our study that the PPC is a good candidate for future clinical applications as it contains signals both overlapping and likely complementary to those found in M1.

REFERENCES AND NOTES

1. R. A. Andersen, C. A. Buneo, *Annu. Rev. Neurosci.* **25**, 189–220 (2002).
2. J. C. Culham, C. Cavina-Pratesi, A. Singhal, *Neuropsychologia* **44**, 2668–2684 (2006).
3. R. Balint, *Monatsschr. Psychiatr. Neurol.* **25**, 51–81 (1909).
4. M. T. Perenin, A. Vighetto, *Brain* **111**, 643–674 (1988).
5. M. A. Goodale, A. D. Milner, *Trends Neurosci.* **15**, 20–25 (1992).
6. A. Sirigu *et al.*, *Science* **273**, 1564–1568 (1996).
7. L. Pisella *et al.*, *Nat. Neurosci.* **3**, 729–736 (2000).
8. S. Musallam, B. D. Corneil, B. Greger, H. Scherberger, R. A. Andersen, *Science* **305**, 258–262 (2004).
9. M. Hauschild, G. H. Mulliken, I. Fineman, G. E. Loeb, R. A. Andersen, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17075–17080 (2012).
10. G. H. Mulliken, S. Musallam, R. A. Andersen, *J. Neurosci.* **28**, 12913–12926 (2008).
11. R. A. Andersen, S. Kellis, C. Klaes, T. Aflalo, *Curr. Biol.* **24**, R885–R897 (2014).
12. W. Truccolo, G. M. Friehs, J. P. Donoghue, L. R. Hochberg, *J. Neurosci.* **28**, 1163–1178 (2008).
13. S.-P. Kim, J. D. Simeral, L. R. Hochberg, J. P. Donoghue, M. J. Black, *J. Neural Eng.* **5**, 455–476 (2008).
14. J. W. Gnadt, R. A. Andersen, *Exp. Brain Res.* **70**, 216–220 (1988).
15. G. H. Mulliken, S. Musallam, R. A. Andersen, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 8170–8177 (2008).
16. D. M. Wolpert, S. J. Goodbody, M. Husain, *Nat. Neurosci.* **1**, 529–533 (1998).
17. L. H. Snyder, A. P. Batista, R. A. Andersen, *Nature* **386**, 167–170 (1997).
18. S. W. C. Chang, A. R. Dickinson, L. H. Snyder, *J. Neurosci.* **28**, 6128–6140 (2008).
19. S. Dangi *et al.*, *Neural Comput.* **26**, 1811–1839 (2014).
20. A. P. Georgopoulos, J. F. Kalaska, R. Caminiti, J. T. Massey, *J. Neurosci.* **2**, 1527–1537 (1982).
21. L. R. Hochberg *et al.*, *Nature* **442**, 164–171 (2006).
22. J. L. Collinger *et al.*, *Lancet* **381**, 557–564 (2013).

ACKNOWLEDGMENTS

We thank EGS for his unwavering dedication and enthusiasm, which made this study possible. We acknowledge V. Shcherbatyuk for computer assistance; T. Yao, A. Berumen, and S. Oviedo, for administrative support; K. Durkin for nursing assistance; and our colleagues at the Applied Physics Laboratory at Johns Hopkins and at Blackrock Microsystems for technical support. This work was supported by the NIH under grants EY013337, EY015545, and P50 MH942581A; the Boswell Foundation; The Center for Neurorestoration at the University of Southern California; and Defense Department contract N66001-10-4056. All primary behavioral and neurophysiological data are archived in the Division of Biology and Biological Engineering at the California Institute of Technology.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/906/suppl/DC1
Materials and Methods
Figs. S1 to S3
Movies S1 to S3
References (*23*–*26*)

20 December 2014; accepted 31 March 2015
10.1126/science.aaa5417

EPIGENETICS

Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing

Darren A. Cusanovich,[1] Riza Daza,[1] Andrew Adey,[2] Hannah A. Pliner,[1] Lena Christiansen,[3] Kevin L. Gunderson,[3] Frank J. Steemers,[3] Cole Trapnell,[1] Jay Shendure[1]*

Technical advances have enabled the collection of genome and transcriptome data sets with single-cell resolution. However, single-cell characterization of the epigenome has remained challenging. Furthermore, because cells must be physically separated before biochemical processing, conventional single-cell preparatory methods scale linearly. We applied combinatorial cellular indexing to measure chromatin accessibility in thousands of single cells per assay, circumventing the need for compartmentalization of individual cells. We report chromatin accessibility profiles from more than 15,000 single cells and use these data to cluster cells on the basis of chromatin accessibility landscapes. We identify modules of coordinately regulated chromatin accessibility at the level of single cells both between and within cell types, with a scalable method that may accelerate progress toward a human cell atlas.

Chromatin state is dynamically regulated in a cell type–specific manner (*1*, *2*). To identify active regulatory regions, sequencing of deoxyribonuclease I (DNase I) digestion products [DNase-seq (*3*)] and assay for transposase-accessible chromatin using sequencing [ATAC-seq (*4*)] measure the degree to which specific regions of chromatin are accessible to regulatory factors. However, these assays measure an average of the chromatin states within a population of cells, masking heterogeneity between and within cell types.

