OREGON HEALTH & SCIENCE UNIVERSITY SCHOOL OF MEDICINE –

GRADUATE STUDIES

NOVEL SINGLE-CELL OMICS ASSAYS OF CORTICOGENESIS

By

Ryan M. Mulqueen

A DISSERTATION

Presented to the Department of Molecular and Medical Genetics

and the Oregon Health & Science University

School of Medicine

in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

February 2021

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD dissertation of

Ryan M. Mulqueen

has been approved

X _____

Co-Mentor

X _____

Co-Mentor

X _____

Member

X _____

Member

X _____

Member

Table of Contents

List of Figures

Acknowledgements

This work to be presented here was carried out as a large collaboration across six years. For that I'd like to thank all those in the Adey and O'Roak labs who have helped me develop into a scientist. This was written during a tough time, a time when truth seemed to matter less and we were more isolated than ever. It speaks volumes that throughout this, I never felt lost or unsupported. I always knew the next steps forward, and for that I'd like to my two mentors: Brian and Andrew have been eternally patient as I pushed myself to learn everything from scratch. Their ability to step in when I needed guidance, and to let me loose when I found my own spark has branded my style of exploration forever. To lead by example, demonstrating what it means to balance a drive for career while maintaining personal perspective, hobbies, and family. For both of them agreeing to take a fresh-out-of-college New Yorker into their new labs, I am grateful. To my committee, I am thankful for them to focus on the context of my work in the field and for other scientists, over getting bogged down in wet lab development for the sake of development.

My development as a scientist is inseparable from my development as a person. I'd like to thank my family. Mike, Anna, Connor and my parents, Dad and Mom. Even when I began frothing at the mouth over biochemical yield, going on and on about reaction efficiencies, or promising to graduate every year for the last three years, my family has inspired me to keep bettering myself in all directions through their support and excitement. I'd like to thank my friends here in Portland, for being a family to me while I play the prodigal son. To my class in the PMCB program, many of them going through the same thing that I am, I thank them for our study sessions as they transformed into gripe sessions. I'd like to thank my lab mates, particularly Andy, Taylor, Brooke, and Marissa. I'd like to thank Kristof, for his ability to grow a community with his kindness and warm-heartedness. And finally, I'd like to thank Casey. You've been the greatest partner through all of this I could ask for; a constant source of joy and sanity. I can't wait to see what our future holds.

Abstract

Single-cell methods have proven to be a powerful tool in the interrogation of complex biology. One such biological system in which single-cell methods have been paramount to the discovery of complex disease and cell development is in the mammalian cortex. In this thesis I present an overview of single-cell method development for chromatin accessibility, chromatin conformation, whole genome sequencing, and whole methylome sequencing. I then proceed to describe the development and application of novel single-cell methylome methods, and apply this to a murine cortical sample to study neuronal methylomes. I present a new generalized chemistry to improve upon the established single-cell combinatorial indexing (sci) flow-through. This leads to substantial improvements of information garnered per cell for chromatin accessibility, chromatin conformation, and whole genome sequencing. I also apply a method to combine two prominent methods for single-cell chromatin accessibility to increase cellular throughput by over 15-fold. I apply this new method for a survey of murine and human mature cortex. Finally, I demonstrate the use of single-cell chromatin assays on the study of chromatin dynamics during corticogenesis in a model system of human forebrain development. Within this system, dynamic changes of enhancer usage for promoter, as well as the transcription factor usage changes as cells develop and mature into the mid-gestation cortex. This body of work bolsters the field of single-cell genomics by introducing novel strategies which address several key hurdles. Further, this work presents the generation of cell type and state atlases of human and mouse cortices.

Introduction

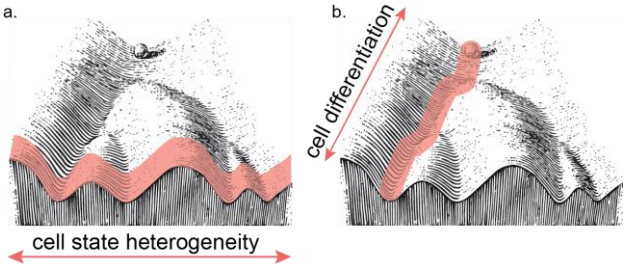


Figure 1 Abstracted landscape of cellular heterogeneity (a.) and differentiation (b.) through Waddington's model.

In 1956, the evolutionary biologist Conrad Waddington posited the concept of an epigenomic landscape. He depicted a ball moving in an uneven landscape, cutting a path through hills and valleys (Figure 1). Just as in this visual aid, a cell follows paths in a contiguous epigenomic landscape, subject to external and internal forces[1]. Coordinated expression of transcription factors, accessibility of promoters and enhancers, covalent tagging of modifying moieties, and genomic compartmentalization all inform cell state and are reflected in the abstract epigenomic landscape. Our understanding of how these forces act in concert is critical for interpretations of development and disorders. One such system where all factors collide is early cortical development[2]. The cortex is the seat of cognition, motor control and sensory perception. It forms in a stereotyped, layered pattern. Early neuronal precursor cells known as radial glia (RG) rapidly and asymmetrically divide, with each division either replenishing the stem cell pool or generating newborn neuronal or glial cell types. The balance between maintaining the progenitor pool and terminal differentiation of neurons and glia is critical[3]. Changes in this balance has been implicated in the roughly three-fold expansion of volume in the human cortex beyond that of other great apes[4]; with the difference apparent as early as mid-gestation[5]. Further, neurodevelopmental disorders such as schizophrenia and autism spectrum, have implicated dysregulation of cortical migration, differentiation, and layering[6–8]. All of this points to the need for a nuanced understanding of cellular diversity across the cortex and epigenomic changes occurring during corticogenesis. Single-cell methods for RNA sequencing have become commonplace for assessing cellular heterogeneity in complex tissues. Just as in bulk assays, RNA is reverse-transcribed, captured and amplified to generate sequencing libraries. This method has cataloged cell types and transcriptomic dynamics through cortical development[9–11]. However many of the genomic sites implicated in human cortical expansion[12] and

neurodevelopmental disorders[13–15] are non-coding, necessitating a deeper understanding of the epigenomic landscape. These lines of inquiry are lost in the limited scope of the transcriptome. To further our understanding of the cortex and its development, it is the goal of this thesis to develop new methods to discretely measure the epigenomic landscape – cognizant of cortical complexity – and thus at single-cell resolution.

Why single-cell analysis?

Bulk methods to capture chromatin accessibility[16], chromatin conformation[17], DNA methylation[18] and others have increased our understanding of cell state diversity and the interplay between these features[19,20]. In actively developing samples, or on samples of complex tissue, bulk assays fall short. By capturing and processing thousands to millions of cells together, signals are merged to an average, covering up cellular diversity and genomic decision points. To specify the region, cell type, or developmental stages linking neurotypical cortical development with disorders, efforts have been made to take sequential samples for a time-course analysis, or to perform micro-dissections of the cortex[21,22]. These two approaches, while able to garner critical information, still lack the granularity to catalog a causative string of events for cell fate decisions, leaving the development of new single-cell methods as a promising recourse.

Single-cell applications, the assessment of genomic or epigenomic profiles from discrete single cells, exists to address the shortcomings of those previous experimental designs. Sampling one cell at a time has two major benefits. The first is a less biased count of heterogeneity in a sample. For instance, previous post-mortem analysis demonstrated a large variance of cell type proportions across individual cortical samples. This heterogeneity has been known to skew analysis in bulk samples[23]. A second approach would be to isolate cells through a marker. However this approach can introduce bias into systems, especially when studying cell state transitions[24]. Through capturing cells in an unbiased manner and subsetting data to cells of interest *post hoc*, this effect can be mitigated and assumptions before data acquisition are limited — isolation of a single cell type allows for a more powerful case-control comparison. For instance, single-cell RNA libraries generated from 48 individuals with Alzheimer's disease pathology uncovered many more

differences using pairwise comparisons respective of cell types than have been previously uncovered in bulk data sets, which compare across all cell types at once. This was driven by the direct comparison between relatively sparse glial subtypes, which are masked in bulk libraries[25]. This same method has been applied to other disorders including major depression[13], autism spectrum[14], and multiple sclerosis[15]. Secondly, single-cell analysis is used for a higher resolution view at dynamic processes. Time course experiments done in bulk are limited by sampling rate, and generally share the same problem with averaging across many cells, or having to synchronize cell cycle or conditional responses prior to sampling. By using a single-cell approach one can capture and order cells through their progression of a state change[26]. Single-cell analysis allows for a higher resolution view into dynamic processes, and the regulatory landscape across complex tissue (Figure 1).

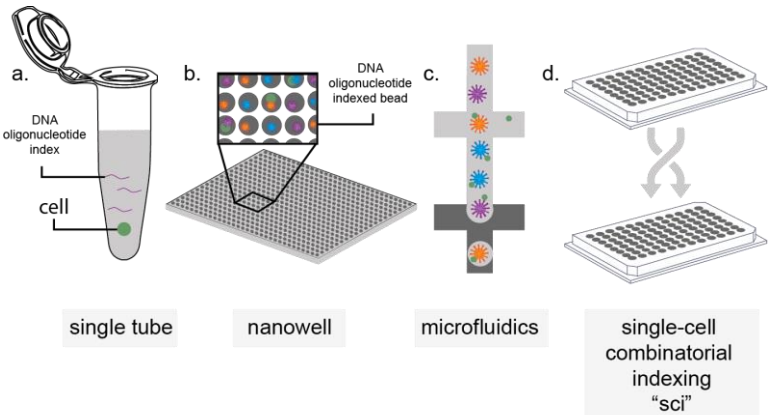


Figure 2. Schematized methods of single-cell isolation. a.) Single-cell single tube isolation, b.) nanowell isolation, c.) microfluidic droplet separation, d.) split-pool labelling (e.g. sci- chemistry). Cells (green) are isolated by various means and co-occupy a space with either DNA oligonucleotide indexes (purple line), oligonucleotide-coated beads (blue, purple, and orange stippled circles), or undergo combinatorial indexing.

Single-cell methods have been developed for many epigenomic and genomic assays, however the analyses share a common through-line. Cells are independently, and specifically labelled with a DNA oligonucleotide index that is shared in every sequence read out generated from that cell, such that each read can be assigned back to a specific reaction condition. Single-cell assays work through isolation; tissue or cell cultures are dissociated and single-cells are placed into a reaction vessel. In its simplest form, the reaction vessel is a single tube (Figure 2a). While this strategy tends to perform well on information captured per cell, a "one cell, one well" strategy is limiting in terms of both cost and effort, and these experimental designs tends to suffer from low cell counts. To address this, commercialized products have been developed to increase throughput. Nanowell platforms increase the throughput of the "one cell, one well" strategy by

shrinking the well and using specialized means of dispersing the dissociated cell suspension and decreasing reagent cost per cell by limiting volume[27] (Figure 2b). Alternatively, microfluidic droplet devices use water-oil emulsions as a means to isolate cells into their own partitioned reaction vessels[28–30] (Figure 2c). In these commercialized reactions, each reaction vessel also contains a microbead coated with a single index identifier that is unique, thus labelling the one cell present within the microfluidic droplet uniquely. A limitation that persists is each assay is that they are still bottlenecked by the one cell, one well strategy. This limits throughput and puts a burden of effort and cost on the experimenter. An alternative to these strategies, popularized by us and others, is the use of a split-pool indexing strategy called combinatorial indexing (Figure 2d). Cells can be uniquely labelled without ever being physically isolated. Instead of a single round of uniquely labelling cells, we perform multiple rounds of labelling, with random sampling in between. In this approach, the combination of indexes becomes the unique identifier for each cell. This process is empirically tailored to account for the random chance that multiple cells may follow the same path through library preparation. This is done by limiting the number of cells in the second round of indexing such that the likelihood of any two cell occupying the same well in the first and second round of indexing is sufficiently low[31]. This allows for multi-cell reactions without physical isolation of single cells. This strategy addresses both low cell count, and low assay efficiency concerns at once, driving down experimental cost and effort.

Regardless of epigenomic or transcriptomic assay, each single cell captured tends to have low information content, with a non-trivial amount of drop-out[32] due to inefficiency in information capture and inherently low input per reaction. To overcome information drop-out, cells are grouped together based on similarities across the measured moiety. Cells are then aggregated together making multiple “pseudo-bulk” libraries of pure cell types or states — agnostically grouping cells for unbiased analyses. The process of single-cell aggregation also highlights a key concern about experimental design: they must balance information per cell with number of cells sampled. Low information content per single cell requires assumptions to be made about cell grouping, as there is high noise in low-information content system. By having low cell count, there is a risk of losing rare cell types, or having insufficient power for pairwise analyses. This key concern informs the

strategy of single-cell capture and the assay used. In the case of cortical development, multiple cellular subtypes are differentiating in parallel, leading to a need for both breadth of cell count and high information per cell to make the most of captured rare events.

Cortex topology and taxonomy

The mature cortex consists of billions of neurons organized into a six-layered (L 1-6) structure. These neurons interact via short and long-range connections to form the complex circuitry which results in the emergent property of cognition. This cortical layout is highly conserved across mammals[33,34]. The neurons of the cortex occupy two major classes, GABAergic (inhibitory; iN) and glutamatergic (excitatory), and from those two major classes, dozens to hundreds of subtypes form, identified through distinct neuronal processes, circuit membership and gene expression patterns[22]. Glutamatergic excitatory neurons possess axonal projections which synapse at various brain regions, and projection properties dominate cell type grouping in single-cell RNA data[22]. These neurons form circuits with their inputs serving to promote action potential firing and downstream neural activity. Typography and topography (i.e. cell type and location) inform neuron function. For instance, glutamatergic pyramidal tract neurons are most common in deep layer L5 (ventrally located) and are associated with executing voluntary movements and planning. These project to subcortical targets like the striatum, thalamus, tectum and pons[35]. In contrast, glutamatergic intratelenecephalic trajectories connect excitatory neurons between cortical layers or cortical brain regions, and span most layers[22]. GABAergic inhibitory neurons have two major subclasses which reflect their point of origin[36]. Adenosine Deaminase RNA Specific B2 (*ADARB2*+) expressing inhibitory neurons are formed in the caudal ganglionic eminence, whereas LIM Homeobox 6 (*LHX*6+) neurons form in the medial ganglionic eminence (CGE and MGE, respectively). Inhibitory neurons produce the small molecule GABA, and modulate neuronal circuitry through dampening neuronal firing. Glia, which have previously been considered to be connective cells, have been found to serve critical roles in maintaining synapse integrity and cortical function. Astrocytes mediate blood brain barrier and have been shown to mediate synapse formation, elimination, and plasticity. Oligodendrocytes enable salutatory

conduction of actionable potentials throughout the brain. Microglia are resident immune cells, participate in phagocytosis and inflammation response. Neuronal circuitry is an active field of study – which benefits from understanding the unique states and types of neurons and glia[37].

The generation of cortical circuitry is critical. Dysfunction during corticogenesis has been implicated in multiple neurodevelopmental disorders[6–8], and evolutionary changes between humans and other apes have been linked to the rapid expansion of the human cortex[4]. In early embryonic development (GW4 in humans; E10 in mice; where "E' is embryonic day post conception and "GW" is gestational week), the ectodermal neural tube expands and compartmentalizes into the prosencephalon (forebrain), the mesencephalon (midbrain) and the rhombencephalon (hindbrain)[38]. The maturing forebrain subdivides along the dorsal-ventral axis into the pallium and subpallium, respectively. The pallium generates the bulk of the cerebral cortex and the subpallium forms the MGE and CGE (Figure 3)[39]. As the pallium develops, neuroepithelial cells differentiate to RG, named as such for their radial projection from the ventricular zone (VZ) towards the dorsal surface of the pallium, and for their combined marker set of neuroepithelial and astroglial expression patterns[40]. RG divide asymmetrically, both producing newborn RG to replenish the pool of stem cells, as well as forming intermediate progenitors (IPs) and a subset newborn Cajal-Retzius neurons (CR) directly. However most CR neurons are born exterior to the developing pallium and migrate tagentially into the marginal zone (MZ). IPs move dorsally to populate the subventricular zone (SVZ) while CRs migrate further to develop in the cortical plate (CP). RG continue to mitotically cycle while their nuclei rhythmically move dorsally up the VZ during G2/M phase (basal RG or bRG), and ventrally for/during/in S-phase along cellular projections in a process known as interkinetic nuclear migration (ventricular RG, or vRG; IKNM)[40]. This process continues through cortical development, with the self-replenishing pool of RG generate IPs and expand the VZ. The process of self-replenishing symmetric divisions and asymmetric neuron generation is partially regulated through the balance of key epigenetic regulators of cells, transcription factors PAX6 and EMX2, respectively[41,42]. IP cells, not anchored to the apical VZ, populate an outer area of the SVZ, split by an inner fiber layer (IFL), forming outer RG cells (oRG). IPs continue to divide and differentiate forming the cortex in an inside out manner, generating deep layer neurons, then the

more superficial layers[43]. Notably, RG and IPs are known to express messenger RNA (mRNA) associated with deep and superficial layer neurons markers prior to differentiation, though they don't express the resultant proteins. This is regulated through post-transcriptional repression mechanism and suggests a priming of RG/IPs throughout maturation[44]. A mature subset of IP, oRG cells form non-neuronal glial cells, such as oligodendrocytes, and astrocytes which permeate across the cortical layers[45]. In humans, these oRG cells are abundant and self-renew, a characteristic that has been postulated to lead to the human specific cortical expansion[9,46,47]. In recent work, differential gene expression of the transcription factor FOXO3 and genes a part of the

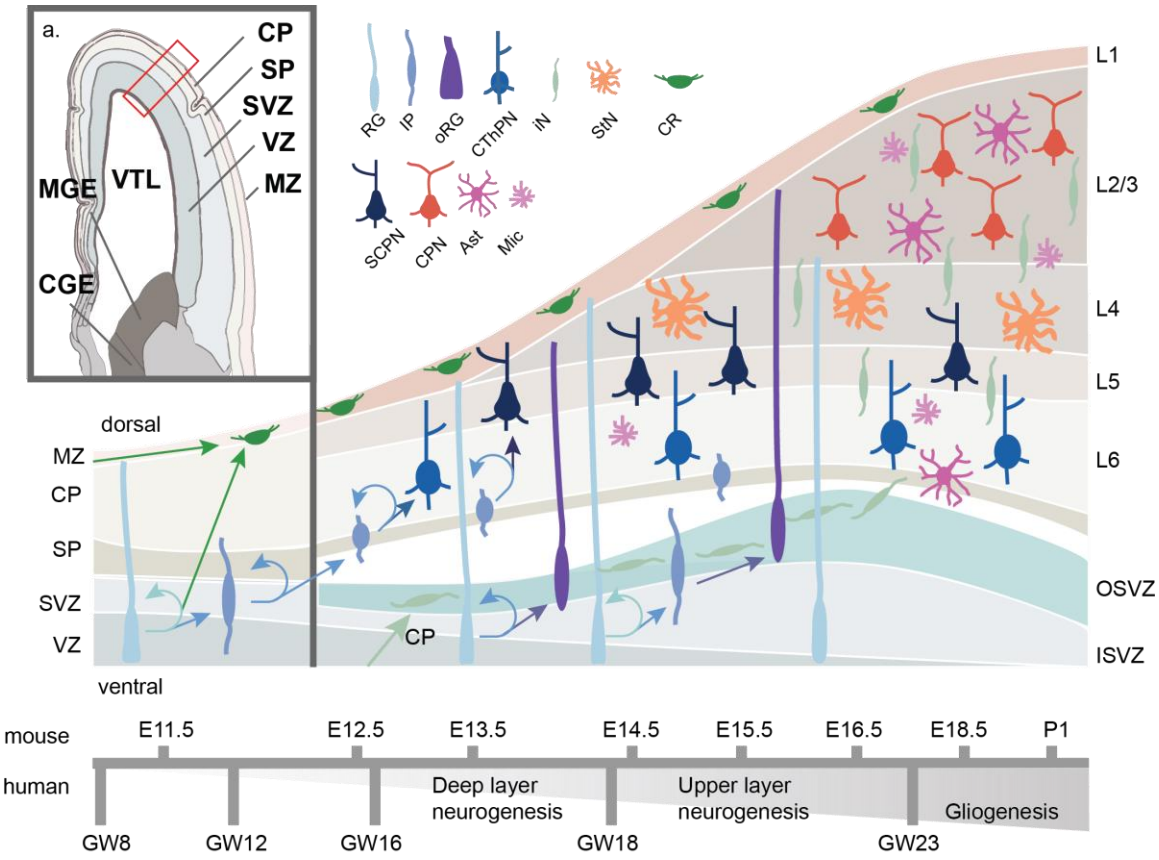


Figure 3. Schematic of human corticogenesis. a) Anatomical view of mid-gestational (GW13) human cortex, adapted from Allen Brainspan imaging (Ziller *et al.* 2015). CP: cortical plate; SP: subplate; SVZ: subventricular zone; VZ: ventricular zone; MZ: marginal zone; VTL: lateral ventricle; CGE: caudal ganglionic eminence; MGE: medial ganglionic eminence.b) Schematic of cell type transistion, lamination and differentiation through corticogenesis. Radial glia (RG) both self-renew and differentiate to Cajal Retzius (CR) or intermediate progenitors (IP). IPs maintain the ability for self-renewal and differentiate to corticothalamic projecting neurons (CThPN) primarily in layer (L6), subcerebral projecting neurons (SCPN) primarily in L5, stellate neurons (StN) primarily in L4, and cortical projecting neurons (CPN) primarily in L2-3. Later born RG may migrate dorsally into the outer subventricular zone (OSV) where they are known as outer RG (oRG) and develop potential to generate glial cells such as astrocytes (Ast) or microglia (Mic), or remain more ventricular in the inner subventricular zone (ISVZ). Estimated timing of corticogenesis in shown below, with mouse corticogenesis timing in shown on top, and human coticogenesis timing on bottom.

mTOR pathway have been implicated in oRG formation and self-renewal[48,49]. After cortical layer formation, RG eventually self-consume into pairs of neurons.

To fully dissect the epigenomic dynamics of corticogenesis, a robust model system is needed. A model system must be both faithful to the subject of study, as well as mutable. Major considerations persist in our ability to understand cortical development both in terms of what may go awry in neurodevelopmental disorders, and what leads to the human-specific expansion of the cortex.[50]. However, the necessary reductionist study to uncover the epigenomic landscape responsible for cortical layer stratification faces major hurdles. First and foremost is sample rarity; human and non-human primate fetal tissue is difficult to obtain. Mouse models lack several key cortical sub-regions and cell types, including the more elaborate organization of progenitors — namely the OSVZ and the oRG found within[45]. In addition, the developing human cortical plate and subplate, containing CR cells, have distinct cell subtypes missing in mouse[33,34]. Regions of accelerated mutation since human divergence from chimpanzees reveal the importance of non-coding regions. 92% of human accelerated regions (HARs; 663/721 HARs) fall outside of transcribed sites, and are enriched for enhancer-like activity or transcription-factor binding motifs[12]. Further, these sites are seen to be active in early embryonic forebrain development[12,51]. Secondly there is also the need for genetic manipulation. Necessity and sufficiency are largely determined through gene knockout and rescue experiments — corticogenesis is no different. An emerging model system must allow for both genetic manipulation and recapitulate human-specific aspects of development.

Cortical organoids as a model of corticogenesis

Cortical organoids are self-organizing three-dimensional cultures that model features of the developing human cerebral cortex[52]. They are an adaptation of a 2D cortical “rosette” method that modelled early polarization of neuroepithelial cells and neural tube formation. Induction of human embryonic stem cells (hESC), or induced pluripotent stem cells (iPSCs) to the ectodermal lineage generates cellular aggregates called embryoid bodies (EBs). Neuroectodermal lineage priming is done through *in vitro* differentiation of stem cells in decreased basic fibroblast growth factor (bFGF) and a high dose of ROCK inhibitors to limit cell death[53,54]. From here the protocol deviates from the

2D cortical rosette method to allow for three-dimensional cortical layering. EBs aggregate and are cultured in suspension in a Neurobasal medium with additives to support neural progenitors and their progeny. Shortly thereafter, EBs are embedded into matrigel, an artificial extracellular matrix, which acts as a scaffold for cell migration. EBs expand in the matrigel to form organoids containing fluid-filled cavities reminiscent of brain ventricles, and buds of neuroepithelium that replicate early to mid-gestation of cortical development. Cortical organoids do not form blood vessels and thus as they expand to up to 4 mm in diameter, the diffusion of oxygen and nutrients to the core decreases. This leads to necrotic centers if grown in culture for multiple months. To mitigate necrosis, cortical organoid protocols all feature a form of agitation to facilitate movement of nutrient rich media through the organoids. To achieve this agitation, organoids are cultured in spinning bioreactors[55], or on orbital shakers, and previous groups have reportedly maintained organoids in culture for excess of 18 months[56]. The original organoid protocol did not include cortical region specification and was largely undirected, showing arealization of both the forebrain (*FOXG1*+), mid brain (*OTX1/2*+), ventral forebrain (*NKX2.1*+) and even retinal tissue. Developments in cortical organoids differentiation have revealed that they can be selectively induced to form different brain regions based on small molecule addition to the culture media. For instance, SMAD inhibitors such as dorsomorphin and SB-4321542 induce rapid neural differentiation to the dorsal forebrain state, while retinoic acid presence in early organoid induction is caudalizing[55].

Cortical organoids develop in a shorter time frame than native corticogenesis occurs, yet they follow the same cell differentiation progression. Exact times vary by protocol, however a generalized timing is as follows. Within 15-20 days *in vitro* (DIV15-20, where DIV0 is the original induction of stem cells to ectodermal lineage), cells form continuous neuroepithelia, surrounding fluid-filled cavities (similar to neural rosettes). Pluripotency markers *OCT4* and *NANOG* begin to diminish, while neural identity markers *SOX1* and *PAX6* increase[57,58]. By DIV30, a radially organized CP begins to form. This region is positive to pre-plate marker *TBR1*, and contains *RELN* expressing CR cells[57]. Bulk RNA analysis shows at this point that organoids closely resemble prefrontal cortex at GW (gestational week) 8-9[55]. Around DIV60-75, organoids exhibit rudimentary separation between early-born deep layer corticofugal neurons (*BCL11B*+) and late-born

superficial layer (*SATB2+*, *POU3F2+*), depending on protocol[52,55,56,59]. Additionally, it is around this time that the human-specific oRG cells (*SOX2+*,*HOPX+*) begin to populate[52]. Around DIV90-100, organoids become more closely correlated to fetal prefrontal cortex at GW17-25[55]. Progeny of oRG begin to form astrocytes (*GFAP+*)[55]. As organoids age further we begin to see the formation of dopaminergic neurons (*TH+*) and mature astrocytes (DIV180)[60]. Organoids in directed protocols that lack a ventral (*NKX2.1+* region) did not exhibit interneuron formation[52,55,56]. However those with the ventral marker showed late formation of interneurons. This is expected, given that interneurons are known to migrate from the ventrally located LGE/MGE during corticogenesis.

When cerebral organoids are generated from mouse embryonic stem cells, they lack both the IFL and oRG[56]. Further supporting evidence that organoid models can recapitulate RG behavior, is a study in which GFP was electroporated, followed by a pulse of BrdU to track lineage divisions in proliferating cells. The authors reported that daughter cells formed after the BrdU pulse chase included both RG and IPs, suggesting this cerebral organoids capture the asymmetric division potential of RG[45].

Organoid cortical models are not without limitations. In RNA comparisons between organoids and primary fetal tissue samples, organoids consistently enrich for genes associated with cellular stress, glycolysis, and electron transport pathways[48,50,52]. However, it has been demonstrated that this can be alleviated with culturing alterations and is likely induced in the early stages of ectodermal lineage priming of pluripotent stem cells. Organoids transplanted into a mouse cortex appropriated mouse oligodendrocytes and astrocytes, which led to decreased cellular stress signals[52]. New protocols have introduced vascularization processes to address glycolysis concerns as well[61]. Xenografted organoids show a higher correlation of radial glia maturation to primary sample age over organoid age[48]. Further the directed differentiation of organoids is not perfect. Mesodermal linage cells have been uncovered, despite early patterning to the neuroectodermal fate[60] and organoids are not homogenous in forebrain cortical area, with many showing both primary visual cortex (V1-like) and prefrontal cortex (PFC-like) signatures[48,52]. This is partially to be expected given the belief that thalamic input helps define areal signature[62,63]. Organoids tend to lack the diversity of cell subtypes that form over time in the human cortex[52]. Despite nuances in organoid

differentiation when compared to native human corticogenesis, this model system remains extremely promising for a battery of previously untestable hypotheses and closely resembles early corticogenesis. In Chapter 4 of this thesis, survey the changing epigenomic landscape of maturing organoids and provide comparisons to what is understood about the epigenome of fetal cortical development.

Single-Cell Chromatin Accessibility Assays

Motivation

The diploid human genome contains roughly 6.4 billion base pairs, totaling a distance of over 3.84 meters of linearized DNA per nucleus. To maintain nuclear integrity and limit search space for DNA binding proteins, cells compact their genomes such that only 1-4% of it is accessible at a given time[64]. This is achieved by the coiling of DNA around histone proteins and condensation of those resultant nucleosomes into larger macromolecular structures. The regions which remain accessible are highly enriched in genomic content relevant to regulating transcription and defining cell type and state, including gene promoters and enhancers[65]. A single-cell approach to measuring chromatin accessibility shares the same two previously stated benefits of single-cell analysis over bulk approaches. Namely, i) single-cell approaches allow for the unbiased interrogation of multiple cell types within a complex tissue sample, and ii) single-cell approaches provide a higher resolution of chromatin reconfiguration in actively differentiating systems than in bulk assays. Most variants uncovered in GWAS studies of neurodevelopmental disorders are located in non-coding regions, thus demonstrating the significance of assessing non-coding regulatory elements.[66]. Using a single-cell chromatin accessibility assay, we are able to uncover which cell types express these non-coding regions that are associated with disease states[67]. By tracking accessible sites in single-cells one can infer the activity of transcription factors, track the opening of enhancers, and infer their recruitment to promoter regions.

Method

In order to catalog the small sections of the genome that are accessible in each cell, several strategies have been developed. All strategies share the common through-line of leveraging the susceptibility of exposed DNA to insult when compared to compacted DNA. All assays are based

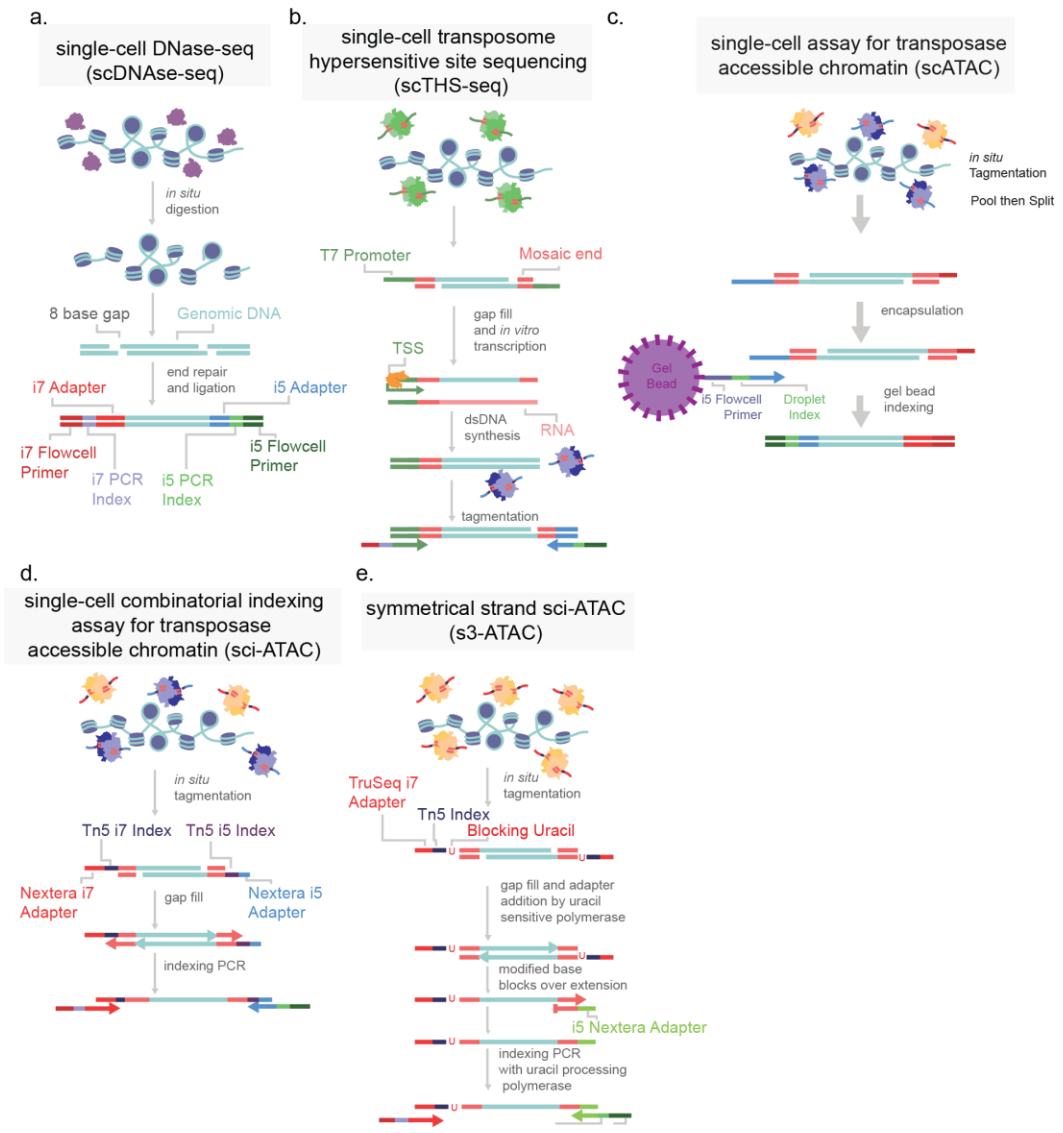


Figure 4. Single-cell chromatin accessibility assays. a) Single-cell DNase-seq uses DNAse I (purple) digests open chromatin, adapters are subsequently added. b) Single-cell transposome hypersensitive site sequencing (scTHS-seq) uses a transposase (green) to introduce a T7 bacterial promoter region to open chromatin and amplify DNA through an RNA intermediate via *in vitro* transcription (orange). c) Single-cell assay for transposase accessible chromatin (scATAC) uses two species of transposase to introduce i5 and i7 adapter sequences. Cells are then encapsulated in droplets with an oligonucleotide coated gel bead to uniquely index each cell. d) Single-cell combinatorial indexing assay for transposase accessible chromatin (sci-ATAC) uses two species of transposase (purple and orange) to introduce two adapters directly into open chromatin. e) Symmetrical strand sci-ATAC uses a single species of transposase and subsequent adapter switching strategy to amplify open regions, further detailed in Chapter 2. Labeled DNA oligonucleotide colors are consistently colored across panels.

on the premise of fragmenting the more vulnerable DNA and the subsequent capture of fragments for sequencing library preparation.

DNase-based methods were until recently the most prominent method, being used in the Encyclopedia of DNA Elements (ENCODE) project. In this approach, the genome, while in its native state, is treated with DNase I, a protein that can digest both single and double stranded DNA (Figure

4). Enzymes are limited by their protein footprint so condensed heterochromatic regions are sterically protected from enzymatic action. The fragmented DNA can then be captured, and sequencing adapters appended for massively parallel sequencing (Figure 4a)[64]. While DNase approaches have been adapted to a single-cell format, this method remains difficult to titer. Changes in DNase I concentration or incubation greatly affect library quality[68]. An alternative method is THS-seq, wherein hyperactive transposase (the protein Tn5) is loaded with a bacterial T7 promoter region and tagmented into regions accessible to the Tn5, again using steric hindrance to select for open regions (Figure 4b). Tagmented DNA is isolated and *in vitro* transcription is used to amplify regions via the added T7 promoter region. RNA intermediates reflecting the open regions of chromatin are then reverse transcribed and sequencing adapter are added[32].

