# Scalable methods for genomic analysis of *in vitro* models of mammalian embryogenesis

Samuel Gabriel Regalado

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Jay Shendure, Chair
Cole Trapnell
Heather Mefford

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

## Abstract

Scalable methods for genomic analysis of *in vitro* models of mammalian embryogenesis

Samuel Gabriel Regalado

Chair of the Supervisory Committee:

Jay Shendure

Department of Genome Sciences

Mammalian development, from the single-celled zygote to a multicellular individual, is an incredible dynamic journey that is marked by many milestones measurable across many scales. In fact, by the end of the first two weeks of human embryogenesis, most precursors of major tissues and organs required for life are already present. This developmental milestone is known as gastrulation. Here the embryo or gastrula undergoes invagination, creating the blastopore and three major layers. For example, the outermost layer, known as ectoderm, gives rise to the nervous system and skin; the middle layer, known as the mesoderm, gives rise to the musculoskeletal system and the heart; the innermost layer, known as the endoderm, gives rise to internal organs such as the lungs and liver. Collectively, these developmental cell types constitute the three major germ layers. Thus it is at the stage of gastrulation that cells of the embryo are specified toward distinct fates, leaving behind their relatively indistinct transcriptional states as pluripotent precursors. The advent of large consortia efforts, like the Human Genome Project or ENCODE, has ushered in new sequencing technologies, e.g. single-cell molecular phenotyping modalities

like scRNA-seq, that are capable of uncovering the individual components or features of the genome that support the blueprint for multicellularity. For example, we now know that the genome can be partitioned into two categories: the coding genome and the non-coding genome. The coding genome is largely made up of genes, including cell-type specifying transcription factors (TFs). While approximately ~22,000 protein-coding genes have been decoded and cataloged, of which ~1600 or so are thought to be TFs, the overall coding proportion only makes up 1-2% of the mammalian genome. The other 98% of the genome is defined by the non-coding genome, where ~1 million non-coding regulatory elements, namely enhancers, are thought to reside. Despite our ever-growing knowledge of the genome, we know very little about the transcription factors or enhancers that are required for the myriad of cell types required for mammalian development. How this remarkable process unfolds at the molecular level is a timely question that remains elusive.

The focus of my PhD has been to elucidate how the process of early development works, particularly when cells undergo cell fate specification during gastrulation. More specifically, I have been intensely focused on understanding the dynamics of germ layer formation through 1) functional characterization of non-coding DNA elements or enhancers, 2) defining key developmental transcription factors, and 3) tracing histories of cell lineages as they are emerging within a multicellular system. To tackle these complex areas of investigation, I have developed scalable methods applied to multicellular *in vitro* embryoid model systems of early development. In the first chapter, I describe current strategies to understand early development and cell fate specification. In the second chapter, I describe efforts to perturb and record lineages using a novel platform for clonal organoid generation. In the third chapter, I describe a highly multiplexed method with single-cell resolution for measuring autonomous activity of non-coding regulatory

DNA in a multicellular context. Finally, in the last chapter, I conclude with my thoughts on the future of *in vitro* models alongside multi-modal measurements.

## TABLE OF CONTENTS

# List of Figures

## Chapter 2

### Main Figures



## Chapter 3

### Main Figures:

**Extended Data Figures:**



**Supplementary:**

## Acknowledgements

As I write this, I am overcome with a flood of emotions. I have fulfilled my childhood dream of becoming a scientist. The memories still come to me when I was a young inner-city kid fantasizing about the prospects of doing cutting-edge science with the world's best minds. If someone had told me that I would someday be doing exactly that, I would have brushed it off as complete nonsense. Very few from background make it this far, and for that, I am grateful to my amazing family, friends, and mentors who have supported me at every step. I know my ancestors would be proud.

It is hard to quantify just how important various aspects of my training have been for me. For example, a primary goal when I began graduate school was to achieve independence, by that I mean, to think creatively, to test hypotheses freely, and to add value in a variety of spaces with my own ideas. The only way I could have reached that point was by being granted the freedom to explore my scientific interests independently, even if it meant starting from scratch or embracing the unknown, regardless of how uncomfortable or slow the progress may have seemed initially. This certainly holds true computationally, as I have enthusiastically embraced the learning experience to acquire the skill set necessary for analyzing big data. I have been so incredibly lucky to have the license to freely explore my path to becoming an independent thinker and scientist. I owe my deepest gratitude to the Department of Genome Sciences, my committee members, the Trapnell and Shendure lab members, and especially to Cole Trapnell and Jay Shendure. I am indebted to this community of role models for shaping me into the scientist I am today.

Debbie Nickerson once told me, “the work we do is inherently hard, and it’s for this reason that we gravitate even closer to our science.” I suspect she said this after sensing some distress in me

for at the time I was a first year navigating the ins and outs of graduate school. Debbie's statement has become my mantra, exemplifying how seriously I take and internalize feedback, hoping it will positively shape me for the better. My committee members, Phil Abitua, Celeste Berg, Heather Mefford, Debbie Nickerson, Cole Trapnell, and Jay Shendure, have imparted so many valuable nuggets of knowledge that I cherish deeply and, more importantly, integrate into my daily practice as a scientist.

The scientist I am today, and the principal investigator I wish to become in the years to follow, is directly influenced by my mentors. Like many, I first learned about Jay Shendure through his science. I was struck not only by the cutting-edge nature of the work coming out of his lab, but by the obviously clear societal implications of the research. As an undergraduate wooed by innovations being concocted in the labs around the Berkeley campus, Jay's work presented a new frontier of science that I was obsessed with exploring. It was only a short time after discovering Jay's work that I was inspired to visit the UW campus and learn more about Genome Sciences, the home department of Jay and Cole. From that point on, I was determined to go back to Seattle for my graduate and medical school training. As soon as I got the acceptance, there was no doubt where I was going and where my first lab rotation would be. I canceled all my other second visits. In what felt like a blink of an eye, I was in Jay's lab. Over the ensuing years Jay has demonstrated what an amazing PI looks like. There are many attributes that I can easily highlight regarding Jay as an incredible PI, for example, his impeccable leadership style or his creative process for thinking through scientific ideas, but that would be stating the obvious. Jay sees you as a full person, and in the process, you see Jay as a full person: an incredible person who is a husband, a father, a brother, an uncle, a son, a friend. It is difficult to meaningfully contextualize Jay's impact on me

as a person and scientist. I am forever grateful to Jay for taking me under his wing and helping to foster my own leadership, mentorship, and scientific style.

My fascination with biology and computer science was sparked by Cole Trapnell. Our paths first crossed in Boston, where I was a summer student and he was a postdoc. Coincidentally, we had also been at Berkeley, possibly at the same time, though unaware of each other, before finally meeting again in Seattle. In Boston, sharing the same wet lab space, I often heard Cole enthusiastically discussing his computer code while conducting wet lab experiments. It was then that the seed of computational biology was planted. Having Cole as a co-mentor has been foundational for me in several ways. Firstly, Cole's profound knowledge of data science is evident in almost every interaction, whether in a one-on-one meeting, a lab meeting, or a casual encounter in the lab. Secondly, Cole helped me imagine myself as a computational biologist. He helped me overcome my own imposter syndrome that was limiting me from realizing my full potential and encouraged me to persevere no matter what. Cole's remarkable mentorship and leadership qualities, as well as his innovative approach to science, has transformed my own thinking as an independent scientist. Above all else, his commitment to excellent training is an example that all leaders in science should follow. His investment in me reflects his unwavering support for his lab's trainees, recognizing that each of us is on a unique journey to becoming a great scientist. I am deeply grateful to Cole for mentoring me on my journey in becoming a scientist.

My scientific journey would not be the same without my incredible collaborators. For me, the ideal collaboration is characterized by a continual exchange of ideas, thoughts, and suggestions, all converging toward a common goal: reaching the finish line as efficiently and productively as

possible. My closest collaborators include Junhong Choi, Silvia Domcke, Beth Martin, and Jean-Benoît Lalanne. I have loved every minute of our collaboration together, even as we traversed rough terrain. I am forever grateful for the science that we did together, and even more grateful to call you my lifelong friends.

To my undergraduate research mentor, Dr. David Weisblat, who inspired me to go to grad school. During my time in the Weisblat lab I worked on dissecting the Hox gene cluster in the understudied lophotrochozoan leech Helobdella. Some of my fondest memories occurred during this time period.

To my family, especially my Mom who catered to my interest in science at a young age. Thank you for investing all that you could to see that I had the best education possible. I also want to thank my sisters, Sylvia and Selena, whose love and support never wavered, even when the challenges of grad school made it difficult for me to always be there for you. To my loving grandparents who invested so much of their love and time in me from the day I was born in hopes that I would fulfill my career ambitions. To Nonna and Baba for always being there, even at a moment's notice. To Auntie Zina, your encouragement and understanding throughout this entire process has helped me in tremendous ways. I am incredibly lucky to have you all in my life.

My PhD work would not have been possible without the consistent and unconditional support of my life partner and wife, Megan. Your patience, cheerleading, and overall emotional and moral support have enabled me to embrace my passion for science and to become the scientist I am today. Despite many long hours in the lab, day or night, weekday or weekend, you always supported me

no matter what. Yet, we have been able to accomplish so much together, both professionally and personally. Despite my desire to pursue a lengthy training opportunity to become a physician scientist, my life has taken on so many exciting turns all thanks to you, my incredible life partner. We traveled, attended concerts, became homeowners, and started a family. Our daughter Lilly was born on February 4th, 2020 at the start of the pandemic. Although it was a turbulent time in world history, Lilly was the constant source of light that shines brighter and brighter with each passing day. Lilly and her soon-to-be baby sibling are the greatest sources of inspiration I could ever ask for. Lilly, you will be an amazing big sister. Thank you Megan for always believing in me and eliminating any semblance of 'giving up' from my vocabulary. You are the air I breathe.

## Dedication

To Megan, Lilly, and Theo:

You remind me every day that life truly is miraculous.

I love you deeply.

# 1. INTRODUCTION:

***How does the single-celled zygote achieve multicellularity?***

The transformation of a single-celled zygote into a multicellular organism is a highly orchestrated process. Throughout embryogenesis, many changes take place in the early embryo, including a key transition point known as gastrulation. At this stage in development, the three germ layers form, which give rise to all the tissue and organ types required for life[1,2]. At the molecular level, these changes are governed by specific transcription factors acting in coordination with enhancer elements to initiate cell-type specific gene expression programs[3]. At the cellular level, relationships between cells establish a hierarchy through which lineages of the various cell types emerge. Despite our understanding of the various genetic contributions influencing the single-celled zygote's transition to multicellularity, it is a mystery as to how this process unfolds. For example, we lack an understanding of the various transcription factors that are essential for germ layer formation. This problem is further exaggerated for developmental enhancers. Deconstructing embryogenesis down to the various genetic components has broad implications for both developmental biology and disease. To that end, a multipronged approach using a variety of tools is required.

***Build a lineage tree to understand development from a single-celled zygote***

Lineage tracing is a powerful technique that has been used in various contexts and model systems. Its strength lies in its ability to unveil the histories of diverse cells within an organism. The invention of the light microscope paved the way for real-time observation of development of various model organisms, including the transparent roundworm C. elegans[4,5]. Here Sir John

Sulston applied lineage tracing to developing C. elegans beginning first as a single cell. Mapping every cell division with the aid of the light microscope allowed Sulston to reconstruct the first lineage map of a multicellular organism. Following this incredible feat, chemical dyes or tracers, such as horseradish peroxidase, and fluorescent proteins, such as GFP, enabled greater specificity to identify distinct cell lineages throughout development[6]. However, such approaches are not comprehensive enough for more complex systems, such as mammalian embryogenesis. Unlike C. elegans, mammalian organisms are non-transparent, indeterministic, comprise complex 3D morphologies, and are typically many orders of magnitude greater in overall cell numbers compared to invertebrates. Microscopy is therefore insufficient to completely trace mammalian embryogenesis. In order to achieve high resolution lineage trees, more advanced methods have been developed to study mammalian development as it unfolds prospectively.

***CRISPR/Cas9 broadens the molecular toolbox***

The advent of CRISPR/Cas9 has ushered in a new suite of genetic tools that are relatively easy to implement yet have facilitated a deeper understanding of mammalian development[7]. In its earliest version, CRISPR/Cas9 comprised an endonuclease that binds to a specific location in the genome defined by a single guide RNA (gRNA) as part of a ribonucleoprotein (RNP) complex. Once locked in on its target sequence, the endonuclease produces a double-strand break that is further repaired by the cellular machinery, either by non-homologous end joining (NHEJ), which repairs by random insertions or deletions, or homology-directed repair, which requires the presence of a repair template. Still, other versions are now available that comprise a catalytically dead endonuclease that is tethered to a transcriptional repressor or activator, and more recently, Dr. David Lui's lab has developed a class of CRISPR-based genome engineering tools, also known as

'prime editors', that allow for precise insertions, deletions, or replacement[8]. Combined, CRISPR modalities have paved the way for next generation genetic manipulations with direct applications to scalable perturbations and lineage tracing.

***CRISPR-based molecular recorders pave way for new lineage tracing technique***

Lineage tracing using a CRISPR-based molecular recorder saw its debut with GESTALT[9]. Here for the first time CRISPR/Cas9 was used to install indels at synthetic target arrays randomly inserted into the genome of zebrafish embryos. Through the unique edit patterns installed at these target arrays, lineage reconstruction became possible with sequencing as a readout. While this proved to be a powerful system, several limitations were identified. Firstly, conventional CRISPR/Cas9 induces double-strand breaks which, in addition to cytotoxic effects, can inadvertently lead to contraction of target arrays and therefore reduce the recording medium. Secondly, edits incorporated are unordered, which complicates the tree building process as the order of the edits must be inferred. As an alternative approach, DNA Typewriter is an ideal lineage tracing tool that circumvents the limitations of conventional CRISPR/Cas9[10]. DNA Typewriter is based on the prime editing machinery – a nickase and a reverse transcriptase tethered to dCas9 – and comprises target arrays known as DNA Tapes, which comprise a tandem array of 5 or 6 protospacers all of which are truncated with the exception of the 5'-most site. This configuration enables the prime editing machinery to bind the first site where it installs a short barcode, usually a dimer, as well as a pam sequence to complete the next protospacer. Overall this process results in sequential editing, ensuring an edit pattern that is ordered. As a result, DNA Typewriter is devoid of double-strand breads (low cytotoxicity), lacks target array contraction (typically induced by double-strand breaks), and simplifies the tree building process.

***Pooled CRISPR/Cas9 genetic screens to probe coding and non-coding DNA***

CRISPR/Cas9 is notably powerful as a genetic perturbation tool[7]. In particular, coupling CRISPR/Cas9 perturbation screens to scRNA-seq, typically referred to as 'perturb-seq', has enabled scalable molecular phenotyping across many different conditions and model systems. Experiments of this kind generally begin with lentiviral transduction of a pool of cells, wherein a library of sgRNAs is introduced, integrated, and expressed in many thousands of single cells. After some period of growth, cells are subjected to a scRNA-seq platform (e.g. 10x Genomics) where global transcriptomics and perturbations are simultaneously read out via high-throughput sequencing. To date, this approach has been applied to a wide variety of contexts, including *in vivo* (e.g. mice) and *in vitro* model systems (e.g. organoids)[11,12]. Moreover, many modalities of CRISPR/Cas9 have been developed, such as knockout, activation, and repression and have been applied to both coding (e.g. genes) and non-coding (e.g. putative enhancers) DNA. Robust data analysis tools are readily accessible for deciphering key genes or gene regulatory networks that are essential for various biological processes. Pooled CRISPR/Cas9 screens have already impacted the fields of cancer, immunology, microbiology, and developmental biology.

***Massively Parallel Reporter Assays to characterize enhancer elements***

The non-coding genome is vast, and to date, over 1 million enhancers have been reported for the mammalian genome[13]. However, most enhancers lack characterization beyond descriptive assays. In addition to CRISPR-based screens targeting non-coding elements, massively parallel reporter assays (MPRAs) can characterize the autonomous activity of tens of thousands of enhancers in a single experiment. Conventional MPRA designs are based on a one-RNA system in which the

transcript measured comprises an enhancer element-linked barcode that is quantifiable with sequencing[14]. Therefore, the one-RNA system of an MPRA is tasked with handling both the detection (is the construct present?) and activity measurements (is the enhancer functional?). MPRAs generally work by introducing constructs into static cell lines which comprise putative enhancers that, if active, assist in driving the expression of the element-linked barcode transcript. Although MPRAs have been an incredible, scalable resource for decoding the non-coding landscape, including testing the functional consequences of disease-relevant mutations, conventional designs limit MPRAs to certain cellular contexts. For example, multicellular models of development, where cell type compositions are easily distinguishable with scRNA-seq, are excellent testing grounds for MPRAs. Yet, the one-RNA design cannot adequately handle the requirements for a single-cell readout in multicellular contexts. We therefore describe a new design for MPRAs that is compatible with scRNA-seq readouts in Chapter 3.

***In vitro models of development with new genetic tools will speed up our discovery rate***

Multicellular models that mimic the development of various mammalian cell lineages have opened up new opportunities to understand cell fate dynamics. These models provide an alternative to resource- or labor-intensive whole animal models, such as mice. Organoids and embryoids are *in vitro* systems that are relatively easy to culture through the aggregation of many hundreds or thousands of induced pluripotent or embryonic stem cells. Under various differentiation culture conditions, organoids as diverse as the brain or kidney can be generated. More recently, embryoids or "stembryos" have also been made possible, giving rise to embryoid bodies, gastruloids, and synthetic embryos, each of which recapitulates various aspects of *in vivo* mammalian embryogenesis[15]. The following work integrates the latest advances in CRISPR/Cas9 technology

and embryoid model generation to decipher the drivers of cell fate decisions, as well as the cellular histories that make mammalian multicellular development possible.

# 2. ESC-DERIVED CLONAL ORGANOIDS FACILITATE MONOPHYLETIC LINEAGE TRACING AND ORGANOID-TO-ORGANOID MEASUREMENTS WITH POOLED CRISPR SCREENS

**Section on TF CRISPR screens:**

Regalado SG, Domcke S, Qiu C, Martin B, Trapnell C, Shendure J.
*Submission planned for 2024*.

**Section on lineage tracing:**

Regalado SG, Choi J, Qiu C, Martin B, Trapnell C, Shendure J.
*Submission planned for 2024*.

This chapter combines two major projects. The first half is on pooled CRISPR screens targeting all known transcription factors while the second half is on lineage tracing with DNA Typewriter. A key innovation that unites both projects, at least from a technical standpoint, is the implementation of monoclonal embryoids as a model system. In this chapter, we show that monoclonal embryoids enable inter-individual measurements, reconstruction of monophyletic lineage trees, and a path forward for multimodal perturbation and lineage tracing within and across individuals.

The first project, beginning in 2018, aims to define the key TFs underlying mammalian germ layer formation. In close collaboration with Silvia Domcke, we combined stem cell biology, genome editing, and single-cell sequencing technology to conduct pooled CRISPR screens with single-cell readouts. With guidance on experimental design and analytical approach from Jay Shendure, Silvia and I carried out mosaic embryoid experiments and analyzed the data strategically and collaboratively. However, no multi-year project is without its trials and tribulations. Despite facing challenges such as transgene silencing and intractable stochastic bottlenecks—two major

confounders that severely limited our work yet are likely applicable to other mosaic organoid screens—we overcome these obstacles. The development of transposon-based piggyFlex constructs and monoclonal organoids were significant accomplishments that were born out of these challenges, and they stand as some of my proudest achievements accomplished in this project. While I am grateful that these advances ultimately helped us in impactful ways, I am even more grateful for how these advances shined a light on the importance that collaborators play at every step, particularly in a multi-year project such as this one. Beth Martin, with whom I worked closely to turn concept-into-construct with piggyFlex, is a shining example of someone whose role in this project was instrumental at various points. Finally, attempting to get this project to the publication finish line within a reasonable timeframe would not be possible without the close collaboration of Chengxiang (CX) Qiu. Very few, if at all, can easily overcome the fear factor of undertaking the lead role of informatician tasked with analyzing, and re-analyzing, mountains of data accumulated over the course of six years with as much grace as CX. Modifying his own obligations and commitments to instead work around the clock on the lineage and TF project is especially noteworthy.

The second project, beginning in 2022, aims to reconstruct the lineage of monoclonal gastruloids that begin their differentiation from a single precursor cell. In close collaboration with Junhong Choi, we developed 1) a new monoclonal gastruloid protocol, 2) a non-invasive measurement of lineage efficiency known as Debris-seq, and 3) an experimental pipeline to understand sources of inter-gastruloid variability. This collaboration began shortly after I conceived of the monoclonal organoid concept, where seeking out Choi's expertise as one of the main inventors of DNA Typewriter might set us on a direct path to a scalable platform for lineage reconstruction in

multicellular *in vitro* models. I was very happy when Choi agreed to collaborate with me on this project. Our immediacy to this collaboration set off a cascade of experiments that never stopped but rather continued around the clock. Even before a dataset had been fully analyzed, we were already on to the next experiment. With Jay's oversight, Choi and I designed experiments, collaborated on the execution of experiments, and analyzed the data. This joint effort between Choi and me was one of the most exhilarating, on-the-edge-of-my-seat experiences that I only hope will be replicated many times over in the future. That said, this project benefited from Beth Martin, who played a key role in helping us pivot to sci-RNA-seq combined with DNA Typewriter to look at inter-gastruloid variability. One result that really got us excited was the diverse cell type composition we observed in monoclonal gastruloids profiled *en masse*. Understanding the cell type complexity required expertise that only CX could provide. CX has therefore become an essential team member in this project, whose efforts will be instrumental, similarly for the TF project, in getting this work to the publication finish line.

Since the inception of these projects, it has been my true honor to work collaboratively with Choi and Silvia, two amazing scientists who have taught me so much while enthusiastically embracing my intellectual ideas and contributions. Moreover, without Beth Martin, a creative who always finds a way no matter how challenging the problem, key milestones may not have been met as efficiently or at all. Additionally, bringing CX on as a major collaborator has always been a bucket list item for me. Since he joined the lab, I have been actively searching for an excuse to work with him. Hopefully, this collaboration lives up to his expectations! Finally, co-mentorship from Cole and Jay, particularly the feedback given during one-on-ones and lab meetings, fostered my own sense of confidence and worthiness in these respective projects. This, in turn, added fuel to my

strongest desire to always be an engaged, respectful, and intellectually forward-thinking collaborator to CX, Choi, Silvia, and Beth. The work described in this chapter has endowed me with a lifetime of excitement for the future of developmental biology, single-cell technology, and monoclonal embryoid or organoid systems.

## Abstract

Organoid models derived from pluripotent stem cells (PSCs) are powerful *in vitro* systems for studying development and disease. The appeal of these models lies in their ease of cultivation, adaptability for genetic manipulation, and significant cell-type overlap with their *in vivo* counterparts. Current protocols involve aggregating numerous PSCs to form 3D spheres or embryoid bodies (EBs), which can then be guided to differentiate into specialized germ layer derivatives that resemble various organ systems or developmental stages, such as brain organoids or 'stembryos'. At the individual organoid level, each cell contributes a unique genotype or a single-guide RNA (sgRNA) in the context of a CRISPR screen, leading to the designation of such organoid systems as 'mosaic'. Through experiments with mosaic EBs, we highlight inherent challenges in current state-of-the-art mosaic organoid screens, including transgene silencing and intrinsic cell-type composition biases resulting from bottlenecking during the differentiation process. To address these issues, we introduce a pipeline for genetically barcoded mouse ESC-derived clonal organoids. Our clonal organoid pipeline establishes key advantages over mosaic organoids, including the capability to assess inter- and intra-organoid variation, robust co-capture of sgRNA and organoid barcode, and lineage reconstruction based on a single phylogeny per individual clonal organoid instead of many phylogenies per individual mosaic organoid. We

validate the utility of clonal organoids across high-content single-cell genomic technologies, including a pooled CRISPR screen and CRISPR-based molecular recording via DNA Typewriter. Together, our work lays the foundation for clonal, genetically homogenous organoids, offering the promising implications for developmental lineage reconstruction, perturbational screens, and investigations into early developmental biology.

## ***Introduction***

How cell lineages are specified during early development is a longstanding question in developmental biology. To better understand the underlying properties of cell fate decisions, various model systems have been used alongside single-cell genomic technologies, for instance, multimodal scRNA-seq for transcriptome and perturbation readouts. *In vitro* models, such as organoids, are genetically tractable systems that have been shown to recapitulate aspects of development, and for these reasons, have become attractive alternatives to more resource-intensive *in vivo* models, such as mice[1].

Organoids, including those modeling the brain[2], heart[3], kidney[4], and "stembryo" models like gastruloids[5], are typically generated by initiating the aggregation of hundreds to thousands of embryonic stem cells (ESCs) or induced pluripotent stem cells (iPSCs), collectively referred to as 'PSCs'. This aggregation is induced by culturing the cells in individual low-adherent wells under defined media conditions. A common feature in these organoid protocols is the formation of a crucial intermediate structure known as the embryoid body (EB)[6]. During this stage of the differentiation process, the three primary germ layers (endoderm, mesoderm, and ectoderm)

emerge within the EB. The formation of the EB intermediate is a key step that enables the development of various organoid systems.

Organoids generated through the current state-of-the-art workflow are commonly characterized as mosaic. For instance, in the context of a CRISPR screen[7], individual organoids consist of numerous cells with different genotypes or sgRNAs. Similarly, in the context of lineage tracing, each organoid comprises numerous starting progenitor cells. Despite the widespread use of mosaic organoids, they present several limitations that we believe influence their utility and robustness. Firstly, the inherent nature of mosaic organoid generation makes it challenging to quantify inter- and intra-organoid variability using a single-cell readout, a crucial consideration for pooled CRISPR screens. Secondly, because many progenitor cells contribute to an individual organoid, lineage relationships are established through many phylogenies per individual organoid, which makes it difficult to infer lineage trees in these organoid systems.

In this study, we utilize mosaic mouse embryoid bodies with pooled CRISPR screens to investigate the role of all transcription factors in germ layer formation. Despite identification of numerous statistically significant cell-type specific TFs, our findings reveal that bottlenecking induces biases in cell type composition, negatively impacting the reproducibility of our screen. To address these limitations associated with mosaic organoids, we introduce a novel pipeline for clonal organoid generation and develop an innovative perturbation construct that co-expresses a sgRNA and an organoid barcode. Demonstrating the suitability of clonal organoids for pooled CRISPR screens, we emphasize their advantages for measuring inter- and intra-organoid variability through single-cell RNA sequencing (scRNA-seq). Furthermore, we extend the clonal organoid pipeline to

lineage tracing using DNA Typewriter, applying this approach to clonal EBs and clonal gastruloids–a novel protocol. Remarkably, DNA Typewriter repurposed as an in-cellular barcoding device can also be used to measure inter-gastruloid variability. Profiles of 150 clonal gastruloids reveal previously unreported cell types compared to the conventional gastruloid protocol. Finally, we underscore the utility of clonal organoid systems and DNA Typewriter by addressing a key developmental biology question: does cell type composition get encoded before the onset of differentiation cues? Altogether our work highlights the advantages of clonal organoid systems in advancing our understanding of developmental processes.

## ***Results***

### *Mosaic EBs pose limitations for large-scale pooled CRISPR screens*

To begin examining developmental gene regulation, we profiled single-cell transcriptomes over a 3-week time course, sampling differentiating mosaic EBs at day 7, day 14, and day 21 (co-embedded UMAP across time course, **Fig. 1a**). Using germ-layer specific marker genes, we identify all major germ layers by day 21 of differentiation. Additionally, fine grained cell types were defined using label transfer and integration with *in vivo* mouse data[8] (**Fig. 1b-c**). Thus, we conclude that mosaic EBs transcriptionally overlap embryo mouse development, thus underscoring their relevance as a model system for studying developmental gene regulation.

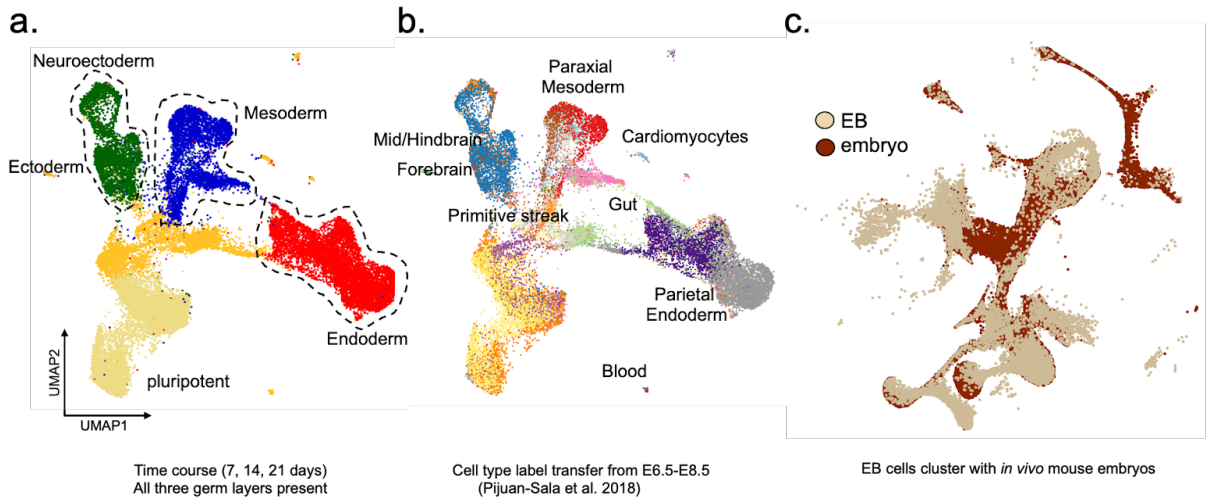



**Figure 1: ESC-derived embryoid bodies recapitulate aspects of mouse development. a.** A time course study profiling single-cell transcriptomes of wild-type mosaic EBs at day7, day14, and day21. All germ layers in UMAP embedding were manually annotated with literature defined marker genes. **b.** For granular cell type annotations, we used Seurat's cell-type label transfer algorithm, where cell types that are assigned on the UMAP plot come from *in vivo* mouse data[8]. Here, we found many cell types that transferred over to our dataset, including foregut, cardiomyocytes, mid and hindbrain, and blood. **c.** Additionally, we co-embedded our EB time course data, colored in beige, alongside embryo data, colored in maroon, and observed a strong correspondence between our EB system and mouse embryos.

We then implemented a pooled CRISPR screening strategy to identify key transcription factors essential for the development of each major germ layer (**Fig. 2**). To achieve this, we established monoclonal mESC lines that stably express either the CRISPRcut or CRISPRi machinery. We validated perturbations in mosaic EBs using lenti-CROP-seq with low multiplicity of infection (without selection) in an initial screen of 31 targets comprising well-known genes crucial for germ layer formation, alongside others important for both EB formation and pluripotency, including

non-targeting controls (NTCs). Our results demonstrate that CRISPR perturbations can be accurately captured in scRNA-seq using the 10x platform (data not shown). Additionally, the differentiation observed aligns with the time course data of wild-type mosaic EBs, indicating that our CRISPR screening approach does not interfere with normal EB development.

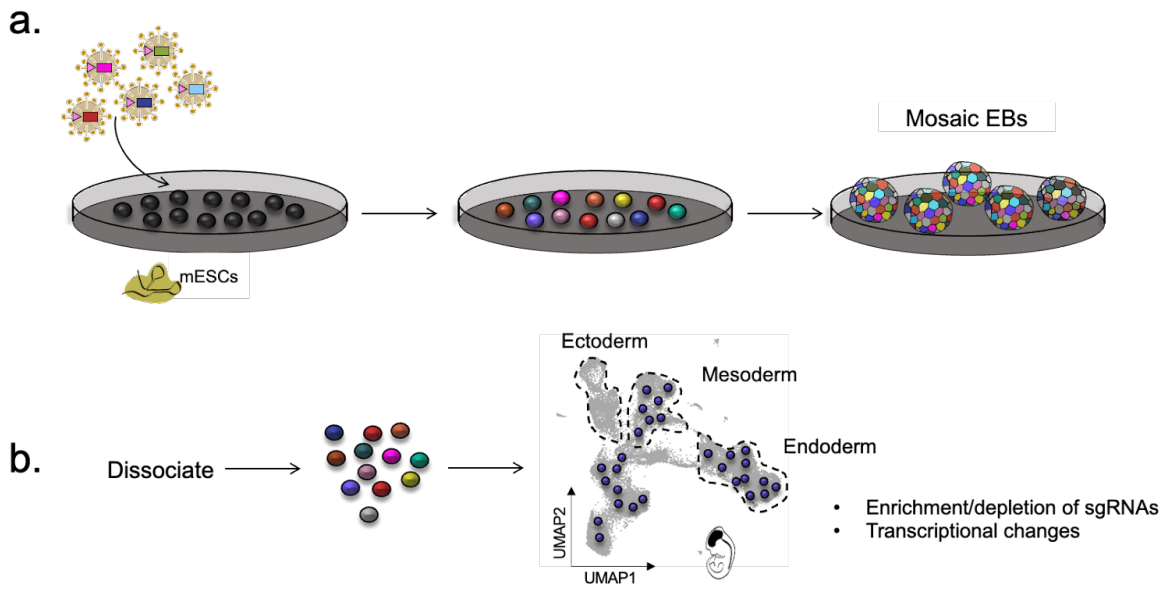



**Figure 2: CRISPR screening approach in mosaic EBs. a.** We first developed two monoclonal mESC lines that stably express the CRISPR machinery (CRISPRi or CRISPRcut). Next, we deliver a sgRNA payload using lentiviral vectors at a low multiplicity of infection so that on average each cell expresses 1 sgRNA. In the case of the pilot screen, no selection was carried out since all targets (except for NTCs) are essential (e.g. factors for pluripotency, EB formation, and differentiation). However, for the large-scale screens, we applied a puromycin selection such that only cells that remain on the dish are those that express a sgRNA, as depicted by the middle dish with different color cells. mESCs are then plated on a low-adheret dish in differentiation media, which leads to spontaneous aggregation and mosaic EB formation. **b.** After differentiation over 21 days, we dissociate mosaic EBs into a single cell suspension, and perform scRNA-seq where we capture

both the transcriptomes and sgRNA simultaneously for each cell. In our analysis we are looking for patterns of sgRNA enrichment or depletion across cell types which allows us to not only infer whether or not a given TF is essential for a given germ layer, but also allows us to measure any transcriptional changes due to the perturbations. As a hypothetical, if we had a sgRNA in our library targeting an important TF for brain development, we would find that cells with this sgRNA tend to be enriched in mesoderm and endoderm but depleted in ectoderm.

Having validated our pooled CRISPR screening approach, we executed two large-scale CRISPRcut screens in mosaic EBs: 1) a comprehensive screen targeting all known TFs, and 2) a follow up screen focusing on TFs identified as statistically significant, carried out across two biological replicates. In the initial screen, we targeted 1644 TFs (library composition: 4932 sgRNAs plus 20% NTCs) using CRISPRcut mosaic EBs (similar to schematic in **Fig. 2a**). A challenge arose due to transgene silencing of the lenti-construct (as shown in **Fig. 4a**), necessitating the flow sorting of GFP+ cells for input into the 10x Chromium. In total, we profiled 117,337 single-cell transcriptomes, each containing a median of 1 sgRNA per cell (**Fig. 3a-b**).

Similar to our pilot screen experiment, we expect cells with a perturbation to a key TF to exhibit a difference in cell type composition compared to wild-type cells. For example, cells with a sgRNA targeting a key TF that is crucial for mesoderm specification would likely follow an ectodermal or endodermal lineage trajectory instead of a mesodermal trajectory. To measure perturbation-driven differences in cell type composition, we used the UMAP representation of cell states to quantify statistically significant cell type differences between perturbed and unperturbed groups of cells. Upon confirming that our dataset captured diverse cell types present in the gastrulating mouse

embryo (**Fig. 3a**), we conducted an analysis to identify differentially distributed perturbations in UMAP space using two different approaches: 1) a chi-square test utilizing contingency tables of sgRNA targets and categorical annotations specific to each cell type, and 2) Moran's I spatial autocorrelation. By considering the union of targets with q-values more significant than the top 5% of NTCs (**Fig. 3c**), we identified 125 TFs that are differentially distributed, e.g. Lmx1B and Runx3 (**Fig. 3e**).

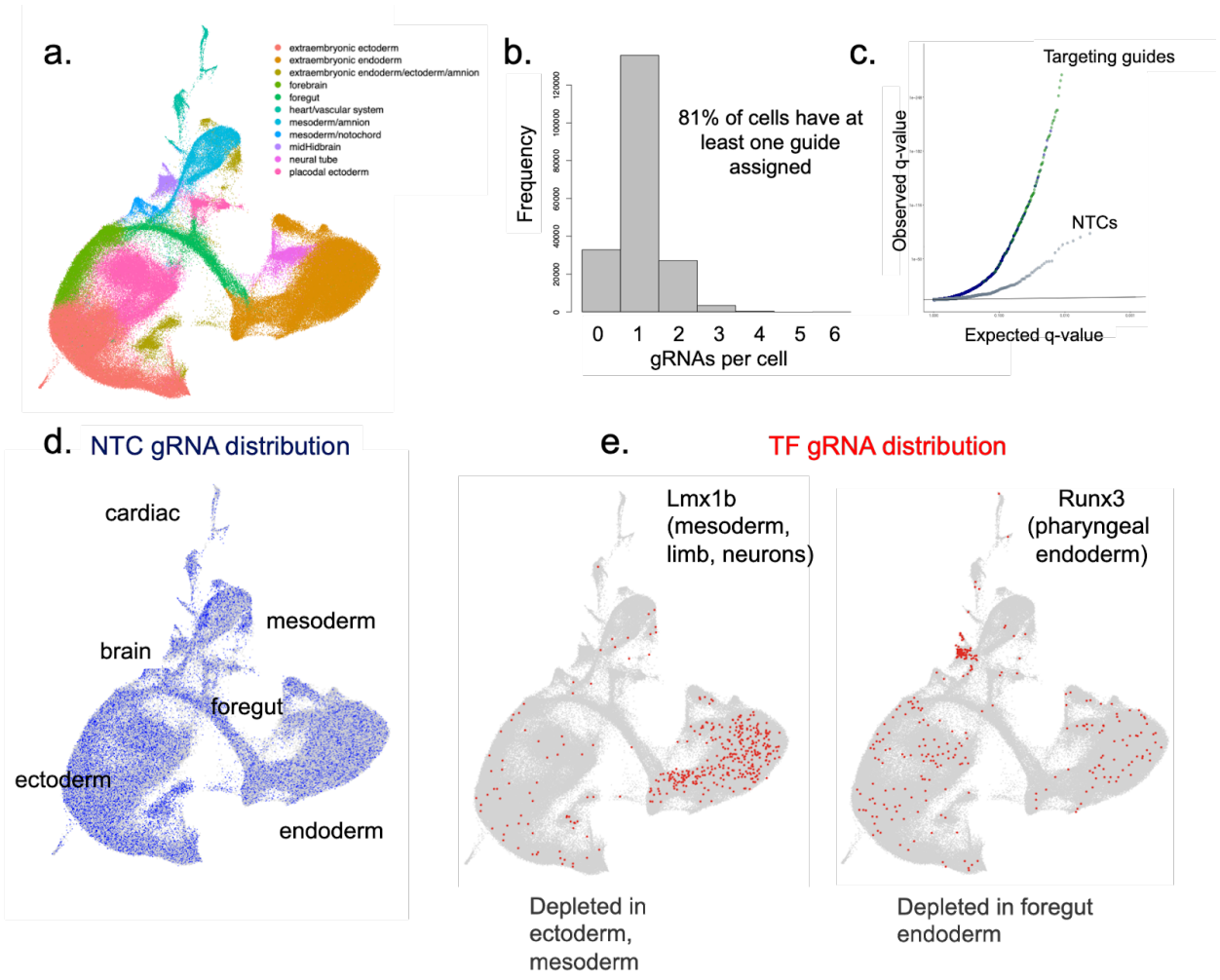



**Figure 3: Large-scale CRISPR screen targeting all known TFs in mosaic EBs. a.** UMAP embedding of ~217K cells with cell-types annotated with marker gene expression analysis and label transferring using mouse embryo data. **b.** Barplot quantifying number of sgRNAs per cell where we find 81% of cells profiled have at least 1 sgRNA. **c.** Q-Q plot of targeting sgRNAs and NTCs. **d.** sgRNA NTC distribution. **e.** Two TFs, Lmx1b and Runx3, with significantly distributed enrichment or depletion patterns.

In a follow up and more highly powered screen, we perturbed 125 TFs (3 sgRNAs per target plus ~15% NTC) using CRISPRcut mosaic EBs (**Fig. 4a**). Despite the transgene silencing with the

lenti-CROP-seq construct, we sorted the non-silenced GFP+ population and carried out scRNA-seq via 10x. Across two independent biological replicates, we profiled 147,007 cells with a median of 1 sgRNA per cell. After defining cell types using marker gene analysis and label transfer with *in vivo* mouse embryos, we looked at whether sgRNA targets are significantly distributed in UMAP space. Q-Q plot visualization shows that while replicate 1 has many significantly distributed sgRNA targets, replicate 2 NTCs are just as significantly distributed as sgRNA targets (**Fig. 4c**). These data underscore the lack of reproducibility in our follow up screen.

We investigated the possibility that bottlenecks at various stages might contribute to the observed reproducibility issues. Initially, we examined the composition of the plasmid library and transduced mESCs used for EB induction, finding a strong correlation between the two (data not shown). We then began to hypothesize that intrinsic properties, such as fewer initiating progenitor mESCs per mosaic EB, stochastic cell death, or clonal expansions over time during the differentiation process, could lead to reduced mosaicism on a per mosaic EB basis (**Fig. 4d**). This reduced mosaicism might then be propagated from culture plate to culture plate or become more pronounced in certain independent replicates compared to others.

To assess our hypothesis, we derived mosaic EBs that were regrown from the same pool of mESCs used in the follow-up screen. Our rationale was that EBs, with a typical cell count ranging between 10,000 and 15,000 cells per mosaic EB, would encapsulate the full complexity of the library (456 total sgRNAs) if they exhibited high mosaicism. To investigate this, we collected individual mosaic EBs on day 21, each deposited into a separate tube. We quantified the total number of cells per EB and conducted amplicon-seq on the sgRNA composition extracted from genomic DNA.

Our findings revealed that all 12 EBs tested (represented in four pairwise scatterplots, each featuring two different EBs) harbored only 50-100 different sgRNAs. Despite 10-fold higher cell counts than sgRNAs, these results suggest that mosaic EBs exhibit low mosaicism. Without the feasibility of extensive sampling, which is often constrained by the cost of 10x reagents, undersampling is likely to introduce bias into measurements. An ideal system that measures biases coming from bottlenecking during EB differentiation is one that could easily readout EB-to-EB heterogeneity. Unfortunately, achieving this with a single-cell readout is not feasible given the 'polyclonal' nature of mosaic EBs.

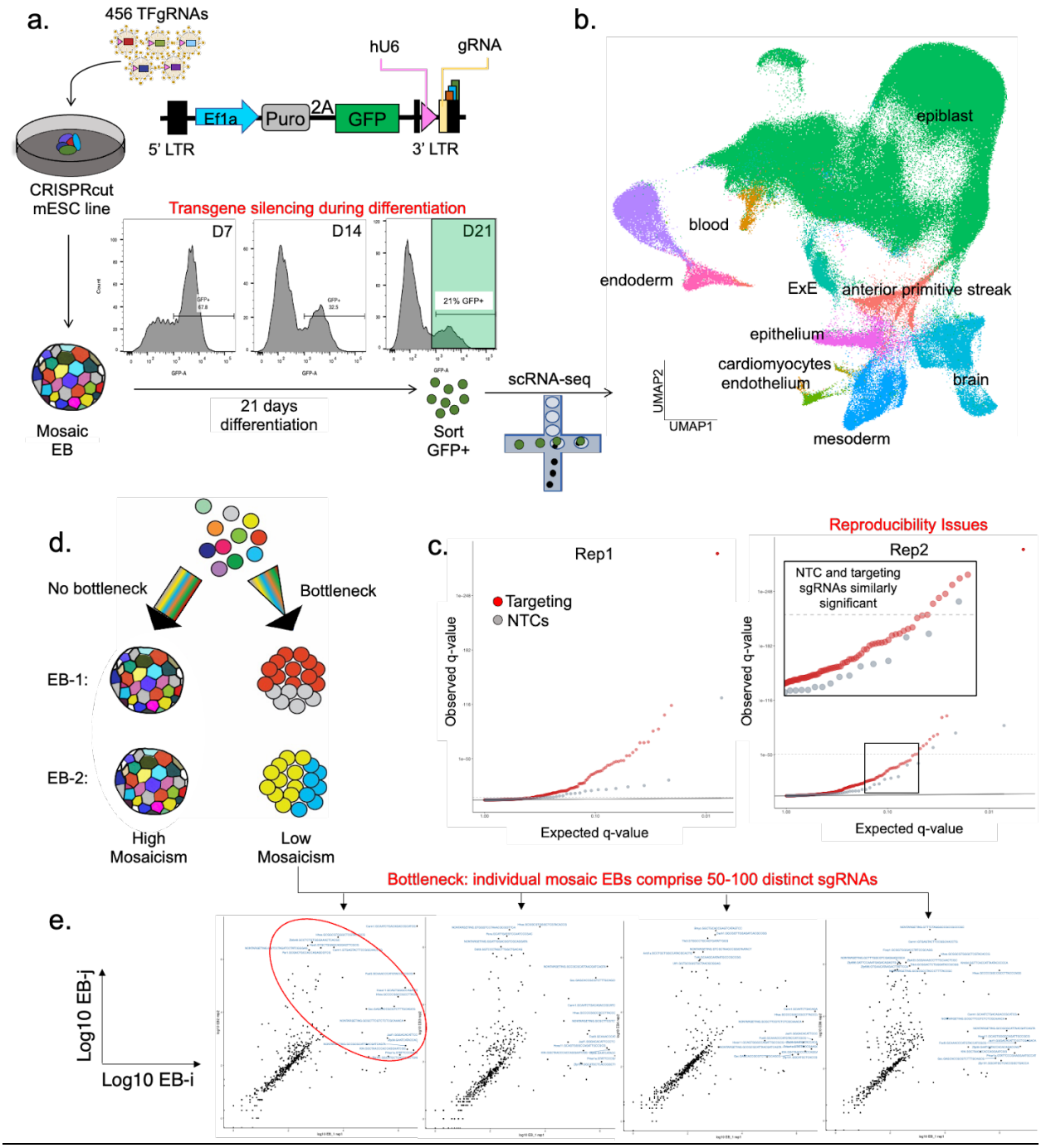



**Figure 4: Large-scale mosaic EB screen highlights key limitations with current state-of-the-art approaches. a.** Schematic of a large-scale mosaic CRISPR screen: a 456 sgRNA library using lenti-CROP-seq is transduced at low MOI into a monoclonal mESC line expressing the CRISPRcut

enzyme. Following puromycin selection, individual mESCs, each carrying a unique sgRNA, undergo expansion, dissociation, and plating in low-adherent dishes, spontaneously aggregating and differentiating into all germ layers by day 21. Notably, transgene silencing of the lenti construct is evidenced by the rapid decline in the GFP population throughout the 21 days of differentiation. Consequently, EBs are sorted for GFP+ cells and profiled by scRNA-seq to capture both transcriptomes and sgRNAs. **b.** UMAP embedding of high-quality transcriptomes passing filters with germ layers or their derivatives. **c.** Q-Q plots illustrating chi-square q-values for cell-type-specific enrichments of sgRNAs. Notably, in biological replicate 2, non-targeting control sgRNAs exhibit a distribution similar to that of target sgRNAs, both being significantly distributed. **d.** After ruling out other sources of bias, such as lower than expected complexity plasmid library or transduced mESCs, we hypothesized that bottlenecking during differentiation results in uneven genotype representation per EB, leading to low mosaicism and reproducibility issues. **e.** Twelve individual mosaic EBs were individually selected, placed into separate PCR tubes, and analyzed via amplicon-seq to identify distinct sgRNAs per EB. Surprisingly, pairwise scatterplots revealed that each EB contained 50-100 unique sgRNAs, demonstrating markedly reduced diversity despite the presence of 10,000 to 15,000 cells per EB.

***Clonal EBs offer key advantages over mosaic EBs***

Two main challenges surfaced during our mosaic EB screens: 1) transgene silencing using lenti-CROP-seq, and 2) bottlenecking during mosaic EB differentiation. To overcome the first challenge, we devised a new construct, 'piggyFlex', a piggyBac transposon with dual RNA expression of a sgRNA and mRNA organoid barcode (orgBC) tucked into a 3' UTR of puro-2A-GFP (**Fig. 5a**). Both the sgRNA and orgBC are capturable with the 10x and sci-RNA-seq platforms

via CS1 primers (sgRNA) and pdT or CS2 (orgBC), respectively. We previously validated robust perturbation and capture of both transcripts in a pooled CRISPR screen in mESCs.

To address the second challenge, we reasoned that measurement bias from EB-to-EB heterogeneity could easily be overcome using clonal EBs, in which each EB derives from a barcoded single cell-of-origin. To test whether it is possible to obtain clonal EBs, we first stably integrated a highly complex library of NTC-orgBC pairs via piggyFlex into mESCs at a low multiplicity of integration. We then generated a homogenous single cell prep via sorting followed by plating at high and low densities on a mouse embryonic feeder layer. After 5-6 days, single cells now turned into clonal colonies are lifted with Collagenase IV followed by gentle agitation and transferring into differentiation media on low adherent plates. The suspended clonal aggregates then become spherical and undergo differentiation as clonal EBs that will ultimately comprise various germ layers (**Fig. 5b**).

Next, we profiled 15 clonal EBs from each condition using scRNA-seq. Cell types were annotated through marker gene analysis and label transfer with *in vivo* embryo mouse data, revealing that clonal EBs exhibited all major germ layers, consistent with wild-type mosaic EBs (**Fig. 5c**). In addition to transcriptomes, both sgRNA and orgBC were co-captured, with a capture rate of approximately 90% determined by post-UMI cutoff for sgRNA transcripts. A significant advantage is that piggyFlex is not silenced during the differentiation process, eliminating the need for flow sorting in our pipeline. Furthermore, since every clonal EB is barcoded, we can isolate the population of cells belonging to each of the profiled EBs. In **Fig. 5d**, a barplot representing total numbers of cells per orgBC for the low seeding density condition is shown. Here, we

recovered 25 unique orgBC despite having picked 15 EBs. However, while most clonal EBs are represented by a single orgBC, two clonal EBs are represented by a double orgBC via two separate integration events. Nevertheless, we attribute other detected barcodes with few cells to EB fragments that may have been inadvertently picked up during the EB picking process. More importantly, the orgBC enables us to examine intra- and inter-EB variation (**Fig. 5e**).

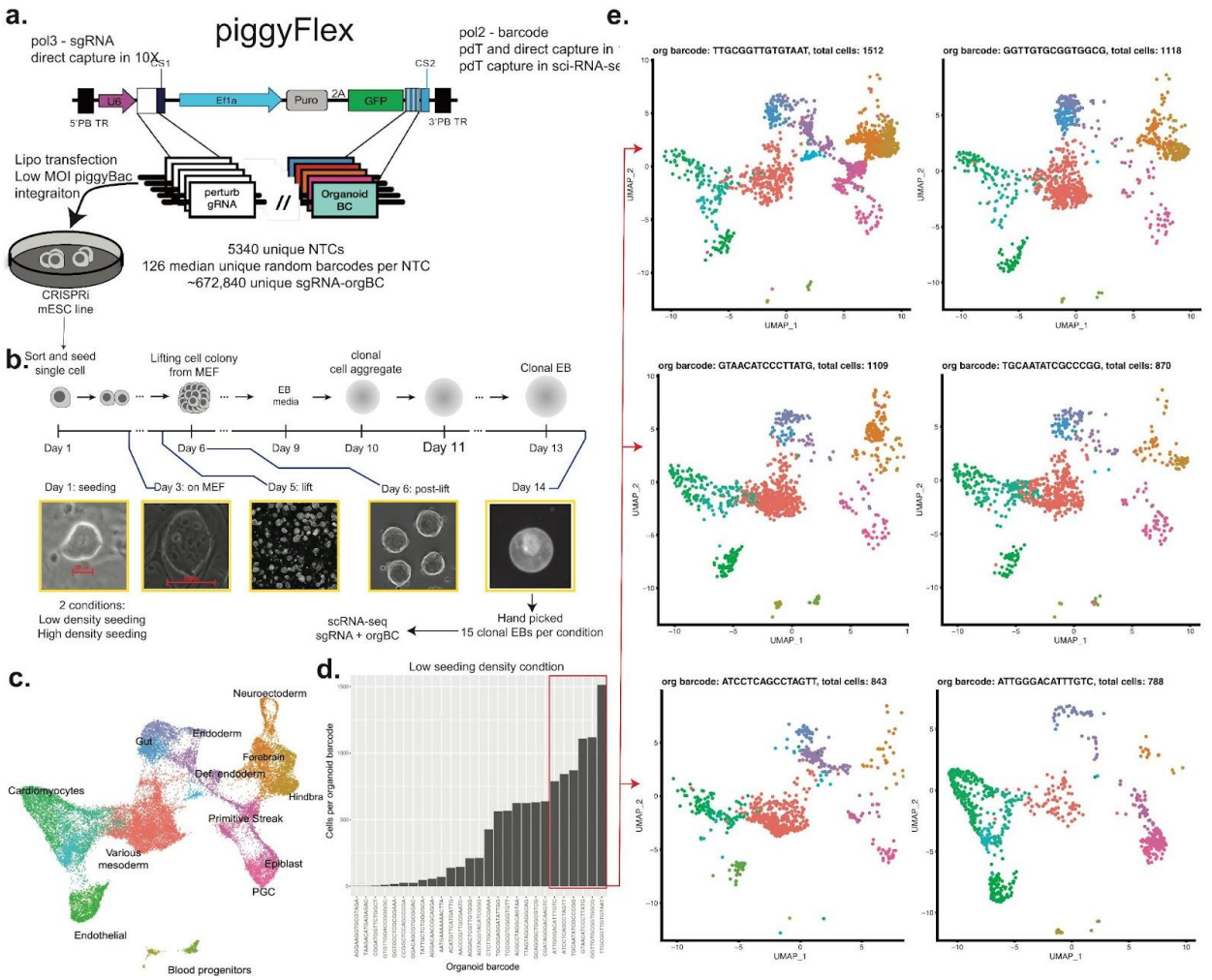



**Figure 5: Clonal EBs with piggyBac 'piggyFlex' cargo enables quantification of EB-to-EB heterogeneity. a.** We initially aimed to address the issue of transgene silencing observed in our

previous mosaic screens with the lenti-CROP-seq vector. To overcome this, we developed a novel construct named 'piggyFlex,' a piggyBac transposon. This dual-RNA expression system facilitates the simultaneous expression of both sgRNA and a unique organoid barcode. The barcode can be captured directly on the 10x bead (CS1) or through the polydT on the 10x bead or in the RT step of sci-RNA-seq, offering flexibility across different modalities. Our results demonstrate that, during 21 days of differentiation, piggyFlex maintains stable expression in EBs, with approximately 90% GFP positivity (data not shown). **b.** To test if we could develop a scalable clonal organoid protocol we took an NTC library of low MOI piggyFlex mESCs, sorted and seeded on a MEF layer at high and low density, and cultured them for 5 days followed by lifting with Collagenase IV treatment and gentle agitation ensuring the colonies remain intact. Colonies, now clonal aggregates, are grown in differentiation media on low adherent plates for 8 days. **c**. UMAP embedding of day 8 clonal EBs (15 clonal EBs picked per low and high seeding density conditions). **d.** Barplot of cells per organoid barcode. **e.** UMAP embedding split by top 6 organoid barcodes.

***Proof-of-principle CRISPR screening in clonal EBs***

Next, we devised a CRISPR screen using the CRISPRi clonal EB pipeline that we developed. Here we used piggyFlex to introduce a complex library of orgBCs each paired to a sgRNA targeting either 1 of 9 TFs or NTCs (3 sgRNAs per target). We set up two conditions. The first condition comprises individual mini pools of clonal EBs in which each mini pool comprises 3 unique sgRNAs to a single target (10 total mini pools). The second condition comprises a pool of all sgRNAs to all targets with the exception of Carm1, which was our strongest statistically significant

hit in the follow up screen that was subsequently over-represented and therefore we intentionally left Carm1 out of the pool to avoid uneven sgRNA representation.

We reasoned that having condition 1 with individual mini pools would allow us to maximize our power to evenly sample single cell transcriptomes from each of the individual mini pools, while condition 2 would allow us to show that clonal EBs could effectively be applied to a pooled screening format. While the analysis is still ongoing, we can qualitatively appreciate that the Hand1 mini-pool, a key TF important for mesoderm germ layer formation, showcases cells with sgRNAs depleted in the mesodermal population (**Fig. 6d**), which is also consistent with the distribution of cells comprising orgBCs from the Hand1 mini-pool (**Fig. 6e**). In addition to this possible hit, we are able to facet the UMAP embedding by individual clonal EBs based on the orgBC. Here we can appreciate the variation in cell type composition across the UMAPs yet depletion of the Hand1-expressing mesodermal population appears to be consistent across the UMAPs.

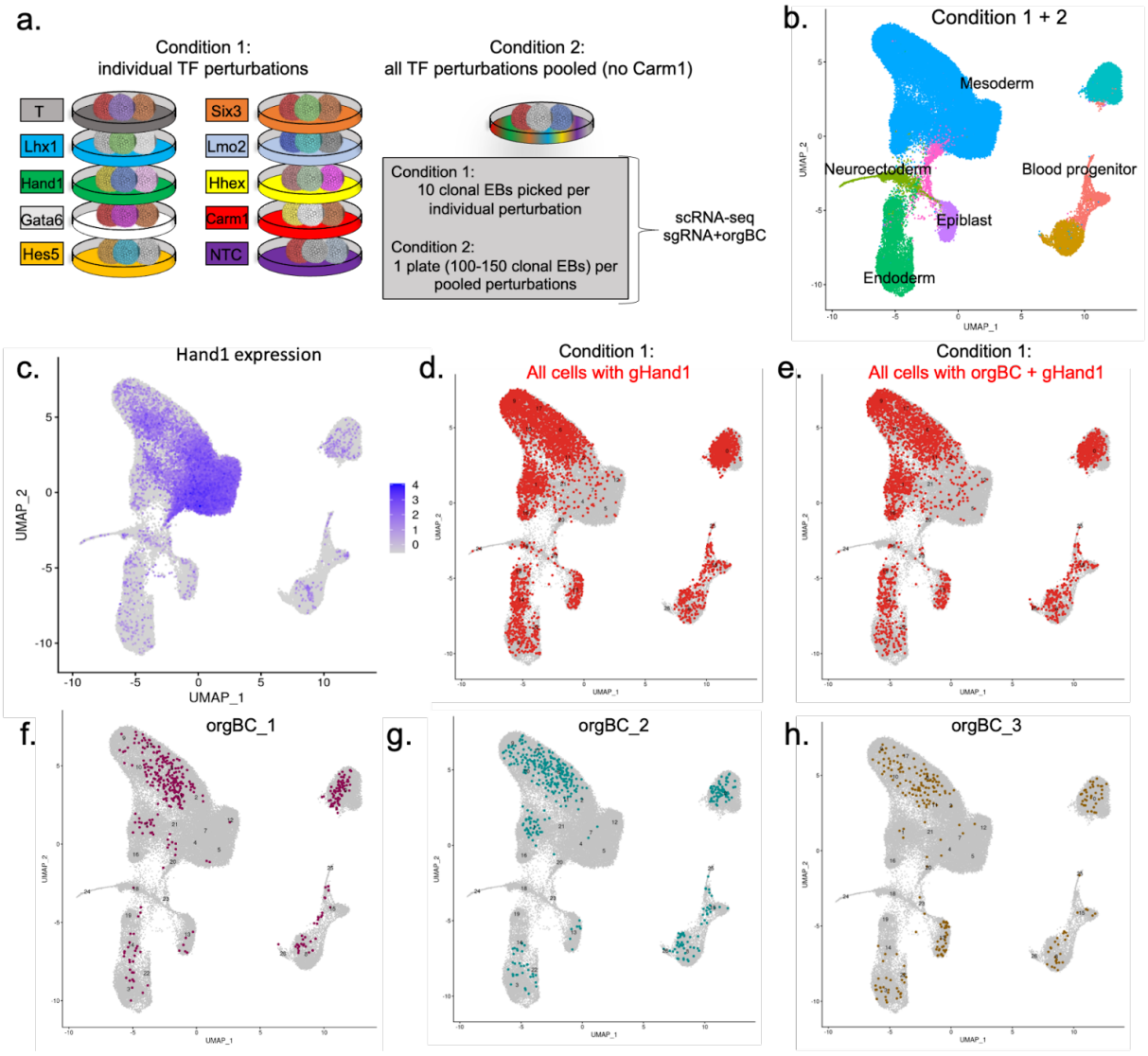



**Figure 6: A proof-of-principle experiment for scalable CRISPR perturbations using clonal EBs. a.** CRISPR screen in clonal EBs perturbing 9 TFs (selected based on DEG in cognate cell type or hit in prior screen) and 1 NTC (3 sgRNAs per target). Two conditions carried out: 1) individual mini pools of separate TF perturbations, wherein each clonal EB comprises a unique sgRNA-barcode (via piggyFlex) yet all sgRNAs target the same TF, and 2) a pool of clonal EBs comprising unique sgRNA-barcode pairs in which all TF perturbations are represented. **b.** UMAP embedding in which the major germ layers are represented. **c.** UMAP embedding of Hand1

expression. **d.** All cells colored red that harbor a sgRNA targeting Hand1 in which depletion of red cells is notable in the mesodermal cluster. **e.** All cells colored red harboring organoid barcodes, previously paired to sgRNAs, for Hand1 perturbation. **f-h.** UMAP embedding is colored by individual clonal EBs via cells sharing the same organoid barcode.

