Massively multiplex single-cell Hi-C

Vijay Ramani[1], Xinxian Deng[2], Ruolan Qiu[1], Kevin L Gunderson[3], Frank J Steemers[3], Christine M Disteche[2,4], William S Noble[1], Zhijun Duan[5,6] & Jay Shendure[1,7]

We present single-cell combinatorial indexed Hi-C (sciHi-C), a method that applies combinatorial cellular indexing to chromosome conformation capture. In this proof of concept, we generate and sequence six sciHi-C libraries comprising a total of 10,696 single cells. We use sciHi-C data to separate cells by karyotypic and cell-cycle state differences and identify cell-to-cell heterogeneity in mammalian chromosomal conformation. Our results demonstrate that combinatorial indexing is a generalizable strategy for single-cell genomics.

Our understanding of genome architecture has progressed through the successive development of new technologies[1]. Advances in microscopy revealed the presence of 'chromosome territories', nuclear regions that preferentially self associate[2]. The advent of chromosome conformation capture (3C) and its derivatives[3] has resulted in a proliferation of data measuring genome architecture and its relation to other aspects of nuclear biology.

3C assays rely on the concept of proximity ligation, a technique that has been used to measure local protein–protein[4], RNA–RNA[5], and DNA–DNA interactions[6]. By coupling an 'all-vs-all' 3C assay with massively parallel sequencing techniques[7,8] (e.g., Hi-C), one can query the relative contact probabilities of DNA genome wide. However, contact probabilities generated by these assays represent ensemble averages of the respective conformations of the millions of nuclei used as input, and scalable techniques characterizing the variance underlying these population averages remain largely underdeveloped. A pioneering study in 2013 demonstrated proof of concept that Hi-C could be performed on single isolated mouse nuclei but relied on the physical separation and processing of single murine cells in independent reaction volumes, with consequent low throughput[9].

The repertoire of high-throughput single-cell techniques for other biochemical assays has expanded rapidly as of late[10–13]. Single-cell RNA-seq (scRNA-seq) was recently paired with droplet-based microfluidics to markedly increase its throughput[11,12]. Orthogonally, we introduced the concept of combinatorial cellular indexing[10]—a method that eschews microfluidic manipulation and instead tags DNA within intact nuclei with successive (combinatorial) rounds of nucleic acid barcodes—to measure chromatin accessibility in thousands of single cells without physically isolating each single cell (single-cell combinatorial indexed ATAC-seq, or sciATAC-seq). Such throughput-boosting strategies have yet to be successfully adapted for single-cell chromosome conformation analysis.

To address this methodological gap, we have developed a high-throughput single-cell Hi-C protocol, termed single-cell combinatorial indexed Hi-C, or sciHi-C (**Fig. 1a**), based on the concept of combinatorial indexing and also building on recent improvements to the Hi-C protocol[14,15] (see Online Methods). A population of 5 to 10 million cells is fixed, lysed to generate nuclei, and restriction digested *in situ* with the enzyme DpnII. Nuclei are then distributed to 96 wells, wherein the first barcode is introduced through ligation of barcoded biotinylated double-stranded bridge adaptors. Intact nuclei are then pooled and subjected to proximity ligation all together, followed by dilution and redistribution to a second 96-well plate. Importantly, this dilution is carried out such that each well in this second plate contains at most 25 nuclei. Following lysis, a second barcode is introduced through ligation of barcoded Y-adaptors.

As the number of barcode combinations (96 × 96) exceeds the number of nuclei (96 × 25), the vast majority of single nuclei are tagged by a unique combination of barcodes. All material is once again pooled, and biotinylated junctions are purified with streptavidin beads, restriction digested, and further processed to Illumina sequencing libraries. Sequencing these molecules with relatively long paired-end reads (i.e., 2 × 250 base pairs (bp)) allows one to identify not only the genome-derived fragments of conventional Hi-C, but also external and internal barcodes (each combination of which is hereafter referred to as a 'cellular index') which enable decomposition of the Hi-C data into single-cell contact probability maps (**Fig. 1b**). Like sciATAC-seq[10], this protocol can process thousands of cells per experiment without requiring the physical isolation of each cell.

As a proof of concept, we applied sciHi-C to synthetic mixtures of cell lines derived from mouse (primary mouse embryonic fibroblasts (MEFs), and the 'Patski' embryonic fibroblast line) and human cells (HeLa S3, the HAP1 cell line, K562, and GM12878; all five experiments and sequenced libraries are summarized in **Supplementary Table 1**, although we focus on ML1 and ML2 biological replicates in the text). All experiments were carried out such that subsets of cell types received specific barcodes during

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Department of Pathology, University of Washington, Seattle, Washington, USA. [3]Illumina Inc., Advanced Research Group, San Diego, California, USA. [4]Department of Medicine, University of Washington, Seattle, Washington, USA. [5]Division of Hematology, University of Washington School of Medicine, Seattle, Washington, USA. [6]Institute for Stem Cell and Regenerative Medicine, University of Washington, Seattle, Washington, USA. [7]Howard Hughes Medical Institute, Seattle, Washington, USA. Correspondence should be addressed to Z.D. (zjduan@uw.edu) or J.S. (shendure@uw.edu).

