shaped by the initial response to BG, and therefore do constitute a form of "memory", although perhaps different in nature from the typical "memory" of retained histone PTMs seen in non-dividing cells. We speculate that the primed state of BG^{exp} cells, which is strongest between D2 and D6, may interface with transcription factors that continued to remain activated at these time points, potentially causing cells to re-active their own primed pathways. Specifically, this model suggests that interferon pathways are temporarily primed between D2 and D6 due to their strong upregulation in response to BG stimulation. Although this primed state itself is not long-lasting (likely to the dilution of histone PTMs across cell divisions), primed interferon pathways are re-engaged by transcription factors that continued to remain activated at D2 and D4 such as NFκB, leading to baseline increases in the expression of genes in these pathways. Although more work is required to further explore this hypothesis, our data demonstrate that stimulus-induced inflammatory programs can remain dynamic for an extended period, even after the initiating stimulus is cleared, which may have important functional implications.

Fundamentally, our model suggests that a fine-tuned series of events with specific timing is essential to induce the phenotypes observed here. Cells must be primed at a timepoint during which the cells still remain marginally active, allowing transcription factors to engage the primed pathways on their own. This model suggests that cells which become deactivated too quickly, or cells which don't become primed during the time frame of this continued activation will likely not induce new gene expression programs or H3K4me1 signatures. The fact that there may be several specific requirements needed to induce long-lasting epigenetic signatures is ultimately consistent with the fact that trained immunity has proven to be incredibly sensitive to experimental design. We propose that the central mechanistic findings in this study – the coupling of transcription factor activity with novel or retained H3K4me1 signatures, passive histone decay, timepoint specific priming, and the ability of priming to drive stimulus-experienced cells down a unique path – may represent general requirements for trained immunity within dividing cell populations devoid of any help from other cell populations of a pro-inflammatory microenvironment. More time course studies in primary cells or transplantation studies in animal models utilizing genomic methods that require only extremely small cell numbers will be necessary to investigate this question further.

MATERIALS AND METHODS

Cells and reagents

All mice were housed in the University of Chicago animal facility in accordance with the policies of, and approved by, the University of Chicago Institutional Animal Care and Use Committee. Mice were housed under SPF conditions.

Bone marrow derived macrophages from the bone marrows of C57BL/6 mice were generated as described in Kaufmann et al.[45] but using DMEM (Gibco) supplemented with 10% heat-inactivated FBS (Gibco) and penicillin-streptomycin-glutamine (Gibco) in place of RPMI and 25 ng/mL M-CSF (Prospec protein specialists) instead of 30% L929-conditioned media.

LM1 cells from were a gift from Dr. Martin Olivier (McGill University) and were cultured in DMEM supplemented with 10% heat-inactivated FBS and pen-strep glutamine unless indicated otherwise in T75 flasks (Falcon). NF-κB GFP reporter LM1 cells were cultured in DMEM supplemented with 10% heat-inactivated FBS and pen-strep glutamine and 2.5 ug/mL puromycin (Sigma) unless indicated otherwise in T75 flasks. DMEM supplemented with 10% heat inactivated FBS and L-glutamine (Gibco) was used during all steps of Salmonella Typhimurium infections.

Beta glucan was provided by D. Williams (East Tennessee State University) and was added to a final concentration of 30 ug/mL in all experiments.

Monoclonal Reporter generation

LM1 cells were transduced with lentiviral particles (Qiagen) containing multiple repeats of the NFκB promoter driving the expression of GFP, using SureENTRY (Qiagen) at an MOI of 10. After 24 hours cells were washed with PBS and given new media. To isolate stably transduced clones, we added 2.5 ug/mL puromycin (Sigma) to the cell culture medium for 24 hours and allowed the cells to reach confluency (~5 days). Cells were harvested and single cells were sorted into 96-well plates. After 2 weeks, reporter activity of individual clones was tested by stimulating clones with beta glucan followed by flow cytometry analysis of GFP fluorescence. One highly responsive clone was chosen and used for all subsequent experiments.

Flow cytometry analysis of reporter activity

To assess NFκB activity at various timepoints following beta glucan washout, cells were harvested by removing media, washing with PBS, and incubating adherent cells in 10 mL Accutase (Sigma) for 5 mins at room temp. Detached cells were pooled, washed 1x with PBS, and resuspended in 1% BSA in PBS. Cells were analyzed immediately on a Fortessa flow cytometer (BD Biosciences) using the FITC channel to measure GFP fluorescence. Data was analyzed using FlowJo version 10.

EdU labelling

EdU incorporation time course assays were performed using Click-iT™ Plus EdU flow cytometry assay kit (Invitrogen) per manufacturer's instructions. Briefly, naïve LM1 cells or LM1 cells stimulated with 30 ug/mL beta glucan 24 hours prior were incubated with 10 μM EdU for 0, 6, 9, 12, 13, 14, or 15 hours as indicated. After the incubation, cells were harvested, washed once with 1% BSA in PBS and fixed in 100 μL of Click-iT™ fixative for 15 minutes at room temperature. Following fixation, cells were washed with 1% BSA in PBS and incubated for

15 minutes in 100 µL of 1X Click-iT™ permeabilization and wash reagent. Cells were stained with Click-iT™ Plus reaction cocktail containing fluorescent Pacific Blue picolyl azide, incubated for 30 minutes at room temp in the dark, washed with 1X Click-iT™ permeabilization and wash reagent, and analyzed on a Fortessa (BD biosciences).