***Clonal organoid concept yields new clonal gastruloid protocol***

We next sought to apply the concept of ESC-derived clonal EBs to a different model system, for example, other stembryo models of development. While EBs recapitulate the formation of germ layers, they lack morphological features that resemble *in vivo* embryos. We therefore considered the possibility of generating clonal gastruloids whose conventional counterpart, mosaic gastruloids, exhibit symmetry breaking, gastrulation, and the formation of the three major axes. An important difference compared to EBs is that gastruloids require a precise concentration of chiron wnt inhibitor, which necessitates their growth in individual wells of 96-well plate. We therefore devised a clonal pipeline strategy that would be amenable to growth requirements for robust gastruloid differentiation.

To enable the generation of a gastruloid that starts from a single cell, we developed a culturing protocol that begins with clonal mESC aggregates (**Fig. 7a**). Similar to the clonal EB protocol, we first dissociate mESCs into single cells and plate them onto a feeder layer of MEFs at a low density (100 to 1000 cells per one 6-well plate). Enmeshed in the MEF layer, each mESC single cell grows into a clonal population that is distinguishable by its domed-shaped morphology and sharp borders. At 4-5 days after mESC seeding, the co-culture is treated with Collagenase IV, which gently lifts the colonies away from the MEF layer with minimal disturbance to the mESC colonies

themselves. mESC colonies are then collected and cultured another day in a non-adherent plate, where each colony forms a spherical cell aggregate. Each cell aggregate is then transferred to an individual well of a 96-well culture plate, thus becoming the substrate for inducing gastruloid differentiation.

To test whether the resulting spherical mESC aggregates derive from a single seeding cell, we constructed a mESC line with random, static DNA barcodes to trace the clonality of each cell. Static DNA barcodes were cloned as a lentiviral vector with 8-mer random DNA sequences and two selection genes, GFP and Puro-resistance gene, which were transduced into a pool of mESCs at a low MOI. We first used flow-sorting to determine the MOI of transduction to be near 0.01 (**Fig. 7b**) and used Puromycin drug selection to retain cells containing a DNA barcode. Based on a Poisson distribution of integration at MOI of 0.01, we estimate that less than 1% of cells harbor more than a single DNA barcode.

Using the mESC line with DNA barcodes, we generated cell aggregates for gastruloid induction following the protocol for clonal mESC cell aggregate production. At the day 6 time point after seeding mESCs on an MEF layer as single cells, we lifted each clonal mESC colony by treating the cell culture with Collagenase IV, pooled cell aggregates, and plated them into a non-adherent plate, allowing them to form a spherical cell aggregate overnight. At the day 7 time point, we transferred each cell aggregate into its own individual well of a 96-well plate by gently pipetting them under a dissecting microscope.

To test whether the clonal cell aggregate can form a gastruloid, we induced cell aggregates into a gastruloid. Typically the conventional (mosaic) gastruloid induction protocol starts by seeding 300 to 500 mESCs into a single non-adherent well containing Neural-differentiation media, which over the ensuing 24-48 hours the cells then form a spheroid cell aggregate. At the 48-hour time point, aggregates are pulsed with the addition of 3 uM CHIR99201 for 24 hours to promote symmetry breaking of cell aggregates and differentiation into both the mesodermal and neuronal cell types. Using our clonal aggregate protocol as the starting substrate rather than many mESC seeding cells, we similarly observed elongation of cell aggregates 48 to 72 hours after the CHIR removal (120- and 144-hour time-point in the conventional protocol), consistent with the conventional protocol, thus suggesting that our clonal protocol is compatible with gastruloid formation.

To see whether the resulting gastruloid-like cell culture is indeed clonal, we sequenced the DNA barcode of each culturing well. Out of 234 wells comprising the initially plated cell aggregates, we reliably collected PCR-amplified DNA barcodes from the genomic DNA of 201 wells (**Fig. 7c**), where some cell aggregates were lost during media exchanges. We sequenced the DNA barcodes and observed that 182 out of 201 cell aggregates (about 90%) had a dominating 8-mer DNA barcode with a frequency above 90% (**Fig. 7d**), considered as a 'monoclonal' culture. In the rest of the 19 cell aggregates, we observed that 5 had the second dominant DNA barcode with a frequency below 10% (mono-clonal), 13 had only the first and second dominant DNA barcode with a frequency above 10% (di-clonal), and only one had the first three dominant DNA barcode with a frequency above 10% (tri-clonal) (**Fig. 7e**). Our result suggests that our protocol not only robustly generates a clonal cell aggregate with a predominantly single ancestor cell, but that differentiation into a gastruloid is feasible with this new protocol.

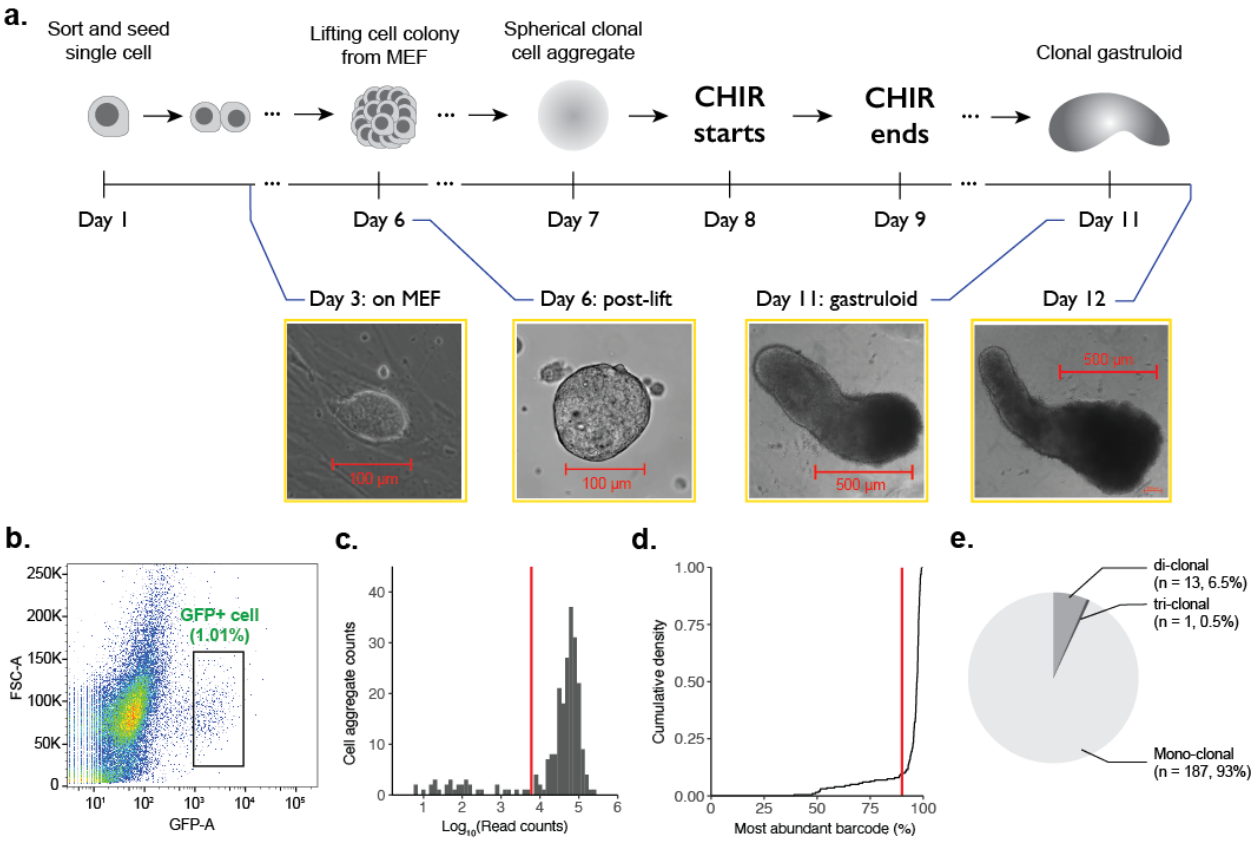



**Figure 7. Clonal gastruloid formation protocol. a.** Schematics of the clonal gastruloid formation protocol. **b.** Flow-sort data for determining the multiplicity of infection for lentiviral DNA barcode. **c.** Distribution of read counts across 234 wells containing clonal cell aggregates. The red vertical line represents a read count of 6000 used to retain clonal cell aggregates with reliable DNA barcode recovery. **d.** Cumulative density plot of clonal cell aggregates with the frequency of their most abundant DNA barcode. The red vertical line marks 90%, where the majority of DNA barcodes from clonal cell aggregates are a single sequence. **e.** Pie chart showing assignments of clonal gastruloids into estimated ancestor cells. The number of seed ancestor cells was inferred as the number of DNA barcodes with a frequency greater than 10%.

***Cell line generation for lineage recording using DNA Typewriter***

After establishing a working protocol to culture the clonal gastruloid, we next tested whether the clonal gastruloid could be efficiently formed using mESC lines engineered to record high-resolution lineage information. The recorded lineage information can be used to reconstruct the lineage tree of the clonal gastruloid, resolving lineage relationships among all recovered cells that share the same ancestor cell. To reconstruct the lineage relationships at a high resolution, we implemented the DNA Typewriter lineage recording system by genetically integrating all necessary components into an mESC line. The DNA Typewriter lineage recording system includes three genetic components: 1) an expression cassette for a prime editor, 2) an expression cassette for a prime editing guide RNA (pegRNA), and 3) an expression cassette for a DNA Tape, which is edited by pegRNA to generate a series of random, synthetic mutations that mark the cell division events. Across three necessary components, high-resolution lineage recording requires multiple copies of pegRNAs and DNA Tapes to ensure highly diverse editing patterns and multiple recording sites, respectively, while we only need a reliable expression of prime editor over time.

To increase the copy number of the Recording Cassette, we performed two rounds of transfections for piggyBac transposon integrations. For the first round of transfection, we mixed three plasmids, PB-RC, PB-PuroR, and HyPBase, at a mass ratio of 90:5:5, intending to select cells with Puromycin resistance after integration. The first round of transfection for piggyBac integration and Puromycin selection generated the cell line mRECv3, which was then subjected to another round of transfection for piggyBac integration with the plasmid mix of PB-RC, PB-iPEmax-BlastR, and HyPBase at a mass ratio of 75:20:5. After transfection, we cultured and selected cells with Blasticidin, which resulted in the cell line mRECv4 that contained diverse clones with different integrations of PB-RC and PB-iPEmax-BlastR. To generate and characterize individual cell lines,

we picked 32 colonies to assess the copy numbers of the Recording Cassettes and Doxycycline-control of prime editing in each cell line. For all 32 lines, we measured the prime editing efficiency for lineage recording with or without the addition of Doxycycline. We identified four cell lines (mRECv4-C5, mRECv4-C23, mRECv4-C25, and mRECv4-C32) that matched our expectation on Dox-inducible prime editing, where we observed a high rate of editing with Doxycycline but very little without it.

In addition to the extent of Dox-inducible prime editing, we characterized these four cell lines in two additional criteria: First, we estimated the copy number of DNA Tape in the Recording Cassette and their relative expression strength in the form of RNA transcripts for scRNA-seq capture. Using the static DNA barcode (TargetBC) associated with each DNA Tape construct, we estimated the copy number to be between 50 to 150 copies, although we had preliminary data of multiple integrations of DNA Tape that share the same TargetBC (data not shown), which may increase the actual number of integration events. Second, we measured the editing rate over twelve days, where we observed a cumulative increase in the overall editing of DNA Tape without an obvious saturation event. Across four cell lines, we observed that the editing consistently accumulates for the twelve days, a time frame that is close to the clonal gastruloid induction protocol including seven days before the CHIR addition and three days after. In mRECv4-C5, for instance, we observed 0.8 edits per DNA Tape in the first 2 days after Dox-induction, and 0.4 edits per DNA Tape in the last 2 days. Assuming we recover tens of DNA Tape per cell, the observed editing rate should generate tens of editing events per day, which is close to the doubling rate of our mESC lines, sufficient to record and reconstruct every cell division during the clonal gastruloid induction protocol.

While the generated mRECv4 lines were tested for inducible Prime Editor activities, it is possible that Prime Editor is silenced or loses its activity over the course of differentiation in each clonal stembryo. We therefore appended fluorescent protein, mCherry, to the Prime Editor construct via P2A sequence to monitor its expression over time, but we anecdotally observed that mCherry fluorescence under the microscope is not always indicative of highly active Prime Editor expression. Therefore, we devised a new assay, "Debris-seq", to assess the recording efficiency without destroying the clonal stembryo. During the clonal stembryo induction protocol, the culturing media of each well harboring a cell aggregate is changed out with a fresh media every day. The replaced media contains cells released from the cell aggregate, and potentially dead cells with its genomic DNA intact. We reasoned that we can collect cellular material from the replaced media, extract genomic DNA via cell lysis, PCR-amplify the DNA Tape region, and sequence the amplicon to assess the overall editing progression of each well. To test Debris-seq, we collected replaced media at the end of CHIR addition (Day 9 in the clonal gastruloid protocol or 72-hr time point in the conventional gastruloid protocol) and the next day (96-hr time point), and sequenced the DNA Tape region after PCR-amplification from salvaged genomic DNA. We compared the frequency of static barcode between two days, which showed high correlation (Pearson correlation 0.98), suggesting that genomic DNA can be reliably collected from the replaced media and inform which clonal stembryo has accrued sufficient editing events without saturation to reconstruct lineage relationships.

***Monophyletic lineage reconstruction of a clonal gastruloid using DNA Typewriter***

Combining the clonal gastruloid protocol and the engineered DNA Typewriter lineage recording cell line, we set out to reconstruct the lineage relationship across cells within individual mouse clonal gastruloids (**Fig. 8a**). While our engineered mRECv4 lines morphologically resembled mESCs, integration of hundreds of piggyBac transposons, when combined with constant exposure to relatively high concentration of doxycycline at 100 ng/mL, could affect its ability to form and differentiate into a gastruloid. Therefore, we mixed and flow sorted three mRECv4 lines – C5, C25 and C32 – onto the MEF layer as a single cell at the start of the clonal gastruloid induction protocol. We followed the clonal gastruloid induction protocol, with the only difference being the addition of 100 ng/mL doxycycline throughout the protocol and the collection of replaced media on Day 10 (one day after the CHIR removal, corresponding to 96-hr time point in the conventional gastruloid induction protocol) for Debris-seq. At Day 11, we selected 3 clonal gastruloids based on their elongated morphology and high degree of editing based on the Debris-seq result, and pooled them together to load on a single lane of 10x Genomics scRNA-seq. We also selected 5 additional clonal cell aggregates, which were subjected to the same protocol but did not show a substantial elongation but a small protrusion.

We processed and sequenced the transcriptomic and the DNA Tape libraries from two lanes of single-cell RNA-seq on the 10x Chromium, and subjected them to a filter for selecting cells with a high threshold based on UMI vs. mito content and doublet scores. To understand whether our clonal gastruloids induced from a highly engineered cell line resembles the published gastruloid data set, we integrated our data with scRNA-seq data generated by Veenvliet et al.[9], which showed a close overlap observed in the UMAP embeddings (**Fig. 8b**). The close overlap between two

datasets suggest that our modified protocol generates clonal gastruloids close to other non-clonal gastruloids in the context of transcriptomically defined cell states.

Next, we processed the DNA Tape libraries and calculated the clonal distance among each pair of cells using a distance matrix. This allowed us to reconstruct the lineage tree of clonal mouse gastruloids. In the first lane, we mixed three clonal gastruloids, which all originated from the mRECv4-C5 cell line. We first used the clonal distance to construct a tree of all cells recovered from the same reaction lane and separated them into 3 groups based on early divergence (**Fig. 8c**).

Among three groups, we focused our initial analysis on the largest group of nearly 1800 cells (**Fig. 8d**), where both neuronal and somitic cell types were well represented. In addition to manual cell-type annotation, informed by prior gastruloid datasets, we integrated our single-cell transcriptomic data with *in vivo* embryo mouse data and appended the cell type labels to the lineage tree (**Fig. 8e-f**). Qualitatively we find many clades along the tree that cluster based on transcriptional similarity. Additionally, we observe dense branching patterns that harbor bifurcations, which further aid in inferring fate decisions of progenitor cells (**Fig. 8f, arrowhead**).

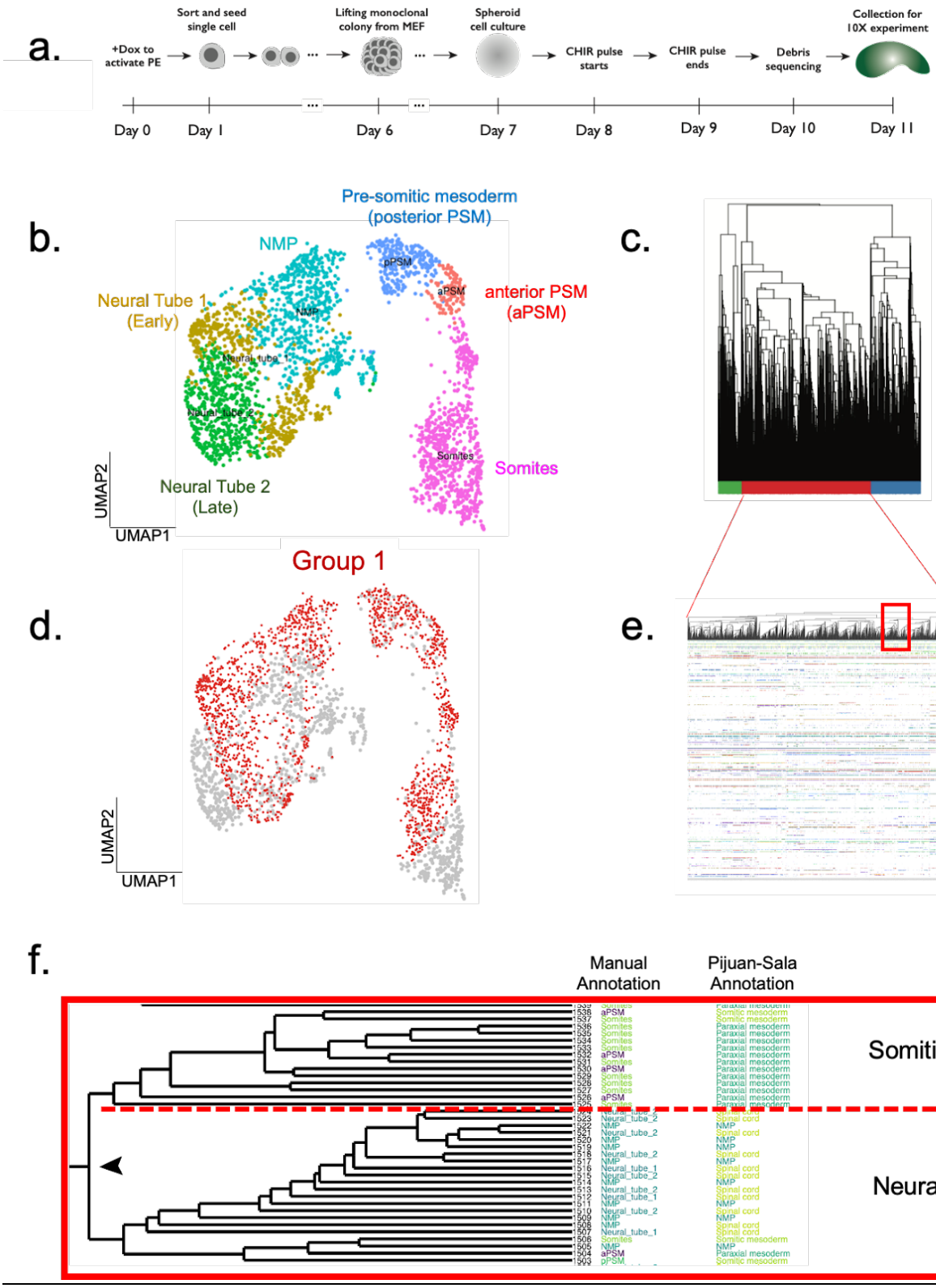

a.
Sort and seed single cell
Lifting monoclonal colony from MEF
Spheroid cell culture
Collection for 10X experiment
+Dox to activate PE
CHIR pulse starts
CHIR pulse ends
Debris sequencing
Day 0
Day 1
Day 6
Day 7
Day 8
Day 9
Day 10
Day 11
b.
Pre-somitic mesoderm (posterior PSM)
NMP
pPSM
aPSM
anterior PSM (aPSM)
Neural Tube 1 (Early)
Neural_tube_1
Neural_tube_2
Somites
Somites
Neural Tube 2 (Late)
UMAP2
UMAP1
c.
d.
Group 1
UMAP2
UMAP1
e.
f.
Manual Annotation
Pijuan-Sala Annotation
Somitic
Neural

**Figure 8: Lineage tracing with DNA Typewriter in clonal gastruloids. a.** Protocol combining lineage tracing using DNA Typewriter and clonal gastruloid formation. **b.** UMAP embedding with cell-type annotations derived from prior literature on conventional mouse gastruloids. **c.** Clonal distances derived from Tapes define individual clonal gastruloids. **d.** Selection of clonal gastruloid, 'Group 1', labeled red in UMAP space. **e.** Lineage tree of Group 1 gastruloid where rows represent cells and columns represent individual Tapes. **f.** Close-up of the lineage tree where a progenitor state (arrowhead) bifurcated into two lineages predominantly made up of somitic or neuronal cells.

***Quantifying lineage relationships in clonal gastruloids***

To understand the performance of lineage recording using DNA Typewriter, we first performed a simple analysis using the calculated clonal distance. For each pair of cells, in addition to their clonal distances, we estimated their transcriptomic distance by calculating the Euclidean distance in the first 50 PCA dimensions. These resulted in two distances across all possible cell pairs: clonal distance and transcriptomic distance, both scaled to be between 0 and 1. We expected that cells that have a small lineage distance, thereby sharing a close lineage, would have a small transcriptional distance, whereas the cells that have a small transcriptional distance might not necessarily have a small lineage distance due to transcriptional convergence during the gastruloid induction protocol. Indeed, cell pairs with small clonal distances (between 0 and 0.1) also had small transcriptional distances (**Fig. 9a**), while cell pairs with small transcriptional distances did not show a significant difference in their lineage distances (not shown). Our result matches well with our expectation that closely related cells share a similar transcriptomic state.

We set out to understand the cell types arising from the gastruloid induction protocol using the inferred lineage relationships among cells. The cell culture for the gastruloid induction is thought to be starting with pluripotent or epiblast-like cells in the beginning, which differentiate into a neuromesodermal progenitors (NMPs) that make a fate-decision between somitic and neural lineages. The bifurcation of NMPs into two different lineages makes the gastruloid one of the simplest models for cell-fate decision. The reconstructed lineage relationships among cells within clonal gastruloid can inform the characteristics such as heritability and relationship of each recovered and annotated cell type.

A new analytic framework referred to as PATH (Phylogenetic Analysis of Transcriptional Heritability) can incorporate the entire tree structure to quantify the heritability and relationships among cell type annotations, in forms of phylogenetic auto-correlation and cross-correlation, respectively, based on the lineage tree[10] (**Fig. 9b**). When applied to our lineage data and cell-type annotation, we observe heritability or autocorrelation that is more consistent with our expectation: In the somitic lineage, somites have higher phylogenetic auto-correlation z scores than pPSM and aPSM, suggesting that the progenies of somites are more likely to be somites rather than transition to other cell types. In the neuronal lineage, we observe a similar trend that neural tube 2 (late) cells have higher phylogenetic auto-correlation z scores than neural tube 1 (early) cells and NMPs. These observations are consistent with our expectation that both neural tube 2 cells and somites, which emanate as opposing trajectories from their NMP precursors, are cell annotations classified in our data as terminally differentiated cells within the clonal gastruloid systems.

We then used the PATH framework to quantify the phylogenetic cross-correlation between cell-type annotations, reasoning that might reveal the relationships between six cell-type annotations used in the gastruloid data (**Fig. 9c**). Looking across the cross-correlation matrix, we can reconstruct the linear relationship among cell-type annotations, in the order of neural tube 2, neural tube 1, NMP, aPSM, pPSM, and Somite, consistent with our previous analysis using clonally mutual nearest cells (not shown).

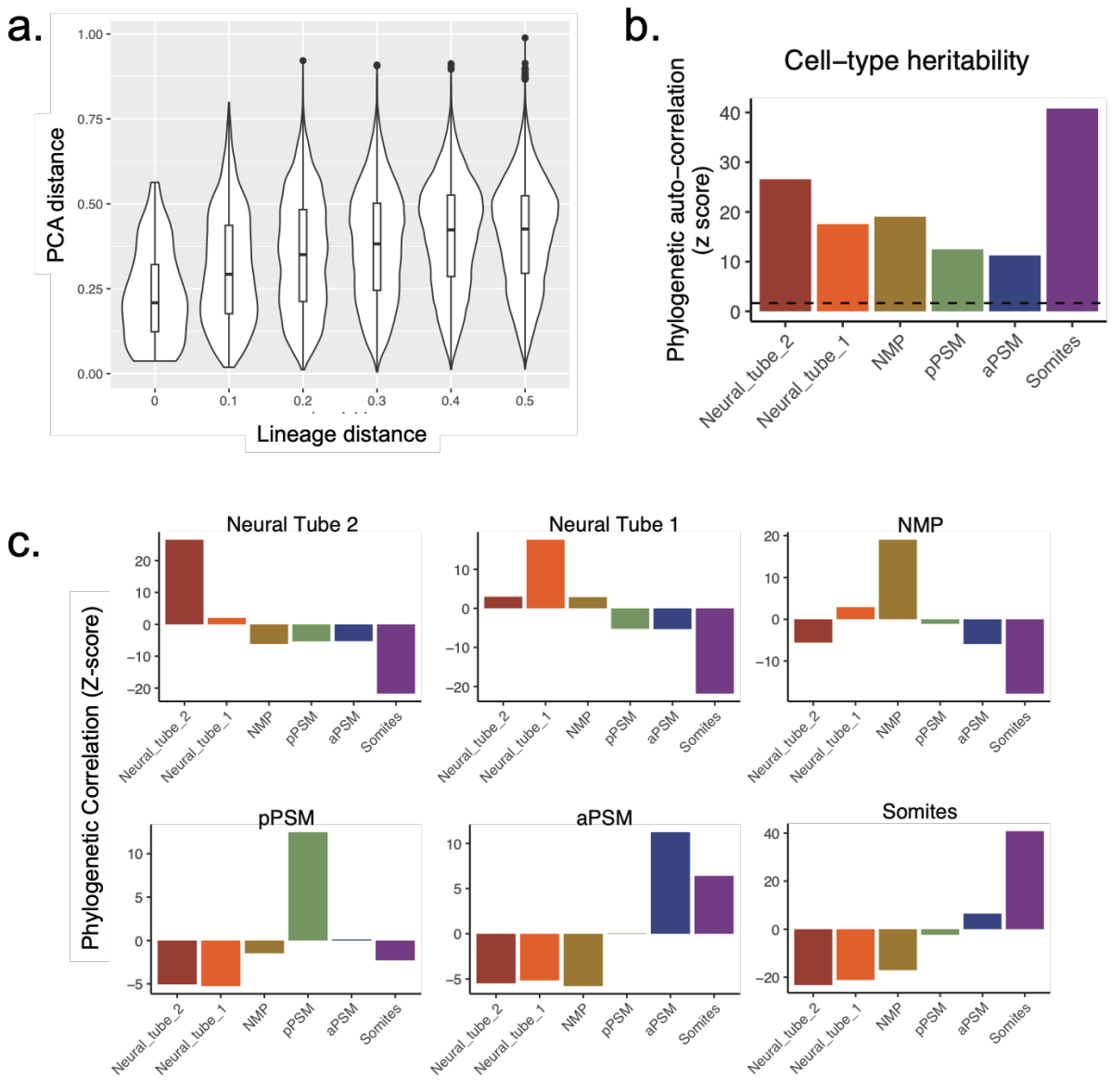


**Figure 9: Lineage-based cell-type relationships. a.** Violin plots of pairwise relationships between lineage distance and transcriptional distance (based on first 50 PCA dimensions). **b.** PATH generated phylogenetic auto-correlation scores to determine degree of heritability within cell-type. **c.** PATH generated phylogenetic cross-correlation to determine degree of heritability within and between cell types.

***Measuring clonal variability using DNA Typewriter for in cellular barcoding***

We reasoned that DNA Typewriter, as a synthetic barcode generator, as applied to lineage tracing above, could be repurposed for *in cellular* barcoding, which we could then leverage to measure inter-gastruloid variation that is compatible with scRNA-seq in which both transcriptomes and Tapes are simultaneously captured. Here we generated 150 clonal gastruloids in which DNA Typewriter is initiated for the first 48 hours during the first few cell divisions. This regimen establishes a unique barcode signature across the DNA Tapes per single cell that is then inherited by all progeny thereafter. At 144 hrs we dissociated gastruloids into single cells, and subjected them to the sci-RNA-seq platform in which we capture transcriptomes and DNA Tapes for all cells profiled. Here we obtained 247,000 high quality transcriptomes. Interestingly, we identify many new cell types, particularly anterior derived, that have not previously been seen in conventional mouse gastruloid protocols. For example, in addition to NMPs, spinal cord, PSMs, and somites, we identify early brain cell types such as hindbrain, midbrain, and motor neurons.

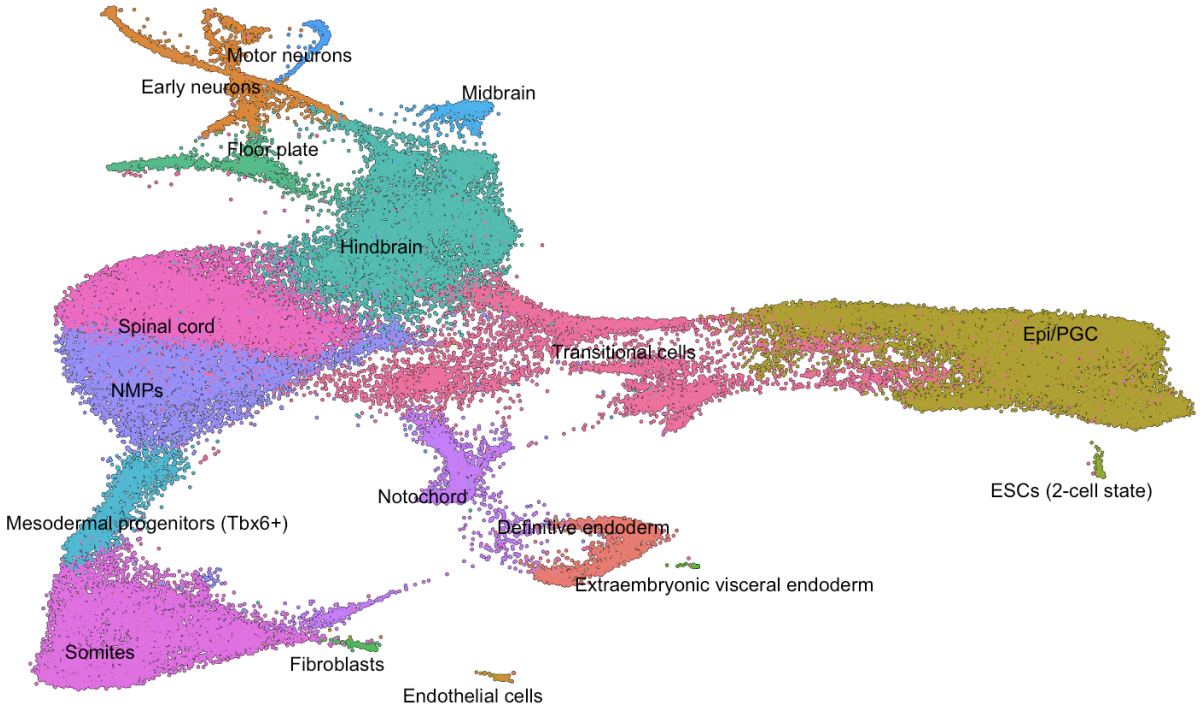



**Figure 7: DNA Typewriter for *in cellular* barcoding in 150 clonal gastruloids.** We profiled 150 clonal gastruloids, each comprising unique barcode compositions via DNA Typewriter, using sci-RNA-seq as a readout. Here we capture a variety of cell types en masse across 100 gastruloids which may not be reflected if only sampling 2 or 3 gastruloids.

## ***Discussion***

Here, we show that mosaic EBs, as a generalizable model for mosaic organoid systems, pose limitations when coupled to pooled CRISPR screens. For example, transgene silencing during differentiation constrains the population of cells with detectable synthetic transcripts, e.g. sgRNAs, which negatively impacts the workflow, requiring additional considerations, such as flow sorting, to achieve reasonable statistical power. Additionally we show that mosaic EBs suffer from bottlenecks, which likely results in biased measurements subsequent to skewed genotype representation that ensues during differentiation. These limitations inspired us to reconsider a new

approach to how we apply developmental *in vitro* multicellular systems to molecular phenotyping with single-cell readouts.

We reasoned that a clonal organoid system, in which each individual arises from a single precursor cell, barcoded upstream, would enable us to profile cells as a pool but then link them back to their original clone group or EB based on unique barcode compositions. We therefore developed a clonal EB generation pipeline and show that with piggyFlex, a transposon-based construct that stably co-expresses sgRNA and organoid barcode, we can robustly devise a pooled CRISPR screen with the added benefit of measuring inter- and intra-organoid variability. As a result, we can statistically account for any bias from inter-organoid variability in order to confidently assess cell type distribution changes that are predominantly driven by perturbations rather than stochastic skewing during the differentiation process.

Next, we sought to extend the clonal organoid concept beyond EBs, and into a system that better recapitulates transcriptional and morphological aspects of early mammalian embryos. For this, we pivoted to the stembryo model known as gastruloids. Here, we devised a new clonal gastruloid derivation protocol in which each gastruloid derives from a single-cell. With this developmentally relevant system, we engineered a mESC line that comprises a highly optimized 'DNA Typewriter' to reconstruct the lineage history of a gastruloid that began its developmental trajectory from a single ES cell. We show that high resolution trees are feasible to generate, as depicted by clear lineage relationships that arise from the lineage tree of the clonal gastruloid model. Moreover, we show that our protocol is scalable and amenable to combinatorial indexing, sci-RNA-seq, where

we capture inter-organoid variability (analysis in progress) using DNA Typewriter as an *in cellular* barcoding system.

Our work lays the groundwork for scalable clonal *in vitro* model synthesis in which we can reliably build lineage trees and imagine overlaying various perturbational pressures, genetic or environmental, with single-cell resolution.

# 3. MULTIPLEX PROFILING OF DEVELOPMENTAL CIS-REGULATORY ELEMENTS WITH QUANTITATIVE, SINGLE-CELL EXPRESSION REPORTERS

**Chapter 3 has been adapted with minimal modification from:**

> Lalanne JB*, Regalado SG*, Domcke S, Calderon D, Martin BK, Li X, Li T, Suiter CC, Lee C, Trapnell C, Shendure J (2023). Multiplex profiling of developmental cis-regulatory elements with quantitative single-cell expression reporters. *Accepted at Nature Methods*
>
> JBL and SGR are co-first authors.

In 2020, I closely collaborated with Jean-Benoît (JB) Lalanne on developing a scalable single-cell MPRA 'scQers' construct that could be ported into a multicellular system and measured at single-cell resolution. Our journey started as a mutual interest in dissecting the cis-regulatory code, leading to a formal collaboration inspired and supervised by Jay Shendure. Early on, I experimented with piggyBac cargoes, including optimal design tests for piggyFlex, a two-RNA expression system that would ultimately inspire the scQers construct. Additionally, I generated mouse embryoid body scATAC-seq data, which JB and I subsequently used to identify putative developmental enhancers to test in a QTL-inspired framework using CRISPRi (not discussed) and in key experiments to validate scQers in the multicellular mouse EB system. Although JB and I conceptualized scQers collaboratively, it was JB's innovative and tech-dev thinking that led to key innovations, such as the implementation of circular RNAs, that enabled highly quantitative measurements with scQers—a critical feature that paved the way forward for the work described in this chapter. JB carried out key proof-of-principle experiments in static cell lines, followed by a collaborative effort between us to decipher the autonomous function of putative enhancers in

mouse EBs. Our work together has been some of the most intellectually stimulating and joyous times in my scientific career.

Formal author contributions are as follows: JBL and SGR conceptualised dual reporters. JBL cloned scQer libraries, planned and carried out experiments in human cell lines and Pol III MPRA. SGR and JBL planned and carried out experiments in mEBs. JBL analysed data, generated figures, and wrote the manuscript with edits from JS and comments from SGR and DC. SGR generated scATAC data in mEBs. SGR and SD generated the mESC line, established mEBs protocols, and performed early profiling of mEBs. BM provided constructs, and protocols for cloning of MPRA cassettes. DC suggested analyses and provided computer scripts for subassembly. XL assisted with cloning of insulatorless piggyBac constructs. TL performed bioinformatic analyses on CREs. CCS provided starting protocols for library subassembly. CT and JS supervised the study.


## Abstract

The inability to scalably and precisely measure the activity of developmental *cis*-regulatory elements (CREs) in multicellular systems is a bottleneck in genomics. Here, we develop a dual RNA cassette that decouples the detection and quantification tasks inherent to multiplex single-cell reporter assays. The resulting measurement of reporter expression is accurate over multiple orders of magnitude, with a precision approaching the limit set by Poisson counting noise. Together with RNA barcode stabilisation via circularization, these scalable single-cell quantitative expression reporters (scQers) provide high-contrast readouts, analogous to classic *in situ* assays, but entirely from sequencing. Screening >200 regions of accessible chromatin in a multicellular *in vitro* model of early mammalian development, we identify thirteen (eight previously uncharacterized) autonomous and cell-type-specific developmental CREs, such as constituents of

the *Sox2* control region exclusively active in pluripotent cells, endoderm-specific CREs including near *Foxa2* and *Gata4*, and a compact pleiotropic CRE in the *Lamc1* locus. We further demonstrate that chimeric CRE pairs generate cognate two-cell-type activity profiles and assess gain/loss-of-function multicellular expression phenotypes from CRE variants with perturbed transcription factor binding sites. scQers can be applied in developmental and multicellular systems to quantitatively characterise native, perturbed, and synthetic *cis*-regulatory elements at scale, with high sensitivity and at single-cell resolution.

## ***Introduction***

Developmental *cis*-regulatory elements (CREs) direct programs of gene expression that unfold with remarkable cell type and spatiotemporal specificity. This tight control underlies the emergence of form and function from a one-cell zygote. Fine-scale regulatory changes in target gene expression, caused by even single nucleotide changes, can both give rise to disease[1–3] as well as drive evolutionary novelty[1,4]. How noncoding DNA encodes the requisite functional information remains incompletely understood even for the best-studied examples[5–8]. More broadly, biochemical marks correlated with enhancer status have now nominated >1M putative CREs in the mouse and human genomes[9]. However, functional profiling of these elements (and variants thereof) across diverse cellular states, particularly in developmental and multicellular contexts, is lagging due to the lack of scalable approaches.

In mammalian systems, most high-throughput functional studies of CREs have been performed in static contexts, typically cancer cell lines[10–13]. The scalability of these biotypes, in conjunction

with massively parallel reporter assays (MPRAs)[14–16] and related techniques[17], have enabled the characterization of complex CRE libraries, leading to accurate sequence-to-function models[11,18–20]. However, new experimental and modelling approaches are needed to extend beyond the scalar activity of cell lines and access dynamic, multi-cell-type regimes. Scalable reporters have been used in directed mammalian differentiation models (e.g., cardiac[21,22], hematopoietic[21,23], neuronal[22,24], naive to epiblast[25]) to discover developmental CREs, but these assays are usually applied to non-branching trajectories with limited cell type heterogeneity. Until now, work on CREs in multicellular systems has predominantly been carried out with transgenic reporters assayed via *in situs*[26–28], approaches which remain semi-quantitative and of limited throughput even with automation[29]. Nonetheless, even at limited scales, these studies reveal the rich phenomenology of metazoan developmental CREs, namely that kilobase-sized DNA sequences can autonomously recapitulate the complex expression patterns of their target genes even when taken out of context.

Two recent innovations are poised to improve the throughput of mammalian regulatory biology in multicellular systems. First, stem-cell-derived models of increasing sophistication, including organoids, gastruloids, and synthetic embryoids[30], enable the scalable delivery of reporters[31] prior to differentiation. Second, single-cell genomics can map cellular states and in principle be combined with multiplex reporter assays to profile CREs in multicellular models (**Fig. 1a**). However, in practice, multiplex reporters in single-cells pose a fundamentally new challenge compared to bulk modalities: in order to measure the activity of any given candidate CRE, one must first determine which reporters are present in which profiled cells. As such, in porting the 'one-RNA' reporter strategy of traditional MPRAs directly to single-cell platforms (**Fig. 1b**), one

relies on the barcoded mRNA for both: 1) per-cell reporter detection; and 2) quantification of expression driven by the candidate CRE. The detection task is challenging for lowly expressed reporter transcripts due to chimeric amplicons (i.e., amplification products spuriously swapping barcodes originally from different molecules), which increase noise in single-cell libraries[32,33]. As such, the simplest adaptation of MPRAs to single-cell assays cannot distinguish between cells in which a given reporter is not expressed vs. cells in which a given reporter is not present (**Fig. 1b**). This confounds the accurate quantification of reporter expression.

To resolve this problem, we developed a dual RNA reporter which separates the detection and quantification tasks (**Fig. 1c**). For reporter detection, we introduce circularized[34] Pol III transcribed barcodes which enable near-complete recovery of the identity of the reporter(s) present in any given cell from single-cell RNA-seq data (scRNA-seq). We demonstrate that these single-cell quantitative expression reporters (scQers) are accurate over multiple orders of magnitude despite the sparsity of scRNA-seq and enable the discovery of lineage-specific regulatory elements with high sensitivity. We anticipate that scQers will enable the scalable, quantitative characterization of CREs in multicellular models of development and otherwise heterogeneous samples.

## ***Results***

### *Decoupling detection and quantification with dual reporters*

We reasoned that detection and quantification could be decoupled via two separate barcoded RNAs linked on individual reporters (**Fig. 1c**). One barcoded RNA, highly and constitutively expressed, serves as the marker for presence/absence of the integrated reporter within any given cell. The

second RNA, a Pol II-expressed mRNA barcoded (hereafter mBC) in its 3' UTR, serves to quantify CRE activity similar to a bulk MPRA reporter. Provided that the two barcodes are *a priori* matched to one another, as well as to distinct CREs, one can separately detect and quantify the activity of reporters in single-cell assays.

Dual RNA reporters require the contiguous production of two separate RNAs. Given that Pol II promoters can act as enhancers[35], we expressed the detection barcode from a Pol III promoter. Interactions are expected to be minimal as a result of the largely orthogonal Pol III and Pol II machineries[36]. To avoid transcriptional collisions[37,38], our reporter architecture (**Fig. 1c**, **Ext. Data Fig. 1a**) places the hU6-driven detection barcode co-directionally upstream of the quantification cassette, which has the CRE immediately upstream of a minimal promoter (allowing for both measurement of enhancer activity and possible enhancer RNA production).

To mitigate the instability of short ectopic Pol III RNAs[39], we embedded the constitutively expressed barcode within the 'Tornado' circularization system[34] (**Ext. Data Fig. 1g-h**). The resulting circular RNA barcodes, hereafter Tornado barcodes (oBC), were expressed >150-fold more highly than their linear equivalent (**Ext. Data Fig. 1g-k**, data from genome-integrated bulk MPRA, minimal impact of random oBC sequence with ≤2.6-fold interquartile range), reaching an estimated >75,000 oBC RNA per cell per cassette[34].

***Benchmarking with a promoter library in human cell lines***

The scQers cassette is defined by three components delivered to cells as a single unit: a detection oBC, a CRE, and a quantification mBC. We first established that scQers report transcriptional expression in single-cells with ≈2% dropout, high accuracy over a large dynamic range ($<10^{-1}$ to $>10^3$ UMI/cell), and high precision (coefficient of variation <1). To do so, we constructed a minimal library of five Pol II promoters spanning a wide activity range[40] (**Fig. 2a**, **Supp. Data 1**), and integrated the payloads by piggyBac[41] transposition at high multiplicity of integration in three human cell lines (HEK293T, HepG2, K562, median MOI of 4, 7, and 6 respectively). Cells were bottlenecked to a few hundred clones, expanded, and then both: 1) hand mixed at 1:1:1 ratios and profiled via scRNA-seq (10x Genomics 3' feature barcoding with optimization, **Ext. Data Fig. 1b-f**); and 2) harvested separately for bulk MPRA (**Fig. 2a**). Thousands of cells per replicate passed standard quality filters, with cell line identity unambiguously mapped from gene expression (**Fig. 2b**, **Ext. Data Fig. 2a**).

***oBCs are near-deterministically retrievable in scRNA-seq***

oBCs were robustly captured on a per-cell basis. In particular, the distribution of oBC unique molecular identifier (UMI) counts displayed bimodality (**Fig. 2c**, **Ext. Data Fig. 2b**) and >30× signal-to-noise. The low count mode corresponds to chimeric amplicons, and the high count mode to expression from valid integration events (≈2500 UMI/cell per barcode, zero-truncated Poisson estimator). To assess oBC dropout, we leveraged redundant measurements across clones (**Fig. 2d**). Consensus integration clonotypes were identified in the bottlenecked population by relying on oBC co-detections[42,43] (**Fig. 2e**, **Ext. Data. Fig. 3a-f**, **Supp. Data 2**). Clonotypes served as ground-truth for precision-recall analysis of detected oBCs in clone-assigned cells, revealing a false negative rate (dropout) of <2% at a false discovery rate of 1% (**Fig. 2h**, **Ext. Data Fig. 3e-f**). This

represents a >10-fold improvement vis-a-vis capture of sgRNAs in single-cell CRISPR screens[43]. In sum, oBCs are transcribed barcodes which nearly eliminate dropout in scRNA-seq.

The high expression of oBCs raises the question of toxicity to cells. In line with original assessments[34], we find little correlation between total oBC RNA expression and markers of apoptosis or immune response (e.g., percent mitochondrial content $R^2<0.03$, p53 expression $R^2<0.02$, RIG-I expression $R^2<0.003$) both in cell lines and mEBs (experiment below).

***Accurate reporter quantification over orders of magnitude***

Comparing reporter expression from single-cell and bulk quantification confirmed the accuracy of scQers. Following detection of reporter integration using oBCs (probability of multiple integrations per cell from the same oBC-promoter-mBC triplet <5%), activity of the associated promoters can be quantified in each cell as the transcriptome-normalised average UMI counts from the matched mBC (**Fig. 2f**, **Ext. Data. Fig. 2c**). Single-cell averaged UMI counts across the different mBCs associated with a given promoter constituted independent measures of activity and spanned over four orders in magnitude for the five promoters (**Fig. 2g**, **Ext Data. Fig. 2d-f**). Bulk MPRA measurements performed on the same cell populations were concordant across the full range of expression levels ($R^2$ log-transformed expression ≥0.87, **Fig. 2g**, **Ext. Data Fig. 2d**). Single-cell measurements of mBCs from as few as 5-10 cells sufficed for accurate quantification (**Ext. Data Fig. 2g**).

Without filtering, spurious read counts can alter reporter quantification. Indeed, library preparation requires a number of amplification steps that can generate 'chimeric' amplicons and lead to erroneous cell-to-barcode connections. In saturated libraries, the signature for these molecular products is a rising frequency of counts below ≈10 UMI/cell (e.g., oBC: **Fig. 2c**, mBC: **Ext. Data Fig. 2e**) which can result in a limit of detection substantially higher than 1 UMI/cell. A dual RNA approach does not abrogate chimaeras, but filters mBC reads based on detection of a matched oBC in the same cell, leading to an average decrease in the tallying of chimeric counts by the proportion of cells harbouring any given oBC-mBC combination. Consequently, lowly expressed mRNAs driven by the minimal and no promoter basal controls (median expression of ≈0.2 UMI/cell below the 1 UMI/cell regime inaccessible from pooled one-RNA reporters, **Fig. 2g**) remained accurately quantified by scQers, suggesting limited zero-inflation[44] in our system. Leveraging our *a priori* matched oBC-mBC pairs, we found a high prevalence of chimeric mBC detections (mBC found in cells without a detected matched oBC: 90% EEF1A1p, 60% Pgk1p, 51% UBCp, 36% no promoter, 52% minimal promoter). As a result, quantifying activity based on Pol II mBC alone (no conditioning on oBC detection) led to biases and increased variability ($R^2$=0.39 for log-transformed single-cell vs. bulk; 1.5 to 25-fold increased variability, **Ext. Data Fig. 2h-i**), highlighting the quantitative advantage of dual RNA reporters.

***Measurement precision approaching Poisson counting noise***

Our clonal pool of cells further allowed us to quantify variability in mBC capture. Multiply represented clones provide internal replicate measurements of the same set of reporters integrated at fixed genomic locations, controlling for an important source of variation from random integration[45–47] (**Fig. 2d**). For a given reporter (mBC) integrated in a specified clone, each clonal

representative sampled provides a measurement of the number of captured reporter mRNA molecules. Clones with multiple cells detected therefore enable sampling of the experimental distribution of the number of mBC UMIs per cell (**Ext. Data Fig. 3g-h**, bottom panels). The variance of this distribution of mBC UMIs can then be determined, providing an estimate of the measurement precision. The minimal variance is expected to be set by Poisson counting noise, reflecting the nature of the measurement as a discrete sampling, with any additional variance corresponding to biological or technical variability.. Across all reporters and clones, we find variability consistent with Poisson counting noise at low expression, and a coefficient of variation substantially below one for two of the promoters (UBCp and EEF1A1p, **Fig. 2i**, **Ext. Data Fig. 3i**). The UBCp promoter in particular displayed detection close to the Poisson scaling (standard deviation/mean = 1/√mean). Variability was not strictly correlated with average expression. For example, the Pgk1p promoter, while expressed more highly than UBCp, exhibited substantially higher cell-to-cell variability (**Ext. Data Fig. 3i**). scQers thus precisely measure reporter mRNA levels in single cells.

Systematic assessment of reporter expression across clones provided estimates of variation due to positional effects (**Supp. Note 1**, **Ext. Data. Fig. 3j**). While insulators[48] in our construct (**Ext. Data Fig. 1a**) substantially reduced context dependence (**Supp. Fig. 1**, **Supp. Data 3**), 41-60% of mBC UMI variability in mBC UMI counts remained attributable to positional context, further confirming the technical precision of our per-cell measurement and the importance of averaging over multiple integration positions.

***Locus-level screen of putative developmental CREs***

Following optimization in cell lines, we sought to apply scQers to discover cell-type-specific CREs in an *in vitro* model of early mammalian development, mouse embryoid bodies[49,50] (mEBs). We drew putative CREs for testing from the neighborhood of prioritized developmental loci (**Fig. 3a-b**). First, by profiling 21-day differentiated mEBs with scRNA-seq and single-cell ATAC-seq[51,52] (scATAC-seq), we established the transcriptional and chromatin accessibility states of various cell types (**Ext. Data Fig. 4**). scATAC-seq data from mEBs was highly correlated to *in vivo* data from matched cell types in E7.5-E8.5 embryos[53] ($R^2$ log-transformed accessibility across top 65k mEB peaks: e.g., parietal endoderm=0.77, neuroectoderm=0.78, mesoderm=0.76), supporting mEBs as a model of gene regulation in early development. Leveraging these data, we nominated 22 developmental genes with germ-layer specific expression and cell-type-specific chromatin accessibility landscapes (**Supp. Data 1**) such as endoderm regulator *Gata4*[54], other lineage-defining transcription factors (*Klf4*, *Foxa2*, *Sox17*), and structural genes (laminins, collagens, tubulin). As a comprehensive set[55] of CREs to profile from these genes, we selected all regions within ±100 kb of their TSS that were reproducibly highly accessible in the expression-cognate cell type (e.g., 13 putative CREs near *Gata4* in **Fig. 3a**, other examples: **Fig. 4a**). As positive controls, we additionally included the four constituents of the core *Sox2* control region[56,57] (**Supp. Data 4**), accessible exclusively in pluripotent cells (**Fig. 3e**). In total, 209 elements were included for profiling (145/209 promoter-distal >1 kb from promoters[58], median element size: 937 bp, 893/956 bp 25$^{th}$/75$^{th}$ percentiles, **Supp. Data 1**). The five exogenous promoters (same as **Fig. 2a**) were also spiked-in as standards. Following library construction and sequential subassemblies (**Supp. Fig. 2**, 204/209 CREs represented with >20 oBC-mBC pairs, 88/145/242 10$^{th}$/50$^{th}$/90$^{th}$ percentile number of valid oBC-mBC pairs per CRE), scQers were integrated in mESCs at high

MOI using piggyBac[59,60] (**Ext. Data Fig. 5c-d,** median MOI = 23, per-cell probability of oBC-CRE-mBC triplet being integrated more than once=1%). Reporter-integrated cells were induced to form mEBs, sampled every 2 days for bulk MPRA quantification across differentiation, and scQered at the three weeks end-point (**Fig. 3b**).

***High performance in a stem-cell derived developmental system***

mEBs reproducibly comprised diverse cell-types unambiguously mappable to *in vivo* germ-layers[61] (**Fig. 3c**, n=43799 pass-filter cells across three biological replicates [replicates 1 and 2: separate transfections; replicate 2B: ~500-clones bottleneck of replicate 2 with 12% identified clonotypes overlap to replicate 2, and thus largely orthogonal; all replicates separate mEB inductions], **Ext. Data Fig. 5e**), confirming successful differentiation despite the presence of reporters at high MOI.

scQers displayed high performance in mEBs. First, oBC were robustly captured (median library complexity=836 UMI/oBC/cell), displaying a bimodal distribution of oBC UMI/cell (**Ext. Data Fig. 5f**). oBC expression was cell-type independent (**Ext. Data Fig. 5g**), enabling uniformly high recovery (<4% oBC dropout at FDR=1% from precision-recall analysis of clonal cells, **Ext. Data Fig. 5i-k**). Second, comparison of end-point bulk and single-cell quantification across profiled CREs confirmed accuracy of reporter expression measurement over the full dynamic range ($R^2$ log-transformed activity=0.81, **Ext. Data Fig. 5a**) and per-cell-type quantification was reproducible ($R^2$ log-transformed across replicates=0.72, **Ext. Data Fig. 5b**). Representation was reasonably uniform across tested CREs (**Ext. Data Fig. 5h**, captured integration events per

element 1597/3153/6197 $10^{th}$/$50^{th}$/$90^{th}$ percentiles, and n=17971 to 34745 for exogenous promoters).

***Single-cell expression maps from Sox2 control regions***

scQers generated high-contrast single-cell maps of CRE activity (**Ext. Data Fig. 6a-b**). As a case study, we considered gene expression control of the pleiotropic regulator *Sox2* (**Fig. 3d**). *Sox2* is a key factor in pluripotency maintenance[57]. Central to *Sox2* control is a distal (≈135 kb from TSS) cluster of CREs necessary for driving high expression in pluripotent cells[56,57], previously shown to function autonomously[57,62]. Of four differentially accessible elements in pluripotent cells from this control region (**Fig. 3e** inset), two displayed robust activity (red **Fig. 3f**, 10-30-fold higher expression vs. basal controls), in agreement with previous characterisation[7,57] (**Ext. Data Fig. 6d**, **Supp. Data 4**) circumscribed to the pluripotent population (e.g., >50-fold higher expression vs. other cell types for *Sox2*:chr3_2007). While *Sox2* was expressed in the pluripotent and ectoderm lineages in mEBs (**Fig. 3d**), CREs from *Sox2* control regions were exclusively active in pluripotent cells (*Essrb*/*Dppa3*-positive[63], **Ext. Data Fig. 4b**). Our results on this previously characterised cluster of regulatory elements confirm that scQers can report cell-type-specific expression in a multicellular system with high sensitivity and contrast. scQer experiments on six additional literature-selected cell type-specific CREs[64–67] further confirmed the robustness of our approach (3/6 with expected activity profiles, 3/6 inactive in mEBs, **Supp. Fig. 3h-i, Supp. Data 4**).

***Systematic identification of active CREs***

We also quantified both activity and cell-type specificity of other tested candidate CREs (n=200), identifying multiple active elements (**Fig. 4a, Ext. Data Fig. 7**). For each CRE, average reporter expression was determined across cells with detections, stratified by cell-type. Activity was defined as the maximum per-cell-type reporter expression, while specificity was taken as the maximum per-cell-type mBC expression divided by the mean expression in all other cells (**Fig. 4a**). We identified 58/204 endogenous CREs with activity in significant excess of the basal controls in all three replicates (**Supp. Data 5**). The elements with the highest expression were the active exogenous promoters (UBCp, Pgk1p, EEF1A1p) at ≈30 to 250 mBC UMI/cell, (levels ≈300× to ≈2500× above basal controls, **Fig. 4a**). Active endogenous CREs spanned a wide range at lower expressions (maximum per-cell-type expression: ≈0.3 to 20 mBC UMI/cell, **Fig. 4a**). Notably, a sizable fraction (19/58) of the active CREs had expression under 1 mBC UMI/cell, and most were below the chimeric read threshold of 10 UMI/mBC/cell, underscoring the usefulness of a high-sensitivity method.

Active CREs displayed distinct expression patterns across mEB cell types. Categorising active CREs as cell-type specific vs. non-specific (permutation test), we found 10/58 developmental CREs with reproducible cell-type-specific activity (red in **Fig. 4a-c**, **Ext. Data Fig. 8a-d**). Singleton validation experiments on the eight most specific CREs confirmed that the elements drove cell-type specific expression (**Supp. Fig. 4-5**). Of the remaining 48 non-specific active elements, 41 (85%) were promoter-proximal (e.g., orange **Fig. 4e**, **Ext. Data Fig. 8d**) compared to 0/10 of cell-type-specific CREs. Conversely, 41/62 tested promoter-proximal elements were found to be active and non-specific (while 0/62 were cell-type specific). Consistent with their

function and distance from TSS, all cell-type-specific CREs showed >10 fold-change in chromatin accessibility in their cognate cell types; in contrast, promoters were constitutively open (<3 fold-change, **Fig. 4f**). Notably, accessibility (rather than change in accessibility) was a poor predictor of activity or specificity (**Ext. Data Fig. 8e**), in line with evidence of the imperfect correspondence between accessibility and function for regulatory elements[55,68]. Single-cell activity maps thus delineated two broad patterns of autonomous function: constitutively active elements (overwhelmingly TSS-proximal, broadly accessible) and cell-type-specific elements (overwhelmingly TSS-distal, differentially accessible).

Our assay relies on high MOI random integration of reporters for scalable multiplexing, raising concerns that genomic positional effects might dominate the signal[45,46]. To assess positional effects, we bottlenecked reporter-integrated mESCs to a few hundred clones in one of the replicate (replicate 2B) prior to mEB induction. Quantifying activity of the 10 cell-type-specific CREs across well-represented clones, we found that most CREs (9/10) retained specificity (>5-fold) across the super-majority (>⅔) (**Supp. Fig. 6**, **Supp. Data 6**), suggesting that positional effects can be averaged over.

***Characterization of lineage-specific, autonomous CREs***

Of the 10 autonomous cell-type-specific CREs identified, two belonged to the core *Sox2* control region (**Fig. 3f**), while the remaining 8, all from distinct parietal endoderm-expressed loci (red **Fig. 4e**, **Ext. Data Fig. 8d**), included a *Gata4* intronic CRE 10 kb downstream of the first exon (chr14_5729, **Fig. 4e** second row) and an CRE 70 kb upstream of *Foxa2* (chr2_13858, **Fig. 4e**

third row). One active element at the *Lamc1* locus (chr1_12189, **Fig. 4e** fourth row) was found to be active in two cell types, with concordant chromatin bi-accessibility (inset **Fig. 4b** fourth row). Identifying mostly endoderm-specific CREs was not unexpected given the uneven sampling of tested elements due to the high proportion of endoderm cells in the scATAC data.

Reporter expression driven by developmental CREs mirrored the predominant pattern of expression of their nearby putatively associated gene (**Fig. 3d** vs. **3f**, **4d** vs. **4e**, **Ext. Data Fig. 8c** vs. **8d**), except for the bi-functional putative *Lamc1* CRE (**Fig. 4d** fourth row, black caret), which drove expression in both parietal endoderm and pluripotent cells, in contrast with endogenous *Lamc1* whose expression was restricted to parietal endoderm. For endoderm-specific CREs, the magnitude of activity induction was on par with endogenous gene induction (**Ext. Data Fig. 8f-g**, **Supp. Note 2**).

