nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-021-00962-z

Check for updates

High-content single-cell combinatorial indexing

Ryan M. Mulqueen[1], Dmitry Pokholok[2], Brendan L. O'Connell[1], Casey A. Thornton[1], Fan Zhang[2], Brian J. O'Roak[1], Jason Link[1,3,4], Galip Gürkan Yardımcı[3,5], Rosalie C. Sears[1,3,4,5], Frank J. Steemers[2] and Andrew C. Adey[1,3,5,6,7] ✉

Single-cell combinatorial indexing (sci) with transposase-based library construction increases the throughput of single-cell genomics assays but produces sparse coverage in terms of usable reads per cell. We develop symmetrical strand sci ('s3'), a uracil-based adapter switching approach that improves the rate of conversion of source DNA into viable sequencing library fragments following tagmentation. We apply this chemistry to assay chromatin accessibility (s3-assay for transposase-accessible chromatin, s3-ATAC) in human cortical and mouse whole-brain tissues, with mouse datasets demonstrating a six- to 13-fold improvement in usable reads per cell compared with other available methods. Application of s3 to single-cell whole-genome sequencing (s3-WGS) and to whole-genome plus chromatin conformation (s3-GCC) yields 148- and 14.8-fold improvements, respectively, in usable reads per cell compared with sci-DNA-sequencing and sci-HiC. We show that s3-WGS and s3-GCC resolve subclonal genomic alterations in patient-derived pancreatic cancer cell lines. We expect that the s3 platform will be compatible with other transposase-based techniques, including sci-MET or CUT&Tag.

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 use single-cell combinatorial indexing (sci)[1], leveraging transposase-based library construction[2] 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 an adapter-switching strategy, 's3', capable of producing one-to-two order-of-magnitude improvements in usable reads obtained per cell for chromatin accessibility (s3-assay for transposase-accessible chromatin, 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 with high-throughput sequencing (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 (PDCLs) from a primary pancreatic tumor.

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 adapters, in the form of forward or reverse primer sequences, are incorporated at each end of the molecule. This occurs only 50% of the time (Fig. 1a). To combat this inefficiency, strategies including the use of larger complements of adapter species[3], incorporation of T7 promoters to enable amplification via in vitro transcription[4–6], or reverse adapter introduction through targeted[7] or random priming[8] or ligation[9] have been developed; however, these methods are often complex and result in limited efficiency improvements. Here we present a strategy of adapter replacement to produce library molecules tagged with both forward and reverse adapters for top and bottom strands. In addition to overcoming the 50% yield limitation, the efficiency of opposite adapter incorporation is also improved when compared to standard tagmentation. This is due to the use of multiple rounds of extension as opposed to a single extension before PCR. This format permits the use of a DNA index sequence embedded within the transposase adapter 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[1,8,10].

Our strategy, symmetrical strand sci (s3; Fig. 1b) 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 adapter 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 adapter sequence. A second template oligo is then introduced that contains a 3′-blocked (inverted dT) locked nucleic acid (LNA) mosaic end reverse complement sequence with a reverse adapter sequence 5′ overhang. This oligo 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 adapter. This results in all library fragments having both a forward and reverse adapter sequence. The LNA-templated extension is carried out over multiple rounds of thermocycling to ensure maximum efficiency of reverse adapter incorporation, which provides an additional improvement over traditional tagmentation workflows where only a single pre-PCR extension is possible. Furthermore, adapter sequences are designed

[1]Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR, USA. [2]Scale Bio, Berkeley, CA, USA. [3]Knight Cancer Institute, Oregon Health and Science University, Portland, OR, USA. [4]Brendan Colson Center for Pancreatic Care, Oregon Health and Science University, Portland, OR, USA. [5]Cancer Early Detection Advanced Research Center, Oregon Health and Science University, Portland, OR, USA. [6]Department of Oncological Sciences, Oregon Health and Science University, Portland, OR, USA. [7]Knight Cardiovascular Institute, Oregon Health and Science University, Portland, OR, USA. ✉e-mail: adey@ohsu.edu

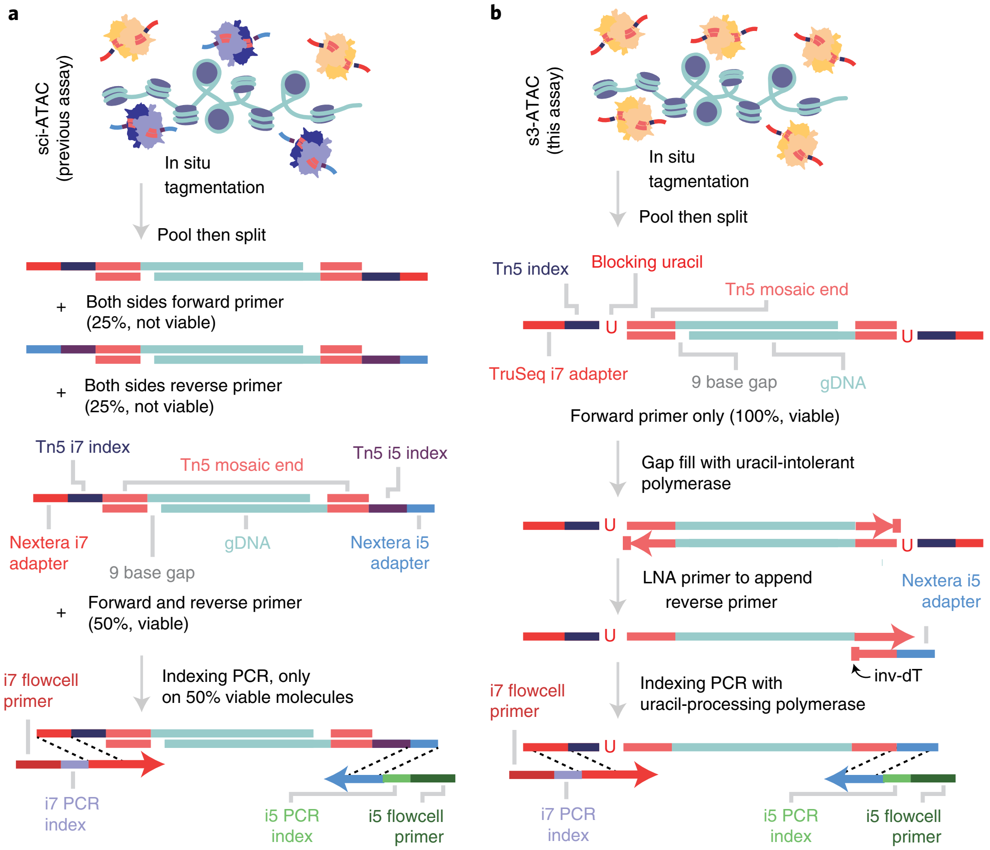

Fig. 1 | Symmetrical strand single-cell combinatorial indexing ATAC-seq (s3-ATAC) improves molecular capture rate. a, Schematic of standard sci-ATAC library construction. **b**, Schematic of s3-ATAC library construction with intermediate steps of adapter switching leading to increased genomic molecule capture rate.

such that standard sequencing recipes can be used instead of the custom workflows and primers that are required for current indexed transposition technologies (Supplementary Tables 1–5)[11,12], making use of the TruSeq read 2 and index read 1 sequencing primer and the standard Nextera read 1 and index read 2 primer.