Cell sorting and post-sort culture of time course experiments

Cell collection for sorting: Prior to sorting, reporter LM1 macrophages either unstimulated or stimulated with beta glucan for 24 hours, were harvested by removing all media, washing with PBS, and incubating adherent cells in Accutase at room temp for 5-10 mins. Detached cells were pooled, washed with PBS, and resuspended in PBS to a concentration of approximately 1 million cells per milliliter. Cells were filtered through a 40 µM sterile filter. Live-dead stain was performed for all samples by addition of 1:10 propidium iodide solution 10 minutes prior to sorting. Cells were sorted on a FACSAria II or FACSAria IIIu instrument. All PI^-, singlet control cells were sorted and all PI^-, singlet, GFP^+ BG-stimulated cells were sorted.

Post-sort cell culture: Sorted control and BG-experienced cells were washed with PBS, resuspended in media, counted, and seeded into T75 flasks (0.6M cells per flask for harvest at D4 and beyond, or ~2M for D2 collection). Cells seeded into flasks for later timepoints (D4 and beyond) were passaged every 48 hours.

Cell harvest: At each harvest timepoint, cells were harvested using Accutase to detach cells, followed by a wash with 1x PBS. Cells were aliquoted into tubes for downstream ChIP-seq or RNA-seq processing. Whenever possible, we aimed to collect 3M cells for ChIP-seq and at least 0.5M cells for RNA-seq. A subset of cells was plated into four wells of a 6-well plate (600,000 cells per well) and stimulated with Pam3CSK4 to assess priming (conditions: control unstim.,

BG unstim., control + pam, BG + pam). After 5 hours of stimulation, plated cells were harvested in Accutase for RNA extraction and sequencing.

RNA extractions

Adherent LM1 macrophages were harvested with Accutase, washed with PBS, and lysed with 1mL of Qiazol. RNA extractions were performed using the miRNeasy mini or miRNeasy micro kits (Qiagen). RNA quality was evaluated with the 2100 Bioanalyzer (Agilent Technologies). For primary BMDMs, cells were harvested by incubating cells in 10-15 mL CellStripper dissociation reagent (Corning) for 20 mins at 37C, and cells were further detached using Cell Lifter (Corning). BMDMs were washed with PBS, lysed, and processed as described above for LM1 cells.

Bulk RNA sequencing (RNA-Seq)

RNA library preparations were carried out on 100-500 ng of RNA with RIN 1.2 to 9.8 using the Illumina TruSeq Stranded Total RNA Sample preparation kit, according to the manufacturer's instructions. The libraries were size-selected using Ampure XP Beads (Beckman Coulter) and quantified using the KAPA Library Quantification kit - Universal (KAPA Biosystems). Sequencing of the RNA-Seq libraries was performed on the Illumina NovaSeq 6000 system using 100-bp single-end sequencing.

RNA-seq data processing

Adaptor sequences and low-quality score bases were first trimmed using Trimmomatic with parameters -phred33 SE ILLUMINACLIP:TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36[100]. The resulting reads were aligned to the mm10 mouse

reference genome using STAR[101]. Read counts are obtained using featureCounts[102] with default parameters.

Differential gene expression analyses

For each RNA-sequencing data set, gene expression levels across all samples were first normalized using the calcNormFactors function implemented in the edgeR R package (version 3.34.0) which utilizes the TMM algorithm (weighted trimmed mean of M-values) to compute normalization factors. Then, the voom function implemented in the limma package (version 3.38.3) was used to log-transform the data and to calculate precision-weights. A weighted fit using the voom-calculated weights was performed with the lmFit function from limma.

Effect of BG stimulation on gene expression: To investigate the impact of beta-glucan stimulation on LM1 cells at various timepoints of stimulation (2, 4, 7, 12, 15, and 18 hours), normalized, log-transformed gene expression levels were fit to the linear model Expression ~ 1 + time + stimulus:time, which corrects for natural changes across time independent of beta-glucan stimulation and therefore captures the independent effect of beta-glucan on gene expression at each time point (BG at 2 hrs., BG at 4 hrs., …, BG at 18 hours). LM1 cells stimulated for 24 hours, and their respective controls were collected in an additional set of extra experiments. Normalized gene expression for these samples was fit to the linear model Expression ~ 1 + experiment + stimulus to correct for "batch" effects between replicate samples while giving an independent estimation of the impact of 24-hours BG stimulation on gene expression.

Effect of previous BG exposure on expression: For samples collected post-BG washout in the time course experiment, the design Experiment ~1 + experiment + time + primary:time +

((Pam:time):primary) was used to correct for batch effects between replicate experiments and natural variations across time while giving an estimate of the lasting impact of previous BG exposure at each time point (effect of BG on expression at D2, D4, etc.).