Leveraging our time-resolved bulk MPRA (**Ext. Data Fig. 9**, **Supp. Data 7**) on the same samples, we found a consistent set of active CREs (53/54 bulk active elements identified as active from scQers, 53/58 scQers identified elements found as bulk active). Importantly, elements found to be cell-type-specific with scQers displayed either temporal increase (red **Ext. Data Fig. 9d**), decrease (core *Sox2* control region, **Ext. Data Fig. 6c**), or non-monotonic behaviour (bifunctional CRE, *Lamc1*:chr1_12189, **Ext. Data Fig. 9d**), supporting their classification as developmental regulatory elements. In contrast, active but non-specific elements displayed little temporal variation across differentiation (e.g., exogenous promoters and endogenous elements, orange **Ext. Data Fig. 9c-d**), as expected for constitutive, promoter-like, CREs. A number of CRE features

(e.g., accessibility, number of transcription factor binding sites, **Supp. Note 3, Supp. Fig. 7**) correlated with measured activity.

Overall, scQers enabled the scaled high-sensitivity characterization of both constitutive promoter-like and lineage-specific autonomously active regulatory elements across diverse cell types of 21-day mouse EBs, with CRE activity profiles matching expression of their putatively associated genes. Additional experiments with synthetic pairs of CREs and elements with optimised/disrupted transcription factor binding sites (**Supp. Note 4**, **Extended Data Fig. 10**, **Supp. Fig. 3**, **Supp. Data 8-9**) confirmed the usefulness of scQers to study regulatory elements.

***Influence of reporter architecture on expression output***

scQers rely on a Pol III cassette in proximity to the Pol II promoter driving reporter mRNAs, raising concerns of interference between the two. To assess possible interaction, we constructed libraries with and without the U6/oBC cassette harbouring the same putative CREs and promoters (**Supp. Fig. 8a, Supp. Data 10**), integrated the reporters in mESCs, differentiated the cells to embryoid bodies, and performed bulk MPRA at various time points. The measured expression driven by the CREs were highly concordant with vs. without the Pol III cassette both for promoters and CREs (**Supp. Fig. 8b-c**-, $R^2$ of log-transformed activities >0.84). Importantly, temporal induction of the cell type-specific CRE did not depend on the presence of the U6-driven RNA (**Supp. Fig. 8d-e**). While these data do not exclude possible interference from the Pol III in all contexts, they suggest that such influence is of limited magnitude for scQers.

Given our reporter architecture, with the CRE directly upstream of the minimal promoter, we also sought to assess whether the measured mBC counts derived from enhancer RNAs (eRNAs[69]) or from initiation at the minimal promoter. To do so, we tested expression from reporters with and without the minimal promoter, as well as constructs placing the CREs downstream (**Supp. Fig. 8a**). Surprisingly, we found little difference in the measured expression comparing reporters with and without minP (**Supp. Fig. 8f-g**), suggesting either cryptic transcription initiation (analogous to transcription initiation within the bacterial ORI in the original STARR-seq assay[70]), or initiation within the CREs themselves (i.e., eRNAs). In addition, although positioning CREs downstream of the reporter cassette compressed the dynamic range of expression (**Supp. Fig. 8h**), in line with previous systematic comparison of different MPRA architectures[13], induction was detectable in 7/13 expected cases (Bonferroni-corrected Wilcoxon test $p<0.05$, **Supp. Fig. 8i-j**), consistent with some of the identified CREs having enhancer activity. Given the possible distance dependence of functional expression outcome to CRE positioning, more experiments will be needed to fully ascertain the molecular origin of the measured mBCs. Despite the prevalence of the CRE-minP-reporter architecture for MPRA assays[15,71–74], there exist no 5' end mapping data to our knowledge in that context. As such, our results draw an important distinction between reporter and enhancer assays. While this does not undermine the unique advantages of scQers to identify elements driving cell-type specific mRNA production, researchers seeking to unambiguously measure enhancement of transcription initiation at a specified site should insulate the enhancer from the promoter, or consider alternative architectures.

***Discussion***

CREs orchestrate the precise unfolding of development in metazoans, enabling the emergence of a species' form and function from a genomic blueprint. However, our ability to study

developmental CREs at scale has been constrained, particularly in mammalian systems. We and others[75–77] have recognized that a simple path forward is to intersect MPRAs with single-cell resolution technologies. Here we overcome key technical challenges of combining these two modalities, resulting in scQers, an MPRA that decouples the detection and quantification of reporters via a dual RNA system and circularization-based enhancement of barcode recovery. scQers extend measurements into a regime fundamentally inaccessible with traditional multiplex reporters, yielding an accurate, precise and high-contrast readout of reporter mRNA levels. Beyond reporter assays, the use of oBCs, and Tornado-based stabilisation more generally, may be of broad utility for robust capture in single-cell and other genomics applications ranging from CRISPR screens to cell lineage tracing.

The relatively low hit rate of our screen (8/200 cell-type specific) suggests that random genome integration followed by differentiation provides a strong filter for elements autonomously competent to reconfigure chromatinized landscapes and drive expression. In addition, lack of activity might be a consequence of our use of a minimal promoter, as opposed to *bona fide* developmental promoters. Recent systematic studies have found promoter choice to be important in scaling the response of regulatory elements[78–80]. Beyond these technical differences, given the complex multi-CRE landscapes considered here, some tested CREs might contribute to regulation, but only in the presence of (or by directly serving as) cooperating elements, in line with recently described facilitators[8] or chromatin-dependent enhancers[11] (e.g., tested but inactive *Sox2*:chr3_2005, which overlaps with facilitator DHS23[7]). While most elements identified here display expression patterns mirroring that of their putatively associated gene, in-genome perturbations will be necessary to confirm their role, if any, in regulation. As they become broadly

available, high-resolution enhancer-to-promoter contact maps[81,82] could be used to prioritise CREs and further strengthen conclusions drawn from reporter measurements.

How many regulatory elements can be profiled with scQers? Based on current measurements, we estimate that 100 detections per CRE per cell-type would robustly detect expressions of 1 UMI/mBC/cell. The number of single-cells that need to be profiled per replicates per CRE is thus estimated to be 100×(# cell types)/MOI (**Supp. Note 5**). The majority of the costs remain on the single-cell assay if using existing commercial droplet-based approaches. With continuous improvement in capture from alternatives, e.g. single-cell combinatorial indexing[83], we anticipate that >10-fold improvement in throughput will soon be achievable.

As predictive models of regulatory activity improve[11,18,19,84–86], quantitative experimental approaches are needed to iterate through design-test-learn cycles and validate underlying mechanistic hypotheses. Benchmarks in cell lines, a proof-of-principle screen in a multicellular stem-cell model, and experiments on synthetic pairs and mutated CREs, establish scQers as a scalable platform for probing gene regulation that should be portable to other developmental systems (e.g., zebrafish[87], *C. intestinalis*[27], the chicken neural crest[88], synthetic embryoids[89,90], *in vivo* neuronal subtypes with AAV derivatives[91]). Although established here with a focus on developmental biology, we envision scQers may also facilitate the identification, optimization, and compactification of highly active cell-type-specific CREs for application in gene therapy and other practical uses[92,93].

**ACKNOWLEDGMENTS:** We thank N. Ahituv, M. Kircher, R. Ziffra, G. Gordon, A. Ellis, J. Tome, and the entire Shendure lab for discussions; participants of the gene regulation subgroup (F. Chardon, W. Chen, T. McDiarmid) for criticisms and advice; T. McDiarmid for noting the high instability of short ectopic Pol III RNAs; CX Qiu for advice on single-cell data annotation; E. Nichols and V. Browning for assistance with the BZ-X810; M. Gailey and D. Miller from the UW Nanopore sequencing core for expert assistance. Plasmid pAV-U6+27-Tornado-Broccoli was a kind gift from S. Jaffrey (Addgene plasmid # 124360). This research is supported by research grants from the National Human Genome Research Institute (NHGRI; UM1HG011966 to JS, R01HG010632 to JS and CT). JBL is a Fellow of the Damon Runyon Cancer Research Foundation (DRG-2435-21). SGR was supported by the NHGRI (F31HG011576). DC was supported by the National Heart, Lung, and Blood Institute (T32HL007828) and NHGRI (F32HG011817). JS is an Investigator of the Howard Hughes Medical Institute.

**AUTHOR CONTRIBUTIONS:** JBL and SGR conceptualised dual reporters. JBL cloned scQer libraries, planned and carried out experiments in human cell lines and Pol III MPRA. SGR and JBL planned and carried out experiments in mEBs. JBL analysed data, generated figures, and wrote the manuscript with edits from JS and comments from SGR and DC. SGR generated scATAC data in mEBs. SGR and SD generated the mESC line, established mEBs protocols, and performed early profiling of mEBs. BM provided constructs, and protocols for cloning of MPRA cassettes. DC suggested analyses and provided computer scripts for subassembly. XL assisted with cloning of insulatorless piggyBac constructs. TL performed bioinformatic analyses on CREs. CCS provided starting protocols for library subassembly. CT and JS supervised the study.

**COMPETING INTERESTS:** JS is a scientific advisory board member, consultant and/or co-founder of Cajal Neuroscience, Guardant Health, Maze Therapeutics, Camp4 Therapeutics, Phase Genomics, Adaptive Biotechnologies, Scale Biosciences, Sixth Street Capital and Pacific Biosciences. All other authors declare no competing interests.

***Figures*:**

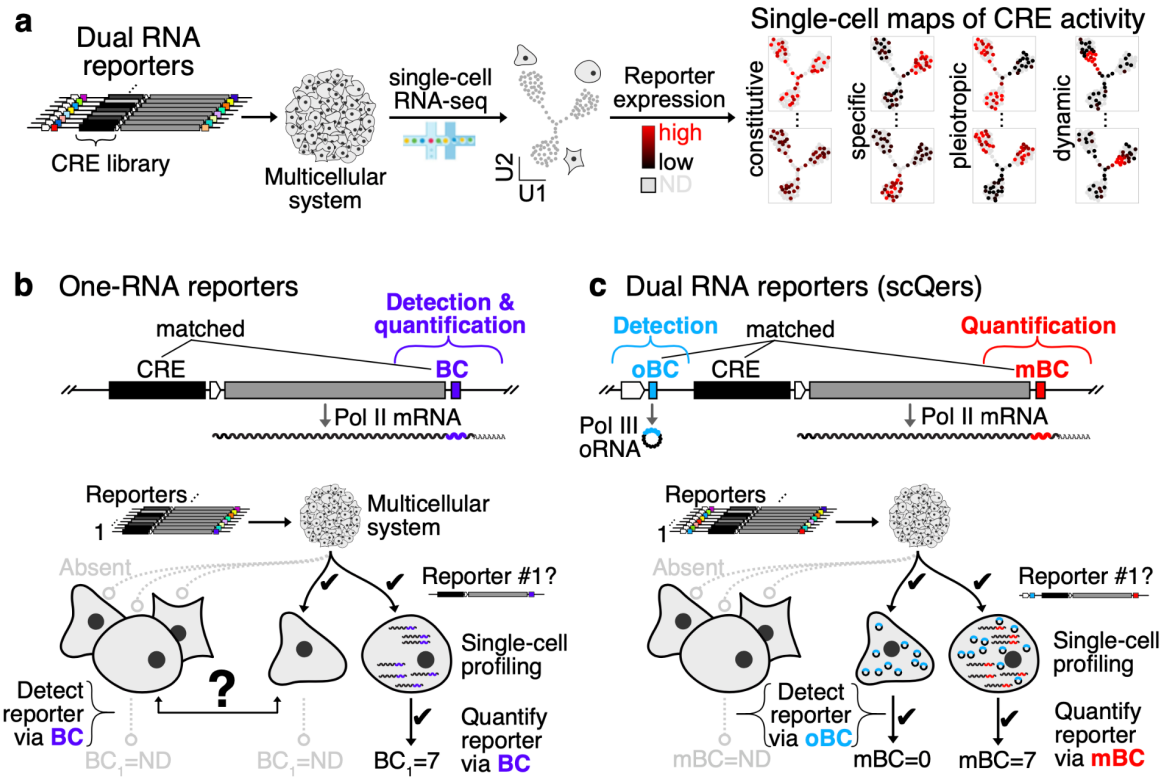



**Figure 1: High-contrast single-cell CRE activity maps with single-cell quantitative expression reporters (scQers)**

**a** Multiplex single-cell reporter assays. Introduction of complex libraries of integrating reporters to multicellular systems followed by scRNA-seq and computational deconvolution of reporter expression.

**b** Traditional multiplex reporters harbour a single barcoded Pol II mRNA (BC, purple) driven by a library of CREs whose activity is to be profiled. In a multiplex single-cell assay, having a single transcript to both detect presence of any given reporter in a profiled cell and measure expression level is biassed. In the extreme case where no mRNA is produced from a CRE in a given cell type, direct detection of the reporter is not possible (left group vs. middle cell).

**c** To resolve this dropout problem, a constitutively and highly expressed Pol III-derived circularised barcoded RNA[34] (Tornado barcodes, oBC, blue), *a priori* matched with the mBC (red) and CRE, is appended co-directionally upstream in a dual RNA cassette. The oBC enables robust detection of reporters in single cells, independent of reporter activity, enabling unbiased measurement of mBCs from the CRE-driven reporter mRNA.

See also **Ext. Data Fig. 1**.

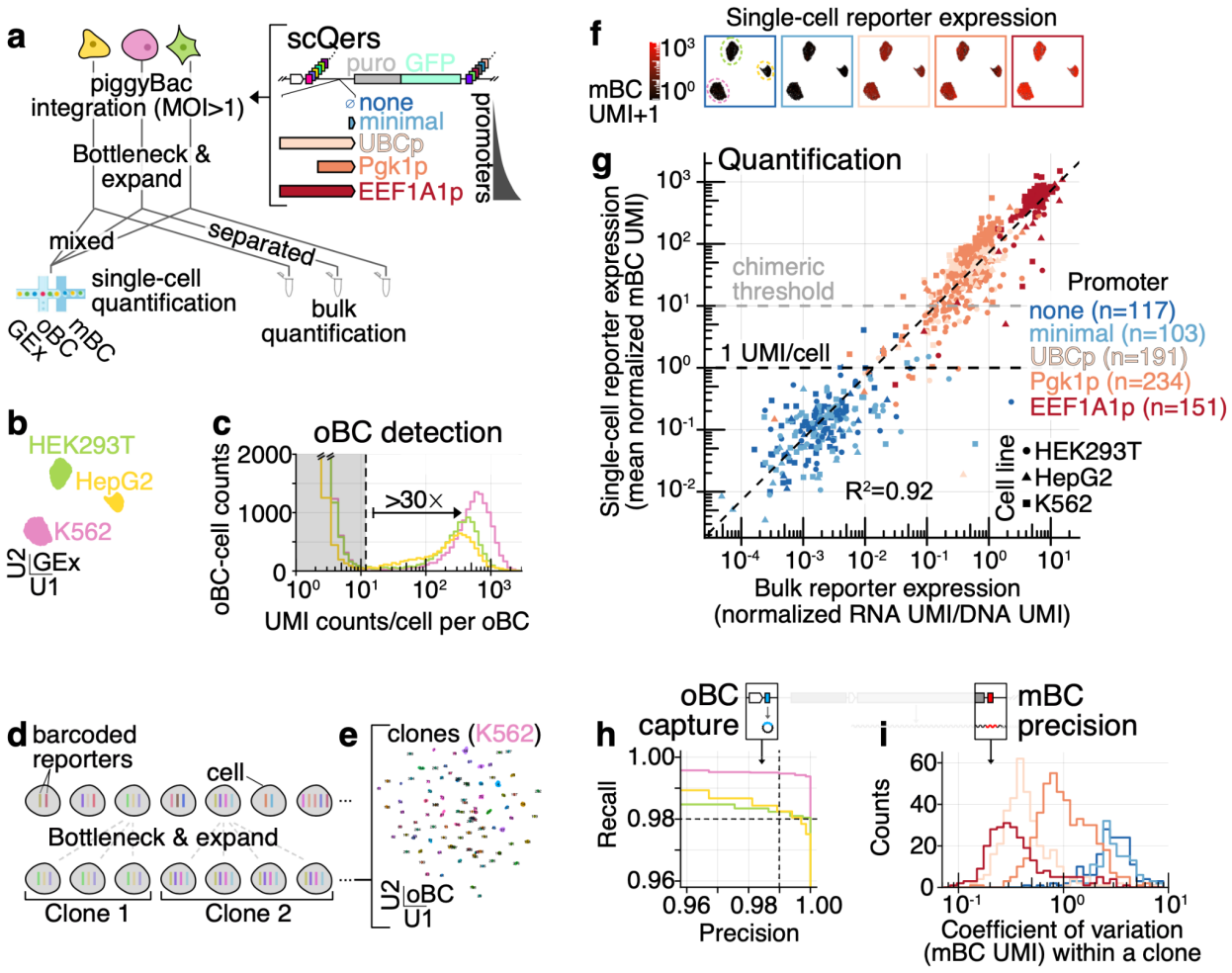



**Figure 2: Benchmarking scQers for accuracy, precision, and capture in human cell lines**

**a** A scQer library of five promoters (n=1122 unique oBC-promoter-mBC triplets, median 205 mBC-oBC pairs per promoter) was integrated in three human cell lines (HepG2, K562, HEK293T) at high multiplicity via piggyBac. Following integration, bottlenecking and expansion, clonal cells were: 1) separately subjected to bulk MPRA; and 2) mixed at 1:1:1 ratio and single-cell profiled.

**b** UMAP projection of quality filtered single-cell transcriptomes . The three well-separated clusters correspond to the three cell lines (replicate A; cell count: K562 n=2184, HEK293T n=2090, HepG2 n=1231).

**c** Bimodal distribution of the UMI counts per oBC per cell, stratified by cell line (low count mode, truncated, grey shading: chimeric amplicons; high count modes: *bona fide* integrations).

**d** Clonally derived cells with a high multiplicity of reporter integrations provide internally controlled replicates of the same measurement for assessing capture of oBC and precision of mBC quantification.

**e** UMAP projection (oBC expression space) for high-confidence-assignment cells assigned to clonotypes for K562 (replicate A; n=1430 cells, n=105 clones).

**f** UMAP projection cells coloured by promoter activity (average normalised mBC UMI count per cell, with pseudocount of 1). Each panel corresponds to a different promoter.

**g** Comparison between the single-cell mBC quantification (y-axis: average normalised mBC UMI over all cells with detected matched oBC) and bulk MPRA quantification (x-axis, RNA over DNA normalised UMI counts). Each point corresponds to an individual mBC (colour: promoter, symbol: cell line). Well-represented mBC are included (>100 bulk DNA UMI, >0 mBC single-cell UMI , and ≥5 single-cell integrations).

**h** Precision-recall curves for retrieval of oBC from cells assigned to clones (consensus clonotypes taken as ground truth, aggregate over all clones with >2 cells ; K562: 195 clones, 2168 cells; HEK293T: 173 clones, 2019 cells; HepG2: 38 clones, 1453 cells). Dashed lines: 99% precision (1% FDR), and 98% recall (2% false negative rate, or dropout).

**i** Distribution of the coefficient of variation (mean over standard deviation) for the normalised mBC UMI counts captured measured across replicate clonal cells profiled (n=946 reporters from n=290 clones, across two biological replicates).

See also **Ext. Data Fig. 2-3**, **Supp. Fig. 1**.

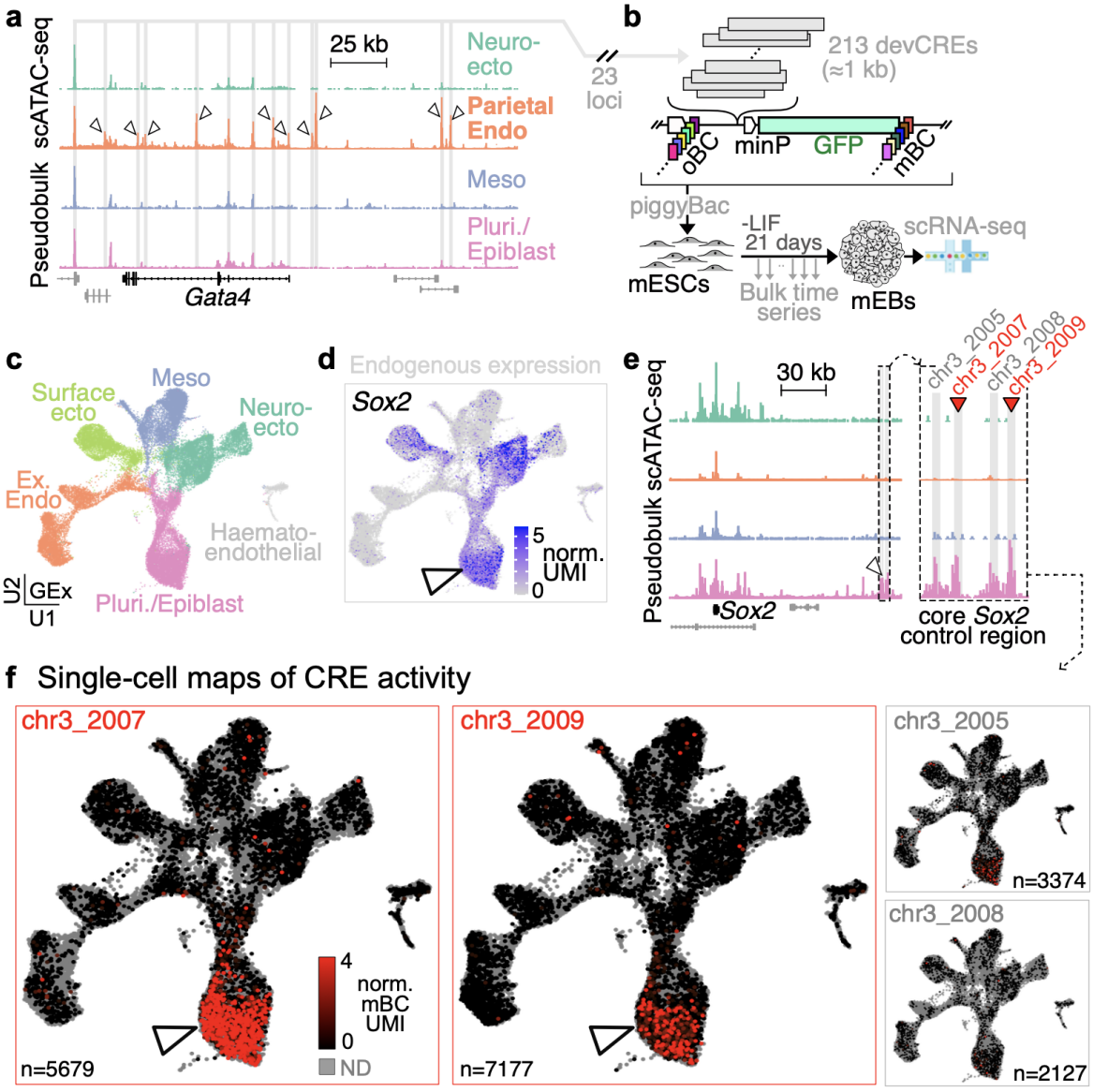



**Figure 3. Locus-level screen of developmental CREs in mouse embryoid bodies**

**a** Pseudo-bulk pileup of scATAC-seq data at *Gata4* (±100 kb from TSS) as a representative selected developmental locus (carets: differentially accessible peaks). *Gata4* is expressed predominantly in parietal endoderm cells (expression **Fig. 4d**, top row). Reproducibly and highly accessible ATAC peaks (in expression-cognate cell-type) within the 200 kb window were profiled (n=13 for *Gata4*, grey shading).

**b** scQers containing 204 putative developmental CREs taken from 23 developmental loci (22 plus *Sox2* control region) were integrated at high MOI in mESC using piggyBac. Transfected libraries included 89% CRE series, 10% exogenous promoters (same as **Fig. 2a**), and 1% EEF1A1p-mCherry (co-transfected for selection to increase MOI[59,60]). Reporter-integrated cells were differentiated to embryoid bodies for 21-days, with bulk sampling every 2 days, and single-cell profiling at three weeks.

**c** UMAP projection of scRNA-seq (n=43799 quality-filtered cells) from three biological replicates of scQer-integrated 21-day mEB cells, with annotation from integration with *in vivo* data[61] (finer annotation **Ext. Data Fig. 4a**).

**d** Endogenous expression (normalised UMI counts) for *Sox2* displayed on UMAP projection, highlighting pleiotropic expression in pluripotent (caret) and ectodermal lineages.

**e** scATAC pseudobulk pileup for *Sox2* locus. Caret points to the *Sox2* control region[56,57], inset zooms in the core. Regions profiled and differentially accessible in the pluripotent population are shaded in grey. Red carets mark the two cell-type-specific CREs.

**f** Single-cell maps of CRE activity for four CREs (separate panels). Each point represents a single cell. Grey indicates cells with no reporter detected for the specified CRE. Colour marks reporter expression (average normalised mBC UMI per cell) from none (black) to high (red) for cells with detected reporters (oBC UMI>10). Colour axis truncated to 4 UMI. Elements chr3_2007 and chr3_2009 have significant expression specific to pluripotent cells (carets) ( **Fig. 4a**, marginal activity from chr3_2005 significant in only 1 of 3 biological replicates), mirroring *Sox2* expression in that cell type (c.f., panel **d**). Number of cells with detected reporter integrations indicated on each panel.

See also **Ext. Data Fig. 4-6**, **Supp. Fig. 2-3**.

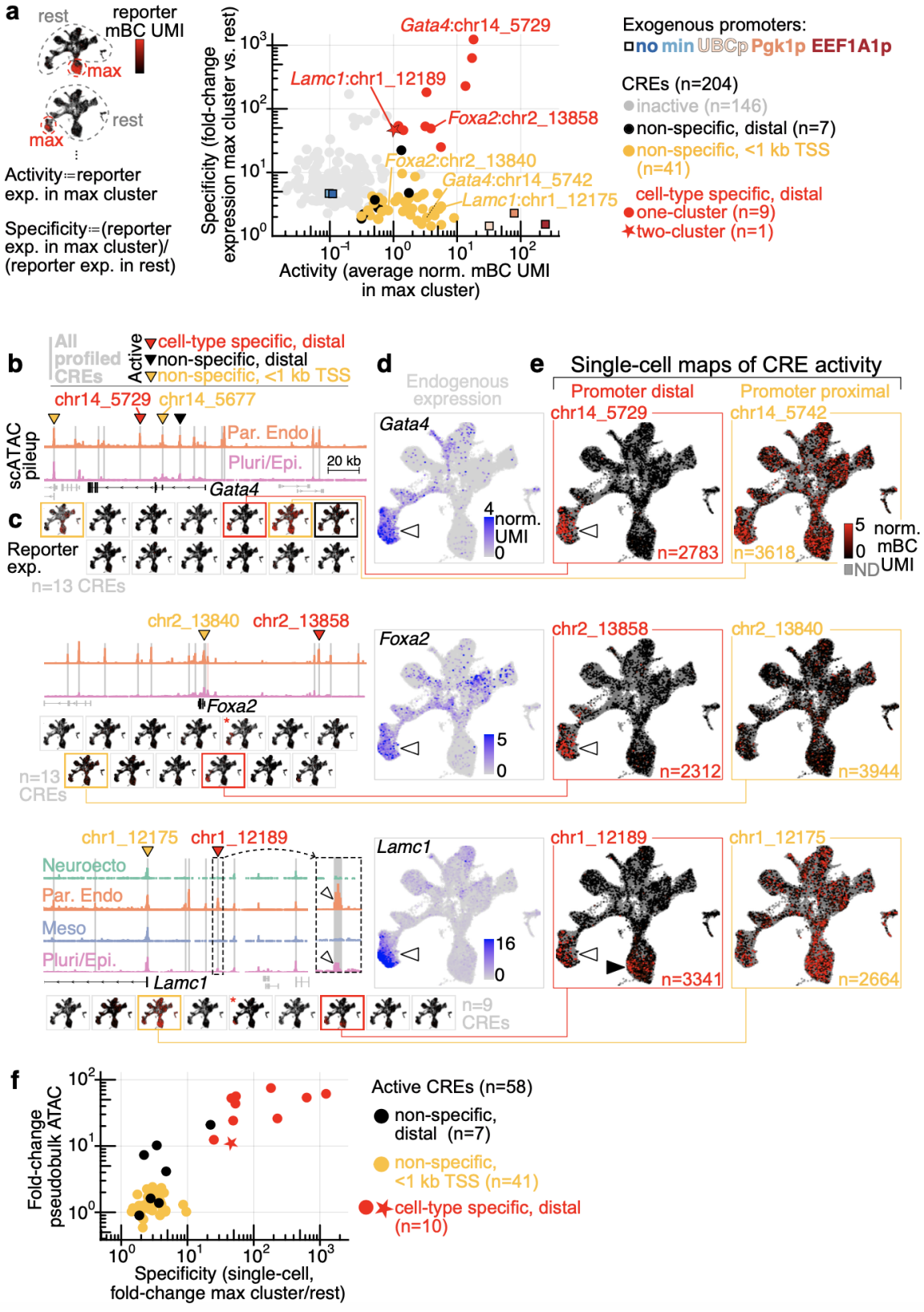


**Figure 4. Multiplexed identification of constitutive and autonomous lineage-specific CREs**

**a** Quantification of CRE function (median from three biological replicates). Activity: reporter expression (average normalised mBC UMI count) in the maximum-expression cell-type (defined from fine clusters of **Ext. Data Fig. 4a**). Specificity: maximum-expression cell-type reporter level over expression in all other cells. Active elements (black: non-specific, distal; orange: non-specific, <1 kb TSS; red: cell-type specific): show excess expression (bootstrap resampling) in all replicates compared to basal controls (no and minimal promoter). ell-type-specific elements (specificity >5 and significantly higher than cell-type permuted sets) are highlighted (red). CRE *Lamc1*:chr1_1218, active in two cell types, is marked with a star. Exogenous promoters (same as **Fig. 2a**) are shown as coloured squares.

Panels **b**-**e** are reproduced for the different loci (top to bottom:*Gata4, Foxa2, Lamc1*).

**b** Pseudobulk pileup of scATAC (pluripotent and parietal endoderm: *Gata4, Foxa2*, also neuroectoderm and mesoderm for *Lamc1*) for 200 kb region centred on gene transcription start site. Grey shading of peaks indicate regions profiled (red shaded peak near *Foxa2* TSS: peak not in the library due to inability to identify specific cloning primers). Carets point to elements identified as active with scQers. Inset for *Lamc1* locus highlights differential accessibility in both pluripotent/epiblast and parietal endoderm cells .

**c** Single-cell CRE activity maps for all tested elements in the locus. Outline indicates activity of element in assay (colouring as panel **a**). Red asterisk mark elements with activity but in <3/3 replicates.

**d** Endogenous expression (scRNA-seq, normalised UMI counts projected on UMAP) for genes corresponding to loci shown. Caret points to the parietal endoderm cells.

**e** Single-cell reporter expression (normalised mBC UMI, projected on UMAP, colormap truncated at 5 mBC UMI/cell for contrast) for putative promoter (orange) and distal CRE (red) associated with the gene in the locus. .. Number of cells with detected reporters per element is indicated. White carets point to parietal endoderm. Black caret (*Lamc1*:chr1_12189) marks reporter expression in pluripotent cells.

**f** Fold-change in ATAC (cognate cluster vs. rest of cells) vs. single-cell reporter expression specificity (definition and colour scheme, panel **a**) for all active elements identified.

See also **Ext. Data Fig. 7-10**, **Supp. Fig. 4-8**.

***Methods***

Primers, oligos, and plasmids are listed in **Supp. Data 11**. Maps of final amplicons and plasmids are on github (shendurelab/scQers). Additional methods details are provided in **Supp. Note 6**.

**scQer single-cell libraries preparation and sequencing**

Each 10x lane provides three scRNA-seq libraries (gene expression, mBC, oBC). Library preparation follows the protocol from the manufacturer (steps number listed in this section refer to: v3.1 manual CG000205 Rev D, 10x Genomics, but likely applicable to updated versions with little modification) until step 2.2a (first cDNA amplifications). At that step, it is critical to spike-in primers specific to the mBC and oBC reporters (oSR38 and oJBL246 respectively) to a final concentration of 0.5 uM. This will ensure higher capture of the reporter RNAs and will help limit the number of PCR cycles overall.

Following cDNA amplification, clean up proceeds per the protocol (with gene expression and mBC components in the pellet fraction 2.3Ax, and oBC in the supernatant fraction 2.3Bxiv). After step 2.3, gene expression libraries are completed following the manufacturer's protocol. We note that gene expression, oBC, and mBC libraries can all be sequenced on the same Illumina Nextseq run with the design described below.

oBC libraries

Final oBC libraries are generated by a semi-nested second PCR using amplified cDNA (55% of fraction 2.3Bxiv as template) in 100 uL using Nextera P5 primers (e.g., NextP5_index1) and

custom indexed P7 primers (e.g., oJBL425-oJBL427). For example: 50 uL 2x KAPA2G Robust HotStart ReadyMix (Roche), 12.5 uL amplified cDNA from step 2.3Bxiv (supernatant), 5 µL 10 µM NextP5_index1 primer, 5 µL 10 µM oJBL425 primer, 0.5 uL SYBr green 100x, and water to 100 µL; run parameters: 3 min at 95C, followed by cycling with 20 s at 95C, 20 s at 60C, 20 s at 72C. To avoid over-amplification, the reactions are tracked by qPCR and stopped at or below the inflection point. Given high expression of oBC, 5-7 PCR cycles are typically sufficient to get high concentration libraries. The resulting amplified libraries are purified by 1.5x Ampure XP beads (Beckman Coulter). To avoid loop-the-loop products, the lowest band (207 bp, amplicon: PCR2_oBC_10x_scQer.gbk on github), can be size-selected prior to sequencing by PAGE purification.

Sequencing of the oBC libraries follows the following structure – read1: primer standard Illumina Nextera read1, ≥28 cycles (cell barcode, UMI), index1: custom oJBL432, 6-10 cycles (sample index), read2: custom oJBL433, ≥16 cycles (oBC).

mBC libraries

Final mBC libraries (here mRNA molecules captured from poly-dT reverse transcription primers) are generated with two steps of PCR, first a semi-nested PCR2 followed by an indexing PCR3. PCR2 conditions: 50 uL 2x KAPA2G Robust HotStart ReadyMix, 12.5 uL amplified cDNA from step 2.3Ax (pellet), 5 µL 10 µM oJBL324 primer, 5 µL 10 µM oJBL529 primer, 0.5 uL SYBr green 100x, and water to 100 µL; run parameters: 3 min at 95C, followed by cycling with 20 s at 95C, 20 s at 65C, 50 s at 72C. To avoid over-amplification, the reactions are tracked by qPCR and stopped at or below the inflection point. 10 PCR cycles are typically sufficient to get high

concentration libraries. PCR2 products are purified by 1x Ampure XP beads. 10% of the PCR2 product then serves as template for an indexing PCR3: same conditions as above, with primers oJBL076 (P5) and custom indexed P7 (e.g., oJBL530-533). Typically, 4-6 cycles are sufficient for indexing. Final libraries are purified by 1x Ampure XP beads (633 bp, amplicon: PCR3_mBC_10x_pdT_scQer.gbk on github).

Sequencing of the mBC libraries follows the following structure – read1: primer standard Illumina Truseq read1, ≥28 cycles (cell barcode, UMI), index1: custom oJBL534, 6-10 cycles (sample index), read2: custom oJBL334, ≥15 cycles (mBC).

**Benchmarking and optimization: promoter series in human cell lines**

Cloning and subassembly of dual-RNA reporters promoter series

To generate the dual-RNA reporter plasmid libraries, we first created a barcoded "cloning dock" plasmid, with restriction sites and homology regions to various cassettes enabling modular addition of 1) Tornado[34] RNAs cargos, 2) cis-regulatory element libraries, and 3) reporter mRNAs. To generate the cloning dock, plasmid p001 containing a piggyBac transposon backbone[94] was digested with XbaI and HpaI (NEB) and the backbone product purified by agarose gel extraction (Zymoclean Gel DNA recovery kit, Zymo Research). To generate the cloning dock insert, a GFP fragment with barcoded 3' UTR was amplified from plasmid pSGR017 with oJBL315+oJBL316 (all primers and oligos are listed in **Supp. Data 11**) and the resulting product gel purified by PAGE. The barcoded 3' UTR was combined with gene block gJBL008 with the piggyBac backbone by isothermal assembly (HiFi NEBuilder, NEB), the resulting plasmid, p022, was electroporated in

*E. coli* (NEB, C3020), and the full complexity of the library maintained. Throughout, constructs were confirmed by colony PCR and Sanger sequencing of multiple clones.

We then added a barcode and capture sequence to the Tornado RNA plasmid pAV-U6+27-Tornado-Broccoli plasmid[34] (Addgene #124360). The Tornado plasmid was digested with NotI and SacII (NEB) and the backbone purified by agarose gel extraction. A barcoded insert fragment was generated by PCR using the pAV-U6+27-Tornado-Broccoli plasmid as template and primers oJBL220+oJBL291. The barcoded insert was assembled with the purified digested Tornado backbone and gene fragment gJBL007 by isothermal assembly and electroporated in *E. coli* (NEB, C3020), maintaining the full complexity of the library. The resulting plasmid, p019, contained the oBC with capture sequence 1 (CS1) cargo inserted in the Tornado cassette. Plasmids p019 was then digested with BamHI and XhoI (NEB) and p022 with BsbI, with the insert and backbone respectively purified by agarose gel extraction. The components were combined by isothermal assembly to generate plasmid library p025, which was electroporated in *E. coli*, maintaining complexity. Plasmid p025 contains the two barcodes (oBC and mBC) separated by 344 bp and is the starting point to clone scQers (**Supp. Fig. 2**).

To construct five libraries (one per promoter in the series, see below), p025 was separately bottlenecked to an estimated 300 clones five separate times, and the oBC and mBC were subassembled from the separate pools. Briefly, amplicons were generated from the bottlenecked p025 as template, and using primers oJBL345 and oJBL337-oJBL341 (indexed primer, one per library). Reactions were carried out in 50 uL volume with 20 ng input plasmid template (25 uL polymerase master mix, 2.5 uL 10 uM oJBL345, 2.5 uL 10 uM indexed primer oJBL337-oJBL341,

0.25 uL 100x SYBr green, water to 50 uL) using Kapa HiFi PCR master mix (Roche) with PCR conditions: 95C 3 min, cycling with 98C 20 seconds, 60C 20 seconds, 72C 30 seconds. Reaction was tracked by qPCR and collected at the inflection point. Amplicons were purified by 1x Ampure.

Libraries were diluted to 2 nM based on the Tapestation D1000 HS quantification, and sequenced on NextSeq 500 with the custom primers: read 1 primer oJBL346 (oBC, 26 cycles), index 1 primer oJBL347 (library index, 6 cycles), read 2 primer oJBL348 (oBC reverse complement, 25 cycles), and index 2 primer oJBL349 (mBC reverse complement, 20 cycles).

Sequencing data was demultiplexed using bcl2fastq. Raw fastq files were processed first by trimming unnecessary cycles from the 3' end (10 cycles from read 1, 5 cycles from read 2, 9 cycles from index 1) using seqtk (https://github.com/lh3/seqtk). Forward and reverse oBC reads were joined and error corrected with PEAR[95] (options -v 16 -m 16 -n 16 -t 16). Using custom python and R scripts, assembled oBC reads were combined with mBC reads, and oBC/mBC pairs were counted. The read count distribution displayed a clear bimodal distribution suggesting a saturated library, and oBC-mBC pairs with >500 reads were retained as valid. In order to further restrict the list of oBC-mBC pairs unique across the five bottlenecked libraries, all oBC/mBC pairs were combined, and any pair containing a oBC or mBC appearing more than once (either within a library, or across different libraries) was discarded to avoid mapping conflicts in the analysis of single-cell reporter data (amounting to 24% of high read count pairs), leaving 1122 unique oBC-mBC pairs across the five libraries (number of oBC-mBC pairs per library ranging from 139 to 306, with a median of 205).

Finally, each bottlenecked p025 library described above was digested with BglII, purified by 1x Ampure, digested with EcoRI (NEB), and the resulting backbone was purified by agarose gel extraction. Inserts comprised of various promoters with puromycin cassette and GFP linked by a P2A element were generated as follows. For the human EEF1A1 promoter (including the first intron), minimal promoter and promoterless cassette, primers oJBL254+oJBL314 were used to amplify respective constructs from plasmids pSGR017, pSGR018, and pSGR019 respectively, yielding a promoter puromycin-P2A-GFP fragment. For the human UBC promoter (including the first intron), puromycin-P2A-GFP fragment was obtained by amplifying from pSGR017 with primers oJBL254+oJBL392, and the promoter fragment was amplified from plasmid pB-rtTA with primers oJBL393+oJBL394. For the mouse Pgk1 promoter (no intron), puromycin-P2A-GFP fragment was obtained by amplifying from pSGR017 with primers oJBL254+oJBL392, and the promoter fragment was amplified from plasmid PGK1p-Cys4-pA with primers oJBL395+oJBL396. Promoter sequences are listed in **Supp. Data 1**. All fragments were gel purified, combined with their respective digested bottlenecked p025 backbones, and electroporated, resulting in five dual-RNA barcode reporter plasmid libraries, one for each promoter: p029 promoterless (noP), p027 minimal promoter (minP), p042 PGK1, p041 UbC, and p028 EEF1A1. Given the a priori subassembly of mBC-oBC pairs for the starting bottlenecked plasmids, and the fact that each library above was assembled separately, each promoter was associated with a list of pairs of oBC and mBC, enabling downstream quantification in a single-cell context.

Plasmid libraries were purified by midiprep (Zymo Research), concentrated by isopropanol precipitation, and pooled at 1:1 ratio by mass. This pooled library of the five promoters was used for both the benchmarking experiment in cell lines (**Fig. 2a**) and was also spiked in the developmental CRE experiment in mESC (**Fig. 3b**).

Cell culture, transfection, bottlenecking, and harvesting

K562 cells (CCL-243, ATCC) were grown in RPMI 1640 medium (ThermoFisher, cat. num. 11875119), supplemented with 10% FBS (Fisher Scientific, Cytiva HyClone™ Fetal Bovine Serum, cat. no. SH3039603) and 1x Penicillin/streptomycin (ThermoFisher, cat. num. 15140122). HepG2 (HB-8065, ATCC) and HEK293T (CRL-3216, ATCC) cells were grown in DMEM (ThermoFisher, cat. num. 10313021) with 10% FBS and 1x Penicillin/streptomycin. Cells were kept at 37C and 5% CO2, and passaged every two days (K562, HEK293T) or when cells reached confluency (HepG2, typically every three days). For clonal expansion, we waited for near confluence from 12-well plates (1-2 weeks) before passaging.

All cells were transfected in mid-exponential phase. K562 cells were transfected using MaxCyte electroporation following manufacturer's protocol (1.5 M cells, with 15 ug reporter scQers promoter plasmid mix (see above), 0.5 ug superPiggybac transposase (SBI) in 50 uL volume). Two replicates of 1 M of HepG2 and HEK293T cells were transfected using lipofectamine 2000 (ThermoFisher, cat. no. 11668030, Gibco Opti-MEM cat. no. 31985) with 4 ug of reporter plasmid mix and 0.2 ug of super PiggyBac transposase (SBI). Medium was changed the next day, and cells passaged as usual thereafter. After 5 days, cells were put on puromycin selection (Gibco, cat. no.

A1113803, concentration: 2 ug/mL), and grown for an additional 10 days to allow complete dilution of non-integrated plasmids. After >15 days of growth post-transfection, populations from each cell line were bottlenecked to an estimated 250 and 500 starting clones, and expanded to large populations. Notably, HepG2 cells displayed less robust growth at low densities, and required longer time for expansion, suggesting an effectively more severe bottleneck, in line with inferred clonal population properties (fewer final clones, **Ext. Data Fig. 3a-b**).

The bulk vs. single-cell quantification experiment (**Fig. 2**) was performed in two replicates. The first replicate (replicate A) with populations bottlenecked at an expected 250 clones, and the second replicate (replicate B) with populations bottlenecked at an expected 500 clones. For each replicate, at the same time, cells from each line were: 1) harvested separately and methanol fixed for bulk quantification, and 2) prepared as single cell suspension, hand-mixed at an expected 1:1:1 ratio, and profiled for single-cell transcriptomics. Briefly, for the bulk methanol fixation, K562 cells (and HEK293T and HepG2 cells following lifting off plate with 0.05% trypsin) were washed once with ice cold PBS, and resuspended in 80% ice cold methanol, to a concentration of 1 M cells/mL, and placed at -80C until further processing. For single-cell processing, cells were washed twice with PBS+BSA (0.04%) and diluted to 1000 cells/uL. Cell dilutions were mixed at estimated equal proportion and loaded to expected 10k recovered cells total on the 10x Chromium platform following manufacturer's protocol (CG000205 Rev D, Single Cell 3' v3.1 with feature barcoding, 10x Genomics), as one lane per replicate (two lanes total). Replicate B showed some evidence of a partial wetting failure, but otherwise displayed a good emulsion.

Bulk MPRA library preparation

Genomic DNA was extracted from methanol fixed cells using the DNeasy kit (Qiagen), and RNA was extracted from cells using TRIzol LS (Thermo Fisher), following manufacturer's instructions in both cases. MPRA amplicon libraries from DNA were generated in two steps of PCR amplification with Kapa HiFi (Roche). 0.5-1 ug of genomic DNA input was used. For low-cycle number PCR1, gDNA was mixed with 50 µL 2× Kapa HiFi master mix, 5 µL 10 µM oJBL039, 5 µL 10 µM oJBL358, and water to 100 µL. Cycling parameters: 1 min at 95C, and 4 cycles of: 20 s at 98C, 20 s at 60C, 30 s at 72C, followed by 4C hold. Primer oJBL358 contains 10 random Ns to serve as a pseudo-UMI (hereafter referred to as UMIs for brevity) to correct for PCR jackpotting. Reactions were cleaned up with Ampure XP beads at 1×, and eluted in 20 µL of 10 mM Tris 8. Illumina adapters and sequencing indices were appended through PCR2, with 4 µL of the eluate from PCR1 taken as input, and 25 µL 2× Kapa HiFi master mix, 0.25 µL 100× SYBr green, 2.5 µL 10 µM oJBL077, 2.5 µL 10 µM indexed primers (oJBL359-oJBL364), and water to 50 µL. Libraries were amplified with tracking by qPCR with: 1 min at 95C, and cycles up to the qPCR inflection point: 20 s at 98C, 20 s at 60C, 30 s at 72C. Libraries were then cleaned up with Ampure XP beads at 1×.

Amplicons libraries for RNA were obtained by first DNase-treating RNA (5 µg RNA, 2 µL TURBO DNase [Thermo Fisher], 2 µL 10× buffer, and water to 20 µL, incubated at 37C for 30 min, cleaned up with RNA clean & concentrator [Zymo Research], and eluted in 11 Tris 7 10 mM). 1 µg of DNase treated RNA was then taken to reverse transcription. Briefly, 2 µL (500 ng/µL) RNA was mixed with 2 µL 1 µM oJBL358, incubated at 65C for 5 min, and placed on ice.

15 µL of reverse transcription master mix was then added (4 µL 5× FS buffer, 1 µL 0.1 M DTT, 1 µL 10 mM dNTP mix, 8 µL water, 1 µL SSIII [Thermo Fisher]), and the reaction incubated at 55C for 60 min, followed by 70C for 15 min. Half of the reverse transcription reaction was then directly amplified for PCR1 (37.5 2× Kapa HiFi master mix, 3.75 µL oJBL039 10 uM, 3.75 µL oJBL077 10 uM, water to 75 µL), with cycling parameters: 1 min at 95C, and 4 cycles of: 20 s at 98C, 20 s at 60C, 30 s at 72C, followed by 4C hold. Reactions were cleaned up with Ampure XP beads at 1×, and eluted in 20 µL of 10 mM Tris 8. PCR2 proceeded as for libraries prepared from genomic DNA, with oJBL077 and indexing primers (oJBL365, oJBL366, oJBL437-oJBL440), and reactions were stopped at inflexion point from qPCR tracking. Libraries were then cleaned up with Ampure XP beads at 1×.

Final libraries were quantified with Qubit dsDNA HS (Thermo Fisher), diluted to 3 nM, run on TapeStation D1000 HS (Agilent) for final quality assessment, and adjusted to final 2 nM based on the Tapestation quantification. Libraries were pooled, paired end sequenced on NextSeq500 with the following primers and cycle numbers: read1 (mBC forward):  28 cycles, primer oJBL369; index1 (UMI): 19 cycles, primer oJBL435; read2 (mBC reverse): 19 cycles, primer oJBL371; index2 (sample index): 10 cycles, primer oJBL370.

Bulk MPRA data processing and quantification

Sequencing data was demultiplexed using bcl2fastq. Raw fastq files were processed first by trimming unnecessary cycles from the 3' end (13 cycles from read 1, 4 cycles from read 2, 9 cycles from index 1) using seqtk (https://github.com/lh3/seqtk). Forward and reverse mBC reads were

joined and error corrected with PEAR [95] (options -v 15 -m 15 -n 15 -t 15). Using custom python and R scripts, successfully assembled barcode reads were combined with UMI reads, mBC-UMI pairs were counted, and the read and UMI counts per mBC determined. The read and UMI counts for the mBC present in the reporter pool (determined *a priori*, see section on reporter cloning and subassembly above) were collected for downstream analysis and comparison to single-cell quantification.

Expression for each mBC from the UMI counts table was computed as follows. First, the total UMI per sample (per cell line and replicate) to the mBC in our list was determined for both RNA and DNA derived libraries. Each mBC UMI count was then normalised by the summed of counts in its respective sample type (DNA and RNA). The normalised RNA UMI count was then divided by the normalised DNA UMI count, to generate the bulk MPRA derived estimate of expression per mBC.

Single-cell reporter data processing

Four different components are needed to perform reporter quantification using our approach: 1) a triplet map connecting cis-regulatory elements with oBC and mBC sequences, 2) single-cell gene expression UMI counts, 3) single-cell oBC UMI counts, and 4) single-cell mBC UMI counts. For this promoter series experiment, the triplet CRE-oBC-mBC map was described above. We briefly describe below how the count data is obtained for the gene expression and barcoded RNAs. In each case, the output is a count table of the form (cell barcode, gene or barcode, UMI count).

*Gene expression libraries*

Data was converted to fastq using bcl2fastq, and fastqs were minimally processed (trimming read 1 to 28 cycles with seqtk, files renamed) to be compatible with cellranger (version 6.0.1, 10x Genomics), which was run using reference GRCh38-2020-A. Each CellRanger count output was processed with Seurat[96]. Briefly, cell barcodes were filtered to those with >700 gene expression RNA UMIs, and between 2 and 15% mitochondrial UMI fraction. This led to 5787, 4278, and 3834, cell barcodes across the replicates A, B1, and B2. 10x data was normalised, scaled and clustered using standard commands (NormalizeData with LogNormalize method, finding 1000 top variable features with FindVariableFeatures, scaling with ScaleData over all genes, RunPCA and retaining top 50 principal components [PCs] calculated on the identified variable features, FindNeighbors on the top PCs, FindClusters with 0.1 resolution, and RunUMAP with n.neighbors of 20 and using the top PCs as input features). The UMAP revealed three clear clusters (**Fig. 2b**, **Ext. Data Fig. 2a**), hypothesised to correspond to the three cell lines profiled. Replicates B1 and B2 also displayed an intermediate cluster, likely as a result of the lane partial wetting failure, found to share marker genes from the neighbouring clusters, which was excluded as plausibly composed of doublets. To confirm the cellular identity of each cluster, in addition to assessment from canonical marker genes (e.g., HBG1/2 in K562, ALB in HepG2), we compared the pseudo-bulked expression (mean across UMI counts for each gene) to bulk expression quantification in the three lines (as assessed from the average of stranded bulk RNA-seq ENCODE[97] datasets in K562 and HepG2, and in HEK293T), finding unambiguous correspondence of each clusters to a single line (average log-transformed $R^2$=0.72 for matches, vs. 0.39 for non-match).

Following preliminary filtering described above, cell barcodes corresponding to putative doublets were further filtered by two stringent methods. First, each large cluster was further sub-clustered using the same method as above, revealing focal subclusters which shared marker genes from large neighbouring clusters, and usually had nearly 2-fold more total RNA UMIs. Cell barcodes contained in these clusters were excluded as likely doublets. Second, scrublet [98] was run on the filtered cell barcode set (>700 RNA UMIs, 2 to 15% mitochondrial RNAs), and a doublet score threshold of 0.25 was selected for filtration based on the separation of the bimodal peaks in the simulated score distribution. Cells either belonging to doublet subclusters or having a scrublet doublet score > 0.25 (we observed high concordance between the two approaches) were filtered out. Finally, cells with anomalously high gene expression UMI (>4000) or anomalously high multiplicity of reporter integration (>100, see below), also likely doublets, were removed, leaving 5505 high confidence cells for replicate A (K562: 2184, HEK293T: 2090, HepG2: 1231), 3533 for replicate B1 (K562: 1303, HEK293T: 1238, HepG2: 992), and 3172 for replicate B2 (K562: 1298, HEK293T: 1056, HepG2: 818).

*mBC libraries*

Data was converted to fastq using bcl2fastq, and fastqs were minimally processed (trimming read 1 to 28 cycles and read 2 to 22 cycles with seqtk, files renamed) to be compatible with cellranger (version 6.0.1, 10x Genomics), which was run to perform error correction on cell barcodes. The resulting position sorted bam files were then parsed for the mBC reads as follows using a custom python script: reads aligning to the reference genome or without either corrected cell barcode or UMI (tags CB and UB in the bam file) were discarded. Only reads with the exact expected 7 nt sequence (TCGACAA) downstream of the mBC (positions 16 to 22) were retained. List of all

UMIs corresponding to a cell barcode and mBC pair were stored, discarding chimeric UMIs (taken to be UMIs for which the proportion of reads associated to a given mBC vs. all other mBC in the specified cell barcode falls below 0.2). mBC composed of all Gs (empty read) were discarded. Importantly, the mBC UMI counts were error corrected as follows. For each given mBC and cell barcode, the Hamming distance between all UMIs was calculated, a graph created by connecting UMIs that with a Hamming distance ≤ 1, and the resulting the number of connected components in the graph was taken as the error-corrected UMI count for a given cell barcode-mBC pair. These error corrected UMI counts were taken as the per single-cell quantification of the reporter mRNA expression (see below for a normalisation strategy to correct for technical factors). Given that cell barcodes derived from capture sequence vs. poly-dT reverse transcription primer are different on the 10x Genomics beads (bases 8 and 9 reverse complemented) on the same bead (and not error corrected by cellranger in our application), we converted the CS2 cell barcodes to their poly-dT counterparts to enable matching across the different libraries.

*oBC libraries*

oBC libraries were processed in an entirely analogous way to the strategy for mBC described above, with the following modifications: two sequencing runs were combined in a single fastq prior to processing, read 2 were trimmed to 23 cycles, and only reads with the GCTTTAA (constant region after the oBC) at positions 17 to 23 were retained. The number of UMIs per oBC per cell barcode was also taken as the error corrected (1 Hamming distance) count and our measure of oBC expression in single cells (see below for a normalisation strategy to correct for gene expression UMIs). Similarly to the CS2 mBC data above, we again converted the CS1 cell barcode to poly-dT cell barcodes.

Quantification of expression in single-cell assay and comparison to bulk

To quantify reporter expression via our single-cell experiment, we first determined the set of valid oBC (present in our oBC-promoter-mBC subassembly table generated *a priori*) detected in each cell. As a tradeoff between specificity and sensitivity (see clonotype precision-recall analysis below), we selected a threshold of ≥12 UMI (**Fig. 2c**) to deem a oBC as present for a given cell barcode. The UMI counts for valid mBC cell-barcode pairs were then joined to the detected oBC in all valid cell barcodes by using the predetermined oBC-mBC (uniquely matchable) association table. In cell barcode/oBC combinations for which there were no detected mBC UMI, a value of 0 was taken (detection of reporter integration from oBC, but no captured reporter mRNA). Importantly: while not detected, given our dual RNA strategy, this represents a "true" zero and contributes to our measurement of expression. mBC UMI counts were normalised by the number of gene expression UMI (from the full transcriptome GEx libraries) detected in each cell, i.e., (mBC UMI)/(GEx UMI)*mean(GEx UMI), where the scaling with the mean gene expression UMI across all cells served to maintain an intuitive unit in the data. Normalisation by simple scaling by gene expression UMI was performed as the mBC UMI counts were correlated ($R^2$ of log-transformed values=0.09) with gene expression UMI with a slope close to 1 (least square fit on log-transformed data, slope: 0.93). We find in both our comparison to bulk data and our clonal analysis (see below) that direct normalisation of mBC by GEx slightly improves the precision of the expression measurement. To quantify single-cell expression for each mBC (**Fig. 2g**), we then directly averaged the normalised mBC UMI counts across all cells with a detected associated oBC.

The averaged normalised mBC UMI described above was directly compared to the bulk expression quantification (from bulk MPRA), **Fig. 2e** and **Ext. Data Fig. 2d**. In these analyses, we only include well-represented barcodes in the comparisons to focus attention on technical noise resulting from the two methods and not noise from sparse sampling of rare barcodes (mBC with 5 or more cells with oBC detected integrations, at least 1 mBC UMI captured across all integrations, and at least 100 DNA UMI from the bulk quantification).

For quantification without conditioning on oBC detection (**Ext. Data Fig. 2h**), the average normalised mBC UMI across all cells with any captured counts was taken. Including an additional step to filter possible chimeric amplicons (removing events for which the number of reads equaled the number of UMIs, unlikely in a saturated library) did not substantially improve performance without oBC detection.

In addition to the accuracy comparison to the bulk quantification, we also directly assessed the number of incorrectly detected mBC (mBC UMI count >0, but not detected as determined by absence of the associated oBC (<12 oBC UMI) in the same cell) for the different promoters. We found the following proportions of valid (oBC matched) mBC detection events (mean proportion from replicates A and B1): no promoter: 60%, minimal promoter: 45.9%, UBCp: 51.4%, Pgk1p: 40.4%, EEF1A1p 10.5%. Spurious detections thus constituted a substantial, and sometimes dominant, proportion of events in all cases.

**Profiling developmental cis-regulatory elements in mouse embryoid bodies**

Cell culture

*Mouse embryonic stem cells*

A low-passage number monoclonal male BL6 (male WD44, ES-C57BL/6 gift from C. Disteche and C. Ware at University of Washington) mouse embryonic stem cell line stably expressing dCas9-BFP-KRAB was used. Cells were grown on gelatin (0.2%) (Sigma, cat. No. G1890) coated plates and cultured in DMEM (ThermoFisher, cat. num. 10313021) supplemented with 15% FBS (Biowest, Premium bovine serum, cat. no. S1620), 1x MEM non-essential amino-acids (ThermoFisher, cat. no. 11140050), 1x Glutamax (ThermoFisher, cat. no. 35050061), $10^{-5}$ beta-mercaptoethanol, and $10^{-4}$ leukemia inhibitory factor (Sigma-Aldrich, ESGRO Recombinant Mouse LIF Protein ESG1107), hereafter referred to serum+LIF medium were necessary, with daily medium changes (aspirate medium, replace with pre-warmed medium), and transfer every two days (aspirate medium, wash with PBS [without $Ca^{2+}$, $Mg^{2+}$], add 2.5 mL [for 10 cm plate] 0.05% trypsin , incubate 2 minutes at 37C, deactivate trypsin and triturate with 10 mL pre-warmed medium, spin down 5 min at 300g, aspirate supernatant, resuspend in pre-warmed medium, and transfer to new gelatinized plate).

*Mouse embryoid bodies induction and maintenance*

Exponentially growing mESCs are lifted from the plate (aspirate serum+LIF medium, wash with PBS, add 2.5 mL [for 10 cm plate] 0.05% trypsin , incubate 2 minutes at 37C, deactivate trypsin and triturate to a single-cell suspension with 10 mL pre-warmed medium). Cells are then counted and spun down (5 min at 300 g). Supernatant is aspirated and cells are resuspended to 2 M/mL in

CA medium (medium for EB induction: DMEM, 10% FBS, 1x MEM non-essential amino-acids, 1x Glutamax, $10^{-5}$ beta-mercaptoethanol). Cells are counted again, and density adjusted to 1 M/mL with CA medium. 3 mL (3 M cells) are added to 12 mL of CA medium in 10 cm plates (suspension plates: non gelatinized, non adherent). One the next day, plates are gently agitated to promote cell aggregation. Following induction, embryoid bodies (mEBs) are passaged every two days (no daily medium change). mEBs are collected using a serological pipette and transferred to a 50 mL conical tube (typically three plates are pooled). Leftover mEBs on plates are recovered by a CA medium wash and pooled in the conical tube. mEBs are left to settle (initially up to 15-20 min, faster as the mEBs grow in size). Once mEBs have settled, medium is aspirated from the top, carefully avoiding disturbing the loose pellet. Fresh, pre-warmed, CA medium is then added to 15 mL/plate and mEBs redistributed to plates.