Results

s3-ATAC creates high-content chromatin profiles. 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 adapter-switching s3 workflow (Fig. 1b). 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 (Fig. 2a). 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 before any processing, that is pretagmentation, we observed a total estimated collision rate of 5.53% (Fig. 2b,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 posttagmentation 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) across four PCR indexing plates (Fig. 2a). We then assessed the total unique sequence reads obtained per cell as a function of the total aligned reads, that is, 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[10]. For this plate, 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, statistics on all plates can be found in Supplementary Table 6). 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 (Supplementary Table 6). When compared to other single-cell ATAC-seq datasets performed on mouse whole-brain tissue, our mouse s3-ATAC libraries contain substantially greater properly paired, unique, nuclear reads per cell with 13.7-, 6.02- and 6.22-fold improvement compared to single-nucleus ATAC-seq (snATAC), 10x Genomics single-cell ATAC (scATAC) and droplet single-cell ATAC-seq (dscATAC), respectively (Fig. 2d and Supplementary Table 7)[13–15]. 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 (Fig. 2e)[16]. We next calculated transcription start site (TSS) enrichment using the approach defined by the ENCODE project (Methods). This produced notable enrichment for both species

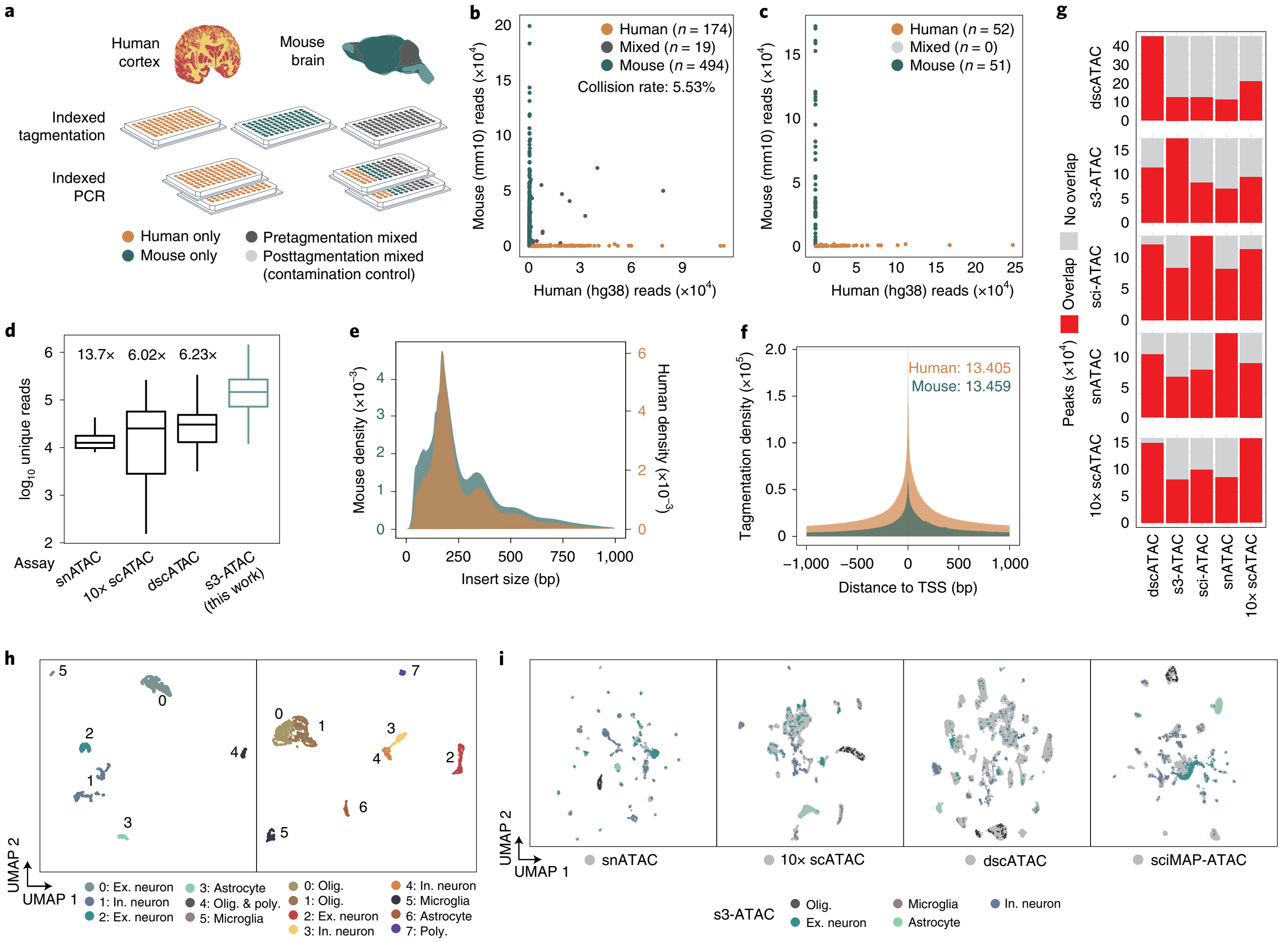

Fig. 2 | s3-ATAC on human cortex and mouse whole brain. a, Experimental flow through and plate layout for the mixed-species experiment, including tagmentation and PCR plate conditions per well. **b**,**c**, Scatter plots of single-cell libraries with counts of unique reads aligned to mouse or human chromosomes in a chimeric reference genome. Points are colored to reflect species assignment (Methods) in both pretagmentation mixing (**b**) and posttagmentation mixing (**c**). **d**, Comparison of unique read counts per cell, restricted to only properly paired reads for s3-ATAC mouse whole-brain sampled cells to unique read counts produced for previously reported datasets ($n = 3{,}034$, 4,117, 46,653 and 298 cells, for snATAC, 10× scATAC, dscATAC and ours, respectively). All comparisons to our data are significantly less (Welch's two-sample *t*-test, *P* values $<5.7\times10^{-42}$, 3.1×10^{-40}, 1.6×10^{-37}, respectively). Fold improvement of our library complexity per method is listed above the method[13–15]. Boxplot represents median and center quartiles with whiskers at the tenth and 90th percentiles. **e**, Insert size distribution of human and mouse libraries reflect nucleosome banding. **f**, Enrichment of reads at TSS for human and mouse libraries with enrichment calculation following ENCODE standard practices. **g**, Stacked bar plot for comparison of peak overlaps across mouse whole-brain datasets. Each row is for a different peak set, with each column showing peak overlap in red. **h**, UMAP projection of mouse whole-brain cell samples ($n = 837$ cells) colored by cluster and cell type assignment (left). UMAP projection human cortex cell samples ($n = 2{,}175$ cells; right). **i**, Integration of s3-ATAC mouse data with other datasets. Points colored by cell type (respective of **h**) or external dataset (gray).

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, Fig. 2f). 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 is 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 (Supplementary Information and Methods). We further compared peak overlap between the matched datasets to assess any systematic bias in s3-ATAC, with all assays performing comparably with respect to the proportion of overlapping reads for each other assay (Fig. 2g).

s3-ATAC resolves cell types in the mammalian brain. 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[17], which we then visualized using uniform manifold approximation and projection (UMAP)[18], performed graph-based clustering at the topic level and processed via Signac[19]. Clear separation of cell types was observed using marker gene signal and differential accessibility profiles (Fig. 1h, Supplementary Information and Supplementary Fig. 1a)[15,20]. Finally, we assessed any systematic bias that may affect biological interpretation by integrating our datasets with snATAC, 10X Genomics scATAC, dscATAC and sciMAP-ATAC (Fig. 2i)[13–15,21]. We observed that our libraries readily integrate across platforms, maintaining cell type discrimination between clusters.