RECEIVED 25 JULY 2016; ACCEPTED 14 DECEMBER 2016; PUBLISHED ONLINE 30 JANUARY 2017; CORRECTED BEFORE PRINT 10 FEBRUARY 2017 (DETAILS ONLINE); DOI:10.1038/NMETH.4155

the first round of barcoding (e.g., in ML1 and ML2, each well during the first round of barcoding contained either HeLa S3 + Patski cells or HAP1 + MEF cells; see Online Methods).

Before deconvolving the resulting data to single cells, we examined the overall distribution of ligation junctions (i.e., contacts). Encouragingly, there were very few contacts between mouse and human cells (ML1, 0.006%; ML2, 0.008%), demonstrating minimal crosstalk between cellular indices and indicating that nuclei remain intact through all ligation steps (confirmed through phase-contrast microscopy; **Supplementary Fig. 1**). We also examined the *cis*:*trans* ratio, defined here as the ratio of long-range (i.e., >20 kb) intrachromosomal contacts to interchromosomal contacts (**Fig. 1c**), and found it to be on par with expectation for high-quality Hi-C data sets (ML1, 4.41; ML2, 4.38).

We next split the Hi-C data to characterize the number of unique read pairs associated with each cellular index, the vast majority of which should correspond to single cells. When examining a histogram of unique index occurrences as a function of read depth, we noted a bimodal distribution reminiscent of patterns seen in sciATAC-seq data sets[10], where low-coverage indices likely represent 'noise' consequent to tags from free DNA in solution (**Supplementary Fig. 2**). After discarding these, we inferred 1,081 cellular indices in ML1, with a median of 9,274 unique read pairs per index (ML2, 841 cellular indices; median of 8,335 unique read pairs per index). Importantly, we also observed minimal barcode bias across replicate experiments (**Supplementary Fig. 3**) as well as similar median *cis*:*trans* ratios per cell (ML1, 4.43 with median absolute deviation (MAD) of 1.66; ML2, 4.34 with MAD of 1.66) (**Fig. 1d** and **Supplementary Fig. 4**).

The only previously published study of single-cell Hi-C data suggests that high single-cell *cis*:*trans* ratios are a hallmark of high-quality single-cell data[9]. The high *cis*:*trans* ratios that we observed are comparable to those of the ten single-cell maps generated in that study, which reported a median value of 6.26 (MAD = 0.74), calculated as the ratio of all intrachromosomal contacts to interchromosomal contacts (i.e., with no cutoff for minimal intrachromosomal distance). Reanalyzing our own data using this more liberal criterion yielded similar ratios of 6.17 (ML1; MAD = 1.99) and 5.96 (ML2; MAD = 1.94). Of note, our ratios are calculated over 1,922 cellular indices (ML1 and ML2 combined), 857 of which have more than 10,000 unique contacts, compared with the ten previously reported single cells—each with at least 10,000 unique contacts. This comparison illustrates the greater scalability of combinatorial methods