Construction of CRE series dual RNA reporter plasmid library

Doubly barcoded backbone p025 was re-cloned at higher complexity (≈1M oBC-mBC pairs), see **Supp. Note 6** for details.

*PCR cloning of putative developmental CREs and assembly in dual RNA plasmid*

Putative CREs selected for profiling (see above) were cloned by PCR from mouse genomic DNA. A compromise amplicon size of 0.9 kb was taken as rough target size in order to balance testing large regions without overly compromising success rate. In order to increase specificity, a nested PCR approach was taken: a first unburdened PCR with selected primers (below), followed by a second nested PCR using primers with homology arm for cloning in the common backbone.

Outer primers for the first PCR (**Supp. Data 11**) were selected by running Primer-BLAST[99] with as PCR templates the 1200 bp sequences for the putative CREs (350 bp symmetric extension on both sides of the ArchR called 500 bp ATAC peak window with bedtools[100] slop, followed with bedtools getfasta to obtain sequences from mm10 genome) with the following run criteria: PCR product size 800 to 1000 bp (forward primer between 0 and 200 bp and reverse primer between 1000 and 1200 bp), primer melting temperature Min 57.0 Opt 60.0 Max 63.0, Max Tm difference 3, no intron junction preference, specificity check to *Mus musculus* (taxid:1009). For certain CREs, Primer-BLAST did not return any specific result with these constraints. Constraints (on product size) were then sequentially relaxed to increase the search space, with ultimately requiring only that the product be at least 500 bp within the window. Five regions (Foxa2_chr2_13861, Sparc_chr11_7210, Lamb1_chr12_2182, Lamb1_chr12_2183, Sox17_chr1_58) were still too repetitive for Primer-BLAST to return results but had non-repeat sequences enabling manual primer selection. Two regions were too repetitive to find any primer pairs whatsoever and were thus not included in the screen (Sparc_chr11_7186, Foxa2_chr2_13842). Overall, primers were ordered to PCR clone 209/2011 CREs from our initial selected set.

Inner primers for the second nested PCR were selected using batch primer3[101] (non default options: GC clamp=1, max poly-X=4) using the first PCR product as a template (but allowing for at most 8 bp overlap between inner primers and the PCR1 product). Primer pairs leading the largest nested PCR product were selected and handles homologous to the backbone were added (forward: 5'accgcatcgatctcgagg[inner forward], reverse: 5'tcccaaagcagatgtagttgac[inner reverse]). Handles were added to the forward/reverse primer so that the orientation of the CRE relative to the promoter matched their relative orientation on the genome relative to the gene.

The first PCR was performed in 20 uL reactions with 40 ng of genomic DNA (harvested from the mESC line used [DNeasy, Qiagen] following manufacturers' instructions) with Kapa Robust (Roche) with following parameters: 95C 3 min; 40 cycles: 95C 15 sec, 60C 20 sec, 72C 1 min 40 sec; final extension 72C 1 min 40 sec; with individual reactions in separate wells of a 96-well plate with primers distributed using a 96-liquidator (Rainin). Products were cleaned up (1x Ampure XP beads) and visually checked on agarose gel (with >95% success rate as judged by presence of ~1kb-sized band, possibly with non-specific products), and eluted in 100 uL of 10 mM Tris 8. 0.5 uL of the purified up PCR1 products was taken as template for the second nested PCR using the same conditions, but with the inner primers. The resulting products were cleaned up (0.6x Ampure XP beads) and visually checked on agarose gel, showing a <2% failure rate and highly clean products (little non-specific bands/smears). The products were quantified with a spectrophotometer (Nanodrop), and pooled to a 1:1 ratio by weight. This pool was used as insert for a pooled Gibson assembly as described below.

Prior to addition of the putative CRE PCR products, the minimal promoter GFP cassette (reporter mRNA) was inserted in the doubly barcoded backbone p025 digested with EcoRI and BglII (NEB) (**Supp. Fig. 2a**) and maintained at highest clonal complexity upon transformation (electroporation without bottleneck) to generate plasmid library p043. The minP-GFP insert was generated by splice PCR (templates: minP fragment: amplification of p027 with primers oJBL314+oJBL416; GFP fragment: amplification of p027 with primers oJBL254+oJBL414) followed by gel extraction. Plasmid library p043 was then digested with NheI/MfeI, combined with the pooled PCR amplified CREs via Gibson assembly, and transformed (electroporation) with a bottleneck

via 100-fold dilution to an estimated complexity of ~50k clones (**Supp. Fig. 2a**). The resulting plasmid library (p055) was then subjected to the final subassembly step to connect oBC to the CRE.

*oBC-CRE subassembly*

Given the length of the inserted CREs (≈1 kb) and diversity of sequences, amplification of the region from minimal promoter to oBC was not a feasible strategy to subassemble oBC to CRE (≈1.3 kb from minP to oBC). We thus relied on tagmentation followed by semi-specific PCR. Briefly, plasmid library p055 was tagmented with Tn5 (Illumina, Nextera Tagment DNA enzyme, cat. no. 15027916) at a concentration such that the expected fragment size would be larger than the oBC to minP distance (≈1.3 kb), determined by a Tn5 titration curve experiment. Following tagmentation (5 uL 2x Tagmentation DNA buffer [Illumina, cat. no. 15027866], 0.4 uL Tn5 enzyme 1, 3.6 uL water, 1 uL 10 ng/uL plasmid library; 30 min at 37C), the tagmented plasmids were cleaned up (Zymo clean and concentrator, 3:1 binding buffer), eluted in 10 uL Tris 8 10 mM. 1 ng (1 uL of the elution) was amplified via semi-specific PCR with a Nextera primer with a P5 handle (oJBL512, binding to all P5 tagmentation events) and a oBC-specific upstream primer (oJBL502, binding to specific portion of the plasmid) in 25 uL (8.9 uL water, 12.5 uL 2x NEBNext master mix, 1.25 uL 10 uM oJBL502, 1.25 uL oJBL512, 1 uL tagmented plasmids, 0.1 uL 200x SYBR green) with the following conditions (gap fill: 72C for 5 min, 98C for 30 sec, then 12 cycles of 98C for 10 sec, 65C for 30 sec, 72C for 1 min). As controls for the non-specific product size distribution, the tagmented plasmids were also amplified with oJBL512 exclusively. Following purification (Zymo clean and concentrators), the amplified libraries were run on PAGE (6% TBE, 180V, 30 min). As anticipated, the amplicons with primers oJBL502+oJBL512 (semi-specific products) displayed reduced size distribution compared to oJBL512 alone amplified (non-specific)

products, with most oJBL512 exclusive amplicons >1.2 kb. Semi-specific oJBL502+oJBL512 products between 450 bp and 800 bp were size selected on the PAGE gel, purified (minimum size size from CRE ≈ 75 bp), and sequenced (read 1: CRE sequence, Illumina Nextera primer [no custom], 34 cycles; index 1: P7-idx, primer oJBL432 15 cycles; read 2: oBC, primer oJBL433, 30 cycles).

Following demultiplexing (from the P7 index), the sequencing data was processed by first aligning read 1 (mapping to CRE) using bowtie2 (v2.4.4)[102] using option '-k 2' to report multi-mapping regions (some of our CRE segments overlapped given the proximity of the called peaks and extension from 500 bp to ≈1 kb tested regions). The resulting alignment sam file was then sorted, converted to bam using SAMtools [103], and merged with the oBC (read 2) using custom scripts into a file storing the oBC, CRE identity of the mapping, position and strand of aligned read within the CRE. Total read counts to each oBC-CRE pair were then summed up with custom scripts, retaining information about distribution of alignment positions and strand within the CRE for downstream processing.

The piled-up count data on oBC-CRE pairs was then filtered to identify *bona fide*, unique pairs. First, pairs with median mapping position outside the expected range from the size selection step (<30 bp and >300 bp) and mapping on the incorrect strand were filtered out. Then, the proportion of oBC reads mapping to any given CRE was calculated across oBC-CRE pairs, and only pairs with >95% of oBC reads mapping to a unique CRE were retained. The read count distribution across all oBC-CRE pairs was bimodal suggesting a saturated library, and only pairs with >30 reads (separating the two modes) were retained. Finally, pairs with anomalously small or large

mapping position dispersal (90th to 10th percentile difference mapping positional spread <30 bp or >300 bp) were filtered out. We note that the positional filters enabled unambiguous discrimination between different but overlapping CREs (given that in all cases one of the CRE would have out of range mapping positions compared to the expected size from the amplicon library). Two elements (*Gata4*:chr14_5749, *Txndc12*:chr4_7975) shared a short identical sequence complicating the mapping, and were treated separately to not confound the fraction of oBC reads mapping to a given CRE. Following these filtering steps, we were left with 43.6k valid oBC-CRE pairs.

*Final oBC-CRE-mBC triplet table*

These oBC-CRE subassembled pairs were then linked with the previously determined oBC-mBC pairs from the starting plasmid library p025. Briefly, oBC (from final oBC-CRE pairs) were joined to mBC via valid oBC-mBC pairs (restricting to the uniquely mapped pairs). The resulting valid triplets oBC-CRE-mBC were then joined with the oBC-promoter-mBC triplets of the exogenous promoter library (experiment from **Fig. 2a**), and any oBC or mBC appearing twice in both libraries were removed from the final triplet list. The final number of valid oBC-CRE-mBC triplets was 33.0k, with a median of 145 valid mBC-oBC pairs per CRE. The resulting triplet map was used to deconvolute single-cell data in reporter quantification. Through the cloning and subassembly process, 5 out of the attempted 209 CREs dropped out (<20 valid mBC-oBC pairs), and consequently could not be quantified (*Col1a1*:chr11_15306, *Col1a2*:chr6_65, *Cited2*:chr10_1265, *Txndc12*:chr4_7952, *Btg1*:chr10_9570).

Experimental details of pooled screen for CRE in mEBs

*Transfection, cell culture, and bottlenecking*

Low passage number mESCs were expanded in serum+LIF medium on gelatin coated plates as described above (passaged every two days, medium change every day) on 10 cm plates. Cells were transfected using Lipofectamine 2000 (Thermo Fisher Scientific) using reverse transfection. Briefly, cells washed with 1x PBS, and lifted by adding 2.5 mL/10 cm plate of trypsin 0.05% (Gibco). Following incubation at 37C for 5 min, cells were triturated with an added 7.5 mL of medium, spun down at 300g for 5 minute, and resuspended by pipetting at an estimated 1.5 M/mL to obtain a single-cell suspension. Following straining (40 um), cells were counted and diluted to 0.5 M/mL with medium. Concurrently, the lipofectamine+opti-MEM (12 uL lipofectamine + 238 uL opti-MEM) and the opti-MEM+DNA (240.4 uL optiMEM + 4 uL 50 ng/uL transposase + 5.6 uL transposon mix) were separately prepared and mixed by pipetting. The 500 uL lipofectamine+DNA+optiMEM mix was then added to a gelatin coated plate, 1 M cells (2 mL) from the single-cell suspension was added to the plate, and gently mixed. No transposase and no DNA controls were included. The transfected transposon was a uneven mix of three components (too boost MOI, see below): 1) 89% of the p055 oBC-CRE-minP-GFP-mBC library, 2) 10% of the oBC-promoter-puromycin-GFP-mBC series (same as for experiment in cell lines, **Fig. 2a**), and 3) 1% of the EEF1A1p-mCherry plasmid (p060, see below). Two biological replicates were transfected in parallel, one with the hyPBase plasmid[41], and one with super PiggyBac (SBI). We did not find substantial difference in MOI in the two replicates (**Ext. Data Fig. 5c**, replicate A vs. B).

Transfected cells were passaged and expanded to allow for integration and unintegrated plasmid dilution. Five days post transfection, cells were split with a portion selected on puromycin (2 ug/mL), and another portion remaining unselected. After 5 days on puromycin, cells from no DNA controls and no transposase controls were dead. While a large proportion of cells in samples with integrated cargos samples died, the puromycin resistant population was expanded for two weeks post transfection to ensure complete dilution of the unintegrated plasmids (maintained on puromycin).

The two replicates were induced to form mEBs in CA medium (no puromycin) on suspension plates as described above (day=0, 14 days post transfection), starting with 24 M cells per replicate (8 10 cm plates with 3 M cells each in 15 mL of CA medium). Replicate A was the sample transfected with hyPBase (and selected on puro), replicate B the sample transfected with the SBI super PiggyBac. Following induction, mEBs were passaged every two days, with sampling 5-10% of EBs at each time point for bulk MPRA (for harvesting, mEBs were pelleted at 5 min at 300 g, medium aspirated, fixed with ice cold 80% methanol, and stored at -80C until processing).

At the 12 day time point, a subset of expanded cells from replicate B were sorted by FACS for mCherry signal, and plated on a MEF monolayer (Thermo Fisher, CF1 Mouse Embryonic Fibroblasts, MitC-treated, cat. no. A34958, plated at 0.4M cells per well) in the wells of 6-well plate at approximately 1000 cells/well for bottlenecking. Following colony expansion for 4 days with daily medium change, colonies were lifted as follows: two washes with 1x PBS, add 750 uL collagenase type IV (0.1%, Stemcell Technologies, cat. no. 07909), 8 min incubation at 37C, lifted colonies aspirated by pipetting. The collagenase treated colonies on MEFs were then gently

washed twice with 1 mL of serum+LIF medium added dropwise to recover additional colonies, and pooled with the previous ones. Lifted colonies were then spun down (400 g, 5 min), medium aspirated, trypsin treated to single-cell suspension (250 uL 0.05% trypsin used to mix the pellet, incubated 3 min at 37C, inactivated and triturated with 2 mL of fresh medium, and plated on gelatin-coated plates for expansion. Counting colonies suggested about half, or 500 clones, were obtained in this way. Following expansion for 8 days, mEB induction with 24 M cells (8 10 cm plates with 3 M cells each) was initiated as above. mEBs were passaged every two days, with sampling 5-10% of EBs at each time point for bulk MPRA as before. The bottlenecked replicate was termed 2B.

*End-point processing and single-cell sequencing*

For both non-bottlenecked and bottlenecked experiments above, mEBs were processed at the three weeks end point as follows (for each replicate): 2 suspension 10 cm plates of mEBs were pooled into a 50 mL conical left to settle. Medium was aspirated and mEBs were washed twice with 1x PBS, resuspended in 3 mL 1x PBS in the second wash, and split in two 1.5 mL aliquots in 2 mL tubes. PBS was aspirated from the tubes, and 500 uL of trypsin 0.25% was added per tube. Tubes were then agitated on a thermomixer at 37C and 650 rpm for 4 minutes. Cells were then gently dissociated by pipetting 10 times, and placed back on the thermomixer for 2 min. 1 mL of medium was then added per sample and pipetted to obtain a single-cell suspensions, the two samples were combined in a 15 mL conical, after passing them through a 100 um strainer. The strained single-cell suspension was counted, and cells were spun down (300 g, 5 min), resuspended to 4 M/mL, and taken to FACS to obtain a clean single-cell suspension (typical gating strategy shown in **Supplementary Figure 1**). >600k cells were then FACS sorted (in <50 min) in pre-warmed

medium to ensure the single-cell nature of the suspension (no gating on fluorescence, only on forward and side scatter) prior to generating the emulsions for single-cell RNA-seq. Sorted cells were then spun down at 400 g at 4C for 5 min, the medium gently aspirated, and cells resuspended to an expected 2.5 M cells/mL (based on FACS sort event counts) in ice cold 1x PBS + 0.04% BSA, cells were further counted and volume adjusted to 1200 k/uL with ice cold PBS+BSA.

Single-cell suspensions in PBS+BSA were taken as the starting point for the 10x Genomics protocol (v3.1 with feature barcoding). Emulsion and reverse transcription were performed per the manufacturer's instruction. Given prior empirical experience with mEBs processing, each 10x lane was slightly overloaded (by an additional 20%) to approach the expected recovery of 10k cells/lane. Each replicate was profiled with 2 lanes of 10x, for a total of 6 lanes.

Single-cell reporter data processing

Processing proceeded in a similar way as described for experiment in cell lines. See **Supp. Note 6** for details.

*Quantification of activity and specificity of CREs and statistical tests*

The following stringent tests were performed to identify active and specific CREs. Each CRE and biological replicate was considered separately.

To assess activity, all integration events (oBC UMI > 10) for the CRE considered were identified, and the total number of such integration events for the CRE recorded. $10^4$ bootstrap resamplings (random sampling with replacement) of the integration events were then performed. In parallel, sampling with replacement of integration events (same number sampled as the CRE considered to control for difference in representation) from both basal promoter controls (minimal and no promoters). For each bootstrap sampling, the average normalised mBC UMI counts (see above), stratified by cell-type clusters (Seurat identified, see **Ext. Data Fig. 4a**), were determined both for the CRE and the basal promoters. The maximum expression cluster identity and expression level in that cluster was stored. Mean expression of the reporter without stratification by cluster identity was also obtained (over all bootstrap resampled integration events irrespective of cell types). Following bootstrapping, an empirical p-value was determined as follows: the null distribution was taken as the maximum cluster expressions across all bootstrap samplings of the two basal promoters. The empirical p-value of expression for the CRE considered to have activity in excess of the basal control (activity p-value) was taken as the probability that maximum cluster bootstrap CRE expression was below that of the basal controls, averaged over all bootstrap sampling for the basal control events (effectively corresponding to a rank-sum test). Empirical activity p-values (over all CREs within a replicate) were Benjamin-Hochberg corrected to obtain a false discovery rate. Corrected empirical p-value without stratification over clusters was similarly performed (mean probability that expression from the CRE over all integration was below that of basal control null bootstrap values). To identify active CREs, we considered elements with either per-cluster maximum expression FDR <10% in all three replicates and/or all cells expression FDR<1% (higher statistical power from more integration events) in all three replicates. 58/204 CREs passed these stringent criteria and were considered active in excess of our basal expression controls.

To assess CRE specificity, a similar approach was taken, but instead of performing comparison to basal promoters, comparisons were performed to datasets with permuted cell cluster identities. For each CRE, $10^4$ repeats were performed were a bootstrapped resampled (no cluster identity permutation) set of integration events was generated, and the fold-change in reporter expression (average normalised mBC UMI) between the maximum expression cluster and the rest of cells was computed. The corresponding quantity, but for a cluster-identity permuted sampling was also performed for each sampling. The specificity empirical p-value for each CRE was taken as the average (over resamplings) probability that the cluster permuted fold-changes in expression (null distribution over all permutations) was higher than the non-permuted one. As before, these empirical p-values were Benjamin-Hochberg corrected (over all CREs, separately for different biological replicates). CREs that were identified as active were further marked as specific if in all biological replicates, the reporter expression fold-change (maximum cluster vs. all other cells) was >5 and the permutation derived FDR < 10%, leading to 9/58 elements.

To systematically assess whether elements had pleiotropic activity (active in multiple cell types), we computed the fold-change in expression in all pairs of clusters vs. the rest of cells, storing the maximum fold-change value and specific cluster pair for each CRE and biological replicate. The median (across biological replicates) fold-changes for pairs vs. individual clusters were compared. Only a single CRE had a paired/single cluster fold-change in excess of 3x was *Lamc1*:chr1_12189 (also elevated: 2.6x for *Foxa2*:chr2_13858 which displayed some activity in visceral endoderm in addition to parietal, **Fig 4e** second row; and 1.5x for *Sox2*:chr3_2007 which had some activity in epiblast cells, **Fig. 3f**). Other elements showed no substantial excess activity in pairs over single

clusters (95% percentile in pair/single fold-changes was at 1.3x and 90% percentile at 1.1x). Permutation tests similar to above confirmed Lamc1 bifunctional activity was highly significant (non permuted fold-change highest in all $10^3$ samplings), leading to a final set of 10/58 active CREs labelled as specific.

To summarise the function of individual CREs, the median activity (defined as the maximum cluster mean reporter expression) and specificity (defined as the fold-change between maximum cluster mean reporter expression vs. mean reporter expression in the rest of cells) across the three biological replicates was determined (shown in **Fig. 4a**).

Some elements were active and/or specific in only a subset of replicates (those marked in **Ext. Data Fig. 7b**, e.g., *Bend5*:chr4_8174, *Foxa2*:chr2_13820, *Sox17*:chr1_77, *Bend5*:chr4_8179, *Lama1*:chr17_7791, *Lamc1*:chr1_12185). These are likely candidates for active elements (falling below our limit of detection possibly because too few integration events were captured due to uneven CRE representation), but were not retained to maintain stringency in our downstream analyses. Quantification summary can be found in **Supp. Data 5**.

Pseudobulk expression in separate cell-types (e.g., **Ext. Data Fig. 5b**) were determined as the average normalised mBC UMI counts over all cells with detected reporters belonging to GEx clusters identified and annotated in **Ext. Data Fig. 4a**.

**Statistics & Reproducibility**

No statistical method was used to predetermine sample size. The experiments were not randomised. The Investigators were not blinded to allocation during experiments and outcome assessment.

Benchmarking experiments and optimization experiments in cell lines were carried out in two independent replicates, with reproducible results. Experiments in mEBs were carried out in biological triplicates, with reproducible results. Bulk MPRA experiments comparing Pol III circular and linear barcodes were carried out in independent biological duplicates, with reproducible results. Singleton validation experiment was performed as a single experiment (with one independent differentiation for the 8 tested constructs). Multiple EBs within each condition however showed expected behaviour (cell-type specific expression).

Detailed statistical tests and quantitative treatment of data are otherwise described at relevant sections in the Methods and supplementary methods (**Supp. Note 6**).

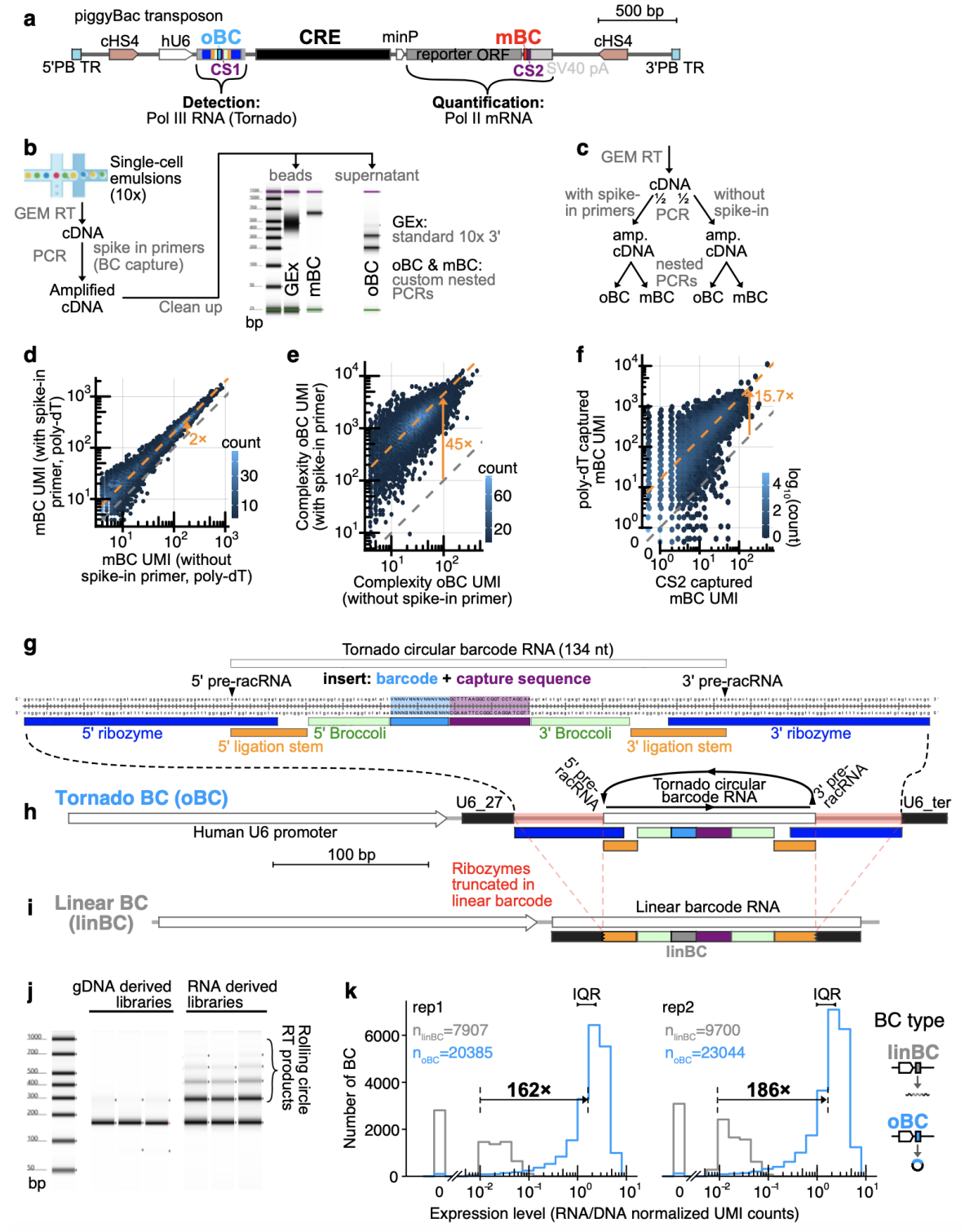
No data were excluded from the analyses apart from a single sample/time point from bulk MPRA in mEBs (day 20, replicate 2B1, first round of experiment). This library had been generated from a lower amount of starting RNA (yield from that extraction had been lower). Inspection of read counts to basal promoters showed drastically higher apparent activity compared to other samples, suggesting that signal in the RNA originated from trace contaminant genomic DNA, which had a

disproportionate weight in that sample due to the low starting RNA quality. This sample was thus excluded from downstream analysis.

**DATA AVAILABILITY:** Raw sequencing data and processed files generated in this study have been deposited to GEO, with accession number GSE217690 and to the IGVF data portal (accession: TSTDS25687808, TSTDS23601776). Published data used: transcription factor binding data (Uniprobe[104]: *Gata4*[105] UP01372, *Sox17*[106] UP00014, *Foxa2*[106] UP00073), mouse embryo *in vivo* scRNA-seq[61] (obtained from R library: "MouseGastrulationData") and scATAC-seq[53] (GEO: GSE205117). Promoter control scQer libraries (p027, p028, p029, p041, p042) and cloning intermediate libraries with pre-associated list of oBC-mBC (p025, p043) have been deposited to Addgene (respective identifiers: 194096, 194097, 194098).

**CODE AVAILABILITY:** Code and scripts used for analyses have been deposited on github (shendurelab/scQers), together with the maps of plasmids and custom sequencing amplicons structures used in this work.

***Extended data figures:***



**Extended Data Figure 1: Dual RNA reporter cassette, single-cell assay, barcode capture optimization, and comparison of circularised vs. linear U6-driven barcodes**

**a** At-scale schematic of the dual RNA reporter cassette in piggyBac transposon (between terminal repeats: PB TR). Flanked by convergent insulators (core chicken hypersensitive site-4 from beta-globin locus, cHS4[48]), the human U6 (hU6) driven Tornado barcode cassette (oBC-CS1, details shown in panel **g**) is co-directionally placed upstream of the CRE library driving an open reading frame-containing reporter transcript (puromycin-P2A-GFP in the case of the promoter series in cell lines, **Fig. 2a**, and GFP alone for mEB experiment, **Fig. 3b**), barcoded in its 3' untranslated region (mBC) upstream of an inserted capture sequence 2 (CS2), and of the SV40 polyadenylation sequence (SV40 pA).

**b** Schematic of the single-cell reporter assay. After 10x Genomics (V3.1, 3' gene expression with feature barcode) GEM reverse transcription, primers (specific to oBC and mBC RNAs) are spiked-in the cDNA amplification mix[107]. Post-cDNA amplification, in addition to standard gene expression (GEx) library generation, nested PCRs from bead fraction (mBC) and supernatant (oBC) are performed to obtain custom single-cell reporter libraries. Amplification of barcodes proceed from different fractions as reporter mRNAs harbouring the mBC are long (>800 bp), purifying with the beads, whereas oBC are short (134 bp), remaining in the supernatant. Example tapestation traces of resulting libraries are shown (showing laddering products from oBC libraries).

**c** Experiment to assess improvement in UMI capture by spiking in primers in initial cDNA amplification. For the experiment with promoter series in cell lines (**Fig. 2a**), replicate B's cDNA was split in two prior to cDNA amplification. One half, replicate B1, received spike-in primers to the oBC and mBC reporters, and the other half, replicate B2, did not. An additional round of

PCR downstream of the first cDNA amplification was performed to obtain libraries in replicate B2.

**d**-**e** Comparison of number of UMIs captured for the same cell barcode and reporter barcodes between replicates B1 (with spike-in primers) and B2 (without spike-in primers) for mBC (panel **d**: 2.0× median increase in UMIs captured. n=8395 mBC-cell barcode pairs with >3 UMI) and oBC (panel **e**: 45× median increase in UMIs captured. n=19323 oBC-cell barcode pairs with >3 UMI), respectively. The higher boost in capture resulting from spike-in primers for the oBC vs. mBC was likely due to the circular nature of the barcode: given the absence of 5' end from which template switching can occur from oBC RNAs, the initial cDNA amplification (primed from the template switching oligo) effectively cannot happen except from the low abundance linear intermediates towards oBC formation; in contrast, the spike-in primers enable directly targeting sequences flanking the barcode in the circular oBC.

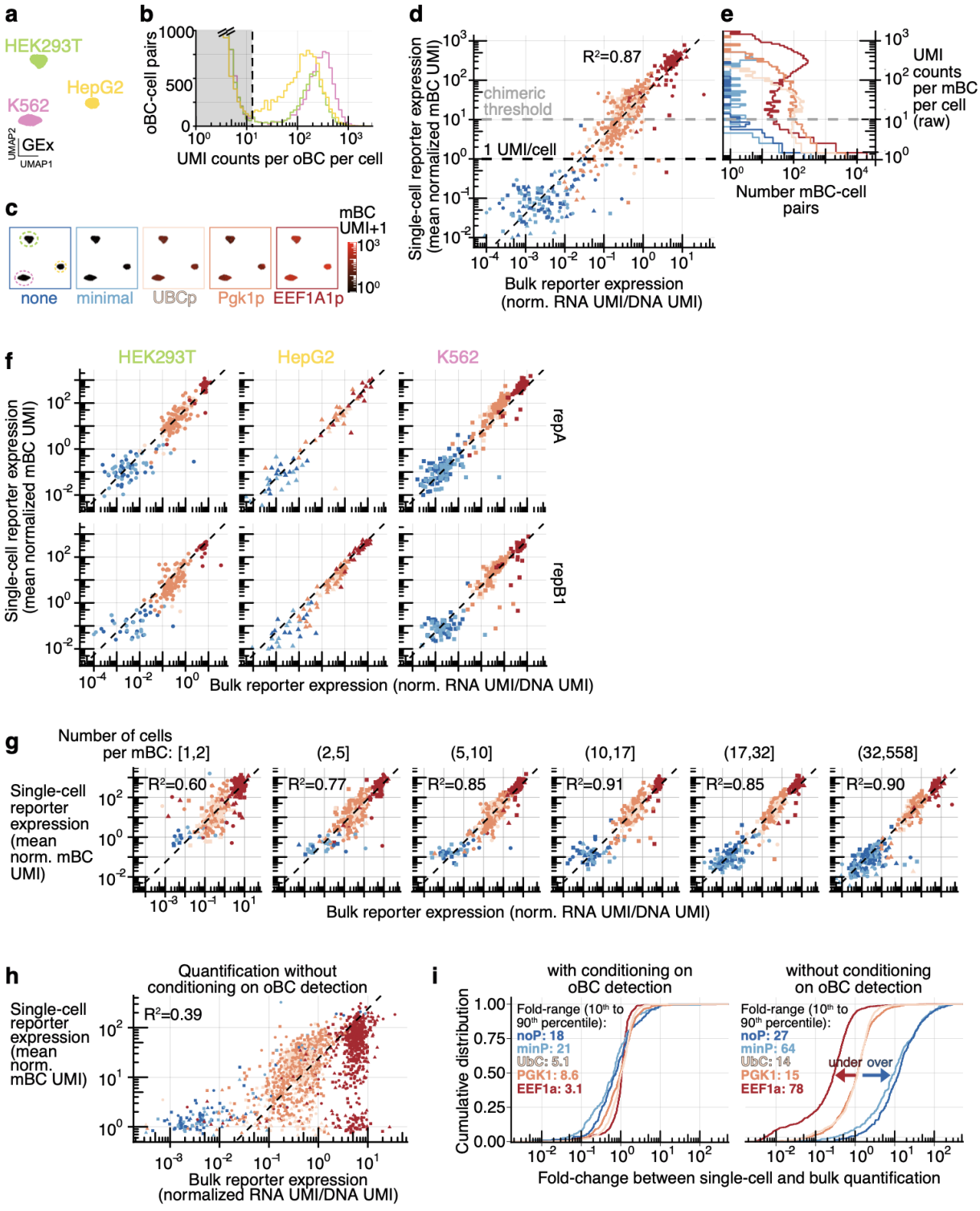
**f** Comparison of captured mBC UMI from poly-dT vs. capture sequence 2 (CS2) on-bead reverse transcription primers (for the same mBC-cell barcode pairs). As expected from primer stoichiometry on beads, >15× increase in captured mBC UMI is seen from the poly-dT vs. CS2 primers (n=21492 mBC-cell barcode pairs with poly-dT and CS2 mBC >0 across both replicate A and B1). CS2 thus adds marginal value for capture of the Pol II-derived polyA-tailed mBC transcripts.

**g** Sequence of the Tornado system[34] with 16 bp barcode (5' VNNNVNNNVNNNVNNN, light blue) and downstream capture sequence 1 (CS1; burgundy) inserted in the loop of Broccoli. 5' and 3' (pre-racRNA) ends cleaved by ribozymes prior to circularization are highlighted (black carets). The circular product is 134 nt long.

**h-i** Schematic of the human U6 promoter driven cassettes tested in a head-to-head MPRA experiments (integrated via piggyBac) to compare expression of the circular version of the barcode (Tornado barcode, or oBC, **h**) to the linear barcode (linear barcode, linBC, **i**), which is the same construct but with 'Twister' ribozymes removed (red highlight in **h**).

**j** Representative tapestation (three out four libraries generated from independent biological replicates shown, two of which were sequenced and shown in panel **k**) traces of genomic DNA-derived vs. RNA-derived amplicon libraries prepared from the oBC vs. linBC MPRA experiment. RNA-derived libraries show clear rolling circle reverse transcription products laddering of the expected periodicity (+134 bp) expected from circular RNAs.

**k** Distribution of MPRA-derived activity estimates (RNA/DNA normalised UMI) for the thousands of different, well-represented (>50 DNA UMI) barcodes of both types (hU6-driven oBC [blue] vs. hU6-driven linBC [grey]) as assessed by bulk MPRA, highlighting both the large difference in steady-state expression (>150× difference in median between linBC and oBC), and tight distribution (interquartile range <3×) for the oBC. Sub-panels correspond to two independent biological replicates.




**Extended Data Figure 2. Assessment of accuracy of single-cell dual RNA reporters**

**a-d** Same as **Fig. 2a-c**, but with data from replicate B1. **a**: Gene expression, **b**: oBC UMI count distribution, **c**: single-cell measure of reporter expression (GEx UMAP projected), **d**: comparison of bulk vs. single-cell quantification of mBC quantification.

**e** Raw distribution of UMI counts per mBC per cell barcode (for valid mBC and cell barcodes pairs, not conditioning on oBC detection) stratified by associated promoter. The 10 mBC UMI/cell threshold ("chimeric threshold") reflects that even for highly expressed promoters, mBC UMI counts rise below that point, as a result of chimeric amplicons generated during library preparation. Without conditioning on oBC detection, these molecular species limit the dynamic range of reliable measurements with one-RNA reporters (see panel **h**).

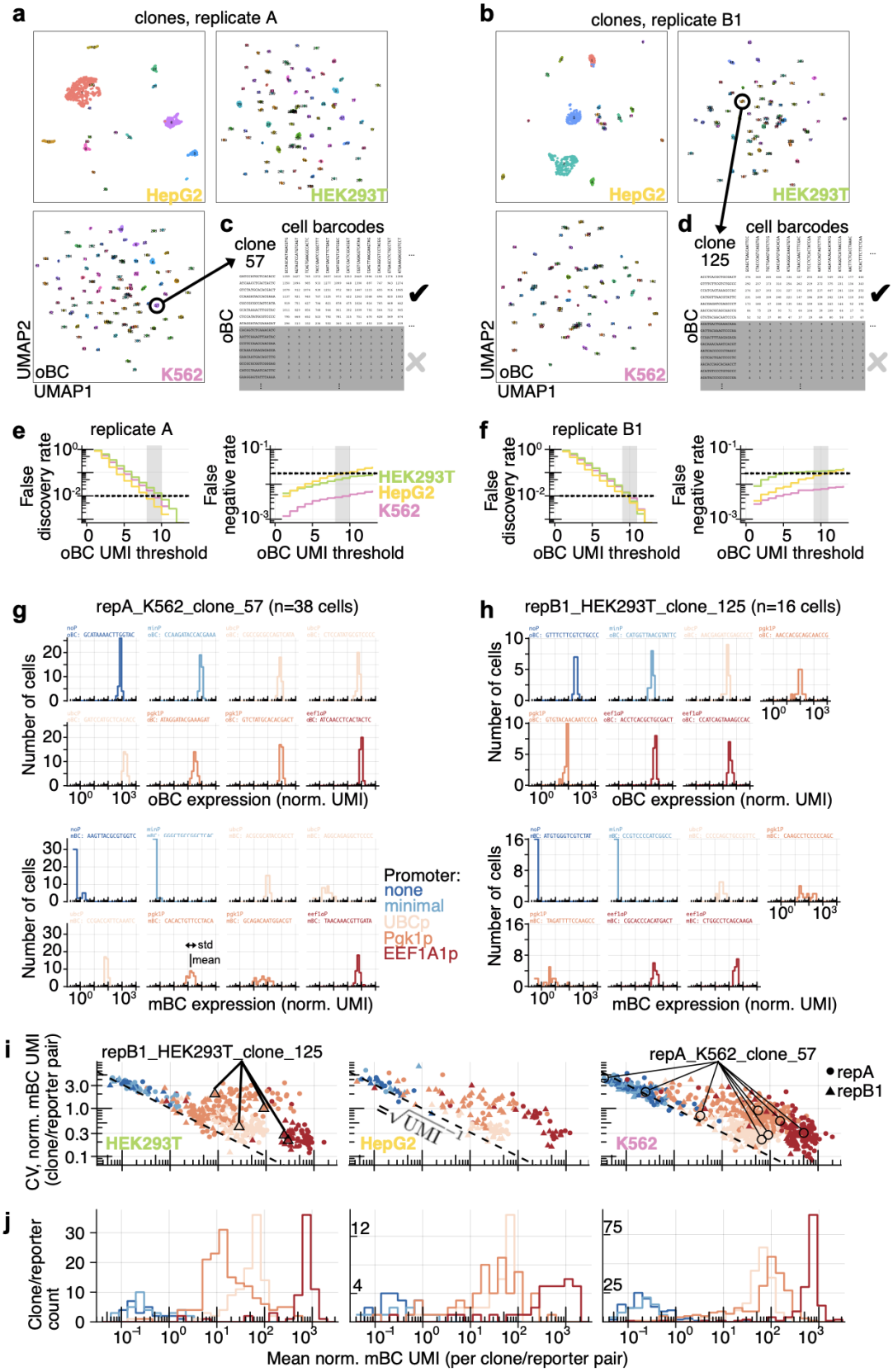
**f** Comparison of bulk MPRA quantification (x-axis, RNA over DNA normalised UMI counts) vs. single-cell quantification (y-axis: average normalised mBC UMI over all cells with detected matched oBC), same as **Fig. 2g**, but stratified by replicates and cell lines. Each point corresponds to an individual mBC, coloured by its associated promoter. Well-represented mBC are included (>100 bulk DNA UMI, >0 measured mBC UMI in single cells, and ≥5 single-cell integrations detected). The diagonal dashed line follows a 1:1 slope.

**g** Assessment of reporter mRNA measurement accuracy vs. number of integration events captured (both replicates). Single-cell vs. bulk quantification (same as **Fig. 2g** and panel **d**), but stratified by the number of cells per mBC over which the single-cell measurement is averaged (split in equal number of mBC bins). Even with as few as 5 to 10 cells captured per mBC, the correspondence with bulk measurement is on par with estimates from more highly represented mBCs ($R^2$ on log-transformed values ≥ 0.85).

**h** Single-cell vs. bulk quantification of mBC expression without conditioning on oBC detection (assuming all mBC capture events are valid, both replicates). In contrast to oBC conditioned

measurements, quantification has a hard floor at 1 UMI/cell (slight variation around 1 from gene expression normalisation) and a limited dynamic range (y-axis spans ≈200× compared to $>10^4$× with oBC conditioning, c.f., **Fig. 2g** and panel **d**). Only well-represented mBC are included (same criterion as **Fig. 2g**: >100 DNA UMI bulk, ≥5 cells with mBC detected). Dashed line marks the 1:1 slope, highlighting systematic biases.

**i** Cumulative distribution of fold-change between single-cell and bulk mBC quantification (median normalised), for both replicates, with (left) and without (right) conditional oBC detection. While the quantification conditioning on oBC is largely unbiased (centred and close to 1), quantification is biassed at the high (underestimation for highly expressed EEF1A1 promoter, red arrow) and low (overestimation for low expression minimal/no promoters, blue arrow) ends of the expression spectrum. In addition to removing systematic biases, conditioning on oBC also reduces variability (quantified as the spread in fold-change, with the range spanned from $10^{th}$ to $90^{th}$ percentile for each promoter displayed on plot).


a
clones, replicate A
HepG2
HEK293T
K562
UMAP2
UMAP1
oBC
b
clones, replicate B1
c
cell barcodes
clone 57
d
clone 125
e
replicate A
False discovery rate
False negative rate
oBC UMI threshold
f
replicate B1
g
repA_K562_clone_57 (n=38 cells)
Number of cells
oBC expression (norm. UMI)
mBC expression (norm. UMI)
Promoter:
none
minimal
UBCp
Pgk1p
EEF1A1p
std
mean
h
repB1_HEK293T_clone_125 (n=16 cells)
i
repB1_HEK293T_clone_125
repA_K562_clone_57
repA
repB1
CV, norm. mBC UMI (clone/reporter pair)
√UMI
j
Clone/reporter count
Mean norm. mBC UMI (per clone/reporter pair)

**Extended Data Figure 3. Benchmarking oBC detection and mBC capture precision with clonal analysis**

**a** and **b** oBC expression space UMAP from cells assigned to high-confidence clones (coloured by mapped clone identity) with at least three cells assigned, separated by cell lines. Panel **c**: replicate A (K562: 105 clones, 1430 cells; HEK293T: 92 clones, 1330 cells; HepG2: 17 clones, 916 cells), Panel **e**: replicate B1 (K562: 90 clones, 738 cells; HEK293T: 81 clones, 689 cells; HepG2: 21 clones, 537 cells).

**c** and **d** Example of raw (error corrected) UMI counts (table truncated) per cell barcode and oBC across assigned cells in clones highlighted respectively in panels **a** and **b** (oBC ordered from high to low counts). Panel **c**: clone repA_K562_clone57 with 38 cells assigned. Panel **d**: clone repB1_HEK293T_clone_125 with 16 cells assigned. Grey shading delineates oBCs not assigned to the clones, highlighting the sharp distinction in UMI counts.

**e** and **f** Systematic analysis of oBC dropout across all high-confidence clones. False discovery rate (left, false positives/[true positives + false positives]), and false negative rate (right panels, false negatives/[false negatives + true positives]) as function of the oBC UMI threshold used for detection. Analyses are performed on high-confidence clones represented by at least 3 cells. Consensus reconstructed clonotypes are taken as ground truth and cells are assigned to these clonotypes with stringent threshold to remove doublets, but loose threshold to allow for up to 50% oBC dropouts per clone. At an FDR of 1% (grey shading), there are about 2% dropout (false negative rate) observed (slightly reduced performance from replicate B1 likely from halved complexity, see **Ext. Data Fig. 1c**). Panel **e**: replicate A, Panel **f**: replicate B1.

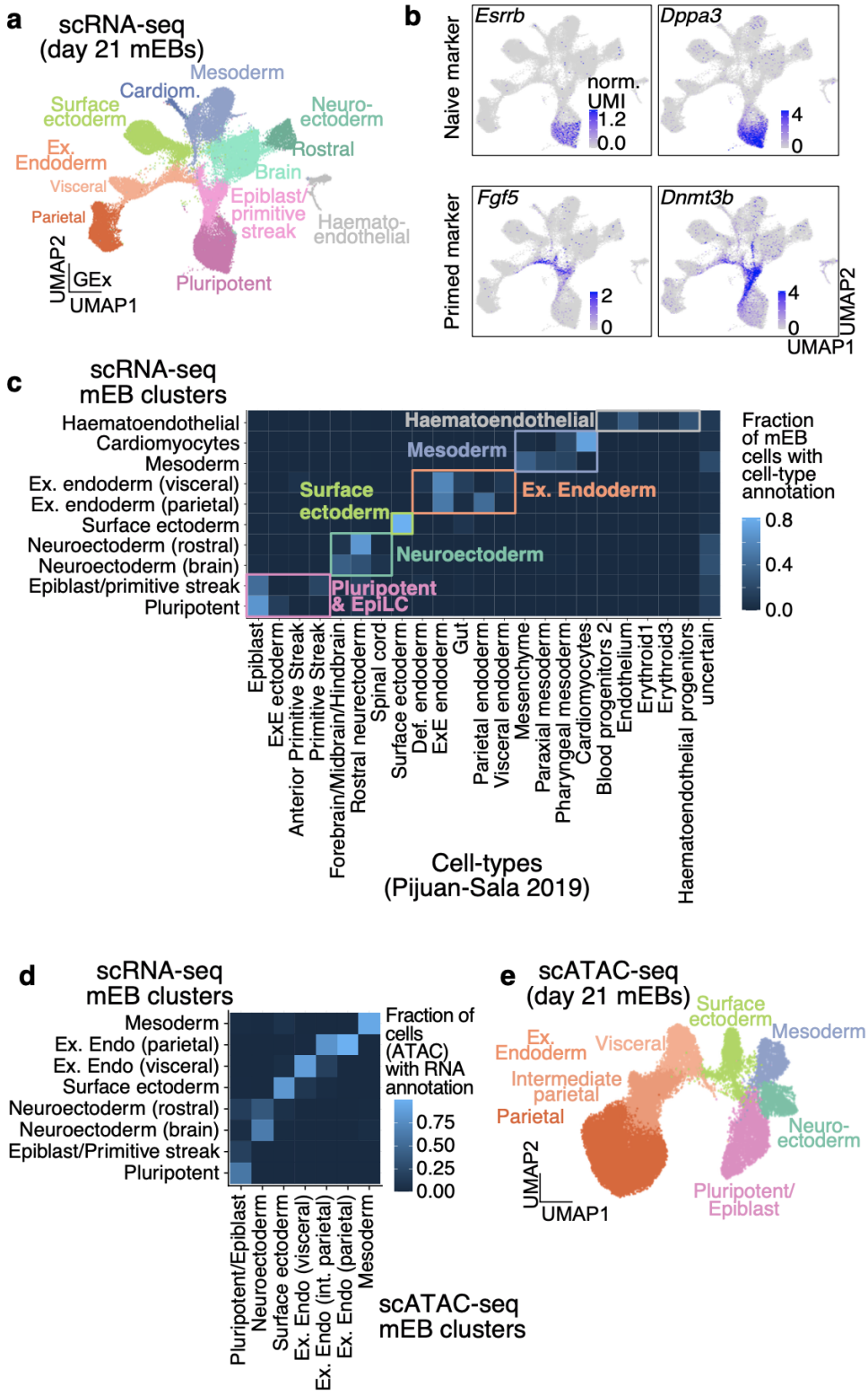
**g** and **h** Example of mBC (top) and oBC (bottom) UMI count distributions across all cells assigned to specific clones (highlighted in panels **a** and **b**). Each sub-panel corresponds to a

reporter integrated in the clone. Panel **g**: clone repA_K562_clone57, with 8 integrated reporters. Panel **h**: clone repB1_HEK293T_clone_125, with 7 integrated reporters. Panels in respective positions within the oBC and mBC set are matched (e.g., in repA_K562_clone57, EEF1A1 promoter with oBC:ATCAACCTCACTACTC and mBC: TAACAAACGTTGATA).

**i** Coefficient of variation analysis of mBC UMI count measurements across all reporter-clone pairs stratified by cell line (left: HEK293T, middle: HepG2, right: K562). Mean over standard deviation (see panel **g** bottom: Pgk1 promoter with mBC:CACACTGTTCCTACA as schematic of both quantities) of normalised mBC UMI counts for reporters in clones as a function of mean normalised mBC UMI (reporters with >0.05 mBC UMI mean expression in clones with >4 cells assigned; replicate A: K562: 392 reporters from 83 clones, HEK293T: 198 reporters from 70 clones, HepG2: 58 reporters from 12 clones; replicate B1: K562: 213 reporters from 58 clones, HEK293T: 123 reporters from 51 clones, HepG2: 95 reporters from 14 clones). Dashed line indicates the Poisson counting scaling $CV=\sqrt{(UMI\ count)^{-1}}$. Each point represents the quantification for a specific reporter within a clone, with point shape marking replicates and colour promoter type. As examples, reporters shown in panels **g** (clone repA_K562_clone57) and **h** (clone repB1_HEK293T_clone_125) are highlighted in black (no and minimal promoter reporters from repB1_HEK293T_clone_125 have 0 mBC UMI and therefore do not appear).

**j** Assessment of position-dependent variability of integrated reporters. Panels show the distribution in mean normalised mBC UMI (expression) across reporters integrated over different clones, stratified by cell line (left: HEK293T, middle: HepG2, right: K562) and promoter type (colour). Same clone/reporter pairs as panel **i**. To account for halved library complexity in replicate B1 (see description in **Ext. Data Fig. 1c**), reporter expression values

from those clones were multiplied by two.



**Extended Data Figure 4. Molecular profiling and integration of single-cell data from 21-day mouse embryoid bodies**

**a** UMAP of scRNA-seq data from quality-filtered cells from scQer-integrated, day 21 mEBs (same as **Fig. 3c**) annotated with fine-resolution cell types derived from label transfer of *in vivo* dataset[61], as shown in panel **c**. These cluster definitions are used to quantify CRE activity over cell types (e.g., **Fig. 4a**, **Ext. Data Fig. 5b**).

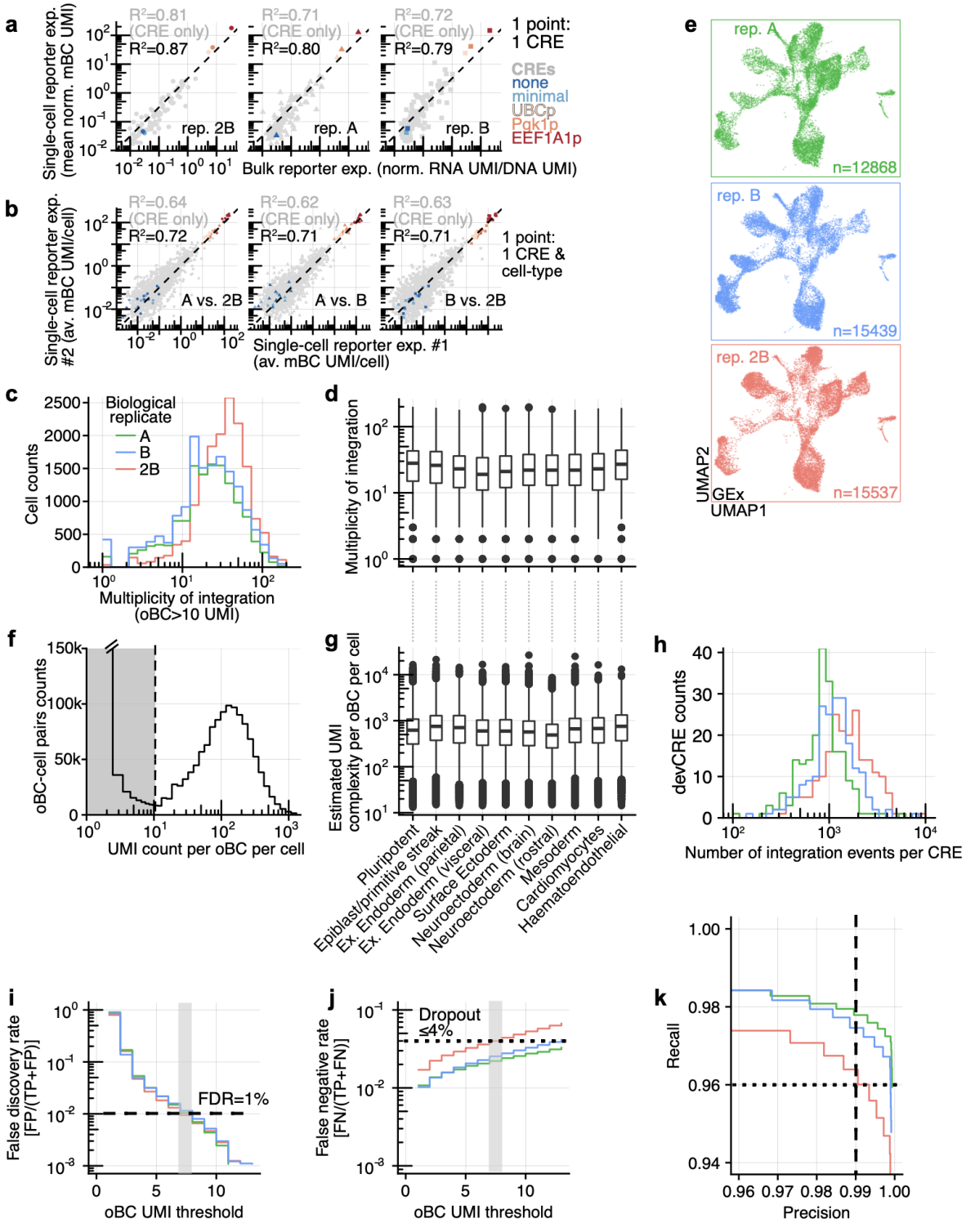
**b** Example of naive and primed pluripotent stem cell marker gene expression (normalised UMI counts) displayed on UMAP, used to annotate the respective cells as pluripotent and epiblast/primitive streak.

**c** Heatmap displaying fraction of mEB-derived cells (from each cluster in panel **a**) with label transferred to cell-types from *in vivo* data from Pijuan-Sala *et al*[61]. Cell types with no associated cells in mEBs (with maximum fraction < 5%) are not included for brevity. Clusters coarse-grained for representation (**Fig. 3c**) are boxed. Uncertain column corresponds to cells that had ambiguous label transfer. The mEB cluster marked as pluripotent was manually annotated from specific expression of canonical marker genes[63] in those cells (panel **b**) as a result of a lack of naive mESCs in the integration dataset.

**d** Integration of scATAC-seq and scRNA-seq for cluster annotation. Heatmap showing fraction of nuclei from scATAC-seq-derived clusters predicted to be from cell-type identified in scRNA-seq data, displaying unambiguous matches. Certain minor cell types (cardiomyocytes, haematoendothelial) were not found at high proportion in the scATAC-seq data.

**e** UMAP of scATAC-seq data from quality filtered cells (n=46408, two biological replicates) from day 21 mEBs. Clusters are labelled based on integration with scRNA-seq data (panel **a**,

panel e).




**Extended Data Figure 5. Quality metrics of single-cell reporter assay in mEBs**

**a** Comparison between single-cell (average normalised mBC UMI count across all cells with detected reporter) and bulk quantification (day 21 samples, RNA/DNA ratio of summed 1% winsorised UMI counts across all barcodes) for well-represented CREs (>100 integrations, >30 total mBC UMI in single-cell assay, and >35 mBC with at least 20 DNA UMI in bulk assay) stratified by biological replicate. CREs (grey) and promoters coloured according to **Fig. 2a**, dashed marks a 1:1 slope. $R^2$ on log-transformed values, including exogenous promoters (black) or not (grey).

**b** Comparison of per-cell type reporter quantification (average normalised mBC UMI over cells in clusters of **Ext. Data Fig. 4a**) for CREs with >0 activity stratified by biological replicates. Each point corresponds to a CRE in a cell-type (10 points per CRE). $R^2$ on log-transformed values, including exogenous promoters (black) or not (grey).

**c** Distribution of multiplicity of integrations (number of oBC with >10 UMI per cell) across individual cells and stratified by replicate (median: repA=20, repB=19, rep2B=31). High MOI in rep2B likely results from further selecting mCherry+ cells (1% co-transfection), not performed for replicates A and B.

**d** Distribution (box plot, centre marks the median, edges of boxes define the 25th and 75th percentiles, whiskers extend to 1.5 times the interquartile range) of multiplicity of integration stratified by cell types (see **Ext. Data Fig. 5a**). Cell type annotations same as in panel **g**. Each box plot is constructed from all cells assigned to a cell type (n=43844 total number of cells over all cell types from three independent experiments).

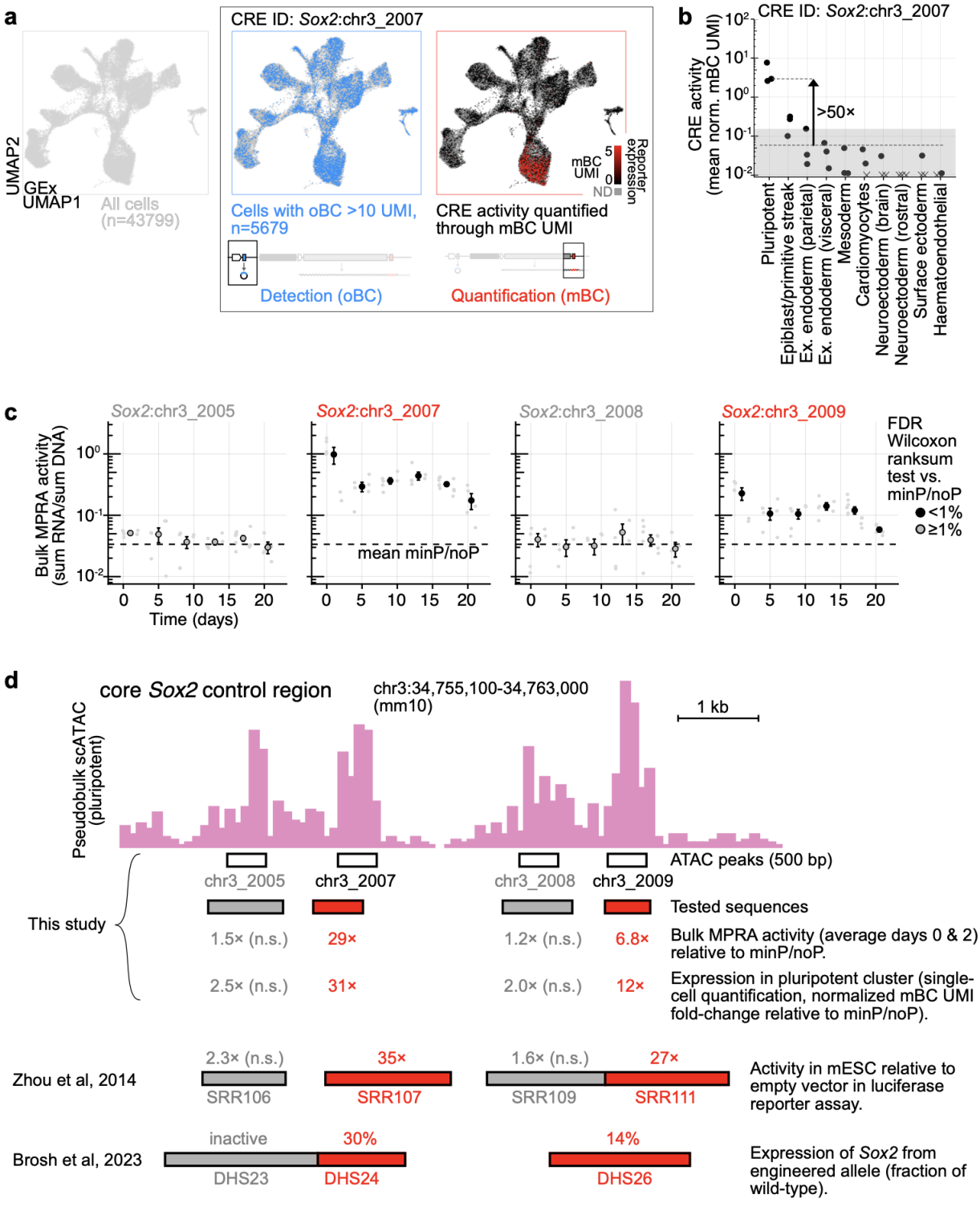
**e** scRNA-seq gene expression UMAP (same as **Fig. 3c**) stratified by biological replicate (no batch correction) showing reproducibility of cell-types obtained in embryoid bodies derived from reporter-containing mESC. Number of cells for each replicate indicated in each panel.