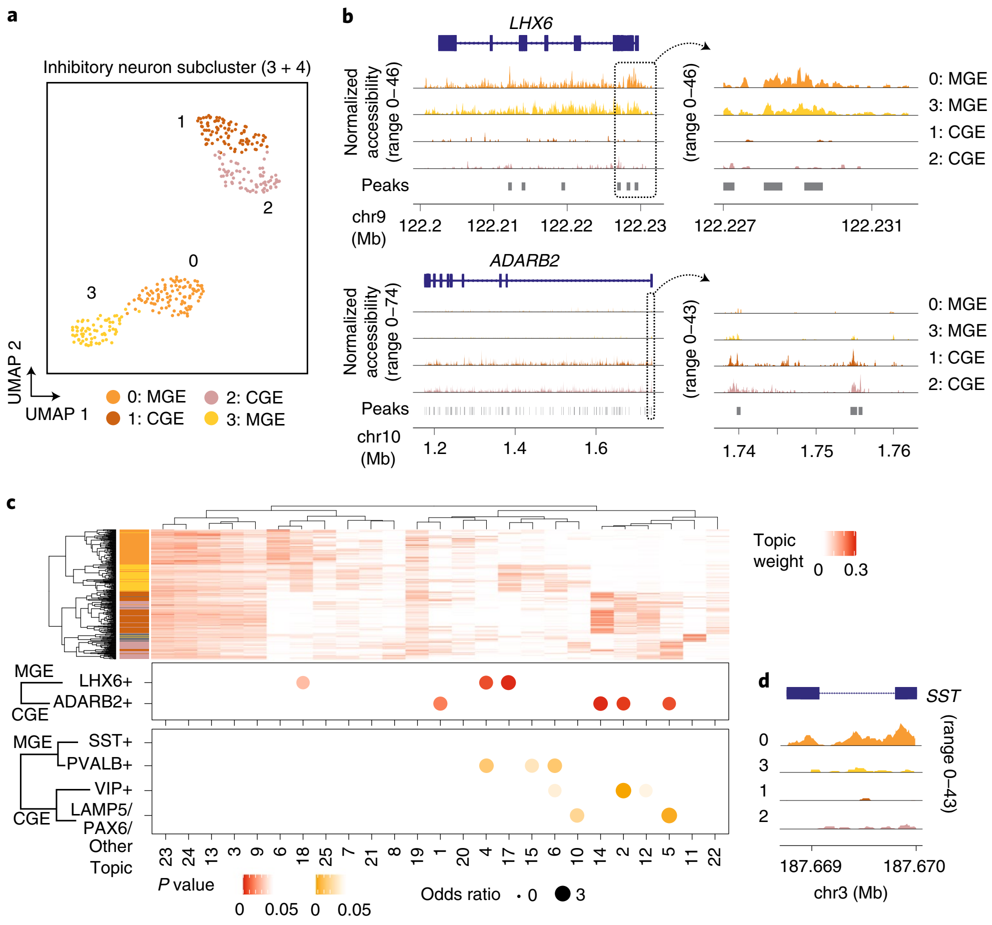

Fig. 3 | s3-ATAC on human cortex inhibitory neurons. a, Subclustering and UMAP projection of human cortical inhibitory neurons (clusters 3 and 4 from **h**, $n = 342$). **b**, Genome coverage track of human inhibitory neurons ($n = 342$) aggregated over four subclusters for genomic locations overlapping MGE and CGE marker genes *LHX6* and *ADARB2*, respectively, with a zoomed-in view of the promoter region (on right). **c**, 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, Methods), with a genome track of *SST* to delineate MGE cell types (right).

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 (Fig. 3a)[22]. 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). From these, we identified 157 GAD1$^+$, ADARB2 + CGE cells and 168 GAD1$^+$, LHX6 + MGE cells (Fig. 3b). We identified 17 cells (subcluster 4) as putative doublets given the coexpression of both LHX6 and ADARB2 and excluded them from subsequent 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 (Supplementary Information). We grouped topics based on cell embeddings (Fig. 3c, top) through hierarchical clustering, and observed topic-based enrichment of sites overlapping cell type specific marker genes previously defined through transcriptomics (Fig. 3c, bottom)[20]. This analysis did not identify a specific topic (or topics) associated with MGE/SST$^+$ cells; however, accessibility at *SST* was elevated specifically in cluster 0 (Fig. 3c, right).

To further assess the impact of increased coverage in our s3-ATAC data, we performed a random read downsampling analysis followed by peak calling, topic modeling and cell type discrimination (Methods and Supplementary Fig. 1b). As expected, the increased reads produced a greater number of called peaks (Supplementary Fig. 1c), although peak calling is a function of the total reads and can be improved by either greater cell number or greater coverage per cell. When examining the downsampled data performance on cell type discrimination, major cell types could be discerned with only 15 or 20% of reads used for human and mouse datasets, respectively (Supplementary Fig. 1d). However, as reads increased, cell types separated more cleanly and produced additional granularity (for example, inhibitory neurons). As a final examination of the advantages provided by improved library complexity, we assessed the impact on sequencing depth required to reach a comparable number of unique, passing reads per cell in our s3-ATAC library when compared to sci-ATAC, which produces libraries with a lower overall complexity (Supplementary Fig. 1e). When targeting a total of 10,000 unique reads per cell, the s3-ATAC library resulted in the removal of only 1.2% of reads as PCR duplicates when compared to 27–38% for the lower-complexity preparations. This directly translates to a reduction in sequencing costs of roughly one-third, enabling studies that want to profile larger cell numbers at a lower depth to do so more efficiently.

s3 for whole-genome and chromatin conformation capture. 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[10] that produces s3-WGS and a strategy to

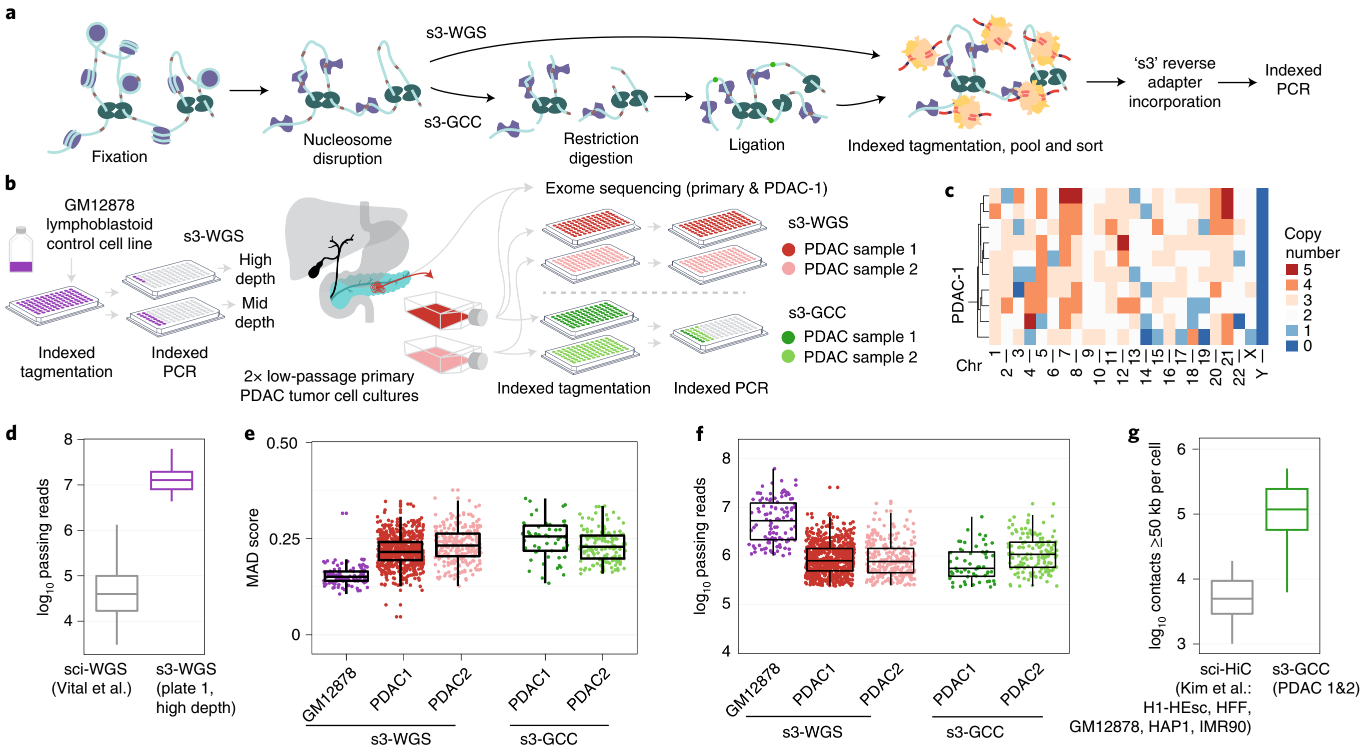

Fig. 4 | s3-WGS and s3-GCC. a, Schematic of sci-WGS and sci-GCC library construction. **b**, Experimental flow through and plate layout for the control GM12878 diploid line (left) and PDAC cell lines (right). **c**, Heatmap summary of chromosome count per cell in PDAC-1 SKY data. Example karyotype of PDAC-1 cell (bottom). **d**, Boxplot of unique read count per cell for the matched GM12878 cell line (n = 3,576, 45 cells for sci-WGS and s3-WGS, respectively)[10]. **e**, Boxplot of MAD score per cell per sample and assay (n = 111, 698, 257, 57 and 145 cells, listed left to right). **f**, Boxplot of reads passing filter per cell. Cell count same as **e**. **g**, Comparison boxplot of s3-GCC and sci-HiC distal contacts (≥50 kb) per cell (n = 2,312, 202 cells for sci-HiC and s3-GCC, respectively)[31]. Boxplots depicts median and center quartiles with tenth and 90th percentile whiskers.