Hierarchical clustering: Normalized, log-transformed expression values were also used for downstream hierarchical clustering analysis. Expression values were corrected to remove the effect of experimental variation and natural across-time changes by subtracting the "experiment" and "time" effects from all samples. Corrected, normalized expression values were scaled and centered such that each gene had a mean = 0 and sd = 1 across all samples. Scaled expression was used to compute a distance matrix between samples using the dist function in R with parameters: method = "Euclidean" and diag=TRUE. Samples were clustered based on calculated Euclidean distances using the function hclust and viewed as a dendrogram using the fviz_dend function implemented in the R package factoextra (version 1.0.7) with parameters: k=7, type=phylogenetic.

Effect of pervious BG exposure on secondary responses to Pam3CSK4: To investigate whether BG-experienced LM1 cells were primed at various time points post-washout, we used the design Expression ~1 + experiment + primary + Pam:primary to estimate priming (the effect of Pam secondary stimulation on naïve/control cells compared to the effect of Pam on BG-experienced cells) independently for each time point using the makeContrasts and contrasts.fit functions implemented in limma. Mashr: To increase our power to detect primed genes shared or unique to each time point we applied Multivariate Adaptive Shrinkage in R (mashr version 0.2.57) to outputs from contrasts.fit. Effect sizes were obtained from contrasts.fit in limma and the standard error of the effect size for each gene was given by multiplying the square root of the posterior

variance by the standard deviation for effect size. Effect sizes and standard errors for each time point (n = 6; D2, D4, D6, D10, D12, D14) were arranged into n x m matrices, n being the number of genes and m being the number of timepoints. We then fit the mash model using canonical and data driven covariance matrices and then stringently defined primed genes as those with an lfsr < 0.01.

Gene ontology analysis

We used the enrichGO function implemented in the clusterProfiler R package (version 4.0.5) to identify gene ontology terms enriched among DE genes (FDR<0.05, abs(logFC)>1) in LM1 cells at 7 hours of beta glucan stimulation. We used the parameters: OrgDb = "org.Mm.eg.db", ont = "BP", pAdjustMethod = "BH", minGSSize = 10, maxGSSize = 500. The same parameters were used to determine gene ontology enrichments among genes grouped into specific trajectories (increasing, constant, and decreasing).

GSEA

Gene set enrichment analyses (GSEA) were performed using the fgsea R package (version 1.18.0) with parameters: maxSize = 500, nperm=100000. To investigate biological pathway enrichments among DE genes during BG stimulation, or at various timepoints after BG washout, genes at each time point were ordered by the rank statistic: -log10(pvalue)*logFC and compared with the Hallmark gene sets from the MSigDB collections.

Enrichments for primed genes after Pam secondary stimulation were performed using the same parameters. For each time point, genes were ordered by the mashr calculated values: -log10(lfsr)*PosteriorMean and compared with the Hallmark gene sets from the MSigDB collections.

ChIP-sequencing

Samples collected at various timepoints of the post-BG time course were crosslinked with 1% v/v formaldehyde for 10 mins shaking at 50 rpm at room temperature and then quenched for 5 minutes by addition of 1.25M Glycine to a final concentration of 125 mM. Formaldehyde fixed samples were sonicated for 12 mins using a Biorupter (Diagenode) with 105 peak power and 200 cycles/burst. ChIP-DNA was prepared by incubating chromatin with antibody for at least 10 hours at 4C on a thermomixer set to 300 rpm, followed by Dynabeads (protein A) incubation for 2 hours and washes with Low Salt wash buffer, High Salt wash buffer, LiCl wash buffer, and TE buffer. The number of input cells used for each ChIP ranged from 0.5 million to 1 million. The following antibodies were used: H3K4me1 (company: CST; ref: 5326S), H3K4me3 (company: CST; ref: 9751S), H3K27Ac (company: Abcam; ref: ab4729). ChIP and input libraries were sequenced on the Illumina NovaSeq 6000 system at the University of Chicago Genomics Facility using 100-bp paired-end sequencing. A subset of samples was sequenced a second time to obtain greater sequencing depth across all samples and histone modifications. These samples were sequenced 100-bp single-end on the Illumina NovaSeq 6000 system at the University of Chicago Genomics Facility.

ChIP-seq data processing

ChIP-seq reads were mapped to the mouse reference genome (GRCm38/mm10) using Bowtie2 with default parameters. Mapped reads were sorted and filtered using the samtools function *view* (with parameter -q 30 to retain only high-quality mapped reads), and sort with default parameters. PCR duplicates were removed using Picard MarkDuplicates program with parameter "REMOVE_DUPLICATES=True". Peaks were subsequently called independently for each

sample using the callpeak function from the MACS2 software suite with parameters -q 0.05 –keep-dup all. For each histone modification, a merged peak file combining called peaks across all samples was generated using the bedops *merge* function with default parameters. Read counts overlapping the merged peak set were obtained using featureCounts[102].

ChIP-seq analysis/differential testing

Initial processing of ChIP-seq count data was performed as described for RNA-seq data: The number of reads overlapping each peak across all samples for a given histone modification were normalized using the calcNormFactors function implemented in the edgeR R package (version 3.34.0). Then, the voom function implemented in the limma package (version 3.38.3) was used to log-transform the data and to calculate precision-weights. A weighted fit using the voom-calculated weights was performed with the lmFit function from limma.

Effect of previous BG exposure on histone modification levels: For samples collected post-BG washout in the time course experiment, the design Experiment ~1 + experiment + time + primary:time was used to correct for batch effects between replicate experiments and natural variations across time while giving an estimate of the lasting impact of previous BG exposure at each time point (effect of BG on histone PTM levels at D2, D4, etc.).