**f** Distribution of oBC UMI counts per cell (similar to **Fig. 2c**) highlighting robust circular barcode RNA capture in differentiated cells. Sharp bimodality and high signal-to-noise enables high-recovery reporter integration detection.

**g** Box plot (centre marks the median, edges of boxes define the 25th and 75th percentiles, whiskers extend to 1.5 times the interquartile range) of estimated total UMI complexity (zero-truncated Poisson) for each captured oBC (>10 UMI) in all cells stratified by cell type, displaying similar levels irrespective of cell type. Each box plot is constructed from all cells assigned to a cell type (n=43844 total number of cells from three independent experiments).

**h** Distribution of number of captured integration events per CRE (not including exogenous promoter series, determined from oBC UMI >10 from oBC-associated CRE) stratified by replicates, showing reasonably uniform coverage across profiled elements.

**i-k** Precision-recall analysis of oBC detection (similar to **Fig. 2h**, **Ext. Data Fig. 3e-f**) for mEB-derived cells. Despite only replicate 2B being directly bottlenecked, replicates A and B also displayed (modest) clonal expansion, which enabled analysis of oBC dropout in these samples as well. High-confidence clones with at least two assigned cells are included (repA: 600 clones, 3977 cells; repB: 635 clones, 6465 cells; rep2B: 325 clones, 8518 cells), with results unchanged if restricting to more highly represented clones. Consensus clonotypes served as ground truth for analysis. Panels **h** and **i** respectively show the false discovery rate (FP/[FP+TP]) and false negative rate (FN/[FN+TP]) as a function of the UMI threshold used to assign barcodes to cells. At 1% FDR, false negative (dropout) is less than 4%. oBC libraries from replicate 2B were not sequenced as deeply (average saturation 6.0% vs. 18.7%), suggesting that part of the dropout is due to incomplete sequencing coverage.


a
CRE ID: Sox2:chr3_2007
UMAP2
GEx UMAP1
All cells (n=43799)
Cells with oBC >10 UMI, n=5679
CRE activity quantified through mBC UMI
mBC UMI
Reporter expression
ND
Detection (oBC)
Quantification (mBC)
b
CRE ID: Sox2:chr3_2007
CRE activity (mean norm. mBC UMI)
>50×
Pluripotent
Epiblast/primitive streak
Ex. endoderm (parietal)
Ex. endoderm (visceral)
Mesoderm
Cardiomyocytes
Neuroectoderm (brain)
Neuroectoderm (rostral)
Surface ectoderm
Haematoendothelial
c
Sox2:chr3_2005
Sox2:chr3_2007
Sox2:chr3_2008
Sox2:chr3_2009
Bulk MPRA activity (sum RNA/sum DNA)
mean minP/noP
Time (days)
FDR Wilcoxon ranksum test vs. minP/noP
<1%
≥1%
d
core Sox2 control region
chr3:34,755,100-34,763,000 (mm10)
1 kb
Pseudobulk scATAC (pluripotent)
ATAC peaks (500 bp)
chr3_2005
chr3_2007
chr3_2008
chr3_2009
This study
Tested sequences
1.5× (n.s.)
29×
1.2× (n.s.)
6.8×
Bulk MPRA activity (average days 0 & 2) relative to minP/noP.
2.5× (n.s.)
31×
2.0× (n.s.)
12×
Expression in pluripotent cluster (single-cell quantification, normalized mBC UMI fold-change relative to minP/noP).
Zhou et al, 2014
2.3× (n.s.)
35×
1.6× (n.s.)
27×
SRR106
SRR107
SRR109
SRR111
Activity in mESC relative to empty vector in luciferase reporter assay.
Brosh et al, 2023
inactive
30%
14%
DHS23
DHS24
DHS26
Expression of Sox2 from engineered allele (fraction of wild-type).

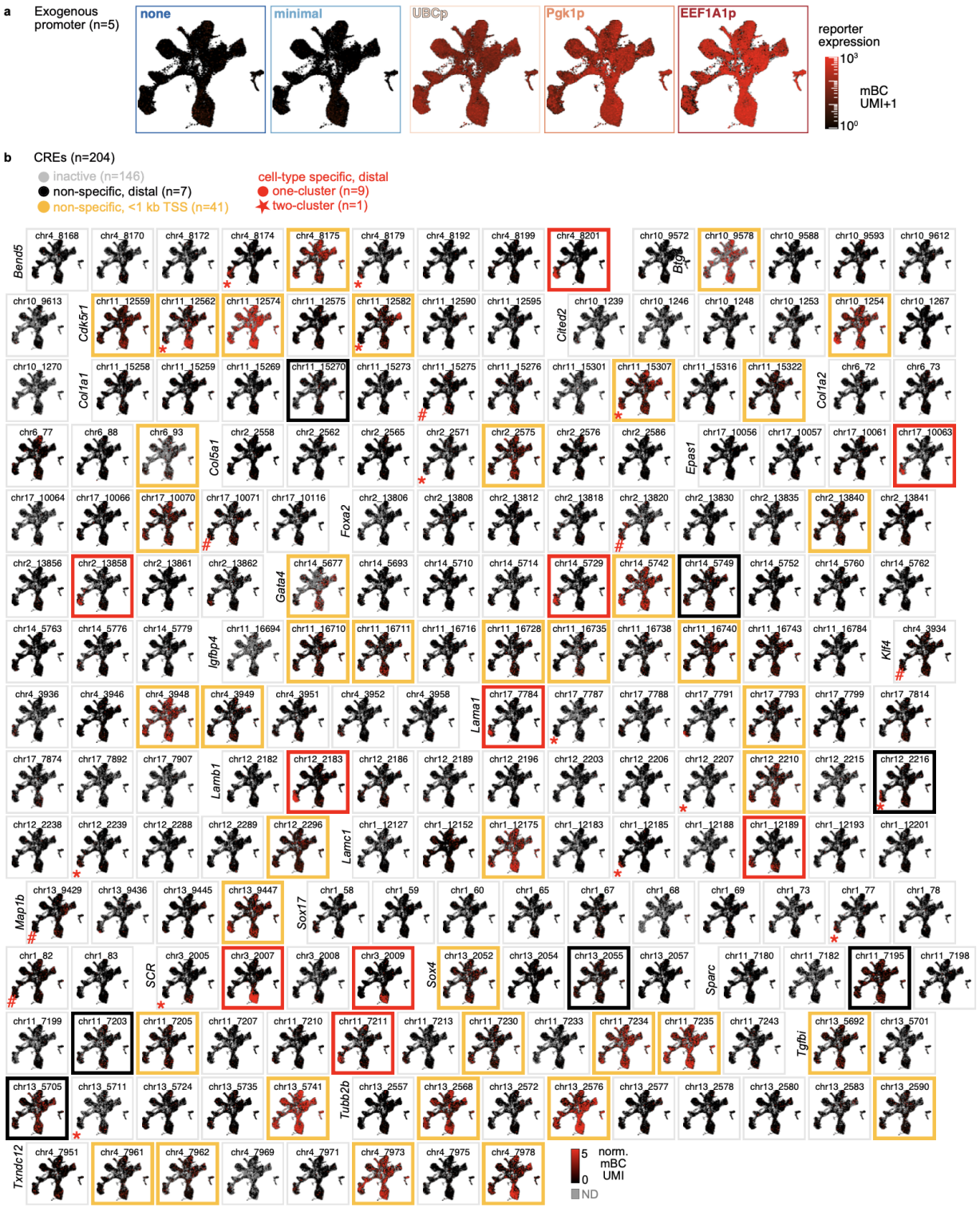
**Extended Data Figure 6. Details on activity of constituent elements of the *Sox2* control region**

**a** Illustration of the steps to construct a single-cell map of CRE activity for a given regulatory element. Left: All cells passing quality filters are initially considered. Middle: Reporter detection. The list of oBCs associated with the CRE of interest (here *Sox2*:chr3_2007, see **Fig. 3f**) from the predetermined oBC-CRE-mBC dictionary are identified. Cell barcodes with one (or more) CRE-associated oBC with >10 UMI are retained (n=5679), shown in blue on the UMAP (grey corresponding to cells with no detected reporters to the CRE of interest). Right: Expression quantification. From the oBC-CRE-mBC dictionary, the UMI counts to CRE-associated mBC are collected. In cases where multiple reporters to the same CRE (but different oBC-mBC pairs) are detected in the same cell, the average mBC UMI is taken. To correct for the fact that some cell types have more RNA (or other technical factors), we normalise the mBC expression by the total UMI to the transcriptome for each considered cell. The resulting single-cell reporter expression can then be layered on the low dimensional projection (black low to high red), enabling visualisation of CRE activity across the manifold of cell states in the system.

**b** Quantification of the average reporter expression (average normalised mBC UMI, see panel **a**) across cells from different cell types (defined as clusters in **Ext. Data Fig. 4a**). Each dot corresponds to a biological replicate. Crosses correspond to cell types/replicates with average expression below 0.01 mBC UMI/cell. Arrow marks the fold change in expression between the maximum cluster (pluripotent) and the rest of cells (defined as specificity in **Fig. 4a**). Grey shading marks the noise floor determined from variability from the basal expression controls (minimal and no promoter).

**c** Bulk MPRA quantification of the four constituents of the core *Sox2* control region (see **Ext. Data Fig. 7** for all CREs), showing consistent results with single-cell quantification (inactive: *Sox2*:chr3_2005, *Sox2*:chr3_2008; active: *Sox2*:chr3_2007, *Sox2*:chr3_2009). Small grey points mark individual replicates and time points. Large points are the average over n= 3 biological replicates from consecutive time points, and are filled if significantly above the basal expression controls (one-sided ranksum test, B-H corrected, <1% FDR). Error bars show the standard error of the mean. Dashed line indicates the mean of basal expression control (minimal and no promoters). The observed decrease in activity over time for *Sox2*:chr3_2007 and *Sox2*:chr3_2009 is consistent with pluripotent cells being progressively depleted from the population, thereby leading to decreased activity when averaged over all cells in bulk.

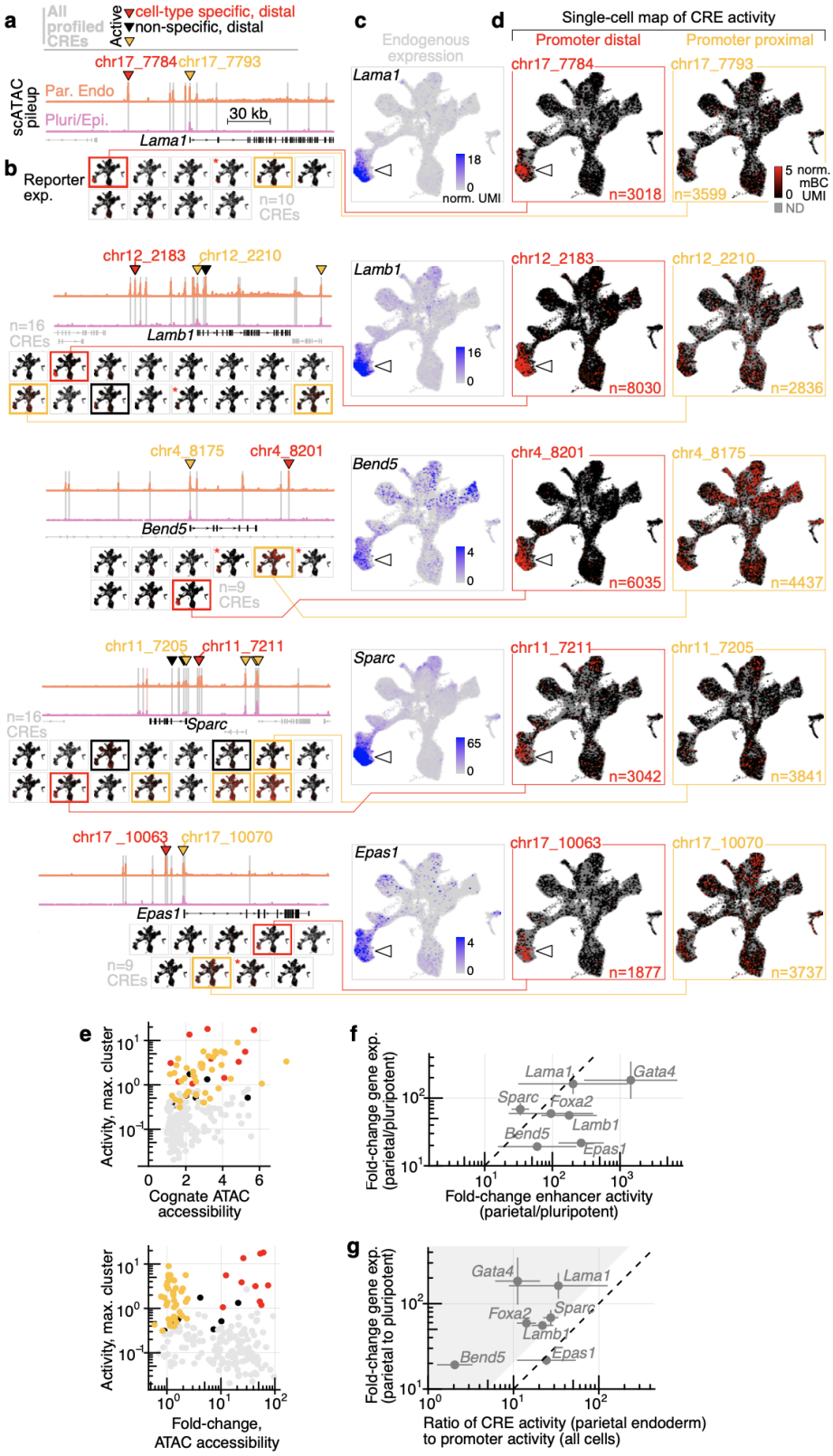
**d** Sox2 control region scATAC pseudobulk pileup in pluripotent/epiblast cluster (zoom in of **Fig. 3e**). Under pileup, elements tested (in the same genomic position reference frame as the pileup, **Supp. Data 4** for positions) are indicated both from this study (top: 500 bp regions peak from ArchR pipeline; bottom: PCR-amplified tested sequences), and two previous studies quantifying reporter activity, Zhou et al[57], and Brosh et al[7]**.** Grey regions were not found to be significantly active. Red regions were found to have activity in pluripotent cells (measured activity is indicated). *Sox2*:chr3_2007 from this study was not entirely nested in previously tested elements (SRR107 and DHS24), suggesting that even higher activity than measured might be achievable with a more inclusive element. The slight misalignment from the ATAC peak for *Sox2*:chr3_2007 resulted from lack of identifiable specific PCR cloning primers in the immediate 3' region.




**Extended Data Figure 7. Systematic characterization of 204 putative CREs in mouse embryoid bodies**

**a** Single-cell reporter expression (average normalised mBC UMI per cell) for the five exogenous promoters used as internal controls. Colour scale is logarithmic (with a pseudocount of 1).

**b** Single-cell reporter expression maps for the 204 profiled CREs. Elements are organised by locus (horizontally). Map outlines indicate the element class as classified in the two-dimensional phenotypic space from **Fig. 4a**. Elements marked with # are found to be active (non-specific) in 2/3 replicates. Elements marked with * are found to be active and specific in at least one replicate with our thresholds. Each map is shown to the same colour scale (normalised mBC UMI from 0 and truncated to 5).




**Extended Data Figure 8. Additional loci with lineage specific distal CREs**

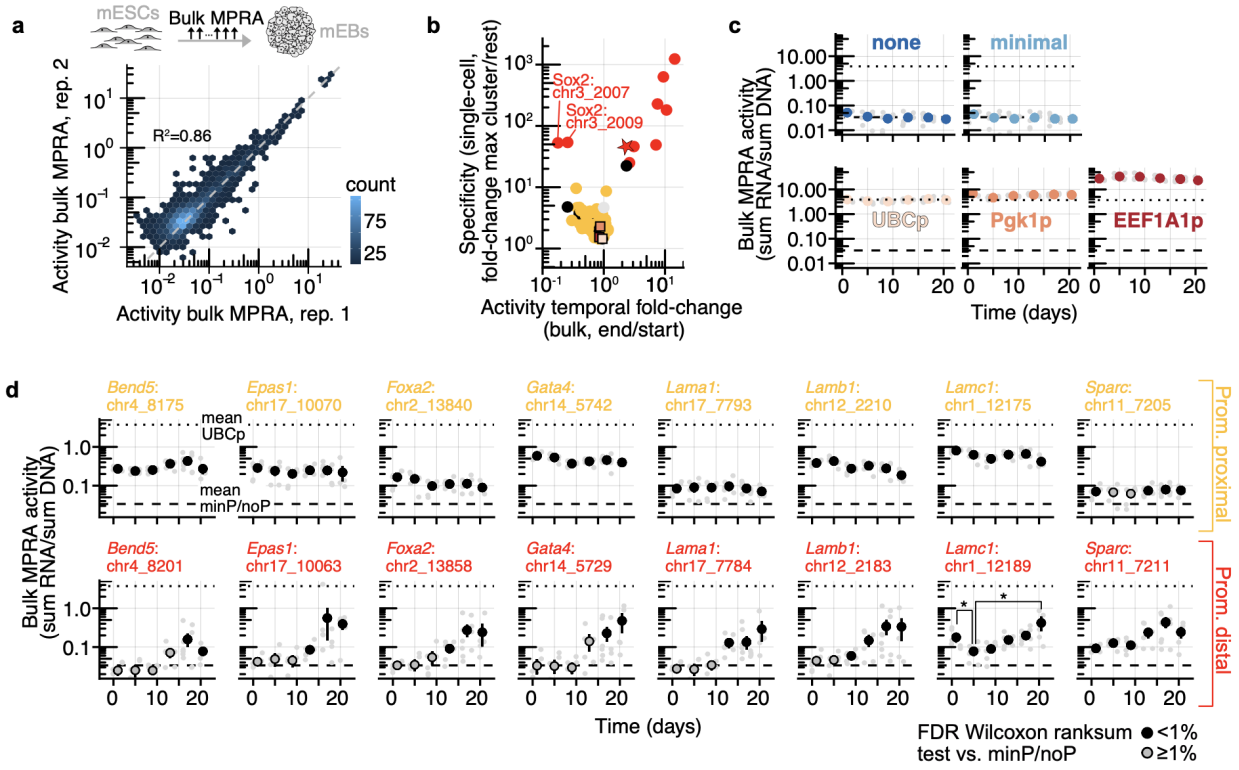
**a-d** Same as **Fig. 4b-e**, but for the additional five loci for which cell-type-specific CREs were identified. Each panel **a-d** is reproduced across rows for the different loci (top to bottom: *Lama1*, *Lamb1, Bend5, Sparc, Epas1*). The pink shaded element at the *Sparc* locus (chr11_7186) could not be cloned by PCR due to inability to identify specific primers in the vicinity.

**e** Measured maximum (across cell-types) activity (y=mean norm. mBC UMI/cell) vs. chromatin accessibility (top, x=number of ATAC reads in peak normalised to in TSS reads$\times 10^{-4}$ in the cognate cell-type) and fold-change in chromatin accessibility (bottom: x=fold-change accessibility in cognate cell type over other cell types). Points are coloured based on their functional categorization (same colours as **Fig. 4a**, grey: inactive, black: non-specific, distal; orange: non-specific, <1 kb TSS; red: cell-type specific).

**f** Fold-change in gene expression (y-axis, ratio normalised UMI in parietal endoderm to pluripotent) vs. CRE induction (x-axis, fold-change reporter levels, average normalised mBC UMI in parietal endoderm over pluripotent) for parietal-endoderm-specific distal CREs. Dashed line is 1:1. Geometric mean over three biological replicates is shown (errorbar: standard deviation of geometric mean).

**g** Assessing recapitulation of endogenous expression from identified autonomous CREs. Each point corresponds to one of 7 parietal endoderm genes with putatively associated identified active CREs and promoters shown in **Fig. 4** and panels **a-d** above (e.g., *Lamb1*: CRE chr12_2183, promoter chr12_2210; CRE associations to genes are putative). Endogenous gene induction (y-axis): fold-change in endogenous gene expression (average in normalised UMI counts) from pluripotent to parietal endoderm. CRE induction over promoter baseline (x-axis): CRE activity in parietal endoderm (reporter level, average normalised mBC UMI parietal endoderm) over mean activity of associated promoter in all cells (reporter level, average

normalised mBC UMI). Dashed line is 1:1. Shaded area corresponds to (CRE induction)<0.5×(gene expression induction). Geometric mean over three biological replicates is shown (errorbar: standard deviation of geometric mean).




**Extended Data Figure 9. Cell-type-specific CREs are temporally dynamic along mEB differentiation**

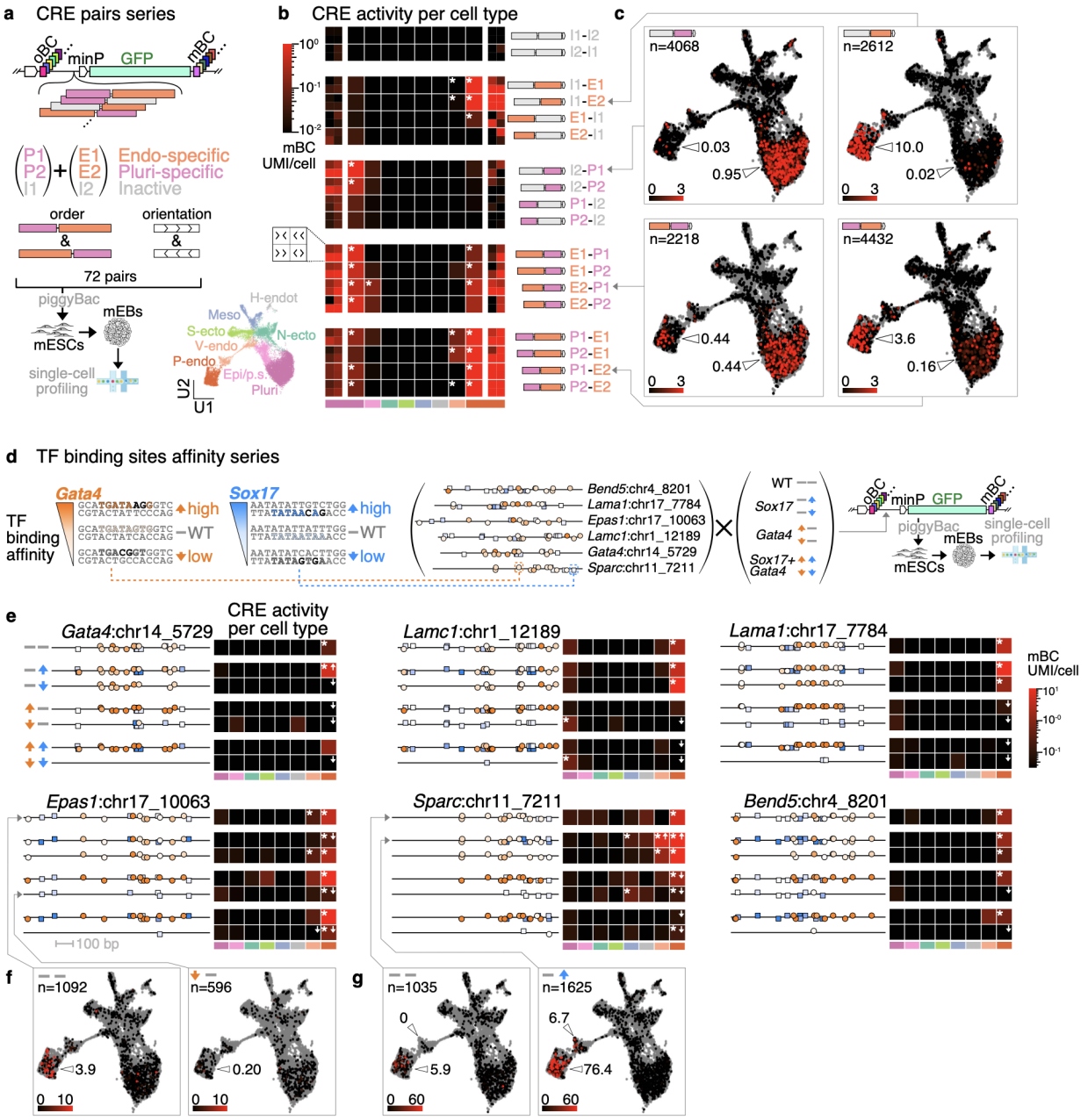
**a** Reproducibility of bulk MPRA measurement. Comparison of bulk MPRA activity (RNA/DNA ratio of summed 1% winsorised normalised UMI counts) for all CREs in two biological replicates (>10 measured barcodes in both replicates, including exogenous promoters) at all time points (n=2508 comparisons, $R^2$ from log-transformed activity).

**b** Differentiating EBs were sampled every two days at passage from all replicates, and bulk RNA/DNA MPRA libraries were generated. Fold-change in bulk MPRA activity across time course (mean activity day 20.5 over mean day 1) was compared to the observed specificity of elements as quantified from the scQer end-point quantification (**Fig. 4a**). Elements shown found to be active in either bulk or single-cell assays are shown and coloured according to class (red: cell-type specific, orange: non-specific, <1 kb from TSS, black: non-specific, distal ≥1 kb TSS).

The one grey point corresponds to the single element found to be active in bulk but not single-cell assay. Active exogenous promoters (UBCp, Pgk1p, EEF1A1p, panel **b**) are shown as squares. There is a correspondence between cell-type specificity and temporal change from the bulk assay. Bulk temporal fold-change is 5-10x smaller compared to single cell quantification likely due to bulk assay averaging activity from all cell-types.

**c** Activity traces of bulk MPRA time quantification for the exogenous promoters included as internal controls. Small grey points correspond to activity (RNA/DNA ratio of summed 1% winsorised normalised UMI counts) from different replicates/time points. Large points are the average of three biological replicates from two adjacent time points, with error bars corresponding to standard deviation of the mean (smaller than symbol size). Average of basal expression controls (no and minimal promoters) is shown as the dashed line, and the dotted line corresponds to the mean UBC promoter activity (reproduced in panel **d** for scale).

**d** Same as panel **c**, but for active cell-type-specific CREs (red) and promoters (orange) from the loci shown in **Fig. 4** and **Ext. Data Fig. 8**. Points are filled when significantly above basal expression controls (one-sided ranksum test, B-H corrected, FDR<1%). Promoters (orange) show largely constant expression over time. CREs (red) show substantial induction over the time course. Bifunctional CRE *Lamc1*:chr1_12189 displays initial decrease followed by and increase consistent with its activity in both undifferentiated and differentiated cells (one-sided Bonferoni corrected ranksum test between day 1 and day 5, p=0.026; and between day 5 and day 20.5, p=0.017).




**Extended Data Figure 10. Additional applications of scQers: pleiotropic activity of synthetic CRE pairs & profiling CREs with disrupted/optimised putative transcription factor binding sites**

**a** Library of pairs of CREs were constructed by joining two pluripotent (P1: *Sox2*:chr3_2007, P2: *Sox2*:chr3_2009) and one inactive sequence (I1: *Cdk5r1*:chr11_12590) with two parietal

endoderm (E1: *Epas1*:chr17_10063, E2: *Gata4*:chr14_5729) and another inactive sequence (I2: *Col5a1*:chr2_2586). Combinatorial libraries with all possible orientations and orders of the 6 components were cloned in scQers, mapped to barcodes with nanopore sequencing, integrated into mESCs and profiled for activity in mEBs. Inset shows UMAP of cells passing QC (n=20477), coloured by the mapped cell type (Pluri: pluripotent, Epi/p.s.: epiblast/primitive streak, N-ecto: neuroectoderm, H-endot: haemato-endothelial, Meso: mesoderm, S-ecto: surface ectoderm, V-endo: visceral endoderm, P-endo: parietal endoderm)

**b** Cell type-specific activity (median norm. mBC UMI per cell over three biological replicates) per cell type per construct. Rows indicate different pairs of CREs (in specified order), and columns correspond to different cell types (based on the colour scheme of the inset in **a**, indicated at bottom). Two outermost columns of the heatmap stratify each CRE pair by relative orientation of its components for their activity in pluripotent (leftmost column) and parietal endoderm (rightmost column), central columns correspond to median over all four relative orientations for a given ordered pair. Stars (*) mark CRE pairs and cell types with activity significantly above negative controls (minP, noP, I1-I2, I2-I1) ($p<0.01$ from one-sided bootstrap resampling of cells with detected constructs with B-H correction).

**c** Example single-cell maps of CRE-pair activity. Number of cells with detected CRE-pairs marked, with norm. mBC UMI/cell shown on a black (low) to red (high) colour scale (grey: CRE of interest not detected). Quantified expression in parietal endoderm and pluripotent cells (median over biological replicates of mean norm. mBC UMI/cell) are indicated.

**d** CRE variants optimising and disrupting the binding affinity of all putative *Gata4* and *Sox17* transcription factor binding sites in combination (variants: WT, *Sox17*-high, *Sox17*-low, *Gata4*-high, *Gata4*-low, *Gata4*-*Sox17*-high, *Gata4*-*Sox17*-low) identified within 6 parietal endoderm-

specific CREs were designed based on UniProbe data[104] (example of approach illustrated in **Supp. Fig. 7d**). Schematics of CREs with mapped TF binding sites are shown (*Gata4*: orange, *Sox17*: blue; hue indicative of binding affinity). *Gata4* and *Sox17* putative binding sites within the *Sparc*:chr11_7211 element and their perturbed instances (affinity optimization or disruption via two mutations per site) are displayed as examples. Variant CREs were cloned as a scQer library, and their activity profiled after integration to mESCs and embryoid body differentiation (same experiment as panels **a-c**). Insert shows UMAP of cells coloured by cell type assignment.

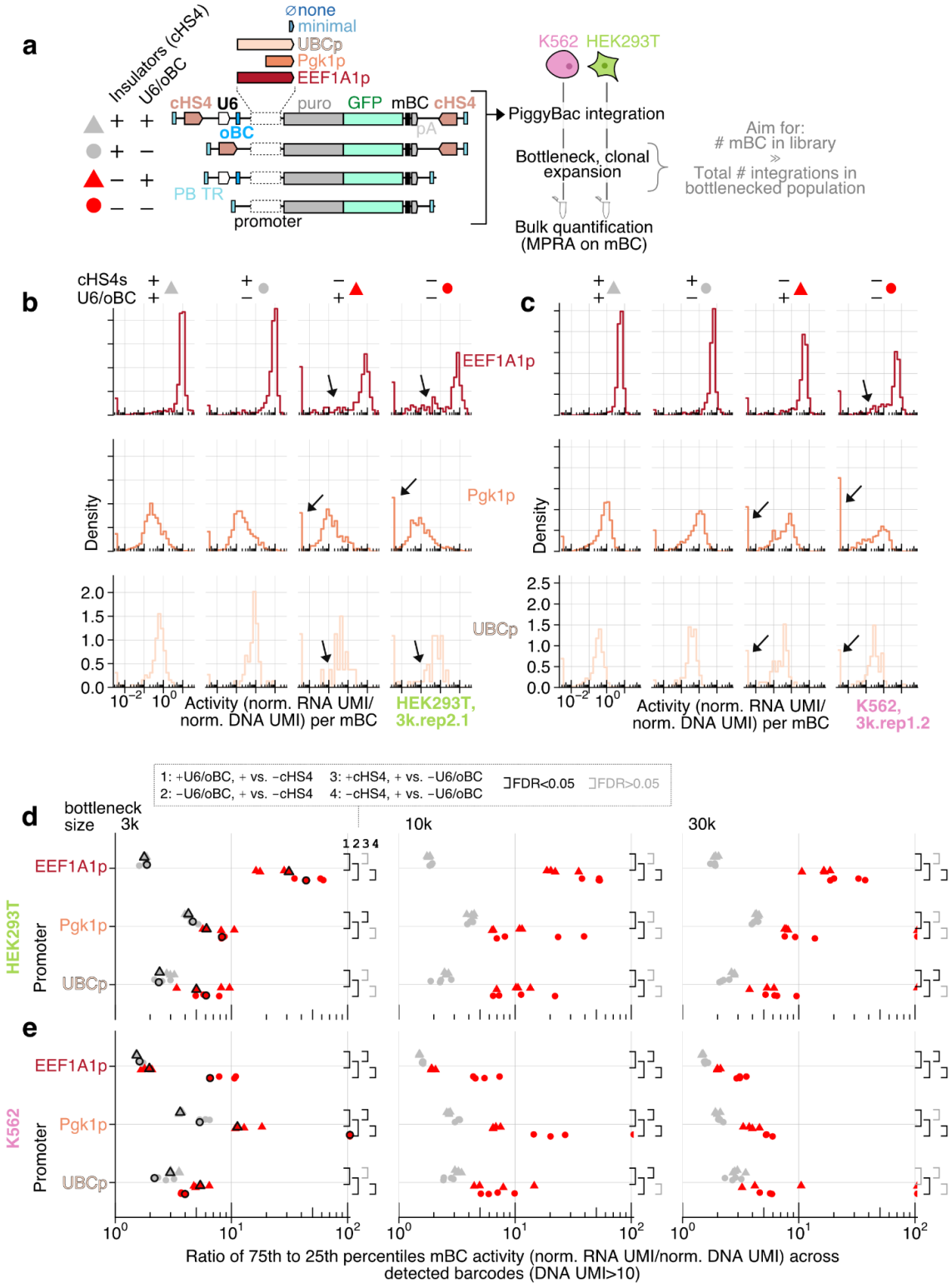
**e** Cell type-specific activity (median norm. mBC UMI/cell over biological replicates) per cell type per CRE. Panels show heatmaps of activity of different CRE series, with CRE TF binding site maps shown (left, rows), and columns correspond to different cell types (based on colour scheme of the inset in **a**, indicated at bottom). Stars (*) indicate significantly higher expression than negative controls ($p<0.01$ from one-sided bootstrap resampling of cells with detected constructs with B-H correction). White arrows mark regions with significantly different expressions than the respective WT CRE variant (one-sided bootstrap resampling of cells with detected constructs with B-H correction; up arrow: increased expression $p<0.01$; down arrow: decreased expression $p<0.01$ if also WT expression > 0.1 mBC UMI/cell).

**f-g** Example single-cell maps of CRE activity (for each respective panel, left: unperturbed CRE, right: perturbed CRE). Number of cells with detected CRE reporters indicated, with norm. mBC UMI/cell shown on a black (low) to red (high) colour scale (grey: CRE of interest not detected). **f** Example of loss-of-activity from disruption of putative *Gata4* sites within CRE *Epas1*:chr17_10063, with mean activity in parietal endoderm indicated. **g** Dramatic instance of gain-of-function, with >10-fold greater expression in parietal endoderm from putative *Sox17* TF

binding site optimization in CRE *Sparc*:chr11_7211, also associated with ectopic expression in the related visceral endoderm.

See also **Supp. Fig. 3** and **7.**

***Supplementary figures:***



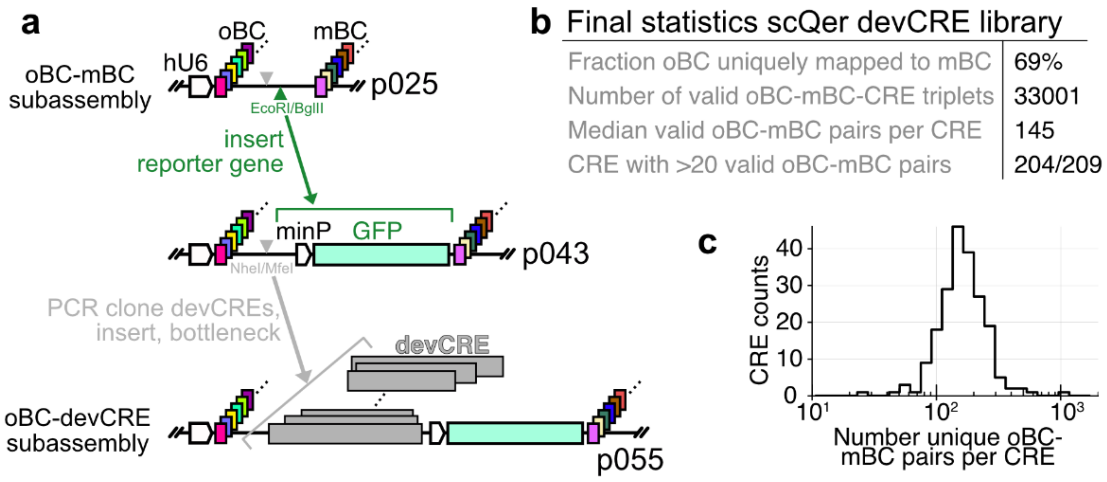
**Supplementary Figure 1. Comparison of cHS4 and Pol III U6/oBC cassette for insulating effects.** (legend on next page)

**a** Schematic of experiment to assess importance of cHS4 and U6/oBC cassettes in mitigating positional effects on randomly integrated promoters. 20 libraries consisting in the 5 ectopic promoters (no promoters, minP, UBCp, Pgk1p, EEF1A1p) in the 4 contexts (+cHS4+U6/oBC, +cHS4-U6/oBC, -cHS4+U6/oBC, -cHS4-U6/oBC), with each library barcoded and pooled prior to integration (piggyBac) in K562 and HEK293T cells, consisting a final pool of 134.1k barcodes uniquely mapped to promoter & reporter architectures. Cells were then bottlenecked (3k, 10k, and 30k estimated starting populations, each in biological duplicates), and MPRA performed on expanded clonal populations (libraries in technical duplicates). In the regime where the number of integrated barcodes is small compared to the total barcode complexity, most barcodes correspond to unique integration events at a specific genomic position. Variability in the per barcode activity across the different reporter architectures provides a measure of the effect of positional variability.

**b** & **c** Example of mBC expression (norm. RNA UMI over norm. DNA UMI count) distributions for active promoters (rows, top: EEF1A1p, middle: Pgk1p, bottom: UBMp) across different reporter architectures (from left to right: +cHS4+U6/oBC, +cHS4-U6/oBC, -cHS4+U6/oBC, -cHS4-U6/oBC). Arrows highlight wider distributions in per mBC expression in promoters (without insulators), reflecting more positional variability. **b** Replicate HEK293T 3k.2.1, **c** Replicate K562 3k.1.2.

**d** & **e** Global quantification of mBC expression spread (ratio of 75th to 25th percentiles) across different biological replicates split by promoters and reporter architectures (**c** HEK293T, **d** K562; grey triangles +cHS4+U6/oBC, grey circles +cHS4-U6/oBC, red triangles -

cHS4+U6/oBC, red circles -cHS4-U6/oBC). Symbols with black outlines correspond to distributions shown in panels in **b** and **c**. For each cell line, bottlenecked population, and promoters, four comparisons were performed (1: +U6/oBC, + vs. -cHS4; 2: -U6/oBC, + vs. -cHS4; 3: +cHS4, + vs. -U6/oBC; 4: -cHS4, + vs. -U6/oBC), with results of statistical test indicated by colour of square brackets (B-H corrected two-sided Wilcoxon test, grey: $p \geq 0.05$, black: $p < 0.05$). Reporters without cHS4 insulates (red) display substantially more variability in nearly all contexts (35/36), and the U6/oBC also reduces variability though only in some promoter/cell line contexts (H293T: 4/18, K562: 11/18)**.**

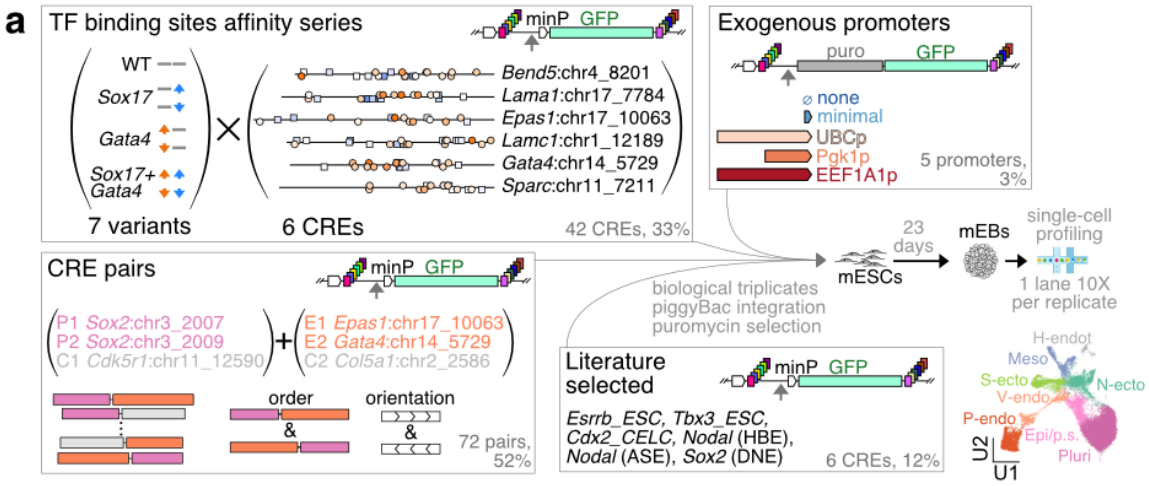


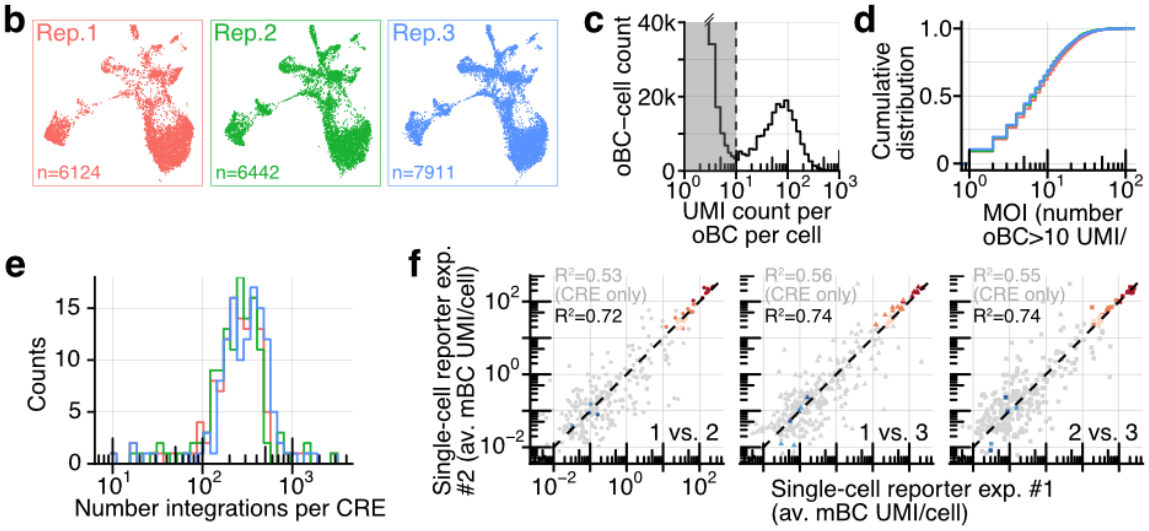

**Supplementary Figure 2. scQer library construction and oBC-CRE-mBC subassemblies**

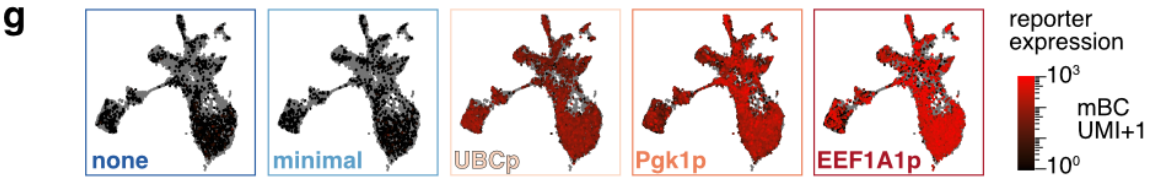
**a** Schematic of procedure to construct doubly barcoded dual RNA reporters. First, a high-complexity (~1 M) library of doubly barcoded (oBC and mBC, separated by multiple cloning site dock) piggyBac transposons is constructed. At this step, oBC and mBC matches are determined (PCR-based library construction). The minimal promoter with GFP cassette is then inserted, and complexity maintained as much as possible. >200 CREs were PCR-cloned, pooled at 1:1 ratios by mass, and inserted in the doubly barcoded minP-GFP backbone by isothermal assembly. The resulting library was bottlenecked to ~50k clones. CRE and oBC matches were then determined on the bottlenecked library (tagmentation with semi-specific PCR). In combination with the initial oBC-mBC pairs, this completes the determination of oBC-CRE-mBC triplets needed to deconvolute single-cell data for reporter activity. Plasmid names (p025, p043, p055) are indicated.

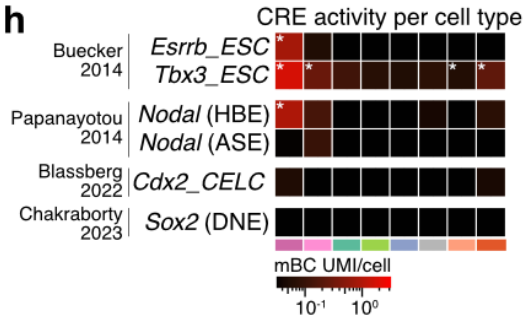
**b** Compilation of statistics from scQers library used to screen putative CREs in mEBs.

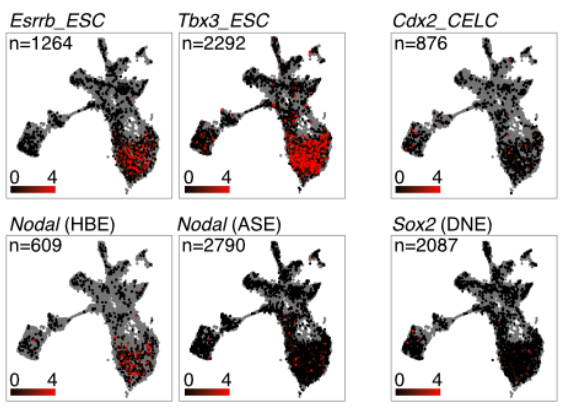
**c** Distribution of number of unique oBC-mBC pairs per CRE following the subassembly and quality filters, displaying largely uniform representation of the >200 putative regulatory elements tested (experiment **Fig. 3b**).


a
TF binding sites affinity series
minP GFP
WT
Sox17
Gata4
Sox17+ Gata4
7 variants
Bend5:chr4_8201
Lama1:chr17_7784
Epas1:chr17_10063
Lamc1:chr1_12189
Gata4:chr14_5729
Sparc:chr11_7211
6 CREs
42 CREs, 33%
Exogenous promoters
puro GFP
none
minimal
UBCp
Pgk1p
EEF1A1p
5 promoters, 3%
CRE pairs
minP GFP
P1 Sox2:chr3_2007
P2 Sox2:chr3_2009
C1 Cdk5r1:chr11_12590
E1 Epas1:chr17_10063
E2 Gata4:chr14_5729
C2 Col5a1:chr2_2586
order
orientation
72 pairs, 52%
23 days
mEBs
mESCs
single-cell profiling
1 lane 10X per replicate
biological triplicates
piggyBac integration
puromycin selection
Literature selected
minP GFP
Esrrb_ESC, Tbx3_ESC, Cdx2_CELC, Nodal (HBE), Nodal (ASE), Sox2 (DNE)
6 CREs, 12%
H-endot
Meso
S-ecto
N-ecto
V-endo
P-endo
Epi/p.s.
Pluri
U2
U1


b
Rep.1
n=6124
Rep.2
n=6442
Rep.3
n=7911
c
oBC−cell count
40k
20k
0
UMI count per oBC per cell
d
Cumulative distribution
1.0
0.5
0
MOI (number oBC>10 UMI/
e
Counts
15
10
5
0
Number integrations per CRE
f
Single-cell reporter exp. #2 (av. mBC UMI/cell)
R²=0.53 (CRE only)
R²=0.72
1 vs. 2
R²=0.56 (CRE only)
R²=0.74
1 vs. 3
R²=0.55 (CRE only)
R²=0.74
2 vs. 3
Single-cell reporter exp. #1 (av. mBC UMI/cell)


g
none
minimal
UBCp
Pgk1p
EEF1A1p
reporter expression
mBC UMI+1


h
CRE activity per cell type
Buecker 2014
Esrrb_ESC
Tbx3_ESC
Papanayotou 2014
Nodal (HBE)
Nodal (ASE)
Blassberg 2022
Cdx2_CELC
Chakraborty 2023
Sox2 (DNE)
mBC UMI/cell


i
Esrrb_ESC
n=1264
Tbx3_ESC
n=2292
Cdx2_CELC
n=876
Nodal (HBE)
n=609
Nodal (ASE)
n=2790
Sox2 (DNE)
n=2087

**Supplementary Figure 3. Quality control metrics for applications of scQer experiment** (legend next page)

**a** Schematic of scQer experiment testing different applications (pairs of CREs, allelic series disrupting/optimising putative TF binding sites, additional CREs selected from the literature). Each library (boxed) was cloned, assembled to barcodes separately, pooled (in proportions shown: 3% ectopic promoter puro-GFP as internal control and to select for high MOI cells, 33% TF binding site allelic series, 52% CRE pairs, 12% literature selected CREs), integrated in mESC, differentiated to mEBs and single-cell profiled (day 23) as described before. Integration to profiling was performed in biological triplicate. Inset shows UMAP of cells passing quality control coloured by assigned cell types.

**b** UMAP of cells passing quality control files split by biological replicates.

**c** Distribution of oBC UMI counts per cell per barcode, showing highly bimodal nature, enabling identification of which reporter was present in which cells (>10 oBC UMI/cell, dashed line).

**d** Cumulative distribution of number of detected reporters per cell (median MOI=7).

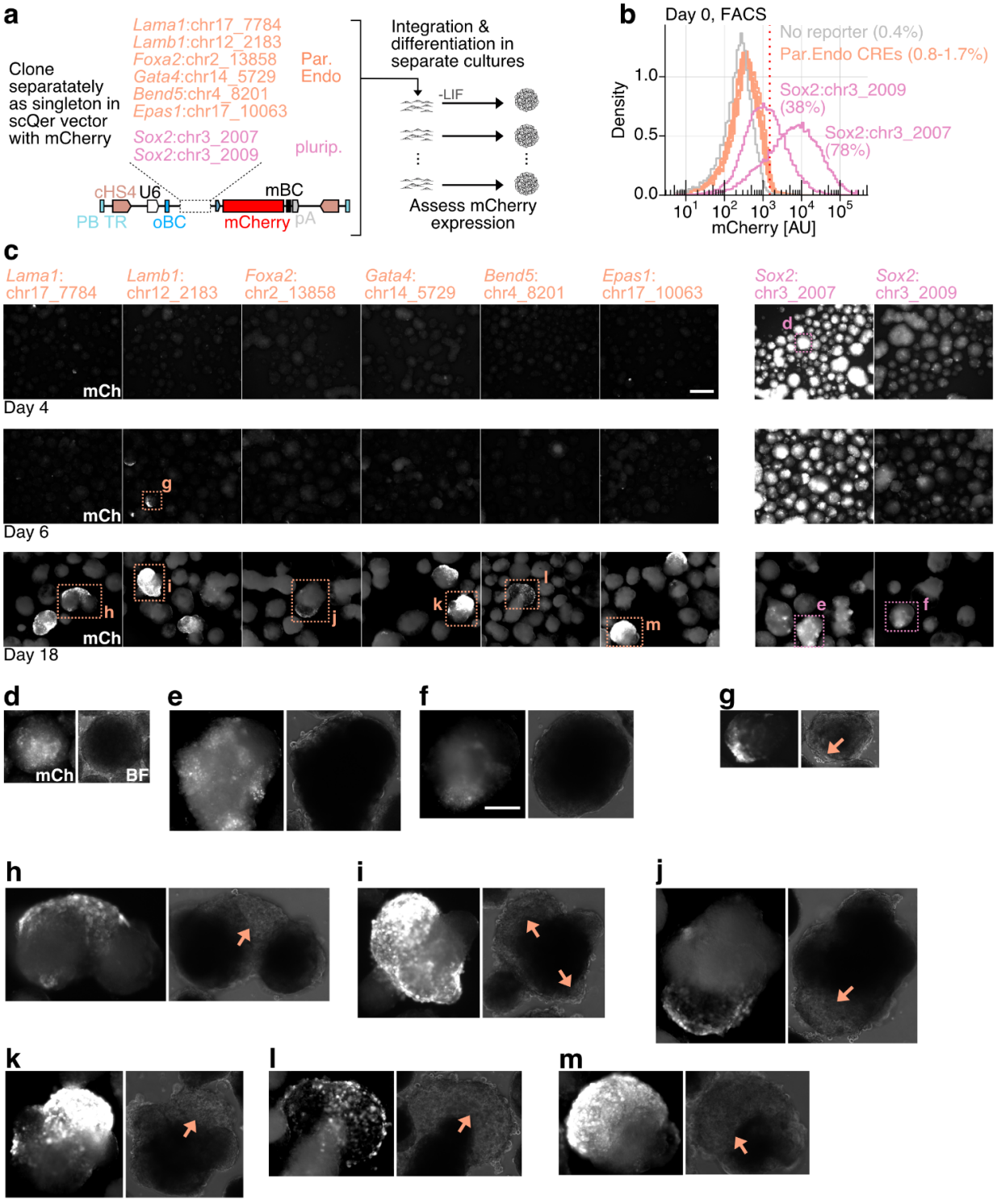
**e** Number of detection events per CRE (or pairs of CRE) across different replicates.

**f** Expression per CRE per cell type (mean norm. mBC UMI/cell; grey: CREs, coloured: exogenous promoters) for CREs with >0 activity stratified by biological replicate. Each point corresponds to a CRE in a cell-type. $R^2$ on log-transformed values, including exogenous promoters (black) or not (grey), are indicated (0.72 to 0.74 including ectopic promoters, 0.53 to 0.56 considering only CREs). Lower reproducibility compared to experiment from **Ext. Data Fig. 5b** likely due to lower representations per CRE. Replicates were pooled for downstream quantification.

**g** Single-cell reporter expression (average normalised mBC UMI per cell) for the five exogenous promoters used as internal controls. Colour scale is logarithmic (with a pseudocount of 1).

**h** Heatmap quantification of expression per cell type for literature-selected CREs[64–67] (rows: CREs, columns: cell types, following colour scheme in panel **a** inset; colour-bar at bottom). Significant expression over negative controls (noP, minP; B-H corrected bootstrap resampling $p<0.01$) are indicated by *. The three CREs with significant activity (the *Esrrb* and *Tbx4* CREs from Buecker et al 2014, and the Nodal HBE from Papanayotou et al 2014) were most strongly expressed in pluripotent cells, as expected based on their original reported activity. Two of the 3 elements found to be inactive (a *Sox2* neural CRE, and a *Cdx2* CRE expressed in caudal epiblast-like cells) were likely not expressed as a result of low representation of cognate cell types in our system. Finally, the *Nodal* ASE was lowly expressed specifically in the expected cell type (epiblast cells), but fell below our stringent significance threshold.

**i** Single-cell map of activity for literature selected CRE (same as panel **h**). Number of detection per CRE indicated. Reporter expression shown from low (black) to high (red) (grey: no detection of CRE of interest).



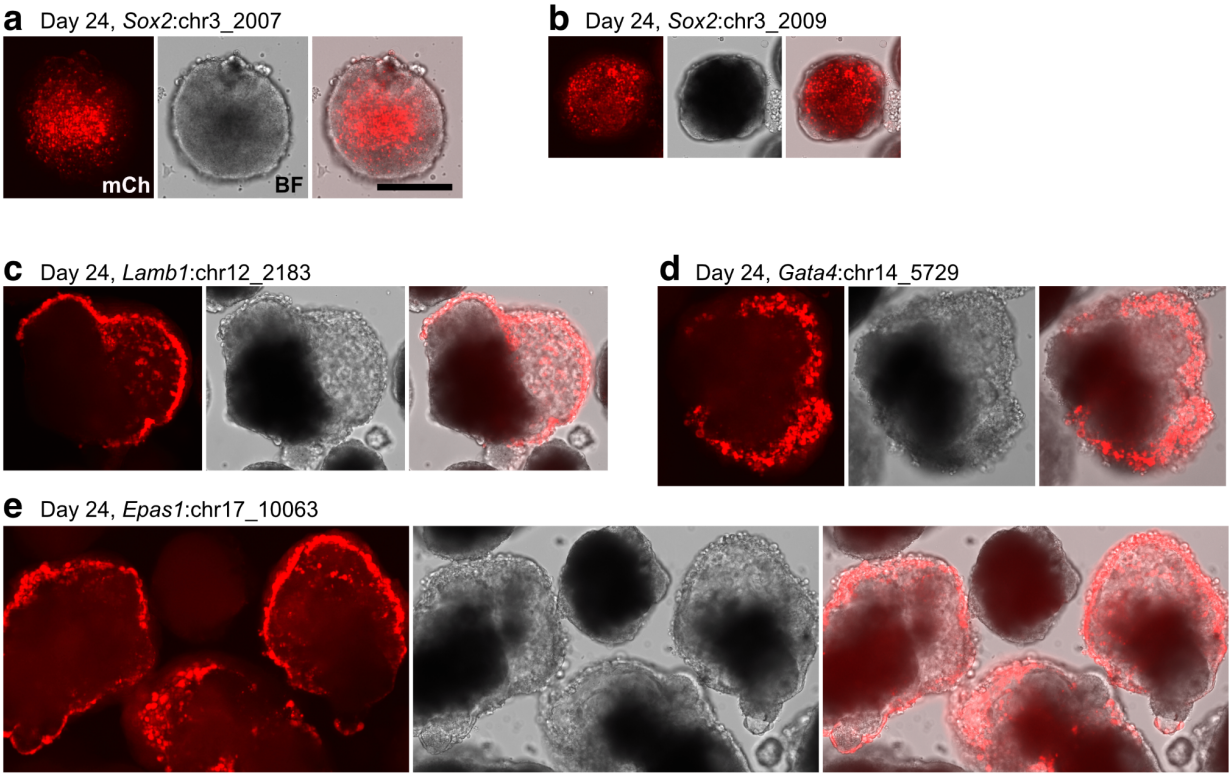

**Supplementary Figure 4. Singleton validation experiment of cell type-specific CREs**

**a** Schematic of singleton validation experiments. The 8 most highly cell type-specific CREs as assessed by our single-cell data (**Fig. 4**) were cloned individually in a scQer backbone with mCherry as the reporter gene. Each construct was transfected separately (co-transfection with 10% promoter puro-GFP series), integrated via piggyBac in mESCs and differentiated in mEBs. Epifluorescence images were acquired on alternate days, and with structured illumination on day 24 (**Supp. Fig. 5**) to assess spatial patterns in expression. One biological replicate per singleton construct was performed (n=8 separate cultures and differentiated samples).

**b** Distribution of mCherry intensity per cell (FACS data) from single-cell suspension (day of embryoid body induction) of scQer-containing singleton lines. Cells harbouring parietal endoderm-specific elements exhibited limited reporter expression (0.8 to 1.7% above background, defined as the top 0.4% of the no-reporter negative control shown in grey). In contrast, pluripotent-specific elements displayed robust mCherry expression in a substantial proportion of cells (mCherry+: 38% for *Sox2*:chr3_2009, 78% for *Sox2*:chr3_2007).