incorporate the core components of HiC library preparation but without ligation junction enrichment to produce s3-GCC (Fig. 4a). Both strategies disrupt nucleosomes to acquire sequence reads uniformly across the genome[10], which we also improved for the s3 assay by optimizing fixation conditions. All experiments were performed using the same number of indexed tagmentation reactions and the same number of nuclei deposited into each PCR indexing well to achieve a comparable expected doublet rate as with the s3-ATAC experiments. We first tested s3-WGS by producing two small-scale libraries on the diploid 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 a high depth (Fig. 4b). 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, Fig. 4d)[10]. This improvement is based on a combination of the s3 workflow, optimization of the nucleosome disruption and the added benefit of thermocycling during s3 adapter switching, which likely improves crosslink reversal before PCR. The second preparation performed comparably, although 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-kilobase (kb) bins, which fell within 0.152 ± 0.025 (mean ± s.d.), comparable to other single-cell genome sequencing techniques (Fig. 4e)[10,23,24].

We performed s3-WGS and s3-GCC on two cell lines derived from a single primary pancreatic ductal adenocarcinoma (PDAC) tumor (Fig. 4b). PDAC is a highly aggressive cancer that typically presents at an advanced stage, making early detection and study of tumor progression key[25]. PDAC studies suffer from a low cancer cell fraction in biopsied samples, thus we used PDCLs maintained at fewer than ten passages. This method allows for multiple modalities of characterization and perturbation, while maintaining a large portion of the heterogeneity present in the tumor sample[26]. We targeted two PDCLs (referred to as PDAC-1 and PDAC-2) derived from the same parent tumor, which had a driver mutation in the oncogene *KRAS* (p.G12D) and profound genomic instability, as indicated by karyotyping (Fig. 4c). 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) for PDAC-1 and -2, respectively (Fig. 4f). MAD scores for the two lines were greater than that of the diploid 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, with a distal contact distribution comparable to bulk HiC datasets (Supplementary Fig 3a). Across both samples, we identified a mean of 118,048 reads per cell that capture genomic contacts at least 50 kb apart from one another, a 14.8-fold improvement over the previous high-throughput single-cell combinatorial indexing technique, sci-HiC[27] (Fig. 4g and Supplementary Information) and comparable to low-throughput scHiC methods that process cells individually[28], with the exception

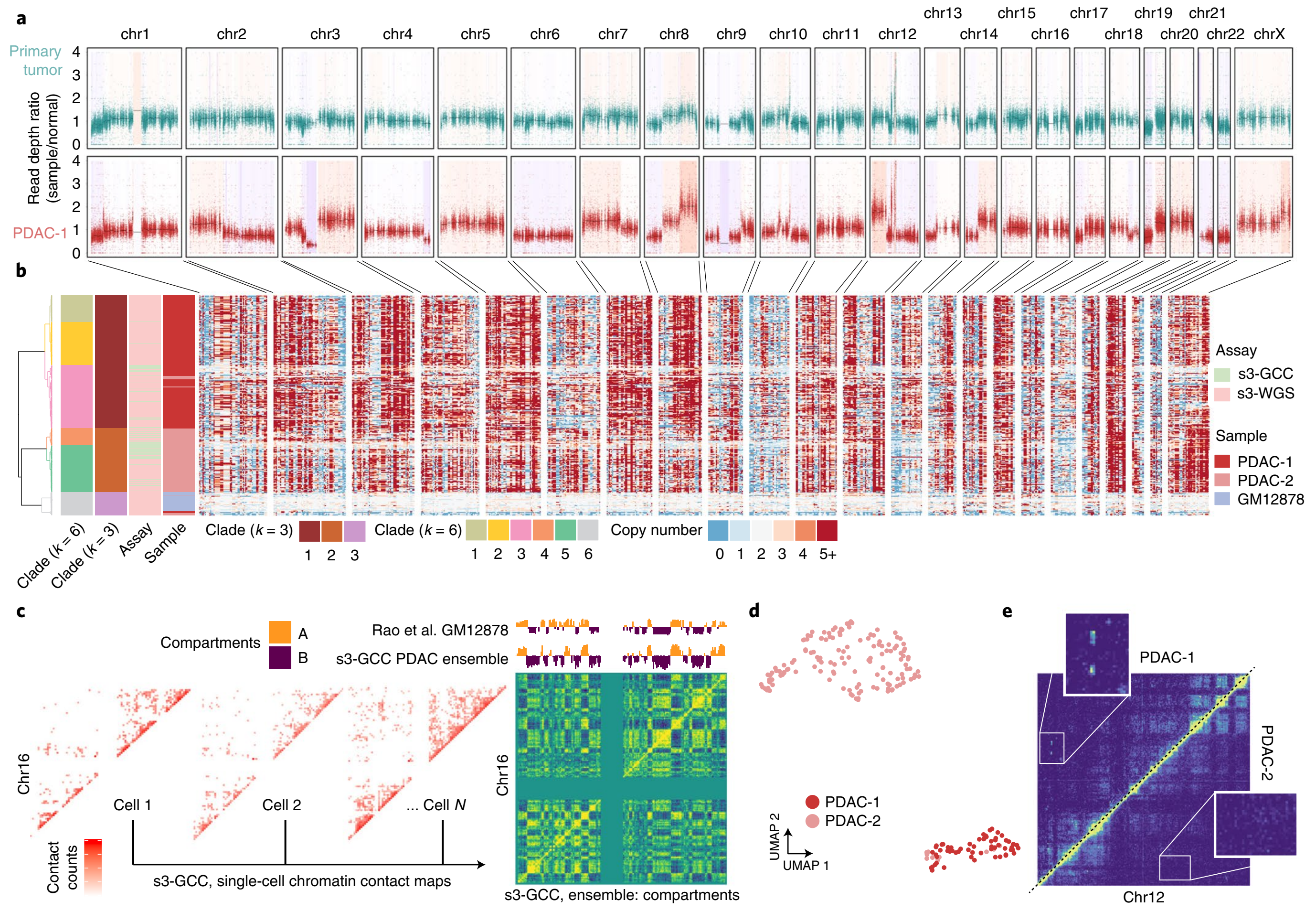

Fig. 5 | s3-WGS for copy number calling and s3-GCC for genome conformation changes. a, WES of primary tumor biopsy and PDAC-1 cell line. Scatterplot of reads per bin with a shading of called copy number variation. **b**, Single-cell whole-genome copy number calling on 500-kb bins genome-wide. Cells (rows) are hierarchically clustered and annotated by assay, sample and assigned clade (left). **c**, Representative single-cell contact maps (raw counts) at 1-Mb resolution for chromosome 16 and ensemble contact map profile at 500-kb resolution with compartment calling. Compartment eigenvectors are plotted above and compared to GM12878 high-depth bulk HiC data from Rao et al.[35]. **d**, UMAP projection of scHiC topic-modeling dimensionality reduction and clustering of single-cell distal contact profiles. **e**, Putative subclonal translocation on chr12 specific to PDAC-1 (top left) when compared to PDAC-2 (bottom right).

of Dip-C, which can achieve exceeding 1×10^6 contact counts[3]. Read pairs spanning ≥50 kb 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 (Supplementary Table 8).