ChromHMM

We used ChromHMM[99] to segment the genome into gene regulatory states for control (n=3 replicates) and BG-experienced cells (n=3 replicates) at D14 using aligned histone PTM bam files for H3K4me1, H3K4me3, and H3K27Ac. Briefly, we first used BinarizeBam with default parameters (binsize=200) to learn chromatin states in 200 bp intervals for each histone PTM. Then we used the LearnModel function (default parameters except n=7 emission states) to

partition the genome into states based on the combinatorial presence or absence of each histone PTM. We manually assigned each emission state to a gene regulatory state as follows: H3K4me1lo/H3K4me3hi/H3K27Achi – "promoter_active", H3K4me1hi/H3K4me3hi/H3K27Achi – "enh/prom_active", H3K4me1hi/H3K4me3lo/H3K27Achi – "enhancer_active", H3K4me1lo/H3K4me3lo/H3K27Aclo – "none", H3K4me1hi/H3K4me3lo/H3K27Aclo – "enhancer_inactive", H3K4me1hi/H3K4me3hi/H3K27Aclo – "enh/prom_inactive".

Motif enrichment analysis

Motif enrichment analyses were performed using the Homer function *findMotifsGenome* with parameters -size 1000 -mask. For motif enrichment analyses performed on subsets of H3K4me1 peaks (i.e., "increasing" or "not-increasing"), we used the entire set of called H3K4me1 peaks (n=122,962) as background.

Gene networks of post-washout increasing genes

We used the interferome database (interferome.org) containing manually curated sets of type I, II, and III interferon genes to predict which genes characterized as "increasing" during the post-BG washout period were predicted to be regulated by IRF7 or STAT1. We defined each increasing gene as belonging to the same network as IRF7, STAT1, or both based on the presence or absence of the respective TF motifs in the gene's promoter region. Network visualization was performed using cytoscape (v3.8.2) in which all circles represent genes in the increasing trajectory, and lines connecting genes to IRF7, STAT1, or both indicate a likely binding of the respective transcription factors to the gene's promoter region.

Relationship between significant H3K4me1 peaks and genes of different trajectories

To investigate the genomic association between H3K4me1 peaks and genes following the same trajectories, we assigned H3K4me1 peaks following “increasing”, “constant”, or “decreasing” trajectories to their nearest genes using the Homer function annotatePeaks with default parameters. Background peaks were defined as all H3K4me1 peaks and foreground peaks were defined as peaks in the “increasing”, “constant”, or “decreasing” trajectories. Gene-peak enrichment was performed using a two-sided Fisher’s exact test (fisher.test in base R) using the frequencies of genes within a given trajectory being assigned to a H3K4me1 peak within the same trajectory.

Salmonella Typhimurium infections

Salmonella culture: Salmonella Typhimurium was grown in TSB media (15g TSB in 500 mL DI water). For infections, the concentration of bacteria was determined by taking the OD600 and bacteria were diluted in media to a concentration of 1.5M bacteria per milliliter of solution.

Bacterial infections: Control and beta-glucan experienced LM1 macrophages were infected with Salmonella Typhimurium at various time points following beta glucan washout to compare functional ability. Approximately 12 hours prior to the desired time of infection, Salmonella Typhimurium cultures were established by incubating a small aliquot of frozen bacterial stock in TSB media at 37C and rotations set to 250 rpm. On the day of infections, cells were seeded at a density of 300,000 cells per well in 6-well plates containing 2 mL media 2 hours prior to infection. Cells were infected for 40 minutes with Salmonella Typhimurium at an MOI=10 by removing supernatant from attached cells, washing with pre-warmed Phosphate-buffered saline, and adding 2 mL per well of bacterial solution (see ‘Salmonella culture’ above). After 40 minutes, bacterial solution was removed, and cells were washed with pre-warmed PBS. Media

containing concentration gentamycin was added to cells and cells were incubated at 37C for 1 hour to kill any remaining extracellular bacteria. Following 1-hour incubation (designated Time0/T0), cells were washed and new media containing gentamycin was added to cells.

Colony forming unit assay

To quantify the kinetics of intracellular bacterial growth in vitro, over a 4-hour period, cells were harvested at 2- and 6-hours post-infection (T2 and T6 respectively) by removing all supernatant and washing cells with PBS. Cells were lysed to release intracellular bacteria by adding 1 mL sterile water supplemented with 1% TritonX-100 and pipetting up and down vigorously 10 times. Serial dilutions made in PBS were plated on TSB plates. Plates were incubated at 37C and counted the next day. Bacterial growth was quantified as the fold-change difference in CFUs between the 6-hour and 2-hour plates.

