Atlas-scale single-cell chromatin accessibility using nanowell-based combinatorial indexing

Brendan L. O'Connell, Ruth V. Nichols, Dmitry Pokholok, et al.

Genome Res. published online February 15, 2023
Access the most recent version at doi:10.1101/gr.276655.122

Method

Atlas-scale single-cell chromatin accessibility using nanowell-based combinatorial indexing

Brendan L. O'Connell,[1] Ruth V. Nichols,[1] Dmitry Pokholok,[2] Jerushah Thomas,[2] Sonia N. Acharya,[1] Andrew Nishida,[1] Casey A. Thornton,[1] Marissa Co,[1] Andrew J. Fields,[1] Frank J. Steemers,[2] and Andrew C. Adey[1,3,4,5]

[1]*Oregon Health & Science University, Department of Molecular and Medical Genetics, Portland, Oregon 97239, USA;* [2]*ScaleBio, San Diego, California 92121, USA;* [3]*Oregon Health & Science University, Knight Cancer Institute, Portland, Oregon 97239, USA;* [4]*Oregon Health & Science University, Cancer Early Detection Advanced Research Center, Portland, Oregon 97239, USA;* [5]*Oregon Health & Science University, Knight Cardiovascular Institute, Portland, Oregon 97239, USA*

Here we present advancements in single-cell combinatorial indexed Assay for Transposase Accessible Chromatin (sciATAC) to measure chromatin accessibility that leverage nanowell chips to achieve atlas-scale cell throughput (>10^5 cells) at low cost. The platform leverages the core of the sciATAC workflow where multiple indexed tagmentation reactions are performed, followed by pooling and distribution to a second set of reaction wells for polymerase chain reaction (PCR)-based indexing. In this work, we instead leverage a chip containing 5184 nanowells at the PCR stage of indexing, enabling a 52-fold improvement in scale and reduction in per-cell preparation costs. We detail three variants that balance cell throughput and depth of coverage, and apply these methods to banked mouse brain tissue, producing maps of cell types as well as neuronal subtypes that include integration with existing single-cell Assay for Transposase Accessible Chromatin (scATAC) and scRNA-seq data sets. Our optimized workflow achieves a high fraction of reads that fall within called peaks (>80%) and low cell doublet rates. The high cell coverage technique produces high unique reads per cell, while retaining high enrichment for open chromatin regions, enabling the assessment of >70,000 unique accessible loci on average for each cell profiled. When compared to current methods in the field, our technique provides similar or superior per-cell information with very low levels of cell-to-cell cross talk, and achieves this at a cost point much lower than existing assays.

[Supplemental material is available for this article.]

The emergence of single-cell assays has reshaped our understanding of cell types and states, with the ability to profile a wide range of molecular properties in thousands of single cells. One of these properties, chromatin accessibility, provides valuable insight into the regulatory state of cells by producing maps of putative regulatory element usage genome-wide. Single-cell assays that capture this property vary, but the large majority leverage some form of the Assay for Transposase Accessible Chromatin (ATAC-seq) (Buenrostro et al. 2013), where a hyperactive transposase is used to simultaneously fragment DNA and append sequencing adapters in a single step referred to as tagmentation. This process is constrained to regions of open chromatin, due to the steric hindrance of nucleosomes, much like the DNase hypersensitivity assays that preceded ATAC-seq. There are a number of strategies to achieve single cell profiles, the majority of which rely on in situ tagmentation followed by cell indexing during polymerase chain reaction (PCR), either in droplets (Lareau et al. 2019; Satpathy et al. 2019) or by the use of multiple tiers of indexing, that is, single-cell combinatorial indexing (sciATAC), where indexes are incorporated at both the tagmentation and PCR stage, without the need to isolate individual cells (Cusanovich et al. 2015). The primary advantages of sciATAC is that individual reactions are not required for each individual cell, reducing overall costs per cell and enabling high cell throughput without handling large numbers of microwell plates.

Similarly, the use of nanoliter-scale chips for single-cell ATAC-seq processing provides several advantages over microwell plate workflows, and has been previously demonstrated for processing individual cells in wells which achieved a nearly 20-fold improvement over existing techniques at the time (Mezger et al. 2018). The first is a reduction in reagent usage, with a 5184-well chip using the same reagent volumes as a single 96-well PCR plate, equating to a 52-fold improvement. This allows for far more reactions to be processed simultaneously. Like microwell plates however, chips can be subjected to additive chemistry, with multiple rounds of reagent additions to perform distinct processing steps, something that is either extremely difficult or not possible with droplet-based workflows. This is particularly valuable for tagmentation workflows, where prior to PCR, nuclei must be ruptured and library fragments freed from the tightly bound transposase, ideally using detergents often not compatible with droplet-based platforms.

We reasoned that the advantages of both combinatorial indexing and nanowell chips can be combined to achieve even greater cell throughput and per-cell cost savings. The flexibility of the ICELL8 instrument for 5184 nanowell chip processing allows for the deposition of an array of 72 × 72 indexing PCR primers and sufficient volumes to distribute hundreds of preindexed nuclei in each well with enough remaining volume for subsequent indexed PCR processing. Here we demonstrate this combined technology

Corresponding author: adey@ohsu.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.276655.122. Freely available online through the *Genome Research* Open Access option.

to produce high-throughput data sets on the mouse brain, as well as a reduced throughput workflow that achieves high coverage per cell. Due to the scaling of the PCR reaction on-chip, the costs for each of these techniques are greatly reduced compared to standard sciATAC or other single-cell ATAC-seq techniques, providing an avenue for atlas-scale single-cell ATAC-seq.

Results

Chip-based sciATAC enables 10^5 cell throughput

To establish proof-of-principle, we implemented a reduced-scale nanowell sciATAC (sciATACv1) experiment on the lymphoblastoid cell line, GM12878, that leveraged 864 indexed tagmentation reactions performed in microwell plates. Reactions were then pooled and distributed across 2304 nanowells (48 × 48) of a chip to a target of 150 preindexed nuclei per well (approximately 350,000 loaded), with each well containing a unique pair of indexed forward and reverse primers in a buffer containing the detergent SDS to facilitate DNA fragment release. The chip was then sealed and incubated to denature nuclei, histones, and Tn5 transposase, and then reloaded onto the instrument for the addition of PCR master mix, which also diluted the SDS to limit interference during amplification. PCR was then performed and all reactions pooled and purified prior to sequencing, alignment and processing using previously established methods (Li and Durbin 2009; Sinnamon et al. 2019; Thornton et al. 2021). In total, we obtained 95,367 sciATAC profiles with a mean passing read count per cell of 5833 (median = 3290) at a median sequencing saturation of 14.8%, approximated as the percentage of duplicate reads observed out of the total reads sequenced. A total of 224,399 peaks were called with a median fraction of reads in peaks (FRiP) of 0.71, suggesting high enrichment for open chromatin regions (Fig. 2E).

We then sought to fully leverage the high index space provided using nanowells by preparing a library from mouse brain tissue using 9 × 96 = 864 indexed tagmentation reactions, facilitated using liquid handling robotics (Agilent Bravo), and the full 72 × 72 = 5184 nanowells at the PCR indexing stage for a total index space of 4.48 M (Fig. 1A; Thornton et al. 2021). To examine doublet rates and the presence of cross-cell-contamination, we included a spike-in of ~5% human cells (K562 cell line). In total, 461,103 mouse cell profiles were produced, with a mean passing read count of 12,445 (median = 3275), equating to a mean properly paired fragment count of 2227 (median = 1116) at a median sequencing saturation of 30%. Using the read data 622,754 peaks were called, with 97.9% falling within previously called mouse DNase I hypersensitivity (DHS) sites (The ENCODE Project Consortium et al. 2020), and with 72.7% of peaks from the droplet single-cell ATAC-seq (dscATAC) data set overlapping with our peak calls (Fig. 2F; Lareau et al. 2019). These peaks produced a median FRiP of 0.59 (0.63 when removing blacklist regions) with an aggregate transcription start site (TSS) enrichment of 18.20 as measured by ENCODE methods and a single-cell enrichment of 5.48 (mean) and 6.26 (median) using ArchR (Granja et al. 2021). Dimensionality reduction and clustering produced 24 distinct clusters that were visualized in two-dimensions and assigned to cell types based on canonical marker gene accessibility profiles and cell type module scoring (Fig. 1B–D; Korsunsky et al. 2019; Granja et al. 2021). Despite the distinct clusters and passing overall quality metrics (Fig. 2A–G; Chen et al. 2018; Preissl et al. 2018; Lareau et al. 2019; Domcke et al. 2020; Mulqueen et al. 2021; Thornton et al. 2021), an assessment of the human cell spike-in revealed an underlying cross-cell contamination rate of 22% (Fig. 2G, left). We suspected that this baseline contamination was due to retained tagmentation adapters that would typically be removed during fluorescence-activated nuclei sorting (FANS), which is used to deposit preindexed nuclei into microwell plates for indexed PCR. The retained adapters can serve as primers during PCR, resulting in index swapping; however, because these random reads are derived from numerous cells spanning all cell types, the result is a low-level background with dominant signal from the true cell type (Fig. 1E).