[1]University of Washington, Department of Genome Sciences, Seattle, WA, USA. [2]Oregon Health and Science University, Department of Molecular and Medical Genetics, Portland, OR, USA. [3]Illumina, Inc., Advanced Research Group, San Diego, CA, USA.
***Corresponding author. E-mail: shendure@uw.edu**

Single-cell methods for genome sequence (*5*), transcriptomes (*6*–*10*), DNA methylation (*11*), and chromosome conformation (*12*) have been reported. However, we presently lack technologies for genome-wide, single-cell characterization of chromatin state. Furthermore, a limitation of most such methods is that single cells are individually compartmentalized, and the nucleic acid content of each cell is biochemically processed within its own reaction volume (*13*–*16*). Processing of large numbers of cells in this way can be expensive and labor intensive, and it is difficult to work with single cells, small volumes, and low nucleic acid inputs.

We recently used combinatorial indexing of genomic DNA fragments for haplotype resolution or de novo genome assembly (*17*, *18*). Here, we adapt the concept of combinatorial indexing to intact nuclei to acquire data from thousands of single cells without requiring their individualized processing (Fig. 1A). First, we molecularly barcode populations of nuclei in each of many wells. We then pool, dilute, and redistribute intact nuclei to a second set of wells, introduce a second barcode, and complete library construction. Because the overwhelming majority of nuclei pass through a unique combination of wells, they are "compartmentalized" by the unique barcode combination that they receive. The rate of "collisions"—i.e., nuclei coincidentally receiving the same combination of indexes—can be tuned by adjusting how many nuclei are distributed to the second set of wells (fig. S1) (*19*).

We sought to integrate combinatorial cellular indexing and ATAC-seq to measure chromatin accessibility in large numbers of single cells. In ATAC-seq, permeabilized nuclei are exposed to transposase loaded with sequencing adapters ["tagmentation" (*4*, *20*)]. In the context of chromatin, the transposase preferentially inserts adapters into nucleosome-free regions. These "open" regions are generally sites of regulatory activity and correlate with DNase I hypersensitive sites (DHSs).

In the integrated method, we molecularly tag nuclei in 96 wells with barcoded transposase complexes (Fig. 1A) (*17*–*19*). We then pool, dilute, and redistribute 15 to 25 nuclei to each of 96 wells of a second plate, using a cell sorter. After lysing nuclei, a second barcode is introduced during polymerase chain reaction (PCR) with indexed primers complementary to the transposase-introduced adapters. Finally, all PCR products are pooled and sequenced, with the expectation that most sequence reads bearing the same combination of barcodes will be derived from a single cell (estimated collision rate of ~11% for experiments described here) (fig. S1).

As an initial test, we mixed equal numbers of nuclei from human (GM12878) and mouse [Patski (*21*)] cell lines, performed combinatorial cellular indexing, and sequenced the resulting

library. Although at least one mappable read was observed for most of the 9216 (96 × 96) possible barcode combinations, most barcodes were associated with very few reads. We used a conservative cutoff of 500 reads per cell (*19*), retaining 533 barcode combinations for further analysis (fig. S2A) (range: 502 to 69,847 reads per barcode combination; median: 2503). A high PCR duplication rate (~73% of mappable, nonmitochondrial reads) confirmed that the library had been sequenced to saturation. We estimate that we recovered 13 to 55% of the molecular complexity that we could expect to recover based on complexity estimates for bulk, 500-cell ATAC-seq experiments (*4*, *19*).

If each barcode combination represents either a mouse or human nucleus, then its corresponding reads should map overwhelmingly to either the mouse or human genome. Indeed, we observe that ~93% of 533 barcode combinations had >90% of their reads mapping to mouse (n = 290) or human (n = 207) (Fig. 1B). In addition, these data retain signals of chromatin accessibility in relation to nucleosome hindrance of insertion events (Fig. 1C). Furthermore, 52% of reads from mouse and 50% of reads from human single cells overlapped reference DHS maps [ENCODE (*19*, *22*)] for these cell lines (20-fold and 34-fold enrichments, respectively) (Fig. 1D and table S1).