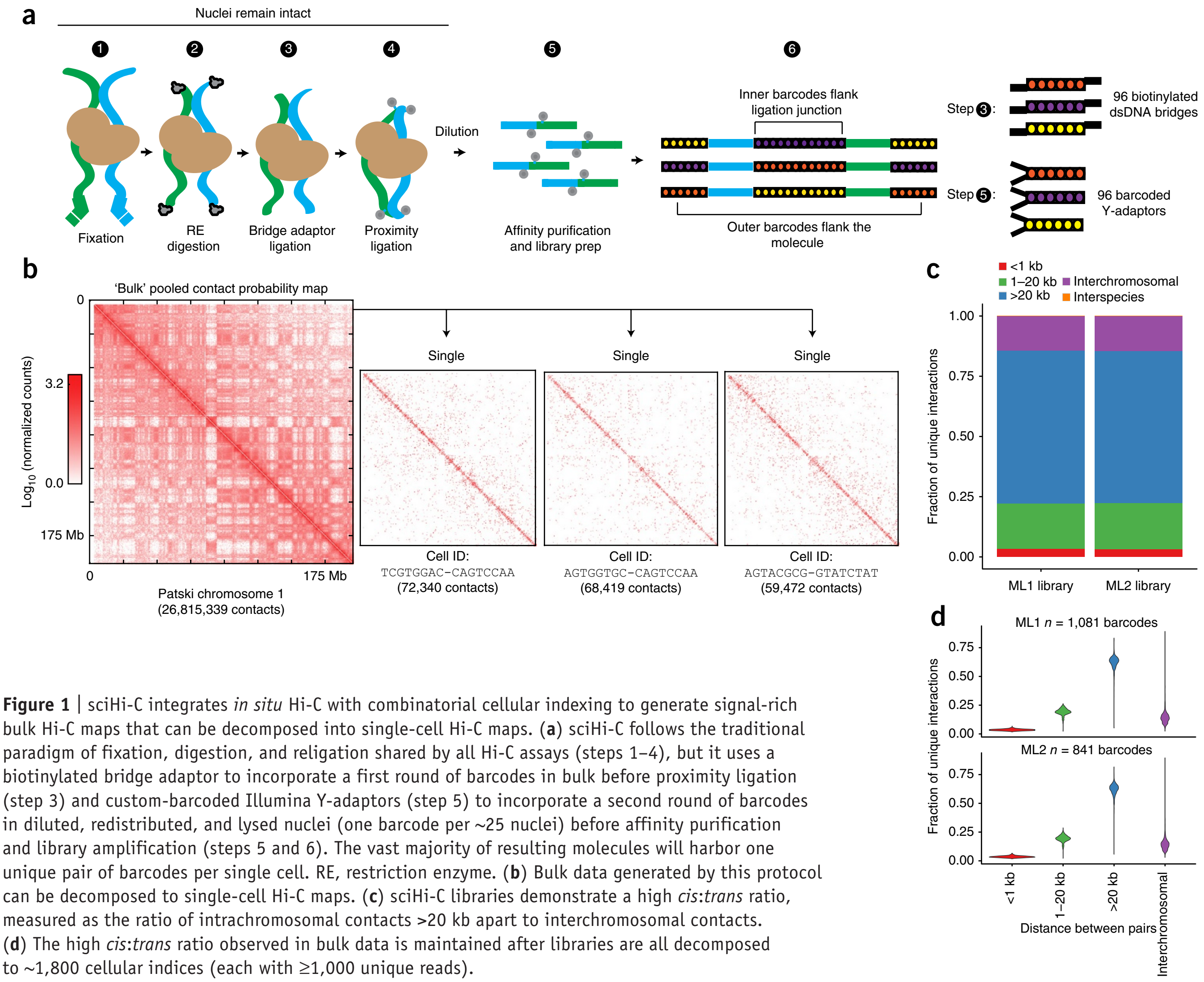

Figure 1 | sciHi-C integrates *in situ* Hi-C with combinatorial cellular indexing to generate signal-rich bulk Hi-C maps that can be decomposed into single-cell Hi-C maps. (**a**) sciHi-C follows the traditional paradigm of fixation, digestion, and religation shared by all Hi-C assays (steps 1–4), but it uses a biotinylated bridge adaptor to incorporate a first round of barcodes in bulk before proximity ligation (step 3) and custom-barcoded Illumina Y-adaptors (step 5) to incorporate a second round of barcodes in diluted, redistributed, and lysed nuclei (one barcode per ~25 nuclei) before affinity purification and library amplification (steps 5 and 6). The vast majority of resulting molecules will harbor one unique pair of barcodes per single cell. RE, restriction enzyme. (**b**) Bulk data generated by this protocol can be decomposed to single-cell Hi-C maps. (**c**) sciHi-C libraries demonstrate a high *cis*:*trans* ratio, measured as the ratio of intrachromosomal contacts >20 kb apart to interchromosomal contacts. (**d**) The high *cis*:*trans* ratio observed in bulk data is maintained after libraries are all decomposed to ~1,800 cellular indices (each with ≥1,000 unique reads).