Optimized nuclei washing retains high-throughput with minimal cross contamination

To address the cross contamination observed in our initial experiments, we developed several strategies to remove excess tagmentation adapters by focusing on optimized washing buffers, along

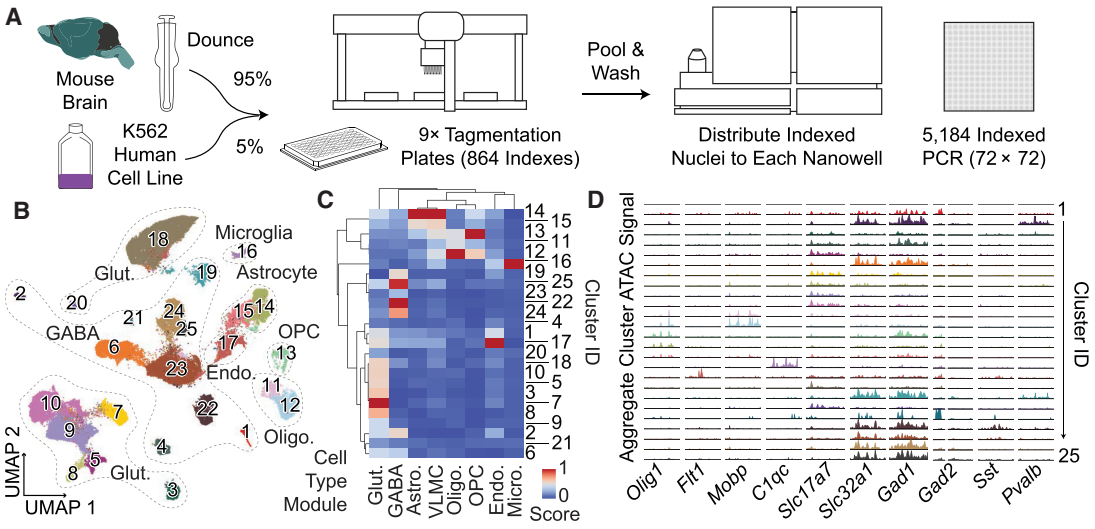

Figure 1. Nanowell-based sciATAC-seq on mouse brain. (*A*) Experimental setup and workflow schematic using liquid handling robotics for high-index tagmentation and nanodispensing into 5184 well chips for PCR-based cell indexing. (*B*) UMAP projection of >420,000 mouse brain single-cell ATAC profiles. (*C*) Cell type module classification of clusters. (*D*) Marker gene accessibility plots for clusters.

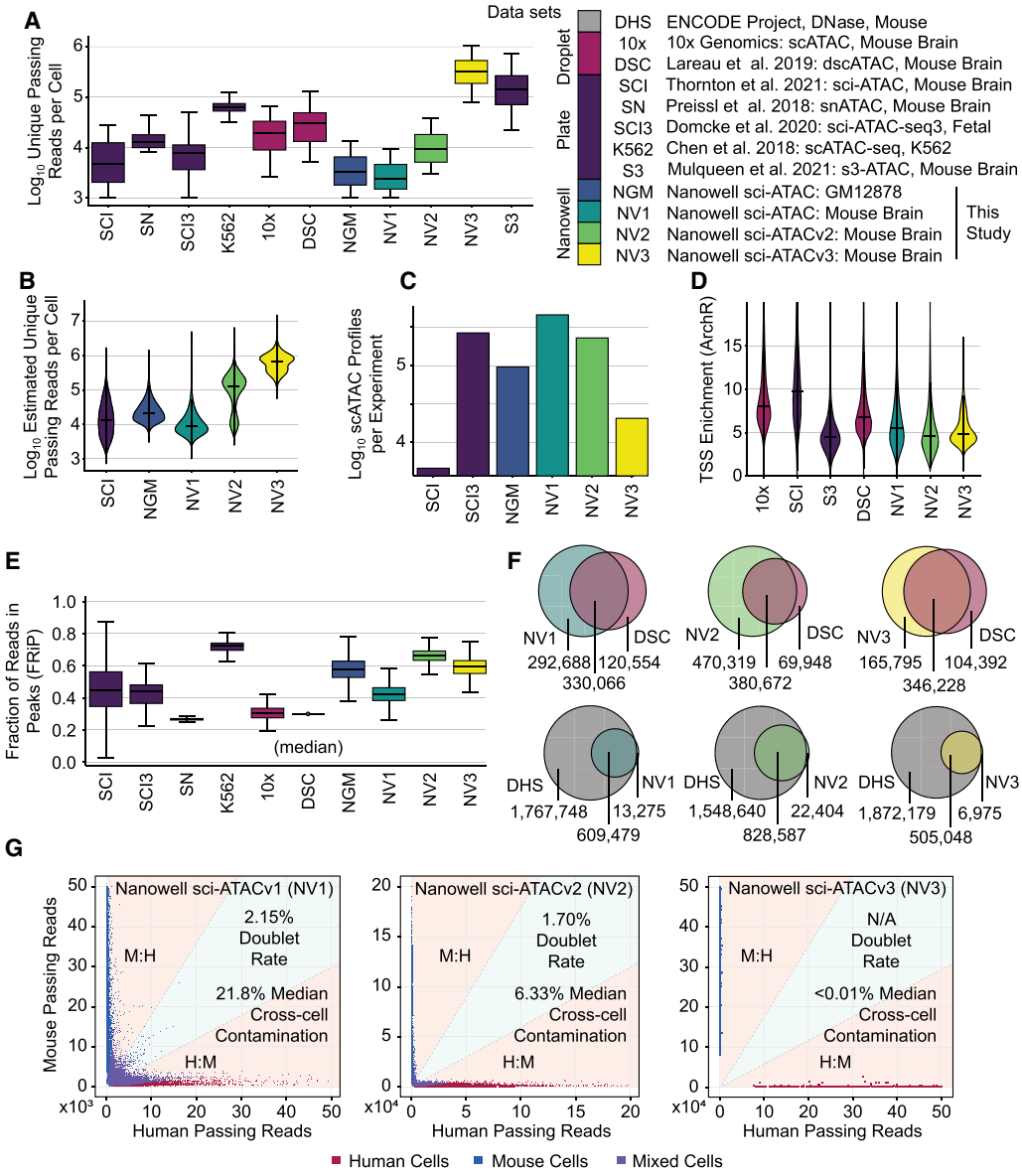

Figure 2. Nanowell-based sciATAC-seq quality metrics. (*A*) Log_{10} unique, passing reads per cell for nanowell sciATAC experiments compared to other single-cell ATAC-seq data sets. Data set legend on *right*. Data from Thornton et al. 2021 are from the nonspatial control experiment. (*B*) Log_{10} projected unique passing reads at 100% sequencing saturation (i.e., total possible unique molecules). (*C*) Log_{10} cell counts produced per preparation. (*D*) TSS Enrichment as determined by ArchR for single-cell ATAC-seq data sets. (*E*) Fraction of reads in called peaks (FRiP) for single-cell ATAC-seq data sets. (*F*) Called peak overlap between mouse brain data sets from this study and dscATAC (*top*) as well as for annotated regulatory elements in mouse (*bottom*). (*G*) Human and mouse aligned read comparisons to assess doublet rates and cross-cell contamination rates for nanowell sciATAC.