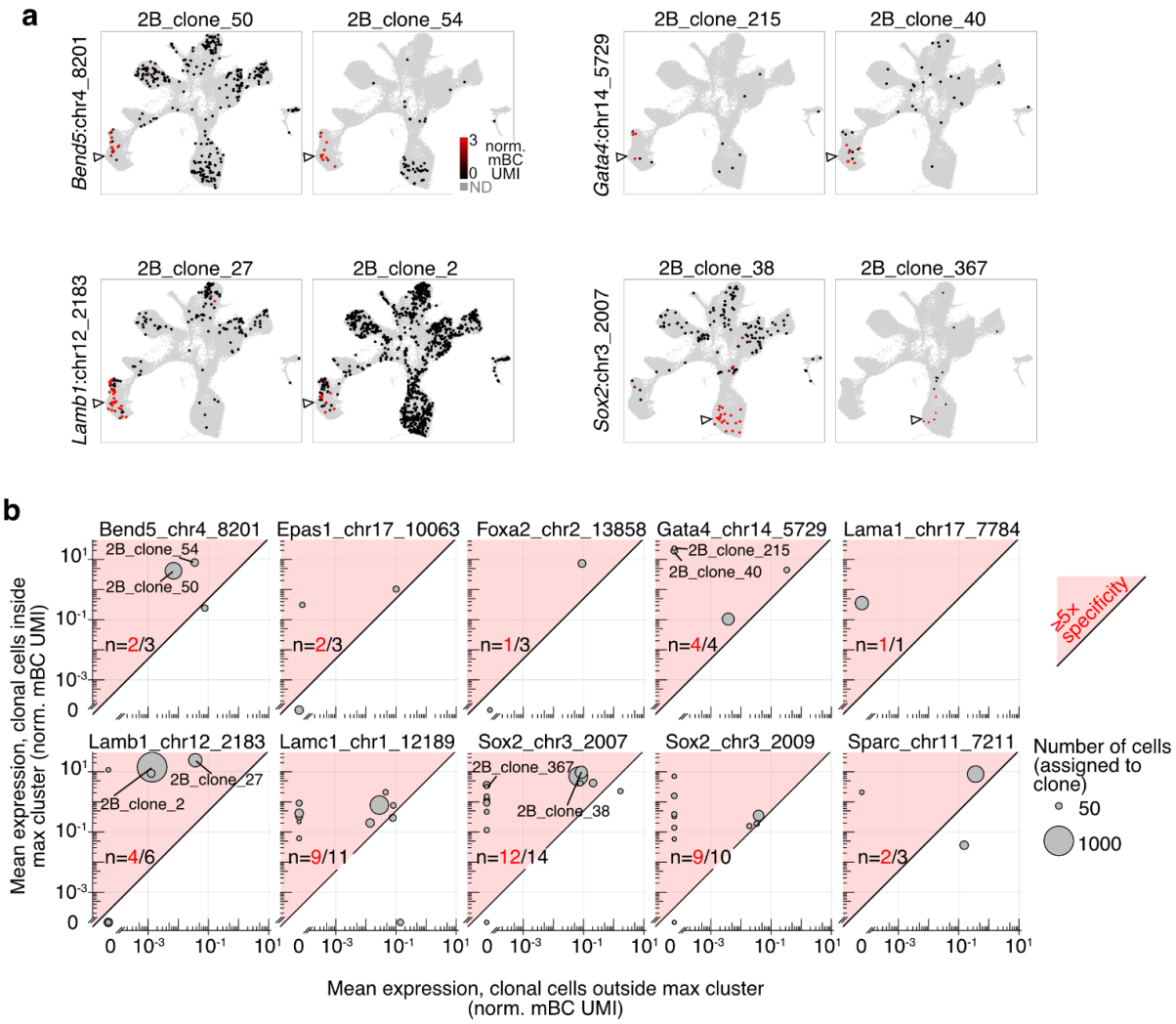
**c** Representative fields of view (mCherry channel, Chroma 594 filter Zeiss Axio observer) from individual examples of embryoid body differentiation (columns: different CREs, rows: different times; for each singleton construct and differentiated sample, many EBs showed characteristic expression patterns displayed). Scale bar: 0.5 mm (same scale for all images). All images on the same day have the same exposure (exposure, day 0 & 4: 5 s, day 18: 2 s) and contrast to allow direct comparison. Over differentiation time courses, mCherry signal emerged in a subset of EBs for all parietal endoderm CREs (left six columns). In contrast, the signal decreased in intensity for pluripotent-specific elements (right two columns). Examples of embryoid bodies with localised expression are indicated and zoomed in panels **d**-**m**.

**d-m** Zoomed-in regions from panels in **c**, highlighting spatial pattern of expression of the reporters (for each panel left: mCherry, right: brightfield). Contrast adjusted differently in each panel. For parietal endoderm-specific elements, expression coincided with cells on the surface with rough morphology (arrows), in accordance with described endodermal cells in embryoid bodies[49]. In contrast, *Sox2* control elements displayed internal expression from largely smooth embryoid bodies. Scale bar: 200 um.




**Supplementary Figure 5. Structured illumination images of mEBs with singleton scQer reporters**

**a**-**e** Images (Keyence BZ-X810) from day 24 singleton mEBs harbouring different CREs (same differentiation experiment as **Supp. Fig. 4**, one biological replicate per singleton construct was performed). Left: mCherry (optical sectioning mode: structured illumination, custom pinhole, slit size 2, slit pitch 6, filter cube: chroma Cy3/R 49004), middle: brightfield, right: merge. Pluripotent elements (**a**, **b**) show distinct internal expression compared to parietal endoderm elements (**c**-**e**) which display signal exclusively in surface/rough cells. Scale bar: 250 um (same scale on all images). Images were acquired with the automatic exposure times on mCherry channel (**a**: 0.33 s, **b**: 1.5 s, **c**: 0.33 s, **d**: 0.33 s, **e**: 0.2 s), brightfield: 4 ms.

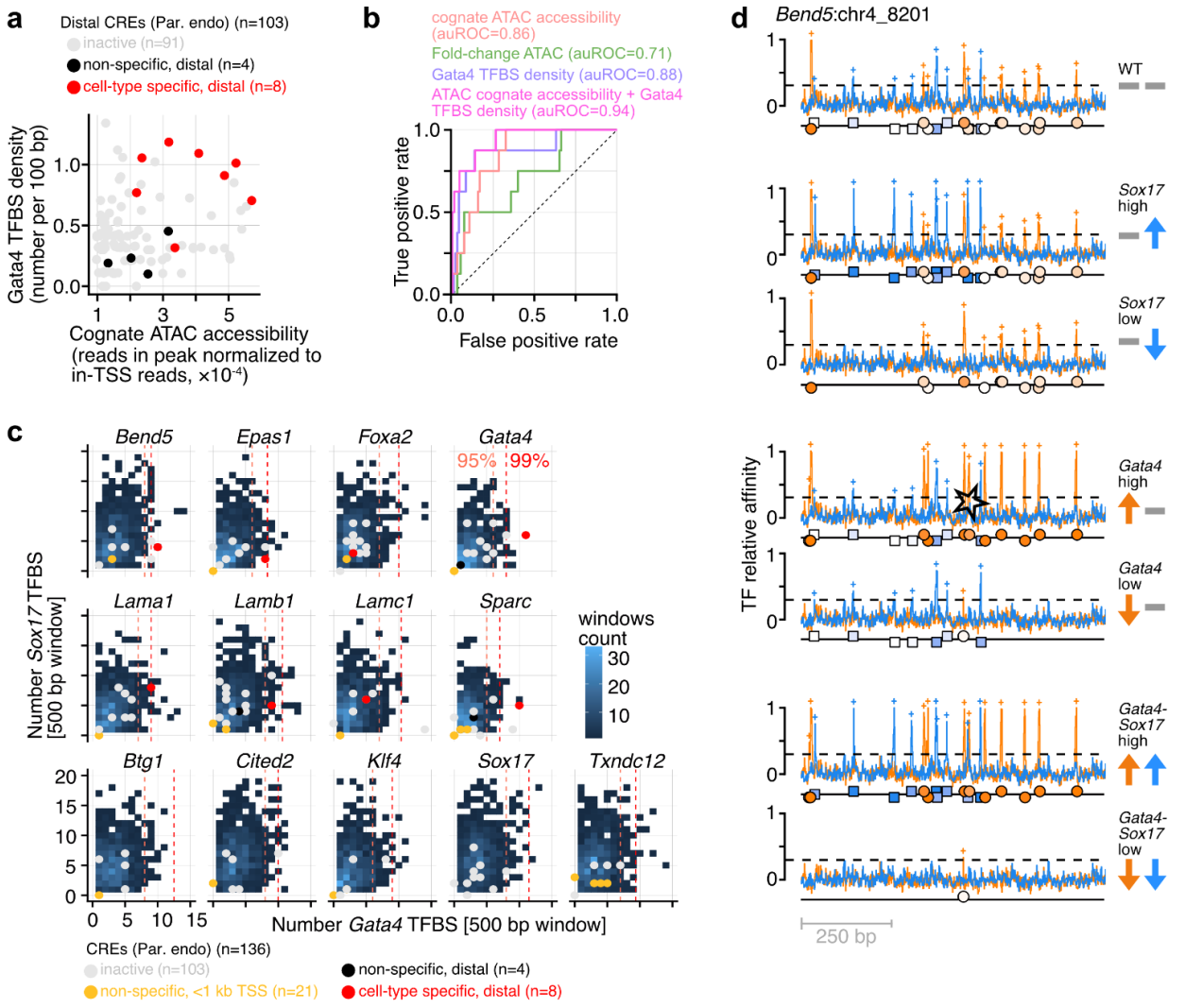



**Supplementary Figure 6. Cell-type-specific CRE expression across clones to assess positional integration effects**

**a** Examples of single-cell map of CRE activity for cells assigned to high-confidence clones for four CREs (two representative clones per element shown, marked in panel **b**). Carets indicate the cluster in which expression is expected based on quantification over all cells. Grey points in the background are all other cells not assigned to the clone.

**b** Systematic quantification of specificity (activity in expected maximum-expression cluster vs. rest of cells, **Fig. 4a**) across all well-represented clones (at least 5 cells in expected maximum

expression cluster(s) and at least 5 cells in other clusters) for the 10 CREs identified as active and specific. Each clone is represented by a circle, whose area corresponds to the number of cells assigned to it. Clones shown in panel **a** are indicated. Red shading delineates the region where specificity is in excess of 5-fold. Fractions of clones meeting this criterion for distinct CRE are indicated on each panel. 9/10 CREs have ≥⅔ of their clones with >5-fold specificity.




**Supplementary Figure 7. CRE features correlated to cell-type-specific activity**

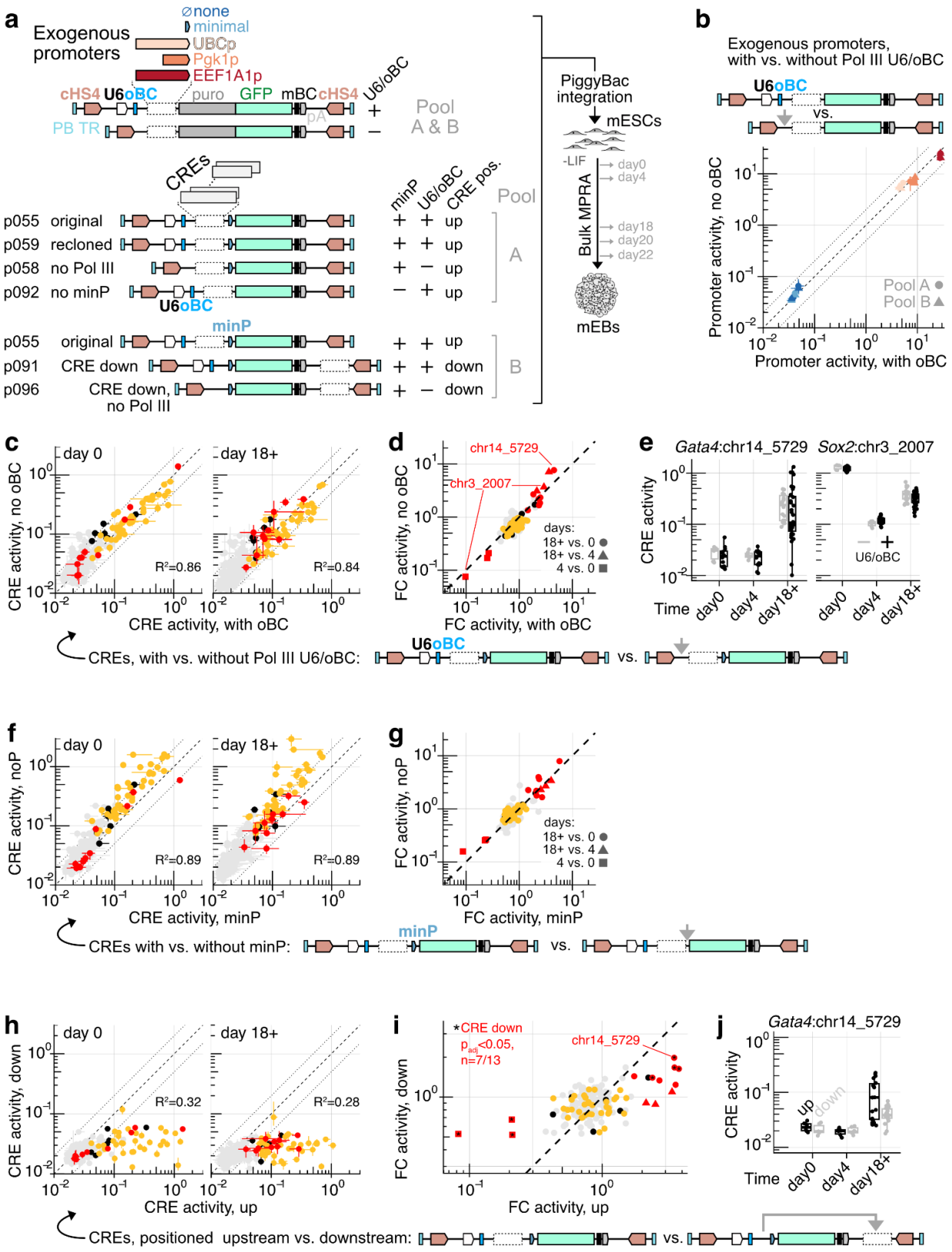
**a** Plot of two features highly enriched for autonomous cell-type-specific CREs: cognate (cell-type corresponding to differential expression of putatively associated gene) ATAC accessibility (x-axis): average in peak reads normalised by reads in TSS (in each cell). y-axis: Density of Gata4 transcription factor binding sites per 100 bp (TFBS with affinity relative to the maximum affinity 8-mer >0.4). Red points mark cell-type-specific CREs. Distal (>1 kb TSS) CREs selected from parietal endoderm loci are shown (n=103).

**b** Receiver operating characteristic (ROC) curves for the classification task (specific vs. non-specific/inactive) from different features. Density of Gata4 TFBS, cognate ATAC accessibility, and fold-change in ATAC signal have good predictive value to discriminate functional elements (auROC >0.7). A logistic regression classifier including only cognate ATAC accessibility and Gata4 TFBS improves performance to auROC=0.94 (precision=0.6 at recall=0.75). Categories are unbalanced (active=8, inactive=95).

**c** Sequence analysis of all 500 bp windows (sliding step 250 bp, excluding any window overlapping with CREs with buffer flank position 500 bp on either sides) for the 13 endoderm-specific developmental loci (±100 kb from TSS of indicated gene). For each genomic sequence window, the number of transcription factor binding sites to *Gata4* and *Sox17* (affinity relative to the maximum affinity 8-mer >0.3) is recorded. Panels show the two-dimensional distribution of binding sites numbers across all windows, stratified by loci (parietal endoderm elements). The number of binding sites is also determined for tested CREs (coloured points; red: cell-type specific, orange: non-specific, <1 kb from TSS, black: non-specific, distal ≥1 kb TSS; grey: inactive) and overlaid on the distributions for comparisons. Cell-type-specific CREs (red points) have an elevated number of Gata4 binding sites compared to other inactive CREs as well as neighboring regions in the locus. Dashed lines mark the 95$^{th}$ and 99th percentile in Gata4 binding site numbers at each locus. 6/8 autonomously active CREs in local top 5%, 4/8 in local top 1% of number of Gata4 binding sites.

**d** Example of bioinformatic approach to identify putative TF binding sites (shown for CRE *Bend5*:ch4_8201). DNA sequences were broken up in overlapping 8-mer (stride length = 1), and the relative affinity of each 8-mer for *Gata4* and *Sox17* obtained from processed UniProbe data. The trace of 8-mer relative affinity to the different TFs is plotted (*Gata4*: orange, *Sox17*: blue).

Putative binding sites are identified as local maxima with relative affinity > 0.3 (dashed line). If multiple local maxima within 3 bp are identified, the maximum affinity position is retained as the putative binding site. Putative sites are marked by crosses (+) above the affinity traces, and are marked below the traces as schematic circles (*Gata4*) and squares (*Sox17*), coloured by the site affinity. Orientation of the putative binding sites is determined relative to a short core PWM (*Gata4*: GATAA, *Sox17*: ACAAT; symbol above line: forward strand, symbol below: reverse strand, symbol on line: no orientation preference). Optimising (disruption) variants are generated by replacing these putative sites by the highest (lowest) affinity 8-mer that is a Hamming distance of 2 away from the putative sites. Affinity traces for the resulting six variants of *Bend5*:ch4_8201 generated and tested experimentally are shown (*Sox17*-high, *Sox17*-low, *Gata4*-high, *Gata4*-low, *Gata4*-*Sox17*-high, *Gata4*-*Sox17*-low). The black star in the *Gata4*-high trace highlights a putative low affinity *Sox17* site disrupted by optimization of the nearby putative *Gata4* site, illustrating possible impacts on other factors' binding of our procedure.


a
Exogenous promoters
none
minimal
UBCp
Pgk1p
EEF1A1p
cHS4 U6oBC puro GFP mBC cHS4
PB TR
pA
U6/oBC
Pool A & B
CREs
minP
U6/oBC
CRE pos.
Pool
p055 original + + up
p059 recloned + + up
p058 no Pol III + − up
p092 no minP − + up
A
U6oBC
minP
p055 original + + up
p091 CRE down + + down
p096 CRE down, no Pol III + − down
B
PiggyBac integration
mESCs
-LIF
Bulk MPRA
day0
day4
day18
day20
day22
mEBs
b
Exogenous promoters, with vs. without Pol III U6/oBC
U6oBC
vs.
Promoter activity, no oBC
Promoter activity, with oBC
Pool A
Pool B
c
day 0
day 18+
CRE activity, no oBC
CRE activity, with oBC
R²=0.86
R²=0.84
CREs, with vs. without Pol III U6/oBC:
U6oBC
vs.
d
chr14_5729
chr3_2007
FC activity, no oBC
FC activity, with oBC
days:
18+ vs. 0
18+ vs. 4
4 vs. 0
e
Gata4:chr14_5729
Sox2:chr3_2007
CRE activity
U6/oBC
Time
day0
day4
day18+
f
day 0
day 18+
CRE activity, noP
CRE activity, minP
R²=0.89
R²=0.89
CREs with vs. without minP:
minP
vs.
g
FC activity, noP
FC activity, minP
days:
18+ vs. 0
18+ vs. 4
4 vs. 0
h
day 0
day 18+
CRE activity, down
CRE activity, up
R²=0.32
R²=0.28
CREs, positioned upstream vs. downstream:
vs.
i
*CRE down
padj<0.05,
n=7/13
chr14_5729
FC activity, down
FC activity, up
j
Gata4:chr14_5729
CRE activity
up
down
Time
day0
day4
day18+

**Supplementary Figure 8. Assessing the impact of different reporter architectures** (legend on next page)

**a** Schematics of constructs included in bulk MPRA assessment of the impact of components of the scQer cassette towards measured expression. Each depicted library was cloned and assembled to barcode dictionaries separately prior to pooling (two pools: A & B) for piggyBac integration in mESCs, differentiation to mEBs and bulk MPRA (three biological replicates, MPRA libraries for each replicate and time point prepared in technical duplicates). Both pools included exogenous promoters with and without U6/oBC (top) together with the same CRE library as **Fig. 3** inserted in reporters with different architectures. Pool A included constructs with and without U6/oBC, and without minP. Pool B included constructs with CREs positioned downstream of the reporter.

**b** Exogenous promoter expression (bulk MPRA activity quantification: sum 1% winsorised normalised RNA UMI over sum 1% winsorised normalised DNA UMI from barcodes associated with promoters over barcodes with >10 DNA UMI) quantified from pool A and B experiments for the with (x-axis) vs. without (y-axis) U6/oBC, highlighting the overall limited influence of the Pol III cassette on the promoter activity. Errorbars correspond to interquartile range over replicates (three biological replicates with two technical replicates each) for a given promoter/time point. Early (days 0, 4) and late time points (days 18, 20, 22) were respectively aggregated for quantification.

**c** Similar to **b**, but for CREs (left: day 0, right: days 18+), coloured according to the activity characterization categories of **Fig. 4** (grey: inactive, black: non-specific, distal; orange: non-specific, <1 kb TSS; red: cell type-specific). Day 18+ panel aggregates data from days 18, 20, and 22.

**d** Temporal fold-change in activity of CREs (circles: day 0 vs. day 18+ [all elements except non-monotonic], squares: day 0 vs. 4 & triangle: day 4 vs. day 18+ [elements with non-monotonic temporal dynamics, i.e., Sox2:chr3_2007, Sox2:chr3_2009, and Lamc1:chr1_12189 have intermediate fold-change displayed]) within reporter cassette with (x-axis) vs. without (y-axis) U6/oBC, highlighting the general lack of influence of Pol III transcription towards measured reporter activity. Fold-changes of elements shown in panel **e** are indicated.

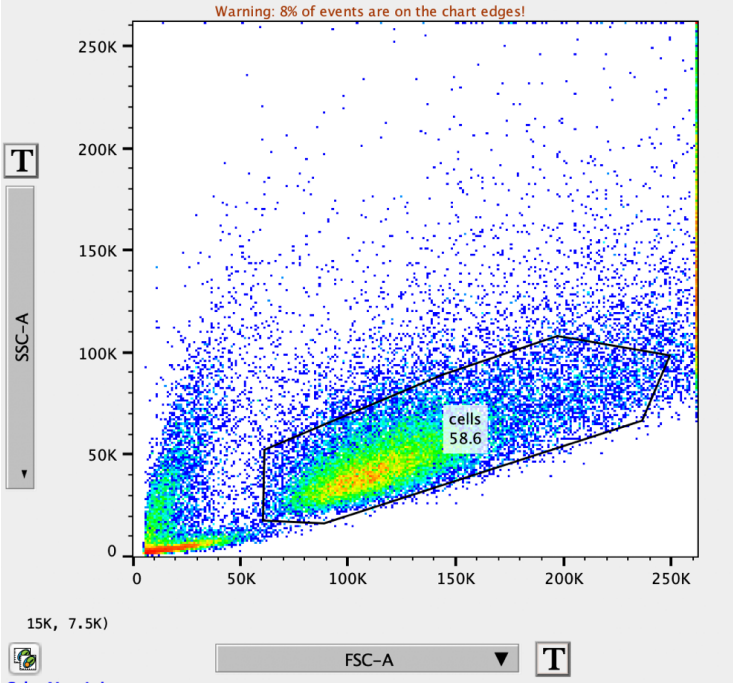
**e** Boxplot (with individual replicate activity shown as beeswarm, whiskers extend to 1.5 times interquartile range, centre median, box extends from 25$^{th}$ to 75$^{th}$ percentiles) of activity of two dynamic regulatory elements (left: *Gata4*:chr14_5729, right: *Sox2*:chr3_2007) at different times, grey: without U6/oBC, black: with U6/oBC. Number of data points: three biological replicates each with two technical replicates per time point (day 18+ quantification includes days 18, 20, and 22). Samples with an oBC cassette (black) also have two separate quantifications originating from cloned libraries p55 and p59 pooled in the experiment (see panel **a**). n=6, 6, 18 (grey, days 0, 4, and 18+ respectively) and n=12, 12, 36 (black, days 0, 4, 18+ respectively).

**f** & **g** Same as **c** & **d**, with comparison with vs. without minP, demonstrating overall lack of importance of the minimal promoter for the measured activity.

**h** & **i** Same as **c** & **d,** with comparison of CREs positioned upstream (original) vs. downstream of the reporter, showing generally muted expression. However, temporal fold-change is still significant for 7/13 dynamic CRE/time point comparisons (Bonferroni adjusted one-sided Wilcoxon test p-value for cell type-specific CRE comparisons, CRE/time comparisons with $p<0.05$ are marked by *).

**j** Same as **e**, demonstrating that element *Gata4*:chr14_5729 still leads to activation of reporter expression even when positioned downstream of the reporter. Number of data points: three

biological replicates each with two technical replicates per time point (day 18+ quantification includes days 18, 20, and 22). n=6, 6, 18 (days 0, 4, and 18+ respectively).




**Supplementary Figure 9. FACS gating strategy**

Example of a typical FACS gating strategy used to obtain high quality single-cell suspension from dissociated cells prior to 10x Genomics library preparation.

***Supplementary Notes***

**Supplementary Note 1: Systematic assessment of integration positional effects**

Clonal analysis (**Fig. 2d**) was also informative with respect to reporter expression variation driven by positional effects (assuming distinct clones harbour reporters integrated at different genomic locations). We observed promoter and cell line-specific effects, with EEF1A1p and UBCp showing remarkably little clone-to-clone variation (interquartile range across clones, UBCp: <2.4 for all cell lines; EEF1A1p: <1.5 in K562 and HEK293T, 4.1 in HepG2, **Ext. Data Fig. 3j**). In contrast, promoter Pgk1p showed both cell line differences in expression and higher variability across clones (IQR 4.8 in HEK293T, 5.9 in K562, 7.2 in HepG2, **Ext. Data Fig. 3j**). Decomposing the mBC UMI variability into positional effects (via clone assignment) vs. the sum of remaining biological and technical noise, showed that precision was limited by genomic context, underscoring the low variability of our capture and the importance of averaging over multiple independent integration positions (fraction mBC UMI variance attributable to clone identity: EEF1Ap=0.60, Pgk1p=0.41, and UBCp=0.57). Still, for the three active promoters considered here, clone-to-clone variability was substantially lower than that of uninsulated reporters[45], suggesting that insulators included in our design (**Ext. Data Fig. 1a**) partially mitigated positional variegation. Integration of promoter-driven reporters with four possible architectures ([+/- U6/oBC]×[+/- flanking cHS4]) provided additional evidence that the cHS4 insulator[48] indeed reduced positional effects across cell lines and promoters. Interestingly, we also found significant though modest and context-specific insulating effects of the U6 cassette (**Supp. Fig. 1**, **Supp. Data 3**).

**Supplementary Note 2: Comparing CRE activity and putatively associated gene inductions**

For endoderm-specific CREs, the magnitude of activity induction (fold-change of mean norm. mBC UMI per cell in parietal endoderm vs. pluripotent) was on par with endogenous gene induction (fold change parietal endoderm to pluripotent, **Ext. Data Fig. 8f**).

What proportion of endogenous regulation do the identified autonomous CREs recapitulate? This question is difficult to directly address because absolute reporter UMI counts cannot be uniformly compared to gene expression UMI counts (i.e. due to gene-to-gene differences in conversion between endogenous mRNA levels and captured UMI counts). Taking activity of the active promoter putatively associated with the induced gene (orange in **Fig. 4e**, **Ext. Data Fig. 8d**) as a baseline (with the caveat that mRNA levels driven by promoters in our reporter system might not be perfectly reflective of endogenous activity), we found that the activity of the autonomous CREs captured a substantial proportion of the expression fold-change, but in 6/7 cases less than a half (shaded **Ext. Data Fig. 8g**), as perhaps expected for multi-CRE landscapes as considered here.

**Supplementary Note 3: Analysis of features of active parietal endoderm CREs**

A number of features were enriched in the 8 active cell-type specific CREs within all 103 tested distal parietal endoderm elements tested. Active CREs displayed higher chromatin accessibility (1.8-fold more accessible, 2.2-fold more differentially accessible, both $p<0.03$ B-H corrected one-sided t-test), but showed no difference in evolutionary conservation (average phyloP score[108]), nor were they significantly closer to the TSS of their putative target gene. Indeed, at all loci, the

autonomously active CRE was not the closest element from the TSS (**Fig. 4b**, **Ext. Data Fig. 8b**). Active elements also showed no evidence of opening earlier than other elements in a pseudotime analysis[109], arguing against them being 'seed enhancers'[110,111]. With regards to finer-level sequence features, active CREs contained a higher density of endodermal regulator Gata4 binding sites, but only if considering binding sites of intermediate-to-high affinities (between 1.3 and 2.2-fold more binding sites for relative affinity lower thresholds between 0.2 and 0.45, $p<0.03$ B-H corrected one-sided t-test, 8-mer affinities from Uniprobe[104–106], binding sites also elevated comparing to all 500 bp windows ±100 kb from TSSs, **Supp. Fig. 7c**). While additional examples are needed to draw general conclusions, this suggests clusters of intermediate affinity binding sites of key regulators might be important for mammalian developmental CRE function, in line with the suboptimization hypothesis[27,112]. Two other endodermal regulators, Foxa2 and Sox17, did not show a higher number of binding sites in active CREs. In short, active parietal endoderm CREs displayed significantly elevated ATAC accessibility and Gata4 transcription factor binding sites (**Supp. Fig. 7a**), with a logistic classifier using these two properties accurately classifying active/inactive elements (auROC=0.94, **Supp. Fig. 7b**, precision=0.6 at recall=0.75).

**Supplementary Note 3: Additional applications of scQers**

Pleiotropic expression from synthetic pairs of CREs

The modularity of CREs, i.e., their ability to function independently and collectively direct complex activity profiles, is a cornerstone of regulatory genomics[4]. To generate pleiotropic expression patterns[113,114] in our *in vitro* system, we created a library of chimeric CRE pairs (pluripotent-specific paired with parietal endoderm-specific and inactive elements as controls, **Ext. Data Fig. 10a**). The pairs were assembled in all possible orders and orientations (72 different

possibilities, combinatorial library construction using Gibson assembly with shared homologies, association of CREs to barcodes using nanopore sequencing). The resulting scQer library was profiled as before in mEBs (QC metrics, **Supp. Fig. 2**, **Supp. Data 9**).

Cell-type-specific CREs displayed expression in their expected cell types when paired with inactive elements (B-H corrected bootstrap $p<0.01$), **Ext. Data Fig. 10b-c**. Importantly, the majority of active pairs led to significant expression in both cognate cell types (7/8, B-H corrected bootstrap $p<0.01$), **Ext. Data Fig. 10b-c**, consistent with the expectation of additive CRE activity. Notably, CRE order had a substantial effect on quantitative activity, with the promoter proximal element displaying higher expression across all constructed pairs (B-H corrected bootstrap $p<0.01$; median fold-change expression decrease in proximal vs. distal-containing pairs: *Epas1*:chr17_10063: 21.8, *Gata4*:chr14_5729: 15.8, *Sox2*:chr3_2007: 4.9, *Sox2*:chr3_2009: 7.4), similar to the decrease observed for reporters with CREs placed downstream of the reporter (**Supp. Fig. 8h**). Identifying the molecular origin of this effect is beyond the scope of this work, but these scQer measurements highlight opportunities and challenges towards engineering complex yet quantitative expression profiles in multicellular systems from synthetic CRE assemblies.

Profiling allelic series of CREs with optimised/disrupted transcription factor binding sites

Transcription factor binding to regulatory DNA is the proximally mechanistic step towards induction of gene expression. Changing TF-DNA binding affinity through perturbations to sequence can lead both to loss and gain-of-function (e.g., by optimising binding affinity[27,112,115]).

To assess importance of binding affinity of developmental transcription factors (TFs) within identified cell-type-specific CREs, we focused on *Gata4* (and *Gata6* given similarity of motifs) and *Sox17* (endodermal TFs[116], differentially expressed, motifs with enrichment in accessible chromatin). For six parietal endoderm elements, wild-type sequence and six variant CREs with optimization and disruption of *Gata4*/*6* and *Sox17* putative binding sites were designed (WT, *Sox17*-high, *Sox17*-low, *Gata4*-high, *Gata4*-low, *Gata4*-*Sox17*-high, *Gata4*-*Sox17*-low). Identification of putative binding sites, together with mutated sequences with high/low affinities, were based on *in vitro* biophysical measurements (UniProbe[104]*Gata4*[105], *Sox17*[106], **Supp. Fig. 7d**, **Supp. Data 8**). The resulting allelic series was cloned as a pool in a scQer library, and its activity profiled in mEBs as before, **Ext. Data Fig. 10d** (same experiment as **Ext. Data Fig. 10a-c**, **Supp. Fig. 2**, **Supp. Data 9**).

Our hypothesis was that disruption and optimization of TF binding sites would respectively ablate and increase activity across the CREs profiled. Data instead revealed a diverse range of effects (**Ext. Data Fig. 10e**). Disruption of *Gata4*/*6* binding sites universally ablated function (lower than WT for 6/6 *Gata4*-low and 6/6 *Gata4*-*Sox17*-low CREs, B-H corrected bootstrap $p<0.01$ compared to WT, median reduction: 77-fold for *Gata4*-low, and 75-fold for *Gata4*-*Sox17*-low), e.g., **Ext. Data Fig. 10f**. Surprisingly, *Gata4/6* binding site optimization also often decreased activity (3/6 CREs), and never significantly increased it, suggesting either that sequences chosen from biophysical measurements did not reflect actual TF affinity in cells, that the sequence changes also altered binding of other TFs nearby (e.g., *Sox17*, **Supp. Fig. 7d**), or that increased binding inhibited function, as has been reported in other contexts[117]. Decrease of *Sox17* binding sites affinity was more innocuous (significant decrease in expression in 1/6 CREs) while optimization had varied

effects (2/6 CRE with increased and 1/6 CRE with decreased activity). Notably, a drastic >10-fold higher expression was observed for *Sparc*:chr11_7211 with optimised *Sox17* sites compared to WT (**Ext. Data Fig. 10g**). This mutated element also displayed an increase in a similar cell type (visceral endoderm). Finally, optimising both *Gata4* and *Sox17* sites tended to decrease expression (3/6 CREs), and never to increase it.

In sum, profiling CREs with systematically perturbed putative binding sites with scQers confirmed the functional importance of these sequence features while underscoring the extent of context-dependence and CRE-specific responses to multi-position changes in sequences.

**Supplementary Note 4: Estimating scale and cost of a scQer experiment**

Synthesis of CREs

One of the factors limiting the scale of the reported experiment was our desire to test sequences of ~1 kb in size (as a rough lower bound on the sequences tested in vivo compiled in the VISTA database[118] and a compromise for PCR cloning). Indeed, we were unsure whether sequences on the scale synthesizable on arrays (200-300 bp) would be fully functional in a developmental system. Since our CREs were PCR-cloned for cost reasons, we aimed for a 96-well plate scale, i.e., hundreds. Given our success rate (97.6%, 204/209 recovered), we conceivably could have gone for a larger number. We anticipate that as longer DNA synthesis becomes more readily available at scale, or the biology of developmental CREs becomes better characterised (such that working with shorter tiles is unequivocally shown to suffice), this will not be a limiting

consideration. We also consider this limitation (cost of synthesis) to be shared with all methods studying regulatory elements and can be circumscribed by other means (e.g., testing fragments of the genome enriched for functional elements, as has extensively been done in previous work, e.g., ATAC-STARR-seq[119], FAIRE-STARR-seq[120]).

Cloning and assembly of barcode dictionary

Assuming a (possibly highly) complex starting library of CREs, our cloning strategy relies on adding the elements in a pre-barcoded backbone. The barcodes in the backbone are generated using random primers, and using standard electroporation and molecular biology, library complexity of millions can routinely be achieved (e.g., our starting backbone included ~1M oBC-mBC pairs). Following CRE addition and bottlenecking, the oBC-CRE-mBC triplet dictionary needs to be generated. Our current approach with tagmentation & semi-specific PCR could be streamlined in future iterations (especially if the CREs profiled are shorter) to be done in a single step of PCR and mapped with paired-end sequencing. In terms of sequencing cost associated with this procedure, we estimate that 25-fold read coverage of the library complexity (number of unique oBC-CRE-mBC) should suffice to obtain high-confidence triplet assignments. For library complexity, we suggest about 100 oBC-mBC pairs per CRE; a larger barcode space decreases the likelihood of spurious chimeric counts being tallied towards quantification. For our library, with ~200 CRE each with ~200 BC pairs, the sequencing coverage necessary for assignment can be estimated to be 200×200×25 = 1M read, which is a modest cost (0.25% of NextSeq2000 P2, ≈4$). More generally, the number of necessary reads for dictionary construction is estimated to be (# CRE)×100 BC pairs × 25 reads/triplet = (# CRE) × 2500. Even for libraries of thousands of

CREs, these sequencing costs amount to a small fraction of the overall sequencing budget, as detailed below.

Single-cell experiment and sequencing

We now turn to the question of scale for the limiting step: the single-cell experiment. We estimate the per-CRE-characterization-cost as a function of a few simple parameters, some set by the biological system of interest and exact nature of the library of elements.

One key variable is how many observations (detections) per CREs are needed to perform a reliable measurement. Our random transgenesis approach comes at the cost of the need to average over barcodes (i.e., over integration positions). Empirically for the cell type-specific elements we discovered (limit of detection at about average 1 mBC UMI per cell) with droplet-based 10X Genomics assay, we find that ~20-40% of mBC had a detected valid UMI in their cognate cell type. Demanding about 20 non-zero observations for averaging positional effects leads to an estimate of 100 detections per cell type. Hence, the total number of cells that need to be profiled per CRE (per biological replicate) for high sensitivity is on the order of 100 detections/cell type × the number of cell types. So, the total number of cells to be profiled can be estimated at around (assuming similar composition of cell types, more cells would need to be profiled for rare cell types etc.):

$N_{cell} = N_{CRE} \times$ (100 detections/CRE/cell type)×(number of cell types)×(replicates)/MOI.

With parameters in our first experiment (MOI ≈20, ≈10 cell types), this amounts to: $N_{cell} \approx N_{CRE} \times 50 \times 3$ rep. A simple way to increase the cost efficiency of scQer is to work with CREs

prioritised by other functional methods (e.g., bulk characterization), increasing the obtained information from the single-cell experiments.

*Sequencing coverage needed:*

GEx: We estimate from experience that $r_{GEx} \approx 4000$ reads per cell suffice in many applications (one could conceivably go lower, especially with diverse cell types) to reliably map cell types and obtain good single-cell resolution.

mBC: we typically found those libraries easy to sequence to saturation, in part because a high proportion of the tested CREs were inactive. If working with an heterogeneous sample (multiple cell types) with CREs tested anticipated to have cell type specificity, this should remain largely valid. As a point of reference with our biotype and libraries, $r_{mBC} \approx 40$ reads per mBC per cell on average led to about 4-fold saturation (4 reads per mBC UMI detected).

oBC: given the high expression (across all cells) and excellent single-cell capture of the oBC, saturating coverage would be costly (estimated UMI complexity per oBC per cell 0.8-2.5k depending on the biotype, which means with an MOI of 20, ≈20k total complexity; so that 5-fold coverage would require 100k reads per cell). Thankfully, reliable reporter detection does not require the full captured oBC complexity to be sequenced. In experiments described in this work, 300 reads per oBC per cell lead to about 5% duplication rate (reads/UMI-1) and near deterministic per cell barcode detection. Importantly, not all oBC UMIs need to be captured to reliably perform

this binary detection task. We estimate that 4-fold less coverage, or $r_{oBC} \approx 75$ reads per oBC per cell, would suffice.

Adding these contributions, we arrive at an estimate of the total number of reads scaled by the number of cells profiled (with MOI=20):

$$R = N_{cell} ( r_{GEx} + MOI \times r_{mBC} + MOI \times r_{oBC} ) \approx N_{cell} (4000 + 800 + 1500)$$

Hence, while the oBC and mBC do not constitute a negligible addition to the estimated sequencing budget, given their lower respective needs compared to GEx, we estimate (ultimately dependent on the exact application) that ~60% more reads are needed to sequence a single-cell reporter experiment (relative to a conventional single-cell GEx experiment).

*Total cost estimate (single-cell + sequencing):*

The costs need to account both for the single-cell library reagents ($c_{sc/cell}$ per cell cost) and the sequencing ($c_{read}$ per read cost). Together with earlier equations, this amounts to:

$$Cost = c_{sc/cell} * N_{cell} + c_{read} R = N_{cell} ( c_{sc/cell} + c_{read} [r_{GEx} + MOI \times r_{mBC} + MOI \times r_{oBC}] )$$

Together with our estimate for the number of cells per CRE above, this constitutes a framework to estimate the scale of the experiments. We can estimate actual costs. For 10X Genomics reagents, the per cell cost $c_{sc/cell}$ is about 0.15$/cell. The cost with a Nextseq2000 P3 100 cycles kit is at about 2.6$ per M reads (Fall 2023). Hence:

Cost (10X Genomics) =

$N_{cell}$ (0.15$/cell + $2.6 \times 10^{-6}$ $/read [4000 + 800 + 1500 reads/cell] ) =

$N_{cell}$ (0.15$/cell + 0.016$/cell)

Above, the sequencing and single-cell library preparation costs are kept separate to illustrate the different components, and shows that the bulk of the cost remains on the single-cell library generation with the droplet-based approach. Importantly, we note that alternatives to droplet-based single-cell approaches are constantly being developed and improved upon, most notably single-cell combinatorial indexing (sci)[83]. Based on this existing optimised sci protocol, the per cell cost decreases by about a factor of 100. Hence, we anticipate that throughput gains on the order of a factor of 10 at similar costs will be achievable in the near future.

**Supplementary Note 5: Additional Methods Details**

## 1 Benchmarking and optimization: promoter series in human cell lines

Single-cell reporter libraries preparation

For single-cell reporters, three libraries are generated: the standard 3' gene expression (GEx) library from 10x, and two custom derived libraries, one for each reporter RNA (oBC and mBC). The latter are obtained from nested PCRs from the amplified cDNA as we detail below.

Briefly, single-cell library preparation proceeded following the manufacturer's protocol (v3.1 manual CG000205 Rev D, 10x Genomics), with some critical modifications listed here. First, one of the replicate's cDNA (replicate B) was split in two equal halves (and brought to same final volume with elution solution 1) after GEM RT cleanup (step 2.1.s) prior to cDNA amplification to allow for a direct comparison the UMIs captured with different enrichment strategy (hereafter replicate B1 and B2). For cDNA amplification, primers specific to the mBC (oSR38) and oBC (oJBL246) reporter transcripts were spiked-in the reaction (similar to TAP-seq[107]) at final concentration of 0.5 uM to boost UMI capture for replicates A and B1 (but not for replicate B2, to allow direct comparison with replicate B1). Following cDNA amplification, both the bead and supernatant derived material (steps 2.3Ax and 2.3Bxiv respectively) were saved for downstream processing. Gene expression libraries for all replicates were prepared following the manufacturer's protocol from 25% of the bead fraction amplified cDNA.

oBC enriched libraries were prepared as follows. For replicate B2 (no primer spiked in), a first outer PCR1 was performed using 25% of the supernatant amplified cDNA with primers oSR40+oJBL246 using Kapa Robust (Roche) and tracking with qPCR until the inflection point (50 uL 2x master mix, 12.5 uL supernatant cDNA, 5 µL 10 µM oJBL246, 5 µL 10 µM oSR40, 0.5 uL SYBr green, and water to 100 µL; run parameters: 3 min at 95C, and cycles 20 s at 95C, 20 s at 60C, 20 s at 72C). Amplicons were cleaned up with 1.75x Ampure XP beads, and 1/10 of the eluate was carried to the inner PCR with the remaining replicates. For replicates A and B1, the outer PCR was performed during the cDNA amplification via the spiked-in primer, and 25% of the supernatant amplified cDNA was taken as input for the next PCR. Semi-nested inner PCR was performed on all samples with primers NextP5_index1 and indexed primers oJBL425-oJBL427, with the same parameters as PCR1 and stopped before the inflection point. Final libraries were purified by 1.5x Ampure XP beads.

As a result of our Pol II reporter construct having a capture sequence (CS2, **Ext. Data Fig. 1a**) downstream of the mBC, reporter mRNAs could be captured from both the poly-dT and CS2 reverse transcription primers on the 10x beads. To systematically compare capture efficiency resulting from the two types of primers, two different libraries were generated (poly-dT captured, and CS2 captured). For poly-dT captured libraries, similar to oBC libraries, we first performed outer PCR on replicate B2 (no spiked-in primers in cDNA amplification) using primers oSR38+oJBL207, using the same PCR conditions as for oBC except for an elongation time of 50 s and an anneal temperature of 65C. 25% of the bead fraction of the purified amplified cDNA was used as template. Following 1x Ampure XP clean up, 10% of the eluate was taken for PCR2. PCR2 was performed on all replicates (directly using 25% of the bead-derived amplified cDNA

for replicates A and B1) using primers oJBL324+oJBL495 and the same parameters as PCR1, tracking by qPCR and purifying by 1x Ampure XP beads. (We note that usage of primer oJBL495 was in the first version of the protocol. Presence of the Nextera Read 2 handle forced sequencing of the mBC libraries on a separate sequencing run from GEx libraries to avoid priming conflicts on read 2. We recommend using the updated set of primers oJBL529 and associated indexing primers as described for the mEB library scQer preparation to enable sequencing all libraries on the same sequencing run). A final PCR was performed to index amplicons with primers oJBL076 and indexed primers (oJBL496-oJBL498), and the resulting amplicons purified by 1x Ampure XP beads. The CS2 libraries were prepared entirely analogously to poly-dT captured libraries, except with the following primers: PCR1 for replicate B2 (SR38+SR40), PCR2 all replicates (oJBL529+oSR40), PCR3 all replicates (NextP5_index1+ indexed primers oJBL530-oJBL532).

We note that for both mBC and oBC libraries, semi-nested PCR is necessary to obtain a clean amplicon library (multiple non-specific amplification products were visible following the outer PCR, but a highly specific product was obtained following the semi-nested inner PCR).

All libraries were diluted to 2 nM per the Tapestation D1000 HS reading, pooled, and loaded on a NextSeq 500 for paired-end sequencing the following custom conditions: read 1: 66 cycles (no custom primer); index 1: 10 cycles (primers spiked in: oJBL432, oJBL494); read 2: 76 cycles (primers spiked in: oJBL433, oJBL334). oBC libraries were resequenced to improve saturation of the highly complex oBC libraries following: read 1: 34 cycles (no custom primers); index 1: 10 cycles (primer oJBL432); read 2 38 cycles (primer oJBL433). CS2 mBC libraries were sequenced separately, with: read 1: 30 cycles (no custom primer); index 1: 15 cycles, primer oJBL534; read

2: 18 cycles, primer oJBL334. For mBC and oBC libraries, read 1 provided the cell barcode and UMI, and read 2 the reporter barcode (sequenced with custom primers).

1.7 Optimization of reporter RNA capture

Two experiments were performed to quantitatively characterise UMI capture in our system. First, as described above, one of the sample (replicate B) cDNA was split in two prior to amplification. This enabled a direct comparison of the number of UMIs captured with vs. without the addition of reporter specific primers during cDNA amplification (as opposed to relying on the template switching oligo). For both oBC and mBC, we compared the UMI counts across cell barcode-mBC pairs (valid cell barcode from gene expression data, valid reporter barcode from subassembly) in replicate B1 (with spike in primers in cDNA amplification) and B2 (without spike in primers in cDNA amplification), **Ext. Data Fig. 1d**. For mBC, we found a median 2.0x increase in UMI counts (for mBC/cell barcode pairs with 4 or more UMIs in both replicates, **Ext. Data Fig. 1d**) in replicate B1 compared to B2, suggesting increased captured resulting from spike-in (both replicates had number of reads per UMI much larger than 1, and in fact larger for replicate B2 [median reads/umi =17.9] compared to B1 [median reads/UMI=6.8], such that this difference cannot be attributed to increase sequencing coverage for replicate B2), consistently with the range previously reported in TAP-seq[107]. Performing the same analysis for oBC led to a much larger boost in the number of UMIs captured (45x increase, also not attributable to high coverage of replicate B1, **Ext. Data Fig. 1e**) as a result of the spiked-in primer. This larger difference was expected from the circular nature of barcode: given the absence of 5' end from which template switching can occur from circular RNAs, the initial cDNA amplification (primed from the template

switching oligo) effectively could not happen except from the linear oBC intermediates (expected to represent a minor fraction) in replicate B2.

In addition, we tested which of poly-dT vs. capture sequence derived primers captured more reporter mRNA. Our reporter cassette (**Ext. Data Fig. 1a**) harbours capture sequence 2 (CS2) downstream of the barcode, enabling a direct comparison, for the same reporter in the same cell, of the different number of UMIs captured from the two different RT primers. Comparing across valid cell barcodes and mBC pairs with at least one UMI captured from both poly-dT and CS2, we found a median of 15.7x more UMIs captured from poly-dT primers, likely a direct reflection of the higher stoichiometry of these primers on the 10x Genomics beads (**Ext. Data Fig. 1f**). As such, in all mRNA expression quantifications, we use the poly-dT captured number of mBC UMIs. In addition, we note that using artificial poly-A sequence in place of CS1 on the oBC would likely result in a similar boost in capture and complexity from this system (RT primer saturation should not be a problem given successful overloading of fluidic emulsions without observed decrease in capture efficiency[121]).

1.8 Estimating per oBC per cell captured library complexity

The single-cell oBC libraries were highly complex and as result were not sequenced to saturation. UMI count distributions shown in **Fig. 2c**, **Ext. Data Fig. 2b** (and similarly in mEB, **Ext. Data Fig. 4f**) were therefore not a measure of the full complexity of the libraries (median duplication rate, i.e., read counts over UMI counts minus 1, of 8.3% and 8.8% respectively for oBCs in the high count mode [≥12 UMIs]). To estimate the total complexity for each oBC in each cell, we used

the maximum likelihood estimator from the zero-truncated Poisson distribution[122], i.e., if for a given oBC with a given cell barcode x:=(reads counts)/(UMI counts), and $\lambda$:=(read counts)/(oBC complexity), then $x = \lambda(1-\exp(-\lambda))^{-1}$. Inverting the relation for each cell barcode and oBC pair in high count mode (provided reads counts is not equal to UMI counts), we find a median complexity of 3652 (K562), 2306 (HEK293T), and 1675 (HepG2) UMI per oBC per cell barcode (replicate A). As expected from splitting the cDNA in two, the estimated complexity was essentially halved for replicated B1 (1731 in K562, 1217 in HEK293T, 850 in HepG2).

1.10 Estimating the probability to have multiple integration per cell for one plasmid

In order to estimate the probability that a unique oBC-promoter-mBC plasmid ends up integrating multiple times in a given cell, we simulated the genomic integration process by drawing multiplicity of integration from the empirically observed distribution (assuming no multi-integration events), and drawing oBC-promoter-mBC combination with replacement with frequency taken as the quantified proportion in our pool (as assessed during the subassembly stage by sequencing the barcodes). We find the probability of multiple integration, regardless of barcode, to be less than 5% in both replicates. This is somewhat higher than expected from the multi-integration probability from a scenario with fixed number of integration sampling exactly equiprobable barcodes (equal <2% under these circumstances). Spread in both MOI (here 2 to 8 integrants per cell interquartile range) and non-even barcode representation (1122 unique oBC-promoter-mBC triplets in our pool, with $10^{th}$ to $90^{th}$ percentile of representation in the pool spanning a 10-fold range: 0.00018 to 0.0017), thus contribute to modestly inflate the likelihood of a multi-integration event, which we nevertheless expect to be rare (<5%) from this empirically derived estimate.

1.11 Clonal cell analysis

Identifying clonal cells harbouring multiple genetic payloads from single-cell data is a non-trivial computational problem, even with high signal to noise ratio, in part due to doublets and barcodes multiply represented across different clones. After assessments of existing approaches[42,43,123] and our own attempt (iterative clustering in high dimensional PCA space from oBC expression), we settled on a modification of the heuristic put forth by Wang and colleagues[43], based on one-sided Fisher's exact test, followed by our own addition of custom quality filtering. Details are provided below.

*1.11.1 Clonotype identification*

We identified high-confidence integration genotypes (hereafter clonotypes) in a two step procedure. First, a raw clonotype identification, followed by a refinement step.

As a first pass identification of clonotypes, we followed Ref[43] and looped through cells (considering cell barcodes assigned to different cell lines from their transcriptome and different replicates separately), assembling a list of clonotypes. Specifically, for each cell, the list of detected oBC was extracted (≥12 oBC UMI, see below for justification for threshold in addition to corresponding to the minimum of the UMI count distribution, **Fig. 2C** and **S3B**). The list of barcodes from the cell was then compared to the oBCs detected in all other stored clonotypes via a one-sided Fisher's exact test, with contingency table given by (number of oBC in cell and clonotype, number of oBC in clonotype but not in cell; number of oBC in cell not in clonotype,

number of oBC from library neither in the clonotype nor in the cell). The test serves to assess the probability that random sampling of oBC leads to as much overlap between cell and clone as observed. A 5% Bonferonni corrected ($0.05/(n_{cells}^2/2)$) p-value was used as a threshold to determine whether a cell was a likely member of a clonotype or not. If oBC from the cell did not overlap significantly from any stored clonotypes, the cell was taken as the representative of a new clonotype. If overlap with multiple existing clonotypes was identified, the cell was marked as a likely doublet. Given the number of cells and barcodes in our experiment (promoter series), *bona fide* clones with only two reporters integrated did not meet the stringency threshold of our test and were thus excluded *de facto*. We note that this heuristic takes the set of oBCs detected from the first representative of a clonotype as the set for all cells (no aggregative correction applied), and as such the resulting cell assignments and clonotypes depend on the order at which cells are considered in the loop. To address this, we implemented an additional downstream filtering step, and returned to the problem of assigning cells to clonotypes after the final list of high-confidence clonotypes was determined.

To refine the raw list of clonotypes identified above, we first exclude clonotypes assigned to a single cell. Then, for a given clonotype, we obtained the union of all detected oBCs ≥12 UMI (clonotype oBCs) from cells assigned to that clonotype. For each of these clonotype oBC, the fraction of cells assigned to the clonotype with detection of that oBC was then determined. We then determined the number of clonotype oBC which were detected between 25% and 75% of cells within the clonotype ($n_{25\% \text{ to } 75\%}$), and the number of clonotype oBC detected in more than 75% of the same cells ($n_{>75\%}$). We also stored the maximum fraction of cells with detection of any one of the oBC within a clonotype (max_frac_detect). We found these quantities to be useful to filter out

likely doublets and clonotypes with too much barcode overlap from valid and easily distinguishable clonotypes. We retained clonotypes for which $n_{>75\%}>2\ n_{25\%\ to\ 75\%}$ and max_frac_detect>0.9. The list of oBC corresponding to a clonotype was then taken as those detected in >50% of cells assigned to that clonotype. Finally, completely nested clonotypes (ones whose set of oBC was a strict subset of another) were eliminated.

*1.11.2 Mapping of cells to clonotypes*

Using the list of clonotypes and associated oBC (described above), we returned to the complete dataset to assign cells to clonotype (thereby avoiding the issue of cell ordering affecting the outcome). Specifically, we obtained the list of detected oBCs (≥12 oBC UMI) in each individual cell. We then computed two quantities across all clonotypes, 1) $f_1$:= fraction of oBCs detected in the cell of interest also present in the clonotype, and 2) $f_2$:= fraction of clonotype oBCs detected in the cell of interest. In words, $f_1$ tracks possible additional barcodes detected in the cell not associated with the clonotype (e.g., doublets), while $f_2$ monitors possible dropouts. For each cell, the top clonotype was taken as the one with the largest $f_1$, and the associated $f_2$ was also retained. Cells were assigned the status of a high confidence singlet if $f_1>0.975$ and $f_2>0.5$. Hence, we stringently filter out possible doublet (require high $f_1$), but remain loose on possible dropouts (allow for low $f_2$, compared to performance, see below). Cells with $f_2\leq0.5$ were considered missed clonotypes.

Through this procedure, we obtained a high proportion of cells assigned as singlets to high-confidence clonotypes (replicate A: K562 67%, HEK293T 66%, HepG2 75%; replicate B1: K562 61%, HEK293T 58%, HepG2 55%), and substantial fraction of the non-singlet cells had an MOI of 2 or lower (replicate A: K562 82%, HEK293T 63%, HepG2 61%; replicate B1: K562 61%,

HEK293T 54%, HepG2 30%). Hence a high proportion of missed clonotypes came from low MOI cells and the high stringency of our p-value threshold. The number of clones was somewhat lower than estimated going in the bottleneck, possibly as a result of clonal competition. HepG2 in particular displayed a more severely bottlenecked population (**Ext. Data Fig. 3a-b**), in line with the longer time necessary for those populations to expand (slow growth was observed at the low plating density). Final clonotypes with cell assignments are listed in **Supp. Data 2**.

Final assignments (with clonotypes with 3 or more cells assigned) were displayed on oBC expression space UMAP (**Ext. Data Fig. 3a-b**) using Seurat (Normalisation method "RC", PCA run on variable features with 100 principal components, UMAP with n.neighbors=10 on top 50 PCs).

*1.11.3 Systematic oBC dropout analysis (precision-recall)*

The high-confidence clonotypes identified through the consensus of co-detected barcodes served as an approximate ground truth to systematically assess the detectability of oBC in our assay. Specifically, for all clonotypes (with 3 or more cells assigned) and singlet-assigned cells as described above, we computed for different oBC UMI count detection threshold the number of true positives (TP:= number of oBC detected in cell also in the clonotype), false positives (FP:= number oBC detected in the cell not present in the clonotype), and false negatives (FN:= number of oBC in the clonotype not detected in the cell). At each oBC threshold, the false discovery rate was taken as FDR:=sum(FP)/(sum(FP)+sum(TP)), and the false negative rate FNR:=sum(FN)/(sum(FN)+sum(TP)), where the sums are over all cells and clonotypes. Results

stratified by cell lines are shown in **Fig. 2h** (stratified by replicates: **Ext. Data Fig. 3e-f**). Direct representative oBC count distributions are shown in the count matrices to two typical clones shown in **Ext. Data Fig. 3c-d**. In order to prioritise stringency, we selected a UMI of 12 as threshold for the expression analysis presented throughout (different threshold for experiment in embryoid bodies, see below). We note that given the loose stringency of our threshold for assignment (tolerating cells with up to 50% dropout in oBC), this analysis should be relatively unbiased given that the FNR is in the few percent range at the threshold of 12 oBC UMI.

*1.11.4 Analysis of reporter barcode expression variability across clones*

In addition to providing assessment of oBC dropout, clonal cells present an opportunity to measure variability in the number of captured reporter mRNAs, while controlling for possible positional effects. For each singlet-assigned cells to high confidence clonotypes, the count distribution of oBC UMI and mBC UMI (GEx normalised) corresponding to the integrated reporter cassettes was obtained (e.g., **Ext. Data Fig. 3g-h** for two example clones). The mean across cells assigned to clonotypes and standard deviation in these quantities was determined (analysis restricted to clones with >4 cells assigned to allow for a robust assessment of the standard deviation). The coefficient of variation (standard deviation over mean) was displayed as a function of the mean (**Ext. Data Fig. 3i**), showing scaling close to the limit set by Poisson counting even for some of the highly expressed promoters, and typically much lower than one. This provides direct evidence that when controlling for positional effects and conditioning on presence of the reporter by orthogonal means (here with oBC detection), single-cell measurements can be highly precise.

To assess the proportion of variance attributable to positional effects vs. other technical and biological factors, we used the law of total variance to decompose in mBC UMI variability. For each separate cell line and promoter, we computed the variance of the mean mBC UMI per clone-reporter pair (explained variance) and the mean variance of mBC UMI across clones-reporter pairs (unexplained variance). We find the proportion of unexplained variance to be (average of replicates A and B1): EEF1A1p K562=0.46; HEK293T=0.37, HepG2=0.37; Pgk1p K562=0.44, HEK293T=0.73, HepG2=0.60; UBCp K562=0.33, HEK293T=0.41, HepG2=0.55. The reporter values in the main text are the average over the cell lines.

Variation of the mean expression for the different promoters across clonotypes also provided estimates for the magnitude of genomic context positional effects. We found restricted variability for most promoters, although with evidence for cell-type-specific differences (interquartile fold-change range, for all promoters listed from K562, HEK293T, and HepG2: UBCp: 2.1, 2.2, 2.4; Pgk1p: 5.9, 4.8, 7.2; EEF1A1p: 1.5, 1.5, 4.1), somewhat smaller than the positional effects observed from the Pgk1 promoter in mES cells[45] which had an observed fold-change interquartile of ≈8, suggesting that positional variegation arising from local the epigenetic environment might be partially mitigated by presence of insulators (core cHS4[48]) in our construct (**Ext. Data Fig. 1a**). See **Supp. Fig. 1** for a direct test confirming the importance of the cHS4 sequences.

Clonal analysis also allowed us to compare the distribution of CV across clones for the raw UMI counts, and the GEx normalised UMI counts. We found small but consistent decreases in variability both for oBC (median GEx normalised CV/raw CV = 0.80) and some promoters (UBCp and EEF1A1p: median GEx normalised CV/raw CV=0.85, no difference for, the no promoter,

minimal, and Pgk1p promoters), justifying our use of this normalisation in our quantification of reporter mRNA.

## 2 Profiling developmental cis-regulatory elements in mouse embryoid bodies

### 2.2 scATAC-seq on mEBs

#### *2.2.1 Experimental method*

Single-nuclei preparation for scATAC-seq were prepared from day 21 mEB as follows: At day 21 of mEB cultures, mEBs (two 10 cm suspension plates) are collected into a 50 mL conical tube and washed 2x with 1x PBS (without $Ca^{2+}$, $Mg^{2+}$). After consecutive PBS washes, mEBs are treated with 1.5mL of 0.25% trypsin and incubated in 37C bath with gentle agitation (steady concentric swirls in 50mL conical tubes) for 3 minutes. For further dissociation, mEBs are then gently triturated 10 times with a P1000 pipette and again incubated at 37C for 3 minutes with gentle agitation. After second incubation, mEBs are gently triturated 10 times with a P1000 pipette. Trypsin digestion is inactivated with CA medium and cells filtered to single-cell suspension through a 100 um filter into a new 50 mL conical. Single cell suspension was counted, and cells spun down at 300 g for 5 minutes. After removing supernatant, wash 1x with 1 mL of 1x PBS + 0.04% BSA and gently pipette mix 5x. Transfer to a 1.5 mL tube and spin at 300g for 5 min at 4C. Wash again with 1mL of 1x PBS + 0.04% BSA. Again, spin at 300 g for 5 min at 4C, and proceeded with 10x Genomics "Nuclei isolation for Single Cell ATAC Sequencing" protocol (V1), with two biological replicates (different mEB differentiation) each with two lanes of 10x (four reactions total).