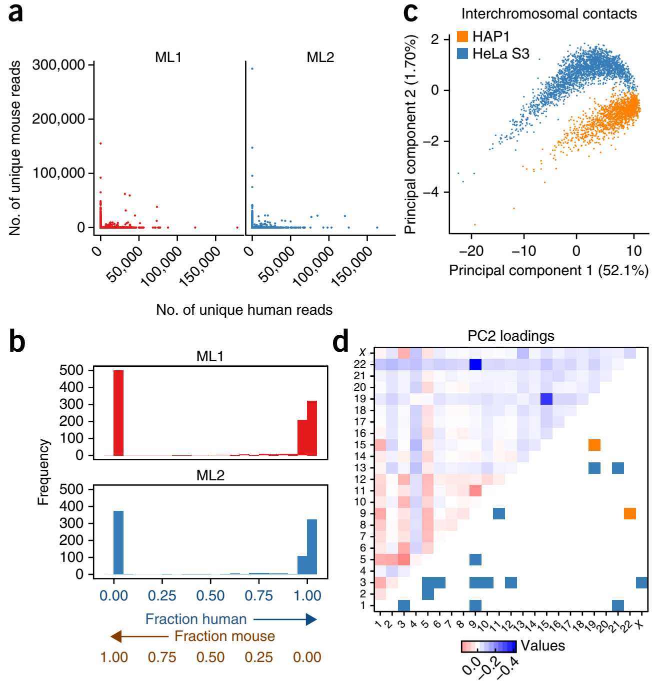

Figure 2 | Cellular indices generated through sciHi-C are overwhelmingly species specific and can be separated by cell type. (**a**) In libraries ML1 and ML2, similar levels of collision (defined as any cellular index with at least 1,000 unique reads but <95% species purity) are observed, and they fall within the expected range. (**b**) Species contamination visualized as a histogram of the fraction of reads mapping to the human genome (only cellular indices with ≥ 1,000 reads shown). (**c**) Projection onto the first two principal components from PCA analysis of interchromosomal contact matrices results in separation of HeLa S3 and HAP1, two karyotypically different cell lines (n = 3,609 cells). Percentages shown are the percentage of variance explained by each plotted component. (**d**) Principal component 2 loadings represent the contribution of each feature (interchromosomal contact) to the observed cell-type separation. Known translocations for each cell type are mirrored against the loading heatmap.

compared with that of methods relying on the physical isolation and serial processing of each single cell.

We designed our experiments to facilitate validation of the single-cell origin of each cellular index. Uniquely tagged cells should be associated with species-specific cellular indices in mixture experiments, with a collision rate broadly defined by a formulation of the 'birthday problem'[10]. Consistent with the expected collision rate, we observed that 4.53% of all ML1 cellular indices (4.40% in ML2) were 'collisions' (i.e., they had less than 95% of reads mapping to either the mouse or human genome) (**Fig. 2a**,**b**). For further analyses we filtered out any cellular indices failing this criterion, while accepting that we remain blind to 'within species' collisions, which likely exist at a similar fraction to that of interspecies collisions. We also filtered out indices where the associated *cis*:*trans* ratio was less than 1 (1.94% of indices in ML1; 1.62% in ML2), which could suggest broken nuclei.

Before continuing, we combined filtered data from ML1 and ML2 with equivalently filtered data from secondary experiments (PL1 and PL2) (**Supplementary Table 1** and **Supplementary Fig. 5**). We then employed a conservative genotype filter[16], which removed 20.4% of human cellular indices (**Supplementary Fig. 6**), leaving us with a combined data set of 3,609 human single-cell Hi-C maps. Together with mouse data (which were filtered for coverage, *cis*:*trans* ratio, and species purity), a total of 8,141 single-cell Hi-C maps were generated across these four experiments.

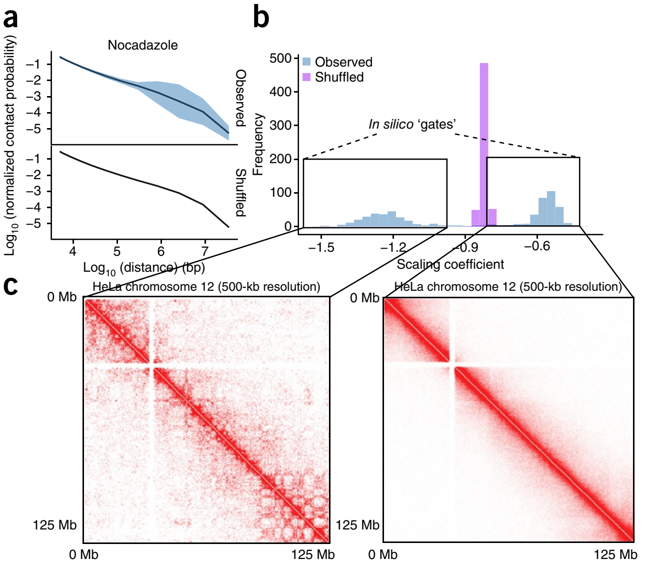

Figure 3 | sciHi-C of nocadazole-arrested HeLa S3 cells enable *in silico* sorting by cell-cycle progression. (**a**) Mean contact probability and s.d. as a function of genomic distance for single HeLa S3 cells from a population treated with nocadazole (n = 588 cells containing at least 5,000 contacts and harboring nocadazole-experiment-specific programmed barcodes), as well as shuffled controls generated by random reassignment of cellular indices. (**b**) Scaling coefficients for 588 single HeLa S3 cells follow a bimodal distribution. (**c**) Cells can be 'sorted' *in silico* to generate two distinct contact probability maps, shown here for HeLa chromosome 12. The labels of both axes indicate chromosomal position.