with numerous other optimization experiments (Methods; Supplemental Note S1). We then deployed the workflow, sciATACv2, on mouse brain tissue with a 5% spike-in of human K562 cells using 864 tagmentation indexes and the full 5185 nanowells for PCR indexing, producing a total of 224,986 passing sciATAC profiles and observed minimal cross-cell contamination at ~6% (Fig. 2G, middle). Cell profiles exhibited a mean passing read count of 14,311 (Fig. 2A, median = 9360) at a sequencing saturation of only 3.8%, implying additional sequencing would yield far greater read counts before the libraries reach a saturation level that leads to excessive diminishing returns. These data were then used to call 850,991 peaks, with 97.4% overlapping with annotated DHS sites, and with 83.8% of dscATAC peaks overlapping with ours. These peaks resulted in a median FRiP of 0.66 (Fig. 2E), and an aggregate TSS enrichment of 13.72 using ENCODE metrics (4.88 mean, 5.28 median per-cell using ArchR). When compared to other methods (Fig. 2D) the ArchR TSS enrichment is similar to dscATAC and S3 (Lareau et al. 2019, Mulqueen et al. 2021). An assessment of marker genes associated with the clusters as well as gene modules confirmed that the clusters represent the major expected cell types (Fig. 3A) which also corresponded to the expected cell types present within an scRNA-seq reference data set which was generated on the isocortex and hippocampus (Fig. 3B–D; Yao et al. 2021). A number of cell types present in the sciATAC data are not present in the scRNA-seq data set (non-neuronal, accounting for <1% of cells), resulting in ATAC-only clusters. These

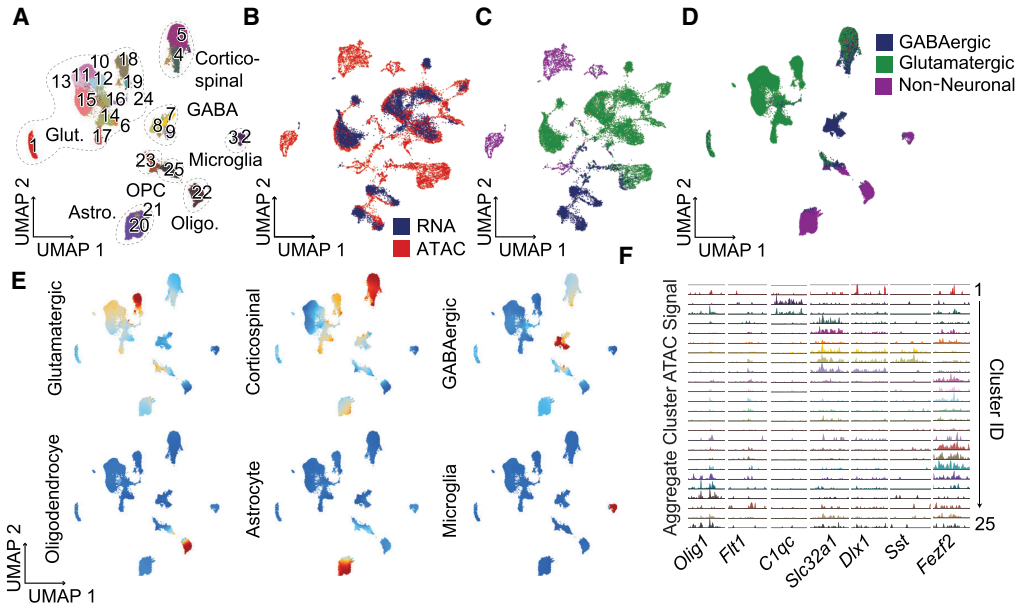

Figure 3. sciATACv2 on whole mouse brain. (*A*) UMAP of 224,986 cells colored by cluster and annotated for cell type class. (*B*) Integration with mouse isocortex and hippocampus scRNA-seq neuron-enriched data. As expected, the whole mouse brain ATAC data set (red) includes cell populations not present in the RNA data set (blue), predominantly for non-neuronal cell types which were depleted in the RNA data set. (*C*) Label transfer of scRNA cell type classes on the integrated UMAP, and shown on the ATAC UMAP (*D*). (*E*) Gene module scores used to confirm RNA assignments and annotate clusters present only in the ATAC data. (*F*) Marker gene tracks used as the final confirmation of cell type assignment.

non-neuronal cell types were instead assigned using marker gene profiles and module scores which produced distinct cell type patterns (Fig. 3E,F; Korsunsky et al. 2019; Granja et al. 2021).

Asymmetrical adapter design produces high-coverage single-cell ATAC data

We reasoned that the improved tagmentation adapter designs used in s3-ATAC to produce high-coverage sciATAC profiles (Mulqueen et al. 2021), may enable improved coverage for our nanowell sciATAC platform (Fig. 4A; sciATACv3). This design includes an index on only one of the two adapter molecules that is immediately adjacent to the transposon recognition sequence as opposed to indexes on both the forward and reverse adapters; it also allows for a shorter adapter sequence that results in an improved tagmentation efficiency. We carried out a preparation using a similar adapter design as s3-ATAC, but without leveraging the full technique that includes single-adapter tagmentation and adapter switching (Fig. 2A). This variant, sciATACv3, was carried out on mouse brain tissue, again including a 5% human cell spike-in (K562) using a single 96-well plate of indexed tagmentation and deposited a target of 15 indexed nuclei into each of the 5184 nanowells after performing our optimized washing protocol. Because of the anticipated complexity increase, we used a single tn5 plate to reduce the total cell count to economize on sequencing. This yielded 21,239 sciATAC profiles (19,781 mouse, 1458 human) with 0% detectable cross talk (Fig. 2G). Although the throughput is lower when compared to previous preparations that utilized more than 96 tagmentation indexes, the mean read count per cell was 413,504 (median = 324,416), equating to a mean fragment count of 141,487 (median = 168,865) at 48% sequencing saturation, achieving a higher coverage than the s3-ATAC technology on the same tissue. The high information content of sciATACv3 makes large-scale experiments with cell counts comparable to sciATACv2 prohibitive if they are to be sequenced to the same depth as the data set presented here, on the order of 1.8×10^{11} sequence reads; however, the improved complexity would enable large experiments sequenced to lower depth to be done so more efficiently (i.e., less PCR duplicates at the same unique fragment count). We therefore focused on this smaller-scale data set to illustrate the unique advantages of the sciATACv3 workflow.

The sciATACv3 data were used to call 512,023 peaks (98.6% overlapping with annotated DHS sites, with 76.3% of dscATAC peaks overlapping with ours), with a TSS enrichment of 13.39 (ENCODE; single-cell mean = 4.89, median = 5.23 via ArchR) and a FRiP of 0.59 (Fig. 2F). This increased efficiency is due to the improved adapter design and optimized pre-PCR processing (similar to s3-ATAC, where distributed nuclei are subjected to a longer incubation prior to PCR to remove the bound Tn5 than can inhibit amplification), coupled with the increased efficiency due to reduced reaction volumes. The high cell coverage and FRiP translate to a mean of 71,891 identified accessible loci per cell (median = 66,867), with the top 10% of cells providing chromatin accessibility information for ≥121,006 peaks. Using these high-coverage data, we identified 25 clusters after topic modeling, again representing the major cell types in the mouse brain (Fig. 4B–D; Supplemental File S1). We then further subclustered the GABAergic neuron populations into 23 subclusters that we first integrated with scRNA-seq data which enabled many clusters to be automatically assigned to subtypes (Fig. 4E,F). The remaining clusters that were not present in the scRNA-seq data set, which was not whole brain, and therefore lacks region-specific subtypes, were assigned using marker gene accessibility to produce 23 clusters that also capture well-defined subtypes (Fig. 5A–D; Supplemental File S1). Finally, we integrated these data directly with data sets using the 10x Genomics platform and dscATAC (Lareau et al. 2019). Most of the resulting clusters were comprised of multiple techniques; however, several were methods-specific for each of the