s3-WGS and s3-GCC resolve subclonal alterations in PDCLs. 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 in the lines. 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 on the PDCL line PDAC-1, derived from the tumor (Fig. 5a). 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 less euploid stromal cell contamination. We then processed all single-cell libraries using SCOPE[23], 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-megabase copy number aberrations when performing breakpoint analysis on 500-kb 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 two PDAC lines, 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 local optima ($k=6$). s3-WGS and s3-GCC cells cluster dependent on cell line, reflecting our ability to capture genome-wide copy number data in our s3-GCC libraries (Fig. 5b). We generated pseudobulk 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-kb bin. This revealed multiple fixed and subclonal genomic arrangements (Supplementary Fig. 2a,b). In PDAC-1 and -2 we see shared copy number loss of tumor supressor genes *CDKN2A*, *SMAD4* and *BRCA2* (refs. [25,29]). 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, tumor node metastasis rate[30]. 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 used the HiC-like component of our s3-GCC libraries. Using read pairs spanning ≥50 kb, we produced chromatin contact maps that produced clear chromatin compartmentalization signal (Fig. 5c)[27]. Single cells were separated by their distal contact information via scHiC topic modeling and observed distinct clusters by PDCLs[31]. Notably, even at this relatively low sequencing depth, we were able to reliably tell PDCL line sparse contact profiles apart (Fig. 5d and Supplementary Fig. 3b,c). Differences between the aggregated contact maps between clusters were then used to assess unique translocation and inversion events across the sampled cells. This identified a putative intrachromosomal translocation between the 8.5–9.5 and 88.5–91.0-megabase (Mb) regions of chromosome 12 (Fig. 5e), which exhibited contact signal comparable to proximal regions in sequence space (Supplementary Fig. 3d). The putative translocation contained *ATP2B1*, which is commonly overexpressed in PDAC[32] and the tumor suppressor gene *DUSP6* (ref. [33]), and is only present in PDAC-1.

Discussion

Our s3 workflow provides 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. 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 use 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[8] or CUT&Tag[34]. Last, unlike sci workflows, the s3 platform does not require custom sequencing primers or custom sequencing recipes, removing one of the main hurdles that groups may face while implementing these technologies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-00962-z.

Received: 8 January 2021; Accepted: 20 May 2021;
Published online: 05 July 2021

References

1. Cusanovich, D. A. et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
2. 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).
3. Tan, L., Xing, D., Chang, C. H., Li, H. & Xie, X. S. Three-dimensional genome structures of single diploid human cells. *Science* **361**, 924–928 (2018).
4. Sos, B. C. et al. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biol.* **17**, 20 (2016).
5. Yin, Y. et al. High-throughput single-cell sequencing with linear amplification. *Mol. Cell* **76**, 676–690.e10 (2019).
6. Chen, C. et al. Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). *Science* **356**, 189–194 (2017).
7. Adey, A. & Shendure, J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res.* **22**, 1139–1143 (2012).
8. Mulqueen, R. M. et al. Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).
9. Wang, O. et al. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res.* **29**, 798–808 (2019).
10. Vitak, S. A. et al. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14**, 302–308 (2017).
11. Amini, S. et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).
12. Adey, A. et al. In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014).
13. 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).
14. *Single Cell ATAC: Official* (10X Genomics Support, 2020); https://support.10xgenomics.com/single-cell-atac
15. Lareau, C. A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility.*Nat. Biotechnol.* **37**, 916–924 (2019).
16. 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).
17. Bravo González-Blas, C. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
18. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2018).
19. Stuart, T., Srivastava, A., Lareau, C. & Satija, R. Multimodal single-cell chromatin analysis with Signac. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.09.373613 (2020).
20. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
21. Thornton, C. A. et al. Spatially mapped single-cell chromatin accessibility. *Nat. Commun.* **12**, 1274 (2021).
22. Domcke, S. et al. A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7721 (2020).
23. 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).
24. Laks, E. et al. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell* **179**, 1207–1221.e22 (2019).
25. Raphael, B. J. et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* **32**, e13 (2017).
26. Lindenburger, K. et al. Drug responses of patient-derived cell lines in vitro that match drug responses of patient PDAc tumors in situ. *Ann. Pancreat. Cancer* **1**, AB024 (2018).
27. Ramani, V. et al. Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017).
28. Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).
29. 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).
30. Wu, H. et al. PRSS1 genotype is associated with prognosis in patients with pancreatic ductal adenocarcinoma. *Oncol. Lett.* **19**, 121–126 (2020).
31. 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.* https://doi.org/10.1371/journal.pcbi.1008173 (2020).
32. Sritangos, P. et al. Plasma membrane Ca^{2+} atpase isoform 4 (PMCA4) has an important role in numerous hallmarks of pancreatic cancer. *Cancers* **12**, https://doi.org/10.3390/cancers12010218 (2020).
33. 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 Biol. Med.* **15**, 14–28 (2018).
34. Kaya-Okur, H. S. et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
35. 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).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

PDCL propagation. Low-passage, PDCLs were propagated from rapidly dissociated PDAC tumors and cultured for continuous propagation in culture medium containing ROCK inhibitor (Y-276320)[36]. Briefly, approximately 50,000 viable, disaggregated tumor cells were plated to a 35 mm diameter, collagen-coated well (Gibco, A11428-02) and passaged 1:3 while subconfluent until reaching 85% confluence on a 10-cm diameter dish. From a fraction of these cells, DNA was extracted to validate the presence of KRAS-G12 mutations by droplet digital PCR (BioRad, 1863506) and to validate a short tandem repeat 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. To ensure that only tumor cells were cultured, G-banded karyotyping was performed by the Knight Diagnostic Research Cytogenetics Laboratory at the Oregon Health and Science University (OHSU). Chromosome spreads from more than 20 cells were observed to ensure that the culture contained 100% tumor cells.

WES and analysis. WES libraries for the patient's blood sample, tumor biopsy and PDCL were carried out by the Knight Diagnostic Research Cytogenetics Laboratory at OHSU. Libraries were prepared using 500 ng of fragmented genomic DNA using KAPA Hyper-Prep Kit (KAPA Biosystems) with Agilent SureSelect XT Target Enrichment System and Human All Exon V5 capture baits (Agilent Technologies), following the manufacturer's protocols. Sequencing was carried out using the Illumina HiSeq 2500 platform by the OHSU Massively Parallel Sequencing Shared Resource. Paired-end reads were aligned with bwa mem (v.0.7.15-r1140) to GRCh38 (hg38, Genome Reference Consortium Human Reference[37] (GCA_000001405.2))[38]. The data were processed following the best practices workflow for the GATK pipeline (v.4.1.9.0)[39]. Exome regions annotated as 'protein-coding' were extracted from GenCode (v.35)[37] 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 (v.3.3.2) in R (v.4.0.0). 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 (Fig. 5a).

s3-ATAC library generation. A formatted stepwise protocol for s3-ATAC is available for review at https://doi.org/10.17504/protocols.io.bd6wi9fe.

Before sample handling, 96 uniquely indexed transposome complexes were assembled using previously described methods[11]. Complexes were diluted to 2.5 µM 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 (Supplementary Table 2). At the time of nuclei dissociation, 50 ml 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, 3 mM 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 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 h postmortem and then placed in a −80 °C freezer for storage. An at-bench dissection stage was set up before nuclei extraction. A petri dish was placed over dry ice, with fresh sterile razors prechilled by dry-ice embedding. Then 7-ml capacity dounce homogenizers were filled with 2 ml 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 15-ml tubes on ice to chill. Immediately before 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 prechilled petri dish and roughly minced with the razors. Razors were then used to transport roughly 1 mg of the minced tissue into the chilled NIB-HEPES buffer within a dounce homogenizer. Suspended samples were given 5 min to equilibrate to the change in salt concentration before douncing. Tissues were then homogenized with five strokes of a loose (A) pestle, another 5 min of incubation and five to ten strokes of a tight (B) pestle. Samples were then filtered through a 35-µm cell strainer (Corning 352235) during transfer to a 15 ml conical tube, and nuclei were held on ice until ready to proceed. Nuclei were pelleted with a 400 relative centrifugal field (r.c.f.) centrifugation at 4 °C in a centrifuge for 10 min. Supernatant was removed and pellets were resuspended in 1 ml 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 10-µl aliquot of suspended nuclei was diluted in 90 µl of NIB-HEPES (1:10 dilution) and quantified on either a Hemocytometer or with a BioRad TC-20 Automated cell counter following the manufacturer's recommended protocols. The stock nuclei suspension was then diluted to a concentration of 1,400 nuclei per µl.