We next explored whether cell types could be separated *in silico* on the basis of single-cell Hi-C signal. We generated matrices where rows represent single cells and columns represent the number of contacts between pairs of chromosomes (**Supplementary Fig. 7**). Principal components analysis (PCA) on this matrix resulted in separation of single HeLa S3 and HAP1 cells (**Fig. 2c**), which was validated by our programmed barcode associations. Principal component 1 (PC1), which strongly correlated with coverage (**Supplementary Fig. 8**), accounted for the majority of the variance (52.1%), while the combination of PC1 and principal component 2 (PC2; 1.07% of the variance) separated HeLa S3 and HAP1 cells. We then analyzed the 'loadings' of our features in PC2, the axis separating HeLa S3 and HAP1 cells, and found that the strongest loadings recapitulated known translocations specific to HAP1 (ref. 17) (namely, translocations between chromosomes 15 and 19, and between chromosomes 9 and 22), while other strong loadings corresponded to documented HeLa S3 translocations[16,18] (**Fig. 2d**). Repeating these analyses by (i) removing specific interactions from the matrices and repeating PCA (**Supplementary Fig. 9**), (ii) using an alternate feature set (interacting 10-Mb intrachromosomal windows; **Supplementary Figs. 7b** and **10**), (iii) separating cells by replicate (**Supplementary Fig. 11**), and (iv) sequencing 908 additional human cells (K562 and GM12878; library ML3 containing 1,175 cells total; **Supplementary Fig. 12**); all steps recapitulated cell-type separation to varying degrees, demonstrating that PCA could potentially be used to separate cell types on the basis of Hi-C signal. The ability to separate such populations could be invaluable, for example, when studying tissue containing a mixture of normal cells and cancerous cells harboring translocations.

We next examined the heterogeneity present in single-cell Hi-C maps in terms of polymer conformation. We plotted contact probability as a function of genomic distance for 769 single cells, each with at least 10,000 unique contacts (**Supplementary Fig. 13a**), finding that contact probability values observed for single cells were markedly more disperse compared with those calculated from a set of shuffled control 'cells', regardless of species analyzed. We then examined the relationship between single-cell power-law scaling coefficients (**Supplementary Fig. 13b**), calculated between distances of 50 kb and 8 Mb[19,20], and single-cell *cis*:*trans* ratios, noting a correlation across four out of five experiments (**Supplementary Fig. 13c** and **Supplementary Fig. 14**) between high *cis*:*trans* ratios and shallow scaling coefficients.

To test whether this variance was related to the relative cell-cycle state of single cells, we arrested HeLa S3 cells using nocadazole, an agent that leads to an enrichment of cells arrested at G2–M phase, and we performed sciHi-C on this population (library ML4; $n = 1,380$ filtered cells). Repeating the above analysis on this data set yielded a strikingly wide variance in single-cell contact probability decay (**Fig. 3a**), and subsequent calculation of scaling coefficients revealed a clear bimodal distribution in the data (**Fig. 3b**). We then performed *in silico* 'sorting' of this data to decompose the aggregate data set into two distinct contact probability maps (**Fig. 3c**), one harboring the 'plaid' compartment pattern expected of interphase chromatin, and another harboring the condensed, compartment-free patterning of mitotic chromatin previously described by Naumova *et al.*[18]. As a control, untreated cells were processed simultaneously (data not shown). Our demonstration of *in silico* cell sorting, as well as the empirical distributions for scaling coefficient in single cycling mouse and human cells, are likely to be highly useful in constraining computational models of mammalian chromosome conformation.

We have shown that sciHi-C is an effective method for profiling chromosome conformation in single cells that relies on combinatorial cellular indexing for rapid scaling to large numbers of cells. As a proof of concept, we applied this method to generate single-cell Hi-C maps for 10,696 cells with at least 1,000 unique contacts. This data set is two orders of magnitude larger than the only published single-cell Hi-C data set, with 3,515 filtered cells containing more than 10,000 unique contacts, compared with the ten existing single-cell maps defined using a similar coverage cutoff.

Given the generally similar workflow of our method and traditional bulk Hi-C, it may be possible to incorporate sciHi-C into routine practice, thus adding a 'single-cell' dimension to Hi-C data production and a means of obtaining single-cell and bulk measurement at once (the latter generated by summing single cells). Furthermore, our demonstration that thousands of single-cell Hi-C maps can be generated in a single workflow, without the need to isolate each cell, demonstrates the power of combinatorial indexing for large-scale single-cell biology. Indeed, as Vitak *et al.*[21] also show in this issue, combinatorial indexing is thus generalizable to additional aspects of single-cell or even intracellular biology, where DNA barcodes can be incorporated *in situ*.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank S. Kasinathan, members of the UW Center for Nuclear Organization and Function, and members of the Shendure lab (particularly M. Kircher), for helpful discussions. HeLa S3 cells were used as part of this study. Henrietta Lacks, and the HeLa cell line that was established from her tumor cells in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. Primary MEF aliquots were a gift from C. Ware (University of Washington), HeLa S3 aliquots were a gift from the Malik Lab (Fred Hutchinson Cancer Research Center), and Patski cell aliquots were a gift from the Disteche lab (University of Washington). This work was funded by grants from the NIH (5T32HG000035 to V.R.; DP1HG007811 and 5R01HG006283 to J.S.; and U54DK107979 to X.D., C.M.D., W.S.N., Z.D., and J.S.). J.S. is an Investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