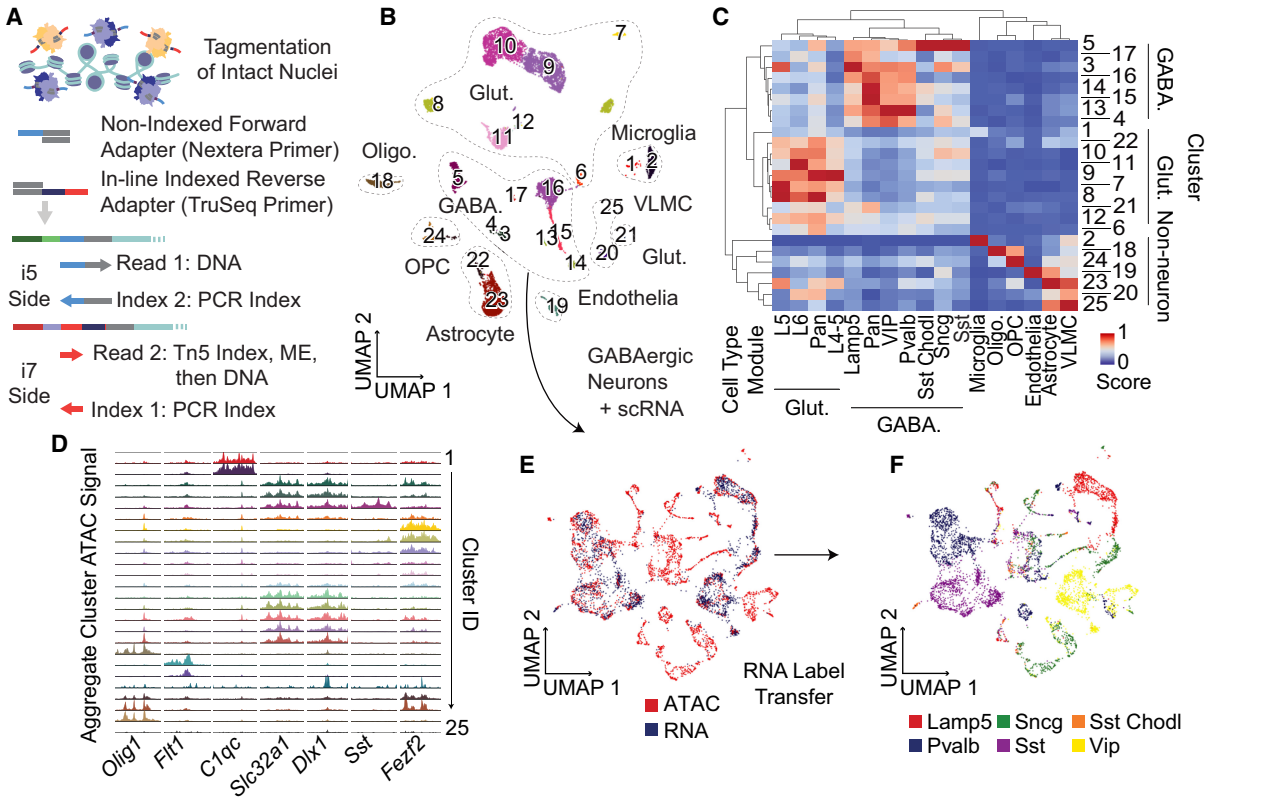

Figure 4. Single-end indexed nanowell sciATAC produces high-coverage profiles. (*A*) Single-end indexed adapter design for nanowell sciATACv3. (*B*) UMAP of 21,239 mouse brain chromatin accessibility profiles produced using nanowell sciATACv3. Dashed line indicates interneuron cell clusters. (*C*) Cell type module scoring for clusters in panel A. (*D*) Marker gene tracks for clusters. (*E*) Integration of GABAergic neurons with scRNA-seq neurons from mouse isocortex and hippocampus. As expected, some clusters are present in whole brain data (ATAC, red) that are not in the RNA data set (blue). (*F*) Label transfer of scRNA cell types used to guide cell type identification. Unmatched clusters were typed using gene set module scores and marker gene profiles.

individual methods, with comparable numbers of method-specific clusters for each comparison (Fig. 5E). When performing clustering on the integrated data set between dscATAC (Lareau et al. 2019) and sciATACv3 data, only 2.3% of cells (n = 579 of 25,713) were present in method-specific clusters (Fig. 5F).

To further explore the benefits of the increased coverage produced by sciATACv3, we first performed a motif analysis on the GABAergic neuron subset, identifying key transcription factor motifs that define the subpopulations (Fig. 6A). We then performed transcription factor footprinting in aggregate over the distinct GABAergic neuron subtypes for each of these motifs and compared profiles to the GABAergic neuron subsets produced from sciATACv1 and v2. Overall, the quality was greatly improved for all factors, including the Vip-specific motif for ARX and Sst-specific motif for TCF12 (Fig. 6B, all motifs: Supplemental File S2). Taken together, although the lower-coverage data from the initial version were able to provide cell type resolution, the improved coverage provided a far greater ability to perform transcription factor footprinting, notably without a significant difference in the FRiP or TSS enrichment of the experiments.

Integration of high-coverage and high-throughput sciATAC GABAergic neuron data sets

We next sought to integrate data sets from all three versions to demonstrate cross–data set cell type and cluster concordance. We isolated all cells classified as GABAergic neurons from all three experiments and performed iterative latent semantic indexing (LSI) followed by batch correction using Harmony (Korsunsky et al. 2019) and visualization using UMAP (Fig. 7A). This produced clear separation of cell groupings comprised of all three experiments. Although some bias remained within each group, cells were largely assigned to the same clusters or to adjacent clusters that were readily merged based on shared cell type markers and module scores (Fig. 7B–E). The resulting subtype clusters produced clear marker gene accessibility profiles, indicating successful integration (Fig. 7F). Integration of the full data sets that include all cell types was attempted; however, the large cell count (>670,000) resulted in substantial computational challenges and iterative LSI could only be performed using a small number of variable features rather than an optimized value, resulting in limited cell type separation. However, there is little added benefit of the entire data set over cell type subsets which is typically the next step in any given analysis workflow.

Discussion

Here, we present a high-throughput platform for profiling single-cell chromatin accessibility that can achieve a cell throughput in the hundreds-of-thousands in addition to a workflow capable of producing high-coverage profiles. The core of the technique is single-cell combinatorial indexed ATAC-seq, or sciATAC, where multiple wells of nuclei are subjected to ATAC-seq library construction