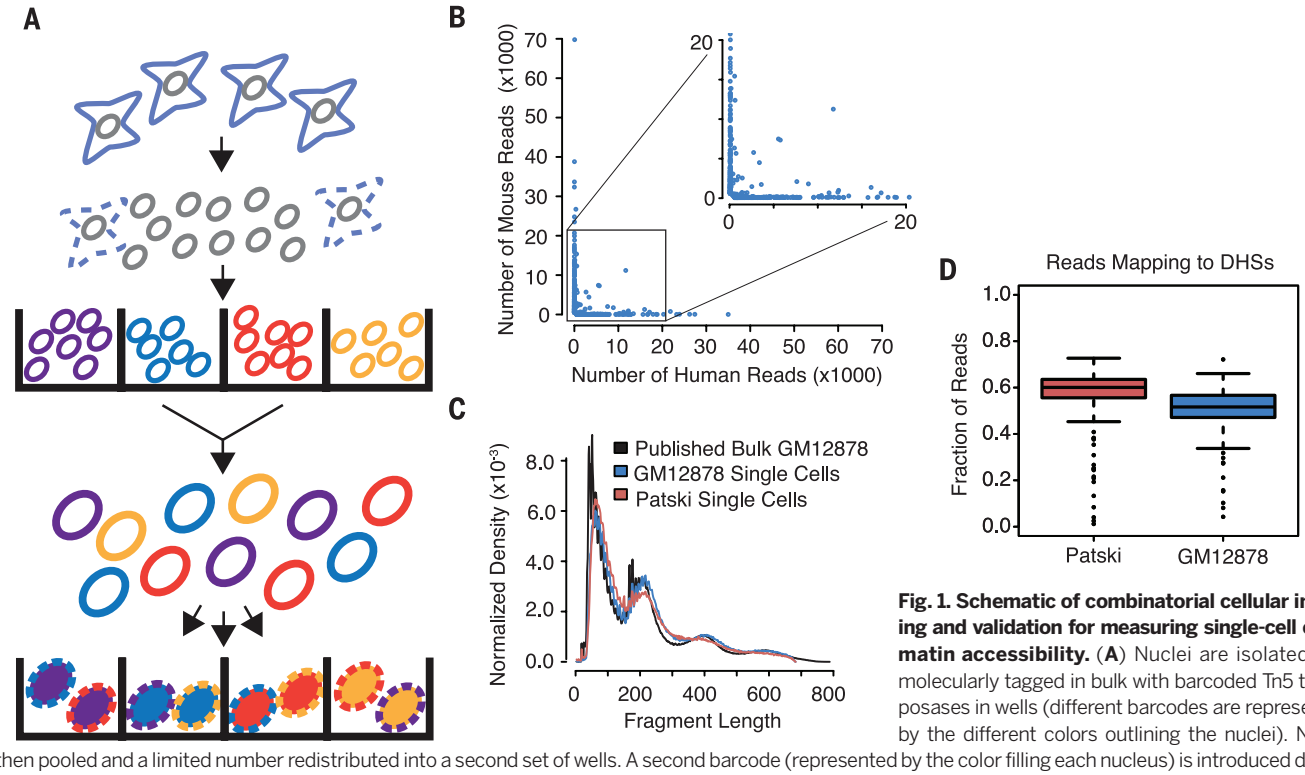

Fig. 1. Schematic of combinatorial cellular indexing and validation for measuring single-cell chromatin accessibility. (**A**) Nuclei are isolated and molecularly tagged in bulk with barcoded Tn5 transposases in wells (different barcodes are represented by the different colors outlining the nuclei). Nuclei are then pooled and a limited number redistributed into a second set of wells. A second barcode (represented by the color filling each nucleus) is introduced during PCR. (**B**) Scatterplot of number of reads mapping uniquely to human or mouse genome for individual barcode combinations. (**C**) Fragment size distribution for single-cell ATAC-seq versus published bulk ATAC-seq (*4*). (**D**) Box plot of the fraction of reads mapping to ENCODE-defined DHSs for individual Patski and GM12878 cells.