By far the most widely used assay for accessible chromatin is ATAC-seq (Assay for Transposase Accessible Chromatin using sequencing)[16]. This method uses the same Tn5 protein used in THS-seq, but uses a simplified workflow. The Tn5 enzyme is, as previously mentioned, sterically limited to regions of open chromatin. ATAC-seq involves loading the enzyme with adapters necessary for PCR and sequencing. At open regions, the Tn5 enzyme both fragments the genomic DNA and appends the PCR adapters in the same reaction (Figure 4c). The excised accessible DNA can then be amplified selectively by using complementary primers. This assay is far more efficient at the capture of open genomic regions than the other approaches and has been adapted to optimize cell isolation and tagmentation conditions[69]. Recently commercialized versions of single-cell ATAC was made available which encapsulates cells or nuclei within microfluidic droplets[29,30]. Another means of single-cell ATAC popularized by us and others is sci-ATAC which uses the aforementioned split-pool barcoding approach (combinatorial indexing) (Figure 4d)[31,67]. Combinatorial indexing is

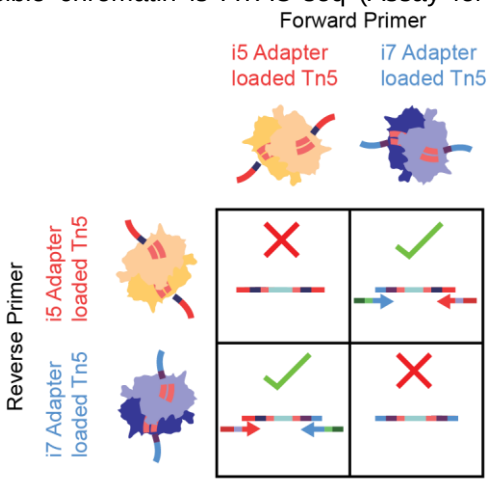


Figure 5. Tagmentation with two separate adapter-loaded Tn5 species has loss in effiency. In a captured molecule, i5 and i7 adapters must be added in the proper orientation for PCR. i5-i5 (top left) and i7-i7 tagmentations (bottom right) are not sequencable, despite the genomic regions being open.

performed through the addition of indexes both at the multiplexed tagmentation stage, and at the final PCR stage. This method is beneficial in that the number of assayable cells per preparation scales exponentially with increasing index combinations. In Chapter 3, I describe our adoption of both sci- and microfluidic platforms to greatly increase the throughput of both systems, essentially capturing multiple cells per droplet leveraging indexed tagmentation. sci-ATAC libraries have uncovered a trove of regulatory information, however the number of captured fragments is inherently limited. To successfully capture a fragment in PCR, it must have the proper tagmentation of both i7 and i5 adapters. This means that ~50% of fragments are lost as i5-i5 or i7-i7 tagmentations (Figure 5). In Chapter 2, I describe a correction to this strategy through the use of single Tn5 species and an adapter switching strategy, named symmetrical strand sci ("s3", Figure 4e). The above summary demonstrates that all protocols show a commonality in the generalized goals of both the fragmentation and capture of unprotected genomic regions. Consequently, the information gathered by all assays is similar in that it is essentially a count of

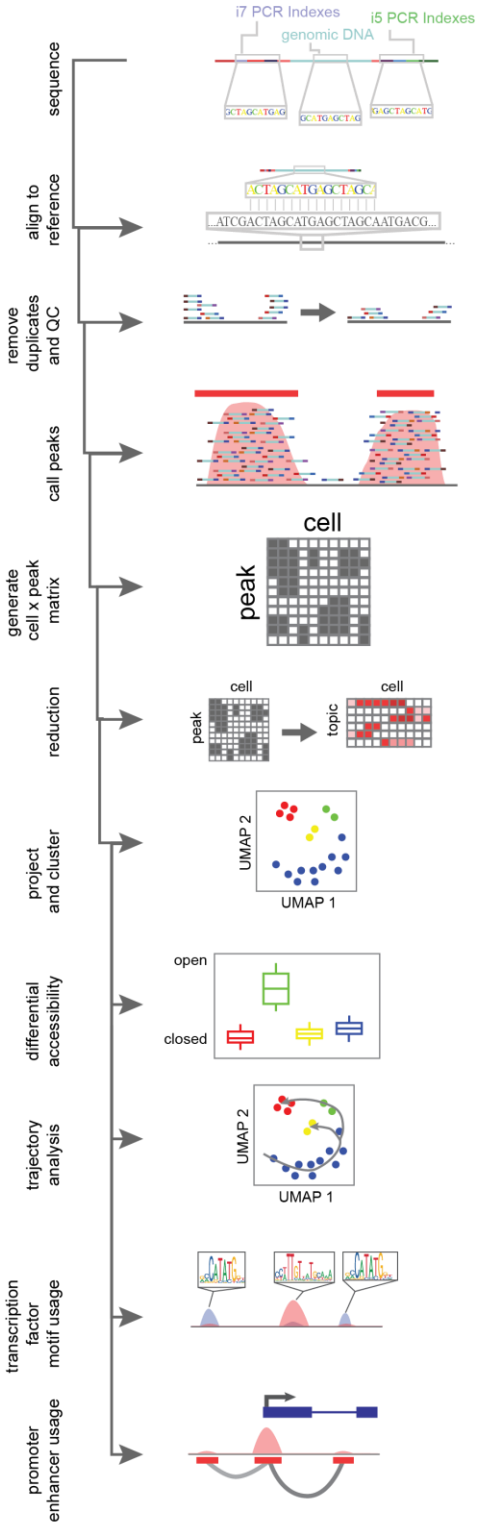


Figure 6 Flow-through of single-cell ATAC-seq data analysis. Reads are generated through sequencing, aligned to a reference, de-duplicated and filtered based on quality control metrics, read pile-ups along the genome are called, then a counts matrix of cell identifier by read count per peak is generated. This counts matrix is then reduced in dimensionality, and clustered and projected into 2D space. From there cluster aggregates (all cells combined within a cluster) have the power for differential accessibility analysis, and can be used for trajectory analysis, transcription factor motif usage and the assessment of cis-coaccessible networks for promoter-enhancer interactions.

captured genomic regions overlapping with a reference genome.

Analysis

Single-cell chromatin accessibility data is count data. Single-cell ATAC-seq methods are by far the most widely used and their analysis will be detailed below; however, similar analysis can be performed with any of the above listed alternative protocols. Genomic DNA fragments captured in the assay are sequenced and aligned to a reference genome by an alignment algorithm (Figure 6)[70,71]. Reads which overlap in alignment ("pile-ups") are used to define discretized regions of open chromatin, essentially assuming that the chromatin accessibility protocol is biasing sequence capture to unprotected regions. The calling of discrete open chromatin regions, or peaks, is done with an peak-calling algorithm, like MACS2[72], which uses a Poisson distribution of reads across the mappable genome in a sliding window of bins. If there are more read counts in a region than expected by this null hypothesis, a peak region is called and an open region of the genome is uncovered[72]. These peaks are then used to "bin" the genome into sites with evidence of accessibility. Single-cell ATAC-seq methods apply peak-calling on the full data set, not accounting for single cells, since any given diploid cell can have at most four captured reads at a given base (two copies of each top and bottom DNA strands). Once peaks are called on the entire data set, cell identity is mapped back to individual reads via the cell identifier (the unique combination of indexes) to generate a sparse cell x peak matrix[31,67], populated by the number of reads per cell aligning to an open region.

Cells are then grouped together based on similarity of peak coverage to overcome single-cell data sparsity. Natural language processing approaches like latent semantic indexing (LSI) apply a weighting schema where peaks more commonly used are decreased in importance[73]. Alternatively, machine-learning approaches such as the latent Dirichlet algorithm (LDA) is used to generate "topics" or groups of peaks commonly seen together within the data. From there the cell x peak matrix is reduced from hundreds of thousands of peaks to a couple of dozen topics, where the number of topics scales with the complexity of the data set. This addresses both the data sparsity of single cells and captures biological information within peaks, wherein shared open sites tend to be enriched in common transcription factor motifs or linked to biological ontology[74].

Following dimensionality reduction, cells are grouped together based on their shared topic weighting by Louvain based clustering algorithms[75]. Cells are projected into two dimensional space via a machine learning algorithm like uniform manifold approximation and mapping (UMAP) or t-distributed stochastic neighbor embedding (tSNE)[76].

Following the unbiased clustering of cells, differences in peak usages between clusters are assessed by use of logistic regression tests. Additionally, the activity of transcription factors can be inferred per cell, based on the expression of transcription factor specific DNA binding motifs. If each peak with reads for a cell is binarized, transcription factor activity can be inferred based on the overrepresentation of motifs present in open sites[77,78]. Given that enhancers and promoters are recruited in a concerted effort to drive transcription, this implies that the accessibility of both promoters and enhancers should co-occur if a site is acting in an enhancer-like function. To assess this agnostically within a data set, we look for the co-occurrence of accessibility in local enhancers linked to a peak region overlapping a known promoter. Cis-co-accessible networks (CCANs) are anchored at the promoter peak, and generated through correlation to other accessible nearby peaks for each cells with proper coverage. This network of enhancers and gene promoters better correlate with gene transcription as compared to either promoter accessibility alone or average gene body accessibility[20,79]. This is possible through the statistical power generated by so many independent samples made in single-cell library preparation. In order to leverage single-cell data to assess cell differentiation or epigenomic shifts, we can order cells in reduced dimensionality space and calculate a minimal spanning tree, or L1-graph, which traverses across cells, minimizing the residual distance from the tree. This allows for ordering of cells in order to infer programmatic shifts in the epigenome during cell state shifts[26].

In recent works, whole organism atlases have been generated on human and mouse development[67,80]. While not focused directly on corticogenesis, these data sets reveal the waves of transcription factor motif accessibility changes as stem cells progress towards maturing neurons. As excitatory neurons mature in the human cortex, there is a marked opening of Rfx and Tal-related transcription factor binding sites (*e.g.* RFX2, TWIST2, NEUROD1) and a closing of early radial glial marker sites like SOX2 and POU factors (*e.g.* POU2F1), reflecting a concordance with known

transcriptomic changes[67]. Further, chromatin accessibility across cortical neurons reflects the spatial organization of cortical layering in the murine brain[80]. However, many questions of chromatin dynamics, RG division, fate specification, and regulatory network formation persist that require a focused approach. In Chapter 4, I detail epigenomic changes during the maturation of cortical organoids as a model system of early human cortex development.

Single-cell Whole Genome Sequencing

Motivation

The genome contained within each nucleus of an organism is exceedingly static, with an estimated frequency of mutation in somatic cells around $5x10^{-10}$ single nucleotide variants per bp per division. This rate of genomic change is bolstered by transposable elements and microsatellite instabilities[81]. However, cancers such as pancreatic ductal adenocarcinoma (PDAC) and triple negative breast cancer (TNBC), undergo marked genomic instability and may possess hypermutator phenotypes that can generate subclonal mutations during a tumor's lifetime[82]. These phenotypes form when DNA repair mechanisms are disrupted. Single-cell whole genome

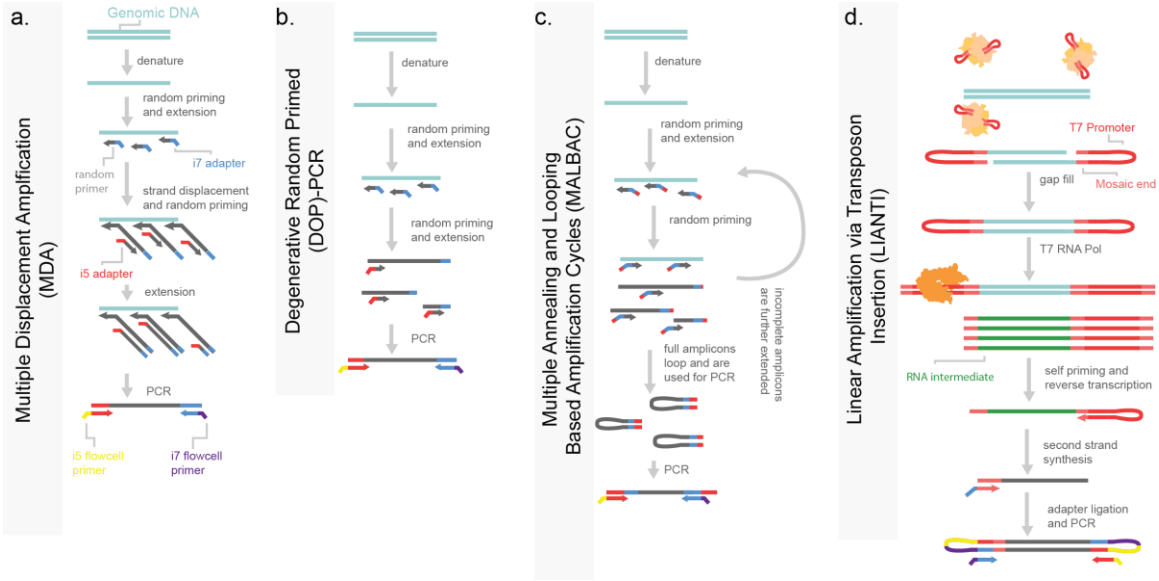


Figure 7. Pre-amplfication based single-cell whole genome sequencing (WGS). a) Multiple displacement amplification (MDA) is performed through the random priming across the whole genome, with subsequent random priming and amplification through a highly-processive polymerase. b) Degenerative random primed PCR, uses random priming to amplify the genome prior to PCR. c) Multiple annealing and looping based amplification cycles uses a random priming strategy with a specialized adapter which will self-hybridize to form intra-molecular hairpins. This self-annealing sequesters these molecules from further amplification. These hairpins are then PCR amplified. d) Linear amplification via transposon insertion (LIANTI) uses a Tn5 enzyme (shown in tan) to tagment DNA and insert a T7 promoter. This promoter is then used to *in vitro* transcription with an RNA T4 polymerase (shown in orange) to amplify the genome through RNA intermediates, which are then processed by reverse transcription, second strand synthesis and adapter ligation prior to PCR. Molecule coloring is consistent across panels.

sequencing is useful in these cases[83]. Heterogeneity in the tumor genome, such a copy number aberrations (CNAs, amplification or deletions of genomic regions) can be used to order the events of cancer progression, identifying prognostic markers and secondary mutations[83]. Darwinian selection can work within rapidly expanding neoplastic tumors, selecting for favorable mutations. Likewise, hypermutator phenotypes have been seen to be lost once cancer cells find local optima in fitness[82]. Additionally, single-cell analysis accounts for tumors with low cancer cell fraction or impure biopsy results, allowing for the distinction between tumor and unaffected somatic cells.

Method

Single-cell whole genome sequencing (scWGS) methods have three major criteria for success. They must capture the genome with high fidelity such that mutations can be faithfully called. They must have high coverage of the genome, in order to provide the statistical power to call copy number changes at high resolution. They must have unbiased coverage, such that there is a large signal-to-noise ratio. With these criteria in mind, the existing methods for scWGS are summarized below. Early protocols were focused on amplifying genomic DNA prior to library generation. Multiple-displacement amplification (MDA) uses random priming via degenerative nucleotides in an attempt to capture the genome in an unbiased fashion (Figure 7a). Amplicons captured across the genome are further amplified by the use of a highly processive polymerase like phi-29 with the generation of 1-2 µg of DNA (far exceeding the 6 pg contained within a cell)[84]. Despite the exceptionally high coverage, MDA produces significant biases due to the multiple rounds of PCR, making the detection of small CNAs difficult. Early examples of MDA were used to support a hypothesis of punctuated evolution in tumor progression in triple-negative breast cancer. This was done by measuring both the cells within the primary tumor and a subsequent liver metastasis[83]. Degenerative oligonucleotide primed polymerase chain reaction amplification (DOP-PCR) is a similar attempt at a random-priming strategy, with limited run-away amplification (Figure 7b). However, this method suffers from low coverage and substantial dropout. A mixture between these two methods, MDA and DOP-PCR, emerged in which PCR amplicons self-sequester after amplification by formation of a thermodynamically stable intramolecular loop in a method named MALBAC (multiple annealing and looping based amplification cycles). This makes PCR

amplification “quasilinear” rather than exponential and leads for high coverage and less capture bias. However, MALBAC is not without limitations, the enzyme used for amplification (*Bst* large fragment) is error prone and the multiple rounds of linear amplification limit cellular throughput and increase cost (Figure 7c)[85]. LIANTI (linear amplification via transposon insertion) uses a transposase to introduce a T7 bacterial promoter region across genomic DNA. The T7 promoter is used for *in vitro* transcription genome wide, increasing the amount of material that can be generated per cell (Figure 7d). The RNA intermediate is then captured and converted to DNA libraries to be sequenced. While this method matches MDA in the amount of library material generated per cell, it has the shared limitation of uneven coverage. Further, error prone transcription exacerbates library quality, lowering the ability to call exacerbating library quality is the error prone intermediate states which decrease fidelity to the genome. This method has been developed for sci-compatibility as well wherein indexes are incorporated with a Tn5 tagmentation step[86].

Direct library preparation (DLP) avoids intermediate molecules by using Tn5 tagmentation to fragment the genome and introduce adapters for PCR, much like ATAC-seq (Figure 8a). In DLP,

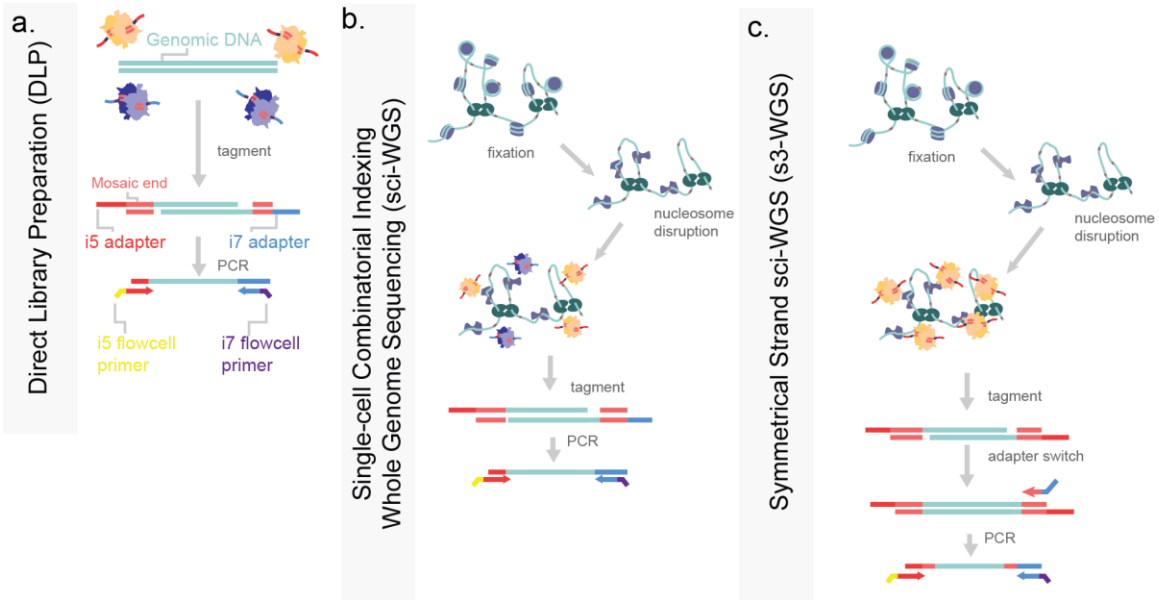


Figure 8. Tagmentation-based strategies of single-cell whole genome sequencing. a) Direct library preparation isolates single-cell genomic DNA in a well prior to full protein degradation. Purified genomic DNA is then tagmented with Tn5 enzymes loaded with i5 and i7 sequencing adapters (tan and purple respectively). b) Single-cell combinatorial indexing for WGS (sci-WGS) performs a fixation and nucleosome disruption *in situ* to render the genome accessible to tagmentation while maintaining nuclear integrity. The resulting tagmented DNA is then PCR amplified. c) Symmetrical strand sci-WGS uses similar pre-processing steps to sci-WGS, with the modification of Tn5 tagmentation such that library capture efficiency is higher (Chapter 2).

proteins associated with the genome are first denatured or digested prior to tagmentation, allowing for full accessibility. This method is simple and does not require pre-amplification, meaning a low rate of error introduction, and even coverage[87,88]. Evolutionary dynamics are prevalent in tumor samples, with known driver mutations often fixed in cancer cells. In parallel our group developed a similar method for higher cell count throughput. In this, we applied a similar approach to DLP using Tn5 for a sci-WGS. In order to denature genomic-associated proteins while maintaining the nuclear integrity needed for split-pool indexing, we used formaldehyde-based cross-linking and denatured with the detergent sodium dodecyl-sulfate (SDS; Figure 8b)[89]. The formaldehyde fixation maintained nuclear integrity while SDS denatured proteins, allowing for even tagmentation across the genome. In Chapter 2, I detail and adaptation to this strategy using the same s3- adapter switching strategy, increase genomic capture rate per cell to over 25% (Figure 8c). This increase in coverage allows for higher resolution assessment of copy number changes, and greater insight into which genes are driving cancer progression.

Analysis

scWGS has the potential to capture both single-nucleotide variants (SNVs) between cell lineages as well as copy number aberrations (CNA). The read out for scWGS, just like scATAC, is count data. Reads captured are aligned to a reference genome (Figure 9). Depending on read count, the genome is then commonly segmented into “bins” or non-overlapping windows. Bins are used to aggregate data, allowing for enhanced statistical power in determining shifts in read counts, as well as to account for genomic biases. In the simplest form bins are a set length[90]. In other approaches bins are defined by a read count threshold, meaning each bin has the same number of reads by different lengths[91,92]. One common instance of genomic bias is the uneven dispersion of Guanine-Cytosine (GC) content. This is known to affect PCR efficiency and thus could bias results if left unaccounted. scWGS CNA callers work to normalize bins by one of two procedures. Either they use a set of cells known to be without CNAs to build a model for expected read counts per bin such as seen in SCOPE[90], or they perform a normalization procedure to estimate ploidy. Normalization procedures include locally weighted linear regression (LOESS), or a modal regression followed by a *post hoc* means to estimate ploidy like in the tools Ginkgo[91] and

HMMcopy[87]. Following normalization across bins, the genome is then segmented to find where CNAs occur. Segmentation across bins occurs through either circular binary segmentation (CBS)[90,91] or hidden Markov model (HMM)[87]. Both methods generate breakpoints, wherein bins shift from one state (*e.g.* diploid) to another (a deletion of amplification). Following this, cells can be hierarchically clustered to infer the phylogeny of CNAs. Further, SNVs can be attained per cell using variant-call tools, such as GATK[93]. From this data, haplotypes can be generated by shared changes. CNAs can be supported by observing shifts in the minor allele fraction; for instance if the minor allele fraction in a clonal population goes to 0, that supports a loss of heterozygosity event. Additionally clones can be hierarchically clustered by shared SNVs, similarly to CNAs[87]. In a breast cancer sample, Laks *et al.* uncovered a fixed amplification of oncogenes *MYC*, *MCL1* and *CCNE1*, clonal loss of heterozygosity of *BRCA2*, and subclonal amplifications of *RAD18* and *RAB18*. Subclonal alterations in tumor suppressors and oncogenes has the potential to inform precision medicine and our understanding of metastases[87]. Though

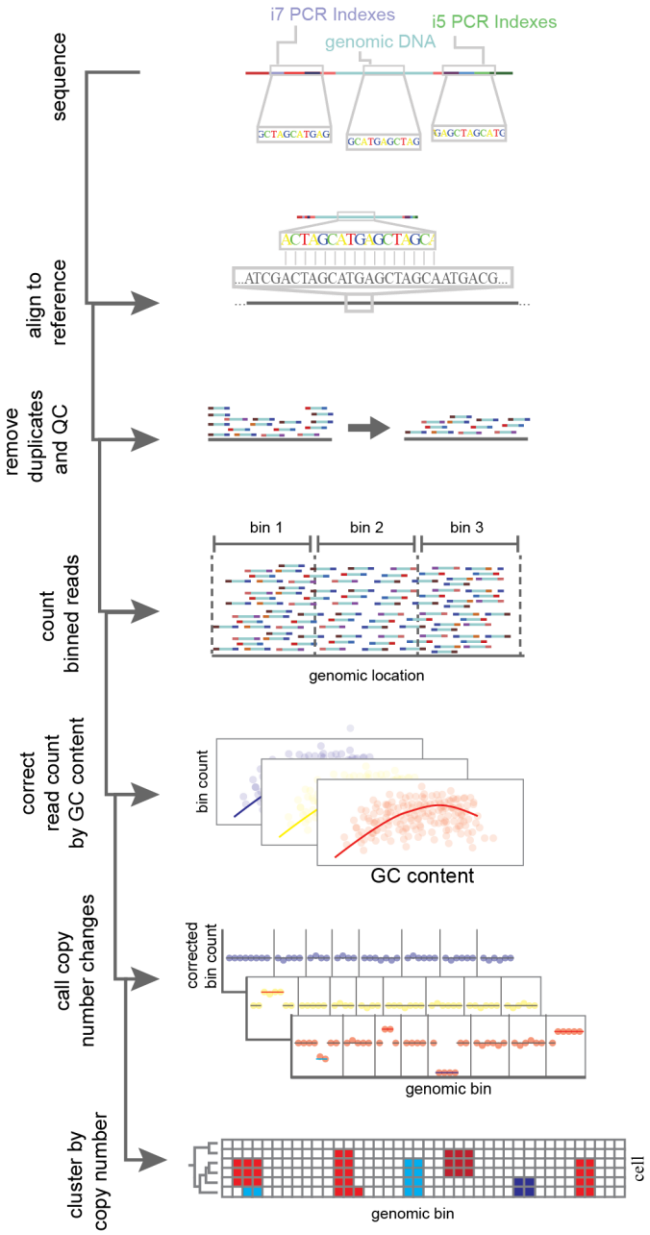


Figure 9. Analysis of copy number aberrations (CNAs) through single-cell whole genome sequencing. Reads are aligned to a reference genome after sequencing. Reads then undergo a quality control filter wherein low confidence mapping of reads and PCR duplicates are filtered out. For each cell, reads passing quality control are then counted within bins across the reference genome. Read count per bin in corrected for confounding factors such as GC content. After that normalization step, genomic bins undergo a segmentation where copy number changes are called based on changes in bin read count. Finally cells are grouped together based on shared CNAs to infer a lineage.

single-cell methods are not necessary to detect these mutations, *per se,* the accuracy of population frequency and the co-occurrence of mutations within cells, provides a more confident picture of cancer progression.

Single-cell Chromatin Conformation Capture

Motivation

Compaction of the human genome is known to play an important role in regulation of transcription. Genomes form a hierarchy of three-dimensional organization, including chromosome territories, A/B compartments associated with epigenomic markers, to smaller topologically associated domains and single chromatin loops[94–96]. Bulk assays for the capture of genome-wide conformation generate contact probabilities – averaging interactions over millions of cells[97]. However, it is now known through FISH that cells display variable genome and chromosome conformations, even when cells are genotypically and phenotypically identical[98]. Additionally, FISH and spectral karyotyping (SKY) remain the most common single-cell methods for uncovering genomic mutations such as translocations and inversions, which are hard to uncover with count data from whole genome sequencing or microarray. Efforts to advance single-cell chromatin conformation capture assays are being made to breach this gap[99]. Therefore, chromatin conformation assays have promising insight to both genomic rearrangement and genomic regulation.

Method

Two critical components for the unbiased capture of chromatin conformation within single-cells are paired-end sequencing and proximity ligation. The majority of single-cell chromatin conformation capture techniques are based on the principles of a bulk implementation of a method termed Hi-C, with adaptations for capture efficiency. In a bulk chromatin capture experiment, the chromatin of a sample is isolated and DNA is cross-linked with genome-associated proteins via a fixative such as formaldehyde[100]. Once covalently linked, the DNA is digested with a promiscuous restriction enzyme leaving “sticky-end” DNA 5’ or 3’ overhangs. Enzymes used are selected for a high frequency of occurrence across the genome and usually have small (and thus more likely occurring) recognition sites. These exposed base pairs are complemented with the addition of

biotinylated nucleotides and the filled-in blunted ends of are ligated together via a ligase. Since DNA is fixed, it is essentially pinned to proteins and nearby DNA, therefore, the ligation of digested DNA is biased towards other strands nearby in three-dimensional space. Following ligation, DNA is fragmented, purified through a streptavidin-biotin purification, and prepared for sequencing[17,101]. Paired-end reads, allow for the mapping of the two ligated chimeric DNA fragments independently. From this, the two fragments are mapped to separate regions of the genome reflecting their physical proximity in three-dimensional space. Increases in the efficiency of Hi-C reactions lowered necessary genomic input, eventually to down to single-cell level. In early proof-of-concept reports of this assay, the reactions for fixation, restriction digestion and ligation were performed *in situ* within nuclei[102]. Use of the nuclear compartment led to increased signal-to-noise as compared to early Hi-C strategies[97]. This was then followed by a purification and second digestion to linearize DNA, prior to a blunt end adapter ligation and PCR[103] (Figure 10a). sci-Hi-C was developed in order to improve on the cell count throughput of chromatin confirmation capture assays. This was done by introducing combinatorial indexing into the sample processing at gap-fill in biotinylation stage[104,105] (Figure 10b). An alternative method, named Dip-C, is aimed at the capture of haplotypes within single cells[106] (Figure 10c). Dip-C omits the biotin-streptavidin pull-down and adds a whole-

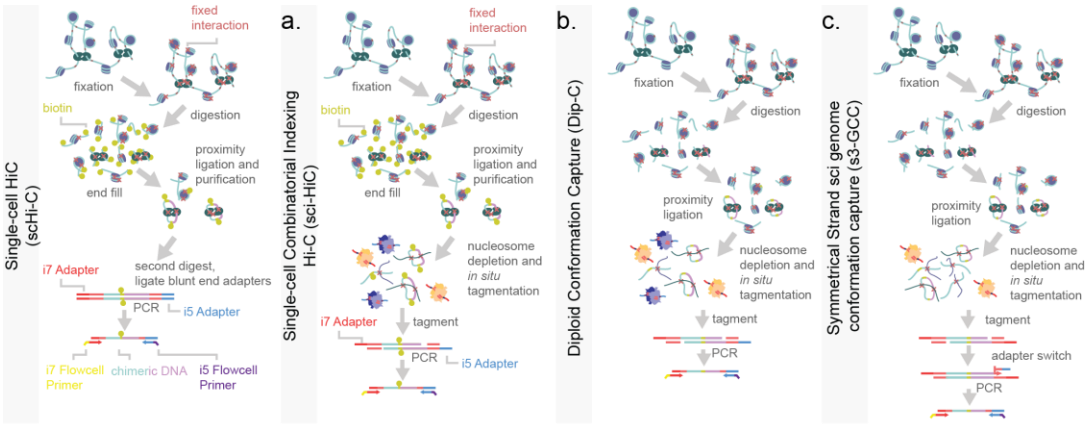


Figure 10. Schematic of single-cell chromatin conformation assays. a) Single-cell Hi-C protocol, including *in situ* fixation and multiple rounds of restriction digest to generate libraries. A biotin-fill in (yellow circles) is performed to allow for both proximity ligation (a low temperature blunt end ligation of nearby fragments) and the selective pull-down of biotinylated DNA during a subsequent clean-up step. Adapters are then blunt-end ligated prior to PCR. b) Single-cell combinatorial indexing Hi-C (sci-HiC) uses *in situ* fixation to set chromatin conformation and restriction enzyme digestion to introduce sticky ends. After biotin fill-in and purification, DNA is then tagmented and PCR amplified. c) Diploid conformation capture (Dip-C) uses a similar strategy without a biotinylation pull-down. d) Symmetrical strand sci- genome conformation capture (s3-GCC) uses a shared strategy with Dip-C wherein there is no selective pull-down for restriction digested DNA. However the more efficient s3 chemistry is used to increase information content garnered per cell (Chapter 2). Color labeling is consistent across panels.

genome amplification similar to LIANTI[107]. This approach captures more of the genome and more contacts per cell than sci-Hi-C, but is limited in cell count throughput. A low cell count but high coverage adaptation of Dip-C has been used to analyze contact maps in murine olfactory bulb and retinal rod receptors. In this study authors used few cells (409) with a relatively high number of chromatin contacts (median of 252,000) to separately cluster gross cell types. By leveraging additional epigenomic knowledge such as methylated regions, they uncovered that methylation frequency correlates to distance from nuclear center. This phenomenon inverts as retinal rod cells mature. Interestingly, because this data was generated on single cells, the authors were able to uncover some of the enhancer-promoter recruitment and variability in the developing population, and track the progress of euchromatic inversion. While promising for insight into gene regulation and chromatin conformation through development, this method is still in its infancy with a need to improve both cell count and contact captures per cell for greater statistical power[106].

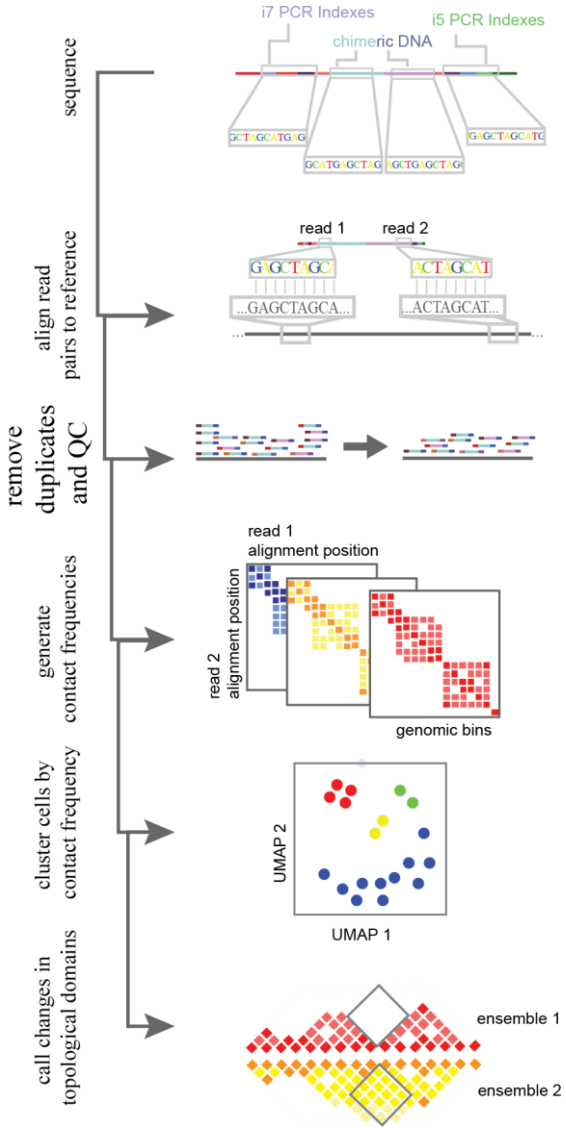


Figure 11. Analysis flow-through for single-cell HiC-like data. Chimeric library molecules are aligned to the genome with read 1 and read 2 (two reads generated through paired end sequencing) are aligned separately. Following alignment, reads are filtered for quality both in alignment confidence and the removal of any PCR duplicates. After this, a square matrix is made per cell for contact frequencies. The position of read 1 and read 2 determine the x and y axis of the matrix, and the resultant bin is the count of occurrences. Cells are then grouped together and projected into 2D space based on the normalized contact frequency matrices through a machine learning algorithm. From this grouped cells are combined for statistical power. Higher resolution ensembles are then used to determine pairwise changes in topological domains, reflecting regulatory changes or large-scale mutations.

Analysis

In chromatin conformation data analysis, each end of paired-end reads are mapped independently to the reference genome and uniquely captured molecules are counted (Figure 11).

The genome is binned as in single-cell whole genome processing. A square counts matrix of contact frequencies (read 1 alignment locus X read 2 alignment locus) is then generated per cell. This matrix is sparse, leading to the construction of algorithms for imputation. Two current methods built specifically for grouping single cells based on their contact frequency matrix, are *scHiCluster* and *HiCRep/MDS*[108]. Contact frequency bins are locally aggregated via a linear convolution or sliding window. In *scHiCluster*, this is then smoothed by a random walk algorithm being treated as weighted network. Alternatively, for *HiCRep/MDS*, the smoothed contact matrices are summarized as weighted similarity measures as in *HiCRep*[108]. The resulting matrices are clustered and projected by dimensionality reduction (with an approach such as multidimensional scaling, MDS) into two dimensional space[109]. Following this topological domain boundary differences between clusters can be attained through merging similar cells and analyzing the "pseudobulk" data through *TopDom*[110]*.* Similar to *HiCRep*, *TopDom* uses a sliding window of up and downstream bins to define genomic regions with fewer locus-locus interactions than other regions around the local genome. This faithfully recapitulates the A/B chromatin compartments seen in bulk data[109]. The variability in topological domains within cell populations requires further study. Early work suggests interesting mechanisms of genomic reconfiguration within 3D space that could provide insight into transcriptional recruitment machinery and the interaction of different epigenomic markers in physical space[106]. To improve on this method, and to provide a means of improved CNA detection, I developed a method of single-cell genome conformation capture using the s3 chemistry described previously in this thesis (s3-GCC, Figure 10c). In Chapter 2, I describe the methodological improvements that lead to both whole genome and HiC-like read-outs from the same cell. This allowed us to identify subclonal translocations in patient-derived cancer cell lines.

Single-cell Methylation

Motivation

DNA methylation, the covalent addition of a methyl group to cytosine, is known to have critical roles in gene regulation and modifying transcription factor binding affinity. Its role in gene silencing and genomic imprinting is also well studied[111,112]. Genomic methylation occurs primarily on the approximately 1 billion cytosines in the genome almost exclusively in the context of cytosine-

guanine dinucleotides (mC, CG) for most cell types[111]. DNA methylation is correlated to gene expression[113,114] and reflects cellular identity[115]. DNA methylation has also been linked to neurodevelopmental disorders in the human frontal cortex[116]. Notably, methylation also occurs at non-CG dinucleotides and is referred to as CH methylation (mCH, H = adenosine, thymine, or cytosine). This is occurs at high levels in embryonic stem cells and mature neurons, though at different trinucleotide patterns, namely CAG for stem cells and CAC for neurons[111,117]. In mature neurons the amount of mCH exceeds that of mCG during synaptogenesis, roughly four weeks after birth in mice or two years after birth in humans[114,118,119]. Remarkably, gene body mCH levels in neurons negatively, yet strongly, correlates with transcriptomic expression and are useful for cell type identification[19]. In bulk methylation profiling of cortical organoids, Luo *et al.* were able to capture the transition of dominant mCH from CAG to CAC during the transition from neuroepithelial cells to mature neurons, suggesting a point of methylome transition from stem-like to neuronal-like[118]. This provides both a model system and a key time-point for future analyses of mCH levels and their regulation[120]. Organoids were observed to have changes in methylome profiling from fetal cortex. These changes manifest as differential methylation across extracellular matrix genes (possibly due to the inclusion of matrigel in culture) and hypomethylation around pericentromeric regions (a previously reported phenomenon for induced pluripotent stem cells)[120].

In the native methylation, reduction of mC is catalyzed by the Tet family of mC hydroxylase proteins, converting the methyl- moiety to hydroxymethyl-, formyl-, and carboxyl- progressively. Hydroxymethylation (5hmC) occurs almost exclusively in the CG context and accumulates in mature neurons. In neurons, 5hmC is known to be enriched near constitutively expressed promoter regions[114]. To this day the role of 5hmC is understudied, likely due to the inability to distinguish mC and 5hmC by the most commonly used assay for methylation, bisulfite conversion. Two reports through alternative assays demonstrate a ratio of 5hmC to mC of 30-50% in mature excitatory neurons[114,121]. Alternatively, new enzymatic methods have been described in which APOBEC3A, a natively expressed deaminase induces direct cytosine deamination in an *in vitro* reaction[122]. To date, this method has not been published as a single-cell protocol, however, it does have a promising adaptation for assaying the understudied moiety 5hmC[121].

Method

Methylation profiling genome-wide is achieved by the selective mutation of non-methylated cytosines. Sodium bisulfite is applied to genomic DNA which effectively deaminates non-methylated cytosine to uracil, through a three-step reaction. Importantly, uracil complements with adenine, which means subsequent library amplification will report non-methylated cytosines as thymine. Through this point mutation in the reference genome, namely thymine where cytosines were expected, methylation profiles can be inferred (bisulfite-sequencing, BS-seq)[18]. The first reported protocol for single-cell methylation was scRRBS (reduced representation bisulfite sequencing, Figure 12a). This method uses a methylation-insensitive restriction enzyme (MspI) to digest genomic DNA prior to bisulfite conversion. MspI is used to enrich for CG-rich regions across the genome, via its cut site (5'C|CGG). The resulting sticky ends enriched at CG-rich genome regions are then adapter ligated, DNA is bisulfite converted, and sequencing libraries are prepared[123].