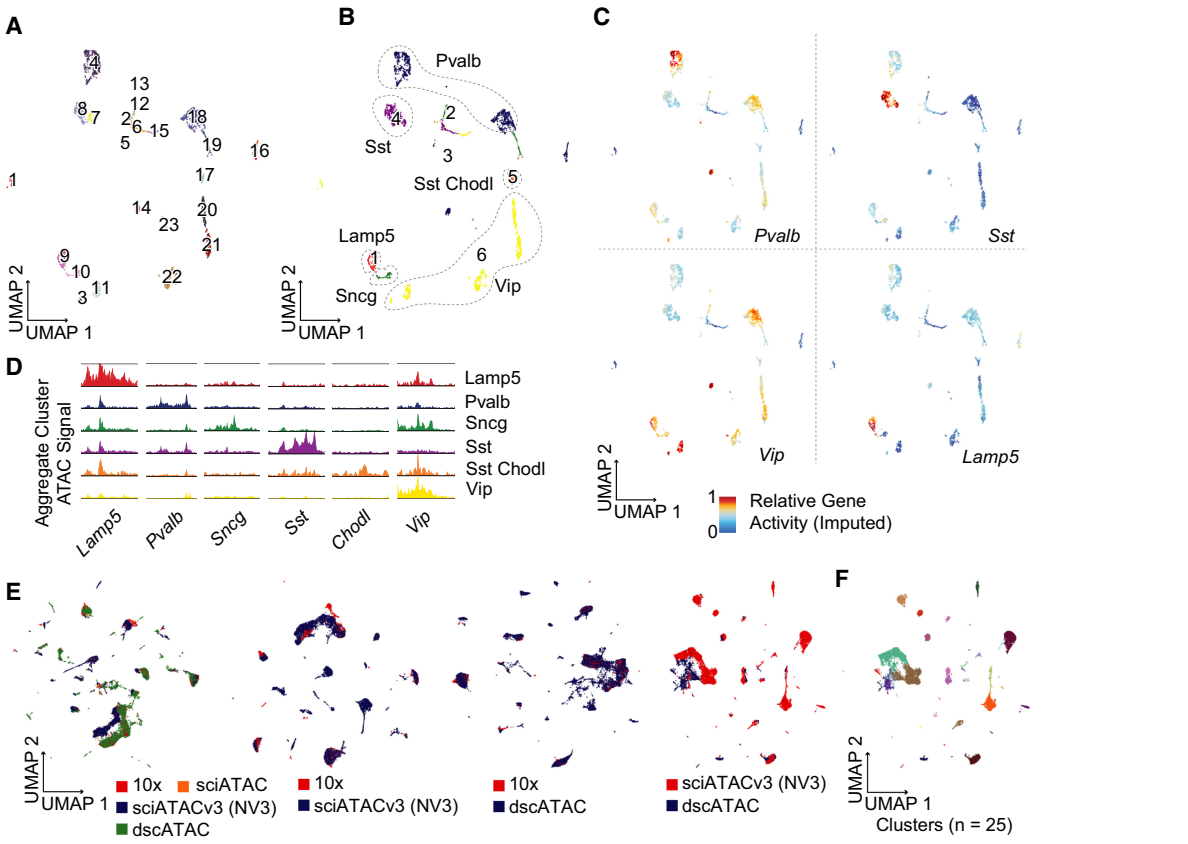

Figure 5. sciATACv3 integration and interneuron subtyping. (*A*) UMAP of GABAergic neurons colored by cluster and by (*B*) subtype. (*C*) Imputed gene activity scores for subtype-defining marker genes, and (*D*) marker gene accessibility tracks. (*E*) Integration with other mouse brain single-cell ATAC-seq data sets as well as paired data set integrations. (*F*) Called clusters on the integrated dscATAC and nanowell sciATACv3 data sets. Despite poor visualization, only 2.3% of cells are in method-specific clusters.

on intact nuclei, leveraging a transposome complex that contains a barcode unique to each well. These nuclei are then pooled and distributed to a secondary set of wells where the probability of two nuclei with the same tagmentation index is low. The secondary wells each contain a unique pair of barcoded PCR primers, thus providing two rounds of cell barcoding: the tagmentation stage and PCR stage; the combination of which forms the full cell barcode. In prior iterations of sciATAC, the secondary PCR indexing was performed on a 96-well plate, requiring relatively large reaction volumes; however, in this study we leverage a 5184 nanowell chip that is processed using the ICELL8 instrument, effectively enabling a 52-fold increase at the PCR indexing stage, while requiring approximately the same PCR reagents as a single 96-well PCR reaction.

We demonstrated the nanowell-based technologies by profiling banked mouse brain tissue and included a human cell line control at a low percentage in order to assess cell cross talk and doublet rates. Our initial workflow produced very high cell counts, though with a high cell-cell cross talk due to the absence of a sorting step that is typical for sciATAC assays that leverage 96-well plates. During tagmentation, a substantial portion of nuclei rupture, releasing tagmented chromatin into solution that is present when all tagmentation wells are pooled. Subsequent spin downs to increase the concentration retain this ambient chromatin which is then disbursed to the subsequent PCR indexing wells. When ambient chromatin has the same index as one of the intact nuclei

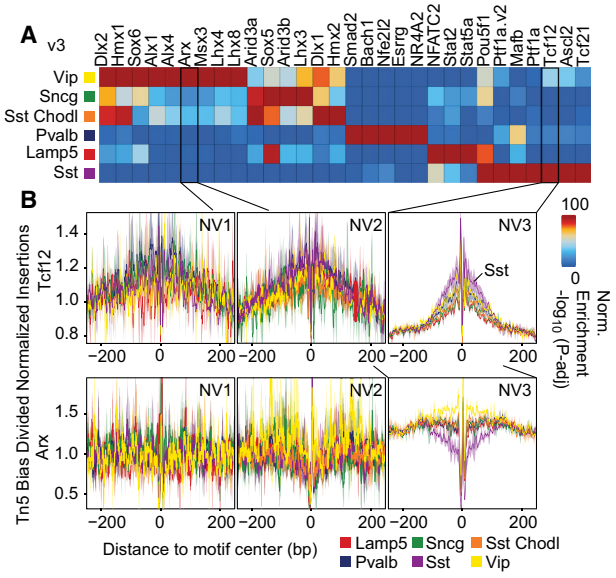

Figure 6. Motif and footprinting analyses and comparison of sciATACv3. (*A*) Motif enrichment analysis for GABAergic neuron subtypes using sciATACv3. (*B*) Transcription factor motif footprinting for select motifs demonstrates cleaner profiles using the higher-coverage sciATACv3. All motifs can be found in Supplemental File S2.

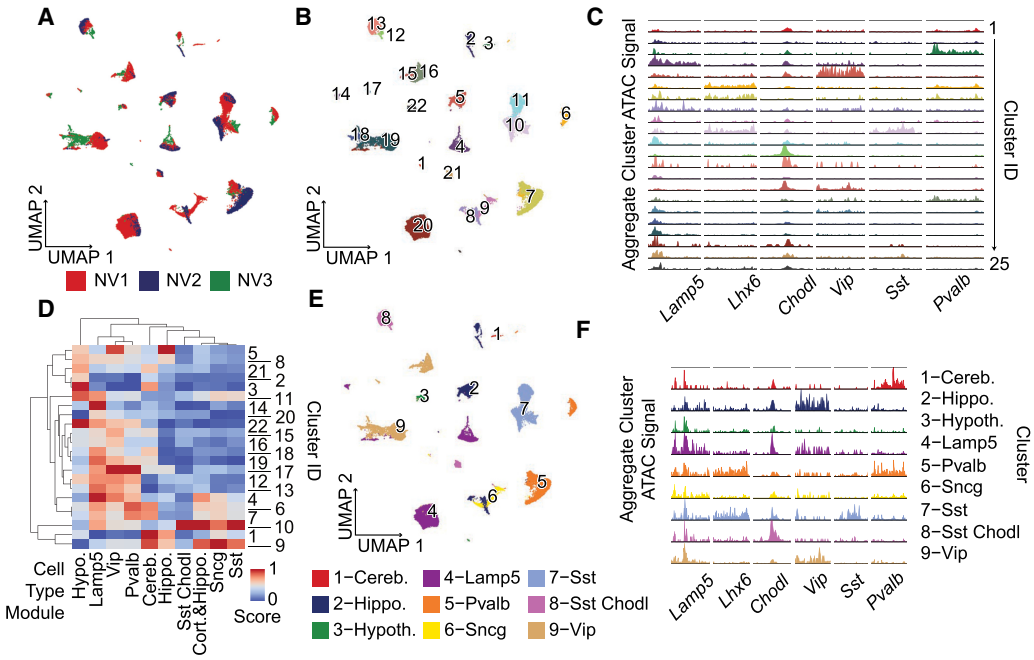

Figure 7. Integration of high-coverage and high-throughput GABAergic sciATAC data sets. (*A*) Integration of GABAergic neurons from all three experiments colored by sciATAC version, and (*B*) cluster. (*C*) Marker gene accessibility profiles for clusters in the combined data set used for subtype identification along with (*D*) module gene scores which include region-specific designations. (*E*) Subtype classifications using both module and marker gene profiles. (*F*) Marker gene accessibility tracks for designated subtypes.

in the well, it becomes associated with that nucleus providing cross talk from other cells. In spite of this, distinct cell type clusters with expected marker gene accessibility profiles were identified. To combat the cross talk, we conducted a series of optimization experiments to remove the ambient chromatin, settling on an improved washing technique that resulted in a high-throughput preparation without the cross talk present using the initial workflow design.