CONCLUSION

In this work, we tackled key questions in the field of trained immunity spanning the spectrum from basic questions about epigenetic heritability to analyses of human patient samples. In chapter II we performed a single cell-level analysis of rare human bone marrow samples collected before and 90 days after BCG vaccination. By integrating a multitude of different computational approaches, combined with in vitro assays and flow cytometry, we showed that BCG vaccination can have a lasting impact on both the gene expression and chromatin accessibility landscape of HSPCs in a way that is linked to the functional properties of immune cells in the periphery. Through our analyses of this data set, we identified key transcription factors that may be critical to this process, validated findings that have been described in mouse models, and proposed a new model for how lasting inflammation within stem cells can shape the downstream epigenetic landscape of progenitors. Our findings lay the groundwork for additional experimental approaches to further test and refine this model, and hopefully underscore the importance of considering how live vaccines, and inflammation more generally, impact the immune system in its entirety. On a more basic level, our analyses continued to raise longstanding questions related to the inherent heritability of trained immunity-associated epigenetic changes across cell divisions in long-lived cell types. In chapter III we pivoted towards a reductionist, more basic model system to further investigate this question. Through dense time course ChIP-seq and RNA-seq we found a tight coupling between transcription factor activity and histone PTMs at all timepoints, and also discovered that new gene expression programs and coupled epigenetic changes emerge after primary stimulus removal and are associated with new changes in cell functionality that only emerge at later timepoints. Ultimately

our results point to the general idea that gene expression and epigenetic rewiring remain coupled in dividing cell types and that the complete loss of TF-driven activation of gene expression would ultimately be expected to lead to the rapid loss of any epigenetic changes. Collectively the findings presented in this thesis suggest that “memory” is more similar to a low-level activation program that is maintained over long periods of time, than to a situation whereby epigenetic marks are deposited and then maintained in a vacuum, which would effectively imply an uncoupling of epigenetics from gene transcription. This work contributes to our mechanistic understanding of innate immune memory in dividing and long-lived cell types, which is ultimately critical to making informed decisions about whether and how innate memory can be used to improve the care of patients.

FUTURE DIRECTIONS

Future directions should focus on a few key questions. First, what is the true duration of BCG-induced inflammation and what are the key players involved in maintaining this state? Assessing human bone marrow samples at time points beyond 90 days would undoubtedly help answer this question. However, as a logistically easier first step, it would also be informative to re-assess some of the pre-existing bone marrow samples by using single cell RNA-sequencing to capture a broader set of cells, including the mature and stromal cells of the bone marrow, which would open the door for investigations into how HSCs interact and potentially remain activated by adaptive immune cells, osteoclasts, and bone marrow resident macrophages, given that we find little evidence for direct pathogen sensing by HSCs themselves. The other key question raised by our work in chapter II revolves around the role of transcription factors in establishing chromatin accessibility within progenitor cells. Our current data implicates a handful of transcription factors as potentially being important in the establishment of DR peaks, however experimental validation will still be critical. It would be interesting to make a few single knockout mice, for example, knockouts for KLF5, KLF6, or EGR1 within LSK cells, to narrow down the current list of TFs to an experimentally validated list of critical TFs which are necessary to establish BCG-induced epigenetic memories.

Chapter III raises critical questions related to the *correlation versus causation* relationship between epigenetic reprogramming and functional reprogramming. It still remains unclear whether the long-lasting changes in H3K4me1 observed in our data are required for the functional reprograming of macrophages or are just a correlate of variation in activity of specific TFs (in our specific setting primarily IRFs). Future studies should move beyond large genome-

wide analyses and towards more precise epigenetic editing (CRISPR dCas9 – based approached for example) to dissect whether particular epigenetic changes alone are necessary and sufficient for the induction of innate immune memory or, if instead, the cornerstone of innate memory is the long-term rewiring of baseline TF activity in response to an initial challenge.

REFERENCES

1. Iwasaki, A. & Medzhitov, R. Control of adaptive immunity by the innate immune system. *Nat Immunol* **16**, 343–353 (2015).
2. Badovinac, V. P., Haring, J. S. & Harty, J. T. Initial T Cell Receptor Transgenic Cell Precursor Frequency Dictates Critical Aspects of the CD8+ T Cell Response to Infection. *Immunity* **26**, 827–841 (2007).
3. Adams, N. M., Grassmann, S. & Sun, J. C. Clonal expansion of innate and adaptive lymphocytes. *Nat Rev Immunol* **20**, 694–707 (2020).
4. Kaech, S. M. & Cui, W. Transcriptional control of effector and memory CD8+ T cell differentiation. *Nat Rev Immunol* **12**, 749–761 (2012).
5. Netea, M. G. *et al.* Trained immunity: A program of innate immune memory in health and disease. *Science* **352**, aaf1098 (2016).
6. Netea, M. G. *et al.* Defining trained immunity and its role in health and disease. *Nat Rev Immunol* **20**, 375–388 (2020).
7. Netea, M. G. & van der Meer, J. W. M. Trained Immunity: An Ancient Way of Remembering. *Cell Host & Microbe* **21**, 297–300 (2017).
8. Quintin, J. *et al.* Candida albicans Infection Affords Protection against Reinfection via Functional Reprogramming of Monocytes. *Cell Host & Microbe* **12**, 223–232 (2012).
9. Cheng, S.-C. *et al.* mTOR- and HIF-1α–mediated aerobic glycolysis as metabolic basis for trained immunity. *Science* **345**, 1250684 (2014).
10. Yoshida, K. *et al.* The transcription factor ATF7 mediates lipopolysaccharide-induced epigenetic changes in macrophages involved in innate immunological memory. *Nat Immunol* **16**, 1034–1043 (2015).