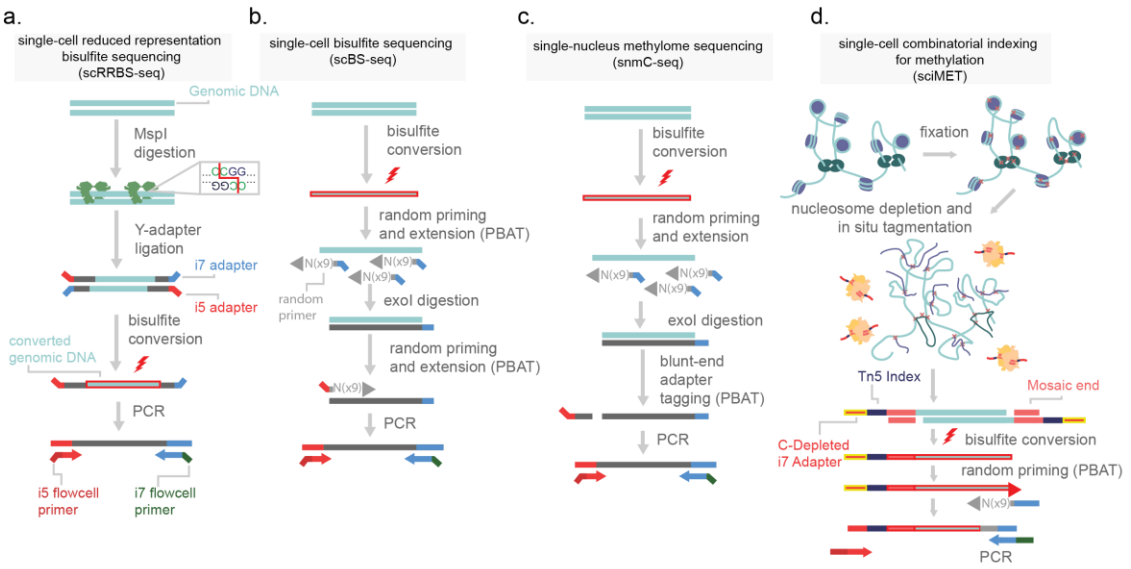


Figure 12. Methods for the generation of single-cell methylomes. a) single-cell reduced representation bisulfite sequencing (scRRBS-seq) digests purified genome DNA with restriction enzyme MspI. This enzyme cuts at a CCGG target sites, fragmenting DNA that is in CG rich regions. Y-adapters (pre-annealed i5 and i7 adapters) are then added on by ligation and the molecule is bisulfite converted. Following this, the DNA is then PCR amplified. b) Single-cell bisulfite sequencing bisulfite converts purified genomic DNA and then uses random priming for post-bisulfite adapter tagging. Prior to the second round of random priming, the reaction is incubated with exonuclease I (exoI) which digests single-stranded DNA. This removes excess primer from the reaction. Following this a second round of random priming is used to introduce the next adapter and the molecules are PCR amplified. c) Single-nucleus methylome sequencing (snmC-seq) is similar to scBS-seq but uses a blunt-end adapter tagging strategy. d) Single-cell combinatorial indexing for methylation (sci-MET) uses a C-depleted oligonucleotide loaded onto a Tn5 enzyme to tagment nucleosome depleted nuclei. Following this, cells are lysed, bisulfite converted, and a post-bisulfite adapter tagging strategy is used prior to PCR amplification (Chapter 1).

BS-seq is harsh and fragments genomic DNA. This is of high concern for scaling the assay to single-cell resolution. To avoid heavy losses of genomic capture, post-bisulfite adapter tagging (PBAT) is used. In PBAT library adapters necessary for PCR and sequencing are added to genomic DNA after BS conversion (Figure 11b)[19,115,124]. In this order of events, BS conversion fragments the genome and denatures DNA to a single-stranded state. Single-cell PBAT strategies such as scBS-seq introduce adapters after conversion through random priming, similar to the single-cell whole genome method DOP-PCR[124,125]. Secondary adapters are then added and libraries can be sequenced. An alternative approach, single-nucleus methylome sequencing (snmC-seq), uses a blunt-end adapter tagging strategy (Figure 12c)[19]. Cells are fully lysed by the bisulfite conversion chemical reaction, making this protocol difficult but not impossible to adapt to higher cell count strategies. In Chapter 1, I detail a new method for high throughput single-cell methylome library generation (sci-MET). In this method I use custom sequencing adapters and indexes depleted in cytosines. The lack of cytosines prevents BS conversion changing the indexes, allowing for the split-pool indexing necessary for sci- chemistry (Figure 12d).

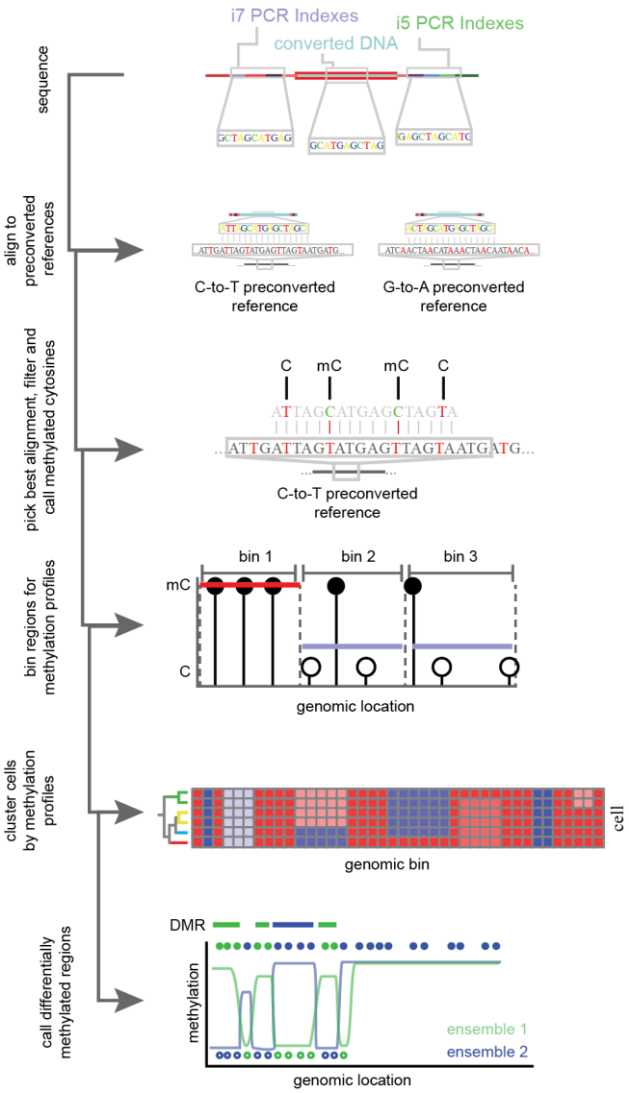


Figure 13. Simplified flow through of single-cell methylation analysis. Bisulfite converted and PCR amplified DNA is sequenced and aligned to pre-converted reference genomes. C-to-T and G-to-A conversions are performed to account for bottom and top strand library capture. For the most confident mapping location and strand, cytosines (C) and methylated cytosines (mC) are called based on point mutations induced in bisulfite conversion. Methylation profiles for cytosines are aggregated over genomic regions and used to group single-cells into clusters. Cells are combined within clusters for increased power and changes in methylation across the genome are calculated.

Analysis

Analysis of single-cell methylation profiles leverage the point-mutations induced through BS conversion. These mutations lead to decreased library complexity and can make reference alignment difficult. To account for this, special considerations must be taken. In one approach, the tool *Bismark*[126] generates four pre-converted reference genomes to account for the full bisulfite treatment of each possible strand of genomic DNA prior to running the short read sequence aligner *Bowtie*[127]. From this, base specific methylation of cytosines can be ascertained. Alignments with greater than 70% methylation of non-CG cytosines reported as methylated are generally removed from analysis as this suggests a read-specific failure of bisulfite conversion[19]. Following filtering, methylation rates (% methylated CG/all CG) are generated across genomic bins and used for dimensionality reduction and clustering. To account for depth of coverage, some strategies apply a *post-hoc* probabilistic binomial model, wherein region methylation rates are weighted by coverage[124]. Notably, for neuronal data, CH methylation rates performs better for discrimination of cell types than CG methylation rates[19]. Differentially methylated regions have been implicated as diagnostic biomarkers[128], and can be calculated between cellular clusters via two-sided t-tests (Figure 13)[129]. High throughout single-cell methods will allow for exploratory analyses of methylome changes across complex systems such as neurodevelopment or tumor progression. It is with this motivation in mind that we developed sci-MET.

Chapter 1: Highly scalable generation of DNA methylation profiles in single cells

This chapter contains a modified version of material that appeared in the author's publication: Mulqueen, Ryan M., et al. "Highly scalable generation of DNA methylation profiles in single cells." Nature biotechnology 36.5 (2018): 428-431. © 2018, Nature Publishing Group, a division of Macmillan Publishers Limited.

Authors collaborating in this work and affiliations

Ryan M Mulqueen[a], Dmitry Pokholok[b], Steven J Norberg[b], Kristof A Torkenczy[a], Andrew J Fields[a], Duanchen Sun[c,d], John R Sinnamon[e], Jay Shendure[f], Cole Trapnell[f], Brian J O'Roak[a], Zheng Xia[c,d], Frank J Steemers[b] & Andrew C Adey[a,h]

a. Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, Oregon, USA
b. Illumina, Inc., San Diego, California, USA
c. Department of Molecular Microbiology & Immunology, Oregon Health & Science University, Portland, Oregon, USA
d. Computational Biology Program, Oregon Health & Science University, Portland, Oregon, USA
e. Vollum Institute, Oregon Health & Science University, Portland, Oregon, USA
f. Department of Genome Sciences, University of Washington, Seattle, Washington, USA
g. Department of Genome Sciences, University of Washington, Seattle, Washington, USA
h. Knight Cardiovascular Institute, Portland, Oregon, USA

Author Contributions

A.C.A. and R.M.M. conceived the sci-MET assay. R.M.M. carried out all sci-MET preparations with contributions from A.J.F. A.C.A., R.M.M., F.J.S., D.P., and S.N. designed the sci-MET adaptors and primers and reduced the assay to practice. R.M.M., F.J.S., D.P., and S.N. carried out all sequencing. R.M.M. led the data analysis. D.S. and Z.X. performed the NMF-tSNE analysis. K.A.T. provided additional analyses. J.R.S. performed mouse cortex dissection. F.J.S., J.S., C.T., and B.J.O. contributed to analysis design and edited the manuscript. A.C.A. supervised all aspects of the study.

Appendix Figures and Tables supplied to the committee.

Abstract

We present a highly scalable assay for whole-genome methylation profiling of single cells. We use our approach, single-cell combinatorial indexing for methylation analysis (sci-MET), to produce 3,282 single-cell bisulfite sequencing libraries and achieve read alignment rates of 68 ± 8%. We apply sci-MET to discriminate the cellular identity of a mixture of three human cell lines and to identify excitatory and inhibitory neuronal populations from mouse cortical tissue.

Main

DNA methylation at cytosine-guanine dinucleotides (CG) and non-CG sites (CH) have cell type-specificity and are subject to active modification during development[130]. This motivates a single-cell approach, which can assess cell-type and developmental-state specificity in complex tissues through methylation profiles. DNA methylation can be probed at base-pair resolution at the whole genome scale using bisulfite sequencing (WGBS)[131]. Recent work optimized whole genome bisulfite sequencing to enable assessment at the single-cell level (scWGBS)[19,115,124,132,133]; these assays provide unique insights into methylation patterning. However, the scWGBS protocol processes each cell in its own reaction vessel, severely limiting cell count throughput. Furthermore, alignment rates for traditional scWGBS libraries are much lower (on the order of 25 ± 20%) than for the equivalent bulk protocol[19,110,119,126], which increases the cost of obtaining sufficient information. A recent study achieved an alignment rate just over 50%, for over 6,000 single cells; however, the study relied on a brute-force strategy that still required an individual reaction well for each cell produced[19].

We have described a strategy for combinatorial indexing that has been extended to multiple applications[73,134–137]. In this platform, DNA (or RNA) within nuclei or cells is modified with an indexed adaptor corresponding to one of 96 (or 384) wells while nuclear integrity is maintained. Reactions are pooled, and a limited number of pre-indexed nuclei are redistributed into each of a new set of wells, such that the probability of two nuclei harboring the same initial index ending up in the same well is low. PCR is then used to incorporate a second index and generate a cell-specific barcode composed of the unique index combinations. We adapted our

single-cell combinatorial indexing strategy, (sci-) to WGBS methylation analysis (sci-MET, Figure 14a) using transposomes with adaptors depleted of cytosines, and thus unaffected by bisulfite treatment (Appendix Tables 1–3). The second adaptor is incorporated after pooling, redistribution, bisulfite conversion, and cleanup by performing multiple rounds of random primer extension, as in traditional scWGBS protocols[124]. This workflow enables the first stage of library construction in one set of wells, followed by the second stage, where each well contains a number of pre-indexed nuclei. We refer to the number of single-cell methylation libraries we expect per experiment as $N \times D$, where N is the number of wells in the second stage of library preparation and D is the number of pre-indexed nuclei in each well (Appendix Fig. 1).

From a 96 × 22 experiment on a B-lymphoblast cell line (GM12878), we generated libraries for which we could identify barcodes corresponding to 708 single cells (33.5% efficiency, defined as the number of libraries generated out of the number expected). Sequencing this library to a low depth (mean 55,129 unique reads per cell; Appendix Figs. 2 and 3) produced methylation profiles that closely matched expectation for the GM12878 cell line (Figure 14b). We next performed sci-MET on a mix of human and mouse cell lines using two alternative nucleosome depletion strategies to estimate the barcode collision rate (i.e., two nuclei of the same transposase barcode ending up in the same PCR well[136]). We observed a high collision rate using a lithium-based approach (22%); however, crosslinking and SDS treatment (xSDS) produced a low collision rate, in line with other combinatorial indexing strategies[73,136,137] at 7.3% (Figure 14c and Appendix Figs. 1–4). We note that the collision rate is tunable by the number of nuclei sorted into each well during the second stage of indexing.

We next profiled pure populations and an uneven artificial mixture of GM12878, primary inguinal fibroblast (Primary Fibro., GM05756), and HEK293 cell lines. In a 40 × 22 experiment using xSDS nucleosome depletion, we characterized genome-wide methylation in 691 single cells passing quality filters (78.5% efficiency; Appendix Figs. 1–3). We achieved a mean alignment rate of 68 ± 8% (Figure 14d), approaching bulk-cell levels, likely due to the efficiency of transposase-based adaptor incorporation[69], and a mean unique aligned read count of 403,265 per cell, with 48 cells producing over one million uniquely aligned reads (Appendix Fig. 2). These data translate to

coverage of mappable[138] CG dinucleotides ranging from 0.05% to 7.0% (mean 1.1 ± 0.9%). Both increased sequencing effort and additional rounds of linear amplification are likely to increase coverage, as libraries were not near saturation (Appendix Fig. 5 and Figure 14e). Based on our projections, sci-MET, in its current form, produces lower per-cell coverage percentages than others have produced[113]; however, sufficient coverage per cell is achievable for cell-type discrimination in a mixed population—the intended goal of low-coverage, high-cell-count strategies.

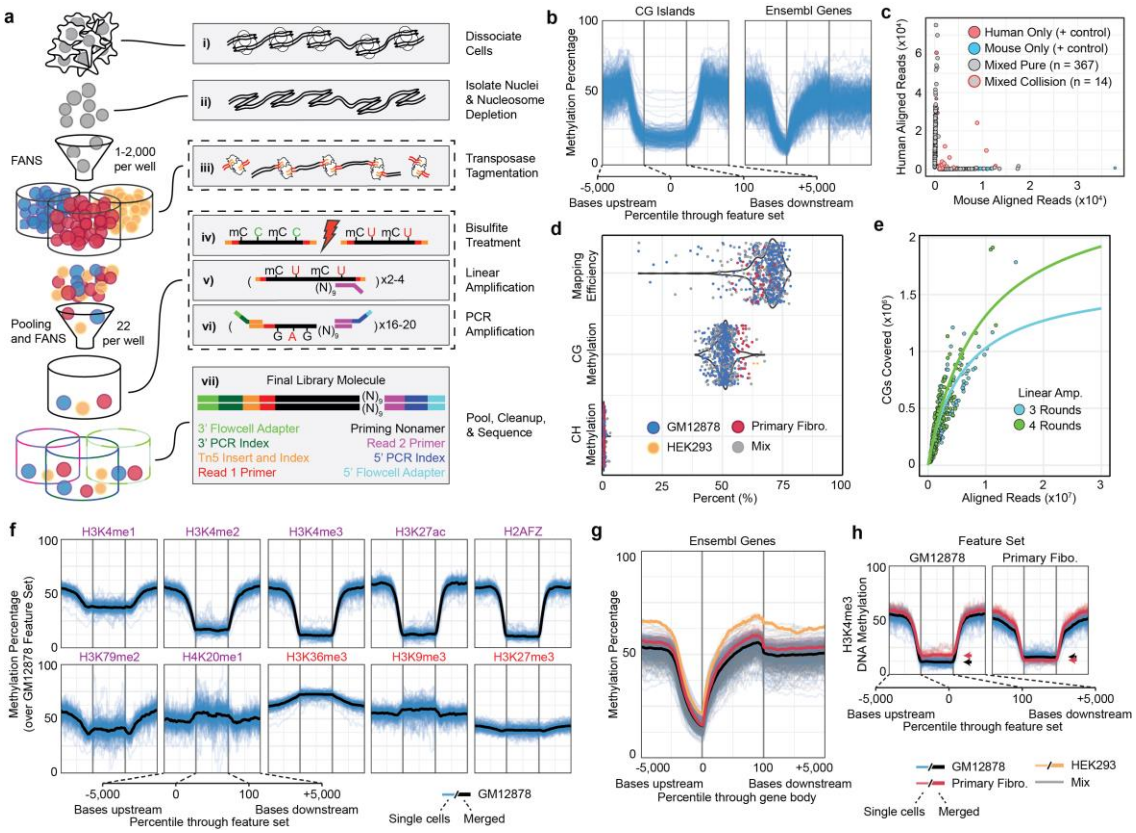


Figure 14. (**a**) The sci-MET workflow. (**b**) Methylation rates for single GM12878 cells (*n* = 283 cells) over CG islands (left) and gene bodies (right). (**c**) sci-MET of mixture of mouse and human cells using xSDS nucleosome depletion to estimate barcode collisions. *n* = 566 cells. (**d**) Mapping efficiency, global CG methylation, and global CH methylation for a mix of human cell lines (*n* = 641 cells). (**e**) The number of CG dinucleotides covered by the total aligned reads per cell. Amp., amplification. (**f**) Methylation rates for GM12878 cells. Purple typeface: generally activating features; red typeface: generally repressive features. (**g**) Methylation rates for the three cell types over annotated genes. (**h**) Methylation rates over GM12878 and Primary Fibroblast ENCODE H3K4me3 ChIP-seq peaks. Arrows indicate the mean for the feature set. Key applies to panels **g** and **h**.

We next summarized methylation status[124] for each cell across autosomal loci of the Ensembl Regulatory Build[139], which contains known transcription factor binding and other regulatory sites. We performed non-negative Matrix Factorization (NMF) followed by t-distributed Stochastic Neighbor Embedding (tSNE) to project cells in two-dimensional space, producing

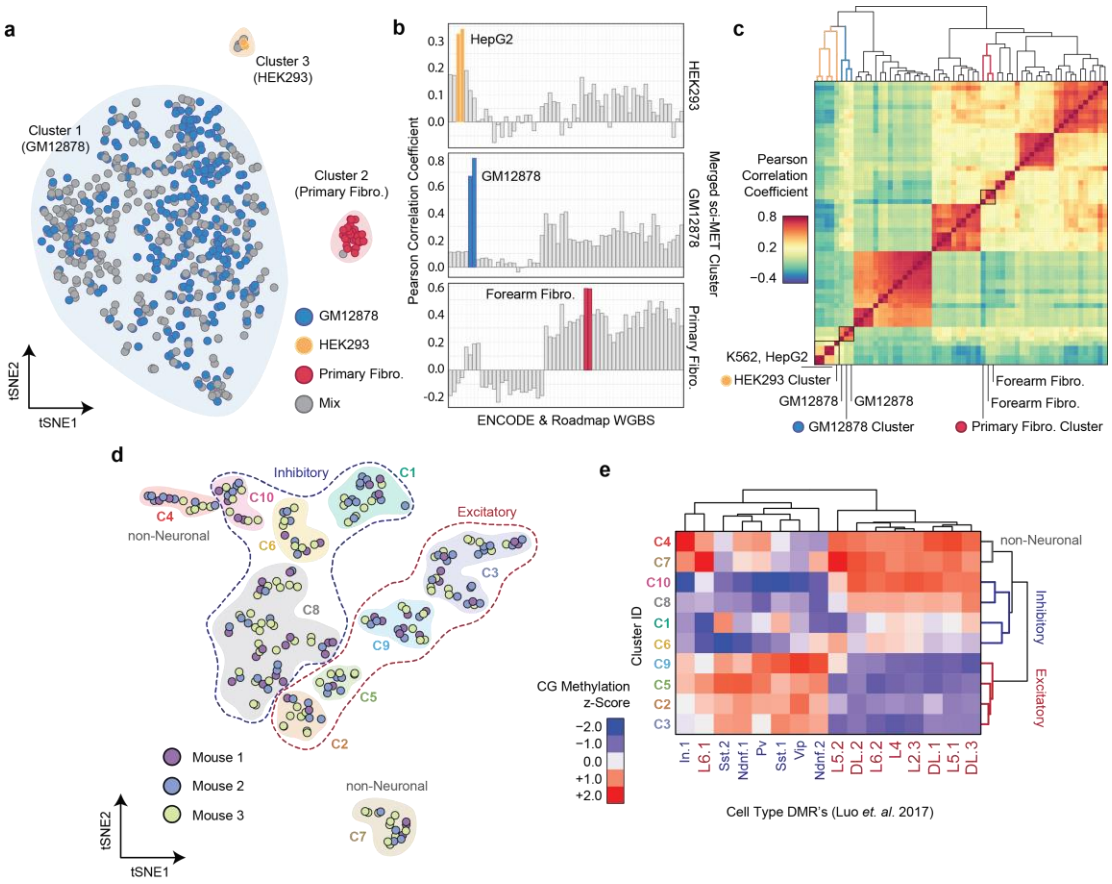


Figure 15. (**a**) NMF-tSNE projections of single-cell methylomes. Clusters are indicated by a shaded background. (**b**) Single-cell methylomes (*n* = 641 cells) were aggregated over the three clusters and then correlated with publically available WGBS data. Closely matched cell types are in color. (**c**) Hierarchical clustering on the Pearson correlation values of HEK293, GM12878, and Primary Fibro. (**d**) NMF-tSNE projection of cortical cells based on CG and CH methylation. Clusters are indicated by a shaded background and grouped by class using dashed lines. (**e**) Methylation *z*-score heatmap of aggregate cell clusters over previously described DMRs (*n* = 285 cells).

clearly defined clusters that were identified using density-based methods (Figure 15a). We correlated the methylation rates of collapsed clusters with publically available WGBS data sets[64,140] for the top 1,000 most-variable regulatory regions. For each merged cluster, the two most highly correlated samples were of the same cell type, or the most similar cell line in the case of HEK293 (Figure 15b,c and Appendix Figs. 6 and 7).

To test whether cell type discrimination is possible in an *in vivo* model, we performed a 96 × 10 preparation from primary cortical tissue of three mice, for a total of 606 single-cell libraries. A subset of the second-stage indexing wells were sequenced to a higher depth than the rest of the plate (186 cells), with the remainder to enough depth to define them as true single-cell libraries (420 cells; Appendix Figs. 1–3). Overall, this preparation produced a mean alignment rate of 59.9 ± 11.9%. In total, 285 cells met a read depth threshold of 30,000 uniquely aligned

reads (mean 186,710) and were carried through subsequent analysis, with the percent of CGs covered genome-wide ranging from 0.10% to 4.5% (mean 0.82 ± 0.85%).

We assessed methylation in the CH context, which has been previously observed at elevated levels and in a distinct patterning in neuronal lineages (Appendix Figs. 8 and 9)[119,141], as well as in the CG context. We processed each matrix (CH over 100 kbp windows, CG over the Ensembl Regulatory Build) individually and combined through NMF-tSNE and clustering (Online Methods and Figure 15d). Two clusters were determined to be likely non-neuronal cell populations, and the remainder neuronal (Appendix Fig. 10). We then aggregated the coverage of each cluster and calculated the methylation percentage over previously described cortical differentially methylated regions (DMRs; Figure 15d)[19]. This revealed a distinct enrichment for each neuronal cluster within sets of excitatory and inhibitory DMRs and allowed us to classify sets of clusters (Figure 15d).

Inherent in our protocol is the ability to scale up to far greater numbers by expanding the number of indexes (Appendix Fig. 11). In addition to the increased throughput, we achieved substantially improved read-alignment rates when compared to existing lower-throughput approaches, dramatically reducing the sequencing burden. Our platform achieves both the throughput and cost-effectiveness (Appendix Table 4) that is required to scale single-cell DNA methylation assessment to levels comparable to other epigenetic and transcriptional properties.

Methods

Preparation of unmethylated control DNA.

100 ng of unmethylated Lambda Phage DNA (Promega, Cat. D1521) was treated with 4 µL of 500 nM transposase-adaptor complex (transposome) pre-loaded with cytosine-depleted custom oligonucleotides in 10 µL of 1× Nextera Tagment DNA (TD) buffer from the Nextera DNA Sample Preparation Kit (Illumina, Cat. FC-121-1031) diluted with nuclei isolation buffer (NIB) to simulate reaction conditions for nuclei. Following incubation for 20 min at 55 °C, this reaction was cleaned with QIAquick PCR Purification Kit (Qiagen, Cat. 28104) and eluted in 30 µL of 10 mM Tris-Cl solution (pH 8.0). The tagmented, cleaned DNA was then quantified via Qubit 2.0 Flourometer dsDNA High Sensitivity Assay (Thermo Fisher, Cat. Q32854).

Tissue culture.

Tissue culture cell lines (GM12878, Coriell; NIH/3T3, ATCC CRL-1658; HEK293, ATCC CRL-1554; Primary Fibro., inguinal fibroblast, GM05756, Coriell, passage 7) were cultured in 5% CO_2 at 37 °C. GM12878 cells were grown in Roswell Park Memorial Institute media (RPMI, Gibco, Cat. 11875093) supplemented with 15% (v/v) FBS (FBS, Gibco, Cat. 10082147), 1× L-glutamine (Gibco, Cat. 25030081), 1× penicillin-streptomycin (Gibco, Cat. 15140122), and gentamicin (Gibco, Cat. 15750060). HEK293 cells were grown in Dulbecco's Modified Eagle's media (DMEM, Gibco, Cat. 11995065), supplemented with 10% FBS (v/v), and 1× L-glutamine. NIH/3T3 cells were grown in the same preparation of DMEM as HEK293 cells. Primary fibroblasts were cultured in a growth medium comprised of DMEM/F12 (with GlutaMax; Thermo Fisher), 10% FBS (FBS; Thermo Fisher, v/v), 1% MEM Non-Essential Amino Acids (Thermo Fisher, v/v), and 1× penicillin-streptomycin (Gibco). Adherent cell lines were grown to ~90% confluency at the time of harvest.

Mouse samples.

All animal studies were approved by the Oregon Health and Science University Institutional Animal Care and Use Committee. C57BL/6J mice were obtained from Jackson Laboratory (stock number 000664). Sixty-day-old C57BL/6J female mice were deeply anesthetized using isoflurane. After decapitation the brain was removed and the entire cortex isolated and placed in ice-cold PBS.

Sample preparation and nuclei isolation.

For library preparation, cells were pelleted if cultured in suspension, or trypsinzed (Gibco, Cat. 25200056), if adherent. Cell were washed once with ice-cold PBS and carried through cross-linking (for the xSDS method) or directly into nuclei preparation using nuclei isolation buffer (NIB, 10 mM Tris HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Igepal (v/v), 1× protease inhibitors (Roche, Cat. 1187358001)). Cortical samples were cut with a sterile razor blade and resuspended in a chilled 5 mL modified nuclei isolation buffer (NIB-HEPES, 20 mM HEPES, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Igepal, 1× protease inhibitors). Cells were given 5 min to equilibrate to the salt solution before five loose strokes in a Dounce homogenizer, another 5 min to

equilibrate, and another five loose strokes and ten tight strokes. Nuclei were then spun in a pre-chilled 4 °C centrifuge for 5 min at 600*g*.

Nucleosome depletion.

Nucleosome depletion and combinatorial indexing strategies were performed similar to previously described, with some variations[136].

Lithium-assisted nucleosome depletion (LAND). Land was performed for generation of GM12878-only and human/mouse libraries. Prepared nuclei were pelleted and resuspended in NIB supplemented with 200 µL of 12.5 mM lithium 3,5-diiodosalicylic acid (Sigma, Cat D3635) for 5 min on ice before addition of 800 µL NIB and then taken directly into the combinatorial indexing protocol.

Cross-linking and SDS nucleosome depletion (xSDS). Cells were cross-linked by incubation in 10 mL of media with 1.5% formaldehyde (v/v) and incubated at room temperature for 10 min with gentle agitation. Cross-linking was quenched with 800 µL 2.5 M glycine and incubated on ice for 5 min. Cells were then spun down, washed with ice-cold PBS, and resuspended in ice-cold NIB for a 20-min incubation on ice with gentle agitation. Cells were then pelleted, washed with 900 µL of 1× NEBuffer 2.1 and resuspended in 800 µL 1× NEBuffer 2.1 with 0.3% SDS (v/v, Sigma, Cat. L3771) and incubated at 42 °C with vigorous shaking for 30 min in a thermomixer (Eppendorf). 200 µL of 10% Triton-X was added to quench, and the solution was incubated at for another 30 min at 42 °C with vigorous shaking. Nuclei were then taken into the combinatorial indexing protocol. We were concerned that the crosslinking might affect the bisulfite conversion reaction; however, based on the methylation rates (particularly for those of nonCG methylation which were very low in concordance with expectations), we determined that not to be the case.

Combinatorial indexing via tagmentation.

Nuclei were stained with 8 µL of 5 mg/mL DAPI (Thermo Fisher, Cat. D1306) and passed through a 35-um cell strainer. A 96-well plate was prepared with 10 µL of 1× TD buffer diluted with NIB in each well. Fluorescence-assisted nuclei sorting (FANS) was performed with a Sony SH800 flow sorter to sort 2,500 single nuclei into each well in fast-sort mode (Appendix Fig. 12). 4

µL of 500 nM transposome, pre-loaded with cytosine-depleted, uniquely indexed, custom oligonucleotides were placed in each well (transposomes assembled as described previously[142]). This cytosine-depleted approach improved downstream PCR amplification and decreased library generation costs compared to previous methylated adaptor attempts[136]. Reactions were incubated at 55 °C for 20 min. All wells were then pooled and stained with DAPI as described for the first FANS sort. A second 96-well plate was prepared with each well containing digestion reagents as described by the manufacturer's protocol for the EZ-96 DNA Methylation MagPrep Kit (Zymo, Cat. D5040) at one-fifth the volume (for a total of 5 µL per well). 22 post-tagmentation nuclei from the pool of all reactions were sorted into each well using the single-cell sorting setting. Some wells were randomly selected to receive only ten nuclei, to allow for unmethylated controls. The plate was then spun down at 600*g* for 5 min at 4 °C.

Library preparation.

Prior to bisulfite conversion, several wells, which only received ten nuclei in the final sort, were spiked with ~35 pg of the prepared unmethylated control DNA, to keep DNA mass constant per well. Nuclei were then processed following manufacturer's protocol for the EZ-96 DNA Methylation MagPrep Kit, with volumes reduced to one-fifth those described by the manufacturer to allow for single-well reaction processing, and other slight modifications. Following the final post-bisulfite library cleanup, each well was eluted in 25 µL of Zymo M-Elution Buffer and transferred to a well in a 96-well plate prepared with the following reaction mixture for linear amplification: 16 µL PCR-clean ddH_2O, 5 µL 10× NEBuffer 2.1 (NEB, Cat. B7202), 2 µL 10 mM dNTP mix (NEB, Cat. N0447), and 2 µL of 10 µM random nonamer primer with a partial sequence of the Illumina Standard Read 2 sequencing primer (9NP, 3′-NNNNNNNNNAGATCGGAAGAGCACACGTCTG-5′). To render libraries single-stranded before linear amplification, reactions were heat-shocked at 95 °C for 45 s and then flash-cooled on ice. Following cooling, 10 U Klenow (3′->5′ exo-) polymerase (Enzymatics, Cat. P7010-LC-L), was added to each reaction, followed by incubation at 4 °C for 5 min, then a slow ramp of +1 °C/15 s, and 37 °C for 90 min. This was repeated for two to four times, depending on library and in accordance with previously described scWGBS protocols (Appendix Fig. 1)[124]. For each

repetition, 1 µL 10 µM 9NP, 1 µL 10 mM dNTP mix, 1.25 4× NEBuffer 2.1, and 10 U Klenow (3′ -> 5′ exo-) polymerase was added after the heat shock and cooling. Following completion of linear amplifications, wells were cleaned with 1.1× (by volume) of 18% PEG SPRI Bead mixture (Sera-Mag SpeedBeads (GE, Cat. 65152105050250) washed and resuspended in 18% PEG 8000 (by mass), 1 M NaCl, 10 mMTris-HCl, pH 8.0, 1 mM EDTA, 0.05% Tween-20), with a 5 min room temperature incubation, then placed on a magnetic rack until the supernatant was cleared. The supernatant was discarded, and beads were washed with 80% ethanol while held in place by magnets. Beads were then dried and libraries were eluted in 21 µL 10 mM Tris-Cl (pH 8.5). The full 21 µL eluate was then placed into a 96-well plate prepared with a PCR reaction mixture containing 25 µL 2× KAPA HiFi HotStart ReadyMix (Kapa, Cat. KK2602), 2 µL each of 10 µM forward and 10 µM reverse uniquely indexed primers (each introducing a 10-nt indexing sequence), and 0.5 µL of 100× SYBR Green I (FMC BioProducts, Cat. FC-121-1031). Real-time PCR was performed on a Bio-Rad CFX thermocycler with the following conditions: 95 °C for 2 min, (94 °C for 80 s, 65 °C for 30 s, 72 °C for 30 s [Image]) for 18–22 cycles. PCR was stopped once libraries reached the inflection point of measured SYBR green fluorescence. Following PCR, libraries were then pooled by column (10 µL/well) and with 0.8× (by volume) 18% PEG SPRI Bead Mixture as described previously. Libraries were eluted off the magnetic beads in 25 µL of 10 mM Tris-Cl (pH 8.5

Library quantification and sequencing.

Libraries were pooled and quantified between the range of 200 bp and 1 kbp using a 2100 Bioanalyzer DNA High Sensitivity kit (Agilent, Cat. 5067-4626; Appendix Fig. 13). Pools were sequenced on either an Illumina NextSeq 500, HiSeq 1000, HiSeq 2500 or HiSeq X, loaded at 0.9 pM, with a 5%, 12%, or 30% PhiX spike-in to improve complexity for the HiSeq 2500, HiSeq 1000 or HiSeqX, and NextSeq 500, respectively. All sequencing runs used a custom locked-nucleic acid (LNA) oligonucleotides for custom sequencing primers to match the standard chemistry temperatures (Appendix Table 3). With the exception of the first GM12878-only library pool, libraries were sequenced with a custom sequencing chemistry protocol (Read 1: 100

imaged cycles; Index Read 1: 10 imaged cycles, 27 dark cycles, 11 imaged cycles; Index Read 2: 10 imaged cycles).

Sequence read processing.

Reads were processed using bcl2fastq (Illumina Inc., v2.19.0) with the "–create-fastq-for-index-reads" and "–with-failed-reads" options to produce fastq files. Fastq reads were then identified by indexes, requiring each index (the two 10-nt indexes introduced by PCR, and the 11-nt index introduced by tagmentation) to independently be within a Hamming distance of two from the expected reference sequences. Reads with all three indexes assigned had the respective reference index sequences concatenated to a barcode and appended to the read name, which served as the barcode identifier. Reads were then trimmed using TrimGalore! (v0.4.0) with option "-a AGATCGGAAGAGC" to identify adapters. Trimmed reads were quality-checked using FastQC (v0.11.3) for adaptor content, percent base across reads for bisulfite conversion biases, and k-mer bias. Alignment to the human (GRCh37), mouse (GRCm38), or a combined human–mouse hybrid genome was performed with Bismark (v0.14.3) using "–bowtie2" and "–unmapped" options[143]. Aligned reads were then de-duplicated based on barcode, chromosome, and starting position.

GM12878-only library development.

GM12878-only libraries were generated as described above with alterations/specifications as follows: library were generated using the LAND method for nucleosome depletion, libraries were generated using four rounds of linear amplification, and were sequenced in a paired-end manner. For the paired-end sequencing strategy the following custom sequencing chemistry protocol was used (Read 1: 50 imaged cycles; Index Read 1: 10 imaged cycles, 27 dark cycles, 11 imaged cycles; Index Read 2: 10 imaged cycles; Read 2: 50 imaged cycles). Sequencing reads were processed using slightly modified read processing pipeline. Trimming was performed with TrimGalore! using the "-paired" option, we observed biases at the start of both read 1 and read 2 sequences, likely due to the random priming strategy, and consequently trimmed the reads with options "–clip_R1 6", "–clip_R2 9". We aligned

reads to the GRCh37 reference genome with Bismark with an added “-p” option for the paired-end alignment.

Human–mouse library development.

Human (GM12878) and mouse (NIH/3T3) cell lines were mixed following nuclei isolation, but before nucleosome depletion in a roughly equal ratio. Nucleosomes were then depleted using the LAND technique and processed as described above. Reads were aligned to a hybrid human–mouse genome. To estimate barcode collision rate we identified putative single-cell libraries with <90% of reads that aligned to a single species which represents approximately half of the total collision rate (Appendix Fig. 4). We also generated a second human–mouse library using a mixture of human (HEK293) and mouse (NIH/3T3) cells which underwent xSDS nucleosome depletion. The human-mouse xSDS library was processed as described above (Figure 14c).

Cell line discrimination library development.

To assess the ability of sci-MET to separate out different cell types using a low-coverage, high-cell count approach, we selected three cell lines: GM12878 (a B-lymphoblastoid cell line), HEK293 (a kidney epithelial cell line), and GM05756 (primary inguinal fibroblast line). We prepared a sci-MET library using xSDS nucleosome depletion that included each cell line on their own in addition to a mix comprised of 40% GM12878, 40% GM05756, and 20% HEK293 where they were combined after nuclei isolation. We suspect that this ratio was dramatically altered owing to the FANS gating that we performed, which likely excluded the majority of the aneuploid HEK293 cells which are difficult to distinguish from euploid doublets (Appendix Fig. 12). Furthermore, for the majority of wells in which the cell identity was known, the cells were GM129878, thus likely favoring the FANS gating to that cell's profile. It is important to note that this challenge would persist for any method of single-cell profiling that requires single-cell sorting, such as all of the existing single-cell methylation assay platforms, and is an important item to consider. Libraries were processed as described above.