We next explored an alternative transposome design (sciATACv3) that leverages an index on only one of the tagmentation adapters as opposed to both, which has shown to provide higher efficiency of tagmentation due to the shorter adapter lengths. Because this technique requires a unique indexed adapter for each individual well (as opposed to a unique pair as in the prior versions: one forward and one reverse), we proceeded with a single 96-plex indexed tagmentation. This produced a preparation with far higher coverage than other techniques, in line with expectations of the optimal adapter design. We also suspect that the reduced surface area of the nanowells compared to traditional 96-well plates also conveys an advantage with less opportunities for material to stick to the walls of the wells. This preparation also produced nearly undetectable cross talk and only a single doublet was identified out of over 20,000 cell profiles. With this low doublet rate (<0.01% total estimated rate), we could achieve a higher cell throughput (est. ~75,000) while maintaining a low collision rate (est. <2%) by simply loading more onto the chip or by the addition of another 96 indexed tagmentation reactions (est. ~100–150,000). This throughput increase would also further reduce the costs per cell for the sciATACv3 assay, compensating for the added costs associated with the use of commercially obtained transposase complexes as opposed to those produced in-house. The increased coverage also meant increased sequencing depth, which would be a barrier if the throughput reached the 10^5 range, requiring on the order of 180 billion raw sequence reads. However, the increased coverage that is possible with the workflow would allow for more efficient shallower sequencing due to a reduce number of PCR duplicate reads at an equivalent unique read count.

The primary advantage of our technique is the low cost, at $0.0026–$0.0053 (USD) per cell for sciATACv1/v2 and $0.056 per cell for sciATACv3, with the primary cost coming from the nanowell chip itself. These cost estimates include the use of transposase complexes produced in-house; however, recently indexed transposase complex of the same design leveraged in sciATACv3 have become commercially available, increasing accessibility to the technique, though likely increasing costs. Even if the cost is doubled, $0.112 per cell is still very economical when compared to other commercially available options.

Taken together, sciATAC performed using a nanowell chip for PCR-based indexing is a viable means of producing atlas-scale single-cell ATAC-seq libraries at a low cost per cell. The technique is capable of producing very high cell counts, or lower counts with much higher coverage, providing flexibility to balance these parameters to the needs of the study. We demonstrated these methods on frozen mouse brain tissue, achieving comparable ATAC-seq metrics as other assays with respect to open chromatin enrichment (as measured by TSS Enrichment); however, it is important to note that the primary factor for determining open chromatin enrichment is the processing of the nuclei prior to the assay, making the downstream components less relevant for this property. The data sets produced were able to be directly integrated with existing single-cell ATAC-seq data sets as well as single-cell RNA-seq data, confirming cell type and neuronal subtype identities, and validating the capabilities of the assay to assess cell type composition in complex tissues. Finally, the assay is directly amenable to sample multiplexing by leveraging subsets of tagmentation indexes for individual samples, a property of all combinatorial indexing workflows that further contributes to the flexibility of the platform.

Methods

GM12878 and K562 cell culture and nuclei isolation

We prepared GM12878 and K562 nuclei by aliquoting 2 mL of cell culture suspension, centrifuging it at 500 rcf for 5:00 to pellet the cells and exchange the supernatant for cold NIB-H (10 mM HEPES at pH 7.5 [Sigma-Aldrich H4034], 10 mM NaCl [Fisher M-11624], 3 mM $MgCl_2$ [Sigma-Aldrich M8226], 0.1% IGEPAL [v/v; Sigma-Aldrich I8896], 0.1% Tween-20 [v/v, Sigma-Aldrich P7949], and 1× protease inhibitor [Roche 11873580001]). The cells were incubated on ice for 15 min in suspension, before being centrifuged at 500 rcf for 5 min, and washed once with 1 mL NIB-H before being resuspended in cold NIB-H and counted on a hemocytometer.

Preparing Takara ICELL8 for sciATACv1

We ran the following custom protocol on the ICELL8: Distribute 35 nL i5 PCR primer (72 barcodes) in 1.7× SDS Tn5 denaturing buffer (Tn5Dn: 15 uM primer, 0.08235% SDS, 5% BSA); seal and spin down chip; 35 nL total. Distribute 35 nL i7 PCR primer (72 barcodes) in 1.7× Tn5Dn; seal and spin down chip; 70 nL total. After tagmenting and washing nuclei, distribute 50 nL nuclei; seal and spin down chip; 120 nL total. Nuclei counts loaded for each experiment and the total output and efficiency can be found in Supplemental Table S1. Incubate at 55°C for 15 min. Set the chip temperature to below RT (10°C–20°C). Distribute 1 × 100 nL dispense of PCR mix (576 µL KAPA 5× High GC Buffer, 57.6 µL 10 µM dNTP mix, 43.2 µL KAPA HiFi DNA polymerase, 938 µL dH_2O) using three filter files. The filter files were necessary to dispense 100 nL across the entire chip; we found that two 50 nL dispenses worked equally well without the additional complexity of filter files. Consequently, all further experiments were performed without the need for filter files. Next is to seal and spin down the chip; the current chip volume is 220 nL total. Finally, distribute the remaining PCR master mix as a 50 nL dispense across the entire chip, without a filter file, then seal and spin down the chip (the final volume is 270 nL). Incubate at 72°C for 10 min, then PCR amplify (see Supplemental Table S2 for ICELL8 dispensing parameters and Supplemental Table S3 for ICELL8-specific cycling conditions).

Mouse brain sciATAC with GM12878 spike-in

Mouse brain samples were sourced from residual samples collected by Thornton et al. (2021). Briefly, male C57Bl/6J mice, aged 8 wk were sacrificed by carbon dioxide primary euthanasia and cervical dislocation secondary euthanasia. Following euthanasia, the mice were immediately decapitated and intact brain tissue was harvested. Harvested tissue was washed in ice-cold phosphate-buffered saline (PBS; pH 7.4) and submerged in TFM (General Data Company TFM-C) within a disposable embedding mold (Electron Microscopy Sciences [EMS] 70183). The embedded brains were flash-frozen in liquid nitrogen cooled isopentane by lowering the sample into the isopentane bath, without submerging, within 5 min of embedding. Samples were immediately transferred to dry ice, paraffin wrapped to delay sample dehydration, and stored in an airtight container at −80°C. We sampled the brain by laterally sectioning the embedded brain behind the olfactory bulb to obtain a coronal section. The brain nuclei were then isolated as described previously by Thornton et al., by adding NIB-H (10 mM HEPES, pH 7.5 [Sigma-Aldrich H4034], 10 mM NaCl [Fisher M-11624], 3 mM $MgCl_2$ [Sigma-Aldrich M8226], 0.1% IGEPAL [v/v; Sigma-Aldrich I8896], 0.1% Tween-20 [v/v, Sigma-Aldrich P7949], and 1× protease inhibitor [Roche 11873580001]) to the section, and homogenizing with a Dounce homogenizer, loose pestle, seven strokes, incubating 10 min on ice, and repeating the homogenization with the tight pestle, seven strokes. The lysate was poured over a 40 µM cell strainer, and the nuclei were then counted. Concurrently, we prepared GM12878 nuclei as described above. Nuclei were then diluted in fresh TAPS-TD (33 mM TAPS pH = 8.5 [Sigma-Aldrich T5130], 66 mM KOAc [Sigma-Aldrich P1190], 10 mM MgOAc [Sigma-Aldrich M5661], 16% DMF [Sigma-Aldrich D4551]) 150 nuclei per µL and transferred to 9 × 96-well plates. One-half of one plate was reserved for the GM12878 spike in, the remainder were mouse brain nuclei. We then tagmented with 1.5 µL of 8 µM combinatorially barcoded transposase, incubating for 15 min at 55°C. Following this, the tagmented nuclei were pooled on ice, washed twice with NIB-H (centrifuging 2 × 5:00 at 500 rcf with a 180° rotation of the tube between centrifugations) and then counted. The whole sample (about 585,000 nuclei) was loaded on a prepared ICELL8 chip as described above, with the exception of using 3 × 50 nL dispensing steps when adding PCR master mix (see Supplemental Table S2 for ICELL8 dispensing parameters). After PCR (see Supplemental Table S3), the library was removed from the ICELL8 chip by centrifugation at 3750 rcf for 10:00, and then purified using SPRI beads at a 1:1 ratio. The library was quantified on an Agilent TapeStation using d1000 reagents, then sequenced on Illumina NovaSeq (two runs).