Tagmentation plates were prepared by the combination of 420 µl of 1,400 nuclei per µl of solution with 540 µl 2× TD Buffer (Nextera XT Kit, Illumina Inc. FC-131-1024). From this mixture, 8 µl (roughly 5,000 nuclei in total) was pipetted into each well of a 96-well plate dependent on well schema (Fig. 1b). Then 1 µl of 2.5 µM uniquely indexed transposase was then pipetted into each well. Tagmentation was performed at 55 °C for 10 min on a 300 r.c.f. 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 2 µl of 5 mg ml^{-1} 4,6-diamidino-2-phenylindole (DAPI) (Thermo Fisher Scientific D1306) were added.

Nuclei were then flow sorted via a Sony SH800 to remove debris and attain an accurate count per well before PCR. A receptacle 96-well plate was prepared with 9 µl of 1× 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[40]. Immediately following sorting completion, the plate was sealed and spun down for 5 min at 500 r.c.f. and 4 °C to ensure nuclei were within the buffer.

Nucleosomes and remaining transposases were then denatured with the addition 1 µl of 0.1% SDS (roughly 0.01% final concentration) per well. Then 4 µl of NPM (Nextera XT Kit, Illumina Inc.) per well was subsequently added to perform gap-fill on tagmented gDNA, with an incubation at 72 °C for 10 min. Next, 1.5 µl of 1 µM A14-LNA-ME oligo was then added to supply the template for adapter switching (Supplementary Table 3). The polymerase-based adapter switching was then performed with the following conditions: initial denaturation at 98 °C for 30 s, ten cycles of 98 °C for 10 s, 59 °C for 20 s and 72 °C for 10 s. The plate was then held at 10 °C. After adapter switching 1% (v/v) Triton-X 100 in ultrapure H_2O (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 µl of sample, 2.5 µl of indexed i7 primer at 10 µM, 2.5 µl of indexed i5 primer at 10 µM, 3 µl of ultrapure H_2O, 25 µl of NEBNext Q5U 2× Master mix (New England Biolabs M0597S) and 0.5 µl of 100× SYBR Green I (Thermo Scientific S7563) for a 50 µl of reaction per well (Supplementary Tables 4 and 5). A real-time PCR was performed on a BioRad CFX with the following conditions, measuring SYBR fluorescence every cycle: 98 °C for 30 s; 16–18 cycles of 98 °C for 10 s, 55 °C for 20 s, 72 °C for 30 s, fluorescent reading, 72 °C for 10 s. After fluorescence passes an exponential growth and begins to inflect, the samples were held at 72 °C for another 30 s then stored at 4 °C.

Amplified libraries were then cleaned by pooling 25 µl per well into a 15-ml conical tube and cleaned via a Qiaquick PCR purification column following the manufacturer's protocol (Qiagen 28106). The pooled sample was eluted in 50 µl 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). Next, 50 µl of vortexed and fully suspended room temperature SPRI beads were combined with the 50-µl library (1× clean up) and incubated at room temperature for 5 min. The reaction was then placed on a magnetic rack and once cleared, supernatant was removed. The remaining pellet was rinsed twice with 100 µl of 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. Next, 31 µl of 10 mM Tris-HCl, pH 8.0 were then used to resuspend the beads off the magnetic rack and allowed to incubate for 5 min 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 double-stranded DNA High-sensitivity assay following the manufacturer's instructions (Thermo Fisher Q32851). Libraries were then diluted to 2 ng µl^{-1} and run on an Agilent Tapestation 4150 D5000 tape (Agilent 5067-5592). Library molecule concentration within the range of 100–1,000 base pairs (bp) 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 the manufacturer's recommendations (Illumina Inc. 20024907, 20024904). For greater sequencing effort, select libraries were also sequenced on a NovaSeq S2 flowcell, again following the manufacturer's recommendations (Illumina Inc. 20028315). For both machines, libraries were sequenced as paired-end libraries with ten cycle index reads and 85 cycles for reads 1 and 2.

s3-WGS library generation. A formatted stepwise protocol for s3-WGS is available for review at https://doi.org/10.17504/protocols.io.beb3jaqn.

Before processing, the following buffers were prepared: 50 ml of NIB-HEPES buffer as described above, as well as 50 ml of a Tris-based NIB (NIB Tris) variant with final concentrations of 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 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 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 before processing. At collection, cells were washed twice with ice-cold 1× PBS (VWR 75800-986) and then trypsinized with 5 ml of 1× TrypLE (Thermo Fisher 12604039) for 15 min at 37 °C. Suspended cells were then collected and pelleted at 300 r.c.f. at 4 °C for 5 min. For suspension-growth cell lines (GM12878), cells were pipetted from growth media and pelleted at 300 r.c.f. at 4 °C for 5 min.

Following the initial pellet, cells were washed with ice-cold 1 ml of NIB-HEPES twice. After the second wash, pellets were then resuspended in 300 µl of 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 r.c.f. centrifugation at 4 °C for 5 min and resuspended in 5 ml of NIB-HEPES. Next, 246 µl of 16% (w/v) formaldehyde (Thermo Fisher 28906) were then added to nuclear suspensions (final concentration 0.75% formaldehyde) to lightly fix nuclei. Nuclei were fixed via incubation in formaldehyde solution for 10 min on an orbital shaker set to 50 r.p.m. Suspensions were then pelleted at 500 r.c.f. for 4 min 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 r.c.f. for 4 min at 4 °C and supernatant was aspirated. The pellet was washed once with 500 µl of 1× NEBuffer 2.1 (NEB B7202S) and then resuspended with 760 µl of 1× NEBuffer 2.1. Then 40 µl 1% SDS (v/v) was added and sample was incubated on a ThermoMixer at 300 r.c.f. set to 37 °C for 20 min. Nucleosome depleted nuclei were then pelleted at 500 r.c.f. at 4 °C for 5 min and then resuspended in 50 µl of NIB Tris. A 5-µl aliquot of nuclei was taken and diluted 1:10 in NIB Tris then quantified as described above. Nuclei were diluted to 500 nuclei per µl with addition of NIB Tris, based on the quantification. Dependent on experimental setup, the 420 µl of nuclei at 500 nuclei per µl were then combined with 540 µl of 2× TD buffer. Following this, nuclei were tagmented, stained and flow sorted, gDNA 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. A formatted stepwise protocol for s3-WGS is available for review at: https://doi.org/10.17504/protocols.io.beb4jaqw.

The same cultured cell line samples were collected 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 (roughly 2–3 million nuclei per sample) were pooled respective of sample. Nuclei were pelleted at 500 r.c.f. at 4 °C for 5 min and resuspended in 90 µl of 1× Cutsmart Buffer (NEB B7204S). Next, 10 µl of 10 U µl^{-1} AluI restriction enzyme (NEB R0137S) were added to each sample. Samples were then digested for 2 h at 37 °C at 300 r.p.m. on a ThermoMixer. Following digestion, nuclear fragments then underwent proximity ligation. Nuclei were pelleted at 500 r.c.f. at 4 °C for 5 min and resuspended in 100 µl of ligation reaction buffer. Ligation buffer is a mixture with final concentrations of 1× T4 DNA Ligase Buffer + ATP (NEB M0202S), 0.01% Triton-X-100, 0.5 mM DTT (Sigma D0632), 200 U of T4 DNA Ligase, diluted in ultrapure H_2O. Ligation took place at 16 °C for 14 h (overnight). Following this incubation, nuclei were pelleted at 500 r.c.f. at 4 °C for 5 min and resuspended in 100 µl of NIB-HEPES buffer. An aliquot of nuclei were quantified as described previously, and were then diluted, aliquoted, tagmented, pooled, DAPI stained, flow sorted, gDNA 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 were 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 and deduplicated using the in-house scitools pipeline[40]. 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 (v.0.7.15-r1140) to hg38 (ref. [38]). 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 deduplicated to remove PCR and optical duplicates by a perl (v.5.16.3) 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 (Fig. 1c). 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 pretagmentation 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%, Fig. 2b). 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 deduplicated to obtain proper estimates of library complexity (Supplementary Fig. 6).

Tagmentation insert quantification. To assess tagmentation insert size, samtools isize (v.1.10) was performed and plotted with ggplot2 (v.3.3.2) in R (v.4.0.0) using the geom_density function (default parameters, Fig. 2e). To assess library quality further, we generated tagmentation site density plots centered around tTSSs. 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 TSSs (v.2.4.36)[41]. From this, we collapsed binned distances (100-bp increments) into a counts table and generated the 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 2,000 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 (Fig. 2f).