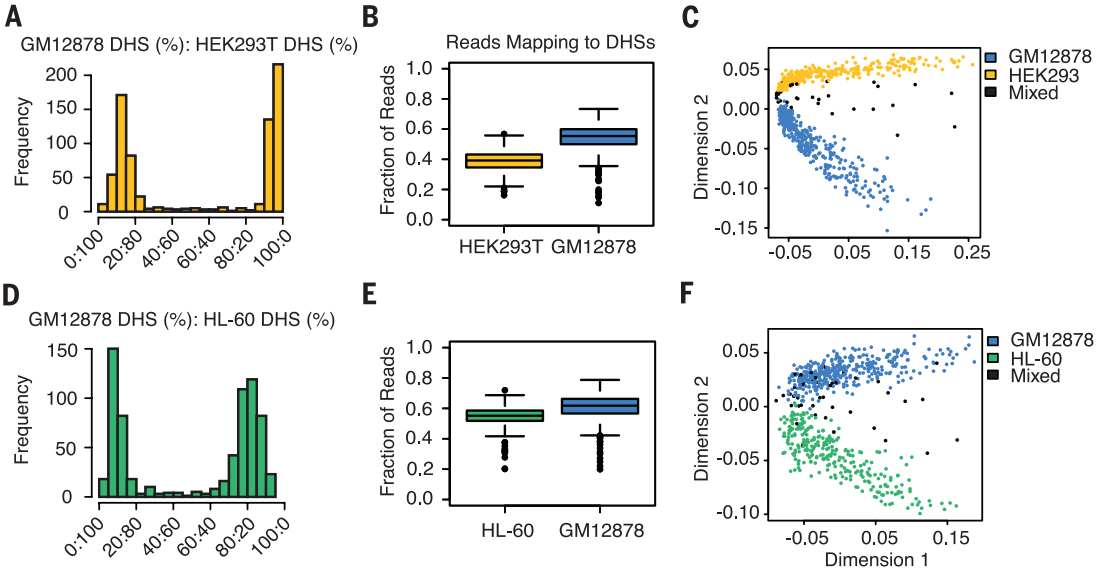

Fig. 2. Single-cell ATAC-seq deconvolutes human cell-type mixtures. (**A** to **C**) GM12878/HEK293T nuclei. (**D** to **F**) GM12878/HL-60 nuclei. [(A) and (D)] Histograms of proportions of reads mapping to cell type–specific DHSs that correspond to one cell type or the other. [(B) and (E)] Box plots of the overall fraction of reads mapping to ENCODE-defined DHSs for individual cells. [(C) and (F)] Multidimensional scaling of single-cell ATAC-seq data using pairwise Jaccard distances between cells based on DHS usage. Cell-type assignments based on proportions shown in (A) and (D).

We next sought to distinguish single cells from the same species. We mixed pairs of cell lines (HEK293T or HL-60 versus GM12878), performed combinatorial cellular indexing, and sequenced the resulting libraries to saturation (65% duplicate rate). For the mixture of HEK293T and GM12878, we recovered 748 cells with ≥500 reads (fig. S2B) (range: 502 to 28,712 reads; median: 1685 reads). Focusing on reads mapping to previously defined cell-type exclusive DHS sites (fig. S3A) (*19*, *22*), we observe a bimodal distribution, with nearly all cells assignable to one of the two cell types (~95% of 748; defined by ≥70% of reads mapping to cell type–specific DHSs corresponding to one cell type or the other) (Fig. 2A). The fraction of reads mapping to reference DHSs in single cells was again strongly enriched [41% (14-fold enrichment) for HEK293T and 52% (18-fold enrichment) for GM12878)] (Fig. 2B and table S1). About 57% of 181,379 distinct sites from the reference DHS maps were observed as accessible in at least one cell. Some fraction of these may be spurious overlaps, but this provides an upper bound on the number of DHSs for which we recovered accessibility information. Individual cells ranged in coverage of this DHS map from 29 to 5890 sites (fig. S4) (median: 429 sites).

For the mixture of HL-60 and GM12878, we recovered 700 cells (fig. S2C) (range: 500 to 21,887 reads; median: 1390 reads; 64% duplicate rate). Although both are representative of the hematopoietic lineage, 94% of cells were assignable based on the same criteria used for HEK293T/GM12878 (Fig. 2D and fig. S3B). The fraction of reads mapping to reference DHSs was again strongly enriched [55% (16-fold enrichment) for HL-60 and 59% (18-fold enrichment) for GM12878] (Fig. 2E and table S1). About 46% of 230,632 distinct sites from the reference DHS maps were observed as accessible in at least one cell, with individual cells ranging in coverage from 72 to 4687 sites (fig. S4) (median: 442 sites).

We next examined whether single cells within a heterogeneous mixture could be clustered in an unsupervised manner. Importantly, at the level of single cells, chromatin accessibility is a nearly binary phenomenon (~2 genome equivalents per cell), in contrast with the dynamic range of mRNA transcripts within single cells. Thus, we reasoned that we would require observations across each of many single cells to generate quantitative estimates for accessibility of a particular site in a particular cell type, within a heterogeneous population.

For each cell-type mixture, we defined the union of ENCODE DHSs [analogous to how RNA-seq transcript quantification relies on a catalog of transcript models (*19*)] and created a binary matrix where DHS sites were scored as "used" or "unused" in each cell. We then calculated Jaccard distances between pairs of cells on the basis of the degree of shared DHS usage. Applying multidimensional scaling to these distances, the first dimension was strongly correlated with the read depth of each cell (fig. S5) (Spearman's rho of ~0.95), whereas the second dimension separated cells consistently with our crude cell-type assignments (Fig. 2, C and F). The extent of discrimination between cell types is proportional to read depth, but even with relatively few reads, individual cells can be clustered on the basis of shared DHS usage alone. To evaluate whether our data provided reproducible and quantitative estimates of the accessibility of DHSs, we used GM12878-assigned cells from all three experiments described above as biological replicates. For each experiment, we summed the number of cells "using" each site and compared these counts between replicates (Spearman's rho's of 0.64 to 0.69, or 0.54 to 0.62 when restricted to sites observed in ≥5 cells in each replicate) and also compared them with bulk ATAC-seq measurements from 500 GM12878 cells (fig. S6) [Spearman's rho's of 0.61 to 0.7 (*4*)]. This positive correlation shows that sites that are more sensitive in bulk experiments are also more commonly observed in single cells. Furthermore, these correlations are not far from the range of 0.64 to 0.72 for replicate bulk measurements from the 500-cell ATAC-seq libraries.