V.R., Z.D., and J.S. conceived of the project. V.R., X.D., R.Q., and Z.D. carried out experiments. C.M.D. and W.S.N. provided invaluable critical input. K.L.G. and F.J.S. were part of initial discussions on novel approaches to single-cell Hi-C. V.R., Z.D., and J.S. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Ramani, V., Shendure, J. & Duan, Z. *Genomics Proteomics Bioinformatics* **14**, 7–20 (2016).
2. Cremer, T. & Cremer, C. *Nat. Rev. Genet.* **2**, 292–301 (2001).
3. van Steensel, B. & Dekker, J. *Nat. Biotechnol.* **28**, 1089–1095 (2010).
4. Söderberg, O. *et al. Nat. Methods* **3**, 995–1000 (2006).
5. Ramani, V., Qiu, R. & Shendure, J. *Nat. Biotechnol.* **33**, 980–984 (2015).
6. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. *Science* **295**, 1306–1311 (2002).
7. Lieberman-Aiden, E. *et al. Science* **326**, 289–293 (2009).
8. Duan, Z. *et al. Nature* **465**, 363–367 (2010).
9. Nagano, T. *et al. Nature* **502**, 59–64 (2013).
10. Cusanovich, D.A. *et al. Science* **348**, 910–914 (2015).
11. Klein, A.M. *et al. Cell* **161**, 1187–1201 (2015).
12. Macosko, E.Z. *et al. Cell* **161**, 1202–1214 (2015).
13. Rotem, A. *et al. Nat. Biotechnol.* **33**, 1165–1172 (2015).
14. Rao, S.S.P. *et al. Cell* **159**, 1665–1680 (2014).
15. Deng, X. *et al. Genome Biol.* **16**, 152 (2015).
16. Adey, A. *et al. Nature* **500**, 207–211 (2013).
17. Essletzbichler, P. *et al. Genome Res.* **24**, 2059–2065 (2014).
18. Naumova, N. *et al. Science* **342**, 948–953 (2013).
19. Imakaev, M. *et al. Nat. Methods* **9**, 999–1003 (2012).
20. Sanborn, A.L. *et al. Proc. Natl. Acad. Sci. USA* **112**, E6456–E6465 (2015).
21. Vitak *et al. Nat. Methods* http://dx.doi/10.1038/NMETH.4154 (2017).

ONLINE METHODS

Cell culture. HeLa S3 (CCL2.2) (gift from Malik Lab), primary MEFs (gift from Ware Lab), and Patski (gift from Disteche lab) cells were cultured at 37 °C, 5% CO_2 in DMEM supplemented with 1× Pen–Strep (Gibco) and 10% fetal bovine serum (FBS; Gibco). HAP1 cells (Haplogen) were cultured at 37 °C, 5% CO_2 in IMDM supplemented with 1× Pen–Strep and 10% FBS. K562 cells were cultured at 37 °C, 5% CO_2 in RPMI-1640 supplemented with 1× Pen–Strep and 10% FBS. GM12878 cells were cultured at 37 °C, 5% CO_2 in RPMI-1640 supplemented with 1× Pen–Strep and 15% FBS. Cells were not tested for mycoplasma.

Cell fixation. Adherent cells (i.e., HeLa S3, HAP1, Patski, MEF) were washed once with 1× PBS (Life Technologies), trypsinized (0.25% trypsin–EDTA, Life Technologies), spun down at 500× g for 5 min, and resuspended in 20 mL serum-free DMEM (IMDM for HAP1). Cells were crosslinked by adding 1.12 mL (2% final concentration, for HeLa S3, HAP1, and MEF) or 1.4 mL (2.5% final concentration, for Patski) 37% formaldehyde (Alcon) and incubated at RT (25 °C) for 10 min, after which crosslinking was quenched using 1 mL 2.5 M glycine. Quenched reactions were incubated on ice for 15 min, spun down at 800× g for 5 min, resuspended in 1× PBS, aliquoted into 10 $\times 10^6$ cell aliquots, pelleted once again at 800× g for 5 min, decanted, snap frozen in liquid nitrogen, and finally stored indefinitely at −80 °C.

Suspension cells (i.e., K562 and GM1878) were spun down at 500× g for 5 min, resuspended in 20 mL serum-free RPMI-1640, crosslinked with a final concentration of 2% formaldehyde, and processed as above.

For nocadazole-arrest experiments, we plated HeLa S3 cells in T75 flasks to ~10% confluency. 24 h later, we replaced media with DMEM containing 10% FBS and nocadazole to a final concentration of 100 ng/mL. We then waited 24 h, then we harvested cells by first harvesting detached cells, then trypsinizing the remaining plated cells. This resulted in a heterogeneous single-cell suspension, which we then fixed as above using 2% formaldehyde.