Mouse cortex library development.

Mouse cortical samples were brought through the sci-MET protocol via xSDS as described above. Notably, we used a modified NIB (NIB-HEPES; described under “Sample

Preparation and Nuclei Isolation"), which substituted the early use of Tris-HCl with HEPES, to avoid quenching formaldehyde during fixation. Three mouse cortical samples were processed in parallel before tagmentation, such that sample identity was maintained. Following this, all nuclei were pooled for downstream library generation. Downstream library construction was processed as described above. Mouse cortex libraries underwent the same quality-control filters, omitting that which removed cell libraries with 5% nonCG methylation.

Single-cell discrimination by unique read count.

We sought to use the unique aligned read count to stratify individual cells from noise (Appendix Fig. 3). First we performed k-means clustering (k = 3) based on the $\log_{10}$ number of unique aligned reads per barcode (the three indexes assigned to a read). We fit a normal distribution to the cluster containing the barcodes with the highest number of unique aligned reads. In case the cluster with the highest aligned reads contained multiple peaks due to low coverage (as in the GM12878-only prep) we used an alternative approach to fitting a normal distribution and fit mixed normal distributions to the clustered data. From the fit distributions, the threshold was then defined based on the 95% confidence interval (CI) of the fitted normal distribution with the highest number of unique reads (mean-(1.96 × s.d.)). We used the kmeans function in R (v. 3.4.2) for clustering and the MASS (v. 7.3-45) and mixtools (v. 1.1.0) packages for fitting the normal and mixed normal distributions.

Methylome coverage estimation.

To provide an accurate measurement of CG dinucleotides covered by sci-MET, we collapsed CG measurements to a single haploid strand using Bismark (v.0.18.2) coverage2cytosine command using the "–merge_CpG" option. We used the recently reported Bismap[138] tool to estimate uniquely mappable regions of the mm10 and hg19 reference genomes. Through this, we determined a total of 27,003,976 CG sites for the haploid hg19 reference and 19,788,681 CG sites for the haploid mm10 reference. These numbers were used for all CG coverage estimates.

Quality control.

We assessed bisulfite conversion efficiency in our preparations through spike-in of unmethylated lambda phage DNA. We aligned fastq reads with the respective 11-nt tagmentation index to the lambda genome (GenBank: J02459.1) using Bismark. We de-duplicated reads, and filtered to high-quality alignments (≥Q30). We observed a highly efficient bisulfite conversion across sci-MET library constructions (>99%; Appendix Table 5).

Individual barcodes per library were assessed for mapping efficiency (calculated as aligned reads/fastq reads assigned to a barcode), and complexity (calculated as de-duplicated, aligned reads/aligned reads assigned to a barcode). Our protocol for library construction both increased the throughput of single-cell generation, and largely increased mapping efficiency compared to previous methods[19,113,115,124,132,133]. Barcodes were filtered by unique read cutoffs (described in "Single-cell discrimination") and subsequently filtered. We required cells which met read threshold cutoff to have a mapping efficiency of ≥5%, a nonCG methylation of ≤ 5% for downstream clustering analysis. We further stratified our library pool to assess the effect of various rounds of linear amplification on single-cell library quality. We found that four rounds of linear amplification significantly increased mapping efficiency (P-value = 7.83×10^{-16}, t = 8.27, Student's two-sided t-test. Transposase complexes showed differences in library construction efficiency (Appendix Fig. 14). Alignment rates and coverage did not correlate strongly with percent methylation per cell (Appendix Fig. 15).

To estimate average library saturation, we fit two-factor saturation curves to single-cell libraries within the human cell line mix experiment using the *drc* (v3.0-1) package's *drm* function in R dependent on rounds of linear amplification. For three and four rounds of linear amplification, our projected upper asymptotes (full sequencing saturation) were 1.66×10^6, and 2.51×10^6, unique CGs per single-cell library, respectively (Figure 14e). All quality assessment data are reported as mean ± s.d. where appropriate.

Individual cell saturation (Appendix Fig. 5) was carried out by projecting the estimated unique read counts per cell to decreasing complexity increments as described previously[136]. We

then calculated the expected CG percent coverage based on the linear relationship between the percent of CG sites covered by the unique read count (Appendix Fig. 16).

Coverage bias across annotations.

We calculated the coverage bias in individual cells across DHS, CG islands, and histone (H2AFZ, H3K27ac, H3K36me3, H3K4me1, H3K4me3, H3K4me3, H3K79me2, H3K9ac, H3K9me3, H4K20me1, H3K27me3) sites using annotated DNase, methylation profiling and CHIP-seq peak data from the publically available UCSC and ENCODE databases[64,140]. We used bedtools multicov (v. 2.22.0) to determine the coverage for each cell across all sites of each annotation bed file. We then determined the fraction of total reads per kilobase pair (kbp) by summing the coverage across all sites in a cell and normalizing by the reads per cell and by the sum of the genomic distance of the peak sites (Appendix Fig. 17).

CG sites covered per n cells analyzed.

We simulated the number of unique CG sites covered in an experiment by an arbitrary number of cells using sci-MET (human cell line experiment data) by performing 100 iterations of sampling of *n* = (10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200,250, 300, 350, 400, 450, 500, 550, 600, 650) cells. We then calculated the aggregate number of unique CG sites covered across all cells for each sampling and fitted a LOWESS curve (using R package *ggplot* v.2.2.1) to the unique CG sites per *n* cells sampled saturation plot (Appendix Fig. 18).

Non-negative matrix factorization, tSNE, and clustering.

We quantified methylation rate across Ensembl Regulatory Build windows using a previously described method[124]. Non-negative Matrix Factorization (NMF) is an unsupervised data decomposition technique and was performed on the summarized windows. Here we used NMF to learn new feature representations[144]. NMF is mathematically approximated by: $\mathbf{A}^{m\times n} \approx \mathbf{W}^{m\times k} \times \mathbf{H}^{k\times n}$, where **A** is the matrix representing the single-cell methylation profiles of *n* samples across *m* features. **W** is a dictionary matrix with a much smaller *k* than *m*. **H** is the activation coefficients on the new basis. All the three of them are non-negative. The column vectors in **W** are called *meta-feature*, which are higher-level abstraction of the original methylation levels and each column in **H** is meta-expression on the new basis of each sample. Here we

set k = 12 to get matrix **A** factorized into low-rank matrix **W** and **H**. In this way, we extracted the uncorrelated basis and the coefficient matrix **H** of the new basis by significantly reducing the dimension of the features. Since relatively few basis vectors are used to represent many data vectors ($k << m$), good approximation can be achieved only if the basis vectors discover structure that is latent in the data, which will aid sample clustering and visualization. Then, given the learned feature representation, Student's t-distributed stochastic neighbor embedding (t-SNE) package *Rtsne* (v.0.13) for R is used to plot the meta-expression matrix $\mathbf{H}^{k \times n}$ with default parameters. Clustering on the NMF-tSNE coordinates was performed using the Density Based Clustering of Applications with Noise (DBSCAN; v.1.1-1) with an epsilon value of 4 and a minimal cell seed threshold of four[145]. This process was performed for cells with ≥30,000 unique aligned reads (Figure 15b). Clusters were assessed for read count and alignment rate bias, as well as validated through Y chromosome read count (Appendix Fig. 19).

Methylation over genomic annotations.

Methylation rates plotted over ChIP-seq and other genomic annotations were generated by aggregating the methylation fractions in percentile windows for 5,000 bp upstream of the feature, through the feature set, and 5,000 bp downstream of the feature and smoothed over three percentile window groups. Methylation rates were carried out for each individual cell as well as for the combination of cells of each specific sample type in the case of the human cell type mix experiment (Appendix Fig. 20).

mCH periodicity.

Two approaches were undertaken to estimate the patterning of CH methylation. First, leveraging our read length (>70 bp on average), we estimated the cis-mCH patterning. For all mCH measurements with both up- and downstream mCHs within the same read, we calculated the distance between the nearest mCH sites. This was performed with a custom Python script on the Bismark alignment file (v. 2.7.9). The minimal distance up or downstream of each mCH site was then plotted using ggplot2 geom_histogram function (v. 2.2.1) in R (v 3.4.2). Second, we assessed all CH measurements around annotated CTCF motif sites (described in more detail in methods section 'Transcription Factor Methylation') to act as a centering point for nucleosome

position. We then normalized the annotated CTCF windows annotated previously[146] and plotted both percent CH methylation using the R packages *GenomicRanges* (v. 1.28.4) and *genomation* (v.1.8.0; Appendix Fig. 9).

Window summaries and correlations over Ensembl regulatory regions.

Using ENCODE and Epigenome Roadmap bulk WGBS samples, we quantified a weighted methylation rate and variance across samples using the Ensembl Regulatory Build loci16. We next took the top 1,000 most variable loci across the bulk samples and summarized methylation rates within single-cell clusters identified above. We performed a Pearson correlation of methylation rates with the bulk WGBS samples using base R *cor* function. Biclustering was performed using the R package *gplots* (v. 3.0.1) *heatmap2* function (Appendix Fig. 21).

Transcription factor methylation.

Transcription factor motifs across the hg19 reference genome were taken from Homer[146]. All sites with a shared transcription factor motif were assumed to be co-regulated, as described[115]. CG sites per cell within the human cell line mix experiment were collapsed and summarized, using *bedtools intersect* and *groupby* commands (v2.22.0). Transcription factor annotations with less than 30 CG measurements were excluded on a per-cell basis. Transcription factor annotations with more than 20% of cells missing a value were excluded, leaving a final count of 237 annotations. The matrix was then clustered using tSNE with package *Rtsne* (v.0.13) in R (Appendix Fig. 22). Additionally, a hierarchical biclustering approach using the R package *ComplexHeatmap* (v1.14.0) was used on the same cell X transcription factor matrix before *Z*-scoring, which failed to appropriately separate out cell types (Appendix Fig. 23).

Non-binary CGs methylation analysis.

To assess CG dinucleotide methylation variability, we collapsed all cells within the GM12878 cluster in the human cell line mix experiment. We defined CG sites with two or more measurements sourced from different cells as either binary (fully methylated or unmethylated across cells) or non-binary (differentially methylated across cells). We then calculated the enrichment of non-binary CG sites overlapping genomic features (chromatin marks, DNase

hypersensitivity regions, CG islands) using *bedtools intersect* (v2.22.0). We compared this enrichment to binary CG sites, calculated in the same manner.

We observed a significant relative enrichment of non-binary sites in repressive marks

$$Relative\,Enrichment = \frac{\frac{CG_{non\text{-}binary}(overlapping)}{CG_{non\text{-}binary}(non\text{-}overlapping)}}{\frac{CG_{binary}(overlapping)}{CG_{binary}(non\text{-}overlapping)}}$$

(H3K27me3) and depletion in activating marks including DNase hypersensitivity regions and CG islands (Appendix Fig. 24). We repeated this analysis on transcription factor motifs described above (Appendix Fig. 25). Finally, we performed a Pearson's chi-squared test for significance of these enrichments (R base function *chisq.test*). False-discovery rate estimation was performed with R package *qvalue* (v.2.8.0).

Clustering of mouse cortex.

NMF was performed as described above for CG methylation over the Ensembl Regulatory Build as well as for methylation in the CH context over 100 kbp windows. We then carried out tSNE and density-based clustering[145] for each of these NMF matrixes independently and then an additional tSNE projection that included a combination of both NMF matrixes weighted equally to produce the projection presented in Figure 15d and then an additional round of density-based clustering (Appendix Fig. 10). The clusters for each case largely agreed with several exceptions where we decided to split the clusters in the joint CG and CH tSNE projection to provide increased granularity.

DMR methylation calculation for mouse cortical clusters.

To identify rudimentary cell types within our low-coverage clustered mouse cortical samples, we collapsed all reads within a respective cluster to increase CG coverage. CG methylated and unmethylated counts which overlapped with neuronal DMRs described by Luo *et al*. 2017 were summed. This was done with the *bedtools intersect* and *groupby* commands (v2.22.0). Percentage methylation of overlapping CG sites was calculated for each Luo *et al*.-defined neuronal subtype. The collapsed neuronal subtype DMR × cluster matrix was Z-scored

using the base R *scale* function. This was then plotted with the R package *ComplexHeatmap* (v1.14.0) with default parameters (Figure 15e).

Chapter 2: High-content single-cell combinatorial indexing

Authors collaborating in this work and affiliations

Ryan M. Mulqueen[a], Dmitry Pokholok[b], Brendan L. O'Connell[a], Casey A. Thornton[a], Fan Zhang[b], Brian J. O'Roak[a], Jason Link[c,d,f], Galip Gurkan Yardmici[c,e], Rosalie C. Sears[a,c,d,e,f], Frank J. Steemers[b], Andrew C. Adey[a,c,e,f,g]

a. Oregon Health & Science University, Department of Molecular and Medical Genetics, Portland, OR
b. ScaleBio, CA
c. Oregon Health & Science University, Cancer Early Detection Advanced Research Center, Portland, OR
d. Oregon Health & Science University, Knight Cancer Institute, Portland, OR
e. Oregon Health & Science University, Department of Oncological Sciences, Portland, OR
f. Oregon Health & Science University, Brendan Colson Center for Pancreatic Care, Portland, OR
g. Oregon Health & Science University, Knight Cardiovascular Institute, Portland, OR

Author Contributions

R.M.M., D.P., F.J.S., and A.C.A. conceived the study. R.M.M. performed all s3 experiments and led all analysis under the supervision of A.C.A.; D.P. and F.Z. performed additional experiments under the supervision of F.J.S.;B.L.O. and G.G.Y. contributed to the design and analysis of chromatin conformation s3-GCC protocol and datasets; B.J.O. provided support for R.M.M. and advice on analysis; C.A.T. contributed to the analysis of cell types in the s3-ATAC datasets; J.L. generated PDCL cell lines and performed characterization of the lines under supervision of R.C.S.; J.L. and R.C.S. contributed to the analysis of PDAC s3-WGS and s3-GCC datasets. Appendix Figures and Tables supplied to the committee.

Abstract

Single-cell genomics assays have emerged as a dominant platform for interrogating complex biological systems. Methods to capture various properties at the single-cell level typically suffer a tradeoff between cell count and information content, which is defined by the number of unique and usable reads acquired per cell. We and others have described workflows that utilize single-cell combinatorial indexing (sci)[147], leveraging transposase-based library construction[135] to assess a variety of genomic properties in high throughput; however, these techniques often produce sparse coverage for the property of interest. Here, we describe a novel adaptor-switching strategy, 's3', capable of producing one-to-two order-of-magnitude improvements in usable reads obtained per cell for chromatin accessibility (s3-ATAC), whole genome sequencing (s3-WGS), and whole genome plus chromatin conformation (s3-GCC), while retaining the same high-throughput capabilities of predecessor 'sci' technologies. We apply s3 to produce high-coverage single-cell ATAC-seq profiles of mouse brain and human cortex tissue; and whole genome and chromatin contact maps for two low-passage patient-derived cell lines from a primary pancreatic tumor (Table 1).

Table 1. Summary of s3 protocols.

Assay	Information Capture	Innovation	Sample
s3-ATAC	Single-cell chromatin conformation	Adapter switching method	Mouse whole brain and human cortex
s3-WGS	Single-cell whole genome	Above, and *in situ* nucleosome depletion optimization.	Patient-derived pancreatic cancer cell line and diploid control line.
s3-GCC	Single-cell whole genome, and genome conformation	Above, and *in situ* genome fixation, digestion and proximity ligation	Patient-derived pancreatic cancer cell line

Main

The core component of many sci-assays, as well as ATAC-seq, is the use of transposase-based library construction. While the transposition reaction itself (tagmentation) is highly efficient, viable sequencing library molecules are only produced when two different adaptors, in the form of forward or reverse primer sequences, are incorporated at each end of the molecule. This occurs only 50% of the time (Figure 16a, left). To combat this inefficiency, strategies including the use of larger complements of adaptor species[99], incorporation of T7 promoters to enable amplification via *in vitro* transcription[148–150], or reverse adaptor introduction through targeted[151] or random priming[152] have been developed; however, these methods are often complex and result in limited efficiency improvements. Here, we present a novel means of adapter replacement to produce library molecules tagged with both forward and reverse adaptors for top and bottom strands, overcoming this efficiency bottleneck. This format permits the use of a DNA index sequence embedded within the transposase adaptor complex, enabling single-cell combinatorial indexing (sci) applications, where two rounds of indexing are performed — the first at the transposition stage, and second at the PCR stage[73,136,152].

Our strategy, symmetrical strand sci (s3; Figure 16b, right) uses single-adapter transposition to incorporate the forward primer sequence, the Tn5 mosaic end sequence and a reaction-specific DNA barcode. As with standard tagmentation workflows, extension through the bottom strand is then performed to provide adaptor sequences on both ends of each molecule; however, the s3 transposome complexes contain a uracil base immediately following the mosaic end sequence. Use of a uracil-intolerant polymerase therefore prevents extension beyond the mosaic end into the DNA barcode and forward adaptor sequence. A second template oligo is then introduced that contains a 3'-blocked locked nucleic acid (LNA) mosaic end reverse complement sequence with a reverse adaptor sequence 5' overhang. This oligonucleotide favorably anneals to the copied mosaic end sequence, due to the higher melting temperature of LNA, and acts as a template for the library molecule to extend through and copy the reverse adaptor. This results in all library fragments having both a forward and reverse adaptor sequence. The LNA-templated extension is carried out over multiple rounds of thermocycling to ensure

maximum efficiency of reverse adaptor incorporation. Furthermore, adapter sequences are designed such that standard sequencing recipes can be used instead of the custom workflows and primers that are required for current indexed transposition technologies (Appendix Tables 6-10)[134,153].

We first sought to establish the s3 technique to assess chromatin accessibility. In s3-ATAC, nuclei are isolated and tagmented using our single-ended, indexed transposomes and carried through the adaptor-switching s3 workflow (Figure 16a). To ensure we attain true single-cell libraries without contamination from other nuclei, and minimal barcode collisions, we performed a mixed-species experiment on primary frozen human cortical tissue from the middle frontal gyrus and frozen mouse whole brain tissue (Figure 16b). We elected to perform this test on primary tissue samples instead of an idealized cell line setting to more accurately capture the rates of cross-cell contamination. Levels of crosstalk were assessed at both points of possible introduction: the tagmentation and PCR stages; by mixing nuclei from the two samples before tagmentation as well as after. Additionally, pure species libraries were produced by leveraging the inherent sample multiplexing capabilities of sci workflows. In the experimental condition where nuclei were mixed prior to any processing, *i.e*. pre-tagmentation, we observed a total estimated collision rate of 5.53% (Figure 17b,c; 2 × 2.77% detected human-mouse collisions), comparable to existing methods and tunable based on the number of nuclei deposited into each PCR indexing reaction. Zero collisions were observed in the post-tagmentation experimental conditions, suggesting no molecular crosstalk during s3 adapter switching or PCR.

In total, we generated 2,175 human and 837 mouse single-cell ATAC-seq profiles passing quality filters (Methods, Appendix Table 11) across four PCR indexing plates (Figure 16b). We then assessed the total unique sequence reads obtained per cell as a function of the total aligned reads, *i.e*. the library complexity. One of our mixed species plates was sequenced to beyond 50% saturation (duplicate reads / total reads), to represent the sequencing depth obtained where diminishing returns of increased sequence depth become excessive[136]. For the mouse cells, the mean sequencing saturation per cell was 63.6% and resulted in a median unique read count per cell of 178,069 (mean = 258,859). The human cells reached a mean

sequencing saturation of 56.6% with a median unique reads per cell of 99,882 (mean = 175,361). We additionally sequenced a plate that contained only human cells to a mean sequencing saturation of 70.4% which produced a median of 100,280 (mean = 146,937) unique reads per cell. When compared to other single-cell ATAC-seq datasets performed on mouse whole brain tissue, our mouse s3-ATAC libraries contain substantially greater reads per cell with 17.1x, 11.6x and 7.8x fold improvement compared to snATAC, 10X Genomics scATAC, and dscATAC, respectively (Figure 16b, Appendix Table 12)[29,154,155]. Read count increases can be indicative of poor ATAC-seq library quality, with increased depth reflecting increased noise and loss of signal at open chromatin regions. To address this, we first assessed read pair insert sizes, revealing the characteristic nucleosome-size banding distribution of ATAC-seq (Figure 16e)[65]. We next calculated transcription start site (TSS) enrichment using the approach defined by the ENCODE project (Methods). This produced significant enrichment for both species at 13.4 for human, well above the 'ideal' standard (>7) and 13.5 for mouse, within the acceptable range and just below ideal (>15). Similarly, the fraction of reads in pile-up genomic regions ("peaks"; FRiP) was comparable to other single-cell ATAC technologies at 31.95% and 29.15% as measured using 292,156 and 174,653 peaks for human and mouse cells respectively. However, FRiP is largely dependent on the number of peaks called, which influenced heavily by cell number and total sequence depth obtained. When expanding to a human cortex high-depth ATAC-seq peak set, a mean of 48.1% of reads were present in peaks, and mean of 78.2% of reads for mouse cells using a high-depth mouse brain ATAC-seq peak set (Methods).

With ample signal, we next sought to discern cell types present within the complex tissues. For each species, we used peaks called on aggregate data to construct a count matrix followed by dimensionality reduction using the topic-modeling tool *cisTopic*[74] which we then

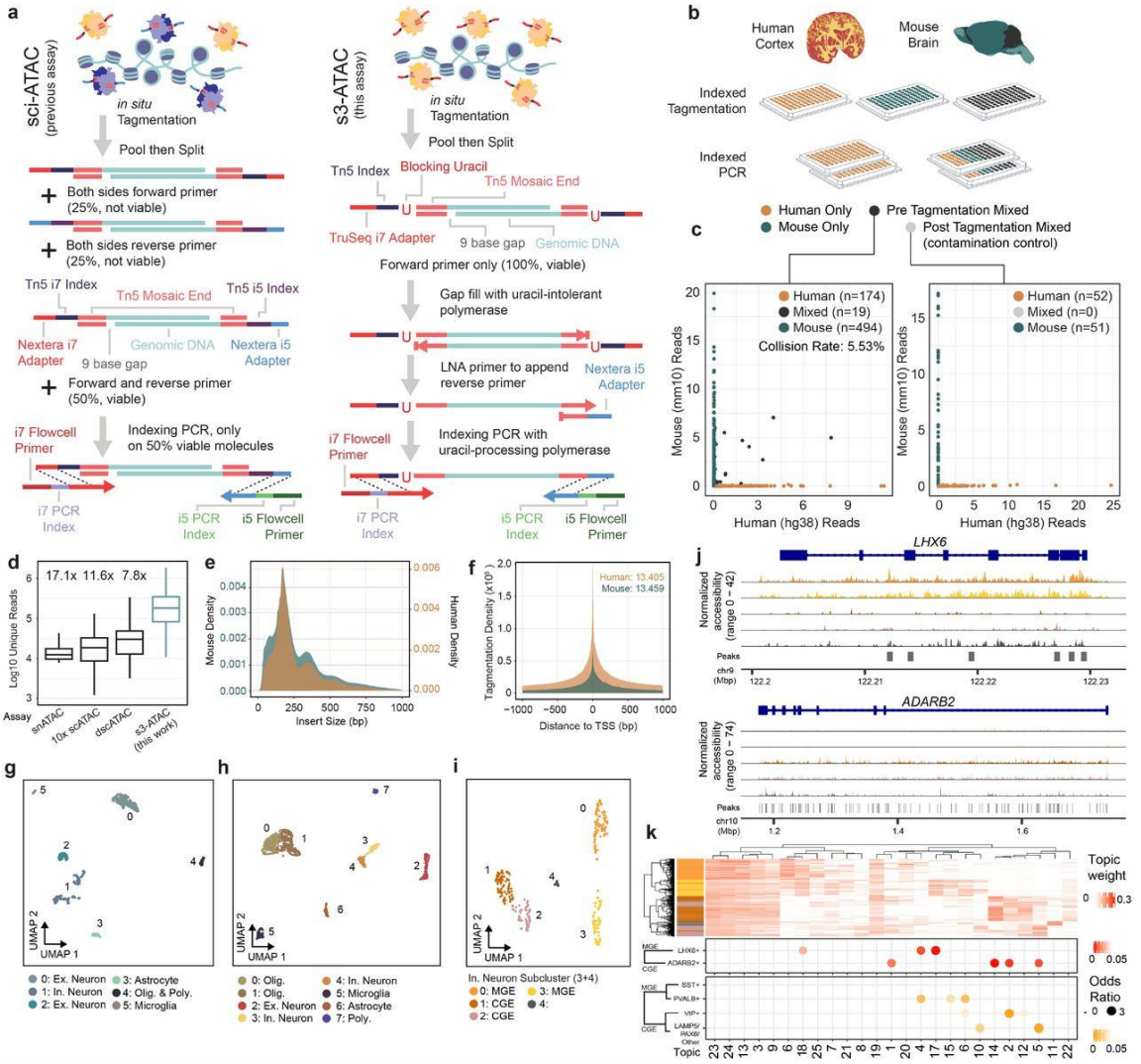


Figure 16. Symmetrical strand single-cell combinatorial indexing ATAC-seq (s3-ATAC). (a) Schematic of standard sci-ATAC library construction (left). Schematic of s3-ATAC library construction with intermediate steps of adapter switching leading to increased genomic molecule capture rate (right). (b) Experimental flow through and plate layout for the mixed-species experiment, including tagmentation and PCR plate conditions per well.(c) Point plots of single cell libraries with counts of unique reads aligned to mouse or human chromosomes in a chimeric reference genome. Points are colored to reflect species assignment (see Methods) in both pre-tagmentation mixing (left) and post-tagmentation mixing (right). (d) Comparison of library complexity for s3-ATAC mouse whole-brain sampled cells to previously reported data sets. All comparisons to our data are significantly less (Welch's two-sample t-test, p value <0.01). Fold improvement of our library complexity per method is listed above the method. (e) Insert size distribution of human and mouse libraries reflexing nucleosome banding.(f) Enrichment of reads at transcription start sites ("TSS") for human and mouse libraries with enrichment calculation following ENCODE standard practices. (g) UMAP projection of mouse whole brain cell samples (n=837 cells) colored by cluster and cell type assignment. (h) UMAP projection human cortex cell samples (n=2,175 cells). (i) Subclustering and UMAP projection of human cortical inhibitory neurons (clusters 3 and 4 from panel h., n=342) (j) Genome coverage track of human inhibitory neurons (n=342) aggregated over 5 subclusters for genomic locations overlapping MGE and CGE marker genes *LHX6* and *ADARB2*, respectively.(k) Hierarchical clustering of topic weight per cell (top). Hypergeometric test of gene set analysis enrichment for human inhibitory neuron marker genes (bottom; Fisher's exact test, see Methods).

visualized using UMAP[76], performed graph-based clustering at the topic level, and processed via *Signac*[156]. Clear separation of cell types was observed using marker gene signal and differential accessibility profiles (Figure 16g-h, Appendix Figure 26, Appendix Table 13)[29,34].

Notably, even with the modest cell count produced by this experiment, the quality improvements allow us to interrogate subclusters of inhibitory neurons previously difficult to distinguish in atlas-level datasets (Figure 16i)[67]. With our improved cell depth, we were able to discern caudal and medial ganglionic eminence inhibitory neurons by marker gene coverage plots across 342 *GAD1*+ cells ("CGE" and "MGE", respectively; Figure 16j). From these, we identified 157 *GAD1*+, *ADARB2*+ CGE cells and 168 *GAD1*+,*LHX6*+ MGE cells. We separated 17 cells (subcluster 4) with apoptotic stress markers likely due in part to post-mortem sampling, which could potentially compound common single cell ATAC analyses. Aggregated genomic signal over our Topic-based dimensionality reduction was used to support our marker gene cell subtype discrimination and describe differentially accessible loci in human cortical inhibitory neurons (Figure 16k).

We then extend the improvements in data quality produced by s3-ATAC to other sci-workflows. This includes our previously-described sci-DNA-seq method[136] that produces single-cell whole genome sequencing libraries (s3-WGS) and a novel strategy to incorporate the core components of HiC library preparation but without ligation junction enrichment to produce whole genome and chromatin conformation information (s3-GCC; Figure 17a). Both strategies disrupt nucleosomes to acquire sequence reads uniformly across the genome[136]. We first tested s3-WGS by producing two small-scale libraries on the euploid lymphoblastoid cell line, GM12878. The first library comprised only four wells at the PCR stage for a target of 60 cells, allowing us to sequence the library to high depth (Figure 17b). This produced a median passing read count per cell of 12,789,812 (mean = 15,238,184), across 45 QC-passing cells (75% cell capture efficiency). With our sequenced library at 72.35% saturation; our complexity is notably higher than the predecessor sci-DNA-seq technology which produced a median of 43,367 reads per cell (mean = 103,138) at the same sequencing saturation (295 and 148 fold improvement in median and mean, respectively; Figure 17b)[136]. The second preparation performed comparably, though sequenced to a lower total depth (15.98% saturation). We also confirmed that the coverage was uniform by assessing the median absolute deviation (MAD) across 500 kbp bins, which fell within 0.152 ± 0.025 (mean ± s.d.), comparable to other single-cell genome sequencing techniques (Figure 17d)[87,90,136].

We performed s3-WGS and s3-GCC on two cultures of a cell line derived from a primary pancreatic ductal adenocarcinoma (PDAC) tumor (Figure 17b). PDAC is a highly-aggressive cancer that typically presents at an advanced stage, making early detection and study of tumor progression key[157]. PDAC studies suffer from a low cancer cell fraction, thus we used a patient derived cell line (PDCL) generated directly from tumor and maintained at fewer than 10 passages. This method allows for multiple modalities of characterization and perturbation, while maintaining the heterogeneity present in the tumor sample[158]. We profiled two PDCL cultures (referred to as PDAC-1 and PDAC-2) to capture variance that may arise during passaging from a parent line derived from a tumor harboring a driver mutation in the oncogene *KRAS* (p.G12D). For our s3-WGS preparations, we produced 773 and 256 single-cell libraries with a mean passing read counts of 1,181,128 and 1,299,949 for PDAC-1 and 2 (at a combined median of 28.46% saturation), respectively. The s3-GCC libraries contained 57 and 145 cells produced a mean passing read count of 973,397 and 1,588,926 (combined median 73.25% sequencing saturation, Appendix Table 13) for PDAC-1 and 2, respectively (Figure 17e). MAD scores for the two lines were greater than that of the euploid karyotype of GM12878, 0.219 ± 0.041 (mean ± s.d.); however, this is expected given the widespread copy number alterations present in the samples. In addition to the WGS component, the s3-GCC libraries also contained reads that were identified as chimeric ligation junctions that provide HiC-like chromatin conformation signal. Across both samples, we identified a mean of 118,048 reads per cell that capture genomic contacts at least 50 kbp apart from one another, a 14.8-fold improvement over the previous single-cell combinatorial indexing technique, sci-HiC[159] (Figure 17f). Read pairs spanning ≥50 kbp accounted for a median of 15.6% and 17.0% of the total reads obtained per cell, which equates to an enrichment of 361- and 402-fold over that of the s3-WGS libraries for PDAC-1 and 2, respectively.

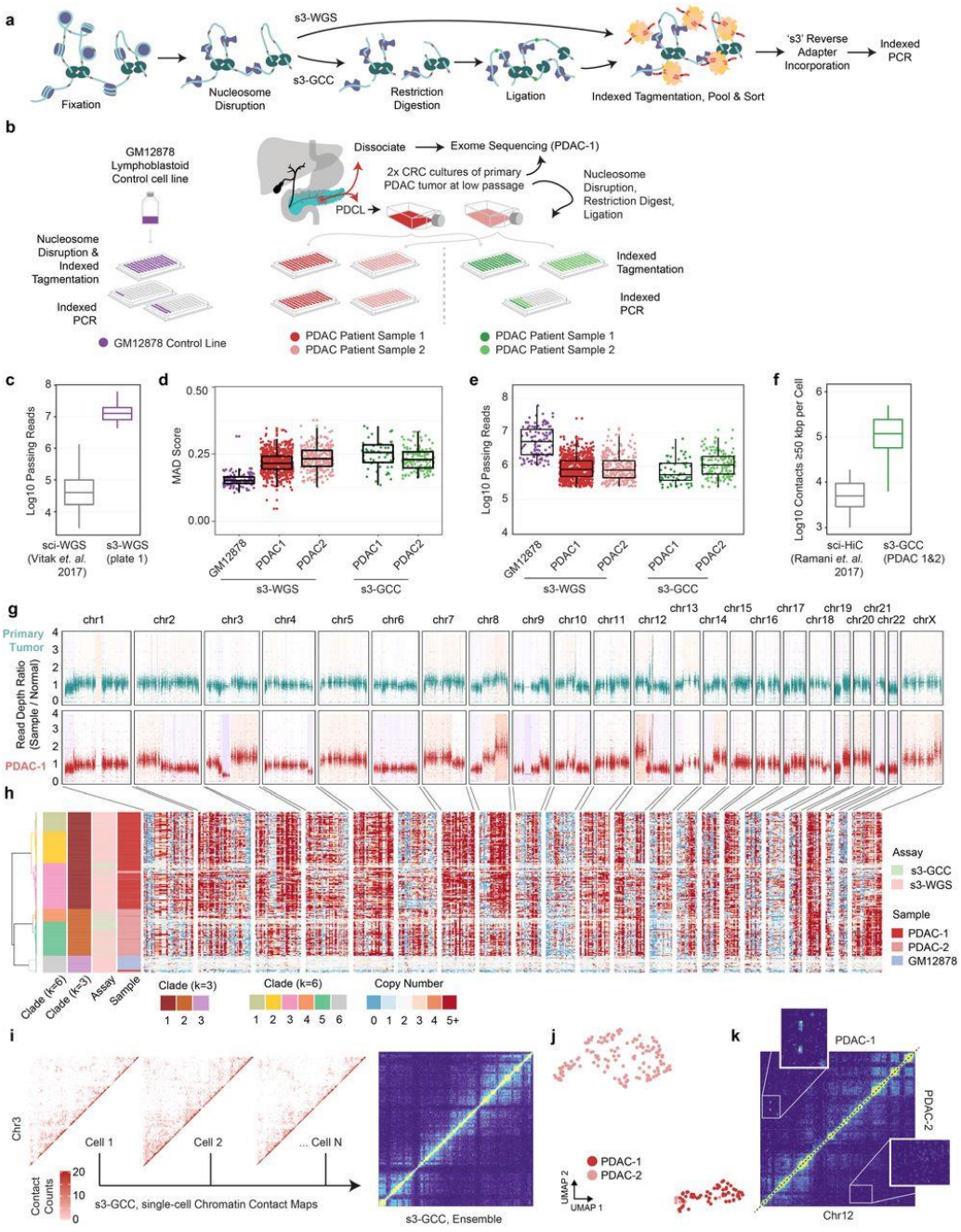


Figure 17. s3 whole genome sequencing (s3-WGS) and genome conformation capture (s3-GCC). (a) Schematic of sci-WGS and sci-GCC library construction. (b) Experimental flow through and plate layout for PDAC and control diploid line. (c) Boxplot of read count per cell for matched GM12878 cell line**9**. (d) Boxplot of MAD score per cell per sample and assay. (e) Boxplot of reads passing filter per cell. (f) Comparison boxplot of s3-GCC and sci-HiC distal contacts (≥50kbp) per cell**33**. (g) Whole exome sequencing of the primary tumor and PDCL. Scatterplot of reads per bin with a shading of called copy number variation. (h) Single-cell whole genome copy number calling on 500 kbp bins genome-wide. Cells (rows) are hierarchically clustered and annotated by assay, sample, and assigned clade (left). (i) Representative single-cell contact maps (raw counts) at 1 Mbp resolution for chromosome 3 and ensemble contact map profile at 500 kbp resolution. (j) scHiCRep dimensionality reduction and clustering of single-cell distal contact profiles. (k) Subclonal translocation on chr12 specific to PDAC-1.

We first focused our analysis on the s3-WGS and the WGS component of the s3-GCC libraries to examine the copy number alterations present. To get a sense of the genomic landscape, we first performed copy number calling on whole exome sequencing (WES) libraries that were generated using primary tumor tissue and in bulk on the PDCL line (Figure 17g). This revealed a profile of copy number aberrations at finer resolution, with a more pronounced profile in the PDCL sample, likely due in part to the absence of euploid stromal cell contamination. We then processed all single-cell libraries using *SCOPE*[90] which revealed a highly altered genomic landscape within each of the two samples. In line with paired karyotyping and bulk exome data, we see a similar pattern per cell of multi-megabasepair copy number aberrations when performing breakpoint analysis on 500 kbp windows, with a median depth per window of 81 reads. Using the inferred copy number profile within genomic windows for the three samples, GM12878 and the two PDCL cultures, we performed hierarchical and K-means clustering on the Jaccard distance between cell breakpoint copy numbers at two different centroid counts. For our optimal centroid value, we found a relatively clean separation between cell lines (k=3), for subclonal analysis we used a higher centroid count at a local optima (k=6). s3-WGS and s3-GCC cells cluster dependent on PDCL culture, reflecting our ability to capture genome-wide copy number data in our s3-GCC libraries (Figure 17h). We generated pseudo-bulk clades from the single-cell read count bins, with an average of 211.3 cells per clade and an average read count of 3,750 per 50 kbp bin. This revealed multiple fixed and subclonal genomic arrangements (Appendix Figure 27). In PDAC-1 and PDAC-2 we see shared copy number loss of tumor suppressor genes *CDKN2A, SMAD4* and *BRCA2*[157,160]. In PDAC-2 we observed a subclonal amplification of *PRSS1*, a mutation that was fixed within our sampling for PDAC-1 and is associated with tumor size and a higher tumor node metastasis (TNM) stage[161]. This suggests that while the lines have the same origin, each culture captured different subsets of tumor clonal populations.

Duplications and deletions are not the sole form of genomic rearrangement that may induce a competitive advantage in cancer cell growth. Genomic inversions are difficult to assess through standard karyotyping and chromosome painting methods, whereas chromosomal

translocations are difficult to uncover in whole-genome amplification methods, since only reads capturing the breakpoint would provide supportive evidence. To address both of these limitations, we utilized the HiC-like component of our s3-GCC libraries. Using read pairs spanning ≥50 kbp, we produced chromatin contact maps that produced clear chromatin compartmentalization signal (Figure 17i)[159]. Single cells were separated by their distal contact information via *HiCRep* and observed distinct clusters by PDCLs[108]. Notably, even at this low sequencing depth, we were able to reliably tell PDCL line sparse contact profiles apart (Figure 17j, Appendix Figure 28). Differences between the aggregated contact maps between clusters were then used to assess unique translocation and inversion events across the sampled cells. We found that our single-cell contact data uncovers an intrachromosomal translocation between the 8.5-9.5 Mbp and 88.5-91.0 Mbp regions of chromosome 12 (Figure 17k), containing *ATP2B1*, which is commonly overexpressed in PDAC[162] and the tumor suppressor gene *DUSP6*[163] that is only present in PDAC-1.