Optimization of post-tagmentation cleanup using modified taguchi methods

Experiments were designed to employ Taguchi's orthogonal arrays for the sake of efficiency that are detailed in Supplemental Note S1. We settled on centrifugation with a cushioning buffer composed of 1× TMG wash buffer (ScaleBio), centrifuging for 10 min at 500 rcf, followed by additional centrifugation if necessary to pellet the nuclei, then resuspending the nuclei in TMG again, and centrifuging one more time. Finally, nuclei are resuspended in TMG with 0.1% Pluronic F-127.

Mouse brain sciATACv2 with K562 spike-in

Mouse brain samples were prepared, and nuclei isolated as above. K562 nuclei were prepared by centrifuging to remove media, and resuspending in 1 mL NIB-H on ice for 20 min, followed by centrifuging at 550 rcf for 5 min, then washing with 1 mL cold NIB-H and re-pelleting, before resuspending at 500 nuclei per µL in 1× TAPS-TD diluted with NIB-H. Both mouse brain and K562 nuclei were tagmented using a total of nine, 96-well plates of combinatorically barcoded transposase in a final volume of 11.5 µL, with 48 wells containing K562 nuclei. Tagmentation was performed at 55°C for 15 min. All nuclei were then pooled on ice, and washed 2× with TMG wash buffer as previously described above. Nuclei were then quantified on a Bio-Rad TC-20, and we estimated a total of approximately 500,000. The whole sample was used for loading to account for dead volume in the ICELL8 system. Approximately 260,000 nuclei were loaded on an ICELL8 chip preloaded with 72 each, barcoded Nextera i5 and i7 PCR primers. PCR was performed as for the previous experiments, with the exception of increasing the number of cycles by 2 to account for the lower nuclei density. Post PCR, we removed the library from the chip by centrifugation at 3750 rcf for 10:00, and then purified using SPRI beads at a 1:1 ratio. The library was then sequenced on a NextSeq 500 High output run to check the quality of the library, then sequenced on Illumina NovaSeq (2 runs).

Mouse brain sciATACv3 with K562 spike-in

Mouse brain samples were prepared, and nuclei isolated as above. K562 nuclei were prepared by centrifuging to remove media, and

resuspending in 1 mL NIB-H on ice for 20 min, followed by centrifuging at 550 rcf for 5 min, then washing with 1 mL cold NIB-H and repelleting, before resuspending at [800] nuclei per μL in 1× TAPS-TD diluted with NIB-H. Both the mouse brain nuclei and the K562 nuclei were then tagmented in a 96-well PCR plate of sciATACv3 tn5 complexes (ScaleBio), where the K562 nuclei occupied column 12, rows 1–8. Additionally, the tagmentation buffer had 10 mM D-Glucosamine added, as we found this addition improves nuclei survival during tagmentation. Tagmentation was performed at 55°C for 17 min, after which nuclei were placed on ice, pooled, then washed 2× with TMG wash buffer as previously described above. Nuclei were loaded on an ICELL8 chip preloaded with 72 Nextera i5 PCR primers and 72 TrueSeq i7 PCR primers. After PCR (see Supplemental Table S3), the library was removed from the ICELL8 chip by centrifugation, and then purified using SPRI beads at a 1:1 ratio. Subsequently, we performed a two-sided SPRI size selection to remove library molecules >1000 bp, as this causes suboptimal clustering on the NovaSeq platform. Libraries were quantified post PCR purification and presequencing with Qubit fluorometer and Agilent TapeStation d1000. Sequencing was performed with standard Illumina chemistry, 2 × 75 bp on Illumina NextSeq 500 and NovaSeq platforms.

Sequencing data preprocessing

The data were prepared using the previously described *scitools* package (Sinnamon et al. 2019). Briefly, the sequencing reads were demultiplexed, and the read names were replaced with the cell barcode and a unique identifier. We then mapped the reads to the human (hg38, GCA_000001405.2), mouse (mm10, GCA_000001635.5), and/or hybrid references (hg38 combined with mm10) using BWA-MEM (Li and Durbin 2009). The resulting BAM files were subjected to various quality controls (see Results; Supplemental File S3), and filtered to remove duplicated reads. Additionally, we removed all barcodes with less than a specified number of reads (see Supplemental Table S1).

Cross talk was calculated by mapping data to a hybrid human/mouse genome. For each cell barcode passing quality/read count filters, we plotted the mouse and human read counts, removed the middle 1/3 of the range (i.e., 1:2 mouse:human to 2:1 mouse:human) as putative doublets, and then calculated median human reads in mouse cells and mouse reads in human cells. The cross talk rate reported is the sum of those two values:

$$\widetilde{CR} = (\tilde{R}_{\text{Mouse,R:[0,0.33]}} \in Cells_{\text{Human}}) + (\tilde{R}_{\text{Human,R:[0,0.33]}} \in Cells_{\text{Mouse}}).$$

Cell type identification with ArchR

We used the ArchR package for additional cell type identification (Korsunsky et al. 2019). Arrow file generation was performed with ArchR version 1.01, as were most analyses. We loaded the duplicate removed, filtered sequences to create Arrow files, using TSSParams = list(window = 101, flank = 200, norm = 10). Subsequently, we processed the data according to the ArchR manual (https://www.archrproject.com/), with the exception of increasing the number of top features used during iterative LSI to between 600,000 and 1,200,000 (for sciATACv3 and combined data), as we found that the resulting dimensionality-reduced data set performed significantly better when clustering and identifying cell types.

We clustered each sample using default ArchR parameters, with the exception of removing the max clusters constraint. For the combined data set, we ran batch correction with the built-in Harmony function in ArchR. For gross cell type identification (i.e., GABA/Glutamatergic neurons vs. non-neuronal cell types) we used a combination of scRNA-seq integration, cross checking with ArchR's GeneScoreModule function, which scores clusters based on aggregate marker gene accessibility. The scRNA-seq was sourced from the Allen Mouse Brain Atlas, with gene lists primarily based on the Allen Brain Atlas for mouse, as well as the Linnarsson Adolescent mouse brain atlas (Arlotta et al. 2005; Zeisel et al. 2018; Tiklová et al. 2019; Samata et al. 2020) (http://hippocampome.org/php/markers.php). Due to the size of the data sets, we randomly down-sampled the scRNA-seq to 10,000 cell profiles to prevent ArchR from crashing during integration. To cross check the scRNA-seq integration results, we used a custom Python script to plot heatmaps of gene scores versus clusters. Additionally, we plotted gene accessibility tracks for various marker genes using ArchR. We found that the scRNA-seq integration with ArchR was generally good for this high-level analysis, when compared to the marker gene results. For the clusters that are represented in the ATAC data and not in the RNA data, we found distinct marker gene profiles as well as gene activities of canonical markers, demonstrating that they are real cells and not technical artifacts. Notably, there are no clusters in our analysis that do not have clear cell type identification by at least two out of five components between (1) marker gene accessibility, (2) integration with RNA, (3) gene module scores, (4) gene activity scores, or (5) integration with other ATAC data. However, we found that our data were sufficiently large enough to be computationally constrained with regards to peak-calling, differential analysis, etc., so we instead chose to subset each of our data sets into three broad types (GABAergic neurons, Glutamatergic neurons, and non-neuronal cells) for further analyses.

For each cell type–specific subset, we reran dimensionality reduction (using the same varFeatures parameter for the subset as we used for the parent library). We then reran clustering and, if necessary, Harmony batch correction where applicable. Next, we plotted UMAP embeddings for each subset. We ran an additional layer of scRNA-seq integration, constraining the scRNA-seq data to the same gross cell type. Of note was our finding that we had additional cell types in our scATAC-seq data when compared to our scRNA-seq reference, specifically subcortical neuron types. We confirmed this using marker genes for various noncortical cell types.