Single-cell combinatorial indexed Hi-C. For the step-by-step combinatorial single-cell combinatorial indexed Hi-C (sciHi-C) protocol, see **Supplementary Protocol** and the Protocol Exchange[22]. Like the recently published scDNase-seq protocol[23], sciHi-C uses carrier plasmid to prevent DNA losses during steps of the protocol where small amounts of DNA are handled. The libraries prepared here each used fixed aliquots of 5 to 10 million cells, which are diluted over the course of the protocol.

All oligonucleotide sequences used in this study were obtained from IDT Technologies (see **Supplementary Data**). All libraries were sequenced on a HiSeq 2500.

Barcode programming. Our primary data sets (library ML1 and biological replicate library ML2) used HeLa S3, HAP1, Patski, and MEFs, with subsets of human and mouse cell types in distinct wells during the first round of barcoding (HeLa S3 + Patski in half of wells; HAP1 + MEFs in half of wells). Our secondary data sets (library PL1 and biological replicate PL2) were generated using the same cell types but a subtly different programming scheme (illustrated in **Supplementary Fig. 15**), wherein each well contained only a single cell type during the first round of barcoding. Finally, we generated and lightly sequenced a fifth library (library ML3), mixing the same murine cell types as before with two new human cell types—GM12878 and K562—in a similar manner to that of sequencing libraries ML1 and ML2 (GM12878 + Patski in half of wells; K562 + MEFs in half of wells).

Bridge adaptor barcode design. Bridge adaptor barcodes were drawn from randomly generated 8-mers, such that the following criteria were met: (i) all adaptors must have a minimum pairwise Levenshtein distance of 3; (ii) adaptors must not contain the sequences TTAA or AAGCTT; (iii) adaptors must contain >60% GC content; (iv) adaptors must not contain homopolymers ≥ length 3; and (v) adaptors must not be palindromic.

Processing sciHi-C data. All code used for sciHi-C data analysis is available as **Supplementary Software** and at https://github.com/VRam142/combinatorialHiC. Below, we describe in detail the analytical pipeline used to process the data. The analytical steps broadly fall under three categories: (i) barcode identification and read trimming; (ii) read alignment, read pairing, and barcode assocation; and (iii) cellular demultiplexing and quality analysis.

Barcode association and read trimming. First, to obtain round 2 (i.e., terminal) barcodes, we use a custom Python script to iterate through both mates, compare the first 8 bases of each read against the 96 known barcode sequences, and then assign barcodes to each mate using a Levenshtein distance cutoff of 2. Reads 'split' in this way are output such that the first 11 bases of each read, which derive from the custom barcoded Y-adaptors, are removed. Mates where either terminal barcode went unidentified, or where the terminal barcodes did not match, are discarded.

For each resulting 'split' pair of reads, the two reads are then scanned using a custom Python script to find the common portion of the bridge adaptor sequence. The 8 bases immediately 5′ of this sequence are isolated and compared against the 96 known bridge adaptor barcodes, again using a Levenshtein distance cutoff of 2. There are cases where the entire bridge adaptor, including both barcodes flanking the ligation junction, is encountered in one mate and not the other. To account for these cases, we also isolate the 8 bases flanking the 3′ end of the common bridge adaptor sequence (when it is encountered within a read), reverse complement it, and compare the resulting 8-mer against the 96 known bridge adaptor barcodes. Output reads are then clipped to remove the bridge adaptor and all 3′ sequence. Barcodes flanking the ligation junction should match; again, mates where barcodes do not match or where a barcode is not found are discarded.

The result of this processing module are three files: filtered reads 1 and 2, and an 'associations' file—a tab-delimited file where the name of each read passing the above filters and their associated barcode combination are listed.

Read alignment, read pairing, and barcode association. As is standard for Hi-C reads, the resulting processed and filtered reads 1 and 2 were aligned separately using bowtie2/2.2.3 to a Burrows–Wheeler index of the concatenated mouse (mm10) and human (hg19) genomes. Individual SAM files were then converted to BED format and filtered for alignments with MAPQ ≥ 30 using a combination of samtools, bedtools, and awk. Using bedtools closest along with a BED file of all DpnII sites in both genomes (generated using HiC-Pro[24]), the closest DpnII site to each read was determined, after which BED files were concatenated, sorted on read ID using UNIX sort, and then processed using a custom

Python script to generate a BEDPE-format file where 5′ mates always precede 3′ mates, and where a simple Python dictionary is used to associate barcode combinations contained in the 'associations' file with each pair of reads. Reads were then sorted by barcode, read 1 chromosome, start, end, read 2 chromosome, start, and end using UNIX sort, and deduplicated using a custom Python script on the following criteria: reads were considered to be PCR duplicates if they were associated with the same cellular index and if they comprised a ligation between the same two restriction sites as defined using bedtools closest.