To identify individual DHSs with significant differences in accessibility between different cell types (based on single-cell data from the GM12878/HL-60 mixture), we performed likelihood ratio tests within the framework of a generalized linear model. We identified 1666 sites [out of 52,479 DHSs tested (*19*)] that were differentially accessible at a false discovery rate (FDR) of 0.05. Interestingly, only about half of these sites are cell-type exclusive in the reference DHS maps (381 GM12878-exclusive and 472 HL-60-exclusive); differentially accessible DHSs are marginally enriched for GM12878-specific sites (hypergeometric $P = 0.04$) and strikingly enriched for HL-60 sites ($P = 2.2 \times 10^{-15}$). They are also larger [1184 base pairs (bp) versus 580 bp

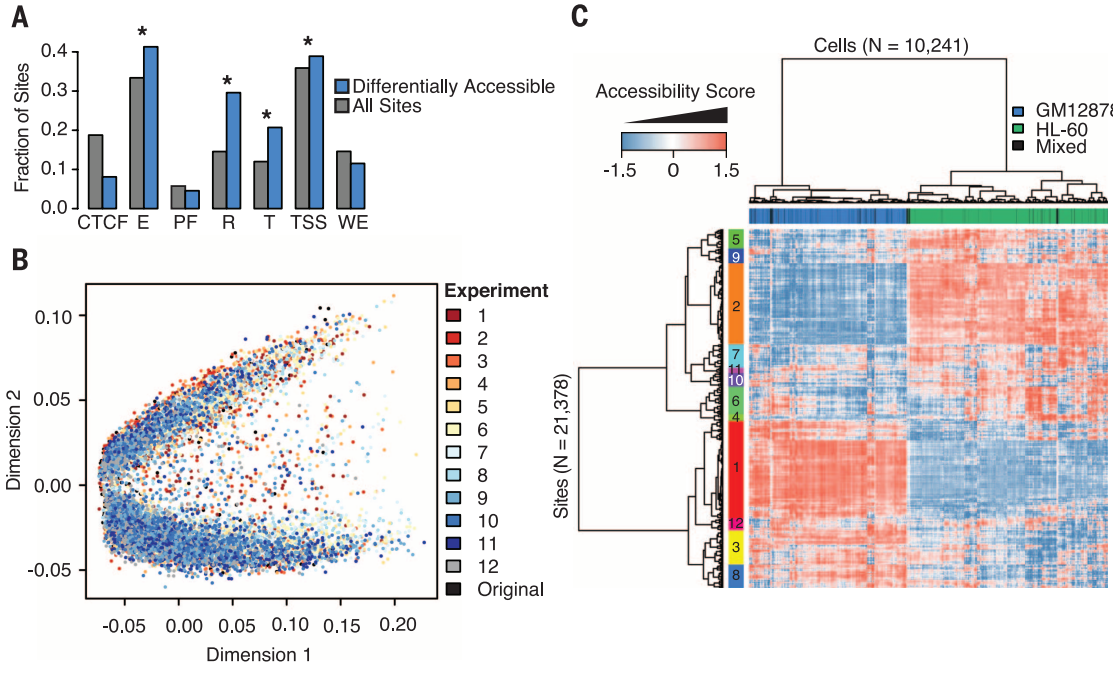

Fig. 3. Single-cell ATAC-seq identifies functionally relevant differences in accessibility between cell types. (**A**) Bar plot for relative fraction of DHSs overlapping each chromatin state (HL-60 versus GM12878). Gray bars show frequencies for all sites tested. Blue bars show frequencies for differentially accessible sites. CTCF, CTCF-enriched element; E, predicted enhancer; PF, predicted promoter flanking region; R, predicted repressed; T, predicted transcribed; TSS, predicted promoter region; WE, predicted weak enhancer. *, significant difference in proportions. Values do not add to 1 because sites can overlap multiple chromatin states. (**B**) Multidimensional scaling of chromatin accessibility data for 14,533 cells (GM12878/HL-60 mixtures from 13 experiments on four dates). (**C**) Heat map of hypersensitive site usage for 10,241 cells (columns) at 21,378 DHSs (rows) (GM12878/HL-60 mixtures). Colors indicate accessibility of sites after latent semantic indexing. Top color bar is coded by cell-type assignments (green, HL-60; blue, GM12878; black, unassigned). Left color bar indicates modules formed by clustering DHSs.

median; Wilcoxon rank sum $P = 3.4 \times 10^{-247}$], observed in more cells (10 cells versus 3 cells median; Wilcoxon rank sum $P \approx 0$), and enriched for "enhancer" (hypergeometric $P = 4.3 \times 10^{-12}$), "repressed" ($P = 1.5 \times 10^{-57}$), "transcribed" ($P = 7.4 \times 10^{-25}$), and "transcription start site" ($P = 5.1 \times 10^{-3}$) annotations in GM12878, relative to sites not identified as differentially accessible (Fig. 3A) (*19*).

We next linked differentially accessible sites defined from single cells to the genes they potentially regulate (*2*) and compared these to genes differentially expressed between GM12878 and HL-60 (*19*). Of 8268 genes linked to ≥1 DHS and expressed in both cell types, 4095 were differentially expressed and 2211 were linked to ≥1 differentially accessible DHS (FDR 0.05). Although the DHS-gene linkages are imperfect, we observe a significant overlap of differentially expressed and differentially accessible genes (1162 genes overlap; hypergeometric $P = 4.8 \times 10^{-4}$). The genes linked to DHSs identified as differentially accessible are enriched for lymphoid and myeloid lineage annotations—e.g., "cytokine signaling" and "antigen processing" (figs. S7 and S8).