11. Kleinnijenhuis, J. *et al.* Bacille Calmette-Guérin induces NOD2-dependent nonspecific protection from reinfection via epigenetic reprogramming of monocytes. *Proceedings of the National Academy of Sciences* **109**, 17537–17542 (2012).

12. Yao, Y. *et al.* Induction of Autonomous Memory Alveolar Macrophages Requires T Cell Help and Is Critical to Trained Immunity. *Cell* **175**, 1634-1650.e17 (2018).

13. Aegerter, H. *et al.* Influenza-induced monocyte-derived alveolar macrophages confer prolonged antibacterial protection. *Nat Immunol* **21**, 145–157 (2020).

14. Roquilly, A. *et al.* Alveolar macrophages are epigenetically altered after inflammation, leading to long-term lung immunoparalysis. *Nat Immunol* **21**, 636–648 (2020).

15. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat Rev Genet* **17**, 487–500 (2016).

16. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**, 207–220 (2019).

17. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacol* **38**, 23–38 (2013).

18. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311–318 (2007).

19. Ostuni, R. *et al.* Latent Enhancers Activated by Stimulation in Differentiated Cells. *Cell* **152**, 157–171 (2013).

20. Saeed, S. *et al.* Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science* **345**, 1251086 (2014).

21. Novakovic, B. *et al.* β-Glucan Reverses the Epigenetic State of LPS-Induced Immunological Tolerance. *Cell* **167**, 1354-1368.e14 (2016).

22. Pacis, A. *et al.* Bacterial infection remodels the DNA methylation landscape of human dendritic cells. *Genome Res.* **25**, 1801–1811 (2015).

23. Ramos-Rodríguez, M. *et al.* The impact of proinflammatory cytokines on the β-cell regulatory landscape provides insights into the genetics of type 1 diabetes. *Nat Genet* **51**, 1588–1595 (2019).

24. Pacis, A. *et al.* Gene activation precedes DNA demethylation in response to infection in human dendritic cells. *Proceedings of the National Academy of Sciences* **116**, 6938–6943 (2019).

25. Foster, S. L., Hargreaves, D. C. & Medzhitov, R. Gene-specific control of inflammation by TLR-induced chromatin modifications. *Nature* **447**, 972–978 (2007).

26. Larsen, S. B. *et al.* Establishment, maintenance, and recall of inflammatory memory. *Cell Stem Cell* **28**, 1758-1774.e8 (2021).

27. Kamada, R. *et al.* Interferon stimulation creates chromatin marks and establishes transcriptional memory. *Proceedings of the National Academy of Sciences* **115**, E9162–E9171 (2018).

28. Cerwenka, A. & Lanier, L. L. Natural killer cell memory in infection, inflammation and cancer. *Nat Rev Immunol* **16**, 112–123 (2016).

29. Arts, R. J. W. *et al.* BCG Vaccination Protects against Experimental Viral Infection in Humans through the Induction of Cytokines Associated with Trained Immunity. *Cell Host & Microbe* **23**, 89-100.e5 (2018).

30. Arts, R. J. W. *et al.* Glutaminolysis and Fumarate Accumulation Integrate Immunometabolic and Epigenetic Programs in Trained Immunity. *Cell Metabolism* **24**, 807–819 (2016).

31. Liu, P.-S. *et al.* α-ketoglutarate orchestrates macrophage activation through metabolic and epigenetic reprogramming. *Nat Immunol* **18**, 985–994 (2017).

32. Kaikkonen, M. U. *et al.* Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription. *Molecular Cell* **51**, 310–325 (2013).

33. Palozola, K. C., Lerner, J. & Zaret, K. S. A changing paradigm of transcriptional memory propagation through mitosis. *Nat Rev Mol Cell Biol* **20**, 55–64 (2019).

34. Audergon, P. N. C. B. *et al.* Restricted epigenetic inheritance of H3K9 methylation. *Science* **348**, 132–135 (2015).

35. Ragunathan, K., Jih, G. & Moazed, D. Epigenetic inheritance uncoupled from sequence-specific recruitment. *Science* **348**, 1258699 (2015).

36. Wang, X. & Moazed, D. DNA sequence-dependent epigenetic inheritance of gene silencing and histone H3K9 methylation. *Science* **356**, 88–91 (2017).

37. Garly, M.-L. *et al.* BCG scar and positive tuberculin reaction associated with reduced child mortality in West Africa: A non-specific beneficial effect of BCG? *Vaccine* **21**, 2782–2790 (2003).

38. Rieckmann, A. *et al.* Vaccinations against smallpox and tuberculosis are associated with better long-term survival: a Danish case-cohort study 1971–2010. *International Journal of Epidemiology* **46**, 695–705 (2017).

39. Giamarellos-Bourboulis, E. J. *et al.* Activate: Randomized Clinical Trial of BCG Vaccination against Infection in the Elderly. *Cell* **183**, 315-323.e9 (2020).

40. Freyne, B. *et al.* Neonatal BCG Vaccination Influences Cytokine Responses to Toll-like Receptor Ligands and Heterologous Antigens. *The Journal of Infectious Diseases* **217**, 1798–1808 (2018).

41. Jensen, K. J. *et al.* Heterologous Immunological Effects of Early BCG Vaccination in Low-Birth-Weight Infants in Guinea-Bissau: A Randomized-controlled Trial. *The Journal of Infectious Diseases* **211**, 956–967 (2015).