Cellular demultiplexing and quality analysis. When demultiplexing cells, we run two custom Python scripts. First, we generate a 'percentages' file that includes the species purity of each cellular index, the coverage of each index, and the number of times a particular restriction fragment is observed once, twice, thrice, and four times. We also include the *cis*:*trans* ratio described above, and, if applicable, the fraction of homozygous alternate HeLa alleles observed. We use these percentages files to filter BEDPE files (see below) and generate, at any desired resolution, single-cell matrices in long format (i.e., BIN1-BIN2-COUNT), with only the 'upper diagonal' of the matrix included to reduce storage footprint. These matrices are then converted to numpy matrices for visualization and further analysis.

Filtration of cellular indices. We applied several filters to our resulting cellular indices to arrive at the cells analyzed in this study. We first removed all cellular indices with fewer than 1,000 unique reads. We next filtered out all indices where the *cis*:*trans* ratio was lower than 1. Finally, for all experiments we removed cellular indices where less than 95% of reads aligned uniquely to either the mouse (mm10) or human (hg19) genomes. For all human cells from HAP1 and HeLa S3 mixing experiments (libraries ML1, ML2, PL1, and PL2) further filtration by genotype was performed. For each cellular index, we examined all reads overlapping with known alternate homozygous sites in the HeLa S3 genome and computed the fraction of sites where the alternate allele is observed. We then drew cutoffs to filter out all cells where this fraction fell between 56% and 99%. We employ this filtering step purely as an additional, conservative measure, and note that this is not strictly necessary. The clear separation of two populations in data derived from library ML4 (nocadazole-arrest experiment), where no genotype filtration was performed, illustrates this.

We do acknowledge that particular applications (e.g., structural modeling) may require more stringent filtration for cellular indices covering single cells. As such, we provide with the raw data files specifying the 'species purity' of each barcode combination in each sequenced library, along with the number of times DpnII restriction fragments are observed in a cell once, twice, thrice, or four times, with the expectation that given some tolerable noise level, one should only observe restriction fragment copy numbers equal to or less than the copy number of that fragment for that cell type. Relatedly, we note that further inspection of the HAP1 cells used in this study revealed that they were not entirely haploid. HAP1 cells, an engineered haploid line, have faster doubling times compared with those of HeLa S3, and have been described as having a relatively large frequency of diploid cells[25]. FACS analysis (data not shown) of the stock used for these experiments showed that ~40% of cells analyzed harbored 2N nucleic acid content, indicating haploid cells in G2 or reverted diploid cells in G1.

Data analysis. *PCA of sciHi-C data.* Single-cell matrices at interchromosomal contact resolution ($\log_{10}$ of contact counts) and 10 Mb resolution (binarized; 0 if absent, 1 if present) were vectorized and concatenated using custom Python scripts. Concatenation was performed such that redundant entries of each contact matrix (i.e., C_{ij} and C_{ji}) were only represented once. Resulting matrices—where rows represent single cells, and columns represent observed contacts—were then decomposed using the PCA function in scikit-learn. For interchromosomal matrices, entries for intrachromosomal contacts (i.e., the diagonal) were set to 0. For 10 Mb intrachromosomal matrices, all interchromosomal contacts were ignored and all entries C_{ij} where $| i - j | < 3$ were set to zero.

Calculation of contact probabilities in single cells. Methods to calculate the scaling probability within single cells were adapted from Fudenberg *et al.*[19] and Sanborn *et al.*[20]. A histogram of contact distances normalized by bin size was generated using logarithmically increasing bins (increasing by powers of 1.12^n). We obtained the scaling coefficient by calculating the line of best fit for the log–log plot of this histogram between distances of 50 kb and 8 Mb. Shuffled controls were generated by randomly reassigning all cellular indices and repeating the above analysis; this importantly maintained the coverage distribution of the new set of simulated 'single cells'.

All plots were generated in R using ggplot2 (http://ggplot2.org/).

Statistics and reproducibility. No sample sizes were predetermined in this study. No *P* values were calculated during this study. All replicates in this study are defined as separate experiments, each carried out using unique starting material (i.e., separate aliquots of fixed cells).

Data availability statement. Raw data containing reads derived from HeLa cells are deposited at dbGaP under accession number phs000640.v4.p1. All processed data and raw reads not containing HeLa data are available at GEO accession GSE84920.

22. Ramani, V., Duan, Z. & Shendure, J. Massively multiplex single-cell Hi-C. *Protocol Exchange* http://dx.doi.org/10.1038/protex.2017.005 (2017).
23. Jin, W. *et al. Nature* **528**, 142–146 (2015).
24. Servant, N. *et al. Genome Biol.* **16**, 259 (2015).
25. Carette, J.E. *et al. Nature* **477**, 340–343 (2011).

Erratum: Massively multiplex single-cell Hi-C

Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan & Jay Shendure

***Nat. Methods*; doi:10.1038/nmeth.4155; corrected online 10 February 2017**

In the version of this article initially published online, the Gene Expression Omnibus (GEO) accession containing all processed data and raw reads (except for HeLa data) was not provided; the correct accession, GSE84920, has now been included. The error has been corrected for the print, PDF and HTML versions of this article as of 10 February 2017.