*2.2.2 Processing of scATAC-seq data*

Fastq files were generated by running makefastq. Fastq files were then processed to fragment files using 10x Genomics cellranger count (cellranger-atac-cs version 1.2.0, reference = refdata-cellranger-atac-mm10-1.2.0), which were processed through the ArchR pipeline[124]. Arrow files were created with function createArrowFiles (minTSS=4, minFrags=1000). Two different double scores were computed with function addDoubletScores (LSI based: k=10, knnMethod="LSI", LSIMethod=1); UMAP based: k=10, LSImet=1, UMAPparam: n_neighbors=40, min_dist=0.4, metric="euclidiean"). In addition, we used AMULET[125] (from its v1.0-beta version, running function ATACDoubletDetector.py and adding as problematic region a union of the ENCODE excluded list, segmental duplications, simple repeats, repeat masker, and microsatellites from mm10 obtained from UCSC). Nuclei were then filtered for: TSS enrichment >8 and >1995 fragment counts. Following dimensional reduction (addIterativeLSI: useMatrix = "TileMatrix", name = "IterativeLSI", iterations = 2, varFeatures = 100000, dimsToUse = 1:30), and clustering (AddClusters: reducedDims = "IterativeLSI", method = "Seurat", name = "Clusters", resolution = 0.5), doublets were stringently removed by inspecting distribution of fragment counts, doublet scores (ArchR derived), and AMULET doublet scores per clusters. All nuclei from clusters with anomalously high doublet scores across metrics were removed. In addition, individual nuclei with either >17782 fragment counts, LSI doublet score > 0, UMAP doublet score > 0, or AMULET score > 0.3 (thresholds assessed from the distribution of anomalous doublet clusters) were filtered out as likely doublets. In the end, 31% of nuclei were removed with these filters, leaving 46408 nuclei passing quality filters (20329 nuclei from replicate 1, 26079 nuclei from replicate 2).

The resulting filtered nuclei were dimensionally reduced and clustered (same parameters as above), leading to 9 clusters with >200 nuclei. Clusters with highly correlated accessibility (determined from pseudobulk averaging over cells in cluster) over all peaks ($R^2$ on log-transformed accessibility >0.55) and proximal in the low-dimensional projection were merged. The intermediate endoderm cluster (connecting the visceral and parietal clusters) was kept separate to avoid diluting the signal from the two otherwise well-delineated extraembryonic endoderm clusters. The final 7 clusters are depicted in **Ext. Data Fig. 4e** (see later section for integration/annotation). scATAC pseudobulk pileup traces (e.g., **Fig. 3a**, **4b**) were generated by first the normalised data using ArchR's function groupRegionSumArrows. A subset of all cell-type pseudboulks was shown due to space limitations in figures.

2.3 Prioritization of developmental loci and putative CRE selection

In order to select regulatory elements possibly implicated in control of gene expression in our system, we prioritized loci on the basis of a number of criteria. First, we identified highly differentially expressed genes within neuroectoderm, endoderm, mesoderm, and pluripotent clusters from mEBs (SGR and SD, in preparation unpublished data) using Seurat FindMarkers function, retaining genes with at least 25% detected expression in the respective clusters, and either a fold-change in expression >1.6x or a fold-change in fraction of cells detected with expression >2. We then identified all peaks from the scATAC data with score (as generated by ArchR) >20 within 100 kb of the TSS of each gene. The resulting gene-peak data table was then augmented with information about the ATAC peaks (accessibility in cell-type cognate to the differential expression, fold-change in accessibility, average phyloP score[108], distance to the nearest gene, overlap with ccRE[9], orthology to a reciprocal human ccRE). Genes were retained for further

assessment if their ±100 kb neighbourhood included 3 or more highly accessible and differentially accessible peaks (top 90$^{th}$ percentile) in the cell-type cognate to the differential expression. In addition, loci harbouring one or more non-exonic peaks with evidence of conservation (average phyloP>0.75 or presence of orthologous human ccRE) and within 50 kb of a very highly differentially expressed genes (>3.8 fold-change in expression or >7.7 fold-change in fraction of cells with detected expression) were retained. Finally, loci with one or more non-exonic peaks with either: 1) strong conservation (>2 average phyloP score) and high differential accessibility (>15x fold change), or 2) evidence of conservation (average phyloP>0.75 or presence of orthologous human ccRE) and very high differential accessibility (>15x fold change), were retained. Filtering on these different criteria led to a list of 89 loci, which were manually evaluated. To arrive at our final list of 22 loci (**Supp. Data 1**), genes were ranked by the number of peaks satisfying the above criteria (conservation and differential activity), and examples from parietal endoderm, neuroectoderm, and mesoderm cell-types with overall low gene density (to avoid the possible complication of neighbouring gene regulation) were selected.

Following loci prioritisation, the final set of putative CREs selected was any peak within 100 kb of the annotated differentially expressed genes above >9.4×10$^{-3}$ accessibility (normalised by TSS reads, average in all cells annotated to differential-expression-cognate cell-type) reproducibly in both scATAC replicates, leading to 206 regions. A strongly differentially accessible peak 2 kb upstream of the gene *Tubb2b* TSS (*Tubb2b*:ch13_2580) which had fortuitously not passed our thresholding criteria was included. We finally added the 4 constituents of the core Sox2 control region as CREs of interest to include, for a total of 211 elements. Robust primers for PCR-cloning could be designed for 209/211 of them (primers: **Supp. Data 11**, list of CRE sequences and positions: **Supp. Data 1**), see section below, and 204/209 were sufficiently represented in our

constructed scQer libraries to allow for quantification (**Fig. 4a**, **Ext. Data Fig. 5h**, **Supp. Fig. 2b-c**).

2.4 Cloning details

*Recloning of oBC-mBC backbone plasmid*

Doubly barcoded backbone plasmid (p025) was recloned in order to increase the complexity of the barcode pairs. Briefly, new barcodes were appended by amplifying the region between the oBC and mBC with primers with random (5'VNNVNNVNNVNN for the oBC) primers (oJBL513+oJBL514) with Kapa HiFi (15 cycles). Following 1.5x Ampure XP beads clean up, the barcoded insert was further amplified (15 cycles Kapa HiFi) to append homology arms for Gibson assembly (oJBL515+oJBL516). The final insert was PAGE purified. The insert-compatible backbone was reconstructed from two PCR products from p025 (oJBL524+oJBL527, oJBL526+oJBL525) of about 2.6 kb each (agarose gel purified). The three pieces were then combined by Gibson assembly, and electroporated in *E. coli* (C3020, NEB). Full complexity of the library was maintained, and estimated to be ≈1M clones by colony counting transformants.

*oBC-mBC subassembly*

Following re-cloning of p025, the oBC-mBC pairs in the library were obtained as described for the promoter series experiment by a single step of PCR (primers oJBL337+oJBL345) to append handles for sequencing (on NextSeq 500, with library structure: read 1 oJBL346 (oBC, 30 cycles); index 1 oJBL347 (library index, 15 cycles); read 2 oJBL334 (mBC reverse complement, 18 cycles); index 2 oJBL348 (oBC reverse complement). Pre-processing also was carried out as

described, resulting in oBC-mBC pairs with each associated with a read count. Given the complexity of the library (unsaturated), a cutoff of at least 5 reads was applied to retain oBC-mBC pairs (1.2M pairs). To mark possible non-uniquely paired barcodes, we computed the proportion of read counts to each oBC and mBC from a given oBC-mBC pair. oBC-mBC pairs with oBC or mBC with read counts proportion belonging to the pair of less than 95% were marked as likely non-unique (78.1% likely unique pairs by this criterion).

*2.5.3 Construction of EEF1A1p-mCherry transposon plasmid*

To obtain an orthogonal selection for co-transfection to boost MOI (see below), we cloned a constitutively expressed red fluorescent protein into the piggyBac transposon. Briefly, p001 (piggyBac transposon backbone) was digested with XbaI and EcoRI (NEB), and size selected on agarose. The EEF1A1 promoter was amplified from p003 using primers oJBL536+oJBL537, and mCherry was amplified from a puro-mCherry containing plasmid with primers oJBL538+oJBL539 with Kapa Robust. The resulting fragments were size selected on agarose, and combined with the digested backbone by Gibson assembly, and transformed. The final plasmid taken from an individual colony was confirmed by Sanger sequencing and used for co-transfection.

*2.5.4 Optimization of high multiplicity of integration with piggyBac in mESCs*

As described above, we used co-transfection of selectable carrier transposon to boost multiplicity of integration following previous successful reports[59,60] of the procedure. Importantly, to prevent any bias on expression of the integrated reporters associated with developmental CREs, we leveraged orthogonal selection modalities not associated with the CREs (puromycin and red

fluorescent protein, not green fluorescent protein). We directly confirmed the increase in MOI by comparing qPCR-estimated cargo DNA doses from genomic DNA of cells at 11 days post transfection with and without puromycin selection. The qPCR was performed as follows: gDNA extraction with DNeasy, dilution to 100 ng/uL, per well reaction with 5 uL 2x PowerUp master mix (Thermo Fisher, cat. no. A25741), 2 uL 10 mM Tris 8, 1 uL 5 uM forward+reverse primer pair, 1 uL 100 ng/uL gDNA. Primer pairs used: GFP: oJBL039+oJBL040, puromycin cassette: oJBL043+oJBL044, *Tfrc1* (endogenous locus for normalisation): oJBL276+oJBL277. We observed on average a 4.0 to 7.0-fold increase in MOI (with vs. without puro selection, cargo dose with per-sample normalisation from endogenous locus) across three biological replicates (with 10% co-transfection of puromycin containing cargo). We note that in our hands, other approaches used to optimise MOI (selecting on higher dose of puromycin, tuning relative and absolute concentration of transposon and transposase, selection on top 10% GFP intensity) did not improve MOI to the same extent as this co-transfection method. Notably, by adding a second selection round on mCherry positive cells on puro selected expanded cells (mCherry plasmid co-transfected at 1% of the cargo DNA), we saw a further increase in MOI in replicate 2B (median MOI ≈20 for replicates A and B, and up ≈50% to median ≈30 for replicate 2B, **Ext. Data Fig. 4c**) suggesting further optimization might be possible to increase median MOI beyond what has been achieved and boost power. We note that transfecting more cells might then be necessary to avoid extensive bottlenecking (already some detectable through clonal analysis, see below, even for the non explicitly bottlenecked populations replicates A and B).

2.6 Single-cell reporter libraries preparation and sequencing

The three single-cell libraries (gene expression GEx, oBC, mBC) for the mEB experiment were prepared as described for the benchmarking experiment in human cell lines (with no splitting of the cDNA, spike-in at 0.5 uM of primers oJBL246 and SR38 at first cDNA amplification to enrich for the reporter barcodes), with the following modifications.

oBC libraries were prepared the same way as for human cell line experiments, but with different P7-indexed primers for the final inner PCR (oJBL501-oJBL506). In addition, to avoid loop-the-loop products in the oBC libraries (anecdotally decreasing sequencing quality), the lowest band in the circularized ladder amplicons (see e.g., **Ext. Data Fig. 1b**, **Ext. Data Fig. 1j**) was size selected on PAGE for each library and used for sequencing.

Given the limited added value of CS2 capture for mBC (**Ext. Data Fig. 1f**), only the poly-dT captured libraries were generated for the mBC, with the following primers for the two rounds of PCRs (following initial cDNA amplification). PCR2: oJBL324+oJBL529. PCR3: oJBL076+ P7-indexed primers (oJBL530-oJBL533).

GEx, oBC, and mBC libraries were sequenced at the same time for replicates A/B on NextSeq500 (read 1: 28 cycles, no custom primers; index 1: 10 cycles, spike in primers oJBL432, oJBL534; read 2: 54 cycles, spike in primers oJBL433, oJBL334). The three libraries for replicate 2B were similarly sequenced, except with 8 cycles on index1 and 56 cycles on read 2. The oBC libraries

for replicates A and B were re-sequenced as part of a NextSeq2000 run, with 28 cycles on read1, 10 cycles on index1, and 20 cycles on read2, with primers SR40+oJBL433 in well 1, and primer oJBL432 in well 2.

## 2.7 Single-cell data processing

*Quality filtering from gene expression libraries*

Fastq files were generated using the makefastq command from cellranger (v6.0.1), and the gene expression count matrices were then generated with cellranger count, with transcriptome reference mm10-3.0.0. Raw count matrices were then imported as a Seurat[96] object (filtering genes expressed in less than 3 cells, and cell barcodes with less than 50 genes measured). Cell barcodes in the high total UMI mode with low mitochondrial RNA proportion were filtered as likely *bona fide* cells (fraction of mitochondrial UMI >1% and <15%, total gene expression UMI > 400 for samples from replicates A, B, and 2B lane1, and >1000 for 2B lane2, which was fortuitously sequenced more deeply). The filtered count matrices were then used to evaluate doublet scores using scrublet[98] (scrub_doublets command, 30 principal components, mean_center=true, normalize_variance=true), and cell barcodes with doublet score > 0.3 (separating the two modes of the simulated doublet distribution from scrublet) were filtered out. Datasets from all replicates were then combined in a single Seurat object, dimensionally reduced and clustered (NormalizeData, normalization.method= "LogNormalize", scale.factor=10000; FindVariableFeatures with selection.method = "vst", nfeatures=1000; ScaleData with all genes as features; RunPCA with identified variable features and 100 principal components; FindNeighbors, dims=1:50; FindClusters, resolution=0.2; RunUMAP, dims=1:50, n.neighbors=50) without batch

correction given the good correspondence between replicates (**Ext. Data Fig. 5e**). The cluster identities were taken as categories for cell-type expression testing (see integration section below).

The following additional quality filtering steps were applied to retain high confidence singlet cells. Clusters comprising less than 1% of cells were considered likely doublets/artifacts, and corresponding cells were removed. Cells members of each cluster identified were separately sub-clustered with the same procedure as above (except resolution 0.5 in FindNeighbors). Any sub-cluster with a median doublet score above 0.15 was deemed composed of likely doublets, and corresponding cells were removed. Cells with anomalously high gene expression UMI counts were removed (with technical lane specific thresholds: >10k for A.1, >9k A.2, >12k B.1, >9k B.2, >15k 2B.2, no anomalous cells in 2B.1). Finally, cells with an estimated MOI > 200 (roughly corresponding to the top 0.1% of the distribution, MOI estimated through oBC UMI > 10, see below) were filtered out. In the end, n=43799 cells passed all these quality filters (12859 replicate A, 15422 replicate B, 15518 replicate 2B).

*mBC and oBC libraries*

Raw data was processed in the same way as for the human cell line promoter experiment to obtain a table of barcodes (mBC or oBC) with read and UMI counts per cell barcode. For oBC libraries of replicates A and B, two sequencing runs (for higher depth) were combined into one by concatenating their fastqs (trimming to the same read size) prior to running in cell ranger for the first processing step. Only cell barcodes passing the QC filters from the GEx analysis were retained in the final count tables.

*Single-cell quantification of reporter expression*

A similar approach as the promoter series experiment was taken to quantify expression in individual cells. For a given CRE of interest, all cells with associated oBC (in the list of valid oBC-CRE-mBC triplets) captured at >10 UMI counts were retained. The associated mBC UMI counts, in the respective cells, was then divided by the depth normalised GEx total UMI counts, and multiplied by the mean normalised GEx total UMI count across all cells (as before, to on average have a normalisation factor with a mean of 1 to not systematically distort the scale of UMI counts while correcting for systematic factors such as cell sizes and overall efficiency of in-emulsion reverse transcription). To correct for slight differences in coverage between replicates, total GEx UMI count across cells was taken per replicate, and used to normalise (by direct division, under the valid assumption that the GEx libraries are far from saturation) the GEx UMI count in individual cells (different scaling factor per replicate). In cells in which multiple reporters (with different oBC-mBC pairs) corresponding to the same CRE were detected (via the oBC), the average across this normalised mBC UMI count was taken to obtain the per-cell estimate (for displaying the single-cell CRE activity maps). For statistical tests and quantification, the average was taken across integration events as opposed to across cells (not first averaging internally within each cell, and then averaging over all cells, instead directly averaging over all integration events, such that each detection event carries the same weight).

## 2.8 Bulk MPRA (CREs, mEB time series experiment)

### *2.8.1 Library preparation and sequencing*

Bulk MPRA libraries for the CRE time series were generated similarly as described for the human cell lines promoter series experiment (**Fig. 2a**) with the following modifications. Genomic DNA and RNA were extracted with the AllPrep kit (Qiagen). 40 samples (different replicate/batch/time points) were processed overall, comprising 13 samples for replicates A and B across 12 time point (day 0, 4, 6, 10, 12, 14, 16, 18, 20, 21; with two technical replicates for day 16) and 14 samples for rep2B (two technical replicates for day 0, 8, 12, 16, 18, 20; one replicate each for day 4, 6). From each sample, 2 libraries (1 gDNA-derived, 1 RNA derived) were constructed, for a total of 80 libraries. Libraries were prepared in three batches (batch1: replicates A and B, days 0, 4, 8, 12, 16; batch2: replicates A and B, days 2, 6, 10, 14, 16, 18, 20, 21; batch3: all samples from replicate 2B). For RNA, DNase treatment was applied to the first batch, but was found to be unnecessary (comparing to no reverse transcription controls), and was consequently not performed on the other two batches. As before, reverse transcription used primer oJBL358. The first PCR using the cDNA for the RNA-derived libraries was with primers oJBL077+oJBL039. The first PCR from genomic DNA was with primers oJBL039+oJBL358. The second PCR (performed on both RNA and gDNA derived samples) using primer oJBL077 and a set of indexed primers (oJBL359-oJBL366, oJBL437-oJBL448, oJBL555-oJBL564).

Each preparation batch was sequenced separately. Batch 1: Nextseq500; read 1: 28 cycles, primer oJBL369 (mBC forward); index 1: 8 cycles, primer oJBL435 (UMI); read 2: 43 cycles, primer oJBL371 (mBC reverse); index 2: 6 cycles, primer oJBL370 (P5-index). Batch 2: Nextseq2000, same set of primers (well 1: oJBL369+oJBL370, well2: oJBL370+oJBL435), read 1: 28 cycles,

index 1: 10 cycles, read 2: 20 cycles, index 2: 6 cycles. Batch 3: Nextseq500: read 1: 18 cycles, index 1: 10 cycles, read 2: 20 cycles, index 2: 6 cycles.

*2.8.2 Data processing and quantification*

Data was pre-processed in the same way as for the human cell line bulk MPRA. Briefly, following demultiplexing with bcl2fastq (v2.20), mBC reads were trimmed to their expected lengths (15 nt) with seqtk's trimfq. mBC reads were then joined/error-corrected using PEAR (v0.9.11, options -v 15 -m 15 -t 15). The correctly assembled barcodes were then reformatted and merged with the (pseudo-)UMI read using custom python scripts, resulting in a list of mBC-UMI paired reads. The read counts for each mBC-UMI pair was determined, and a final pileup performed to generate a table of total UMI and read counts for each mBC.

From these raw mBC UMI counts, only mBC sequences from valid oBC-CRE-mBC triplets (including the promoter series) from our subassembly (34121 mBC total, 33001 from CREs, and 1120 from exogenous promoters) were retained, and appropriate metadata information (sample, time point, RNA/DNA, etc) was appended. DNA UMI counts for each barcode (across each sample) were then normalised for sequencing depth dividing by the summed UMI counts from that sample. RNA UMI counts were similarly normalised. To obtain an activity for each CRE, we first only included well-represented mBC (requiring >20 DNA reads) for quantification. Then, we 1% winsorised DNA and RNA normalised UMI counts (to mitigate extreme outliers) across all barcodes from a given CRE. The winsorised normalised UMI counts were then summed across mBCs (for a given CRE) for DNA and RNA, and the ratio was taken to be the activity of the CRE

in that sample. For a given CRE, the averaged activity from all samples from two adjacent time points (days 0 & 2, 4 & 6, 8 & 10, 12 & 14, 16 & 18, 20 & 21) were shown in **Ext. Data Fig. 9**, with error bar the standard deviation of the mean across these samples.

As a statistical test of activity, we used a Wilcoxon rank-sum test (one-sided). At each aggregate time point, the activity from all samples from the CRE of interest was compared to the activity of basal expression controls (minimal and no promoter) from all samples/time points. The resulting p-values (across all time points and CREs) were Benjamin-Hochberg corrected. Activity displayed as significant when the false discovery rate was below 1% (**Ext. Data Fig. 6c**, **Ext. Data Fig. 9c**, summary of quantification in **Supp. Data 7**). The fold-change in activity over time (**Ext. Data Fig. 9b**) was taken as the mean activity for day 20.5 (all samples from days 20 and 21) over day 1 (all samples from day 0 and 2).

## 2.9 Single-cell data integration

### *2.9.1 Integration between scRNA-seq and Pijuan-Sala et al in vivo scRNA-seq*

We compared our day 21 mEB (containing scQers) scRNA-seq data to available *in vivo* data from mouse development[61] (E6.5 to E8.5) to annotate identified clusters from low dimensional projections of our data. Samples spanning time points E6.5 to E8.5 were obtained (using the R library 'MouseGastrulationData', function EmbryoAtlasData with all samples except ids 11, 22, and 23). The count matrix was extracted together with the metadata, and a Seurat object was created after converting the gene names for compatibility, and merged with the mEB dataset. We performed integration as previously described [126]. Briefly, a list of objects was generated from the

merged Seurat object, the two datasets were separately normalised and features identified (NormalizeData; FindVariableFeatures, selection.method="vst", nfeatures=2000). Functions SelectIntegrationFeatures, FindIntegrationAnchors, and IntegrateData were sequentially applied to the list, and the integrated data was then dimensionally reduced via scaling and PCA (ScaleData, RunPCA with 30 principal components). The PCA embedding space from the integrated dataset was used to identify neighbours using a method adapted from[127]. For each cell in the mEB dataset, the top 10 closest distance neighbours in the dataset-integrated PCA space from the *in vivo* dataset were identified, and their cell-type annotation stored. Cell annotation from the *in vivo* data was transferred if >6/10 nearest neighbours had the same cell-type label, and taken as 'uncertain' otherwise. This provided an annotation label for each cell in our mEB dataset. To aggregate the annotation across clusters in the mEB data, we determined the fraction of cells per mEB derived clusters with *in vivo* cell-type annotation, shown in the heatmap of **Ext. Data Fig. 4c**. In that representation, *in vivo* cell types with a maximum fraction across all mEB clusters <5% were not displayed for brevity. Final mEB cluster annotations were determined by inspection, and coarse-grained clusters (**Fig. 3c**) naturally combined cell-types from the same lineage. One important distinction was the label of pluripotent cells, not present in the *in vivo* dataset given that the earliest time point covered was E6.5. The putative cluster of pluripotent cells was closest to epiblast cells within this constrained label-transfer assignment (**Ext. Data Fig. 4c**), but inspection of key marker genes of naive pluripotency [63,128] such as *Esrrb*, *Dppa3* (**Ext. Data Fig. 4b**) were sharply expressed in that cluster, in contrast to markers of primed pluripotency (*Fgf5*, *Dnmt3b*) which were expressed in other clusters (**Ext. Data Fig. 4b**). These justified our identification of this cluster as pluripotent cells.

Performing the label transfer on coarse-grained annotations (grouping all endodermal cells, ectodermal cells, etc.) decreased the proportion of the 'uncertain' label, which in some instances was spuriously created by mEB cells associated with mixed populations from otherwise well-defined lineages (e.g., the multiple different mesodermal cell types). We verified that the final label transfer was robust to the number of neighbours (5 to 20) considered in the PCA integrated embeddings.

*2.9.2 Integration between scRNA-seq and scATAC-seq and correlation with in vivo data*

The scRNA-seq and scATAC-seq in mEBs was not performed on the same set of cells or as a co-assay (but samples were derived from the same mESC line). We therefore relied on computational approaches to relate the clusters of the low dimensional representations from the two modalities. To that end, we performed unconstrained integration using ArchR[124] function addGeneIntegrationMatrix which uses the functionalities of Seurat[126]. The resulting assignments unambiguously mapped clusters from the RNA to the ATAC (**Ext. Data Fig. 4d**), with some of the finer resolution achievable in the scRNA-seq (e.g., different mesodermal and neuroectodermal clusters) not distinguishable in the scATAC possibly as a result of the fewer number of nuclei sampled from these cell types.

As additional verification for the validity of these cell-type assignments on the scATAC data, we compared the data to available scATAC datasets from mouse embryos at E7.5 and E8.5[53]. We downloaded pileup bigWig scATAC files from all cell types (GEO: accession GSE205117), and generated bigWig pileup from our mEB datasets (ArchR's getGroupBW function, tileSize=50,

maxCells=100000, ceiling=10, normMethod="ReadsInTSS"). We then computed the average accessibility across all peaks called by ArchR using UCSC utility function[130] bigWigAverageOverBed. Restricting to the top 25% scoring peaks called in the mEB scATAC dataset (ArchR score >20, corresponding to 65k peaks), when then computed the $R^2$ on log-transformed peak accessibility in the *in vivo* and mEB datasets across all cell-types/clusters. The overwhelming majority clusters in the mEB scATAC data assigned from the comparison to scRNA-seq had their highest correlations to corresponding *in vivo* cell types: mEB parietal endoderm vs. *in vivo* parietal endoderm $R^2$=0.77; mEB mesoderm vs. *in vivo* mesenchyme $R^2$=0.76, vs. Pharyngeal_mesoderm $R^2$=0.72, vs. Paraxial mesoderm $R^2$=0.72; mEB neuroectoderm vs. *in vivo* Forebrain, Midbrain, Hindbrain $R^2$=0.78, spinal cord $R^2$=0.75; mEB pluripotent/epiblast vs. *in vivo* epiblast $R^2$=0.86; mEB visceral endoderm vs. *in vivo* ExE endoderm $R^2$=0.61, vs. visceral endoderm $R^2$=0.48. Only mEB surface ectoderm had a higher correlation with another in vivo cell-type (highest *in vivo* correlation to gut, $R^2$=0.67, we note that the label-transfer from scRNA-seq datasets suggests partial recognition of surface ectoderm as gut, **Ext. Data Fig. 4c**), but still had accessibility highly correlated to the expected cognate cluster (second highest correlation $R^2$=0.60 to *in vivo* surface ectoderm). Taken together, these show the mEBs harbor complex epigenetic states broadly representative of *in vivo* gene regulation.

## 2.10 Clonal cell analysis

### *2.10.1 Clonotype identification, refinement, cell assignments, basic metrics, and dropout assessment*

Analysis proceeded as described above for the human cell line experiment with minor modifications. First, only oBC associated with CREs (not exogenous promoters) were considered

for clonal assignment. That was to minimise the likelihood of spurious doublets being called as a result of the lower complexity of the promoter library (increasing the likelihood of co-integration of the same pair of barcodes in otherwise unrelated clones). The UMI cutoff for oBC detection per cell was set to >10. Other parameters for the procedure (raw clonotype identification with Fisher exact test, clonotype refinement, cell assignment to high confidence clonotype) were as before.

Across the three replicates, the fraction of cells assigned to high confidence clonotype was 4535/12859 (29%, 896 clonotypes) for replicate A, 6854/15422 (53%, 866 clonotypes) for replicate B, and 8406/15518 (54%, 360 clonotypes) for replicate 2B. The mean numbers of cells assigned for these high confidence clonotypes were respectively 5.0, 7.9, and 23.4 for replicates A, B, and 2B, consistent with replicate 2B having been directly bottlenecked. We note that evidence of substantial clonal expansion in the non explicitly bottlenecked replicates (A and B) suggests that our procedure to select high MOI cells (selection on puromycin from ≈5% of plasmid transfected containing the expressed resistance cassette) did severely reduce the complexity of the cell population. While replicate 2B corresponded to a sub-sampling of replicate B (we were not aware at the time of substantial bottlenecking in our population), most clones and cells did not overlap between the two samples (44 clonotypes identified in both samples, or 44/866 of clonotypes comprising 499/6854 clonotype-assigned cells for replicate B; 44/360 of clonotypes 2533/8406 clonotype-assigned cells for replicate 2B). Summary of clonotypes and assigned cells can be found in **Supp. Data 6**.

oBC dropout analysis (**Ext. Data Fig. 5i-k**) from the clonotypes and cell assignment was performed as described for the human cell line experiment.

*2.10.2 CRE expression pattern across clones*

From the assignment of cells to high confidence clonotypes, we sought to characterize how the expression of CREs varied across clones, with the assumption that different clones correspond to different genomic positions of integration of the reporter driven by the CRE. For each of the 10 active cell-type-specific CREs identified, we obtained the list of high confidence clones harboring at least one reporter integration corresponding to the CRE. In order to obtain sufficient statistical power to estimate expression, we then restricted the analysis to clones with 5 or more assigned cells in both the cell type of expected CRE expression (e.g., pluripotent for *Sox2*:chr3_2007, parietal endoderm for *Gata4*:chr13_5729, etc.) from the analysis over all cells, and 5 or more cells assigned to the rest of cell-types. We then computed the fold-change in mean reporter expression (average normalised mBC UMI) over cells in these two compartments (cognate vs. rest of cells), and calculated the number of clones per CRE for which fold-change was >5. For 9/10 CRE, more than ⅔ of clones retained a >5 specificity (**Supp. Fig. 6**).

2.11 Analysis of features of profiled putative developmental CREs

Various features of the profiled CRE were considered for correlation with cell-type-specific activity. These were determined as follows:

ATAC accessibility: for each peak in each cell, the corresponding read count was normalised by the total number of TSS reads in that nucleus. The average overall cells assigned to a given cluster was then taken as the mean accessibility for the given peak in that cluster. Fold-change in

accessibility was taken as that measure of accessibility over the mean accessibility averaging over cells from well-delineated clusters (pluripotent/epiblast, neuroectoderm, mesoderm, and extraembryonic endoderm/parietal), such that for example fold-change accessibility for parietal endoderm was: mean accessibility in parietal endoderm divided by mean accessibility in all cells from epiblast/pluripotent, neuroectoderm, and mesoderm clusters. The visceral and intermediate parietal endoderm clusters were not considered for the fold change computation to not have cells from the same lineage be included in the comparison, which could have artificially decreased the effect size for parietal endoderm.

Pseudotime opening: In the absence of a time series scATAC-seq dataset, we considered pseudo-time trajectories[109]. First, scATAC data was clustered at higher resolution (resolution=2) using ArchR's addClusters function. A trajectory from pluripotent to parietal endoderm, passing through these more highly resolved clusters, was then defined and created with the addTrajectory function. Accessibility information along the trajectory was extracted with function getTrajectory (useMatrix="PeakMatrix", log2Norm=TRUE, and smoothWindow=10). Pseudotime smoothed accessibility values for each considered peak (distal parietal endoderm) was then obtained. To estimate the pseudotime at which a peak became accessible, we fit an exponential sigmoid (logistic function, using SSlogis and nls in R) to each accessibility vs. pseudotime trace. The pseudotime at which the sigmoid reached 20% of its maximum from baseline was selected as the heuristic value to compare opening times of the different peaks.

Evolutionary conservation: to assess evolutionary conservation, we calculated the average phyloP [108] score (mm10.60way.phyloP60way.bw) over ArchR-defined 500 bp ATAC peaks using

function bigWigAverageOverBed[130] similarly to previous assessment of non-coding element conservation[9].

For single-feature classifiers, a simple thresholding on the feature was used to generate the ROC curves (**Supp. Fig. 7b**). To combine cognate ATAC accessibility and number of Gata4 binding sites, we used scikit-learn[131] function LogisticRegression with an l1 penalty (mean=0 and standard deviation=1 input variables) and the roc_curve function to compute the performance metric.

## 3 Pol III driven circular vs. linear barcode MPRA experiment

### 3.1 Cloning of plasmids

The Tornado cassette was first cloned in a piggyBac transposon. The piggyBac cloning dock p022 was digested with BbsI (NEB) and the U6-Tornado-Broccoli insert excised from the pAV-U6+27-Tornado-Broccoli plasmid[34] (Addgene #124360) with BamHI and XhoI (NEB) digestion. Both backbone and insert were purified by agarose gel extraction (Zymoclean Gel DNA recovery kit, Zymo Research), the fragments combined by isothermal assembly (HiFi NEBuilder, NEB) into plasmid p051, and transformed in *E. coli* (NEB, C3040H). A single clone was selected and the plasmid confirmed by Sanger sequencing. A truncated version of the Tornado cassette, excluding the 5' and 3' portion of the ribozyme not overlapping with the final circular RNA sequence, was generated by digesting p051 with XbaI and SalI, and combining with gblock linear_TB_CS by isothermal assembly. The resulting plasmid, p052, was transformed in *E. coli* (NEB, C3040H). A single clone was selected and the plasmid confirmed by Sanger sequencing.

To generate complex libraries barcodes, barcoded inserts (5'VNNNVNNNVNNNVNNN) with downstream capture sequence 1 (CS1, 5'GCTTTAAGGCCGGTCCTAGCAA) were amplified from ultramer uJBL519. For circular barcodes, primers oJBL520+oJBL521 were used for amplification, the resulting product PAGE purified, and inserted in NotI+SacII digested p051 purified by agarose gel extraction by isothermal assembly. The plasmid library, p053 (**Ext. Data Fig. 1h**), was concentrated and eluted in water (Zymo Clean and Concentrator, Zymo research), and electroporated in *E. coli* (NEB, C3020) following manufacturer's instruction. A similar procedure was taken for linear barcodes, except that primers oJBL522+oJBL523 were used to amplify uJBL519, and integrated in NotI+SacII digested p052, resulting in plasmid library p054 (**Ext. Data Fig. 1i**). Of note, both circular and linear barcode constructs were compatible for reverse transcription and amplification from the same primers for library preparation to minimise biases. Following outgrowth post electroporation, a dilution series was plated to assess library complexity, and populations estimated at 50k clones were expanded, and resulting plasmids libraries purified (ZymoPure II plasmid Midiprep kit, Zymo Research), and further concentrated to 1 μg/μL by isopropanol precipitation.

3.2 Transfection, cell culture, and cell harvesting

2.75 μL each of plasmids libraries p053 and p054 were mixed to 0.6 μL (0.3 μg) of SBI super piggyBac, and transfected 2.5 M of exponentially growing K562 cells in duplicates using a Nucleofector following manufacturer's protocol for K562 (kit V4XC-2024, Lonza BioResearch) in duplicates. After two weeks of exponential growth with ⅕ split every two days to allow for dilution of unintegrated plasmids, cells were harvested in exponential phase (<0.75 M/mL), and methanol fixed. Briefly, cells were pelleted at 500 g for 5 min, washed with ice cold 1x PBS to 2

M/mL, pelleted at 500 g for 5 min, resuspended to 15 M/mL, and ice cold methanol was added drop by drop to 80%. Aliquots of 4 M fixed cells were stored at -80C until DNA or RNA extractions.

3.3 Massively parallel reporter assay library generation and sequencing:

Bulk MPRA for Pol III barcodes proceeded similarly as described before. Fixed cells were split, and genomic DNA was extracted from methanol fixed cells using the DNeasy kit (Qiagen), and RNA was extracted from cells using TRIzol LS (Thermo Fisher), following manufacturer's instructions in both cases.

Amplicon libraries from DNA were generated in two steps of PCR amplification with Kapa HiFi (Roche). For genomic DNA, 500 ng of input was used, and for plasmids (to map barcodes present in both constructs), 3 ng was used. For low-cycle number PCR1, 500 ng of DNA was mixed with 50 µL 2× Kapa HiFi master mix, 5 µL 10 µM oJBL246, 5 µL 10 µM oJBL424, and water to 100 µL. Cycling parameters: 1 min at 95C, and 4 cycles of: 20 s at 98C, 20 s at 60C, 30 s at 72C, followed by 4C hold. Primer oJBL424 contains 10 random Ns to serve as a pseudo-UMI (hereafter referred to as UMIs for brevity) to correct for PCR jackpotting. Reactions were cleaned up with Ampure XP beads (Beckman Coulter) at 1.75×, and eluted in 20 µL of 10 mM Tris 8. Illumina adapters and sequencing indices were appended through PCR2, with 4 µL of the eluate from PCR1 taken as input, and 25 µL 2× Kapa HiFi master mix, 0.25 µL 100× SYBr green, 2.5 µL 10 µM oJBL076, 2.5 µL 10 µM indexed primers (DNA rep1: oJBL501, DNA rep2: oJBL502, plasmid p053: oJBL427, plasmid p054: oJBL504), and water to 50 µL. Libraries were amplified with

tracking by qPCR with: 1 min at 95C, and cycles up to the qPCR inflection point (typically 15-17 cycles) of: 20 s at 98C, 20 s at 60C, 30 s at 72C. Libraries were then cleaned up with Ampure XP beads at 1.75×.

Amplicons libraries for RNA were obtained by first DNase treating the RNA (5 µg RNA, 2 µL TURBO DNase [Thermo Fisher], 2 µL 10× buffer, and waster to 20 µL, incubated at 37C for 30 min, cleaned up with RNA clean & concentrator [Zymo Research], and eluted in 11 Tris 7 10 mM), and taking 1 µg of DNase treated RNA to reverse transcription. Briefly, 2 µL (500 ng/µL) RNA was mixed with 2 µL 1 µM oJBL424, incubated at 65C for 5 min, and placed on ice. 15 µL of reverse transcription master mix was then added (4 µL 5× FS buffer, 1 µL 0.1 M DTT, 1 µL 10 mM dNTP mix, 8 µL water, 1 µL SSIII [Thermo Fisher]), and the reaction incubated at 55C for 60 min, followed by 70C for 15 min. ¼ of the reverse transcription reaction was then directly amplified for PCR1 (37.5 2× Kapa HiFi master mix, 3.75 µL oJBL246, 3.75 µL oJBL076, water to 75 µL), with cycling parameters: 1 min at 95C, and 4 cycles of: 20 s at 98C, 20 s at 60C, 30 s at 72C, followed by 4C hold. Reactions were cleaned up with Ampure XP beads (Beckman Coulter) at 1.75×, and eluted in 20 µL of 10 mM Tris 8. PCR2 proceeded as for libraries prepared from plasmids and genomic DNA, with indexing primers oJBL508 and oJBL509 for replicates 1 and 2 respectively, and reactions stopped at inflexion point from qPCR tracking (cycle 7). Libraries were then cleaned up with Ampure XP beads at 1.75×. Notably, given the circular nature of the Tornado barcodes, rolling-circle loop-the-loop RT products were prominently visible (at least 4-loops products detectable) at the expected size laddering from repeats of the circular RNA length (**Ext. Data Fig. 1j**). To prevent possible phasing issues on the sequencer, the product of the lowest size, which was the same for both linear and circular barcodes, was purified by PAGE extraction.

Final amplicon libraries were quantified with Qubit dsDNA HS (Thermo Fisher), diluted to 3 nM, run on TapeStation D1000 HS (Agilent) for final quality assessment, and adjusted to final 2 nM based on the Tapestation quantification. Libraries were pooled, loaded as a fraction of a NextSeq500 lane, and paired end sequenced with the following parameters: read1 (barcode forward): 25 cycles with primer oJBL431, index1 (index): 20 cycles with primer oJBL432, read2 (barcode reverse): 20 cycles with primer oJBL433, index2 (UMI): 10 cycles with primer oJBL434.

3.4 Data pre-processing and quantification

Sequencing data was demultiplexed using bcl2fastq. Raw fastq files were processed first by trimming unnecessary cycles from the 3' end (9 cycles from read 1, 4 cycles from read 2) using seqtk (https://github.com/lh3/seqtk). Forward and reverse barcode reads were joined and error corrected with PEAR[95] (options -v 16 -m 16 -n 16 -t 16). Using custom python and R scripts, successfully assembled barcode reads were combined with UMI reads, barcode/UMI pairs were counted, and the read counts and UMI count per barcode was determined. These barcode count files served as the processed inputs for downstream analysis.

For downstream processing, the identity of barcodes present in each transfected plasmid library (p053 and p054) was first determined by inspecting the distribution of barcode UMI counts from separate libraries directly prepared from the respective plasmids. The count distribution displayed clear bimodal nature, and barcodes in the high count mode (>9 UMIs for p053, >5 UMIs for p054)

were retained as valid. Following removal of barcodes present in both libraries (34 out of 193705), we were left with a list of barcodes for expression analysis (59.0k for p053, 134.6k for p054).

For quantifying steady-state expression of linear and circular barcodes, we tallied the UMI counts for all valid barcodes for genomic DNA and RNA derived libraries. DNA UMI counts were reasonably correlated from genomic DNA to plasmid ($R^2$ of log-transformed BC UMI counts= 0.42, and 0.43 respectively for replicate 1 and 2). Steady-state expression (referred to as "activity") was defined as the normalised RNA UMI counts over the normalised DNA UMI counts, with normalised UMI counts defined as UMI counts over all UMIs mapping to valid barcodes in the respective libraries. For BC well represented in the library (>50 DNA UMI counts), the activity was >150-fold higher for Tornado barcodes compared to linear barcodes (**Supp. Fig. 2e**, median activity fold-change 162× in replicate 1 and 186× in replicate 2). Difference in activity was largely insensitive to threshold selection on DNA UMI, and the summed RNA/DNA UMI counts across all linear vs. circular barcodes irrespective of DNA UMI counts confirmed >100-fold higher in activity for circular over linear barcodes (107× for rep1, 113× for rep2). Circular barcodes had a tight range in activity across barcode sequences (interquartile range in activity spanning 2.5-fold and 2.6-fold for replicates 1 and 2 respectively).

3.5 Estimating expression levels of oBC per cell per integrated cassette

To estimate the relative steady-state expression level of oBC driven by human U6 Pol III promoters, we used two different reverse-transcription qPCR quantifications, from K562 cells with genome-integrated dual reporter constructs. First, following cell harvesting and RNA extraction as previously described, 1 μg of DNase-treated RNA was combined with 100 pmole of random hexamer in 2 μL of 10 mM Tris 7 buffer, incubated at 65C for 5 min, and placed back on ice. 8 μL

of MuLV mix (1 µL 10× buffer, 0.5 µL 10 mM dNTP mix, 6 µL DEPC treated water, 0.5 µL MuLV [NEB]) was added to the RNA and random hexamer mix, and incubated at 25C for 5 min, 42C for 60 min, and 65C for 20 min. RNA was hydrolyzed from the reverse transcription mix by adding 2 µL of 1M NaOH and heating to 95C for 5 min. The cDNA was subsequently neutralised by adding 2 µL of 1M HCl, and diluted ten-fold by adding 86 µL of 10 mM Tris 8. For each primer pair, 2 µL of diluted cDNA was directly used for qPCR, and run by adding 2 µL of 10 mM Tris 8 and 5 µL of PowerUp master mix (ThermoFisher) and 1 µL forward+reverse 5 µM primer mix per well. Each primer pair/sample was run in technical triplicate wells, with PCR conditions (2 min at 50C, 2 min at 95C, and cycles: 15s at 95C, 15s at 60C,, 15s at 72C). qPCR primers targeting both the reporters (Pol III oBC: oJBL246+oJBL247, Pol II GFP mRNA: oJBL039+oJBL040), and highly expressed endogenous genes EEF1A1 (oJBL001+oJBL002, with these primers obtained from [132]) were used. oBC expression was normalised to EEF1A1 level using a ΔCt method. To normalise for multiplicity of integration in the genome, we performed qPCR from extracted genomic DNA extracted with DNeasy (Qiagen), using 100 ng gDNA input per triplicate (5 µL PowerUp SYBr mix, 1 µL forward+reverse 5 µM primer mix, and 10 mM Tris 8 to 10 µL), using the same cycling parameters as for RT-qPCR, and primers targeting the piggyBac reporter payload (GFP: oJBL039+oJBL040, puromycin cassette: oJBL043+oJBL044) in addition to endogenous genes for normalization (RPPH1: oJBL085+oJBL086, TERT: oJBL091+oJBL092). Relative levels of oBC RNA and Pol II GFP mRNA compared to the endogenous EEF1A1 mRNA were normalised by DNA dose per cell inferred from qPCR, leading to 5.2±0.8 for oBC and 0.12±0.04 for GFP (± standard error of the mean from 4 biological replicates), as the estimated expression per integrated copy expression. GFP level was the average produced by the five promoters included in the exogenous library (no promoter, minimal promoter, UBCp, Pgk1p, EEF1A1p),

most of the expression coming from EEF1A1 promoter, and indeed close to the expected level of the endogenous EEF1A1 mRNA level when correcting for this factor (5×0.12=0.6≈1).

As an additional measurement, which might be not affected by possible systematic underestimation given the fact that oBCs are short (134 bp), leaving fewer space for priming from random hexamers, we used the qPCR cycle number obtained from preparation of sequencing libraries, which involves reverse transcription from target specific primers instead of random hexamers. For oBC, the same approach described for the linear vs. circular barcode MPRA was taken. For the mBC (the Pol II reporter mRNA), we used the same procedure, except with the following primers: reverse transcription with primer oJBL358, PCR1 with primers oJBL077+oJBL039, PCR2 with primers oJBL077 and one of oJBL359-oJBL366. Comparing the qPCR Ct value for oBC vs. mBC libraries, we estimated a ΔCt of 10.4±1.2 (± spread across two biological replicates) corresponding to a relative abundance fold-change of ≈1300 for oBC vs. mBC (we note that different RT or PCR primer efficiency could drive part of this difference), which is internally controlled for multiplicity of integration as both are part of the reporter construct. Correcting for the difference between the endogenous EEF1A1 mRNA and GFP reporter (per integrated copy) seen with random hexamers led to an estimate of 1300×0.12=156-fold higher oBC expression compared to the EEF1A1 mRNA, which is one of the most highly expressed mRNA in K562 cells (as assessed from the average of stranded bulk RNA-seq datasets from ENCODE [97] in K562). Given the presence of multiple rolling-circle reverse transcription products (**Ext. Data Fig. 1j**), we note that this quantification can be considered a slight overestimate. Taking the geometric mean of the random hexamer and target specific primer as an estimate of oBC abundance leads to ≈32-fold higher expression of oBC relative to the EEF1A1

mRNA. Given that EEF1A1 comprises 1.2% of mRNAs in K562 (estimated from bulk RNA-seq TPM), and taking 200,000 total mRNAs per cell (BNID109916 [133]), this leads to an estimate of 1.2%×200,000×32 >75,000 oBC RNAs per cell per integrated copy of the cassette in the genome, which converted to concentration assuming a radius of 10 um for K562 cells, leads to ≈30 μM.

## 4 Assessing the influence of cHS4 and U6/oBC on genomic integration positional effects

### 4.1 Experiment description and statistical rationale

We sought to quantify whether the Pol III promoter cassette (U6 promoter driven Tornado barcode) affected the sensitivity of the Pol II expression to genomic integration positions in our random transgenesis approach, especially in comparison to a canonical insulator element (cHS4, present in our original reporter architecture). To do so, we constructed a series of four different piggyBac transposon backbone with the four possible combinations of +/- insulators and +/- U6/oBC. The five exogenous promoters (**Fig. 2a**) were then cloned to drive the expression of a barcoded mRNA within the context of all four different reporter architectures (**Supp. Fig. 1a**).

We reasoned that starting from a complex library of redundantly barcoded reporters and following integration of the library in a polyclonal population, bottlenecking of the population such that the number of total integrations ($N_{integrations}$ := number of clones × average number of integration per clones) was much smaller than the total number of barcodes in the starting plasmid library (defined as $N_{mBC}$) would ensure that each barcode would be associated to a probabilistically unique genomic integration position. Measuring the expression level of these reporters using conventional bulk MPRA and assessing variability in the expression across barcodes would then provide a measure of variation coming from positional effects for the different reporter architecture.

Statistically, the likelihood of having a barcode integrated at multiple positions can be estimated. Assuming equal representation of the barcodes in the library (best case scenario), the distribution of number of integration in the population of any specific barcode follows a Bernoulli process with trial number equal to the total number of integrations $N_{integrations}$, and success probability equal to $1/N_{mBC}$. The probability to have more than one integration for any specific barcode is then:

$$1 - \left(1 - \frac{1}{N_{mBC}}\right)^{N_{integrations}} - \frac{N_{integrations}}{N_{mBC}}\left(1 - \frac{1}{N_{mBC}}\right)^{N_{integrations}-1} \approx \frac{1}{2}\left(\frac{N_{integrations}}{N_{mBC}}\right)^2 \; if \; N_{integrations} \ll N_{mBC}$$

Hence, if the sampled barcode complexity in a bottlenecked population is much less than that of the full library, each detected barcode will indeed be probabilistically uniquely integrated. In the context of bulk MPRA (without positional mapping), the number of independent integrations is not directly measurable. Instead, what is observable is the number of detected barcodes from the library, which equals the number of different unique barcodes sampled (above a representation threshold). In the above approximate Bernoulli process, the correction is proportional to the fraction of barcodes sampled twice or more (above), and is small in the considered limit, and the proportion of detected barcodes is then to leading order $N_{integrations}\ (N_{mBC})^{-1}$. The fraction of detected barcodes that are not uniquely integrated can then be estimated as the proportion of barcodes with more than one integration (above) over the proportion of detected barcodes, or approximately $N_{integrations}\ (2\ N_{mBC})^{-1}$.

## 4.2 Construction of reporter libraries with/without cHS4 and U6/oBC

### *4.2.1 Cloning of reporters*

The 20 promoter libraries (five promoters × four reporter architectures) were cloned separately and pooled prior to transfection. Briefly, in order to construct the libraries, four high complexity

(large number of mBCs in library) barcoded backbones were used as starting points: p22 (with insulators, without U6/oBC), p25 (with insulators, with U6/oBC), p93 (without insulators, without U6/oBC), and p94 (without insulators, with U6/oBC) (**Supp. Fig. 1a**). The re-cloned high complexity p25 backbone used to construct scQer libraries was used directly for this purpose. A high-complexity p22 library was generated by a three fragment Gibson assembly, each obtained from PCR of the original p22 (backbone fragment1 with primers oJBL524+oJBL527; backbone fragment2 with primers oJBL526+oJBL528; barcoded insert fragments obtained with two rounds of PCR with primers oJBL514+oJBL518 followed by oJBL516+oJBL518; all fragments were size selected on agarose for backbone fragments and PAGE for barcoded insert), and electroporated in in 25 uL of C3020 cells (NEB). p93 and p94 were obtained by Esp3I digest of the insulator free piggyBac transposon plasmid pXL005, followed by addition of barcoded inserts with Gibson assembly (obtained from PCR with primers oJBL679+oJBL680 respectively from p22 and p25) and electroporation in 25 uL of C3020 cells (NEB). All backbone were at high complexity (>1M transformants as estimated by plating a ) and were confirmed by Sanger sequencing. We note that the barcode sets of p22/p93 and p25/p94 overlap as a result of the cloning strategy, but that given bottlenecking of the final libraries, the number of colliding barcodes was minimal.

From these barcoded backbones, ectopic promoters were introduced (separately for each backbone/promoter pair) as described before, with BglII+EcoRI digest and Gibson with PCR products containing the different promoters with homologies to the backbone (see section 1.1 for details). The resulting plasmids were transformed individually in 25 uL of C3040 cells (NEB), *de facto* leading to a bottleneck (estimated complexity of transformant ~500 to 20k depending on the library). Cells were outgrown overnight and plasmids purified by midiprep (Zymo).

*4.2.2 Association of mBC to promoters*

To obtain the list of mBC corresponding to each plasmid library, we prepared amplicons for sequencing similarly to the DNA arm of bulk MPRA, as described before. Starting from 5 ng of each plasmid library, two rounds of PCR were performed (4 cycles PCR1 with primers oJBL039+oJBL753 to append a pseudo 10 bp UMI, followed by 8 cycles with oJBL361 and Nextera v2 P7 indexed primers). Amplicons were cleaned up with Ampure XP beads (1x), and pooled for sequencing. Sequencing was performed on a Nextseq2000 using custom set of primers (read 1: oJBL369, 15 cycles to read mBC; index 1: oJBL335 [Nextera index 1], 10 cycles to read the sample index; read 2: oJBL494 [Nextera read 2], 10 cycles to read the pseudo-UMI; index 2: oJBL371, 15 cycles to read the reverse complement of the mBC). The data was processed to a piled up file (counting number of reads and UMI per barcode per library) as described before. The count distributions per barcode per library were inspected and found to be bimodal. The *bona fide* barcodes present in the libraries were taken to be those in the high count mode (count threshold the minimum of the bimodal distribution), leading to 153.9k barcodes across the 20 libraries. To ensure no inter-library barcode collision (given that the libraries were pooled for the experiment), the final list of barcodes used was filtered to only have barcodes present in a single library out of the 20 pooled for the experiment, leading to a final set of 134.1k barcodes (13% multiply represented barcodes removed). Barcode complexity per library spanned 588 to 22.1k with interquartile range 3.6k to 11.3k.

## 4.3 Transfection and bottlenecking of cell population

Libraries were pooled in accordance with the estimated number of barcodes per construct to ensure similar representation of individual barcodes (i.e., pooled not equimolar per library, but equimolar per barcode). This was to ensure that barcodes of the less complex libraries remained predominantly uniquely integrated. Pooled libraries were transfected in K562 and HEK293 (6.5 M K562 cells with nucleofector, kit V4XC-2024 (Lonza BioResearch) program FF120, with 8 ug libraries pool [transposon] and 400 ng hyPBase transposase; 2M HEK293 cells, lipofectamine2000 with 8 ug libraries pool [transposon] and 400 ng hyPBase transposase). Cells were allowed to recover for five days and passaged upon confluence (HEK293) or reaching 1 M/mL (K562). Following five days, 2 ug/mL puromycin selection was applied for another 7 days to select for piggyBac-mediated genomic integration of reporter transposons (which contained the GFP-P2A-puromycin resistance ORF). After dilution of the unintegrated plasmids (12 days post transfection), polyclonal cell populations were seeded at an estimated starting number of clones of 3k, 10k, and 30k each in biological duplicates. After expansion of populations to approximately 5M cells (7 to 9 days), cells were harvested, fixed in 80% methanol, and stored at -80C until extraction.

## 4.4 Bulk MPRA experiment and data analysis

### *4.4.1 Bulk MPRA library construction*

RNA and DNA was extracted as before from methanol fixed cells (AllPrep kit, Qiagen). MPRA libraries were constructed as previously described, with slight modifications. Briefly, for RNA libraries, 5 ug of RNA was reverse transcribed with UMI containing primer oJBL753 using

SuperScript IV (10 uL 500 ng/uL RNA mixed with 2 uL 1 uM oJBL753; 5 min at 65C, ice for >2 min; followed by addition of 4 uL 5x buffer, 1 uL 0.1M DTT, 1 uL 10 mM dNTPs, 1 uL water, 1 uL SSIV as a master mix for 20 uL total reaction; 55C for 60 min, 80C for 10 min). The cDNA was taken directly as template for PCR1 (25 uL reactions, 10 uL cDNA, primers oJBL039+Nextera v2 P7 indexed primers; 4 elongation cycles). Following the first PCR, reactions were cleaned up with 1x Ampure XP beads, eluted in 12 uL 10 mM Tris 8. 4 uL of PCR1 eluates were taken to PCR2 (10 uL reactions, primers oJBL077+oJBL359, tracked with qPCR and SYBr green, stopped at cycles 9-15 depending on reaction's inflection point). Reactions were again purified with 1x Ampure XP and eluted in a final volume of 10 uL 10 mM Tris 8. For DNA, approximately 10 ug genomic DNA was amplified in 25 uL reactions (primers oJBL039+oJBL753; 4 cycles). Following 1x Ampure XP cleanup and elution in 12 uL 10 mM Tris 8, 4 uL of the eluate was amplified in a second round of PCR (10 uL reactions, 15 elongation cycles; primers oJBL360+Nextera v2 P7 indexed primers), cleaned up with 1x Ampure XP, and eluted in 10 uL 10 mM Tris 8. Each biological replicate was process in technical duplicate, for a total of 48 libraries (2 cell lines × 2 library types RNA/DNA × 3 bottlenecking factors × 2 biological replicates × 2 technical replicates). All reactions were performed in 96-well plates and processed at the same time as the bulk MPRA experiment from EBs to assess different aspects of reporter architecture (see section 5). To quantify final amplicon concentrations, a real-time qPCR experiment was performed using P5 and P7 primers (oJBL076, oJBL077 respectively). Samples were pooled according to their concentration as determined by qPCR and sequenced on a Nextseq2000 using custom set of primers (read 1: oJBL369, 15 cycles to read mBC; index 1: oJBL335 [Nextera index 1], 10 cycles to read the sample index; read 2: oJBL494 [Nextera read

2], 10 cycles to read the pseudo-UMI; index 2: oJBL371, 15 cycles to read the reverse complement of the mBC).

*4.4.2 Bulk MPRA analysis*

Bulk MPRA data was processed as previously described. Libraries were sequenced at an average 2M reads/library (interquartile range 1.5M to 2.75M). A threshold of 10 DNA UMI per barcode was selected to deem a barcode as present in the cell populations. For the majority of conditions, the set of barcodes detected in cells were indeed only a fraction of those present in the libraries (median recovery across 20 libraries: HEK293_03k: 6.0%, HEK293_10k: 10.4, HEK293_30k: 17.8%, K562_03k: 13.1%, K562_10k: 24.8%, K562_30k: 31.0%), as necessary for statistical assumptions of probabilistically unique integrations (see section above). We confirmed that the variability measure was stable over the different populations (except for Pgkp promoters in K562, which showed evidence of lower variability at higher starting population). Within all samples (bottlenecked populations), for each barcode detected (DNA UMI>10), the activity was computed as the normalised RNA UMI (RNA UMI count over the summed RNA UMI count for the sample) over the normalised DNA UMI. Example activity score distributions from active promoters for replicate HEK293 bottlenecked at 30k (rep1.1) are shown in **Supp. Fig. 1b**. Variability across activity scores measured from the mBCs is quantified as the ratio between the 75th and 25th percentile (shown for all replicates in **Supp. Fig. 1c** for HEK293, and **Supp. Fig. 1d** for K562). A two-sided Wilcoxon test (Bonferonni correction) was used to compare this metric of variability between the following pairs of reporter architecture for all active promoters (UBCp, Pgk1p, EEF1A1p): (U6/oBC+, cHS4+) vs. (U6/oBC-, cHS4+), (U6/oBC+, cHS4+) vs. (U6/oBC+, cHS4-

), (U6/oBC-, cHS4+) vs. (U6/oBC-, cHS4-), and (U6/oBC+, cHS4-) vs. (U6/oBC-, cHS4-) (**Supp. Fig. 1c**).

## 5 Control experiments assessing impact of reporter architecture (bulk MPRA in mEBs)

In order to test the influence of various components of the scQer architecture, we constructed a series of libraries harbouring different components/positions of the regulatory elements (summarised in **Supp. Fig. 8a**). We then directly compared the activity of the CREs within the different reporter architecture using bulk MPRA across mouse embryoid body differentiation series.

### 5.1 Construction of reporter libraries with different architectures

#### *5.1.1 Cloning of reporter libraries*

Given that some reporter architectures lacked oBC (to explicitly test the influence of the Pol III cassette), which we had previously used for barcode dictionary generation, we had to adapt our cloning strategy to be able to directly subassemble the CREs to the mBC. We describe in turn the different approaches taken.

Libraries p058 & p059 (with vs. without oBC): to clone the standard scQer cassette with and without the Pol III cassette, we started from the high complexity barcoded plasmid docks p022 (no oBC, mBC only) and p025 (with both oBC and mBC). In contrast to the early strategy (highlighted in **Supp. Fig. 2**), we first inserted the library of CREs (same inner PCR pool as previously) using Gibson assembly from the NheI and MfeI digested barcoded backbone, leading to intermediate

plasmids p056 (no oBC) and p057 (with oBC) respectively. These libraries were bottlenecked to 1% post transformation (electroporation) to be in the appropriate complexity range (~100 mBCs per CRE). p056 and p057 served as template for subassembly (see next section for details). To complete the library, we then integrated the minP-GFP cassette. As another difference from the previous approach (linearization of backbone with EcoRI+BglII), to avoid losing a substantial fraction of our CRE due to these frequent cutters, we used spCas9 (NEB) to perform *in vitro* digestion at a specific site on the plasmid using a specifically designed crRNA (roJBL677). Following duplex formation with the tracrRNA (Alt-R® CRISPR-Cas9 tracrRNA, IDT) (5 uL 100 uM of each crRNA and tracrRNA, 95C for 5 min, then cool to room temperature), 2 ug of p056 and p057 plasmids were digested (150 uL reactions: 15 uL r3.1 buffer, 5 uL spCas9, 15 uL 300 nM crRNA:tracrRNA duplex, 100 uL nuclease free water, incubate 10 min at 25C, addition of plasmid, incubate 15 min at 37C, addition of 5 uL 100 mg/uL ProK, incubate 10 min at room temperature, 0.5x Ampure XP clean up and elution in 20 uL prior to run on gel) and size selected on agarose gel. The minP-GFP insert with compatible homologies for Gibson assembly was obtained from PCR with primers oJBL254+oJBL676, and assembled with the p056 and p057 libraries (serving as template for CRE-mBC subassembly, see below), and electroporated in a high efficiency transformation to maintain complexity. The resulting libraries, p058 (no oBC) and p059 (with oBC), were purified (midiprep, Zymo) and mixed in as parts of the final transfected pools of the bulk MPRA experiments. We note that library p059 effectively consisted in a very similar construct compared to the original p055 (only difference: a short 41 bp added sequence [containing the original BglII+EcoRI insertion site] between minP and the CRE due the different integration strategy for p058), and was used to assess internally the reproducibility of our bulk MPRA measurements.