To optimize combinatorial cellular indexing, we tested 12 conditions on 3 days, always with GM12878/HL-60 mixtures. We collected as many as nearly 1500 cells in a single experiment, and we improved the median read depth to >3000 per cell in some experiments (figs. S9 to S11). We merged chromatin accessibility maps for 14,533 single cells (all GM12878/HL-60) and conducted multidimensional scaling. Although the actual mixture proportion varied between experiments, the clustering of the two cell types was highly robust to experimental condition (Fig. 3B). With this full complement of cells, ~96% of 230,632 potential sites in our DHS reference map are observed as accessible in at least one cell (individual cells covering between 4 and 12,333 sites (median: 664 sites) (fig. S4).

We used latent semantic indexing to reduce the dimensionality of this matrix [after filtering out low coverage cells and rarely used sites (*19*)], yielding a heat map of chromatin accessibility for 10,241 cells at 21,378 DHSs (Fig. 3C and fig. S12). This resulted in two large clades corresponding to the two cell types, while also identifying the subset of sites underlying that separation. Additionally, we observe a number of smaller modules of DHSs that exhibit coordinately regulated chromatin accessibility. Linking these sites again to the genes they potentially regulate (*2*), the major modules are enriched for gene ontology terms consistent with the two cell types (e.g., "osteoclast differentiation" for a module more open in HL-60) (Fig. 3C and figs. S13 and S14).

To evaluate cell-to-cell variation within a cell type, we took the subset of cells classified as GM12878 and repeated latent semantic indexing (*19*), yielding a heat map of chromatin accessibility for 4118 cells at 22,755 DHSs. Hierarchical clustering identified four major subgroups of single cells and seven modules of coordinately regulated chromatin accessibility (Fig. 4A). These modules of DHSs are enriched for binding by particular transcription factors (hypergeometric FDR 0.10) (fig. S15), in some cases quite strongly, and are linked to genes associated with immune response, cell cycle regulation, and other processes (figs. S16 and S17). Importantly, although we included samples from experiments conducted on different days, the cell subtypes do not cluster by experiment (figs. S18 and S19), and the enrichments for transcription factor binding within subtype-defining modules are apparent even with subsets of the data (figs. S20 and S21). Sites in modules 1 and 2 are highly enriched for binding by transcription factors such as nuclear factor κB (NF-κB) and other factors downstream of the B cell receptor (*19*). The four GM12878 subtypes appear principally defined by the activation status of these two modules, suggesting that variability across the cells is driven by NF-κB activity. These results indicate that even within an apparently homogeneous cell type, we are able to identify subsets of cells with differences in their regulatory landscape related to cell cycle and possibly environmental signals. Focusing on individual loci within GM12878,