Library complexity analysis. To project library complexity through sequencing effort, pre-deduplicated 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 datasets 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 that have been applied to postnatal mouse whole brain (n = 3,034, 5,336 and 46,653 for snATAC, 10× Genomics scATAC and dscATAC, respectively)[13–15]. Our data were filtered to just unique reads that were properly paired via the samtools view '–f 2' argument to allow for proper read-count comparison across datasets that report read pairs or fragments. Data passing self-reported filters were used for comparison and plotted with the ggplot geom_boxplot function (Fig. 2d). 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. For peak overlap comparison, we added an additional sci-ATAC low sequencing effort dataset[21] performed on adult mouse flash frozen brain tissue. We then counted unique peak overlaps between datasets and plotted as stacked bar plots via ggplot geom_bar function (Fig. 2g).

For assessment on sequencing effort necessary to reach a median unique reads per cell threshold, we compared our s3-ATAC mouse data to the publically available sci-ATAC matched sample data[21]. We randomly subsampled the bam files pre-deduplication and calculated per cellID library complexity, as described above[10,21]. The resulting model was plotted with geom_line. We then calculated amount of PCR duplicate reads at that threshold for sequencing effort comparison (Supplementary Fig. 1e).

Dimensionality reduction. Pseudo-bulked data (agnostic of cellID) were then used to call read pile-ups or peaks via macs2 (v.2.2.7.1) with option --keep-dup all[42]. Narrowpeak bed files were then merged by overlap and extended to a minimum of 500 bp 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 (v.1.7). The counts matrix and tabix files were then input into a SeuratObject for Signac (v.1.0.0) processing[19,43]. We performed LDA-based dimensionality reduction via cisTopic (v.0.3.0) with 27 topics for mouse cells and 24 topics for human cells[17]. The number of topics were selected after generating 25 separate models per species with topic counts of five, ten, 20–30, 40, 50, 55 and 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 UMAP by the function umap in the uwot package (v.0.1.8, Fig. 2h)[44]. *Cis*-coaccessibility networks were generated through the Signac wrapper of cicero (v.1.3.4.10)[45]. Genome track plots with *cis*-coaccessibility network linkages were generated through Signac function CoveragePlot for marker genes previously described (Supplementary Fig. 1a)[19]. 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 (Supplementary Tables 6 and 7and Supplementary Information).

Subsampling. For subsampling analysis, the processed, deduplicated bam files were split by cellID into single-cell bam files. Each bam file was then subsampled randomly using samtools view --s argument for 0.5, 1, 2, 5, 10, 15, 20, 40, 50, 60 and 80%, respectively. Following this single-cell subsampled bams were collated respective of downsampling percentage and processed through peak calling, dimensionality reduction and projection as described above with the following exception. Topic model generation was limited to ten and 20–30 topics. Number of peaks callable per downsampled dataset were plotted via geom_bar. The final projections were plotted via geom_point with the color of the cell type assignment in the full dataset (Supplementary Fig. 1b–d).

Cross-platform integration. Data used in library complexity and peak overlap comparisons were integrated using the Signac package as follows[46]. Counts matrices were generated using the platform-defined peak regions and formatted as Seurat objects. For each integration, our s3-ATAC mouse data, we generated a new counts matrix was generated from the platform-defined peaks. Following this, counts matrices were merged and latent semantic indexing (lsi) was used for reduction. Harmony was used to integrate, and a UMAP projection was generated using dimensions 2–30 for sci-ATAC, scATAC and snATAC datasets and 2–40 for dscATAC[47]. Integrated plots were generated using our defined cell types (Fig. 2i).

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[19,44,48]. CisTopic was performed on the full set of human peaks (292,156) with those 342 subset cells. Twelve 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 five subclusters. One cluster was removed based on putative doublets (Fig. 3a). Coverage plots were generated as reported above for *ADARB2* and *LHX6* (Fig. 3b). Peaks were then assigned to topics using the cisTopic binarizecisTopics function with argument thrP = 0.975 (mean count per topic, 2,429 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[20]. 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 ($P \leq 0.05$) and used ggplot geom_point with color reflecting the reported P value and size proportional to odds ratio to generate a bubble plot (Fig. 3c).

s3-WGS and s3-GCC analysis. *Quality control.* s3-WGS and s3-GCC cellIDs were initially filtered to samples with either $\geq 1\times10^5$ or $\geq 1\times10^6$ unique reads (PDCL and GM12878 samples, respectively). CellIDs were split after deduplication into single-cell bam files. They were then processed via the pipeline in the package SCOPE (v.1.1)[23]. The genome was split into 500-kb bins with each bin being assigned a GC content and mappability score (generated through CODEX2)[49]. Reads with a mapping quality of $Q \geq 10$ were counted in bins per cellID. Bins with a mappability score <0.9 or GC content ≤20 or ≥80% were removed (5,449 bins passing filter). Additionally, cellIDs with low coverage were removed (1,268 samples passing filter). MAD scores were calculated per cell on 500-kb bins of cells passing filter as previously described[23]. Briefly, let $Y_{i,j}$ be the raw read count for the ith cellID of the jth 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

$$\text{MAD score}_i = \text{median}\left(|d - \text{median}(d)|\right)$$

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

MAD scores were then plotted using the ggplot geom_jitter and geom_boxplot functions (Fig. 4e).

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 (Fig. 4c). 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 (v.2.5.5) by function Heatmap[50]. Pairwise distance between cells was generated by Jaccard distance through the R library philentropy (v.0.4.0)[51] 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 three and six clades based on the two independent optimal k value searches using the find_k function in the R library dendextend (v.1.14.0) given a range of two to ten and five to ten clusters, respectively (Fig. 5b)[52]. Cells with shared clade membership were then combined into pseudobulk clades for higher resolution copy number calling. After combining counts data across 50-kb bins (and filtered as described above), we had six clades with 154, 250, 363, 100, 268 and 133 cells, with mean reads per bin of 1,207, 2,442, 4,662, 2,071, 2,700 and 9,416, 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 (Supplementary Fig. 2a). Select genomic locations[25] of recurrently mutated genes were visualized and plotted using IGV with five bins (250 kb) up- and downstream from the TSSs (Supplementary Fig. 2b)[53].

Generation of GCC contact profiles. 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 ≥50 kb via pysam[54] and output as upper-triangle triple-sparse format at 1-Mb bin sizes. Raw contact matrices were then plotted with R and ComplexHeatmap (Fig. 5c, left). Merged ensemble plots were generated by summing single-cell contact matrices generated as described above for 500-kb 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 20 kb) as traditional distal interactions. We analyzed these cells using our previously established topic model for analysis and characterization of single-cell HiC data[31]. 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 probabilistic 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 20 Mb or less, as opposed to original parameter where we used interactions that are separated by distances lower than 10 Mb. After training, we found that the number of topics that best describe the data is 15. We visualized cells using UMAP and found that most cells from two lines cluster separately (Fig. 5c). Overall, these results validate the HiC-like characteristics of GCC data and further show that we can capture the subtle differences in chromatin organization of the two lines.

Compartment calling. We called compartments from pseudobulk HiC contact matrices and calculated contact probabilities as described in the original HiC paper by Lieberman-Aiden et al.[55]. We briefly describe these methods here. To obtain compartment calls, we first normalized using iterative correction and eigenvector decomposition (ICE) and removed the distance effect in HiC matrices for each chromosome at 500-kb resolution. For each normalized HiC contact matrix, we calculated the Spearman Correlation Coefficient matrix from the normalized HiC matrix, which yields a matrix with clear plaid pattern. Resulting matrices were reduced to one dimensional representation by performing eigendecomposition; the first eigenvector typically yields the compartment calls, which closely tracks the two-dimensional plaid pattern in one dimension. For comparison we used compartment calls from Rao et al.[35]. on the GM12878 cell line (Fig. 5c, right).