For the GABAergic neuron subsets, we also called peaks using ArchR's built-in, Tile-matrix-based peak caller. These peak sets allowed us to look at marker peaks, as well as analyzing motif enrichment. We calculated motif enrichment for each library individually, as well as for the combined data set. Finally, we performed motif footprinting for a number of enriched motifs in our data. Note: We attempted to use ArchR's integrated MACS2 for peak calling, however, there is a known ArchR bug where some installations of ArchR and MACS2 are incompatible for unknown reasons. Note that the same MACS2 install worked fine when run outside of ArchR for our peak-based analyses. Additionally, while we tried calling peaks on the complete, nonsubset libraries, we ran into an issue where the resulting matrix was simply too large for R to handle.

Peak-based analysis

We called peaks with MACS2 and used the results to create a cells-by-peaks matrix (Zhang et al. 2008; Sinnamon et al. 2019). These data were used as input for the cisTopic3 pipeline (Bravo González-Blas et al. 2019). We used the topic-cell distribution to cluster the cells into phenogroups using the Louvain method as previously described (Levine et al. 2015; Sinnamon et al. 2019). Using the per-phenogroup gene accessibility data with neuronal marker genes, we identified putative cell type to phenogroup

Table 1. Peak and cell counts for sciATAC experiments

Library	Peaks	Cells (in subset)
sciATACv1 GABAergic	106,877	275,394
sciATACv2 GABAergic	194,108	42,483
sciATACv3 GABAergic	265,644	3461
Combined v1, 2, and 3 data sets	343,196	288,682

mappings (note, the markers were a subset of the set described below in Archer analysis). This analysis was used to cross check the ArchR results (Table 1).

Software availability

All code and analyses are available as Supplemental Code and at GitHub (https://github.com/adeylab/nanowell_sciATAC).

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE218881.

Competing interest statement

D.P., J.T., and F.J.S. are employees of ScaleBio.

Acknowledgments

We are grateful for helpful suggestions and feedback from other members of the Adey Lab as well as Rachel Fish, Sally Zhang, Maggie Bostic, Mike Young, and Andrew Farmer at TakaraBio USA for initial access to the ICELL8 instrument and support with protocol implementation. This work was funded by RF1MH128842 (National Institutes of Health [NIH]/National Institute of Mental Health), R01DA047237 (NIH/National Institute on Drug Abuse), and R35GM124704 (NIH/National Institute of General Medical Sciences) to A.C.A.

Author contributions: B.L.O. and A.C.A. conceived the study and wrote the manuscript with input from all authors. B.L.O., D.P., J.T., F.J.S., and A.C.A. developed the technical approach and workflow. B.L.O. carried out all experiments and analysis with assistance and input from R.V.N., C.A.T., M.C., S.N.A., A.N., and A.J.F. All authors reviewed and contributed to the manuscript.

References

Arlotta P, Molyneaux BJ, Chen J, Inoue J, Kominami R, MacKlis JD. 2005. Neuronal subtype-specific genes that control corticospinal motor neuron development in vivo. *Neuron* **45:** 207–221. doi:10.1016/j.neuron.2004.12.036

Bravo González-Blas C, Minnoye L, Papasokrati D, Aibar S, Hulselmans G, Christiaens V, Davie K, Wouters J, Aerts S. 2019. cisTopic: *cis*-regulatory topic modeling on single-cell ATAC-seq data. *Nat Methods* **16:** 397–400. doi:10.1038/s41592-019-0367-1

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10:** 1213–1218. doi:10.1038/nmeth.2688

Chen X, Miragaia RJ, Natarajan KN, Teichmann SA. 2018. A rapid and robust method for single cell chromatin accessibility profiling. *Nat Commun* **9:** 5345. doi:10.1038/s41467-018-07771-0

Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J. 2015. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348:** 910–914. doi:10.1126/science.aab1601

Domcke S, Hill AJ, Daza RM, Cao J, O'Day DR, Pliner HA, Aldinger KA, Pokholok D, Zhang F, Milbank JH, et al. 2020. A human cell atlas of fetal chromatin accessibility. *Science* **370:** eaba7612. doi:10.1126/science.aba7612

The ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, et al. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583:** 699–710. doi:10.1038/s41586-020-2493-4

Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, Greenleaf WJ. 2021. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* **53:** 403–411. doi:10.1038/s41588-021-00790-6

Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S. 2019. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* **16:** 1289–1296. doi:10.1038/s41592-019-0619-0

Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, Pokholok D, Aryee MJ, Steemers FJ, Lebofsky R, et al. 2019. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat Biotechnol* **37:** 916–924. doi:10.1038/s41587-019-0147-6

Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, et al. 2015. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162:** 184–197. doi:10.1016/j.cell.2015.05.047

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760. doi:10.1093/bioinformatics/btp324

Mezger A, Klemm S, Mann I, Brower K, Mir A, Bostick M, Farmer A, Fordyce P, Linnarsson S, Greenleaf W. 2018. High-throughput chromatin accessibility profiling at single-cell resolution. *Nat Commun* **9:** 3647. doi:10.1038/s41467-018-05887-x

Mulqueen RM, Pokholok D, O'Connell BL, Thornton CA, Zhang F, O'Roak BJ, Link J, Yardımcı GG, Sears RC, Steemers FJ, et al. 2021. High-content single-cell combinatorial indexing. *Nat Biotechnol* **39:** 1574–1580. doi:10.1038/s41587-021-00962-z

Preissl S, Fang R, Huang H, Zhao Y, Raviram R, Gorkin DU, Zhang Y, Sos BC, Afzal V, Dickel DE, et al. 2018. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nature Neurosci* **21:** 432–439. doi:10.1038/s41593-018-0079-3

Samata B, Takaichi R, Ishii Y, Fukushima K, Nakagawa H, Ono Y, Takahashi J. 2020. L1CAM is a marker for enriching corticospinal motor neurons in the developing brain. *Front Cell Neurosci* **14:** 31. doi:10.3389/fncel.2020.00031

Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, Olsen BN, Mumbach MR, Pierce SE, Corces MR, et al. 2019. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37:** 925–936. doi:10.1038/s41587-019-0206-z

Sinnamon JR, Torkenczy KA, Linhoff MW, Vitak SA, Mulqueen RM, Pliner HA, Trapnell C, Steemers FJ, Mandel G, Adey AC. 2019. The accessible chromatin landscape of the murine hippocampus at single-cell resolution. *Genome Res* **29:** 857–869. doi:10.1101/gr.243725.118

Thornton CA, Mulqueen RM, Torkenczy KA, Nishida A, Lowenstein EG, Fields AJ, Steemers FJ, Zhang W, McConnell HL, Woltjer RL, et al. 2021. Spatially mapped single-cell chromatin accessibility. *Nat Commun* **12:** 1274. doi:10.1038/s41467-021-21515-7

Tiklová K, Björklund ÅK, Lahti L, Fiorenzano A, Nolbrant S, Gillberg L, Volakakis N, Yokota C, Hilscher MM, Hauling T, et al. 2019. Single-cell RNA sequencing reveals midbrain dopamine neuron diversity emerging during mouse brain development. *Nat Commun* **10:** 581. doi:10.1038/s41467-019-08453-1

Yao Z, van Velthoven CTJ, Nguyen TN, Goldy J, Sedeno-Cortes AE, Baftizadeh F, Bertagnolli D, Casper T, Chiang M, Crichton K, et al. 2021. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell* **184:** 3222–3241.e26. doi:10.1016/j.cell.2021.04.021

Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, Häring M, Braun E, Borm LE, La Manno G, et al. 2018. Molecular architecture of the mouse nervous system. *Cell* **174:** 999–1014.e22. doi:10.1016/j.cell.2018.06.021

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9:** R137. doi:10.1186/gb-2008-9-9-r137

Received January 31, 2022; accepted in revised form December 8, 2022.