Table 2. Summary of results using s3 protocols.

Assay	Cells Captured	Fold improvement of information per cell over previous protocol	Utility demonstration
s3-ATAC	3,012	7.8X reads per cell over dscATAC-seq	Subclustering of cell subtypes in the cortex at low cell count
s3-WGS	1074	295X reads per cell over sci-DNA-seq	Assessment of high resolution copy number changes in a tumor-derived sample
s3-GCC	202	14.8X distal contacts captured over sci-HiC	Identification of putative translocations in a tumor-derived sample

Taken together, our s3 workflow represents marked improvements over the predecessor sci platform with respect to passing reads obtained per cell without sacrificing signal enrichment in the case of s3-ATAC, or coverage uniformity for s3-WGS (Table 2). We also introduce another variant of combinatorial indexing workflows, s3-GCC to obtain both genome sequencing and chromatin conformation, with improved chromatin contacts obtained per cell when compared to

sci-HiC. We demonstrate the utility of these approaches by assessing two patient-derived tumor cell lines with genomic instability. Our analysis reveals patterns of focal amplification for disease-relevant genes, and uncover wide-scale heterogeneity at a throughput not attainable with standard karyotyping. Additionally, we highlight the joint analysis of our protocols for uncovering the chromatin compartment disrupting effect of copy number aberrations. Furthermore, the s3 workflow has the same inherent throughput potential of standard single-cell combinatorial indexing, with the ability to readily scale into the tens and hundreds of thousands of cells by expanding the set of transposome and PCR indexes. We also expect that this platform will be compatible with other transposase-based techniques, including sci-MET[152], or CUT&Tag[164]. Lastly, unlike sci workflows, the s3 platform does not require custom sequencing primers or custom sequencing recipes, removing one of the major hurdles that groups may face while implementing these technologies.

Methods

PDCL propagation

Low-passage, patient-derived cell lines (PDCLs) were propagated from rapidly dissociated PDAC tumors and cultured for continuous propagation in culture medium containing ROCK inhibitor (Y-276320)[165]. Briefly, approximately 50,000 viable, disaggregated tumor cells were plated to a 35mm diameter, collagen-coated well (Gibco, A11428-02) and passaged 1:3 while subconfluent until reaching 85% confluence on a 10cm diameter dish. From a fraction of these cells, DNA was extracted to validate the presence of KRAS-G12 mutations by ddPCR (Bio-Rad, 1863506) and to validate an STR profile that matches normal leukocyte DNA from the same patient (Genetica). PDCLs exhibited morphologies consistent with epithelial tumor cells and abundant KRT expression was detected by immunocytofluorescence using the monoclonal antibodies: AE1/AE3, C-11, and Cam5.2.

Whole Exome Sequencing and Analysis

Whole exome sequencing libraries for the patient blood sample, tumor biopsy, and PDCL were carried out by the Knight Diagnostic Research Cytogenetics Lab at OHSU. Libraries were prepared using 500 ng of fragmented gDNA using KAPA Hyper-Prep Kit (KAPA Biosystems) with

Agilent SureSelect XT Target Enrichment System and Human All Exon V5 capture baits (Agilent Technologies), following manufacturer's protocols. Sequencing was carried out using the Illumina HiSeq 2500 platform by the OHSU Massively Parallel Sequencing Shared Resource (MPSSR). Paired-end reads were aligned with *bwa mem* (v0.7.15-r1140) to GRCh38 ("hg38",Genome Reference Consortium Human Reference 38 (GCA_000001405.2))[166]. The data was processed following the best practices workflow for the GATK pipeline (v4.1.9.0)[93]. Exome regions annotated as "protein-coding" were extracted from GenCode (v35)[167] and used as the intervals for processing. The following commands were then used for WES data normalization and segmentation with additional options were specified: *PreprocessInvertals, CollectReadCounts, AnnotateIntervals, FilterIntervals, CreateRedCountPanelOfNormals* (using the matched blood sample as the normal, with *minimum-interval-median-percentile* set to 5.0), and finally *PlotDenoisedCopyRatios*. The output was then plotted with *ggplot2* (v3.3.2) in *R* (v4.0.0). The *geom_rect* function was used to shade the genomic region based on the relative copy number with segmentation interval, and *geom_point* was used to plot normalized bin reads.

s3-ATAC Library Generation

Prior to sample handling, 96 uniquely indexed transposome complexes were assembled using previously-described methods[134]. Complexes were diluted to 2.5uM in a protein storage buffer composed of 50% (v/v) glycerol (Sigma G5516), 100 mM NaCl (Fisher Scientific S271-3), 50 mM Tris pH 7.5 (Life technologies AM9855), 0.1 mM EDTA (Fisher Scientific AM9260G), 1 mM DTT (VWR 97061-340), and stored at −20°C. At the time of nuclei dissociation, 50mL of nuclei isolation buffer (NIB-HEPES) was freshly prepared with final concentrations of 10 mM HEPES-KOH (Fisher Scientific, BP310-500 and Sigma Aldrich 1050121000, respectively), pH 7.2, 10 mM NaCl, 3mM MgCl2 (Fisher Scientific AC223210010), 0.1 % (v/v) IGEPAL CA-630 (Sigma Aldrich I3021), 0.1 % (v/v) Tween (Sigma-Aldrich P-7949) and diluted in PCR-grade Ultrapure distilled water (Thermo Fisher Scientific 10977015). After dilution, two tablets of Pierce(tm) Protease Inhibitor Mini Tablets, EDTA-free (Thermo Fisher A32955) were dissolved and suspended to prevent protease degradation during nuclei isolation.

For s3-ATAC tissue handling, primary samples of C57/B6 mouse whole brain were extracted and flash frozen in a liquid nitrogen bath, before being stored at −80°C. Human cortex samples from the middle frontal gyrus were sourced from the Oregon Brain Bank from a 50-year-old female of normal health status. Tissue was collected at 21 hours post-mortem and then placed in a −80°C freezer for storage. An at-bench dissection stage was set up prior to nuclei extraction. A petri dish was placed over dry ice, with fresh sterile razors pre-chilled by dry-ice embedding. 7mL capacity dounce homogenizers were filled with 2mL of NIB-HEPES buffer and held on wet ice. Dounce homogenizer pestles were held in in ice cold 70% (v/v) ethanol (Decon Laboratories Inc 2701) in 15mL tubes on ice to chill. Immediately prior to use, pestles were rinsed with chilled distilled water. For tissue dissociation, mouse and human brain samples were treated similarly. The still frozen block of tissue was placed on the clean pre-chilled petri dish and roughly minced with the razors. Razors were then used to transport roughly 1 mg the minced tissue into the chilled NIB-HEPES buffer within a dounce homogenizer. Suspended samples were given 5 minutes to equilibrate to the change in salt concentration prior to douncing. Tissues were then homogenized with 5 strokes of a loose (A) pestle, another 5 minute incubation, and 5-10 strokes of a tight (B) pestle. Samples were then filtered through a 35 µm cell strainer (Corning 352235) during transfer to a 15mL conical tube, and nuclei were held on ice until ready to proceed. Nuclei were pelleted with a 400 rcf centrifugation at 4°C in a centrifuge for 10 minutes. Supernatant was removed and pellets were resuspended in 1mL of NIB-HEPES buffer. This step was repeated for a second wash, and nuclei were once again held on ice until ready to proceed. A 10uL aliquot of suspended nuclei was diluted in 90uL NIB-HEPES (1:10 dilution) and quantified on either a Hemocytometer or with a BioRad TC-20 Automated cell counter following manufacturer's recommended protocols. The stock nuclei suspension was then diluted to a concentration of 1400 nuclei/uL.

Tagmentation plates were prepared by the combination of 420 uL of 1400 nuclei/uL solution with 540 uL 2X TD Buffer (Nextera XT Kit, Illumina Inc. FC-131-1024). From this mixture, 8uL (~5000 nuclei total) was pipetted into each well of a 96 well plate dependent on well schema (Figure 16b). 1uL of 2.5uM uniquely indexed transposase was then pipetted into each well.

Tagmentation was performed at 55°C for 10 minutes on a 300 rcf Eppendorf ThermoMixer. Following this incubation, plate temperature was brought down with a brief incubation on ice to stop the reaction. Dependent on experimental schema pools of tagmented nuclei were combined and 2uL 5mg/mL DAPI (Thermo Fisher Scientific D1306) was added.

Nuclei were then flow sorted via a Sony SH800 to remove debris and attain an accurate count per well prior to PCR. A receptacle 96 well plate was prepared with 9uL 1X TD buffer (Nextera XT Kit, Illumina Inc. FC-131-1024, diluted with ultrapure water), and held in a sample chamber kept at 4°C. Fluorescent nuclei were then flow sorted gating by size, internal complexity and DAPI fluorescence for single nuclei following the same gating strategy as previously described[38]. Immediately following sorting completion, the plate was sealed and spun down for 5 minutes at 500 rcf and 4°C to ensure nuclei were within the buffer.

Nucleosomes and remaining transposases were then denatured with the addition 1uL of 0.1% SDS (~0.01% f.c.) per well. 4uL of NPM (Nextera XT Kit, Illumina Inc) per well was subsequently added to perform gap-fill on tagmented genomic DNA, with an incubation at 72°C for 10 minutes. 1.5 uL of 1uM A14-LNA-ME oligo was then added to supply the template for adapter switching. The polymerase based adapter switching was then performed with the following conditions: initial denaturation at 98°C for 30 seconds, 10 cycles of 98°C for 10 seconds, 59°C for 20 seconds and 72°C for 10 seconds. The plate was then held at 10°C. After adapter switching 1% (v/v) Triton-X 100 in ultrapure H2O (Sigma 93426) was added to quench persisting SDS. At this point, some plates were stored at −20°C for several weeks while others were immediately processed.

The following was then combined per well for PCR: 16.5 ul sample, 2.5uL indexed i7 primer at 10 uM, 2.5uL indexed i5 primer at 10 uM, 3 uL of ultrapure H2O, and 25 uL of NEBNext Q5U 2X Master mix (New England Biolabs M0597S), and 0.5uL 100X SYBR Green I (Thermo Scientific S7563) for a 50 uL reaction per well. A real time PCR was performed on a BioRad CFX with the following conditions, measuring SYBR fluorescence every cycle: 98°C for 30 seconds; 16-18 cycles of 98°C for 10 seconds, 55°C for 20 seconds, 72°C for 30 seconds, fluorescent

reading, 72°C for 10 seconds. After fluorescence passes an exponential growth and begins to inflect, the samples were held at 72°C for another 30 seconds then stored at 4°C.
Amplified libraries were then cleaned by pooling 25 uL per well into a 15 mL conical tube and cleaned via a Qiaquick PCR purification column following manufacturer's protocol (Qiagen 28106). The pooled sample was eluted in 50 uL 10 mM Tris-HCl, pH 8.0. Library molecules then went through a size selection via SPRI selection beads (Mag-Bind® TotalPure NGS Omega Biotek M1378-01). 50 uL of vortexed and fully suspended room temperature SPRI beads was combined with the 50 uL library (1X clean up) and incubated at room temperature for 5 minutes. The reaction was then placed on a magnetic rack and once cleared, supernatant was removed. The remaining pellet was rinsed twice with 100 uL fresh 80% ethanol. After ethanol was pipetted out, the tube was spun down and placed back on the magnetic rack to remove any lingering ethanol. 31 uL of 10 mM Tris-HCl, pH 8.0 was then used to resuspend the beads off the magnetic rack and allowed to incubate for 5 minutes at room temperature. The tube was again placed on the magnetic rack and once cleared, the full volume of supernatant was moved to a clean tube. DNA was then quantified by Qubit dsDNA High-sensitivity assay following manufacturer's instructions (Thermo Fisher Q32851). Libraries were then diluted to 2ng/uL and run on an Agilent Tapestation 4150 D5000 tape (Agilent 5067-5592). Library molecule concentration within the range of 100-1000bp was then used for final dilution of the library to 1 nM. Diluted libraries were then sequenced on High or Mid capacity 150 bp sequencing kits on the Nextseq 500 system following manufacturer's recommendations (Illumina Inc. 20024907, 20024904). For greater sequencing effort, select libraries were also sequenced on a NovaSeq S2 flowcell, again following manufacturer's recommendations (Illumina Inc. 20028315). For both machines libraries were sequenced as paired-end libraries with 10 cycle index reads and 85 cycles for read 1 and read 2.

s3-WGS Library Generation

Prior to processing the following buffers were prepared: 50mL of NIB HEPES buffer as described above, as well as 50mL of a Tris-based NIB (NIB Tris) variant with final concentrations of 10 mM Tris HCl pH 7.4, 10 mM NaCl, 3mM MgCl2, 0.1 % (v/v) IGEPAL CA-630, 0.1 % (v/v) Tween and diluted in PCR-grade Ultrapure distilled water. After dilution, two tablets of Pierce(tm)

Protease Inhibitor Mini Tablets, EDTA-free were dissolved and suspended to prevent protease degradation during nuclei isolation.

s3-WGS library preparation was performed on cell lines as follows. For patient derived PDCL cell lines, cells were plated at a density of 1×10^6 on a T25 flask the day prior to processing. At harvest, cells were washed twice with ice cold 1X PBS (VWR 75800-986) and then trypsinized with 5mL 1X TrypLE (Thermo Fisher 12604039) for 15 minutes at 37°C. Suspended cells were then collected and pelleted at 300 rcf at 4°C for 5 minutes. For suspension-growth cell lines (GM12878), cells were pipetted from growth media and pelleted at 300 rcf at 4°C for 5 minutes. Following the initial pellet, cells were washed with ice cold 1mL NIB HEPES twice. After the second wash, pellets were then resuspended in 300 uL NIB HEPES. Nuclei were aliquoted and quantified as described above, then aliquots of 1 million nuclei were generated based on the quantification. The aliquots were pelleted by a 300 rcf centrifugation at 4°C for 5 minutes and resuspended in 5 mL NIB HEPES. 246 uL 16% (w/v) formaldehyde (Thermo Fisher 28906) was then added to nuclear suspensions (f.c. 0.75% formaldehyde) to lightly fix nuclei. Nuclei were fixed via incubation in formaldehyde solution for 10 minutes on an orbital shaker set to 50 rpm. Suspensions were then pelleted at 500 rcf for 4 minutes at 4°C and supernatant was aspirated. Pellet was then resuspended in 1 mL of NIB Tris Buffer to quench remaining formaldehyde. Nuclei were again pelleted at 500 rcf for 4 minutes at 4°C and supernatant was aspirated. The pellet was washed once with 500uL 1X NEBuffer 2.1 (NEB B7202S) and then resuspended with 760 uL 1X NEBuffer 2.1. 40 uL 1% SDS (v/v) was added and sample was incubated on a ThermoMixer at 300 rcf set to 37°C for 20 minutes. Nucleosome depleted nuclei were then pelleted at 500 rcf at 4°C for 5 minutes and then resuspended in 50 uL NIB Tris. A 5 uL aliquot of nuclei was taken and diluted 1:10 in NIB Tris then quantified as described above. Nuclei were diluted to 500 nuclei/uL with addition of NIB Tris, based on the quantification. Dependent on experimental setup, the 420 uL of nuclei at 500 nuclei/uL were then combined with 540 uL 2X TD buffer. Following this, nuclei were tagmented, stained and flow sorted, genomic DNA was gap-filled and adapter switching was performed as described for the s3-ATAC protocol. Library amplification was performed by PCR as described above with fewer total cycles (13-15) likely due

to more initial capture events per library. Libraries were then cleaned, size selected, quantified and sequenced as described previously.

s3-GCC Library Generation

The same cultured cell line samples were harvested as described for s3-WGS library generation, and processed from the same pool of fixed, nucleosome depleted nuclei. Following quantification of nuclei, the full remaining nuclear suspensions (~2-3 million nuclei per sample) were pooled respective of sample. Nuclei were pelleted at 500 rcf at 4°C for 5 minutes and resuspended in 90 uL 1X Cutsmart Buffer (NEB B7204S). 10 uL of 10U/uL AluI restriction enzyme (NEB R0137S) was added to each sample. Samples were then digested for 2 hours at 37°C at 300 rpm on a ThermoMixer. Following digestion, nuclear fragments then underwent proximity ligation. Nuclei were pelleted at 500 rcf at 4°C for 5 minutes and resuspended in 100uL ligation reaction buffer. Ligation buffer is a mixture with final concentrations of 1X T4 DNA Ligase Buffer + ATP (NEB M0202S), 0.01 % TritonX-100, 0.5mM DTT (Sigma D0632), 200 U of T4 DNA Ligase, diluted in ultrapure H2O. Ligation took place at 16°C for 14 hours (overnight). Following this incubation, nuclei were pelleted at 500 rcf at 4°C for 5 minutes and resuspended in 100 uL NIB HEPES buffer. An aliquot of nuclei were quantified as described previously, and were then diluted, aliquoted, tagmented, pooled, DAPI stained, flow sorted, genomic DNA was gap-filled and adapter switching was performed as described for the s3-ATAC protocol. Library amplification occurred at the same rate as the s3-WGS libraries (13-15 cycles) and libraries were subsequently pooled, cleaned, quantified and sequenced as described above.

Computational Analysis

Preprocessing

The initial processing of all library types was the same. After sequencing, data was converted from bcl format to FastQ format using *bcl2fastq* (v 2.19.0, Illumina Inc.) with the following options *with-failed-reads, no-lane-splitting*, f*astq-compression-level=9, create-fastq-for-index-reads*. Data were then demultiplexed, aligned, de-duplicated using the in-house *scitools* pipeline (ref [31]). Briefly, FastQ reads were assigned to their expected primer index sequence allowing for sequencing error (Hamming distance ≤2) and indexes were concatenated

to form a “cellID”. Reads that could be assigned unambiguously to a cellID were then aligned to reference genomes. For s3-WGS and s3-GCC libraries, paired reads were aligned with *bwa mem* (v0.7.15-r1140) to hg38[166]. For s3-ATAC libraries, reads were first aligned to a concatenated hybrid genome of hg38 and GRCm38 (“mm10”, Genome Reference Consortium Mouse Build 38 (GCA_000001635.2)). Reads were then de-duplicated to remove PCR and optical duplicates by a *perl* (v5.16.3) script aware of cellID, chromosome and read start, read end and strand. From there putative single-cells were distinguished from debris and error-generated cellIDs by both unique reads and percentage of unique reads.

s3-ATAC Analysis

Barnyard Analysis

With single-cell libraries distinguished, we next quantified contamination between nuclei during library generation. We calculated the read count of unique reads per cellID aligning to either human reference or mouse reference chromosomes (Figure 16c). CellIDs with ≥90% of reads aligning to a single reference genome were considered *bona fide* single-cells. Those not passing this filter (2.7%,19/687 cells for pre-tagmentation barnyard) were considered collisions. Collision rate was estimated to account for cryptic collisions (mouse cell-mouse cell or human cell-human-cell) by multiplying by two (final collision rate of 5.5%). *Bona fide* single-cell cellIDs were then split from the original FastQ files to be aligned to the proper hg38 or mm10 genomes with *bwa mem* as described above. Human and mouse assigned cellIDs were then processed in parallel for the rest of the analysis. After alignment, reads were again de-duplicated to obtain proper estimates of library complexity.

Tagmentation Insert Quantification

To assess tagmentation insert size, *samtools isize* (v. 1.10) was performed and plotted with *ggplot2* (v3.3.2) in *R* (v4.0.0) using the *geom_density* function (default parameters, Figure 16e). To assess library quality further, we generated tagmentation site density plots centered around transcription start sites (TSSs). We used the alignment position (chromosome and start site) for each read to generate a bed file that was then piped into the BEDOPS closest-feature command mapped the distance between all read start sites and transcription start sites (v

2.4.36)[168]. From this, we collapsed binned distances (100bp increments) into a counts table and generated percentage of read start site distances within each counts table. We plotted these data using R and *ggplot2 geom_density* function (default parameters) subset to 2000 base pairs around the start site to visualize enrichment. TSS enrichment values were calculated for each experimental condition using the method established by the ENCODE project (https://www.encodeproject.org/data-standards/terms/enrichment), whereby the aggregate distribution of reads ±1,000 bp centered on the set of TSSs is then used to generate 100 bp windows at the flanks of the distribution as the background and then through the distribution, where the maximum window centered on the TSS is used to calculate the fold enrichment over the outer flanking windows.

Library Complexity Analysis

To project library complexity through sequencing effort, pre-de-duplicated cellID read sets were used to build a projection as follows**8**. Reads were randomly subsampled starting at 1% of the total reads with 5% of data added in increasing increments to build a simple saturation curve per cellID. A summarized saturation curve per species was generated and plotted in *ggplot2* using the *geom_smooth* function, descripting the curves mean, median and standard error. For comparison to publicly available data sets of a matched tissue type, we focused our analysis on the mouse brain libraries. We plotted our PCR plate sequenced to 36.4% ± 17.4% unique reads/total reads for comparison to three other single-cell ATAC-seq methods which have been applied to post-natal mouse whole brain[29,154,155]. Data passing self-reported filters were used for comparison and plotted with *ggplot geom_boxplot* function. Welch's two-sample T test comparisons between unique reads per cell were calculated with the *t.test* function in base *R* for a one-sided alternative hypothesis.

Dimensionality Reduction

Pseudo-bulked data (agnostic of cellID) was then used to call read pile-ups or "peaks" via macs2 (v.2.2.7.1) with option –keep-dup all[72]. Narrowpeak bed files were then merged by overlap and extended to a minimum of 500bp for a total of 292,156 peaks for human and 174,653 peaks for mouse. A scitools perl script was then used to generate a sparse matrix of **peaks × cellID** to

count occurrence of reads within peak regions per cell. FRiP was calculated as the number of unique, usable reads per cell that are present within the peaks out of the total number of unique, usable reads for that cell for each peak bed file. Cells with less than 20% of reads within peaks were then filtered out. Tabix formatted files were generated using *samtools* and *tabix* (v1.7). The counts matrix and tabix files were then input into a SeuratObject for *Signac* (v1.0.0) processing[156,169].We performed LDA-based dimensionality reduction via *cisTopic* (v0.3.0)with 27 topics for mouse cells and 24 topics for human cells[74]. The number of topics were selected after generating 25 separate models per species with topic counts of 5,10,20-30,40,50,55,60-70 and selecting the topic count using selectModel based on the second derivative of model perplexity. Cell clustering was performed with Signac *FindNeighbors* and *FindClusters* functions on the **topic weight × cellID** data frame. For *FindClusters* function call, resolution was set to 0.3 and 0.2 for human and mouse samples, respectively. The respective **topic weight × cellID** was then projected into two dimensional space via a uniform manifold approximation and projection ("UMAP") by the function *umap* in the *uwot* package (v0.1.8, Figure 16g-h)[76]. Cis-coaccessibility networks (CCANs) were generated through the Signac wrapper of *cicero* (v1.3.4.10)[20]. Genome track plots with CCAN linkages were generated through *Signac* function *CoveragePlot* for marker genes previously described[156]. Differential accessibility between clusters in one by one, and one by rest comparisons were generated using *Signac* function *FindMarkers* using options: test.use = 'LR', and only.pos=T, with latent.vars = 'nCount_peaks', to account for read depth. Cell type per cluster was assigned based on genome track plots and differentially accessible sites.

Subclustering

After gross cell type assignment of mouse and human cell lines, human inhibitory neurons (GAD1+) clusters 3 and 4 were subset from the SeuratObject. Those 342 cells were then iteratively clustered by performing the same cisTopic, UMAP, and Signac processing with the following changes[74,76,156]. CisTopic was performed on the full set of human peaks (292,156) with those 342 subset cells. 12 Topic models were constructed (5, 10, 20-30 topics) and the 25 topic model was chosen on the second derivate of the model perplexity. A resolution of 0.5 was used in the Signac *FindClusters* on the **topic weight × cellID** call to attain 5 subclusters. Coverage plots

were generated as reported above for *ADARB2* and *LHX6* . Peaks were then assigned to topics using the cisTopic *binarizecisTopics* function with argument thrP=0.975 (mean count per topic: 2429 peaks). We then performed a simple gene set enrichment analysis on human cortical inhibitory neurons and subtypes based on RNA-identified marker genes defined previously[34].We used a Fisher's Exact test with the function *fisher.test* with function alternative.hypothesis = "greater" to look for enrichment of topic-assigned peaks in marker gene bodies for inhibitory neuron subclasses relative to all topic-assigned peaks. We filtered results to those with nominal enrichment (p value ≤ 0.05) and used *ggplot geom_point* with color reflecting the reported p-value and size proportional to odds ratio to generate a bubble plot (Figure 16k).

s3-WGS and s3-GCC Analysis

Quality Control

s3-WGS and s3-GCC cellIDs were initially filtered to samples with either ≥1×10^5 or ≥1×10^6 unique reads (PDCL and GM12878 samples, respectivley). CellIDs were split after de-duplication into single-cell bam files. They were then processed via the pipeline in the package *SCOPE* (v1.1)[90]. The genome was split into 500 kbp bins with each bin being assigned a GC content and mappability score (generated through CODEX2)[170]. Reads with a mapping quality of Q ≥ 10 were counted in bins per cellID. Bins with a mappability score < 0.9 or GC content ≤ 20% or ≥ 80% were removed (5449 bins passing filter). Additionally, cellIDs with low coverage were removed (1268 samples passing filter). Median absolute deviation (MAD) scores were calculated per cell on 500kb bins of cells passing filter as previously described[90]. Briefly, let $Y_{i,j}$ be the raw read count for the i^{th} cellID of the j^{th} bin (from 1.. n bins). Let N_i be a cell-specific scaling factor (total read depth) and B_j be a bin-specific normalization, output as *beta.hat* from the function *normalize_codex2_ns_noK*. Such that MAD scores were then plotted using the ggplot *geom_jitter* and *geom_boxplot* functions.

$$where\ d = \frac{\frac{Y_{i,j}}{N_i B_j} - \frac{Y_{i,j+1}}{N_i B_{j+1}}}{\left(\sum_{j=i}^{n} \frac{Y_{i,j}}{N_i B_j}\right)/n}$$

$$MAD\ score_i = median(|d - median(d)|)$$

Copy Number Calling

SCOPE assumes diploid cells within the sample for normalization steps. To this end we used GM12878 lymphoblastoid cell line as our normal diploid samples and used an *a priori* estimate of 2.6N based on averaged PDCL karyotyping results (Figure 17c). We then used the *SCOPE* function *normalize_scope_foreach* with the following options: K=5, T=1:6 to normalize read distributions per cell. We segmented the genome into breakpoints per chromosome and inferred copy number per breakpoint per cell by *segment_CBScs* allowing for a simple nested structure of copy number changes (*max.ns=1*). To plot inferred copy number per cell, we used the *R* library *ComplexHeatmap* (v2.5.5) by function *Heatmap*[171]. Pairwise distance between cells was generated by Jaccard distance through the R library *philentropy* (v0.4.0)[172] on windows categorized as "neutral" (2N), "amplified" (>2N) or "deleted" (<2N). Cells then underwent hierarchical clustering by the "ward.D2" argument in the function *hclust*. The resultant dendrogram was then cut into both 3 and 6 clades based on the two independent optimal k value searches using the *find_k* function in the R library *dendextend* (v1.14.0) given a range of 2 to 10 and 5 to 10 clusters, respectively (Figure 17i)[173]. Cells with shared clade membership were then combined into "pseudobulk" clades for higher resolution copy number calling. After combining counts data across 50 kbp bins (and filtered as described above), we had 6 clades with 154, 250, 363, 100, 268 and 133 cells, with mean reads per bin of 1207, 2442, 4662, 2071, 2700, and 9416, respectively. These pseudobulk sampled were then normalized as described above with clade 6, containing 83.45% GM12878 cells (111/133 cells) as the normal diploid sample. The genome per sample was then segmented as described above and normalized reads per bin as well as segmentation calls were plotted with *ggplot2 geom_point* and *geom_rect* functions. Select genomic locations[157] of recurrently mutated genes were vizualized and plotted using IGV with 5 bins (250kbp) up and downstream from the transcription start sites. (Appendix Figure 28)[174].

s3-GCC contact profile raw counts were generated for cellIDs passing the read count and SCOPE filters (215 cells) as follows. For initial plotting of single-cell profiles, paired-end read bam files were filtered for an insert length of ≥50kbp via *pysam*[175] and output as upper-triangle triple-sparse format at 1mbp bin sizes. Raw contact matrices were then plotted

with *R* and *ComplexHeatmap* (Figure 17j, left). Merged ensemble plots were generated by summing single-cell contact matrices generated as described above for 500 kbp bins. Following this, we performed dimensionality reduction and clustering analyses using a topic modeling approach. We treated the GCC portion of single-cell sequencing fragments (read pairs separated by a genomic distance higher than 1kb) as traditional distal interactions. We analyzed these cells using our previously established topic model for analysis and characterization of single-cell Hi-C data[176]. In the topic modeling framework, each cell is treated as a mixture of "topics" where each topic corresponds to a set of distal interactions. The model is trained in an unsupervised manner to find the optimum number of topics that best describe the data and associates each distal interaction with a probablistic mixture of topics.

We trained a topic model using the GCC data with the default parameters in Kim et al. However, we altered one parameter, which is the range of distal interactions that are input into the model. Due to high coverage of s3-GCC assays, we opted for distal interactions that are separated by a genomic distance of 20Mb or less, as opposed to original parameter where we used interactions that are separated by distances lower than 10Mb. After training, we found that the number of topics that best describe the data is 15. We visualized cells using UMAP and found that the majority of cells from two lines cluster separately. Overall, these results validate the Hi-C like characteristics of GCC data and further show that we can capture the subtle differences in chromatin organization of the two lines.

Chapter 3: sciDROP Single-cell chromatin assay at one hundred thousand cell output

Authors collaborating in this work and affiliations

Ryan M. Mulqueen[a*] & Hao Zhang[b*], Dmitry Pokholok[c], Frank J. Steemers[c], Andrew C. Adey[a,d,e,f,g] & Darren Cusanovich[h,i]

a. Oregon Health & Science University, Department of Molecular and Medical Genetics, Portland, OR
b. University of Arizona, Graduate Program in Molecular Medicine, Tucson, AZ
c. ScaleBio, CA
d. Oregon Health & Science University, Cancer Early Detection Advanced Research Center, Portland, OR
e. Oregon Health & Science University, Department of Oncological Sciences, Portland, OR
f. Oregon Health & Science University, Brendan Colson Center for Pancreatic Care, Portland, OR
g. Oregon Health & Science University, Knight Cardiovascular Institute, Portland, OR
h. University of Arizona, Cellular and Molecular Medicine, Tucson, AZ
i. University of Arizona, Asthma & Airway Research Center, Tucson, AZ

*Authors contributed equally to this work

Author Contributions

R.M.M., D.P., F.J.S., D.C., and A.C.A. conceived the study. R.M.M. and H.Z. performed all sci-DROP experiments and led all analysis under the supervision of A.C.A. and D.C.; H.Z. performed additional optimization under D.C. R.M.M., D.P., F.J.S., D.C., and A.C.A. contributed to the design of sci-DROP protocol. R.M.M. designed and implemented the analysis under advice from A.C.A. The manuscript was written by R.M.M. and A.C.A.

Appendix Figures and Tables supplied to the committee.

Abstract

Chromatin accessibility has demonstrated a powerful ability to identify cell types and states. Performed at single-cell resolution, this assay has generated catalogs of genome-wide DNA regulatory sites, dynamic chromatin reorganization through development[67], and whole organism cell atlases on model species[80]. Single cells in isolation are ineffectual for explaining complex and/or developing tissue, so instead methods look to generate hundreds to thousands of single cell libraries in parallel at once to provide cell-to-cell context. There are two general methods for achieving single-cell data with hundreds to thousands of cells in a single experiment. First, cells can be isolated into a single reaction vessel, be it tube or nanoliter droplet on a microfluidics platform, as seen in the commercialized products of 10X Genomics[30] and Bio-Rad[28]. Captured cells share the microfluidic droplet with a gel bead synthesized with a set of unique oligonucleotides, used to specifically label the cell. Second, iterative split-pool labelling as is seen in single-cell combinatorial indexing (sci), can identify single cells while never truly dropping down to single cell single reaction conditions[147]. Here we demonstrate a method to increase the throughput of single cell ATAC seq by combining these two approaches; tagmenting nuclei with unique indexes and saturating the nuclei loading within the 10X Chromium platform to maximize the throughput of single cell library generation. We use this strategy to generate up to 100,000 cells per reaction on the 10X Chromium controller (~20X increase in throughput), and describe novel biology at atlas-level cell counts. We demonstrate this method on human cortex and mouse whole brain samples.

Main Text

The most robust approach to identify chromatin patterns at a single-cell scale is through ATAC-seq (assay for transposase accessible chromatin by sequencing), in which a hyperactive transposase enzyme inserts sequencing adapters into sterically open regions of chromatin. The resulting pile-up of genome aligned reads identifies loci that are putatively active in expression or regulation[177]. The efficiency of this process has allowed generation of ATAC-seq libraries from single cell inputs. Various methods of single-cell ATAC-seq generation have been reported,

however there is an upper limit for cell-specific library generation. When cell input exceeds indexing throughput, index collisions occur, leading to multiple cells sharing indexes and conflating analysis[178]. To address this issue we combined a 96-well indexed tagmentation approach with a microfluidic gel bead encapsulation approach to combinatorially introduce indexes at both stages (Fig. 18a). As multiple rounds of fragment capture on the gel beads occur within the droplet, there is a possibility for index-switching[178]. With this in mind, we modified our protocol from a linear amplification to an exponential amplification strategy to mitigate necessary cycles. High cell count strategies can lead to a sequencing burden during early quality control; requiring billions of reads prior to *post hoc* identification of cells through separation in library complexity. To address this, we developed a modified protocol to sequester some cells for quality control prior to additional sequencing effort on the full libraries. After encapsulation, we split the pool of encapsulated cell-beads for 10% and 90% of the volumes.

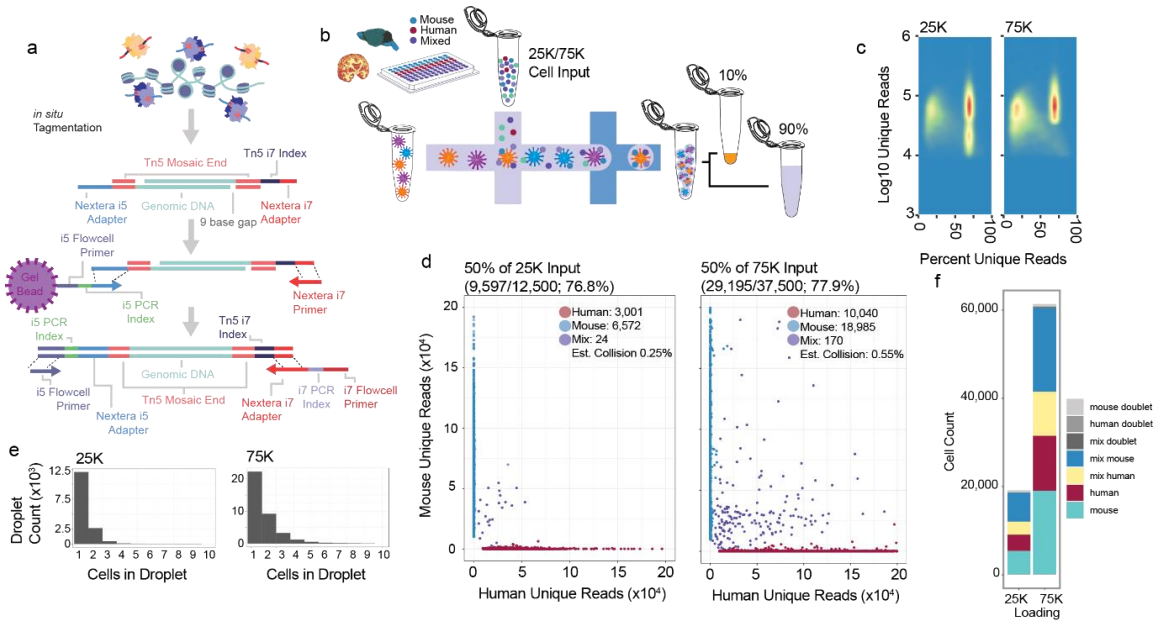


Figure 18. sciDROP generates high quality single-cell ATAC libraries at high throughput. a) Molecular details of sciDROP library generation, schematized. b) Experimental flow through for sciDROP human cortex and mouse full brain library generation. After 96-plex tagmentation, cells are loaded into a 10X Chromium microfluidics device at either 25,000 or 75,000 target cells (25K/75K, respectively). Following cell encapsulation in the formed emulsion, libraries are split for quality control into 10% and 90% pools for quality control. c) Two-dimensional density map of cells passing initial read filters for percent unique reads (library saturation) and unique read counts. d) Mixed-species tagmentation wells were subject to alignment in both human and mouse reference genomes. e) Number of cells per droplet quantified on a histogram, showing a majority of droplets are still contain only a single cell. f) Quantification of cells in 25K and 75K library pools. Conditions include doublets uncovered either through cross-species alignment ("mixed doublet") or through reduced dimension detection strategy (see Methods). Other cells passing these filters are colored by identified species and tagmentation condition.

To test this strategy, we performed a multiplexed tagmentation of human cortex and mouse whole brain samples. A mixed species experiment such as this (Fig. 18b.) allows for an accurate estimation of collision rate since each index is expected to align uniquely to either the human or mouse reference genome. Indexes with cross-alignment indicate collisions and allow us to empirically scale cells loaded during droplet formation. We performed two separate experiments following the same tagmentation scheme and loaded either 25,000 (~5X recommended loading volume) or 75,000 (~15X). Libraries were sequenced to an average depth 44,865 unique reads, with a saturation rate of 69.1% unique reads (Fig 18c). Using a species purity cutoff of 90%, we uncovered an estimated collision rate of 0.25% and 0.55%, respectively; consistent with our estimated average cell loading per droplet of 1.26 and 1.66 (Fig 1e. see Methods). This suggests that even at an exponential increase, over 2.09×10^6 cells can be loaded within a single lane before a 5% collision rate is attained (3.47 nuclei per droplet). Physical constraints such as device clogging or cell suspension density limits are likely to occur first. We captured 19,141 and 61,388 cells, respectively (Fig 18f.). To uncover sample complexity, iterative dimensionality reduction[74] and clustering[169] was performed on human and mouse cells separately from the single 15X loading lane (Fig 18f, Fig 19a, Fig 20a). We ran an analysis for cryptic doublets within species to remove barcode collisions passing our initial species alignment filter (Methods, Appendix Tables 14-15)[179]. We then used previously published single-cell RNA data to predict cell types within our data set, using cis-coaccessible networks (CCANs) to generate gene activity scores for comparison to transcription[34,156,180–182]. Prediction labels were confirmed with canonical marker genes per cell type (Fig 19b, Fig 20b)[181,182]. We performed an *a priori* marker identification analysis across cell subclusters on gene activity, and transcription factor motif accessibility(Fig 19c-d, Fig 20c-d)[20,75,78]. Notably in both our human and mouse samples, we were able to detect and identify many expected cell types with high confidence.