42. Liggett, L. A. & Sankaran, V. G. Unraveling Hematopoiesis through the Lens of Genomics. *Cell* **182**, 1384–1400 (2020).

43. Doulatov, S., Notta, F., Laurenti, E. & Dick, J. E. Hematopoiesis: A Human Perspective. *Cell Stem Cell* **10**, 120–136 (2012).

44. Mercer, E. M., Lin, Y. C. & Murre, C. Factors and networks that underpin early hematopoiesis. *Seminars in Immunology* **23**, 317–325 (2011).

45. Kaufmann, E. *et al.* BCG Educates Hematopoietic Stem Cells to Generate Protective Innate Immunity against Tuberculosis. *Cell* **172**, 176-190.e19 (2018).

46. Redelman-Sidi, G., Glickman, M. S. & Bochner, B. H. The mechanism of action of BCG therapy for bladder cancer—a current perspective. *Nat Rev Urol* **11**, 153–162 (2014).

47. Naik, S. *et al.* Inflammatory memory sensitizes skin epithelial stem cells to tissue damage. *Nature* **550**, 475–480 (2017).

48. Gonzales, K. A. U. *et al.* Stem cells expand potency and alter tissue fitness by accumulating diverse epigenetic memories. *Science* **374**, eabh2444 (2021).

49. Ordovas-Montanes, J. *et al.* Allergic inflammatory memory in human respiratory epithelial progenitor cells. *Nature* **560**, 649–654 (2018).

50. Cirovic, B. *et al.* BCG Vaccination in Humans Elicits Trained Immunity via the Hematopoietic Progenitor Compartment. *Cell Host & Microbe* **28**, 322-334.e5 (2020).

51. Majeti, R., Park, C. Y. & Weissman, I. L. Identification of a Hierarchy of Multipotent Hematopoietic Progenitors in Human Cord Blood. *Cell Stem Cell* **1**, 635–645 (2007).

52. Manz, M. G. & Boettcher, S. Emergency granulopoiesis. *Nat Rev Immunol* **14**, 302–314 (2014).

53. Kwak, H.-J. *et al.* Myeloid Cell-Derived Reactive Oxygen Species Externally Regulate the Proliferation of Myeloid Progenitors in Emergency Granulopoiesis. *Immunity* **42**, 159–171 (2015).

54. Zhu, H. *et al.* Reactive Oxygen Species–Producing Myeloid Cells Act as a Bone Marrow Niche for Sterile Inflammation–Induced Reactive Granulopoiesis. *The Journal of Immunology* **198**, 2854–2864 (2017).

55. Kimura, A. *et al.* The transcription factors STAT5A/B regulate GM-CSF–mediated granulopoiesis. *Blood* **114**, 4721–4728 (2009).

56. Brook, B. *et al.* BCG vaccination–induced emergency granulopoiesis provides rapid protection from neonatal sepsis. *Science Translational Medicine* **12**, eaax4517 (2020).

57. Williamson, S. L., Gadd, E., Pillay, T. & Toldi, G. Non-specific effects of BCG vaccination on neutrophil and lymphocyte counts of healthy neonates from a developed country. *Vaccine* **39**, 1887–1891 (2021).

58. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nat Methods* **19**, 159–170 (2022).

59. Hirai, H. *et al.* C/EBPβ is required for 'emergency' granulopoiesis. *Nat Immunol* **7**, 732–739 (2006).

60. Sasaki, Y. *et al.* IL-6 Generated from Human Hematopoietic Stem and Progenitor Cells through TLR4 Signaling Promotes Emergency Granulopoiesis by Regulating Transcription Factor Expression. *The Journal of Immunology* **207**, 1078–1086 (2021).

61. Zhao, J. L. *et al.* Conversion of Danger Signals into Cytokine Signals by Hematopoietic Stem and Progenitor Cells for Regulation of Stress-Induced Hematopoiesis. *Cell Stem Cell* **14**, 445–459 (2014).

62. Grassi, L. *et al.* Dynamics of Transcription Regulation in Human Bone Marrow Myeloid Differentiation to Mature Blood Neutrophils. *Cell Reports* **24**, 2784–2794 (2018).

63. Ai, Z. & Udalova, I. A. Transcriptional regulation of neutrophil differentiation and function during inflammation. *Journal of Leukocyte Biology* **107**, 419–430 (2020).

64. Geest, C. R. *et al.* p38 MAP Kinase Inhibits Neutrophil Development Through Phosphorylation of C/EBPα on Serine 21. *Stem Cells* **27**, 2271–2282 (2009).

65. Croker, B. A. *et al.* SOCS3 Is a Critical Physiological Negative Regulator of G-CSF Signaling and Emergency Granulopoiesis. *Immunity* **20**, 153–165 (2004).

66. Notta, F. *et al.* Isolation of Single Human Hematopoietic Stem Cells Capable of Long-Term Multilineage Engraftment. *Science* **333**, 218–221 (2011).

67. Mitroulis, I. *et al.* Modulation of Myelopoiesis Progenitors Is an Integral Component of Trained Immunity. *Cell* **172**, 147-161.e12 (2018).

68. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome Biol* **20**, 45 (2019).

69. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083–1086 (2017).