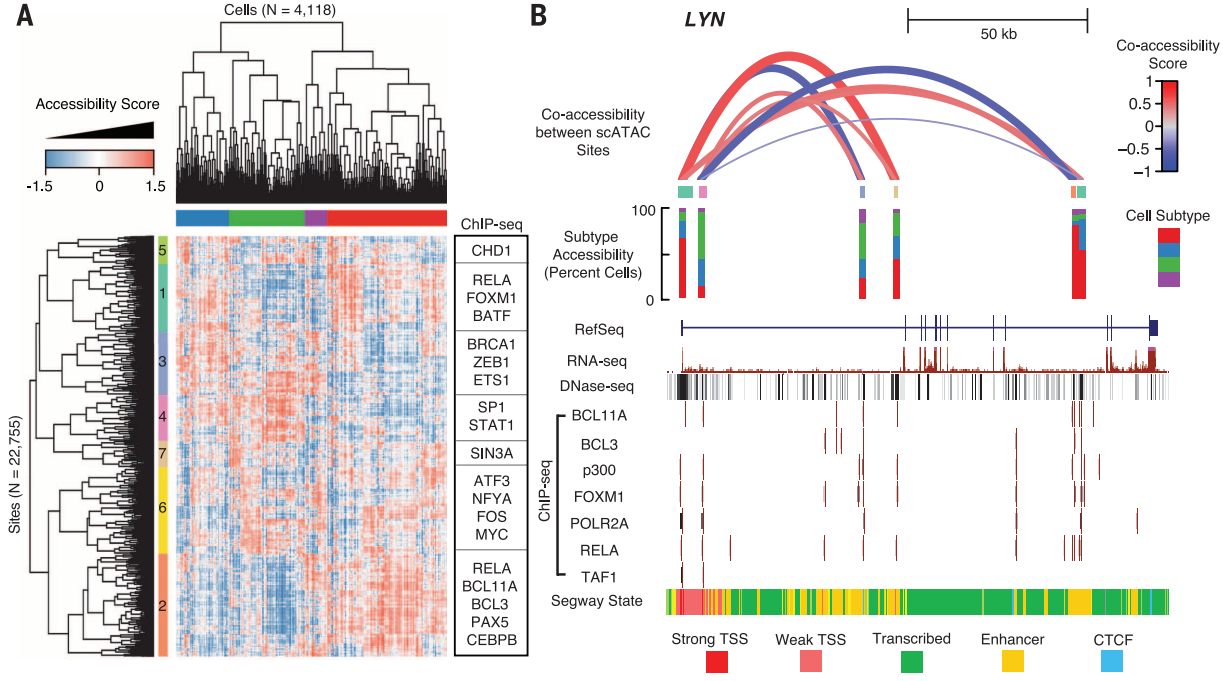

Fig. 4. Single-cell ATAC-seq identifies GM12878 subtypes. (**A**) Heat map of chromatin accessibility measures after latent semantic indexing of DHS usage shows that GM12878 cells cluster into subpopulations. Modules of coordinately accessible chromatin accessibility are significantly enriched for binding of selected transcription factors (TFs) (examples on right). (**B**) Detailed depiction of *LYN* locus. The top shows coaccessibility scores between the transcription start sites and four putative enhancers in the region, which are Pearson correlation values of latent semantic indexing–based accessibility scores between cells, for six DHSs present in this region. Height and thickness of each loop indicates the strength of correlation (red, positive; blue, negative). Middle shows in which subtypes [defined in top bar of (A)] these elements are most often accessible. Bottom shows ENCODE data for this region from the University of California–Santa Cruz browser, including transcript model, DHS peaks, chromatin immunoprecipitation sequencing (ChIP-seq) binding profiles for several TFs, and predicted chromatin state.

we observe sets of regulatory sites that exhibit patterns of coordinated regulation (e.g., *LYN*, encoding a tyrosine kinase involved in B cell signaling) (Fig. 4B), although reproducibility of these patterns across biological replicates was modest (fig. S22). Given the sparsity of the data, identifying pairs of coaccessible DNA elements within individual loci is statistically challenging and merits further development.

We report chromatin accessibility maps for >15,000 single cells. Our combinatorial cellular indexing scheme could feasibly be scaled to collect data from ~17,280 cells per experiment by using 384-by-384 barcoding and sorting 100 nuclei per well (assuming similar cell recovery and collision rates) (fig. S1) (*19*). Particularly as large-scale efforts to build a human cell atlas are contemplated (*23*), it is worth noting that because DNA is at uniform copy number, single-cell chromatin accessibility mapping may require far fewer reads per single cell to define cell types, relative to single-cell RNA-seq. As such, this method's simplicity and scalability may accelerate the characterization of complex tissues containing myriad cell types, as well as dynamic processes such as differentiation.

REFERENCES AND NOTES

1. A. B. Stergachis *et al.*, *Cell* **154**, 888–903 (2013).
2. R. E. Thurman *et al.*, *Nature* **489**, 75–82 (2012).
3. A. P. Boyle *et al.*, *Cell* **132**, 311–322 (2008).
4. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, *Nat. Methods* **10**, 1213–1218 (2013).
5. N. Navin *et al.*, *Nature* **472**, 90–94 (2011).
6. A. R. Wu *et al.*, *Nat. Methods* **11**, 41–46 (2014).
7. D. A. Jaitin *et al.*, *Science* **343**, 776–779 (2014).
8. Q. Deng, D. Ramsköld, B. Reinius, R. Sandberg, *Science* **343**, 193–196 (2014).
9. A. K. Shalek *et al.*, *Nature* **498**, 236–240 (2013).
10. C. Trapnell *et al.*, *Nat. Biotechnol.* **32**, 381–386 (2014).
11. S. A. Smallwood *et al.*, *Nat. Methods* **11**, 817–820 (2014).
12. T. Nagano *et al.*, *Nature* **502**, 59–64 (2013).
13. J. Gole *et al.*, *Nat. Biotechnol.* **31**, 1126–1132 (2013).
14. H. C. Fan, G. K. Fu, S. P. A. Fodor, *Science* **347**, 1258367 (2015).
15. A.-E. Saliba, A. J. Westermann, S. A. Gorski, J. Vogel, *Nucleic Acids Res.* **42**, 8845–8860 (2014).
16. X. Pan, *Single Cell Biol.* **3**, 106 (2014).
17. A. Adey *et al.*, *Genome Res.* **24**, 2041–2049 (2014).
18. S. Amini *et al.*, *Nat. Genet.* **46**, 1343–1349 (2014).
19. Materials and methods are available as supplementary materials on *Science* Online.
20. A. Adey *et al.*, *Genome Biol.* **11**, R119 (2010).
21. F. Yang, T. Babak, J. Shendure, C. M. Disteche, *Genome Res.* **20**, 614–622 (2010).
22. The ENCODE Project Consortium, *Nature* **489**, 57–74 (2012).
23. A. Regev, The Human Cell Atlas; www.genome.gov/Multimedia/Slides/GSPFuture2014/10_Regev.pdf.