In the human data set, cell types identified through single-cell RNA-seq experiments[34,181] were identified through the use of gene activity scores. Inhibitory neuron subtypes were readily discerned through the bias in use of transcription factor motifs. Notably, in the human sample, *GAD1+* inhibitory neurons (clusters 4 and 5) displayed a clear separation between medial and

medial ganglionic eminence markers (*LHX6*+ and *ADARB2*+, respectively) similar to the work described in Chapter 2. However, with increased cell count, we were able to further separate the *LHX6*+ cluster 4 into *SST*+ and *PVALB*+ cells. The *PVALB* gene does not show any read pile-ups along the gene length, leading *PVALB*+ cells to be inferred by known co-expressed markers such as *CUX2*. Likewise, we can separate *ADARB2*+ inhibitory neurons (cluster 5) into *VIP*+ and *LAMP5*+ cells (Fig 19b)[34]. The separation of these cell types was not solely in gene activity, but we also observed a clear bias in transcription factor motif usage. *LHX6*+ cells in cluster 4, showed a specific enrichment in nuclear receptors with C4 zinc fingers, such as RORB, RORC, and ESRRB[77]. Conversely, *ADARB2*+ inhibitory neurons (cluster 5) showed an enrichment in POU-domain factors. We saw a similar separation between inhibitory neuron gene activity and transcription factor motif usage in the mouse sample (Fig 20).

In summary the work described here displays a combination of two techniques for single-cell ATAC-seq library generation. By combining the reaction isolation of microfluidic emulsion used in the 10X chromium system, and the multi-cell reactions possible with combinatorial indexing, we improve cell count throughput by over an order of magnitude. In two proof-of-principle experiments, we demonstrate the high quality libraries attainable. We then use a single tube reaction to generate two high cell count assessments of human and mouse brain tissue samples (22,565 and 38,606 cells, respectively). To do this we leveraged the inherent multiplexing ability of multiple tagmentation reactions, meaning number of independent samples need not be sacrificed, overcoming a major hurdle of the 10X platform[30]. The read depth and cell count attainable within this single reaction corresponds well with transcriptome profiles from large-scale single-cell RNA atlases[34,181,182] and allows for the assay of genome-wide peaks and transcription factor motif usage. The ease and availability of this method will lead to substantially higher quality data sets and more nuanced chromatin accessibility studies.

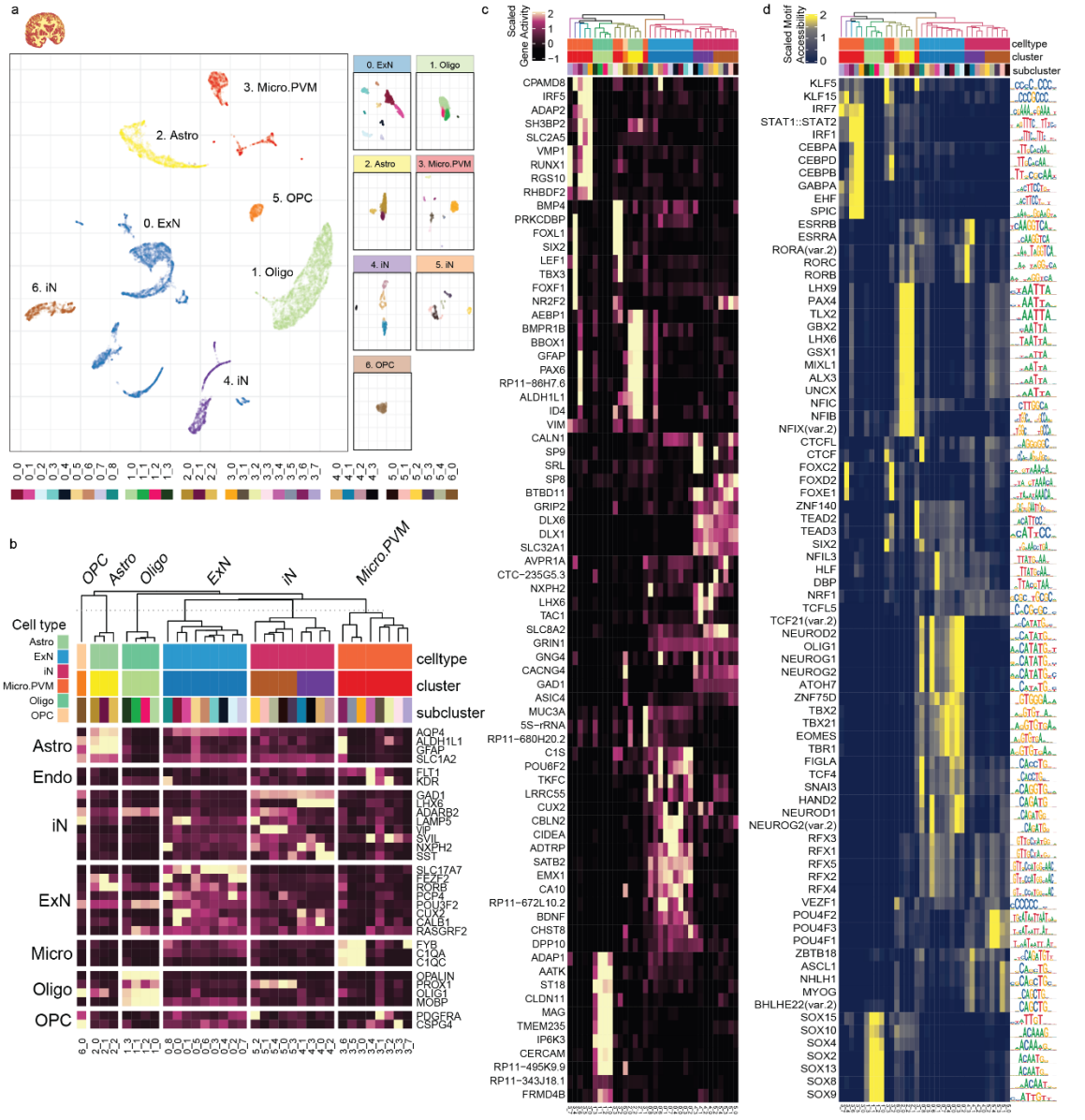


Figure 19. Cell type identification and marker assessment in human cortex sample. a) UMAP projection of 75K loading human cells (n= 22,565). Identified clusters then underwent a second round of reduction and UMAP projection in parallel (right panels). Subcluster coloring is consistent throughout the figure. b) Cell type and subtype identification through canonical marker gene sets. Z-scored average gene activity score per subcluster is plotted as a heatmap. Subclusters are hierarchically clustered and labelled by cell type. Astro: astrocytes; Endo: endothelial cells; iN: inhibitory neurons; ExN: excitatory neurons; Micro: microglia; Oligo: oligodendrocytes; OPC: oligodendrocyte progenitor cells; Micro.PVM: microglia and perivascular macrophages. c) A priori determination of marker genes through chromatin accessibility-derived gene activity. Z-scored average gene activity score per subcluster is plotted, with the top 3 markers per subcluster shown. Subclusters are hierarchically clustered based on all differentially accessible gene activities. d) A priori determination of marker transcription factor motifs through genome-wide transcription factor motif accessibility. Z-scored average motif accessibility per subcluster is plotted, with the top 3 markers per subcluster shown. Subclusters are hierarchically clustered based on all differentially accessible gene activities, consistent with panel c. SeqLogos are plotted alongside heatmap rows.

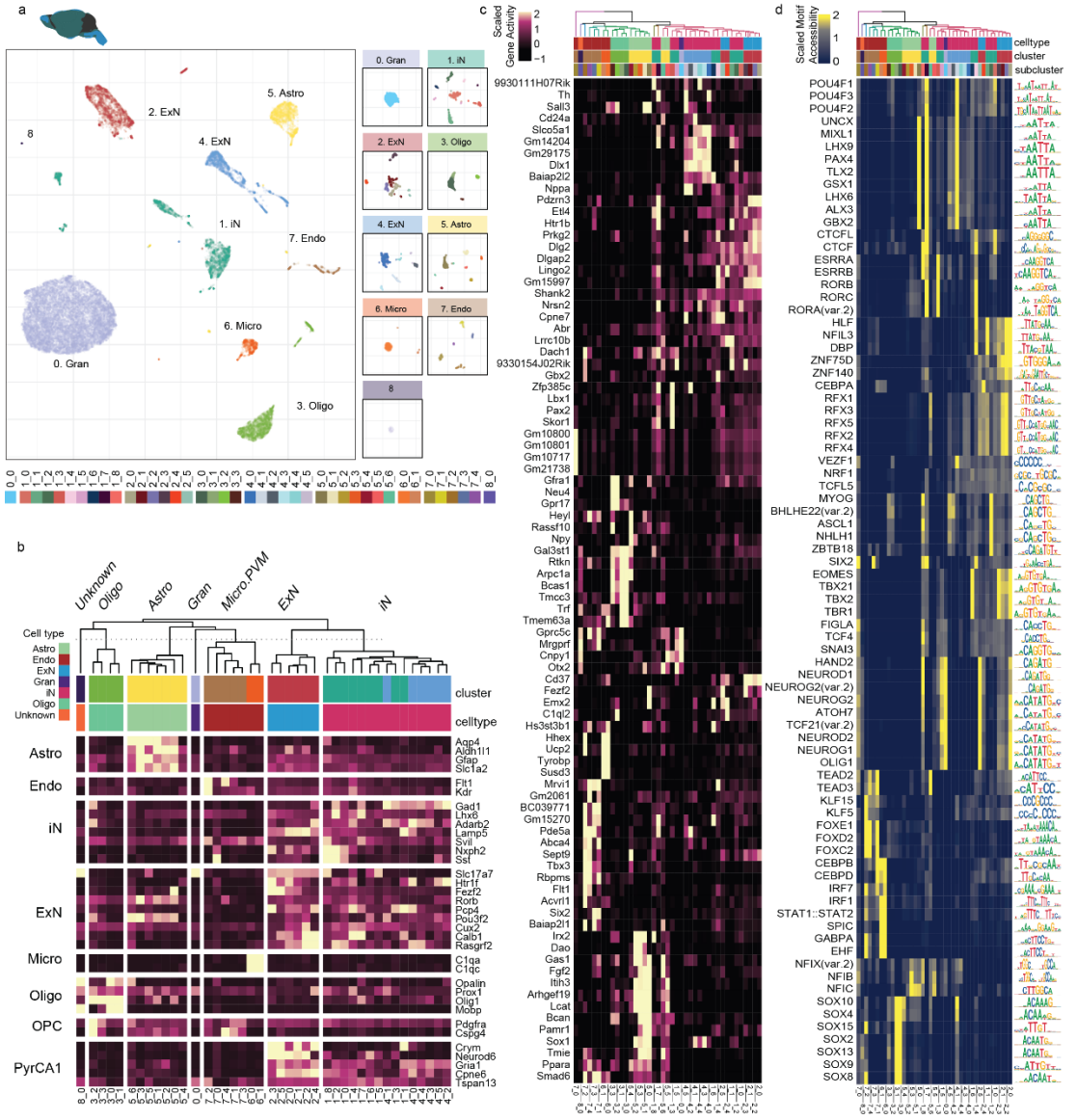


Figure 20. Cell type identification and marker assessment in mouse whole brain sample. a) UMAP projection of 75K loading human cells (n= 38,606). Identified clusters then underwent a second round of reduction and UMAP projection in parallel (right panels). Subcluster coloring is consistent throughout the figure. b) Cell type and subtype identification through canonical marker gene sets. Z-scored average gene activity score per subcluster is plotted as a heatmap. Subclusters are hierarchically clustered and labelled by cell type. Abbreviations are consistent with Figure 2. PyrCA1: pyrimidal CA1 neurons. c) A priori determination of marker genes through chromatin accessibility-derived gene activity. Z-scored average gene activity score per subcluster is plotted, with the top 3 markers per subcluster shown. Subclusters are hierarchically clustered based on all differentially accessible gene activities. d) A priori determination of marker transcription factor motifs through genome-wide transcription factor motif accessibility. Z-scored average motif accessibility per subcluster is plotted, with the top 3 markers per subcluster shown. Subclusters are hierarchically clustered based on all differentially accessible gene activities, consistent with panel c. SeqLogos are plotted alongside heatmap rows.

Methods

Sample preparation

At the time of nuclei dissociation, 50mL of nuclei isolation buffer (NIB-HEPES) was freshly prepared with final concentrations of 10 mM HEPES-KOH (Fisher Scientific, BP310-500 and Sigma Aldrich 1050121000, respectively), pH 7.2, 10 mM NaCl (Fisher Scientific S271-3), 3mM MgCl2 (Fisher Scientific AC223210010), 0.1 % (v/v) IGEPAL CA-630 (Sigma Aldrich I3021), 0.1 % (v/v) Tween-20 (Sigma-Aldrich P-7949) and diluted in PCR-grade Ultrapure distilled water (Thermo Fisher Scientific 10977015). After dilution, two tablets of Pierce™ Protease Inhibitor Mini Tablets, EDTA-free (Thermo Fisher A32955) were dissolved and suspended to prevent protease degradation during nuclei isolation.

Primary samples of C57/B6 mouse whole brain were extracted and flash frozen in a liquid nitrogen bath, before being stored at -80°C. Human cortex samples from the middle frontal gyrus were sourced from the Oregon Brain Bank from a 50-year-old female of normal health status. Tissue was collected at 21 hours post-mortem and then placed in a -80°C freezer for storage. An at-bench dissection stage was set up prior to nuclei extraction. A petri dish was placed over dry ice, with fresh sterile razors pre-chilled by dry-ice embedding. 7mL capacity dounce homogenizers were filled with 2mL of NIB-HEPES buffer and held on wet ice. Dounce homogenizer pestles were held in ice cold 70% (v/v) ethanol (Decon Laboratories Inc 2701) in 15mL tubes on ice to chill. Immediately prior to use, pestles were rinsed with chilled distilled water. For tissue dissociation, mouse and human brain samples were treated similarly. The still frozen block of tissue was placed on the clean pre-chilled petri dish and roughly minced with the razors. Razors were then used to transport roughly 1 mg the minced tissue into the chilled NIB-HEPES buffer within a dounce homogenizer. Suspended samples were given 5 minutes to equilibrate to the change in salt concentration prior to douncing. Tissues were then homogenized with 5 strokes of a loose (A) pestle, another 5 minute incubation, and 5-10 strokes of a tight (B) pestle. Nuclei were transferred to a 15mL conical tube and pelleted with a 400 rcf centrifugation at 4°C in a centrifuge for 10 minutes. Supernatant was removed and pellets were resuspended in 5mL of ATAC-PBS buffer

(APB) consisting of 1X PBS (Thermo Fisher 10010) and 0.04mg/mL (f.c.) of bovine serum albumin (BSA, Sigma Aldric A2058). Samples were then filtered through a 35 µm cell strainer (Corning 352235). A 10uL aliquot of suspended nuclei was diluted in 90uL APB (1:10 dilution) and manually counted on a hemocytometer with Trypan Blue staining (Thermo Scientific T8154). The stock nuclei suspension was then diluted to a concentration of 2,857 nuclei/uL in APB. Dependent on experimental schema pools of tagmented nuclei were combined to allow for the assessment of pure samples and to test index collision rates (Appendix Tables 14-15).

Tagmentation buffer solution (TB1) was prepared with a final concentration of 1.92X concentration TD buffer (Nextera XT Kit, Illumina Inc. FC-131-1024), 0.0192% (f.c) Digitonin (Bivision 2082-1), 0.192% Tween-20 and diluted in PCR-Grade Ultrapure distilled water. Tagmentation plates were prepared by the combination of 1430 uL of TB1 with 770 uL nuclei solution. This mixture was mixed briefly on ice. 20uL of the mixture was placed into ready-to-use 96-well iTSM plate containing 5uL of 100nM pre-indexed transposase (ScaleBio). Tagmentation was performed at 37°C for 60 minutes on a 300 rcf Eppendorf ThermoMixer with a lid heated to 65°C. Following this incubation, plate temperature was brought down with a 5 minute incubation on ice to stop the reaction. Tagmented nuclei were then pooled into a single 5mL conical tube. 5mL of tagmentation wash buffer (TMG) was prepared consisting of a final concentration of 10mM Tris Acetate pH 7.5 (Sigma 93352 and Sigma A6283, respectively), 5mM MgAcetate (Sigma M5661) and 10% (v/v) glycerol (Sigma G5516), diluted in PCR grade water. 1mL of TMG was added on top of the chilled tagmented nuclei. Nuclei were pelleted at 500 x g for 10 minutes. Most of the supernatant was removed with care not to disturb the pellet. Then 500uL of TMG was added to the pellet and the tube was once again spun at 500 x g for 5 minutes. 490 uL was removed leading a low volume of concentrated nuclei. Loading buffer was prepared consisting of 10% (v/v) glycerol, 20 mM NaCl, 10 mM Tris-Cl pH 7.5 (Life technologies AM9855), 0.02 mM EDTA (Fisher Scientific AM9260G), 0.2 mM DTT (VWR 97061-340), and 0.2X TB1 (v/v). The nuclear pellet was resuspended with an additional 30uL of loading buffer. An aliquot of 2uL of sample was dilute 20-50X and quantified with Trypan Blue on a hemocytometer. Depending on experiment, a 14uL nuclei

solution containing the desired amount of nuclei in loading buffer was then combined with 1uL of 75 uM oligo SBS12 (5'CGTGTGCTCTTCCGATCT in TE buffer).

The 10X Chromium was then run with the custom nuclei solution as per manufacturer's instructions (10x Document CG000209 Rev D) with the following adaptations. At step 2.4e during GEM aspiration and transfer, 100uL GEM volume was split into two tubes, with one receiving 10uL and the other 90uL (henceforth referred to as 10% and 90% samples). At step 2.5.a, GEM incubation cycles were limited to 6. For Pre-PCR wash elution (Step 3.2.j) the 10% sample was eluted in 8.5uL whereas the 90% sample was eluted in 32.5uL. For step 3.2.n, the 10% sample had 8uL transferred to a new strip, while the 90% sample had 32uL transferred to a new strip. At step 4.1.b, the sample Index PCR mix was split with 11.5uL and 46uL being combined with the 10% and 90% samples, respectively. For step 4.1.c, 1uL and 2uL of a 10uM i7 TruSeq primer was used, respectively. For step 4.1.d, 8 and 7 PCR cycles were used, respectively. Libraries were then checked for quality and quantified by Qubit DNA HS assay (Agilent Q32851) and Tapestation D5000 (Agilent 5067-5589) following manufacturer's instructions. Libraries were then diluted and sequenced on a NextSeq 500 mid-capacity or NovaSeq 6000 S4 flow cells (Illumina Inc.).

Computational Analysis

Raw code is available at https://mulqueenr.github.io/scidrop/

Preprocessing

After sequencing, data was converted from bcl format to FastQ format using bcl2fastq (v 2.19.0, Illumina Inc.) with the following options with-failed-reads, no-lane-splitting, fastq-compression-level=9, create-fastq-for-index-reads. Data were then demultiplexed, aligned, de-duplicated using the in-house scitools pipeline[31]. Briefly, FastQ reads were assigned to their expected primer index sequence allowing for sequencing error (Hamming distance ≤2) and indexes were concatenated to form a "cellID". Reads that could be assigned unambiguously to a cellID were then aligned to reference genomes. Paired reads were aligned with bwa mem (v0.7.15-r1140)[183] first aligned to a concatenated hybrid genome of hg38 and GRCm38 ("mm10", Genome Reference Consortium Mouse Build 38 (GCA_000001635.2)). Reads were then de-duplicated to remove PCR and

optical duplicates by a perl (v5.16.3) script aware of cellID, chromosome and read start, read end and strand. From there putative single-cells were distinguished from debris and error-generated cellIDs by both unique reads and percentage of unique reads.

Barnyard Analysis

With single-cell libraries distinguished, we next quantified contamination between nuclei during library generation. We calculated the read count of unique reads per cellID aligning to either human reference or mouse reference chromosomes (Figure 18). CellIDs with ≥90% of reads aligning to a single reference genome were considered bona fide single-cells. Those not passing this filter were considered collisions. Collision rate was estimated to account for cryptic collisions (mouse cell-mouse cell or human cell-human-cell) by multiplying by two. Bona fide single-cell cellIDs were then split from the original FastQ files to be aligned to the proper hg38 or mm10 genomes with bwa mem as described above. Human and mouse assigned cellIDs were then processed in parallel for the rest of the analysis. After alignment, reads were again de-duplicated to obtain proper estimates of library complexity (Appendix Tables 14-15).

Tagmentation Insert Quantification

We generated tagmentation site density plots centered around transcription start sites (TSSs). We used the alignment position (chromosome and start site) for each read to generate a bed file that was then piped into the BEDOPS closest-feature command mapped the distance between all read start sites and transcription start sites (v 2.4.36)[168]. From this, we collapsed binned distances (100bp increments) into a counts table and generated percentage of read start site distances within each counts table. We plotted these data using R and ggplot2 geom_density function (default parameters) subset to 2000 base pairs around the start site to visualize enrichment. TSS enrichment values were calculated for each experimental condition using the method established by the ENCODE project (https://www.encodeproject.org/data-standards/terms/enrichment)[184], whereby the aggregate distribution of reads ±1,000 bp centered on the set of TSSs is then used to generate 100 bp windows at the flanks of the distribution as the background and then through the distribution, where the maximum window centered on the TSS is used to calculate the fold enrichment over the outer flanking windows.

Dimensionality Reduction

Pseudo-bulked data (agnostic of cellID) was then used to call read pile-ups or "peaks" via macs2 (v.2.2.7.1)[72] with option –keep-dup all. Narrowpeak bed files were then merged by overlap and extended to a minimum of 500bp for a total of ### peaks for human and ### peaks for mouse. A scitools perl script was then used to generate a sparse matrix of peaks × cellID to count the occurrence of reads within peak regions per cell. Fraction of reads in peaks (FRiP) was calculated as the number of unique, usable reads per cell that are present within the peaks out of the total number of unique, usable reads for that cell for each peak bed file. Tabix formatted files were generated using samtools and tabix (v1.7). The counts matrix and tabix files were then input into a SeuratObject for Signac (v1.0.0) processing[75,169]. We performed LDA-based dimensionality reduction via cisTopic (v0.3.0)[74] with 28 and 30 topics for human and mouse cells, respectively. The number of topics were selected after generating 25 separate models per species with topic counts of 5,10,20-30,40,50,55,60-70 and selecting the topic count using selectModel based on the second derivative of model perplexity. Cell clustering was performed with Signac FindNeighbors and FindClusters functions on the topic weight × cellID data frame. For FindClusters function call, resolution was set to 0.01 and 0.02 for human and mouse samples, respectively. The respective topic weight × cellID was then projected into two dimensional space via a uniform manifold approximation and projection ("UMAP") by the function umap in the uwot package (v0.1.8)[185]. To check for putative doublets within-species, we then ran scrublet analysis and removed scrublet-identified doubles from further analysis[179]. A second iteration of subclustering was performed on each cluster to better ascertain cell type diversity. This was done as described above with the data subset to just the cells within the respective cluster for both cisTopic model building and UMAP projection. Resolution per subcluster was set *post hoc* based on cell separation in UMAP projection. Cis-coaccessibility networks (CCANs) and the resulting gene activities were generated through the Signac wrapper of cicero (v1.3.4.10)[20]. Genome track plots with CCAN linkages were generated through Signac function CoveragePlot for marker genes previously described. Genome-wide accessibility of known transcription factor motifs was calculated per cell using the JASPAR database (release 8)[77] via chromVAR[78]. Differential

accessibility between subclusters in one by all other comparisons were generated using Signac function FindMarkers using options: test.use = 'LR', and only.pos=T, with latent.vars = 'nCount_peaks', to account for read depth.

Cell Type Identification

For cell type identification we used previously existing single-cell RNA data sets of the human M1 cortex, and mouse whole cortex and hippocampus. We applied the Signac label transfer strategy between the annotated single-cell RNA with our gene activity scores at the level of our subclustered cell groups. For cell type refinement, we plotted the average gene activity score per subcluster for a set of RNA-defined marker genes, as well as markers defined within our data sets on the gene activity scores using the Signac FindMarkers function as described above. Subcluster dendrograms were generated by using base R functions *dist* and *hclust* through running Z-scored average gene activity on internally-defined markers and based on “ward.D2” clustering of Euclidean distance. The resultant dendrogram was used for both pre-defined and internally defined marker sets. Results were plotted via ComplexHeatmap (v2.5.5).

Chapter 4: Single-cell ATAC-seq reveals chromatin dynamics of *in vitro* corticogenesis

Authors collaborating in this work and affiliations

Ryan M. Mulqueen[a*], Brooke A. DeRosa[a*], Casey A. Thornton[a], Zeynep Sayar[b], Kristof A. Torkenczy[a], Andrew J. Fields[a], Kevin M. Wright[c], Xiaolin Nan[b,d,e], Ramesh Ramji[f], Frank J. Steemers[g], Brian J. O'Roak[a], Andrew C. Adey[a,b,d,h,l,j]

a. Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, Oregon, USA
b. Knight Cardiovascular Institute, Portland, Oregon, USA
c. Vollum Institute, Oregon Health & Science University, Portland, Oregon, USA
d. Knight Cancer Institute's Cancer Early Detection Advanced Research Center (CEDAR), Oregon Health & Science University, Portland, OR, USA.
e. Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA
f. Illumina, Inc., San Diego, California, USA
g. ScaleBio, CA
h. Oregon Health & Science University, Department of Oncological Sciences, Portland, OR
i. Oregon Health & Science University, Brendan Colson Center for Pancreatic Care, Portland, OR
j. Oregon Health & Science University, Knight Cardiovascular Institute, Portland, OR

*These authors contributed equally: Ryan M. Mulqueen, Brooke A. DeRosa.

Author Contributions

A.C.A. and F.J.S supervised all aspects of Pitstop 2 assessment on in situ tagmentation. F.J.S. and R.R. initiated Pitstop 2 experiments. R.M.M. led Pitstop 2 assessment with C.A.T. ATAC-see experiments were performed by Z.S. and X.N. A.J.F. aided in scip-ATAC-seq protocol development. B.A.D. performed all organoid differentiation, culture maintenance, and immunostaining. Organoid single-cell ATAC-seq experiments were performed by R.M.M., who led

the analysis with input from K.A.T and C.A.T. Biological interpretation was performed by R.M.M., B.A.D, K.M.W., B.J.O., and A.C.A. The manuscript was written by R.M.M., B.A.D., B.J.O., and A.C.A. with input from all authors.

Appendix Figures and Tables supplied to the committee.

Abstract

Development is a complex process that requires the precise modulation of regulatory gene networks controlled through dynamic changes in the epigenome. Single-cell -omic technologies provide an avenue for understanding the mechanisms of these processes by capturing the progression of epigenetic cell states during the course of cellular differentiation using *in vitro* or *in vivo* models[186]. Single-cell combinatorial indexing (sci-) has been applied as a strategy for identifying single-cell -omic originating libraries and removes the necessity of single-cell, single-compartment chemistry[80]. Here, we apply a sci- assay for transposase accessible chromatin by sequencing (ATAC-seq; sci-ATAC) to characterize the chromatin dynamics of developing forebrain-like organoids, an *in vitro* model of human corticogenesis[55]. Using these data, we characterized novel putative regulatory elements, compared the epigenome of the organoid model to human cortex data, generated a high-resolution pseudotemporal map of chromatin accessibility through differentiation, and measured epigenomic changes coinciding with a neurogenic fate decision points. Finally, we combined transcription factor motif accessibility with gene activity (GA) scores to directly observe the dynamics of complex regulatory programs that regulate neurogenesis through developmental pseudotime.

Main Text

Recent methodical advances have enabled the preparation of thousands of single-cell -omics libraries simultaneously. Using the general sci- framework, single-cell library generating methods have been developed to measure accessible chromatin[73], genomic sequence variation[187], transcription[137], chromatin folding[159] and DNA methylation[188]. Specifically, sci-ATAC enables the interrogation of open chromatin regions, which are predominantly active promoters and enhancers, and make up between 1-4% of the genome[189]. In sci-ATAC, generation of

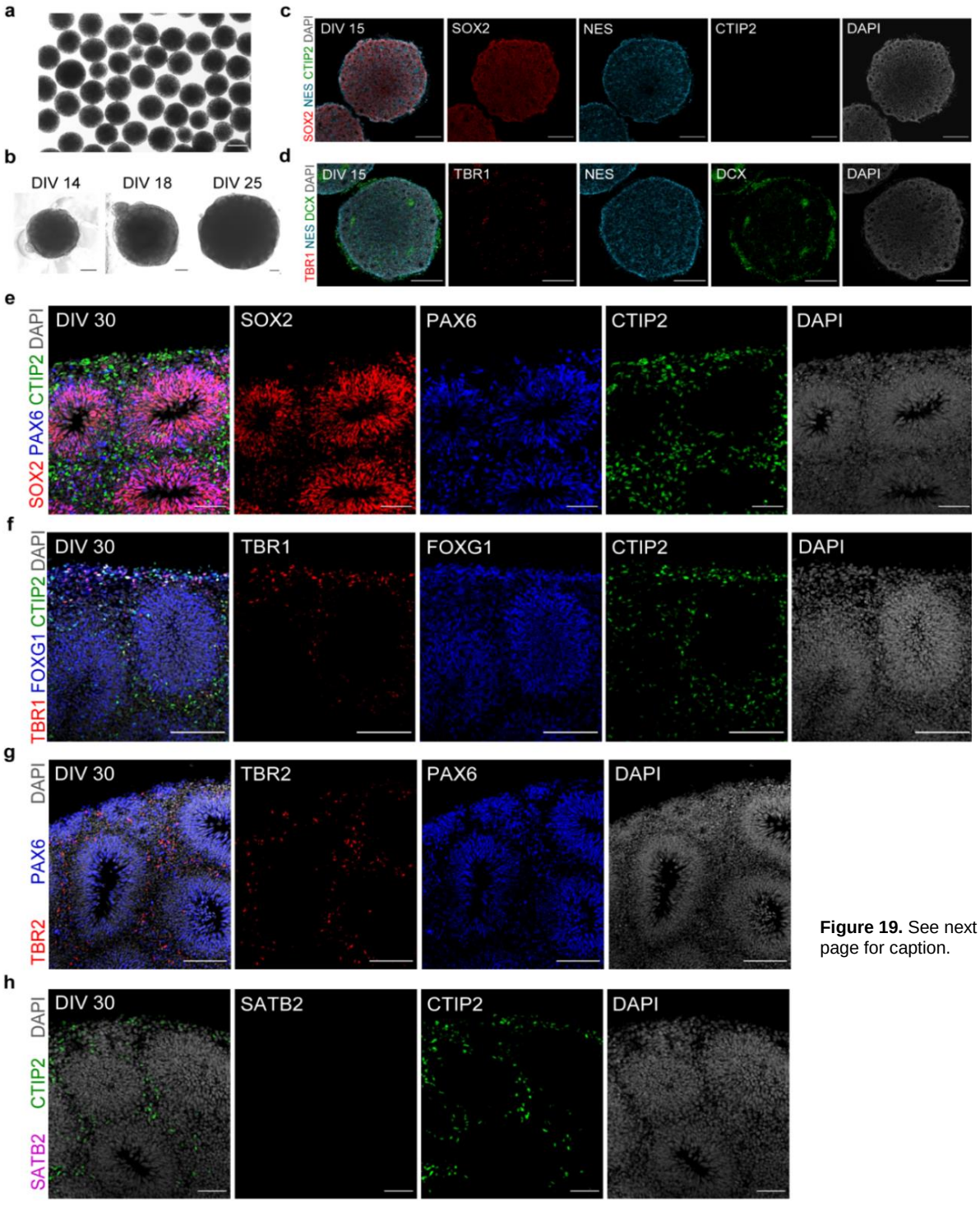


Figure 19. See next page for caption.

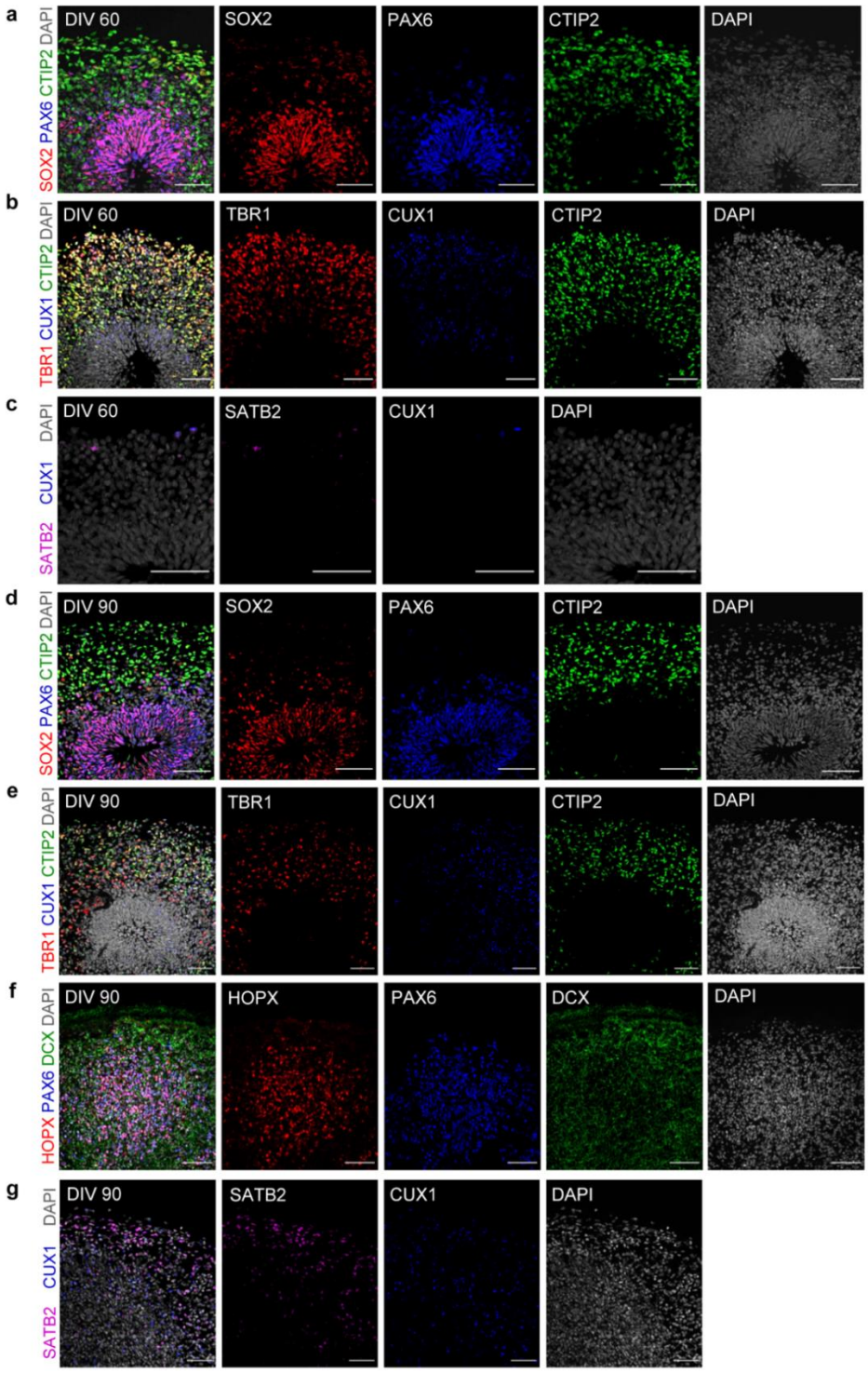
sequencing library molecules is selective towards regions of open chromatin due to the steric hindrance caused by DNA-bound proteins such as histones, on the hyperactive derivative of the

Figure 19. Characterization of earlier stage forebrain-like organoids. a, Brightfield image of days in vitro (DIV) 7 organoids showing uniformity in size and shape. Scale bar: 200 µm. b, Brightfield images of organoids at DIV 14, 18, and 25 showing growth over time. Note the increased number of neuroepithelial buds around the perimeter of DIV 25 organoids. Scale bars: 200 µm. c-d, Immunohistochemical characterization of organoids at DIV 15. Scale bars: 200 µm. c, At DIV 15, the majority of cells stain positive for the progenitor markers SOX2 and Nestin (NES) and do not yet express the layer 5 marker CTIP2. Scale bars: 200 µm. d, DIV 15 organoids stained for NES, DCX, and TBR1. Scale bars: 200 µm. At DIV 15, organoids are mainly comprised of SOX2+/NES+ progenitors (c) with patches of newly born DCX+/TBR1+ layer 6 neurons (d), which are the first to be born during cortical neurogenesis. e-h, Expanded immunohistochemical characterization of organoids at DIV 30. e, DIV 30 organoid with image panels show individual staining of SOX2, PAX6, and CTIP2. Scale bars: 50 µm. As seen in Fig 2a. f, DIV 30 organoid immunostained for the deep layer neuron markers TBR1 and CTIP2, in addition to FOXG1, a general marker of forebrain development. The subsequent expression of CTIP2 (layer 5) after TBR1 (layer 6) mimics the stepwise order of deep layer neurogenesis in vivo. Scale bars: 100 µm. g, DIV 30 organoid immunostained for EOMES and PAX6, markers of progenitors in the subventricular zone and ventricular zone, respectfully. Scale bars: 100 µm. h, DIV 30 organoid immunostained for SATB2 and CTIP2. Note that at DIV 30, organoids do not yet express SATB2, a marker of upper layer neurons. Scale bars: 50 µm.

cut-and-paste Tn5 transposase[65]. The sci-ATAC platform was recently utilized to produce whole-organism maps[67,80], demonstrating the throughput and power of the technique.