Calculation of contact probabilities. To calculate intrachromosomal contact probabilities for each genomic distance, we first calculated the mean HiC signal at a given genomic distance for a given resolution. For a HiC matrix binned at 500-kb resolution, the mean HiC signal for distances less than 500 kb is the mean of the diagonal, the mean HiC signal for distances between 500 and 1,000 kb is the mean of the first off-diagonal, and so forth. After calculating the mean HiC signal vector for the every intrachromosomal matrix, we obtain a mean HiC signal vector at 500-kb resolution that contains 500 elements, ranging from 0 to around 250 Mb, since the largest chromosome (Chr1) is approximately 250 Mb long. To convert this vector into probabilities, we simply divide the vector by the sum. When plotting contact probabilities, we typically omit the visualizing the contact probabilities for distances larger than 200 Mb, as the mean HiC signal at such long distances is both sparse and noisy. External bulk HiC datasets have been downloaded from the ENCODE consortium's data portal, https://www.encodeproject.org/ via accession codes ENCSR194SRI, ENCSR346DCU, ENCSR444WCZ and ENCSR079VIJ (Supplementary Fig. 3a).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data discussed in this publication have been deposited in the National Center for Biotechnology Information's (NCBI's) Gene Expression Omnibus (GEO) and are accessible through GEO Series accession number GSE174226. External single-cell ATAC datasets were downloaded from GEO sample accession number GSM2668124 for snATAC, and external sites for dscATAC (https://github.com/buenrostrolab/dscATAC_analysis_code/blob/master/mousebrain/data/mousebrain-master_dataframe.rds) and 10X Genomics scATAC (https://cf.10xgenomics.com/samples/cell-atac/1.1.0/atac_v1_adult_brain_fresh_5k). The external single-cell WGS dataset was downloaded from NCBI BioProject PRJNA326698 (https://www.ncbi.nlm.nih.gov/sra/SRX2005587). Single-cell HiC datasets were downloaded from the 4D Nucleosome project (https://data.4dnucleome.org/publications/048d4558-2cac-41d2-ac6e-ff2ac3f007c4/#expsets-table). External bulk HiC datasets have been downloaded from the ENCODE consortium's data portal, https://www.encodeproject.org/ via accession codes ENCSR194SRI, ENCSR346DCU, ENCSR444WCZ and ENCSR079VIJ. Source data are provided with this paper.

Code availability

Code and custom scripts used in this study are available at https://github.com/adeylab/scitools and https://mulqueenr.github.io/.

References

36. Liu, X. et al. Conditional reprogramming and long-term expansion of normal and tumor cells from human biospecimens. *Nat. Protoc.* **12**, 439–451 (2017).
37. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
38. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
39. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at *bioRxiv* https://doi.org/10.1101/201178 (2017).
40. Sinnamon, J. R. et al. The accessible chromatin landscape of the murine hippocampus at single-cell resolution. *Genome Res.* **29**, 857–869 (2019).
41. Neph, S. et al. BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
42. Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
43. 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).
44. Mcinnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection software review repository archive. *J. Open Source Softw.* https://doi.org/10.21105/joss.00861 (2018).
45. Pliner, H. A. et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* https://doi.org/10.1016/j.molcel.2018.06.044 (2018).
46. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
47. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
48. González-Blas, C. B. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
49. Jiang, Y. et al. CODEX2: Full-spectrum copy number variation detection by high-throughput DNA sequencing. *Genome Biol.* **19**, 202 (2018).
50. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data.*Bioinformatics* **32**, 2847–2849 (2016).
51. Drost, H.-G. Philentropy: information theory and distance quantification with R. *J. Open Source Softw.* **3**, 765 (2018).
52. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718–3720 (2015).
53. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
54. pysam. https://github.com/pysam-developers/pysam (GitHub, 2020).
55. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

Acknowledgements

We thank other members of the Adey Laboratory as well as J. Shendure and C. Trapnell for helpful suggestions and feedback. This work was funded by grants R01DA047237 (NIH/NIDA) and R35GM124704 (NIH/NIGMS) to A.C.A. and R01MH113926 (NIH/NIMH) to B.J.O. We also thank the Oregon Brain Bank for the donated biological sample used in this study.

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. The paper was written by R.M.M. and A.C.A. All authors reviewed and contributed to the paper.

Competing interests

D.P., F.Z. and F.J.S. are employees of Scale Bio. R.M.M., D.P., F.Z., F.J.S. and A.C.A. are authors on licensed patents that cover components of the technologies described in this paper. This potential conflict of interest for A.C.A. and R.M.M. has been reviewed and managed by OHSU.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-00962-z.

Correspondence and requests for materials should be addressed to A.C.A.

Peer review information *Nature Biotechnology* thanks Kun Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Ryan Mulqueen, Andrew Adey

Last updated by author(s): May 11, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

□ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: CFX Manager v3.1 for real-time PCR, and NextSeq500 NCS v4.0 or NovaSeq 6000 for sequencing.

Data analysis: Bcl2Fastq (Illumina Inc, v2.19.0) was used to generate fastq files from bcl files. Bwa mem (v0.7.15-r1140) for sequencing read alignment. Macs2 (v2.1.1.20160309) for chromatin accessible peak calling. Scitools suite of single-cell analysis tools for data pre-processing (github.com/adeylab/scitools). BEDOPS (v2.4.36) was used for transcription site enrichment calculations. R version 4.0.0 was used for analysis. The following packages were used for analyses cisTopic(v0.3.0), EnsDb.Mmusculus (v79_2.99.0), EnsDb.Hsapiens (v86_2.99.0), monocle3 (v0.2.3.0), ggplot2 (v3.3.2), Seurat (v3.2.1) and Signac (v1.1.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE174226 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174226). External single-cell ATAC data sets were downloaded from GEO sample accession number GSM2668124 for snATAC (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2668124), and external sites for dscATAC (https://github.com/buenrostrolab/dscATAC_analysis_code/blob/master/mousebrain/data/mousebrain-master_dataframe.rds), and 10X Genomics scATAC (https://cf.10xgenomics.com/samples/cell-atac/1.1.0/atac_v1_adult_brain_fresh_5k). External single-cell WGS dataset was downloaded from NCBI BioProject PRJNA326698 (https://www.ncbi.nlm.nih.gov/sra/SRX2005587). Single-cell Hi-C data sets were downloaded from the 4D Nucleosome project (https://data.4dnucleome.org/publications/048d4558-2cac-41d2-ac6e-ff2ac3f007c4/#expsets-table). External bulk Hi-C datasets have been downloaded from the ENCODE consortium's data portal, https://www.encodeproject.org/ via accession codes ENCSR194SRI, ENCSR346DCU, ENCSR444WCZ, ENCSR079VIJ.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical analyses were performed to pre-determine sample size. The number of single-cell libraries to be generates was seletected for comparison with other single-cell chromatin accessibility data sets. These external data sets were also used to infer expected effect sizes, technical limitations and budget considerations. Single-cell whole genome and GCC librariy sizes were chosen to demonstrate the technology.
Data exclusions	Reads were excluded if indexes did not match predetermined barcode sequences. Barcodes (used to identify cellIDs) were filtered as described in the manuscript and required reads to have a Q-score of >= 10.
Replication	s3-ATAC was used to generate a total of 3 independent experiment replicates. All attempts at replication were successful. Single-cell libraries were treated as independent measurements of the biological systems used.
Randomization	Organism allotment was not random, as data generated was used to demonstrate primarily quality metrics. Cell libraries were inherently randomized during the s3 protocol after a mixing and diluting step.
Blinding	Cell libraries were inherently blinded during the PCR phase of the s3 protocol after a mixing and diluting step.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☐	☒	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	GM12878, Coriell; Patient-derived cell lines (PDCLs)
Authentication	From a fraction of these PDCL 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. To ensure that only tumor cells were cultured, G-banded karyotyping was performed by the Knight Diagnostic Research Cytogenetics Lab at OHSU.
Mycoplasma contamination	Lines were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used in this experiment.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	The mouse used for whole brain experiments were C57BL/6J, aged within 8-12 weeks. All mouse cages were kept on a 12 h light/dark cycle at a temperature of 70F and within a humidity range of 30-70%.
Wild animals	No wild animals were used in this study.
Field-collected samples	No field-collected samples were used in this study.
Ethics oversight	All work overseen and approved by OHSU Institutional Animal Care and Use Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.