ACKNOWLEDGMENTS

We thank the Trapnell and Shendure laboratories, particularly R. Hause, C. Lee, V. Ramani, R. Qiu, Z. Duan, and J. Kitzman, for helpful discussions; D. Prunkard, J. Fredrickson, and L. Gitari in the Rabinovitch laboratory for their exceptional assistance in flow sorting; and C. Disteche for the Patski cell line. This work was funded by an NIH Director's Pioneer Award (1DP1HG007811 to J.S.) and a grant from the Paul G. Allen Family Foundation (J.S.). C.T. is supported in part by the Damon Runyon Cancer Research Foundation (DFS-#10-14). All sequencing data are available from the NIH National Center for Biotechnology Information Gene Expression Omnibus (accession number GSE68103). L.C., K.L.G., and F.J.S. declare competing financial interests in the form of stock ownership and paid employment by Illumina, Inc. One or more embodiments of one or more patents and patent applications filed by Illumina may encompass the methods, reagents, and data disclosed in this manuscript. All methods for making the transposase complexes are described in (*18*); however, Illumina will provide transposase complexes in response to reasonable requests from the scientific community subject to a material transfer agreement. Some work in this study is related to technology described in patent applications WO2014142850, 2014/0194324, 2010/0120098, 2011/0287435, 2013/0196860, and 2012/0208705.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/910/suppl/DC1
Materials and Methods
Figs. S1 to S22
Tables S1 and S2
References (*24*–*39*)

19 March 2015; accepted 24 April 2015
Published online 7 May 2015;
10.1126/science.aab1601

VIROLOGY

A virus that infects a hyperthermophile encapsidates A-form DNA

Frank DiMaio,[1]* Xiong Yu,[2]* Elena Rensen,[3] Mart Krupovic,[3] David Prangishvili,[3]† Edward H. Egelman[2]†

Extremophiles, microorganisms thriving in extreme environmental conditions, must have proteins and nucleic acids that are stable at extremes of temperature and pH. The nonenveloped, rod-shaped virus SIRV2 (*Sulfolobus islandicus* rod-shaped virus 2) infects the hyperthermophilic acidophile *Sulfolobus islandicus*, which lives at 80°C and pH 3. We have used cryo–electron microscopy to generate a three-dimensional reconstruction of the SIRV2 virion at ~4 angstrom resolution, which revealed a previously unknown form of virion organization. Although almost half of the capsid protein is unstructured in solution, this unstructured region folds in the virion into a single extended α helix that wraps around the DNA. The DNA is entirely in the A-form, which suggests a common mechanism with bacterial spores for protecting DNA in the most adverse environments.

Extreme geothermal environments, with temperatures above 80°C, are the habitat of hyperthermophilic DNA viruses that parasitize Archaea (*1*). These viruses have more than 92% of genes without homologs in databases (*2*, *3*), distinct protein folds (*4*), and distinct mechanisms of viral egress (*5*). The high diversity of virion morphotypes may underpin virion morphogenesis and DNA packaging, which could determine the high stability of the virions. Viruses from the family *Rudiviridae* (*6*) consist of a nonenveloped, helically arranged nucleoprotein composed of double-stranded DNA (dsDNA) and thousands of copies of a 134-residue protein. To understand the mechanisms stabilizing rudiviral DNA in natural habitats of host cells, which involve high temperatures (~80°C) and low pH values (~pH 3), we used cryo–electron microscopy (cryo-EM) to analyze the rudivirus SIRV2 (*Sulfolobus islandicus* rod-shaped virus 2) (*6*), which infects the hyperthermophilic acidophilic archaeon *Sulfolobus islandicus* (*7*) (see supplementary materials and methods). Members of the archaeal genus *Sulfolobus* maintain their cytoplasmic pH neutral at pH 5.6 to 6.5 (*8*, *9*). SIRV2 is therefore exposed to a wide range of pH values: from about pH 6 in the cellular cytoplasm, where it assembles and maturates (*10*), to pH 2 to 3 in the extracellular environment. We performed our studies at pH 6. SIRV2 is stable over a wide range of temperatures: from −80°C, the temperature at which the virus can be stored for years without loss of infectivity, to 80°C, the temperature of its natural environment. The overall morphology of the virion is maintained, regardless of the use of negative-stain imaging at 75°C (*11*) or cryo-EM with a sample at 4°C before vitrification (Fig. 1A).

Electron cryo-micrographs of SIRV2 (Fig. 1A) showed strong helical striations in most of the virions with a periodicity of 42 Å. We performed three-dimensional (3D) reconstruction using the iterative helical real space reconstruction method (*12*), after first determining the helical symmetry. Only one solution (with 14.67 subunits per turn of the 42 Å pitch helix) yielded a reconstruction with recognizable secondary structure, almost all α helical, and a resolution of ~3.8 Å in the more-ordered interior, which surrounds the DNA (fig. S2). The asymmetric unit was a symmetrical dimer, the α helices of which were wrapping around a continuous dsDNA. The DNA

[1]Department of Biochemistry, University of Washington, Seattle, WA 98195, USA. [2]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908, USA. [3]Institut Pasteur, Department of Microbiology, 25 rue du Dr. Roux, Paris 75015, France.
*These authors contributed equally to this work. **†Corresponding author. E-mail: egelman@virginia.edu (E.H.E.); david.prangishvili@pasteur.fr (D.P.)**