We sought to characterize a complex sample with actively forming cell types through differentiation. Brain organoids are a powerful model system to study human neurodevelopment *in vitro*[55]. Data from bulk and single-cell RNA-seq, H3K4me3 ChIP-seq (chromatin-immunoprecipitation and sequencing), and bulk DNA methylation analysis, demonstrate that these models are strongly correlated with similar data from primary human fetal brain samples ranging from the early to mid-gestational period (post-conception weeks 9-24)[48,52,55,120,190]. Specifically, forebrain-like organoids derived from induced pluripotent stem cells (iPSCs) mimic the early stages of human corticogenesis and lamination, wherein proliferating radial glia cells in the ventricular zone generate a pool of progenitors[45,50,55]. From these radial glia cells, tightly regulated transcription factors drive either continued proliferation, or neurogenesis, when the radial glia or its intermediate progenitor differentiates into neurons[191]. However, our understanding of the temporal dynamics of non-coding regions and regulatory sites within this critical timeframe is lacking. For these reasons, we chose a forebrain-like organoid model system for leveraging our sci-ATAC method on studying epigenomic dynamics[6755].

We differentiated forebrain-like organoids from human iPSCs for up to 90 days *in vitro* (DIV) using a previously described miniature bioreactor protocol with modifications to increase organoid uniformity (Methods, Fig. 19 and 20, Appendix Tables 16-18)[55]. Subsets of organoids were collected from two separate differentiation experiments and characterized by their expression of cortical markers at multiple time points. Similar to previous results[55], these



forebrain-like organoids mimicked the *in vivo* developmental processes by developing multi-layered structures resembling the ventricular zone comprised of SOX2+/PAX6+ progenitors, subventricular zone comprised of EOMES+ (aka TBR2) intermediate progenitors and cortical plate, where we observe almost exclusive expression of layer-specific neuronal markers such as

Figure 20 Characterization of later stage forebrain-like organoids. a-c, Expanded immunohistochemical characterization of organoids at days in vitro (DIV) 60 a, DIV 60 organoid immunostained for SOX2, PAX6, and CTIP2. b, Expanded immunohistochemical characterization of a DIV 60 organoid as seen in Fig. 2a. Image panels show individual staining of TBR1, CTIP2, and CUX1. c, DIV 60 organoid immunostained for CUX1 and SATB2. d-g, Expanded immunohistochemical characterization of organoids at DIV 90. d, DIV 90 organoid immunostained for SOX2, PAX6, and CTIP2. e, DIV 90 organoid immunostained for TBR1, CUX1, and CTIP2. f, DIV 90 organoid immunostained for PAX6, DCX, and HOPX, a marker outer radial glia. g, DIV 90 organoid as shown in Fig. 2a. Image panels show individual staining of SATB2 and CUX. Scale bars: 50 µm.

TBR1, BCL11B (aka CTIP2), and SATB2 (Fig 21a)[55]. Additionally, formation of post-mitotic neurons followed the expected step-wise temporal order of layer-specific neurogenesis, indicated by generation of TBR1+ layer 6 neurons prior to layer 5 CTIP2+ neurons, followed by SATB2+ and CUX1+ upper layer neurons (Fig 20).

We performed sci-ATAC on two separate differentiation experiments derived from the same IPSCs of a neurodevelopmentally normal individual. From this we generated 35,590 quality control (QC)-passing single-cell ATAC profiles from four DIV 30 organoids, four DIV 60 organoid, and eight DIV 90 organoids (Fig. 21a,d). We tested methods of dissociation and nuclear isolation to determine possible increases to transposase activity during *in situ* tagmentation. Though we found increased nuclear occupancy of Tn5 with addition of a nuclear pore complex inhibitor, Pitstop 2, these data failed to reliably replicate in later experiments, suggesting confounding sample variables we were unable to ascertain (Appendix Note 2). We addressed this orthogonally by later improving sci-ATAC with a novel adapter switching strategy (described in Chapter 2).

We used the full set of peaks and performed dimensionality reduction through the use of cisTopic[74], a machine learning approach which defined an optimal 28 “topics” based on shared peak accessibility, producing a matrix of cells by topic weights. This matrix was then used to identify eight clusters of cells based on similarity[169]. We next projected the clusters into 2-dimensional space for visualization using uniform manifold approximation positioning[76] (umap). We saw a strong bias of clustering and projection from DIV, suggesting we were capturing changes to the epigenome through organoid differentiation. Conversely, we did not see strong bias in clustering due to differentiation, transcription start site enrichment, or read counts in similarly aged organoids (Fig 21b). We noted clusters were also unbiased to Pitstop 2 treatment did not interfere with DIV 90 organoid clustering position (Appendix Note 2). To look at cluster proportion per organoid, we measured the proportion of cluster-membership per organoid. We observed some clear patterns of differentiation, namely an increase in cluster 2 and 5 proportion

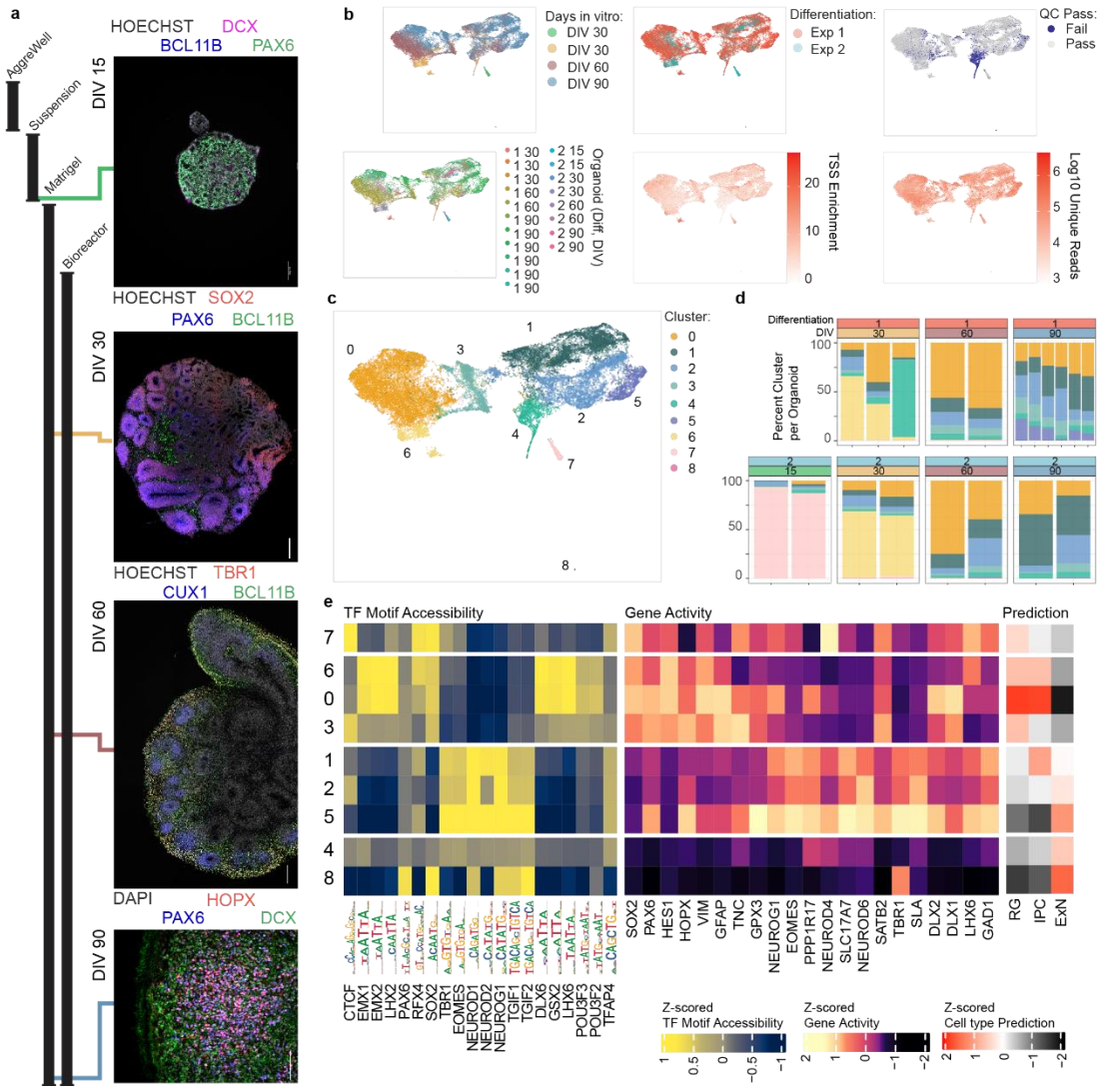


Figure 21. a) Representative images of organoid immunohistochemical staining. b) UMAP projection of sci-ATAC profiles into two-dimensional space, cells are colored by days *in vitro* (DIV), differentiation experiment, quality control passing filter, organoid sampled, transcription start site (TSS enrichment), and Log10 unique reads per cell. c) Cells were clustered on cisTopic reduced dimension matrix and colored by cluster over the same UMAP projection as panel b. d) Stacked bar plots showing the percentage of cells in each cluster per organoid. Bar plots are split by differentiation experiment (top ribbon) and DIV (bottom ribbon). Colors coincide with clusters in panel c. e) (left) Z-scored heatmap of transcription factor (TF) motif accessibility from previously defined markers for primary sample radial glia (RG), intermediate progenitor cell (IPC) and excitatory neurons (ExN). JASPAR reported TF motifs are shown per column. (middle) Gene activity scores calculated for RNA-seq defined marker genes in primary cortical samples. (right) Label-transfer prediction values acquired through cannonical correlation analysis (CCA, Methods) on single-cell RNA-seq data on primary samples. All rows are in the same order throughout panel e.

in organoids from DIV 60 to DIV 90, and a sharp decrease in cluster 6 membership after DIV 30 (Fig 21d). This suggests an ability to capture active chromatin differentiation dynamics of cells through organoid differentiation. We next sought to characterize the cell type and differentiation state of the eight clusters by several epigenetic marks. We leveraged bulk epigenomic data of putative enhancer regions from purified primary human cortical cell types[24] and a single-cell RNA

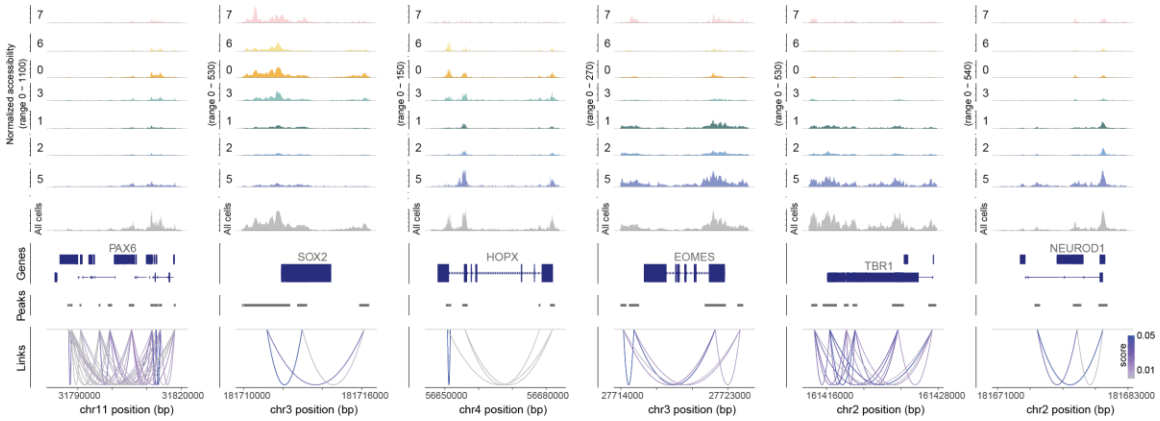


Figure 22. Coverage plots showing normalized read pile-ups per cluster. Select canonical cortex development markers are shown. Gene track and cis-coaccessible networks (CCANs) used for the generation of gene activity scores are shown below.

data set[52]. To infer transcription factor usage in our ATAC profiles, we calculated the genome-wide transcription factor motif accessibility of all validated transcription factors in the publically available JASPAR data base[77,78]. To infer transcription rate, we also generated gene activity scores, through looking at local cis-coaccessible sites anchored at open promoters (Fig 22)[20]. We found that our data followed the expected patterns of transcription factor motif usage and putative marker gene transcription, allowing us to grossly order cell type progression by use of the marker genes, transcription factors, and DIV. To confirm our assessment, we performed a label transfer via co-assay integration on the single-cell RNA data set (Fig 21e)[52]. This demonstrated concordance of our ordering, showing a cluster progression from RG to IPC to excitatory neurons (ExN, Fig 24a). We observed additional statistically significant (q value < 0.05) enrichment of previously reported marker genes and transcription factor activities[191]. For example, cluster 6 and 0 showed enrichment in motifs for NKX6-2, a proliferating radial-glia associated transcription factor, Clusters 1, 2, and 5 showed motifs associated with the NEUROD and T-box family factors, such as NEUROD2, EOMES (aka TBR2) and TBR1, suggesting they were populated by post-mitotic neurons (Fig. 21e)[9,49,55]. To bolster our analysis against read drop-out, we also looked at enrichment across transcription factor modules (TF modules), collections of gene networks known to be co-expressed with transcription factors across cortex development[11]. We generated module scores per cell for 782 modules with 285 ± 91 genes (mean ± standard deviation, Fig 23). This analysis further supports our delineation of radial glia to excitatory neurons, with later

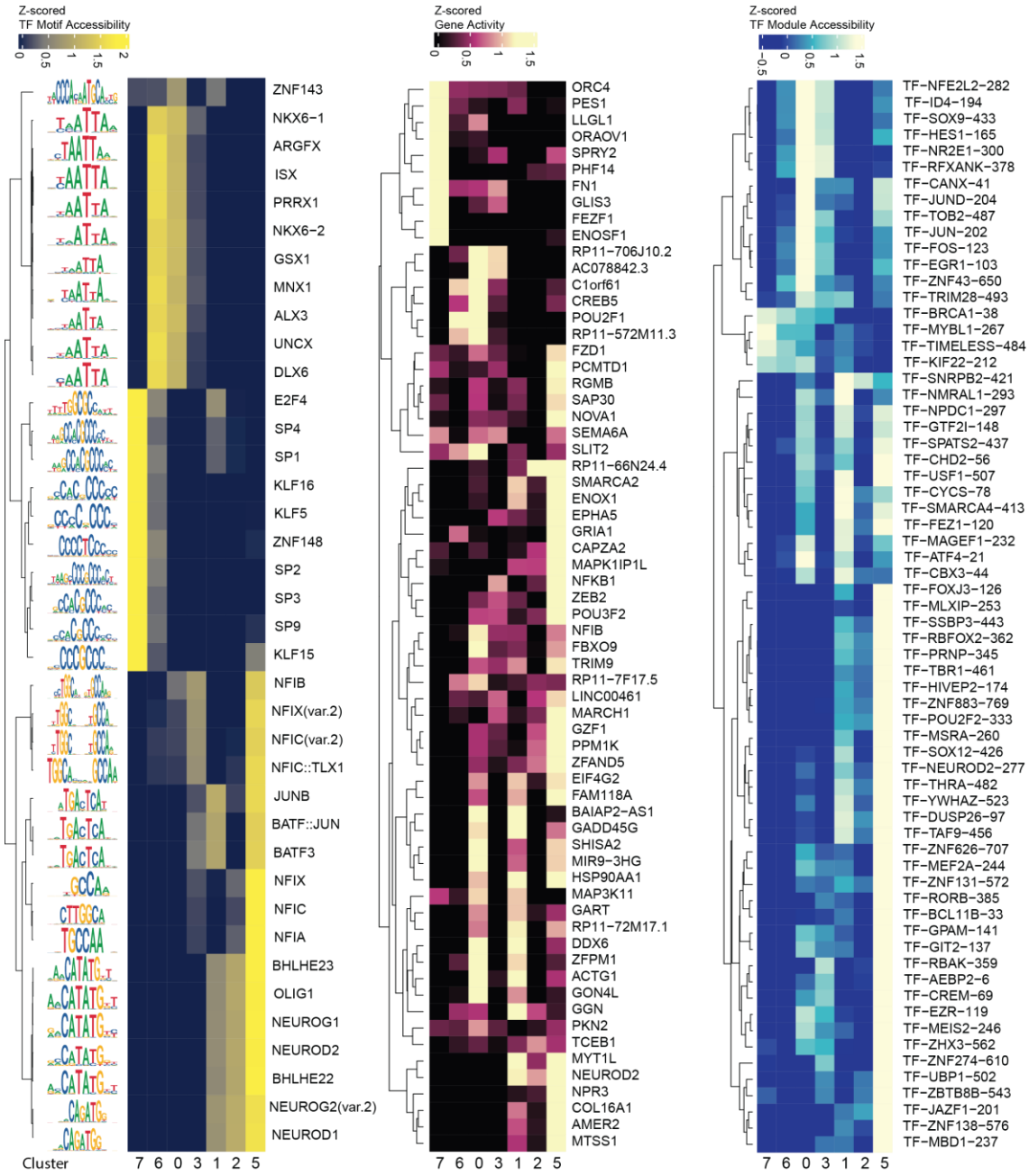


Figure 23 (Left) Differential TF motif accessibility across clusters calculated through logistic regression. Heatmap shows non-redudant top 5 TF motifs based on q value. Values are then Z-scored and plotted, JASPAR–reported motifs are shown to the left. (Middle) Top 5 per cluster non-redudant gene activity score significant differences. (Right) TF module signficant differences per cluster.

clusters showing enrichment in TF modules defined by factors such as NEUROD2, BCL11B, and RORB, whereas radial-glia like clusters show enrichment in HES1, SOX9 and NFE2L2[11,50]. Notably, many of the gene activity scores uncovered as statistically different between clusters are

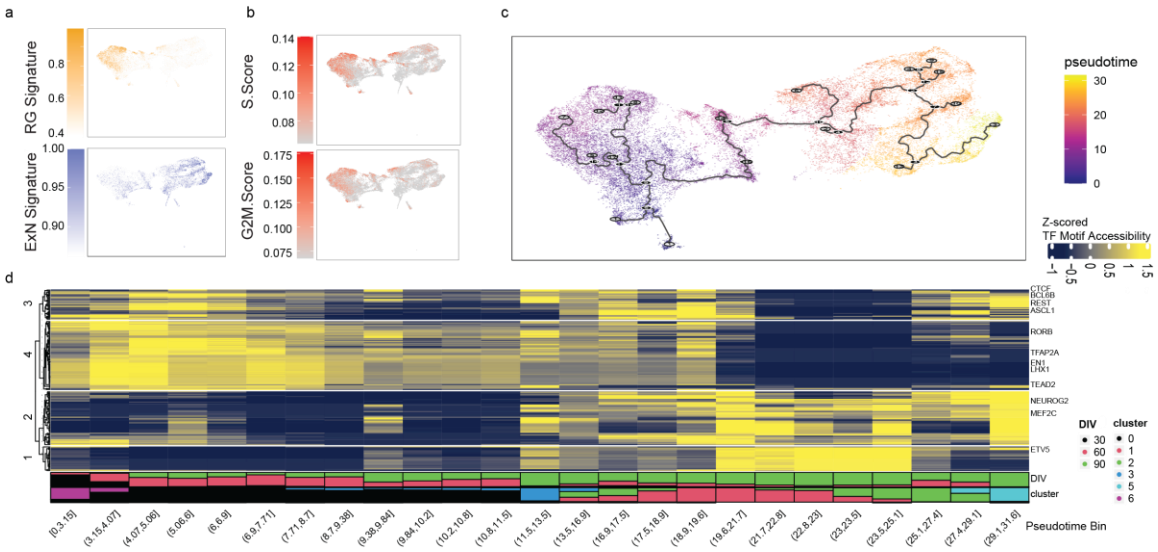


Figure 24 a) Projected values of radial glial (RG) and excitatory neuron (ExN) signatures over UMAP projection. Values are same as Figure 21e, and demonstrate a gradient of values to justify a pseudotime trajectory. b) Cell cycle scores calculated from primary RG cell cycle markers, suggesting the root of the ExN clusters through the still cycling intermediate progenitor cells. c) Trajectory (line) and assigned pseudotime value (color gradient) across organoid cells. d) Transcription factor motif accessibility changes through pseudotime from binned values. (Below) stacked bar plots showing amount of DIV and cluster assignment of cells per bin.

not commonly used as marker genes in organoids, suggesting a either a difference in organoid protocol, or a difference in the regulation of transcriptomic and epigenomic modalities[48,50,52,67].

We further examined the accessibility levels at proximal elements for a set of genes with a known window of activity in corticogenesis based on the assumption that *cis*-acting elements should have increased accessibility when a gene is active[31,192]. We examined accessibility levels at proximal elements for a set of genes with a known window of activity in corticogenesis. Clusters 6, 0, and 3 showed a higher density of proximal reads for neural progenitor genes such as SOX2 or PAX6[9,49,59]. Cluster 5 showed increased read density around deep layer neuronal cortex markers such as TBR1 and NEUROD2[9,49,55], indicating that these marks are activating as cells are exiting the basal-progenitor like state, with an overlap of the two marker sets during the transition (Fig. 22). These assessments further confirmed our progressive cluster assignments through corticogenesis. It is notable that due to similar or shared DNA binding motifs of many classes of transcription factors, the combination of promoter region accessibility and genome-wide transcription factor motif presentation jointly inform interpretation.

Reasoning that all cells within the organoid samples were sourced from a shared stem-like origin and that clustering recapitulates the stereotyped patterns of corticogenesis, we sought to capture epigenome dynamics through generating a pseudotemporal ordering of cells – a trajectory of epigenomic changes during differentiation[26]. We observed a gradient of signal for S-phase and G2/M-phase defined through single-cell RNA analysis of radial glia previously (Fig 24b)[11]. We constructed a trajectory across all cells as done previously for primary cortical samples[11,67]. We rooted the trajectory within cluster 6, populated almost solely by DIV 30 organoid-sourced cells most similar to the primodial neuroepithelia. We then used both transcription factor motif accessibility and TF modules to measure changes in epigenomic regulation through differentiation. We found that multiple waves of transcription factor motif opening occur during organoid differentiation, including known patterns that have been observed in single-cell transcriptomic studies of murine corticogenesis, and bulk ATAC-seq in fetal human samples[10,193]. The earliest waves can be explained by a neuroepithelial-like state, showing relative increases in transcription factor motifs associated with telencephalic commitment or symmetric division in proliferating radial glia cells, such as OTX2 or EMX2[194]. Progressive waves follow known programs of transcription factors linked to corticogenesis, with many transcription factor waves spanning cluster boundaries (Fig. 24d). Transient increases in transcription factors associated with radial-glial proliferation, e.g., POU3F3 (aka BRN1), EOMES and NFIX, precede the waves of transcription factors such as MEF2C, NEUROD6, NEUROG2, and BACH2, which are linked to neuronal migration and maturation[10,193].

In conclusion, we utilized our sci-ATAC assay to characterize a burgeoning model of neurodevelopment and provide, to our knowledge, the first single-cell chromatin accessibility profile of this model. Our findings not only recapitulate known waves of epigenomic reprogramming, but produce maps of regulatory element usage, revealing cascades of co-accessible patterns that incorporate novel loci.

Methods

Sample generation

iPSC culture

Forebrain organoids were differentiated in two parallel rounds from a human control iPSC line (CW20043) obtained from the California Institute for Regenerative Medicine (CIRM) repository at the Coriell Institute for Medical Research. The iPSCs were maintained feeder-free on 100 mm^2 dishes coated with 5 µg/mL vitronectin (Thermo Fisher, Cat. A14700) in StemFlex Medium (Thermo Fisher, Cat. A3349401) and kept inside a 37°C incubator with 4% O_2 and 5% CO_2. Prior to passaging or thawing, StemFlex Medium was supplemented with the following ROCK inhibitors to promote viability: 10 µM Y27632 (Stemgent; Cat. 04-0012-10) and 1 µM Thiazovivin (Stemgent, Cat. 04-0017). The culture medium was exchanged daily on weekdays; a double volume of media was provided on Fridays. The iPSCs were passaged at ~80-90% confluency as follows: cells were washed once with PBS then incubated with Versene (Thermo Fisher, Cat. 15040066) for 4 minutes at 37°C. After Versene was aspirated, StemFlex medium containing 10 µM Y27632 and 1 µM Thiazovivin was added to the culture dish to collect cells. The iPSC suspension was further diluted in StemFlex with Y27632 and Thiazovivin then transferred to new vitronectin-coated dishes using a 1:12-1:25 split ratio depending on colony density prior to passaging. CryoStor10 freezing medium (STEMCELL Technologies, Cat. 07930) was used for cryopreservation of iPSCs.

Differentiation of forebrain organoids

Differentiation of forebrain organoids from human iPSCs was carried out using a previously described protocol with certain modifications[55,58]. In this modified protocol, cortical neurogenesis is initiated in AggreWell800 plates (STEMCELLTechnologies; Cat. 34850) to produce uniformly-sized embryoid bodies that are chemically induced to develop into forebrain-like tissues through the addition of small molecule SMAD inhibitors. Refer to Appendix Table 16 for the regimen of factors to add to the differentiation medium used during feedings. AggreWell800 plates were prepared for aggregate culture according to manufacturer instructions using STEMdiff Neural Induction Media (NIM; STEMCELL Technologies, Cat. 05835)

supplemented with 10 µM Y27632, 2 µM Thiazovivin, 2 µM Dorsomorphin (Tocris, Cat. 3090), and 2 µM A83-01 (Tocris, Cat. 2939) then set aside until needed. Of note, all supplemental factors (e.g. small molecules, recombinant proteins) added to differentiation medium were done so immediately prior to feeding.

To begin neural induction (day 0 of *in vitro* differentiation; DIV 0), two 80-90% confluent 100 mm^2 dishes of iPSCs were washed once with 10 mL PBS then treated with 5 mL Accutase (STEMCELL Technologies, Cat. 07920) for 8 minutes at 37°C. Colonies were disaggregated into a single cell suspension by pipetting up and down for ~10 seconds with a 5 mL serological pipet, then for an additional 3-5 seconds with a P1000 pipet. The single iPSCs in Accutase were transferred to 50 mL conical tube then the 100 mm^2 dishes were immediately washed twice with 10 mL of DMEM/F12 with GlutaMax (Thermo Fisher, Cat. 10565018) that was added to the same 50 mL tube containing cells. Residual clumps of cells were removed by passing the suspension through a 40 µm cell strainer into a new 50 mL conical tube. Filtered iPSCs were centrifuged at 200 x g for 5 minutes, after which, the supernatant was aspirated and the pellets resuspended in 1 mL NIM containing 10 µM Y27632, 2 µM Thiazovivin, 2 µM Dorsomorphin, and 2 µM A83-01. Cells were kept on ice while counting and performing calculations.

Each well in the AggreWell800 plate contains 300 microwells. Each of these microwells is used to form a single organoid initially comprised of approximately 10,000 cells. To achieve this, we seeded 3,000,000 iPSCs per well of the AggreWell800 plate, centrifuged the plate at 100 x g for 3 minutes to collect ~10,000 cells into each of the 300 microwells, then incubated at 37 °C with 5% CO_2. Organoids were cultured in the AggreWell800 plate for 5 days with a daily 75% medium change. On day 2, Y27632 and Thiazovivin were omitted from the differentiation medium. On day 5, organoids were harvested from the AggreWell800 plate according to manufacturer instructions using wide-bore P1000 tips that were prepared by cutting off tips with a clean pair of scissors then autoclaved to sterilize. Organoids collected from a single well of the AggreWell800 plate were transferred to one 60 mm^2 ultra-low attachment dish (Corning, Cat. 3261) in Forebrain Differentiation Medium I (FDM I) comprised of DMEM/F12 with GlutaMax, 1% N2 Supplement (Thermo Fisher, 17502048), 1% MEM Non-Essential Amino Acids (Thermo

Fisher, Cat. 11140050), 1% Penicillin/Streptomycin (Thermo Fisher, Cat. 15140122), and 1 µg/mL heparin (Sigma, Cat. H3149) supplemented with 1 µM CHIR99021 (Stemgent, Cat. 04-0004) and 1 µM SB431542 (Stemgent, Cat. 04-0010). On day 6 (the following day), organoids were embedded in 20 µL droplets of Growth Factor Reduced (GFR) Matrigel (Corning, Cat. 354230) using wide-bore P200 tips as previously described[57], then returned to ultra-low attachment suspension culture in FDM I with 1 µM CHIR99021 and 1 µM SB431542 and for another 8 days with medium changes every 2 days.

On day 14, GFR Matrigel-embedded organoids were transferred to culture in either a nylon (laser sintered) or ULTEM 9085m 3D-printed Spin Omega 12-well miniature bioreactor[55,58] (refer to Qian *et al*[55]*.* for detailed instructions on bioreactor 3D-printing and assembly) in Forebrain Differentiation Medium II (FDM II) consisting of a 1:1 mix of DMEM/F12 with GlutaMax and Neurobasal (Thermo Fisher, Cat. 17504044) with 1% N2 Supplement, 2% B27 Supplement (Thermo Fisher, Cat. 17504044), 0.5% GlutaMax (Thermo Fisher, Cat. 35050061), 1% MEM Non-Essential Amino Acids, 1% Penicillin/Streptomycin, 2.5 µg/mL Insulin (Sigma, Cat. I9278), and 50 µM 2-Mercaptoethanol (Sigma, Cat. M3148) with medium changes every 2-3 days. Approximately every two weeks, organoids were transferred to different wells in a new 12-well bioreactor plate to avoid position effects[55,58]. On day 70, we began adding 0.2 mM L-Ascorbic Acid (Sigma, Cat. A4403), 0.5 mM cAMP (Sigma, Cat. A9501), 20 ng/mL BDNF (Peprotech, Cat. 450-02), and 20 ng/mL GDNF (Peprotech, Cat. 450-10) to the FDM II and continued exchanging the medium every 2-3 days up until day 90 when the experiment ended.

Organoid freezing protocol

Single organoids were transferred to individual 1.5 mL tubes using a P1000 pipet equipped with a wide-bore tip then pelleted by centrifugation at 500 x g for 2 minutes at 4 °C. After removing the supernatant, organoid pellets were flash frozen by placing the tubes in a slurry of ethanol and dry ice for approximately 2 minutes then transferred to a -80 °C freezer for storage.

Organoid characterization

Immunohistochemistry

Methods used to prepare organoids for cryosectioning were adapted from a previously described protocol[57]. In brief, wide-bore P200 or P1000 tips were used to transfer two to three organoids to single wells in a 24-well plate containing 250 µL medium for each cryosection block to be embedded. Organoids in the 24-well plate were washed once with 1 mL PBS then fixed with 1 mL of 4% PFA (Sigma, Cat. 158127) for 15 minutes at 4°C. Organoids were then washed three times with 1 mL PBS for 10 minutes at room temperature. The final PBS wash replaced with 30% sucrose (Sigma, Cat. S7903) with 0.02% sodium azide (Sigma, Cat. S2002) then the plate was incubated for 24-72 hours at 4°C. Subsequently, organoids were equilibrated in 1 mL of a 7.5% gelatin (Sigma; Cat. G1890)/10% sucrose embedding solution inside a 37°C incubator for 15-30 minutes. During this period, biopsy cryomolds (Tissue-Tek, Cat. 4565) for each well with organoids by coating the bottoms with a thin layer of the gelatin/sucrose embedding solution. Organoids were transferred to the cryomolds pre-coated with embedding solution using wide-bore tips then incubated at 4°C for 5 minutes. Cryomolds were then filled with embedding solution and allowed to solidify for 20-30 minutes at 4°C. The gelatin/sucrose blocks with organoids were frozen in a -30°C to -50°C isopentane bath for 2 minutes, after which, the cryomolds with frozen blocks were tightly wrapped with Parafilm and stored at -80°C until sectioned.

For immunohistochemistry, frozen blocks with embedded organoids were cut into 20 µm sections using a cryostat (Leica). Sections were serially collected onto Superfrost Plus microscope slides (Fisherbrand, Cat. 22-037-246), allowed to dry for ≥30 minutes, and then stored at -20°C or 4°C until ready for staining. Prior to staining, slides were equilibrated to room temperature then sections were circumscribed with a hydrophobic barrier pen (Invignome, Cat. GPF-VPSA-V). Sections were washed twice with PBS for 10 minutes then blocked for 1 hour at room temperature in permeabilization/blocking buffer comprised of PBS with 10% normal goat serum (Jackson ImmunoResearch, Cat. 005-000-121), 1% bovine serum albumin (BSA, Millipore, Cat. 126626), 0.3% Triton X-100 (TX-100, Sigma, Cat. 11332481001), 0.05% Tween-20 (Sigma, Cat. P1379), 0.3 M glycine (Sigma, Cat. G7126) and 0.01% sodium azide (Sigma, Cat. S2002) for 1 hour at room temperature. During the blocking step, primary antibodies (Appendix Table 16)

were diluted in a buffer containing PBS, 2% NGS, 1% BSA, 0.01% TX-100, 0.05% Tween-20, and 0.01% sodium azide. The diluted primary antibodies were applied to sections then incubated overnight at 4°C inside a StainTray (Simport Scientific, M922-2). Primary antibodies were washed from the sections five times with PBS for 5 minutes at room temperature. During wash steps, secondary antibodies (Appendix Table 17) were prepared by diluting 1:1000 in the same buffer used to dilute primary antibodies. Sections were incubated with the diluted secondary antibodies inside a StainTray for 1 hour in the dark at room temperature (sections were protected from light following secondary staining). Secondary antibodies were washed from the sections three times with PBS for 5 minutes, then nuclei were counterstained with DAPI (Thermo Fisher, Cat. D1306) for 10 minutes at room temperature. After DAPI staining, sections were washed an additional two times then glass coverslips were mounted with ProLong Diamond Anti-Fade Mounting Medium (Thermo Fisher, Cat. P36961).

Microscopy and image processing

Live organoid images were taken with a Nikon Ts2 inverted microscope. Optical sections of organoids were acquired with a Zeiss ApoTome AxioImager M2 fluorescent upright microscope and processed using Fiji software.

sci-ATAC on Organoids

Nuclei isolation and Tagmentation

sci-ATAC libraries were prepared on 16 organoids (differentiation 1: 3 DIV30, 3 DIV60, and 4 DIV90; differentiation 2: 2 DIV30, 2 DIV 60 and 2 DIV90) and two bulk preparations of DIV15 pooled cells from differentiation 2. Nuclei were isolated from the flash frozen pellets by resuspension with Nuclei Isolation Buffer (NIB; 10mM Tris HCl, pH 7.5 [Fisher, Cat. T1503 and Fisher, Cat. A144], 10mM NaCl [Fisher, Cat. M-11624], 3mM MgCl2 [Sigma, Cat. M8226], 0.1% IGEPAL [v/v; Sigma, I8896], 0.1% Tween-20 [v/v, Sigma, Cat. P7949] and 1x protease inhibitor [Roche, Cat. 11873580001]). For DIV 30 organoids, 300 µL NIB was used; for DIV 60 and DIV 90 organoid tube, 600 µL of NIB was used. Resuspension was done by 10-20 triturations of NIB solution to break up the pellet of cells. Cells were then incubated on ice for 10 minutes and then triturated another 10 times. The full volume of each sample was then each run through a 35 µm

cell strainer (BD Biosciences, Cat. 352235) and nuclei were stained with 3 µL of DAPI (5mg/mL, Thermo Fisher, Cat. D1306) for DIV30 samples, and 5 µL DAPI for DIV60 and DIV90 samples. To test library complexity improvement, one DIV 90 organoid sample was dissociated with 70 µM Pitstop2 NIB solution and another with 70 µM Pitstop2 NIB and the OMNI-ATAC digitonin nuclear isolation (Bivision 2082-1)[195], prepared as described above.

2X TD buffer (Illumina, Cat. FC-121-1031) aliquots were prepared with a supplementation of 3 mM Pitstop 2, to the appropriate final concentration (0 µM or 70 µM, see further described in Supplementary Note). 5µL of the appropriately supplemented 2X TD buffer was added to 5 µL (5,000 nuclei total) samples, respective of test condition (10 µL final volume). A Sony SH800 FACS machine was used to sort 5,000 nuclei (identified through DAPI gating) into each well of multiple 96 well plates. The wells containing sorted nuclei (n=432 wells) were then tagmented in parallel (Appendix Table 19). For each reaction, 1µL of 8 µM loaded indexed transposase was added (See Picelli *et al.* for loading protocol)[142]. The nuclei treated with Pitstop 2 prior to sorting were once again treated with a final concentration of 70 µM Pitstop 2 added to the TD buffer. All reactions were pooled respective of Pitstop 2 condition and 3µL of DAPI (5mg/mL) was added to each.

Sorting nuclei

96-well plates were prepared, each well containing 8.5 µL of protease buffer (PB; 30 mM Tris HCl, pH 7.5 [Fisher, Cat. T1503 and Fisher, Cat. A144], 2 mM EDTA [Ambion, Cat. AM9261, 20 mM KCl [Fisher, Cat. P217 and Fisher, Cat. A144], 0.2% Triton X-100 [v/v, Sigma, Cat. 9002-93-1], 500 ug/mL serine protease [Fisher, Cat. NC9221823). To each well, a combination of 1 µL 10 mM i5 and 1µL 10 mM i7 PCR primers (Appendix Table 19, IDT) containing a well-specific index combination was added. DAPI-stained nuclei pools were then sorted using a Sony SH800 FACS machine with sample and sorting chambers held at 5°C. Gating was performed to isolate a clean population of singlet nuclei.100 tagmented nuclei were deposited into each well of the PCR plates by FANS.

Transposase denaturation and PCR

Following sorting, plates were spun at 500 rcf for 5 minutes at 4°C to ensure nuclei were within the reaction buffer. Transposomes and any other proteins within the reaction were then degraded by the serine protease by holding the samples at 55°C for 20 minutes, the protease was then denatured by heating to 70°C for 30 minutes. Following this, 13.5 µL of PCR Master Mix (13 µL 2X KAPA Hotstart HiFi [Fisher, Cat. NC0465187], 0.25 µL (2 units) Bst3.0 [NEB, Cat. M4374] and 0.25 µL 100X SYBR Green I [FMC BioProducts, Cat. 50513]) was added to each well. Real-time (RT)-PCR was performed on a BioRad CFX machine, for the following temperatures and times: 72°C for 5 minutes, 98°C for 30 seconds, and then multiple rounds of 98°C for 30 seconds, 63°C for 30 seconds, 72°C for 1 minute and a SYBR plate read before starting the next cycle. PCR reactions were stopped when the SYBR readout for a majority of wells plateaus (19 cycles).