Library p092 (no minP): to construct scQer reporters without the minimal promoter, we first inserted a no promoter GFP cassette into high oBC-mBC complexity library p025 (Gibson assembly, backbone: BglII+EcoRI selected on agarose, insert: noP-GFP constructed by fusion PCR [primers oJBL254+oJBL314, final product size selected on agarose] with two shorter fragments, themselves obtained by amplifying p029 with primers oJBL314+oJBL417, oJBL254+oJBL414). The resulting library (obtained from high efficiency electroporation), p045, was then digested with MfeI and NheI (NEB) and the CRE pool (same inner PCR pool as before) inserted by Gibson assembly, electroporated (no bottleneck was implemented, instead diluting the electro-competent cells in 10% glycerol 10-fold prior to electroporation). The resulting plasmid library p092, was used for subassembly between CRE and mBC (via the oBC as before), and for the final pool transfected into cells.

Libraries p091 and p096 (CRE downstream, without and with oBC): to create constructs in which the CREs were positioned downstream of the reporter cassette and promoter, we first introduced the minP-GFP fragment in high complexity plasmids p022 and p025 as before (Gibson; backbone: BglII+EcoRI digest with agarose size selection, insert minP-GFP as previously; electroporation). The resulting high complexity libraries, respectively p090 and p095, were then digested using new sites which by happenstance happened to be paired unique cutters between the SV40 poly-A signal and the downstream cHS4 insulator: EcoNI+BtsXI (NEB). To append compatible homology handles to this new portion of the backbone, the inner PCR pool of CREs was re-amplified in a low-cycle PCR reaction with primers oJBL674+oJBL675, and following 1x Ampure XP clean up, inserted by Gibson assembly. The resulting high complexity libraries, p091 (no oBC, CRE

downstream) and p096 (with oBC, CRE downstream), were each bottlenecked to an estimated 50k clones, and plasmid libraries were purified and served as template for CRE-mBC subassembly (new approach, see below), and used directly for transfection into cells as part of final pools.

Cloning of the promoter series without oBC libraries (p033, p034, p035, p039, p040: exogenous promoters without oBC cassette) proceeded exactly as described in a previous section (different pool).

*5.1.2 mBC to CRE subassembly strategies*

As alluded to above, we took different strategies for the different reporter architectures to map mBC to CREs. We describe below the various strategies used for the different cassettes.

Subassembly of libraries p058 & p059: the connection between mBC and CRE was obtained using a similar strategy as that connecting oBC to CREs. Briefly, the plasmid library was tagmented as described before (section 2.4.4). Instead of using primers upstream of the oBC, primers downstream of the mBC were used. Following 13 cycles of semi-specific PCR (primers: indexed Nextera P5 + oJBL358). Fragments in the range 350-700 bp were size selected on PAGE, and paired-end sequenced on a Nextseq500 (read1 32 cycles Nextera_read1 primer: tagmented CRE; read 2 32 cycles primer oJBL371: mBC; index 2 10 cycles Nextera_index2 primer: sample index). Bioinformatic processing of the data was similar as for oBC-CRE subassembly, yielding 15.5k mBC with median 65 mBC/CRE for p058 and 27.8k mBCs with median 117 mBC/CRE for p059

passing the individual library controls (see below for additional filter to avoid inter-library collisions in the pool).

Subassembly of library p092: given the architecture of the reporter was similar to the original p055 (except without the minimal promoter), we used the same strategy (oBC-CRE) as previously described obtaining 49.4k valid mBCs (via the original oBC-mBC pairing) and a median of 207 mBCs/CRE.

Subassembly of libraries p091 & p096: since the CREs were inserted downstream of the GFP reporter, the CREs could be directly subassembled to the mBCs in this context. To do so, we again used tagmentation followed by 13 cycles of semi-specific PCR (primers: indexed Nextera P7 + oJBL708) and a PAGE size selection (600 bp to 900 bp). The library was paired-end sequenced on a Nextseq500 (read1 32 cycles oJBL707 primer: start of inserted CRE; index1 18 cycles Nextera_index1 primer: sample index, read 2 32 cycles primer oJBL371: mBC). Data processing for subassembly was similar as for the oBC-CRE mapping in p055, and leading to identification of 32.0k and 47.3k valid mBCs, and a median of 162 and 241 mBCs/CRE for libraries pJBL091 and pJBL096 respectively.

Subassembly of promoters without oBC (p033 series) was performed by obtaining the list of mBC (PCR amplification and sequencing of the product) from sequencing of PCR products (two steps: PCR1 with Kapa HiFi in 20 uL, 4 cycles, primers oJBL039+oJBL358, Ampure 1x cleanup, PCR2 with Kapa HiFi in 20 uL, primers oJBL077+oJBL362-o366 indexed series for 10 cycles). The

product was sequenced as a spike-in on Nextseq 2000 with custom primers (read 1: 148 cycles, primer oJBL369; index 1: 10 cycles, empty [blank read]; index2: 10 cycles, primer oJBL370). Read 1 was sufficiently long to cover both the mBC and the pseudo-UMI installed by PCR1 of the library preparation. Read1 fastq was then trimmed to separate files for the two pieces (mBC and UMI) of information for downstream processing. mBC and UMI where the piled-up as for MPRA amplicons. The bona fide set of mBC present in each respective libraries was then determined from the high-count mode of the bimodal distribution of UMI count per mBC, yielding a median of 2.8k mBCs per promoter (span of 0.3k to 4.3k mBC for the different promoter, with UBCp being less well represented).

Given that all the libraries cloned were generated as bottlenecked versions of starting high complexity plasmids (p022 and p025), mBCs shared between the final pooled libraries were expected to (rarely) occur. In order to obtain the final list of unique and valid CRE-mBC pairs, we therefore excluded from the final table, any mBC that was present across multiple libraries pooled for the experiment (see next section), as these would have been impossible to interpret (if associated with multiple CREs).

We provide here additional information on quality control checks that were performed as part of these subassemblies. First, oBC-CRE subassemblies were performed on libraries p057 and p059 (only difference being the addition of minP-GFP between the CRE and mBC in going from p057 and p059), to assess possible loss in library complexity in the cloning step. We found an identified oBC overlap of >98% between the two libraries, and an $R^2$ of the log-transformed oBC counts of 0.91, both arguing in limited loss in complexity in the step of adding the reporter in the library

backbone. Second, as a way to test the faithfulness of our triplet dictionary, we performed oBC-CRE and CRE-mBC subassembly on library p057 (feasible because the reporter is not yet integrated as described above). We found 99.5% concordant associated CREs (fraction of agreement in associated CREs between detected predetermined valid oBC-mBC pairs).

5.2 Transfection of final pooled libraries and mEB induction

Two pools of libraries were generated for the purpose of experiments, with the following final compositions (by mass). Pool A: 30% p058, 17.5% p059, 30% p092, 17.5% p055 (original library from the first round of experiments), 2.5% p27 series (original exogenous promoter pool), 2.5% p33 series (exogenous promoters without oBC cassette, itself equal mass pool of p033, p034, p035, p039, p040). Pool B: 37.5% p091, 37.5% p096, 20% p055 (original), 2.5% p27 series (original), 2.5% p33 series.

The two respective pools were transfected (lipofectamine 2000) in mESCs as previously described, except each reaction scaled up by 2-fold. Specifically, each pool was transfected in 6 separate reactions (2M cells transfected and plated in 6 cm plates) with 400 ng hyPBase plasmid, 8 ug of reporter pool. Following recovery of cells for 2 days, cells from pairs of transfection replicates from each pool were combined (from 6 to 3 plates of cells), and the resulting three set of cells were hereafter maintained separately, constituting our biological triplicates. Puromycin (2 ug/mL) selective pressure was applied 3 days post transfection and until induction of mEBs, which happened 11 days post transfection. mEB induction proceeded as previously described, with 15M of cells split in 5 plates (3M/plate) for each biological replicate. 3M of cells per replicate were also

sampled for the day 0 time point. Plates from the same replicates were mixed at each medium change (once every two days). One plate's worth of mEB was harvested on days 4, 18, 20, and 22. At harvest, cells/mEBs were fixed in 80% methanol and placed at -80C until RNA/DNA extraction for library preparation.

5.3 Bulk MPRA experiment and analysis

RNA/DNA extraction, MPRA library preparation (with the updated amplicon design performed in technical duplicate for each sample), sequencing, preprocessing and aggregation to count tables proceeded as described in section 4.4, with the only modification that the threshold for including mBCs in the derived MPRA activity was taken to be >3 DNA reads per barcode. The per-CRE activity derived from MPRA counts was taken as the summed normalised RNA UMI over summed normalised DNA UMI from all mBCs corresponding to a CRE (from the subassembly). The activity score for each CRE was calculated separately for the different libraries (reporter architectures) for each sample (corresponding to sample time, biological/technical replicate). The displayed scores (**Supp. Fig. 8b-c**, **f**, and **h**) corresponded to the median across replicates, and the error bars mark the interquartile range (25th to 75th percentiles).

## 6 Singleton validation experiments

In order to obtain orthogonal (not relying on single-cell genomics) evidence of the autonomous & cell-type-specific activity of the CREs identified in our initial screen (experiment of **Fig. 3-4**), we cloned active elements individually. These constructs were then respectively transfected and genome-integrated in separate cultures of mESCs, following which each culture was differentiated to embryoid bodies (as described above) over >3 weeks. Epifluorescence pictures were taken throughout the time course and with structured illumination (end point) to assess the resulting domains of reporter expression within mEBs.

### 6.1 Cloning of singleton mCherry scQer reporters

To allow for co-transfection and selection on puromycin with the promoter series (reporter ORF: puromycin-P2A-GFP), we constructed a scQer reporter with mCherry instead of GFP. Briefly, barcoded backbone p025 was digested with EcoRI and BglII and size selected on agarose. The minP-mCherry cassette was generated by splice PCR from two PAGE size selected PCR fragments (fragment minP: primers oJBL314+oJBL416, template p027; fragment mCherry: primers oJBL254+oJBL414, template p060) using primers oJBL254+oJBL314. The resulting minP-mCherry insert was size selected on PAGE, inserted by Gibson assembly into the digested p025 backbone to generate plasmid p062, and the library electroporated in *E. coli* as before (NEB, C3020). Plasmid library p062 was digested with MfeI and NheI and size selected on agarose. The resulting linear backbone was compatible with the inner PCR products used to clone the initial library of CREs (see section 2.4.3), and the products for the 8 most specific CREs were separately inserted by Gibson assembly to generate 8 distinct plasmid libraries (p065 *Lama1*:chr17_7784, p066 *Lamb1*:chr12_2183, p067 *Foxa2*:chr2_13858, p068 *Gata4*:chr14_5729, p069

*Sox2*:chr3_2007, p070 *Sox2*:chr3_2009, p071 *Bend5*:chr4_8201, p072 *Epas1*:chr17_10063). Assembled plasmids were transformed into *E. coli* (NEB, C3040). The singleton plasmids were separately purified for transfection into cells. The number of unique barcode pairs per construct was estimated to be around 100-400 (by counting plated colonies). In this case, we however did not sub-assemble the barcode dictionary as a sequencing-based readout was not used for these experiment. All plasmids were verified by Sanger sequencing.

6.2 Singleton mEB differentiation experiment

We transfected the plasmids in mESC grown as before with lipofectamine 2000 individually in separate cultures (0.5 M cells in 1 well of 12-well plate; 1 ug scQer mCherry singleton plasmid, 100 ng promoter puromycin-GFP series plasmid pool [from **Fig. 1**], 100 ng hyPBase plasmid). Following cell recovery, puromycin selection (2 ug/mL) was applied (day 3) until EB induction (day 10). On EB induction day, cells from the single-cell suspension were profiled by FACS to quantitatively measure mCherry expression in the pluripotent state (**Supp. Fig. 4b**). Two plates of mEBs (3 M/plate) per singleton construct were initiated as described before. On every medium change (every two days), epifluorescence images were taken (examples in **Supp. Fig. 4c**) from each culture to assess mCherry expression. While the *Sox2* elements both showed high activity at day 0, consistent with their activity in the pluripotent state, both in FACS and from epifluorescence (**Supp. Fig. 4b-c**), mEBs harbouring parietal endoderm elements initially displayed essentially no expression above background, with a fraction of them (for all elements) inducing expression over the time course (**Supp. Fig. 4c**). Domains of expression between the Sox2 (pluripotent-specific) elements sharply contrasted (internal, spotted) with that of parietal-specific elements (all on surface, **Supp. Fig. 4d-m** and **Supp. Fig. 5**). These observed domains of mCherry expression were

in line with early observations in embryoid bodies reporting endodermal cells on the surface with a rough morphology[49] indicative of basement membrane component production, and similar to spatial patterns of expression directly observed for *Gata4*[134].

## 7 Example applications of scQers: CRE pairs and TF binding sites allelic series

To illustrate the usefulness of scQers to study questions in regulatory genomics, we constructed three new libraries (literature-selected, paired CREs, perturbed TFBS). These were then profiled as before using scQers (integration at high MOI with piggyBac in mESC, differentiation to mEBs over 3 weeks and single-cell endpoint profiling).

### 7.1 Identification of putative transcription factor binding sites

We hypothesised that putative transcription factor binding sites within the identified cell-type-specific CREs would be important sequence features to perturb and lead to large changes in activity.

#### *7.1.1 Putative Transcription factor binding sites identification*

To characterise the transcription factor binding composition of tested elements using a biophysically grounded empirical approach (in the absence of high resolution ChIP-seq data in our system), we took an approach inspired by Farley and colleagues[27,115]. Briefly, we obtained protein array binding data from endodermal transcription factors *Gata4*, *Foxa2*, and *Sox17* from Uniprobe[104–106], which provides affinity measures for all DNA 8-mers. We converted the raw measurements ("Median" column in the raw data files) to relative affinities. To do so, we treated

the mode of affinities as the experimental noise floor, and computed the relative affinity as: (affinity-baseline)(max(affinity)-baseline). We note that the final list contains a relative affinity for an 8-mer and its reverse complement (as the protein binding arrays hold double stranded DNA). Before computing the maximum affinity, we divided the score of palindromic 8-mers by two, as we found those to be anomalously high. The resulting relative affinity 8-mer table was then used to scan all regulatory elements and genomic regions, yielding a value for each 8-bp stretch (in a strand agnostic manner). We then identified local maxima in the relative affinity trace. For a given relative affinity threshold, the local maxima above threshold were retained. To collapse maxima close to each other, we generated a graph between maxima (one node per maximum) with an adjacency matrix determined by distance (connect maxima less than 3 bp apart). Connected components of the graphs were identified, with one putative TF binding site assigned per connected component (typically a single maximum) at the highest-affinity position. The procedure was applied across a range of affinity thresholds, and both on individual CREs (e.g., **Ext. Data Fig. 10** and **Supp. Fig. 7d**) across full genomic loci (±100 kb from TSS, 500 bp windows with 250 bp sliding step and excluding tested CREs and surrounding 500 bp, **Supp. Fig. 7c**).

*7.1.3 TF binding site optimization and disruption*

To generate CRE variants to test (**Ext. Data Fig. 10d**), we identified putative binding sites for transcription factors *Gata4* and *Sox17* as described above (relative affinity threshold = 0.3) within 6 parietal endoderm CREs. Then, for each class of perturbation (optimization/disruption) and target TF (*Gata4*, *Sox17*, or both), we cycled through the corresponding putative sites. For disruption, the set of 8-mer within a Hamming distance of 2 of the binding site were identified, and the nearby 8-mer with lowest affinity was selected. For optimization, to comprehensively

identify the local maximum in sequence, we searched for the highest affinity 8-mer within Hamming distance of 2 of the seven 8-mers within ±3 bp of the local maximum (buffer distance = -3, -2, -1, 0, 1, 2, 3 bp), replacing the optimal 8-mer in its original position within the sequence. For perturbations to both *Gata4* and *Sox17*, we first mutated the *Sox17* sites, and then the *Gata4* sites. For simplicity, we did not prevent optimization/disruption mutations that would affect the other TF (e.g., not preventing mutations to Gata4 sites overlapping with Sox17 sites and vice versa). An example of the affinity traces for original and mutated variants are shown in **Supp. Fig. 7d**. Variant CRE sequences and information on mutations can be found in **Supp. Data 8**.

## 7.2 Cloning of scQer libraries

### *7.2.1 Literature selected elements*

In order to provide additional tests for the ability of scQers to detect cell-type-specific activity, we search for additional CREss with evidence for function in pluripotent or differentiated mouse stem cells. We selected two regulatory elements from Buecker et al[64] (*Tbx3* and *Esrrb* pluripotent-specific), two *Nodal* CREs from Papanayotou et al[65], one neural *Sox2* element[67], and one *Cdx2* intronic CRE[66]. These were PCR cloned as previously (unburdened outer PCR followed by inner PCR with cloning handles, oJBL632-641 and oJBL644-657; see **Supp. Data 4** for description of primers and the sequences and genomic coordinates of CREs tested). Cloning handles were compatible for Gibson assembly with p043 (digested with MfeI+NheI), were integrated as a pool as before, electroporated in *E. coli* (NEB, C3020). The resulting library was bottlenecked to ~20k clones, and the plasmid purified for subassembly and transfection in mESCs.

*7.2.2 Paired CREs combinatorial assembly*

To assemble pairs of CREs on the same construct, we used the common handles (homology arms used to clone in the original barcoded backbone p043) as primer binding sites to append new homology arms compatible with pairwise combinatorial assembly. Briefly, a new barcoded backbone with updated homology arms for integrating CREs was created. p043 was digested with HindIII and NcoI, size selected on agarose and a new replacement insert with updated homology arms, gene block gJBL009, was integrated by Gibson assembly. The resulting library, p063, was electroporated (NEB, C3020) and expanded while maintaining complexity. Plasmid library p063 was the same as p043, but now with new homology arms: upstream 5' AGGACTCTACCAACGCTAGTCCAAGCAAGG and downstream 5' gtgcagcgcgatgtatagcagtgcgcgaag. For each of the six selected CREs (pluripotent specific: P1 *Sox2*:chr3_2007, P2 *Sox2*:chr3_2009; parietal endoderm specific: E1 *Epas1*:chr17_10063, E2 *Gata4*:chr14_5729; inactive: I1 *Cdk5r1*:chr11_12590, I2 *Col5a1*:chr2_2586), four PCR products were generated, corresponding to targeted order and orientation in the final assembly (oJBL578+oJBL579 upstream forward Uf, oJBL580+oJBL581 upstream reverse Ur, oJBL584+oJBL585 downstream forward Df, oJBL582+oJBL583 downstream reverse Dr), using the previous inner PCR products as template (all with same common handle). The 8 above primers were designed to either have one homology to the backbone, or one homology to a central joining region (to pair two CREs, central junction sequence: 5'TGACGAAGCTATACTCGGTCGCGAGGACGT), such that the resulting PCR products would be assembled obligately as a pair. The 24 different products were size selected on agarose. Two Gibson assemblies were performed for combinatorial assembly with the MfeI & NheI digested updated backbone p063 described above. With the shorthand notation just introduced,

assembly 1 pluripotent upstream/endoderm downstream: [P1_Uf, P1_Ur, P2_Uf, P2_Ur, I1_Uf, I1_Ur] assembled with [E1_Df, E1_Dr, E2_Df, E2_Dr, I2_Df, I2_Dr]. Assembly 2 endoderm upstream/pluripotent downstream: [E1_Uf, E1_Ur, E2_Uf, E2_Ur, I2_Uf, I2_Ur] assembled with [P1_Df, P1_Dr, P2_Df, P2_Dr, I1_Df, I1_Dr]. All 72 (9 possible CRE pairs × 2 orders × 4 relative orientations) possible combinations of order and orientation of CREs would then be represented across these two pools, for example E1_Uf::P2_Dr, etc. An example annotated plasmid with combination P1_Ur::E1_Dr is included. The two libraries were each bottlenecked to ~20k and ~1k constructs, and purified for subassembly and transfection.

### *7.2.3 CREs with mutated putative transcription factor binding sites*

The mutated CREs with optimised and disrupted putative binding sites to *Gata4* and *Sox17* (see above) were synthesised (eBlocks, IDT) with the same homology handles for Gibson assembly in the barcoded backbone (p043) as the original cloning strategy. The eBlock CREs were pooled at equimolar ratio and inserted in p043 digested with MfeI & NheI as before. The resulting library was electroporated in *E. coli* (NEB, C3020), bottlenecked to about 75k clones, and purified for barcode subassembly and transfection.

## 7.3 Generation of barcode to CRE dictionaries

Subassembly of the CRE to barcodes for the literature selected elements was performed as before (tagmentation with semi-specific PCR, size selection on PAGE, sequencing of oBC & tagmented CRE on a paired-end Illumina run). On the other hand, given the nature of the libraries, the paired CREs and mutated regulatory elements required slightly different approaches, as detailed below.

*7.3.1 Connecting pairs of CREs to barcodes with long read (Nanopore) data*

The minimum distance between the oBC and the second CRE in a construct with two ~1kb size CRE is about 1.4 kb (shorter if the distal CRE is less than 1 kb). Based on quantification of clustering efficiency (~100-fold lower for amplicons of 1.5 kb) on patterned flow cells[135], we attempted to perform the subassembly in the same way as for previous scQer constructs, except size-selecting a larger region on the gel to try to capture the CRE junction and obtain information about the oBC and the two CREs in a single paired-end read (with the oBC read as one of the indexed reads). However, despite selecting for sizes exceeding the first CRE in principle, the recovered reads predominantly were tagmented in the first CRE, underscoring the severe length bias at these amplicon sizes (1-3-1.5 kb) on the short read platforms. While coverage was insufficient to map the oBC-CRE1-CRE2, we obtained enough reads per oBCs in the library to be able to error-correct the slightly error prone and shallowly covered long-read data (see below).

To circumvent the size limitation of the short-read platform, we sequenced the four libraries (pluripotent up::endoderm down high complexity, pluripotent up::endoderm down low complexity, endoderm up::pluripotent down high complexity, endoderm up::pluripotent down low complexity) on a MINion flow cell of the Nanopore GridION platform, using sparse tagmentation to linearize the plasmid libraries prior to adapter ligation (Rapid Sequencing Kit, Nanopore).
The nanopore data was processed with custom scripts with the following set of heuristics. First, constant 30 bp 'signposts' sequences flanking variable regions were selected to establish as reference positions within the nanopore read:

CRE upstream: 5' TGGCGAGGACTCTACCAACGCTAGTCCAAG,

CRE junction: 5’ TGACGAAGCTATACTCGGTCGCGAGGACGT,

CRE downstream: 5’ GTGCAGCGCGATGTATAGCAGTGCGCGAAG,

mBC upstream: 5’ CGAGCTGTACAAGTGAACGCGTTAAGTCGA,

mBC downstream: 5’ TCGACAAGCTCACCTATTAGCGGCTAAGGC.

We then performed local alignments with the Smith-Waterman algorithm (leveraging a fast implementation[136]) using the above signpost sequences as query, and the nanopore reads as targets. We retained reads with at least 4 of 5 signpost alignments with score ≥50 (≥80% match). Further, only reads with sequences with roughly correct sizes were kept (distance [start to start] in read between CRE upstream and CRE junction, and CRE junction and CRE downstream >500 and <1500, and distance between mBC upstream and mBC downstream >42 and <52). The sequences of the regions intervening between signposts (CRE upstream, CRE downstream, mBC) were then extracted from the reads. The extracted CRE sequences were then aligned by Smith-Waterman (again with the fast implementation) to the expected set cloned in the library as query, and the identity (with the relative orientation) of the maximum scoring CRE was stored (alignment score ≥750 [threshold determined from the distribution of scores], otherwise treated as unmapped). Finally a pile-up table counting the number of reads supporting any given combination of oriented CRE upstream/downstream and mBC was compiled and served as starting point for error correction from the Illumina sequencing (this step was useful given that the libraries were not fully saturated from the Nanopore data and each construct not supported by multiple reads, especially for the high complexity libraries).

We leveraged the high coverage of the oBC from our attempt at subassembly from Illumina data to error-correct the Nanopore assembled CREup-CREdown-mBC triplets. Briefly, high read count

oBC-CREup pairs (libraries saturated, threshold coverage determined from bimodal distribution of counts to select the high count mode) was determined from the Illumina run, and the associated set of mBCs determined from the pre-determined pairs obtained from the starting oBC-mBC p025 library. Then, for each mBC from the Nanopore pile-up file, the Levenshtein distance to all mBC in the Illumina-identified set was determined, and the minimum distance mBC was stored if unique and within a distance of 2 at most. If the oriented upstream CRE (oBC proximal) identified by the Illumina subassembly matched the oriented upstream CRE from the Nanopore data, the error corrected mBC was considered valid. A Nanopore read count threshold was further applied for the low complexity libraries (>1 read for pluripotent up::endoderm down low complexity, >4 reads for endoderm up::pluripotent down low complexity). Following removal of non-duplicated mBC sets and non-unique oBC-mBC pairs, the oBC-CREup-CREdown-mBC tables were saved for downstream analysis.

*7.3.2 CRE with allelic series of perturbed transcription factor binding sites*

Associating oBC to CRE was also more challenging for the library of variants because of the similarity between the different sequences, requiring overall better coverage (not only in terms of read counts, but also positionally across the full element) and a refined computational strategy. Following an analogous approach as before for oBC-CRE association (tagmentation & semi-specific PCR), we however size selected three different ranges on the PAGE gel (360 to 600 bp, 600 to 900 bp, and 900 to 1.4 kb). Following separate purifications, the different size ranges were then re-pooled prior to sequencing with a preference towards the longer products (ratio 1:3:9 for short:mid:long) to mitigate the size bias for bridge amplification on the Illumina flow cell and therefore get more uniform coverage of tagmentation event along the variant CREs, helping with

sequencing single-nucleotide variants introduced by design to optimise/disrupt putative TF binding sites.

Following sequencing, we applied the following computational strategy to map oBC to variant CRE. First, following alignment (to the set of unperturbed WT CRE as query) of the tagmentation-based read (in CRE), a pile-up table compiling for each pair of detected oBC and associated WT CRE (prior to variant call) the read count and set of all identified mutations within the reads (with associated read counts per mutation). Only oBC-(WT CRE) with sufficient coverage (empirically set to 90 reads) were retained. From a list of mutations per CRE in the allelic series (see **Supp. Data 8**) and the pile-up table, the following metrics were calculated for every oBC-(WT CRE) pair: total read coverage, mean read coverage per in-variant mutation (for each possible variant), total number in-variant mutations covered by reads, total number of out-variant mutations covered by reads. These metrics were finally used to classify each oBC-(WT CRE) pair to its most likely CRE variant as follows. Non perturbed CREs were identified as sequences with low normalised read coverage per mutation (maximum [across all possible variants] mean in-variant mutation coverage divided by total read counts), with threshold set by the bimodality of the distribution. In cases where the maximum [across variant types] total in-variant mutation read counts was <1.5-fold the second maximum, the variant type with highest per-mutation mean coverage was maintained. In cases where the maximum was >1.5 the second maximum, the corresponding maximum variant was retained. We note that these criteria were selected on the basis of certain mutation sets per CRE being strict subset of larger sets (e.g., *Sox17*-high $\subset$ *Gata4*-*Sox17*-high), further complicating the variant call. oBC-(WT CRE) pairs with more out-variant mutations than in-variant mutations were excluded as too heavily mutated and thus unusable.

*7.3.3 Final barcode dictionary for pooled experiment*

After identification of barcodes in the separate libraries above, we ensured that there were no collisions between barcodes. More specifically, a table compiling all oBC to CRE dictionaries from the libraries used in the experiment (literature-selected CRE, paired CRE, allelic series of mutated putative transcription binding sites, and exogenous promoter set). Then, barcodes duplicated across the libraries were removed as ambiguous for interpretation of the single-cell data. The final number of unique (unambiguously usable) scQer constructs was 114.3k, with the following construct coverage per library:

Paired CREs: pluripotent up::endoderm down high complexity: 14.8k

Paired CREs: pluripotent up::endoderm down low complexity: 2.0k

Paired CREs: endoderm up::pluripotent down high complexity: 21.1k

Paired CREs: endoderm up::pluripotent down low complexity: 0.3k

Literature-selected CREs: 13.7k

Mutated CREs (optimised/disrupted binding sites): 61.3k

Exogenous promoter series: 1.1k

7.4 scQer experiment, data processing, and analysis

The 7 libraries above were pooled with the following proportions (pluripotent up::endoderm down high 13%, pluripotent up::endoderm down low 13%, endoderm up::pluripotent down high 13%, endoderm up::pluripotent down low 13%, mutated CREs 33%, literature-selected CREs 12%, exogenous promoters 3%) prior to transfection in mESCs (biological triplicates, 2M cells per transfection, lipofectamine 2000). Puromycin selection (2 ug/mL) was applied following cell post-

transfection recovery (day 3), and mEB induction (4 plates of mEB per replicate, one set of mEB per transfection biological replicate) initiated on day 10 post-transfection after unintegrated plasmid dilution. On day 23 post mEB induction, 2 plate's worth of mEBs cells were processed and sequencing libraries constructed as described before (dissociation of mEB to single-cell suspension, FACS sorting for individual cells, 10x library prep [3' v3.1 with feature barcoding, one lane per biological replicate], scQer library generation: GEx, mBC, and oBC).

Sequencing data was processed as before (oBC & mBC: cell barcode error correction through cellRanger, extraction of barcodes from bam file, UMI error correction and pile-up; GEx: cellRanger 6.0.1 with mm10-3.0.0 transcriptome, retaining high read count cells [>450 transcriptome UMI, >1% and <12.5% fraction of reads mapping to mitochondrial genes], doublet removal with scrublet [<0.3 doublet score], initial processing and dimensional reduction with Seurat [NormalizeData, normalization.method ='LogNormalize'; FindVariableFeatures, nfeatures=1000; ScaleData; RunPCA; FindNeighbors with 50 top principal components; FindClusters, resolution=0.2; RunUMAP]. Putative doublets were further removed by sub-clustering (processing from full data applied to clusters, sub-clusters identified with FindClusters, resolution=0.5; cells from sub-clusters with median scrublet score ≥0.15 were removed). Cells from clusters making up <1% of all cells were not considered. Cells with outlier transcriptome counts (GEx UMI count > 8000) and MOI (number of detected oBC [>10 UMI/cell] > 110) were also not considered further as possible doublets. After these quality control thresholds were applied, we were left respectively with n=6124, 6442, and 7911 cells across our three biological replicates. Clusters were annotated by inspection of marker genes and comparison with previously systematically integrated data (i.e., **Ext. Data Fig. 4a**).

Bootstrap resampling was used to assess activity of CREs. Specifically, for each cell-type and CRE whose activity was to be tested, the set of normalised mBC UMI counts corresponding to detected reporter events (oBC UMI>10/cell) were collated for tests and controls. Given the lower representation (resulting from smaller MOI & less differentiated cells), the bootstrapping analysis was performed on all replicates pooled. These normalised mBC UMI counts were then sampled with replacement for $10^4$ bootstraps (number of sampling per bootstrap equal number of detections). For each bootstrap, the 1% winsorised mean was calculated. The bootstrap p-value was taken as the proportion of bootstraps in which the mean from test ≤ control (to assess higher expression, respectively ≥ to assess lower expression). The bootstrap p-values were adjusted to an FDR by the method of Benjamin-Hochberg.

# 4: SUMMARY AND FUTURE DIRECTIONS

The genome is the blueprint for multicellular mammalian development. The work described in my thesis attempts to uncover the various genetic contributions to cell fate decisions that are foundational for developing embryos. In pursuit of these efforts, we developed new technological approaches, and while they will surely pave the way to new biological insights, it is worth noting that a number of technical challenges were met:

- Transgene silencing upon differentiation
- Stochastic bottlenecks in mosaic EBs
- Deciphering origins of inter-gastruloid heterogeneity
- Limitations of conventional MPRA designs

Overcoming these hurdles improved our measurements, inspired new pipelines for embryoid generation, and set the stage for exciting follow up experiments.

In our pooled CRISPR screens using mosaic embryoid bodies, we encountered the issue of transgene silencing. We initially used lenti-viral constructs to deliver and stably express sgRNAs. Robust expression of the lenti-viral constructs was noted in the mESC state; however, upon differentiation into mEBs, lenti-viral constructs were silenced. The silencing could easily be observed with FACS analysis of the GFP population wherein ~80% of cells were negative for GFP by day 21 of EB differentiation. Because of silencing, we would be underpowered in terms of measuring functional consequences of sgRNAs if single cells from EBs were profiled *en masse*. Instead, we sorted out the remaining GFP positive population, thereby enriching cells to be profiled that express sgRNAs throughout differentiation. Although this workaround was a sufficient short-term solution, it meant that we were largely confined to transcriptional profiling with the 10x

Genomics platform where limits on numbers of cells per lane, typically around 10K cells per lane for the v3.1 standard kit, could reasonably be met with FACS sorting. To capitalize on scalable approaches like transcriptional profiling with high throughput technologies such as sci-RNA-seq, high cell numbers (typically 1-2 million cells) are often required. Sorting on a low GFP positive population is infeasible due to the time-intensive nature of achieving the required input. We therefore generated a new construct, 'piggyFlex', a piggyBac transposon that is not silenced. The dual-RNA architecture of this construct enables perturbations that can easily be readout across multiple scRNA-seq platforms, including droplet based 10x platform and combinatorial indexing based sci-RNA-seq.

Another challenge we encountered was stochastic bottlenecks during mosaic EB differentiation. Mosaic EBs are derived from many starting precursor cells, which prevents a straightforward approach to track heterogeneity between EBs. To circumvent this issue, we developed a clonal EB system that enables EB-to-EB variability to be measured with a single-cell readout. In short, the clonal EB system works by first integrating piggyFlex at a low multiplicity of integration in mESCs in which a paired sgRNA and organoid barcode are co-expressed. Single mESCs are seeded on a MEF layer and will clonally expand over 5 days after which the single-cell turned colony is lifted and placed in differentiation media. The end result is a pool of clonal EBs each traceable via the organoid barcode. In two proof-of-principle experiments, we show that we can easily measure EB-to-EB heterogeneity and devise CRISPR perturbations to ascertain cell fate dependency on key TFs. Such a system paves the way for further investigations that aim to understand the heterogeneity that is intrinsic to organoid and stembryo models.

Moreover, the clonal EB system underscores the powerful possibility to reconstruct a monophyletic tree of differentiation proceeding from a single precursor cell. In Chapter 2, we apply DNA Typewriter to clonal gastruloids and succeed in producing a high resolution tree. In these initial experiments, we observe inter-gastruloid variability, which we set out to investigate further. Repurposing the DNA Typewriter into an in cellular barcoding system, we generated ~150 clonal gastruloids, each uniquely barcoded by the DNA Typewriter. We therefore profiled all clonal gastruloids en masse where we captured both transcriptomes and Tape barcodes. We leverage the barcode information to group cells based on their clonal origins which enables us to then measure inter-gastruloid variability. We observed a greater variety of cell type composition as compared to conventional gastruloids, including more anterior derived cell types such as the brain. A notable finding in this experiment is that more heterogeneity is captured when profiling clonal gastruloids en masse as compared to profiling only a handful of gastruloids as was done in the prior experiment. While we measured a high degree of gastruloid-to-gastruloid variability we wondered if such variation is a heritable trait. We therefore developed a pipeline in which we record lineage of ~100 clonal gastruloids, each of which descends from a common epiblast-like aggregate, itself derived from a single cell. This system allows us to reconstruct a 'tree-of-trees' where we can ask whether cell type variation is a heritable trait. While the analysis is still ongoing, we are encouraged by preliminary results suggesting that cell type composition is encoded early on at pre-differentiation, which for *in vivo* development implies that seemingly identical precursor epiblast cells in the embryo may already encode cell lineage information.

Functionally characterizing the nearly 1 million mammalian enhancers that have been nominated via mostly descriptive assays requires multi-pronged approaches that are highly multiplexable.

Massively Parallel Reporter Assays are a promising approach that is highly scalable. However, conventional designs limit their utility to static cell lines despite the fact that a greater repertoire of enhancer elements are actually active in contexts that exhibit dynamical processes, such as differentiated multicellular systems that mimic development. Studying enhancer activity in multicellular contexts requires a single-cell platform in which the enhancer measurements are compatible with single-cell readouts. In order to overcome the limitations of conventional MPRA designs, which are primarily based on one-RNA systems in which a single transcript is used for both detection of the construct (is it in the cell or not?) and activity measurements (is the enhancer active or not?). Therefore, one-RNA systems make it impossible to accurately interpret the meaning of no activity as either absence of a construct or inactive enhancer element, which is particularly important in the context of a multicellular system in which enhancer activity will vary in a cell-type specific manner. In the context of scQers (Chapter 3), the goal was to characterize the autonomous functions of putative enhancer elements in a multicellular context. The novelty of the scQers construct lies in its dual-RNA system, where the detection of the construct (i.e., its presence or absence) is driven by highly expressed and stable polIII barcoded circular RNAs. This system offers a key advantage: near-deterministic detection of the construct in scRNA-seq. On the other hand, quantification (i.e., enhancer DNA activity) is assessed by the expression of a polII barcoded mRNA, itself driven by the queried non-coding DNA sequence. The dual-expression construct therefore manages the detection and quantification tasks that one-RNA constructs, such as conventional MPRA designs, fail to do. Using mosaic EBs, we screened hundreds of putative regulatory elements and discovered numerous DNA sequences with cell-type specific activity. With this result, we demonstrate that scQers not only overcomes the limitations of a one-RNA MPRA design in a single-cell context, but also paves the way for scalable testing in multicellular

systems.

Looking to the future, I envision three key areas that will be wide open for investigation using organoid models that span mosaic and clonal pipelines. Firstly, our pilot experiment using clonal EBs in a pooled CRISPR screen highlights key advantages over mosaic organoid screens. Although the clonal EB TF screen, discussed in Chapter 2, was a success, we only targeted 9 TFs. A next step is to scale this up, minimally to the 125 TFs identified as statistically significant in our first large-scale pooled CRISPR screen in the mosaic EB context. Secondly, clonal and mosaic models may, in some limited cases, be interdependent with each other to address key questions that neither can address independently. For example, addressing key biological questions related to cell autonomous and non-autonomous processes may require both clonal and mosaic approaches. Finally, as shown in Chapter 2, clonal organoid models are a robust choice for pooled CRISPR screens and lineage tracing experiments. Yet, there is a strong precedent for combining these two modalities in a single experiment. Achieving this will have wide-ranging implications, including deeper molecular descriptions of genotype-to-phenotype linkages for a wide spectrum of disease mutations.

Clonal EBs highlight two key advantages over mosaic systems: 1) clonal EBs are not prone to stochastic bottlenecks as is the case for mosaic EBs, and 2) EB-to-EB heterogeneity can be assessed in a systematic way. In our clonal EB TF screen we showed that perturbations for individual clonal EBs can be detected with single-cell readouts. However, our screen was limited to 9 TFs. The next steps are therefore to scale up the number of TFs in a single clonal EB TF experiment. In our large-scale mosaic EB TF screen targeting all known TFs (~1600), we identified

125 statistically significant hits. The follow up large-scale mosaic EB TF screen targeting these 125 TFs was not reproducible across two biological replicates, that we predict was due to stochastic bottlenecks intrinsic to each of the replicates in the mosaic context. Thus returning to this same set of TFs in a clonal EB TF perturbation screen may more definitively disentangle a true effect versus bias from EB-to-EB heterogeneity.

The decision to work with clonal or mosaic organoid models depends on the goal or biological insight. Yet, there's at least one potential area where using both is desirable. For example, deciphering processes that are cell autonomous versus non-autonomous. For example, I envision large-scale clonal EB perturbation screens yielding new key TF targets that will require further follow up for in-depth characterization. In a completely clonal context, a perturbation, experienced by all cells that make up the individual or clonal EB, may either be lethal to the extent that it is missing in the data or skews cell composition in a significant way. Yet, the effect of the TF perturbation may also have a phenotype that presents differently if in a mosaic context that is mixed with wild-type and perturbation cells. Understanding the dynamics of key TFs on processes throughout development (e.g. pluripotency, differentiation, lineage specification etc.).

While the choice to pursue screens in a mosaic *in vitro* system may be obvious in the context of non-coding DNA screens, we make a strong case for clonal systems in Chapter 2, especially in the context of lineage tracing or pooled CRISPR screens targeting cell-type specific regulators. We are at a particularly exciting time in genomics where many new technologies are improving our discovery rate. Yet, a comprehensive understanding of the regulators of cell type specification or lineage reconstruction across all cell-types that emerge throughout development remains sparse,

let alone the combination of the two in a developmental system. Even once we have a complete lineage reference or have characterized all TFs for all of mammalian development, understanding how lineage skews away from normal in the context of a perturbation will require clonal organoid systems in which both perturbation and lineage modalities are ascertained for the same individual. This will substantially speed up a major goal to develop a reference tree that represents the many manifolds of developmental cell fate trajectories required for mammalian organogenesis.

# References


**Chapter 1:**

1. Rossant, J., and Tam, P. P. Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse. *Development* **136**, 701–713 (2009).
2. Lim, J., and Thiery, J. P. Epithelial-mesenchymal transitions: insights from development. *Development* **139**, 3471–3486 (2012).
3. Spitz, F., Furlong, E. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).
4. Stent, G. S. Developmental cell lineage. *Int. J. Dev. Biol.* **42**, 237–241 (1998)
5. Sulston J. E., Schierenberg E., White J. G., and Thomson J. N., The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**, 64–119 (1983).
6. Kretzschmar K. and Watt F. M., Lineage tracing. *Cell* **148**, 33–45 (2012).
7. Bock, C. *et al.* High-content CRISPR screening. *Nat Rev Methods Primers* **2**, 8 (2022).
8. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
9. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
10. Choi, J. *et al.* A time-resolved, multi-symbol molecular recorder via sequential genome editing. *Nature* **608**, 98–107 (2022).
11. Jin, X. *et al.* In vivo Perturb-seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science* **370**, aaz6063 (2020).
12. Kuhn, M., Santinha, A. J. & Platt, R. J. Moving from in vitro to in vivo CRISPR screens. *Gene Genome Editing* **2**, 100008 (2021).
13. Bernstein BE. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol*. **28**, 1045–1048 (2010).
14. Klein, J.C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* **17**, 1083–1091 (2020).
15. El Azhar Y., Sonnen K. F. Development in a Dish-In Vitro Models of Mammalian Embryonic Development. *Front. Cel Dev. Biol.* **9**, 655993 (2021).

**Chapter 2:**

1. El Azhar Y., Sonnen K. F. Development in a Dish-In Vitro Models of Mammalian Embryonic Development. *Front. Cel Dev. Biol.* **9**, 655993 (2021).
2. Lancaster, M.A. *et al.* Cerebral organoids model human brain development and microcephaly. Nature. 2013;501:373–9.
3. Drakhlis, L. *et al.* Human heart-forming organoids recapitulate early heart and foregut development. *Nat. Biotechnol.* 39, 737–746 (2021).
4. Takasato, M. *et al.* Kidney organoids from human iPS cells contain multiple lineages and

model human nephrogenesis. *Nature* **526**, 564–568 (2015).

5. van den Brink, S.C. *et al.* Symmetry breaking, germ layer specification and axial organisation in aggregates of mouse embryonic stem cells. Development. Nov;141(22):4231-42 (2014 ).
6. Doetschman, T.C. *et al*. The *in vitro* development of blastocyst-derived embryonic stem cell lines: formation of visceral yolk sac, blood islands and myocardium. *Development* 87:27–45 (1985).
7. Bock, C. *et al.* High-content CRISPR screening. *Nat Rev Methods Primers* **2**, 8 (2022).
8. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
9. Veenvliet, J.V. *et al.* Mouse embryonic stem cells self-organize into trunk-like structures with neural tube and somites.*Science* **370**, eaba4937 (2020).
10. Schiffman, J.S. *et al.* Defining ancestry, heritability and plasticity of cellular phenotypes in somatic evolution *bioRxiv* (2023) doi.org/10.1101/2022.12.28.522128.

**Chapter 3:**

1. Kvon, E. Z. *et al.* Comprehensive In Vivo Interrogation Reveals Phenotypic Impact of Human Enhancer Variants. *Cell* **180**, 1262–1271.e15 (2020).
2. Claringbould, A. & Zaugg, J. B. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends Mol. Med.* **27**, 1060–1073 (2021).
3. Lim, F. *et al.* Affinity-optimizing enhancer variants disrupt development. *Nature* 1–9 (2024).
4. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
5. Hay, D. *et al.* Genetic dissection of the α-globin super-enhancer in vivo. *Nat. Genet.* **48**, 895–903 (2016).
6. Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell* **167**, 633–642.e11 (2016).
7. Synthetic regulatory genomics uncovers enhancer context dependence at the Sox2 locus. *Mol. Cell* **83**, 1140–1152.e7 (2023).
8. Super-enhancers include classical enhancers and facilitators to fully activate gene expression. *Cell* **186**, 5826–5839.e18 (2023).
9. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
10. Weingarten-Gabbay, S. *et al.* Systematic interrogation of human promoters. *Genome Res.* **29**, 171–183 (2019).
11. Sahu, B. *et al.* Sequence determinants of human gene regulatory elements. *Nat. Genet.* **54**, 283–294 (2022).

12. Vockley, C. M. *et al.* Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Research* vol. 25 1206–1214 Preprint at https://doi.org/10.1101/gr.190090.115 (2015).

13. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* **17**, 1083–1091 (2020).

14. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).

15. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).

16. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).

17. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

18. de Almeida, B. P., Reiter, F., Pagani, M. & Stark, A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat. Genet.* **54**, 613–624 (2022).

19. Gosai, S. J. *et al.* Machine-guided design of synthetic cell type-specific -regulatory elements. *bioRxiv* (2023) doi:10.1101/2023.08.08.552077.

20. Agarwal, V. *et al.* Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. *bioRxiv* (2023) doi:10.1101/2023.03.05.531189.

21. Wilkinson, A. C. *et al.* Single site-specific integration targeting coupled with embryonic stem cell differentiation provides a high-throughput alternative to in vivo enhancer analyses. *Biol. Open* **2**, 1229–1238 (2013).

22. Dickel, D. E. *et al.* Function-based identification of mammalian enhancers using site-specific integration. *Nat. Methods* **11**, 566–571 (2014).

23. Edginton-White, B. *et al.* A genome-wide relay of signalling-responsive enhancers drives hematopoietic specification. *Nat. Commun.* **14**, 267 (2023).

24. Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N. & Yosef, N. Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. *Cell Stem Cell* **25**, 713–727.e10 (2019).

25. Thomas, H. F. *et al.* Temporal dissection of an enhancer cluster reveals distinct temporal and functional contributions of individual elements. *Mol. Cell* **81**, 969–982.e13 (2021).

26. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).

27. Farley, E. K. *et al.* Suboptimization of developmental enhancers. *Science* **350**, 325–328 (2015).

28. Kvon, E. Z. *et al.* Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* **512**, 91–95 (2014).

29. Fuqua, T. *et al.* Dense and pleiotropic regulatory information in a developmental enhancer. *Nature* **587**, 235–239 (2020).

30. Simunovic, M. & Brivanlou, A. H. Embryoids, organoids and gastruloids: new approaches to understanding embryogenesis. *Development* **144**, 976–985 (2017).

31. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).

32. Hill, A. J. *et al.* On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* **15**, 271–274 (2018).

33. Dixit, A. Correcting Chimeric Crosstalk in Single Cell RNA-seq Experiments. Preprint at https://doi.org/10.1101/093237.

34. Litke, J. L. & Jaffrey, S. R. Highly efficient expression of circular RNA aptamers in cells using autocatalytic transcripts. *Nat. Biotechnol.* **37**, 667–675 (2019).

35. Dao, L. T. M. & Spicuglia, S. Transcriptional regulation by promoters with enhancer function. *Transcription* vol. 9 307–314 Preprint at https://doi.org/10.1080/21541264.2018.1486150 (2018).

36. Vannini, A. & Cramer, P. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Mol. Cell* **45**, 439–446 (2012).

37. Yeganeh, M., Praz, V., Cousin, P. & Hernandez, N. Transcriptional interference by RNA polymerase III affects expression of the gene. *Genes Dev.* **31**, 413–421 (2017).

38. Lukoszek, R., Mueller-Roeber, B. & Ignatova, Z. Interplay between polymerase II- and polymerase III-assisted expression of overlapping genes. *FEBS Lett.* **587**, 3692–3695 (2013).

39. Ma, H. *et al.* CRISPR-Cas9 nuclear dynamics and target recognition in living cells. *J. Cell Biol.* **214**, 529–537 (2016).

40. Qin, J. Y. *et al.* Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PLoS One* **5**, e10611 (2010).

41. Yusa, K., Zhou, L., Li, M. A., Bradley, A. & Craig, N. L. A hyperactive piggyBac transposase for mammalian applications. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1531–1536 (2011).

42. Ribeiro-Dos-Santos, A. M., Hogan, M. S., Luther, R. D., Brosh, R. & Maurano, M. T. Genomic context sensitivity of insulator function. *Genome Res.* **32**, 425–436 (2022).

43. Wang, Y., Xie, S., Armendariz, D. & Hon, G. C. Computational identification of clonal cells in single-cell CRISPR screens. *BMC Genomics* **23**, 135 (2022).

44. Svensson, V. Droplet scRNA-seq is not zero-inflated. Preprint at https://doi.org/10.1101/582064.

45. Akhtar, W. *et al.* Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* **154**, 914–927 (2013).

46. Maricque, B. B., Chaudhari, H. G. & Cohen, B. A. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nature Biotechnology* vol. 37 90–95 Preprint at https://doi.org/10.1038/nbt.4285 (2019).

47. Moudgil, A. *et al.* Self-reporting transposons enable simultaneous readout of gene expression and transcription factor binding in single cells. Preprint at https://doi.org/10.1101/538553.

48. Chung, J. H., Bell, A. C. & Felsenfeld, G. Characterization of the chicken beta-globin insulator. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 575–580 (1997).

49. Martin, G. R. & Evans, M. J. Differentiation of clonal lines of teratocarcinoma cells: formation of embryoid bodies in vitro. *Proc. Natl. Acad. Sci. U. S. A.* **72**, 1441–1445 (1975).

50. Doetschman, T. C., Eistetter, H., Katz, M., Schmidt, W. & Kemler, R. The in vitro development of blastocyst-derived embryonic stem cell lines: formation of visceral yolk sac, blood islands and myocardium. *J. Embryol. Exp. Morphol.* **87**, 27–45 (1985).

51. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

52. Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).

53. Argelaguet, R. *et al.* Decoding gene regulation in the mouse embryo using single-cell multi-omics. *bioRxiv* 2022.06.15.496239 (2022) doi:10.1101/2022.06.15.496239.

54. Fujikura, J. *et al.* Differentiation of embryonic stem cells is induced by GATA factors. *Genes Dev.* **16**, 784–789 (2002).

55. Mannion, B. J. *et al.* Uncovering Hidden Enhancers Through Unbiased In Vivo Testing. *bioRxiv* 2022.05.29.493901 (2022) doi:10.1101/2022.05.29.493901.

56. Li, Y. *et al.* CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One* **9**, e114485 (2014).

57. Zhou, H. Y. *et al.* A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes Dev.* **28**, 2699–2711 (2014).

58. Horlbeck, M. A. *et al.* Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife* **5**, (2016).

59. Gam, J. J., DiAndreth, B., Jones, R. D., Huh, J. & Weiss, R. A 'poly-transfection' method for rapid, one-pot characterization and optimization of genetic systems. *Nucleic Acids Research* vol. 47 e106–e106 Preprint at https://doi.org/10.1093/nar/gkz623 (2019).

60. Kalhor, R. *et al.* Developmental barcoding of whole mouse via homing CRISPR. *Science* **361**, (2018).

61. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).

62. Peng, T. *et al.* STARR-seq identifies active, chromatin-masked, and dormant enhancers in pluripotent mouse embryonic stem cells. *Genome Biol.* **21**, 243 (2020).

63. Weinberger, L., Ayyash, M., Novershtern, N. & Hanna, J. H. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nat. Rev. Mol. Cell Biol.* **17**, 155–169 (2016).

64. Buecker, C. *et al.* Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell* **14**, 838–853 (2014).

65. Papanayotou, C. *et al.* A novel nodal enhancer dependent on pluripotency factors and smad2/3 signaling conditions a regulatory switch during epiblast maturation. *PLoS Biol.* **12**, e1001890 (2014).

66. Blassberg, R. *et al.* Sox2 levels regulate the chromatin occupancy of WNT mediators in epiblast progenitors responsible for vertebrate body formation. *Nat. Cell Biol.* **24**, 633–644 (2022).

67. Chakraborty, S. *et al.* Enhancer-promoter interactions can bypass CTCF-mediated boundaries and contribute to phenotypic robustness. *Nat. Genet.* **55**, 280–290 (2023).

68. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* **21**, 292–310 (2020).

69. Sartorelli, V. & Lauberth, S. M. Enhancer RNAs are an important regulatory layer of the epigenome. *Nat. Struct. Mol. Biol.* **27**, 521–528 (2020).

70. Muerdter, F. *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141–149 (2018).

71. Zahm, A. M. *et al.* Discovery and Validation of Context-Dependent Synthetic Mammalian Promoters. *bioRxiv* (2023) doi:10.1101/2023.05.11.539703.

72. Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.* **34**, 1180–1190 (2016).

73. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **172**, 1132–1134 (2018).

74. McAfee, J. C. *et al.* Systematic investigation of allelic regulatory activity of schizophrenia-associated common variants. *Cell Genom* **3**, 100404 (2023).

75. Mangan, R. J. *et al.* Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell* **185**, 4587–4603.e23 (2022).

76. Hrvatin, S. *et al.* A scalable platform for the development of cell-type-specific viral drivers. *Elife* **8**, (2019).

77. Zhao, S. *et al.* A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. *Nat. Genet.* **55**, 346–354 (2023).

78. Martinez-Ara, M., Comoglio, F., van Arensbergen, J. & van Steensel, B. Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome. *Mol. Cell* **82**, 2519–2531.e6 (2022).

79. Bergman, D. T. *et al.* Compatibility rules of human enhancer and promoter sequences. *Nature* **607**, 176–184 (2022).

80. Martinez-Ara, M., Comoglio, F. & van Steensel, B. Large-scale analysis of the integration of enhancer-enhancer signals by promoters. *bioRxiv* 2023.08.11.552995 (2023) doi:10.1101/2023.08.11.552995.

81. Goel, V. Y., Huseyin, M. K. & Hansen, A. S. Region Capture Micro-C reveals coalescence of enhancers and promoters into nested microcompartments. *Nat. Genet.* **55**, 1048–1056 (2023).

82. Hua, P. *et al.* Defining genome architecture at base-pair resolution. *Nature* **595**, 125–129 (2021).

83. Martin, B. K. *et al.* Optimized single-nucleus transcriptional profiling by combinatorial indexing. *Nat. Protoc.* **18**, 188–207 (2023).

84. Minnoye, L. *et al.* Cross-species analysis of enhancer logic using deep learning. *Genome Res.* **30**, 1815–1834 (2020).

85. Taskiran, I. I. *et al.* Cell-type-directed design of synthetic enhancers. *Nature* 1–9 (2023).

86. de Almeida, B. P. *et al.* Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo. *Nature* 1–5 (2023).

87. Wong, E. S. *et al.* Deep conservation of the enhancer regulatory code in animals. *Science* **370**, (2020).

88. Williams, R. M. *et al.* Reconstruction of the Global Neural Crest Gene Regulatory Network In Vivo. *Dev. Cell* **51**, 255–276.e7 (2019).

89. Tarazi, S. *et al.* Post-gastrulation synthetic embryos generated ex utero from mouse naive ESCs. *Cell* **185**, 3290–3306.e25 (2022).

90. Amadei, G. *et al.* Synthetic embryos complete gastrulation to neurulation and organogenesis. *Nature* (2022) doi:10.1038/s41586-022-05246-3.

91. Graybuck, L. T. *et al.* Enhancer viruses for combinatorial cell-subclass-specific labeling. *Neuron* **109**, 1449–1464.e13 (2021).

92. Mich, J. K. *et al.* Enhancer-AAVs allow genetic access to oligodendrocytes and diverse populations of astrocytes across species. *bioRxiv* (2023) doi:10.1101/2023.09.20.558718.

93. Psatha, N. *et al.* Large-scale discovery of potent, compact and lineage specific enhancers for gene therapy vectors. *bioRxiv* (2023) doi:10.1101/2023.10.04.559165.

94. Calderon, D. *et al. Trans*MPRA: A framework for assaying the role of many *trans*-acting factors at many enhancers. Preprint at https://doi.org/10.1101/2020.09.30.321323.

95. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).

96. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).

97. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

98. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems* vol. 8 281–291.e9 Preprint at https://doi.org/10.1016/j.cels.2018.11.005 (2019).

99. Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134 (2012).

100. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

101. You, F. M. *et al.* BatchPrimer3: A high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* vol. 9 Preprint at https://doi.org/10.1186/1471-2105-9-253 (2008).

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

103. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

104. Hume, M. A., Barrera, L. A., Gisselbrecht, S. S. & Bulyk, M. L. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **43**, D117–22 (2015).

105. Mariani, L., Weinand, K., Vedenko, A., Barrera, L. A. & Bulyk, M. L. Identification of Human Lineage-Specific Transcriptional Coregulators Enabled by a Glossary of Binding Modules and Tunable Genomic Backgrounds. *Cell Systems* vol. 5 654 Preprint at https://doi.org/10.1016/j.cels.2017.12.011 (2017).

106. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).

107. Schraivogel, D. *et al.* Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat. Methods* **17**, 629–635 (2020).

108. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).

109. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

110. Shin, H. Y. *et al.* Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nat. Genet.* **48**, 904–911 (2016).

111. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103–1116.e20 (2020).

112. Crocker, J. *et al.* Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* **160**, 191–203 (2015).

113. Samee, M. A. H. *et al.* Quantitative Measurement and Thermodynamic Modeling of Fused Enhancers Support a Two-Tiered Mechanism for Interpreting Regulatory DNA. *Cell Rep.* **21**, 236–245 (2017).

114. Visel, A. *et al.* Functional autonomy of distant-acting human enhancers. *Genomics* **93**, 509–513 (2009).

115. Lim, F. *et al.* Affinity-optimizing variants within the ZRS enhancer disrupt limb development. *bioRxiv* 2022.05.27.493789 (2022) doi:10.1101/2022.05.27.493789.

116. Artus, J., Piliszek, A. & Hadjantonakis, A.-K. The primitive endoderm lineage of the mouse blastocyst: sequential transcription factor activation and regulation of differentiation by Sox17. *Dev. Biol.* **350**, 393–404 (2011).

117. White, M. A. *et al.* A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors. *Cell Rep.* **17**, 1247–1254 (2016).

118. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–92 (2007).

119. Hansen, T. J. & Hodges, E. ATAC-STARR-seq reveals transcription factor-bound activators and silencers across the chromatin accessible human genome. *Genome Res.* **32**, 1529–1541 (2022).

120. Glaser, L. V. *et al.* Assessing genome-wide dynamic changes in enhancer activity during early mESC differentiation by FAIRE-STARR-seq. *Nucleic Acids Res.* **49**, 12178–12195 (2021).

121. Datlinger, P. *et al.* Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat. Methods* **18**, 635–642 (2021).

122. David, F. N. & Johnson, N. L. The Truncated Poisson. *Biometrics* vol. 8 275 Preprint at https://doi.org/10.2307/3001863 (1952).

123. Simeonov, K. P. *et al.* Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. *Cancer Cell* **39**, 1150–1162.e9 (2021).

124. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).

125. Thibodeau, A. *et al.* AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. *Genome Biol.* **22**, 252 (2021).

126. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).

127. Rhodes, K. *et al.* Human embryoid bodies as a novel system for genomic studies of functionally diverse cell types. *Elife* **11**, (2022).

128. Mohammed, H. *et al.* Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation. *Cell Rep.* **20**, 1215–1228 (2017).

129. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

130. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).

131. Garreta, R. & Moncecchi, G. *Learning scikit-learn: Machine Learning in Python*. (Packt Publishing Ltd, 2013).

132. David, F. P. A., Rougemont, J. & Deplancke, B. GETPrime 2.0: gene- and transcript-specific qPCR primers for 13 species including polymorphisms. *Nucleic Acids Res.* **45**, D56–D60 (2017).

133. Milo, R., Jorgensen, P., Moran, U., Weber, G. & Springer, M. BioNumbers--the database of key numbers in molecular and cell biology. *Nucleic Acids Res.* **38**, D750–3 (2010).

134. Leahy, A., Xiong, J. W., Kuhnert, F. & Stuhlmann, H. Use of developmental marker genes to define temporal and spatial patterns of differentiation during embryoid body formation. *J. Exp. Zool.* **284**, 67–81 (1999).

135. Gohl, D. M. *et al.* Measuring sequencer size bias using REcount: a novel method for highly accurate Illumina sequencing-based quantification. *Genome Biol.* **20**, 85 (2019).

136. Zhao, M., Lee, W.-P., Garrison, E. P. & Marth, G. T. SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. *PLoS One* **8**, e82138 (2013).