70. Van de Sande, B. *et al.* A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat Protoc* **15**, 2247–2276 (2020).

71. de Laval, B. *et al.* C/EBPβ-Dependent Epigenetic Memory Induces Trained Immunity in Hematopoietic Stem Cells. *Cell Stem Cell* **26**, 657-674.e8 (2020).

72. Kalafati, L. *et al.* Innate Immune Training of Granulopoiesis Promotes Anti-tumor Activity. *Cell* **183**, 771-785.e12 (2020).

73. King, K. Y. & Goodell, M. A. Inflammatory modulation of HSCs: viewing the HSC as a foundation for the immune response. *Nat Rev Immunol* **11**, 685–692 (2011).

74. Nagai, Y. *et al.* Toll-like Receptors on Hematopoietic Progenitor Cells Stimulate Innate Immune System Replenishment. *Immunity* **24**, 801–812 (2006).

75. Seeley, J. J. & Ghosh, S. Molecular mechanisms of innate memory and tolerance to LPS. *Journal of Leukocyte Biology* **101**, 107–119 (2017).

76. Medzhitov, R., Schneider, D. S. & Soares, M. P. Disease Tolerance as a Defense Strategy. *Science* **335**, 936–941 (2012).

77. Sato, N. *et al.* MyD88 But Not TRIF Is Essential for Osteoclastogenesis Induced by Lipopolysaccharide, Diacyl Lipopeptide, and IL-1α. *Journal of Experimental Medicine* **200**, 601–611 (2004).

78. Hayashi, S.-I. *et al.* Distinct Osteoclast Precursors in the Bone Marrow and Extramedullary Organs Characterized by Responsiveness to Toll-Like Receptor Ligands and TNF-α. *The Journal of Immunology* **171**, 5130–5139 (2003).

79. Winkler, I. G. *et al.* Bone marrow macrophages maintain hematopoietic stem cell (HSC) niches and their depletion mobilizes HSCs. *Blood* **116**, 4815–4828 (2010).

80. Li, X. *et al.* Maladaptive innate immune training of myelopoiesis links inflammatory comorbidities. *Cell* **185**, 1709-1727.e18 (2022).

81. Zsebo, K. *et al.* Vascular endothelial cells and granulopoiesis: interleukin-1 stimulates release of G-CSF and GM-CSF. *Blood* **71**, 99–103 (1988).

82. Hestdal, K. *et al.* In vivo effect of interleukin-1 alpha on hematopoiesis: role of colony-stimulating factor receptor modulation. *Blood* **80**, 2486–2494 (1992).

83. O'Connor, L., Gilmour, J. & Bonifer, C. The Role of the Ubiquitously Expressed Transcription Factor Sp1 in Tissue-specific Transcriptional Regulation and in Disease. *Yale J Biol Med* **89**, 513–525 (2016).

84. Patel, S., Xi, Z. F., Seo, E. Y., McGaughey, D. & Segre, J. A. Klf4 and corticosteroids activate an overlapping set of transcriptional targets to accelerate in utero epidermal barrier acquisition. *Proceedings of the National Academy of Sciences* **103**, 18668–18673 (2006).

85. Liao, X. *et al.* Krüppel-like factor 4 regulates macrophage polarization. *J Clin Invest* **121**, 2736–2749 (2011).

86. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* **36**, 89–94 (2018).

87. Cortal, A., Martignetti, L., Six, E. & Rausell, A. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. *Nat Biotechnol* **39**, 1095–1102 (2021).

88. Sandelin, A. & Wasserman, W. W. Constrained Binding Site Diversity within Families of Transcription Factors Enhances Pattern Discovery Bioinformatics. *Journal of Molecular Biology* **338**, 207–215 (2004).

89. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1–22 (2010).

90. Ghisletti, S. *et al.* Identification and Characterization of Enhancers Controlling the Inflammatory Gene Expression Program in Macrophages. *Immunity* **32**, 317–328 (2010).

91. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences* **107**, 21931–21936 (2010).

92. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).

93. Aygün, O., Mehta, S. & Grewal, S. I. S. HDAC-mediated suppression of histone turnover promotes epigenetic stability of heterochromatin. *Nat Struct Mol Biol* **20**, 547–554 (2013).

94. Escobar, T. M. *et al.* Active and Repressed Chromatin Domains Exhibit Distinct Nucleosome Segregation during DNA Replication. *Cell* **179**, 953-963.e11 (2019).

95. Escobar, T. M., Loyola, A. & Reinberg, D. Parental nucleosome segregation and the inheritance of cellular identity. *Nat Rev Genet* **22**, 379–392 (2021).

96. Prentice, S. *et al.* BCG-induced non-specific effects on heterologous infectious disease in Ugandan neonates: an investigator-blind randomised controlled trial. *The Lancet Infectious Diseases* **21**, 993–1003 (2021).

97. Forget, G. *et al.* Role of host phosphotyrosine phosphatase SHP-1 in the development of murine leishmaniasis. *European Journal of Immunology* **31**, 3185–3196 (2001).

98. Lavin, Y. *et al.* Tissue-Resident Macrophage Enhancer Landscapes Are Shaped by the Local Microenvironment. *Cell* **159**, 1312–1326 (2014).

99. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215–216 (2012).

100. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

101. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

102. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).