Library pooling, cleanup and sequencing

For the PCR plate, 10 µL of each well was pooled for clean-up and quantification. First, the 960 µL pool was concentrated on a PCR purification column (Qiagen, Cat. 28106) following manufacturer's protocol. DNA was eluted off the column in 50 µL 10mM Tris HCl, pH 8.0 (Fisher, Cat. T1503 and Fisher, Cat. A144). Library pools were then cleaned and size selected using SPRI beads generated as describe previously[188]. An equal volume of prepared SPRI beads (1X) were added to the library pools, and incubated at room temperature for 15 minutes. Beads were then pelleted on a magnetic rack and washed twice with 150 µL of freshly prepared 80% ethanol (v/v, Decon, Cat. 2705). Following the second wash, all remaining liquid was carefully removed from the tube without disrupting the pellet. Pellets were then allowed to dry for 10 minutes, before being resuspended in 50 µL 10mM Tris HCl, pH 8.0. This 1:1 volume SPRI bead clean-up was repeated a second time. For the final elution of the second cleanup, libraries were eluted in 27 µL 10mM Tris HCl, pH 8.0. 2 µL of this eluate was used for quantification with a Qubit HS Assay (Thermo Fisher, Cat. Q32851). Following this, libraries were diluted to 4 ng/µL and 1µL was run on an Agilent DNA HS BioAnalyzer (Agilent, Cat. 5067-4626). Libraries were then diluted based on BioAnalyzer-reported molarity in the 150-1000bp range and loaded on a NextSeq 500

(Illumina Inc.) sequencer High Capacity kit with a loading concentration of 1.2 pM with a custom sequencing protocol, for 75 cycles of read 1, 30 cycles of index 1 and index 2, and 75 cycles of read 2[31].

Computational Analysis

Raw code is available at mulqueenr.github.io/organoid

FastQ generation, index assignment, single-cell read set definition

Following sequencing libraries on the Illumina NextSeq 500, bcl files were converted to FastQ format using bcl2fastq (v2.19.0, Illumina Inc.) with the option “with-failed-reads". FastQ reads were then combined across sequencing runs. Read sets were allocated to index barcodes with an allowance of two Hamming distance from possible index combination (perl v5.16.3, custom script). FastQ-format reads were modified so as to have the accepted indexing barcode (cell ID) as the read name. FastQ files were then aligned to the hg38 reference genome (GRCh38, NCBI) via bwa-mem (v0.7.15-r1140)[183]. The resulting aligned reads then underwent a removal of duplicate reads based on unique cell ID, chromosome and start sites of reads. The number of total reads and persisting unique reads was used for a comparison of library complexity (unique reads/total reads respective of cell ID). Single-cell libraries were defined by cell ID read sets containing a unique read cut-off of at least 1,000 reads with Q>=10 mapping quality. In total we generated 35,590 sci-ATAC libraries with a mean unique read count per cell of 18,939.

Generation of counts matrix and cisTopic dimensionality reduction

We called peaks as read-pileup regions using MACS2 (v.2.2.7.1)[72]. In total, we uncovered 183,391 peaks. We used the filtered peaks from our data set and the cell ID-associated, deduplicated reads to generate a **cell ID by peak read count matrix**, wherein each element within the matrix describes the number of reads from the respective cell ID overlapping with the peak feature. For each single-cell library an average of 30.80 ± 11.85 % (mean ± s.d.) of reads overlapped with peaks. Tabix formatted files were generated using samtools and tabix (v1.7). The counts matrix and tabix files were then input into a SeuratObject for Signac (v1.0.0)

processing[75,169]. Peak-set motif analysis was performed by chromVAR[78] on JASPAR2020 motif elements (release 8)[77]. With the *Signac* functions: *getMatrixSet*, and *CreateMotifMatrix*.

We performed LDA-based dimensionality reduction via cisTopic (v0.3.0)[74] on our cell ID by peak matrix. We used 27 topics. The number of topics were selected after generating 10 parallel models topic counts of 20-30 and selecting the topic count using *selectModel* based on the second derivative of model perplexity. Cell clustering was performed with Signac *FindNeighbors* and *FindClusters* functions on the topic weight × cellID data frame. For *FindClusters* function call, resolution was set to 0.2. The respective topic weight × cellID was then projected into two dimensional space via a uniform manifold approximation and projection ("UMAP") by the function umap in the uwot package (v0.1.80)[185]. To check for putative doublets within-species, we then ran scrublet analysis (Appendix Table 20) and removed scrublet-identified doubles from further analysis[179]. We plotted the projection with various coloring schemes via *ggplot2* (v3.3.2) matching annotations for cluster assignment, organoid source, differentiation experiment, DIV, and Pitstop 2 treatment, scrublet identification of doublets and cells passing quality control filters (Fig 21b). We found an unequal proportion of DIV-sourced cells through the clusters, suggesting shifting cell type populations through differentiation (Fig. 21b). We plotted this change in cluster proportion using *ggplot2 geom_bar* function with arguments *position="fill",* and *stat="identity"*.

Cell type assignment

Since the ATAC-seq signal is an indirect measurement of genomic active regions, and a large portion of existing data on corticogenesis and organoid differentiation assay transcript counts, we sought to better correlate our ATAC data with RNA-seq. To do this, we generated cis-coaccessible networks (CCANs) anchored at promoter regions to incorporate putative enhancer activity. This was done via the *cicero* package (v1.3.4.10)[20]. Per cell we generated a gene activity score based on the read counts across the promoter-anchored CCANs. We then used this to infer cell type similarities with known corticogenic marker genes. We took the mean value per cluster and Z-scored across clusters via the *scale* function (base R). This was performed for the markers of neuroepithelia, radial-glia, intermediate progenitors, excitatory and inhibitory neurons defined in

ref [24]. This was then plotted via *ComplexHeatmap* (v2.5.5). We performed the same scaling and plotting for *chromVAR* defined motif-accessibility for transcription factors defined in ref [24] and plotted along the corresponding motifs plotted through the *Signac* function *MotifPlot*. Finally, we performed a label-transfer method on our gene activity scores compared to single-cell RNA data. We took single-cell RNA data generated in ref [52] and the self-reported marker gene set, and performed cross-modality integration through canonical correlation analysis (CCA) as described previously[75]. The mean predicted value was summarized per cluster, scaled and plotted as described above.

Addition of Module Scores from Gene Sets

To assess the role of regulatory networks on organoid differentiation we performed several analyses of pre-defined gene sets on our gene activity scores per cell. To measure cell cycle scoring, we used genes listed as important for S-phase or G2/M-phase in radial glia, defined in ref [11] (573 and 462 genes, respectively). To calculate cell-cycle scores per cell, we used the *Seurat* function *CellCycleScoring* supplied with the lists of corticogenic specific S-phase and G2/M-phase genes on our gene activity matrix[169]. Additionally transcription factor gene networks scores were defined through neocortical development mid-gestation in humans[11]. We calculated the enrichment of gene activity in these sets, by the *Seurat AddModuleScore* function. We further calculated module scores for eigengenes across primary tissue and organoid differentiation described by Pollen *et al*[48] using the same method.

Differential Motif Accessibility and Gene Activity Scores

To calculate differential motif accessibility, differential TF module accessibility, and differential gene activity scores across clusters, we used the Signac function *FindMarkers* using logistic regression and using the read count within peaks as a latent variable[75].

Monocle Trajectories and Pseudotime Analysis

A pseudotime trajectory was generated through the use of monocle3[26] with the without partitioning or closing loops additionally, minimal branch length of 20. Following this, cells were assigned a pseudotime through residual values to the trajectory. Calculation of bias across pseudotime was performed through Moran's I test, using monocle3's *graph_test* function. Cells

were then split into 25 bins and chromVAR motif values were summarized per bin. Motifs found to be nominally significant (q value <0.1) were plotted as a heatmap and rows were hierarchically clustered by Euclidean distance.

Summary and Discussion

The work reported here contains a framework for several new methodologies for single-cell library generation and analysis. To address concerns of the balance between low information content per cell, and rare cell type dropout, I describe improvements on both cell throughput and information content.

The role of the methylome in genome-wide regulation is understudied and requires the jump to a commonplace single-cell methylation protocol. In the novel protocol, sci-MET (described in Chapter 1), I developed a method for high-cell count DNA methylation analysis. In this work I both demonstrate the validity of our results, replicating bulk methylome profiles from single cells, and demonstrating the ability to discriminate mixed cell types. Notably, within our mixture of tissue culture lines, we are able to reliably discern specific methylation profiles of fibroblast cells from lymphoblast cells when fibroblasts are less than 5% of the cell population. This suggests a power to discern rare cell types in complex mixtures. Additionally, we demonstrate this strategy on the mouse cortex, displaying an ability to separate out neuronal subtypes. However, single-cell methylation profiling thus far has focused on few idealized cases such as the mouse cortex[19] or embryonic stem cells[113,124,133], which are known to have unique methylation profiles compared to other somatic tissues[111]. Methods for clustering neurons largely rely on the methylation of CH sites, rarely seen, or completely absent in most somatic tissues[19]. As methylation profiling becomes further recognized for biomarkers of cancer[206,207], sci-MET will likely play an important role in building an atlas of differentially methylated regions. Such an atlas is critical for our understanding of progressive changes to the methylome in cancer samples or methylation changes in the “cancer field effect”[208], and atlas-scale single-cell methylation data sets are required for our generalized understanding of cell type and state methylome changes. This is the first and still only non- “single-cell, single well” method of single-cell methylome

profiling and remains a promising protocol for larger scale future studies[19,115,124,132,133]. Further, since it follows the same premise of sci protocols, it is directly adaptable to a spatial analysis used in the micro-biopsy punch derived sci-MAP protocol[209]. This opens new avenues of investigation in methylation profiling.

One constant limitation of single-cell experiments is the low capture rate of targeted genomic regions. Lower coverage of target regions leads to noisy signal, thus making cell-to-cell comparisons underpowered. In Chapter 2, I described a generalized adaptation to sci- protocols: s3. This change to the molecular design greatly increases captured molecules per cell. I describe improvements of single-cell ATAC libraries by over an order of magnitude, while maintaining cell count throughput. Through this, I described the chromatin profiles of both human cortex and mouse whole brain samples at an information density per cell never before achieved, to the best of my knowledge[29,154,155]. I uncovered differentially accessible genomic regions, and cis-coacessibility networks (CCANs) centered on genomic regions that will inform future cell type discrimination in the cortex. The increased coverage of accessible regions per cell allowed for robust CCANs which were used to infer cell activity. Even with samples as low as 50-80 cells, I could discern unique cell states with proper statistical power. This demonstrates that s3-ATAC is a powerful tool for discrimination of cell types in highly complex samples, and to make the most of rare state capture events.

I also adapt the s3 strategy to whole genome sequencing and genome conformation capture. In these assays, I improve coverage per cell to analogous assays by over 100 fold and over 10 fold, respectively. I apply these protocols in a proof-of-concept study to a patient-derived model of pancreatic ductal adenocarcinoma and survey the wide spread genomic instability. In this, I see reproducible profiles of dramatic genomic-changes that are masked in bulk whole-exome analysis. I uncover subclonal copy number changes that include genes important for PDAC invasion potential[161]. I also use the genome conformation information to uncover a subclonal translocation across a previously reported PDAC-associated locus[163]. This is, to my knowledge, the first time a single-cell Hi-C read-out has been used to uncover subclonal translocations in a sample. We demonstrate that coverage across the genome for non-distal

reads is still sufficient for copy-number calling. The study of subclonal genomic conformation changes, as well as cryptic structural variations like inversions and translocations is an interesting view into genome regulation changes that could be associated with treatment-targetable cancer-drivers[200]. This protocol has the potential to not only inform studies on cancer-derived genomic instability from a general biological sense, but also serve as an easily automated replacement of karyotyping or array CGH that far exceeds the cell count and resolution currently attainable[201,202]. This could prove invaluable to the study of non-random translocation and inversion events during cancer progression, a stated goal of present large scale atlas efforts[203].

One open question in the field of single-cell omics is the relation between epigenomic and transcriptomic regulation. Cross-assay integration, the co-emedding of epigenetic and transcriptomic measures, remains difficult, but is currently gaining traction[75]. This difficulty reflects our incomplete understanding of how epigenomic and transcriptomic programs intersect. New methods allow for the assessment of both chromatin and transcript counts within a single-cell, providing a valuable truth-state for our understanding[113,196,197]. Currently, many studies remain underpowered, with data too sparse for true generalization[198]. However, s3 protocol development is a promising route forward for high depth capture of single-cell transcriptomics with epigenetics, or even an additional bisulfite conversion step for methylome analysis. This adaptation to sci protocols captures more information per cell than previously described by over an order of magnitude for multiple assays. This level of depth per cell has huge potential for high quality data sets and unique analyses in future experiments.

As ATAC-seq library generation becomes more commonplace, a need for simple high-cell count library generation increases. In Chapter 3, I describe sci-DROP, an adaptation to a widely available commercialized droplet-based single-cell ATAC product. This improvement increases cell throughput by >15-fold, generating single-cell atlas level cell counts while both driving down the cost per cell, and maintaining the ability to multiplex samples[80]. In this work we deeply sequence ~80,000 cells from the human cortex and mouse brain. We describe an atlas of mature human cortical and mouse whole brain cell types, describing marker sites both at the genomic locus and co-accessible network level. Interestingly, we find that co-accessible networks

sometimes fail to discriminate cell types at known RNA-described markers. This reflects the movement of standard single-cell analysis pipelines away from a handful of marker genes to a more nuanced, holistic label-transfer methodology[75]. The method described here shows minimal cross-talk between cells in the same capture droplet, and could theoretically scale significantly higher. Additionally, because tagmentation reactions are multiplex, we were able to run multiple samples on a single reaction downstream in a single tube. To show this and the cell count throughput possible, in a single tube reaction, I was able to catalog both the complex epigenomic landscape of a human cortical sample, and across a mouse whole brain preparation. Following the work I described, generating a million cells libraries can now only take a single day experiment and cost the price of a 10X Chromium kit[30]. The cost burden becomes on almost entirely on sequencing effort. With this expansion on throughput, organismal cell atlases are attainable — expanding the possibility of single-cell assay use in building new species reference atlases for evolutionary biology comparisons and multi-tissue effect studies.

Finally, in Chapter 4, I also describe chromatin dynamics during a mid-gestational model of human corticogenesis. I generated, to my knowledge, the first single-cell ATAC profile of organoid differentiation. Through this, I was able to both assess chromatin structure relative to what is expected from primary bulk and single-cell samples. I used the active differentiation in this model system to uncovered dramatic changes in transcription factor motif usage, and the progressive recruitment of enhancers to mature glutamatergic marker gene promoters. In terms of corticogenic model systems, forebrain organoids remain a promising method. Efforts are underway to address cellular stress which is commonly seen in RNA profiling, and that we have observed through our ATAC data. This issue reflects poor nutrient and oxygen transfer to cells, driving them to a glycolytic metabolism. Recent work has introduced a functional vascularization that addresses this issue[204], while another has demonstrated a xenografted approach[52]. In both approaches, organoids displayed an improved transcriptomic correlation to human mid-gestational cortex. The question of how valid these models can become and their utility is still an open question. Thus far, there are few cases of cortical organoids modelling neurodevelopmental and neurodegenerative disorders. Recently, An AAV1-based gene knockout model of *GLB1* in

cortical organoids have been used recently for the study of the neurodegenerative disorder, GM1 gangliosidosis[205]. In addition cortical organoids were infected with the Zika virus to assess its role in microcephaly[58]. Improvements to organoid fidelity is a necessary component of model development, especially in generalizing results from perturbation experiments to our collective understanding of brain development. The work described here is a necessary step for future work in organoid model systems, to account for cell stressors present in organoid generation and understand the timing of differentiation in culture.

Taken together, this body of work provides relatively low-cost, open-source, easily adoptable and adaptable methods as a resource for the field of single-cell omics. This is necessary to democratize single-cell assays for application on new samples, and for the further development of new methods. Beyond that, the assays described here are scalable. Commercialized products for single-cell technologies are set to discrete sample sizes (~5000 cells dependent on the assay). However, the small experiments to test biochemical reaction efficiencies are absolutely critical to moving the field forward. To help in this effort, I have developed and described methods that prioritize i) scalability, ii) avoiding the use of specialized equipment, and iii) affordability. All of these criteria are crucial to the adoption of a method. The sci- protocols described here can be limited to as few as a couple dozen cells and expanded to hundreds of thousands in their output. This allows for external research groups to perform small-scale tests and avoid prohibitive sequencing costs. Since these biochemical reactions have been shown to increase in efficiency through scaling upwards, small scale tests remain directly applicable as larger data sets are needed. This allows for new research groups to adapt these methods for their own use, without having to brute-force cell throughput at prohibitively high cost and effort. Further, many single-cell assays require specialized equipment for cell isolation such as the 10X Genomics of BioRad microfluidics controllers; cell isolation in sci protocols can be performed with just a pipette to dilute nuclei into a 96-well plate. Finally, sci- protocols are relatively affordable. Costs have been scaled to the point where they cost less than a cent to generate a cell library, but more critically the barrier to entry is small. Enzymes and reagents are commercially available, with high quality unloaded Tn5 enzyme becoming recently more

accessible. Together, all of these considerations allow for protocol development in a method that can be readily adopted.

Single-cell omics has proven a powerful tool in capturing cellular heterogeneity and are rapidly gaining popularity for their unbiased approach. This collective body of work demonstrates the open-source development of adaptive novel methods and their applications to several single-cell omic assays. This not only advances our understanding of basic cellular biology of cortical development, but provides valuable new avenues of inquiry for future use across biological questions.

References

1. Waddington, C. H. Genetic Assimilation of the Bithorax Phenotype. *Evolution (N. Y).* **10**, 1 (1956).
2. Mo, A. *et al.* Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron* **86**, 1369–1384 (2015).
3. Tuoc, T. C. *et al.* Chromatin regulation by BAF170 controls cerebral cortical size and thickness. *Dev. Cell* **25**, 256–69 (2013).
4. Herculano-Houzel, S. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 10661–10668 (2012).
5. Sakai, D., Dixon, J., Dixon, M. J. & Trainor, P. A. Mammalian Neurogenesis Requires Treacle-Plk1 for Precise Control of Spindle Orientation, Mitotic Progression, and Maintenance of Neural Progenitor Cells. *PLoS Genet.* **8**, e1002566 (2012).
6. Mariani, J. *et al.* FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. *Cell* **162**, 375–390 (2015).
7. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
8. Nardone, S. *et al.* ORIGINAL ARTICLE DNA methylation analysis of the autistic brain reveals multiple dysregulated biological pathways. *Transl. Psychiatry* **4**, e433-9 (2014).
9. Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018).
10. Telley, L. *et al.* Sequential transcriptional waves direct the differentiation of newborn neurons in the mouse neocortex. *Science* **351**, 1443–6 (2016).
11. Polioudakis, D. *et al.* A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron* **103**, 785-801.e8 (2019).
12. Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. R. & Pollard, K. S. Many human accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. B Biol. Sci.* **368**, 20130025 (2013).
13. Nagy, C. *et al.* Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci.* **23**, 771–781 (2020).
14. Velmeshev, D. *et al.* Single-cell genomics identifies cell type–specific molecular changes in autism. *Science (80-.).* **364**, 685–689 (2019).
15. Schirmer, L. *et al.* Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature* **573**, 75–82 (2019).
16. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* 21.29.1-21.29.9 (2015). doi:10.1002/0471142727.mb2129s109
17. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).* **326**, 289–293 (2009).
18. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 1827–1831 (1992).
19. Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science (80-.).* **357**, 600–604 (2017).
20. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **71**, 858-871.e8 (2018).
21. Raphael, B. J. *et al.* Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* (2017). doi:10.1016/j.ccell.2017.07.007
22. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
23. McGregor, K. *et al.* An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome Biol.* **17**, 84 (2016).
24. Song, M. *et al.* Cell-type-specific 3D epigenomes in the developing human cortex. *Nature* **587**, 644–649 (2020).

25. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
26. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
27. Mezger, A. *et al.* High-throughput chromatin accessibility profiling at single-cell resolution. *Nat. Commun.* **9**, 1–6 (2018).
28. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).
29. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* (2019). doi:10.1038/s41587-019-0147-6
30. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
31. Sinnamon, J. R. *et al.* The accessible chromatin landscape of the hippocampus at single-cell resolution. doi:10.1101/407668
32. Sos, B. *et al.* Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biol* **17**, 20 (2016).
33. Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).
34. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
35. Economo, M. N. *et al.* Distinct descending motor cortex pathways and their roles in movement. *Nature* **563**, 79–84 (2018).
36. Lim, L., Mi, D., Llorca, A. & Marín, O. Development and Functional Diversification of Cortical Interneurons. *Neuron* **100**, 294–313 (2018).
37. Jäkel, S. & Dimou, L. Glial cells and their function in the adult brain: A journey through the history of their ablation. *Frontiers in Cellular Neuroscience* **11**, 24 (2017).
38. Stiles, J. & Jernigan, T. L. The basics of brain development. *Neuropsychol. Rev.* **20**, 327–348 (2010).
39. Fan, X. *et al.* Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis. *Cell Res.* **28**, 730–745 (2018).
40. Götz, M. & Huttner, W. B. The cell biology of neurogenesis. *Nature Reviews Molecular Cell Biology* **6**, 777–788 (2005).
41. Heins, N. *et al.* Emx2 promotes symmetric cell divisions and a multipotential fate in precursors from the cerebral cortex. *Mol. Cell. Neurosci.* **18**, 485–502 (2001).
42. Manuel, M. N., Mi, D., Mason, J. O. & Price, D. J. Regulation of cerebral cortical neurogenesis by the Pax6 transcription factor. *Front. Cell. Neurosci.* **9**, 70 (2015).
43. Noctor, S. C., Flint, A. C., Weissman, T. A., Dammerman, R. S. & Kriegstein, A. R. Neurons derived from radial glial cells establish radial units in neocortex. *Nature* **409**, 714–720 (2001).
44. Zahr, S. K. *et al.* A Translational Repression Complex in Developing Mammalian Neural Stem Cells that Regulates Neuronal Specification. *Neuron* **97**, 520-537.e6 (2018).
45. Lancaster, M. A. *et al.* Cerebral organoids model human brain development and microcephaly. *Nature* **501**, 373–379 (2013).
46. Rakic, P. A small step for the cell, a giant leap for mankind: a hypothesis of neocortical expansion during evolution. *Trends Neurosci.* **18**, 383–388 (1995).
47. Otani, T., Marchetto, M. C., Gage, F. H., Simons, B. D. & Livesey, F. J. 2D and 3D Stem Cell Models of Primate Cortical Development Identify Species-Specific Differences in Progenitor Behavior Contributing to Brain Size. *Cell Stem Cell* **18**, 467–480 (2016).
48. Pollen, A. A. *et al.* Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell* **176**, 743-756.e17 (2019).
49. Pollen, A. A. *et al.* Molecular Identity of Human Outer Radial Glia during Cortical Development. *Cell* **163**, 55–67 (2015).
50. Camp, J. G. *et al.* Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15672–7 (2015).
51. Pollen, A. A. *et al.* Establishing Cerebral Organoids as Models of Human-Specific Brain

Evolution. *Cell* **176**, 743-756.e17 (2019).
52. Bhaduri, A. *et al.* Cell stress in cortical organoids impairs molecular subtype specification. *Nature* **578**, 142–148 (2020).
53. Zhang, S. C., Wernig, M., Duncan, I. D., Brüstle, O. & Thomson, J. A. In vitro differentiation of transplantable neural precursors from human embryonic stem cells. *Nat. Biotechnol.* **19**, 1129–1133 (2001).
54. Hu, B. Y. & Zhang, S. C. Directed differentiation of neural-stem cells and subtype-specific neurons from hESCs. *Methods Mol. Biol.* **636**, 123–137 (2010).
55. Qian, X. *et al.* Brain-Region-Specific Organoids Using Mini-bioreactors for Modeling ZIKV Exposure. *Cell* **165**, 1238–1254 (2016).
56. Lancaster, M. A. & Knoblich, J. A. Generation of cerebral organoids from human pluripotent stem cells. *Nat. Protoc.* **9**, 2329–40 (2014).
57. Lancaster, M. A. & Knoblich, J. A. Generation of cerebral organoids from human pluripotent stem cells. *Nat. Protoc.* **9**, 2329–40 (2014).
58. Qian, X. *et al.* Generation of human brain region–specific organoids using a miniaturized spinning bioreactor. *Nat. Protoc.* **13**, 565–580 (2018).
59. Dominguez, M. H., Ayoub, A. E. & Rakic, P. POU-III transcription factors (Brn1, Brn2, and Oct6) influence neurogenesis, molecular identity, and migratory destination of upper-layer cells of the cerebral cortex. *Cereb. Cortex* **23**, 2632–43 (2013).
60. Quadrato, G. *et al.* Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* **545**, 48–53 (2017).
61. Shi, Y. *et al.* Vascularized human cortical organoids (vOrganoids) model cortical development in vivo. *PLOS Biol.* **18**, e3000705 (2020).
62. Cadwell, C. R., Bhaduri, A., Mostajo-Radji, M. A., Keefe, M. G. & Nowakowski, T. J. Development and Arealization of the Cerebral Cortex. *Neuron* **103**, 980–1004 (2019).
63. Simi, A. & Studer, M. Developmental genetic programs and activity-dependent mechanisms instruct neocortical area mapping. *Current Opinion in Neurobiology* **53**, 96–102 (2018).
64. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
65. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
66. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
67. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370**, (2020).
68. Cooper, J., Ding, Y., Song, J. & Zhao, K. Genome-wide mapping of DNase I hypersensitive sites in rare cell populations using single-cell DNase sequencing. *Nat. Protoc.* **12**, 2342–2354 (2017).
69. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, R119 (2010).
70. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
71. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–9 (2012).
72. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
73. Cusanovich, D. A. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (80-.).* **348**, 910–914 (2015).
74. González-Blas, C. B. *et al.* Cis-topic modelling of single cell epigenomes. *bioRxiv* 370346 (2018). doi:10.1101/370346
75. Stuart, T., Srivastava, A., Lareau, C. & Satija, R. Multimodal single-cell chromatin analysis with Signac. *bioRxiv* 2020.11.09.373613 (2020). doi:10.1101/2020.11.09.373613
76. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2018).
77. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor

binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).
78. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
79. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
80. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324.e18 (2018).
81. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
82. Loeb, L. A. A mutator phenotype in cancer. *Cancer Res.* **61**, 3230–3239 (2001).
83. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–95 (2011).
84. Paez, J. G. *et al.* Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.* **32**, e71 (2004).
85. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science (80-.).* **338**, 1622–1626 (2012).
86. Yin, Y. *et al.* High-Throughput Single-Cell Sequencing with Linear Amplification. *Mol. Cell* **76**, 676-690.e10 (2019).
87. Laks, E. *et al.* Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. *Cell* **179**, 1207-1221.e22 (2019).
88. Zahn, H. *et al.* Scalable whole-genome single-cell library preparation without preamplification. *Nat. Methods* **14**, 167–173 (2017).
89. Vitak, S. A. *et al.* Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14**, 302–308 (2017).
90. Wang, R., Lin, D. Y. & Jiang, Y. SCOPE: A Normalization and Copy-Number Estimation Method for Single-Cell DNA Sequencing. *Cell Syst.* **10**, 445-452.e6 (2020).
91. Garvin, T. *et al.* Interactive analysis and assessment of single-cell copy-number variations. *Nature Methods* **12**, 1058–1060 (2015).
92. Wang, X., Chen, H. & Zhang, N. R. DNA copy number profiling using single-cell sequencing. *Brief. Bioinform.* **19**, 731–736 (2018).
93. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2017). doi:10.1101/201178
94. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics* **2**, 292–301 (2001).
95. Bickmore, W. A. The spatial organization of the human genome. *Annual Review of Genomics and Human Genetics* **14**, 67–84 (2013).
96. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164**, 1110–1121 (2016).
97. Rao, S. S. P. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
98. Osborne, C. S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.* **36**, 1065–1071 (2004).
99. Tan, L., Xing, D., Chang, C. H., Li, H. & Xie, X. S. Three-dimensional genome structures of single diploid human cells. *Science (80-.).* **361**, 924–928 (2018).
100. Hoffman, E. A., Frey, B. L., Smith, L. M. & Auble, D. T. Formaldehyde crosslinking: A tool for the study of chromatin complexes. *Journal of Biological Chemistry* **290**, 26404–26411 (2015).
101. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
102. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
103. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).

104. Ramani, V. *et al.* Sci-Hi-C: A single-cell Hi-C method for mapping 3D genome organization in large number of single cells. *Methods* **170**, 61–68 (2020).
105. Ramani, V., Qiu, R. & Shendure, J. High-throughput determination of RNA structure by proximity ligation. *Nat. Biotechnol.* **33**, 980–984 (2015).
106. Tan, L., Xing, D., Chang, C. H., Li, H. & Xie, X. S. Three-dimensional genome structures of single diploid human cells. *Science (80-.).* **361**, 924–928 (2018).
107. Chen, C. *et al.* Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science (80-.).* **356**, 189–194 (2017).
108. Liu, J., Lin, D., Yardımcı, G. G. & Noble, W. S. Unsupervised embedding of single-cell Hi-C data. *Bioinformatics* **34**, i96–i104 (2018).
109. Zhou, J. *et al.* Robust single-cell Hi-C clustering by convolution- And random-walk–based imputation. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 14011–14018 (2019).
110. Shin, H. *et al.* TopDom: An efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44**, e70 (2015).
111. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
112. Bird, A. DNA methylation patterns and epigenetic memory. *Genes and Development* **16**, 6–21 (2002).
113. Clark, S. J. *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* 1–9 doi:10.1038/s41467-018-03149-4
114. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science (80-.).* **341**, (2013).
115. Farlik, M. *et al.* Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Rep.* **10**, 1386–1397 (2015).
116. Jaffe, A. E. *et al.* Mapping DNA methylation across development , genotype and schizophrenia in the human frontal cortex. **19**, 4–7 (2016).
117. Lee, J. H., Park, S. J. & Nakai, K. Differential landscape of non-CpG methylation in embryonic stem cells and neurons caused by DNMT3s. *Sci. Rep.* **7**, 1–11 (2017).
118. Luo, C. *et al.* Cerebral Organoids Recapitulate Epigenomic Signatures of the Human Fetal Brain Resource Cerebral Organoids Recapitulate Epigenomic Signatures of the Human Fetal Brain. *CellReports* **17**, 3369–3384 (2016).
119. Guo, J. U. *et al.* Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.* **17**, 215–22 (2014).
120. Luo, C. *et al.* Cerebral Organoids Recapitulate Epigenomic Signatures of the Human Fetal Brain. *Cell Rep.* **17**, 3369–3384 (2016).
121. Schutsky, E. K. *et al.* Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat. Biotechnol.* **36**, 1083–1090 (2018).
122. Feng, S., Zhong, Z., Wang, M. & Jacobsen, S. E. Efficient and accurate determination of genome-wide DNA methylation patterns in Arabidopsis thaliana with enzymatic methyl sequencing. *Epigenetics and Chromatin* **13**, 42 (2020).
123. Guo, H. *et al.* Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nat. Protoc.* **10**, 645–659 (2015).
124. Smallwood, S. a *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–20 (2014).
125. Miura, F., Enomoto, Y., Dairiki, R. & Ito, T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res.* **40**, e136–e136 (2012).
126. Krueger, F. & Andrews, S. R. Bismark : a flexible aligner and methylation caller for Bisulfite-Seq applications. **27**, 1571–1572 (2011).
127. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
128. Jensen, S. Ø. *et al.* Novel DNA methylation biomarkers show high sensitivity and specificity for blood-based detection of colorectal cancer- A clinical biomarker discovery and validation study. *Clin. Epigenetics* **11**, 158 (2019).
129. Assenov, Y. *et al.* Comprehensive analysis of DNA methylation data with RnBeads. **11**, (2014).

130. Varley, K. E. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* **23**, 555–567 (2013).
131. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
132. Farlik, M. *et al.* Resource DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation Resource DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. 808–822 (2016). doi:10.1016/j.stem.2016.10.019
133. Angermueller, C. *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. **13**, (2016).
134. Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).
135. Adey, A. & Shendure, J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res.* **22**, 1139–43 (2012).
136. Vitak, S. A. *et al.* Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14**, 302–308 (2017).
137. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
138. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: Quantifying genome and methylome mappability. *Nucleic Acids Res.* **46**, e120 (2018).
139. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The Ensembl Regulatory Build. *Genome Biol.* **16**, 56 (2015).
140. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).
141. Lister, R. *et al.* Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science (80-.).* **341**, 1237905–1237905 (2013).
142. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–40 (2014).
143. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
144. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
145. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. (1996).
146. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).
147. Cusanovich, D. a *et al.* Epigenetics. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–4 (2015).
148. Sos, B. C. *et al.* Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biol.* **17**, 20 (2016).
149. Yin, Y. *et al.* High-Throughput Single-Cell Sequencing with Linear Amplification. *Mol. Cell* **76**, 676-690.e10 (2019).
150. Chen, C. *et al.* Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science (80-.).* **356**, 189–194 (2017).
151. Adey, A. & Shendure, J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res.* **22**, 1139–1143 (2012).
152. Mulqueen, R. M. *et al.* Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).
153. Adey, A. *et al.* In vitro , long-range sequence information for de novo genome assembly via transposase contiguity. 2041–2049 doi:10.1101/gr.178319.114.24
154. Preissl, S. *et al.* Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439 (2018).
155. Datasets -Single Cell ATAC -Official 10x Genomics Support. Available at: https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_adult_brain_fresh_5k. (Accessed: 31st December 2020)

156. Stuart, T., Srivastava, A., Lareau, C. & Satija, R. Multimodal single-cell chromatin analysis with Signac. *bioRxiv* 2020.11.09.373613 (2020). doi:10.1101/2020.11.09.373613
157. Raphael, B. J. *et al.* Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* **32**, 185-203.e13 (2017).
158. Lindenburger, K. *et al.* AB024. S024. Drug responses of patient-derived cell lines in vitro that match drug responses of patient PDAc tumors in situ. *Ann. Pancreat. Cancer* **1**, AB024–AB024 (2018).
159. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017).
160. Ahmed, S., Bradshaw, A.-D., Gera, S., Dewan, M. & Xu, R. The TGF-β/Smad4 Signaling Pathway in Pancreatic Carcinogenesis and Its Clinical Significance. *J. Clin. Med.* **6**, 5 (2017).
161. Wu, H. *et al.* PRSS1 genotype is associated with prognosis in patients with pancreatic ductal adenocarcinoma. *Oncol. Lett.* **19**, 121–126 (2020).
162. Sritangos, P. *et al.* Plasma membrane Ca2+ atpase isoform 4 (PMCA4) has an important role in numerous hallmarks of pancreatic cancer. *Cancers (Basel).* **12**, (2020).
163. Ahmad, M. K., Abdollah, N. A., Shafie, N. H., Yusof, N. M. & Razak, S. R. A. Dual-specificity phosphatase 6 (DUSP6): a review of its molecular characteristics and clinical relevance in cancer. *Cancer Biology and Medicine* **15**, 14–28 (2018).
164. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
165. Liu, X. *et al.* Conditional reprogramming and long-term expansion of normal and tumor cells from human biospecimens. *Nat. Protoc.* **12**, 439–451 (2017).
166. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
167. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
168. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
169. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
170. Jiang, Y. *et al.* CODEX2: Full-spectrum copy number variation detection by high-throughput DNA sequencing. *Genome Biol.* **19**, 202 (2018).
171. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
172. Drost, H.-G. Philentropy: Information Theory and Distance Quantification with R. *J. Open Source Softw.* **3**, 765 (2018).
173. Galili, T. dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718–3720 (2015).
174. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24–26 (2011).
175. pysam-developers/pysam. Available at: https://github.com/pysam-developers/pysam. (Accessed: 8th January 2021)
176. Kim, H. J. *et al.* Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data. *PLoS Comput. Biol.* **16**, (2020).
177. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
178. Lareau, C. A., Ma, S., Duarte, F. M. & Buenrostro, J. D. Inference and effects of barcode multiplets in droplet-based single-cell assays. *Nat. Commun.* **11**, 1–9 (2020).
179. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**, 281-291.e9 (2019).
180. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
181. Allen Institute for Brain Science. Allen Human Brain Atlas. (2020).

182. Allen Institute for Brain Science. Allen Brain Cell Types Database. (2019).
183. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–95 (2010).
184. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
185. Mcinnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection Software • Review • Repository • Archive. (2018). doi:10.21105/joss.00861
186. Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: Recording the past and predicting the future. *Science* **358**, 69–75 (2017).
187. Vitak, S. A. *et al.* Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14**, (2017).
188. Mulqueen, R. M. *et al.* Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).
189. Stergachis, A. B. *et al.* Developmental Fate and Cellular Maturity Encoded in Human Regulatory DNA Landscapes. *Cell* **154**, 888–903 (2013).
190. Amiri, A. *et al.* Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science* **362**, eaat6720 (2018).
191. Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science (80-.).* **358**, 1318–1323 (2017).
192. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).
193. De La Torre-Ubieta, L. *et al.* The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. (2018). doi:10.1016/j.cell.2017.12.014
194. Ziller, M. J. *et al.* Dissecting neural differentiation regulatory networks through epigenetic footprinting. *Nature* **518**, 355–359 (2015).
195. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
196. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
197. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science (80-.).* **361**, 1380–1385 (2018).
198. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).
199. Yuan, Y. & Bar-Joseph, Z. Deep learning for inferring gene relationships from single-cell expression data. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 27151–27158 (2019).
200. Singhi, A. D. *et al.* Identification of targetable ALK rearrangements in pancreatic ductal adenocarcinoma. *JNCCN J. Natl. Compr. Cancer Netw.* **15**, 555–562 (2017).
201. Vermeesch, J. R. *et al.* Guidelines for molecular karyotyping in constitutional genetic diagnosis. *Eur. J. Hum. Genet.* **15**, 1105–1114 (2007).
202. Kearney, H. M., Thorland, E. C., Brown, K. K., Quintero-Rivera, F. & South, S. T. American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet. Med.* **13**, 680–685 (2011).
203. Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell* **181**, 236–249 (2020).
204. Cakir, B. *et al.* Engineering of human brain organoids with a functional vascular-like system. *Nat. Methods* **16**, 1169–1175 (2019).
205. Latour, Y. L. *et al.* Human GLB1 knockout cerebral organoids: A model system for testing AAV9-mediated GLB1 gene therapy for reducing GM1 ganglioside storage in GM1 gangliosidosis. *Mol. Genet. Metab. Reports* **21**, 100513 (2019).
206. Huang, C. C., Du, M. & Wang, L. Bioinformatics analysis for circulating cell-free DNA in cancer. *Cancers (Basel).* **11**, 1–15 (2019).
207. Liu, M. C. *et al.* Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745–759 (2020).
208. Luo, Y., Yu, M. & Grady, W. M. Field cancerization in the colon: A role for aberrant DNA methylation? *Gastroenterology Report* **2**, 16–20 (2014).

209. Thornton, C. A. *et al.* Spatially mapped single-cell chromatin accessibility. *Nat. Commun.* **12**, 1–16 (2021).