nature biotechnology

Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state

Assaf Rotem[1,2,7], Oren Ram[2–4,7], Noam Shoresh[2,7], Ralph A Sperling[1,6], Alon Goren[5], David A Weitz[1] & Bradley E Bernstein[2–4]

Chromatin profiling provides a versatile means to investigate functional genomic elements and their regulation. However, current methods yield ensemble profiles that are insensitive to cell-to-cell variation. Here we combine microfluidics, DNA barcoding and sequencing to collect chromatin data at single-cell resolution. We demonstrate the utility of the technology by assaying thousands of individual cells and using the data to deconvolute a mixture of ES cells, fibroblasts and hematopoietic progenitors into high-quality chromatin state maps for each cell type. The data from each single cell are sparse, comprising on the order of 1,000 unique reads. However, by assaying thousands of ES cells, we identify a spectrum of subpopulations defined by differences in chromatin signatures of pluripotency and differentiation priming. We corroborate these findings by comparison to orthogonal single-cell gene expression data. Our method for single-cell analysis reveals aspects of epigenetic heterogeneity not captured by transcriptional analysis alone.

The diversity of cells and tissues in an organism depends on chromatin organization, which controls access to genes and regulatory elements[1]. Regulatory proteins that catalyze post-translational histone modifications, remodel nucleosomes or otherwise alter chromatin structure are implicated in a wide range of developmental programs and are frequently mutated in cancer and other diseases[2]. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a widely used method for mapping histone modifications, transcription factors and other protein-DNA interactions genome-wide. Complementary methods have also been established for mapping accessible DNA, chromosomal loops and higher-order structures and interactions. The various data types can be integrated into genome-wide maps that provide systematic insight into the locations and cell type specificities of promoters, enhancers, noncoding RNAs, epigenetic repressors and other fundamental features of genome organization and regulation[1,3,4].

A limitation of chromatin mapping technologies is that they require large amounts of input material and yield ‘averaged’ profiles that are insensitive to cellular heterogeneity. This is a major shortcoming given that cell-to-cell variability is inherent to most tissues and cell populations. Cellular heterogeneity may be evident histologically, functionally (for example, in self-renewal assays) or in gene expression measurements, which have revealed striking heterogeneity within apparently homogeneous samples[5–7]. However, despite some initial progress[8–11], the extent and significance of chromatin-state heterogeneity remains largely uncharted.

Although single cell genomic technologies are evolving rapidly and challenging traditional views of biological systems[6] enabling the study of genetic mutations and transcriptomes at single cell resolution, and revealing marked heterogeneity in tissues, cellular responses and tumors[5,12–15], single cell analysis of chromatin states has remained elusive so far.

In parallel, advances in microfluidics are affecting chemistry, biology and medical diagnostics[16]. Miniaturized lab-on-chip devices enable precise control of fluidics in increasingly sophisticated configurations. Drop-based microfluidics (DBM) is a further innovation in which micron-sized aqueous drops immersed in an inert carrier oil are rapidly conducted through a microfluidics device[17]. The drops are ideal microreactors and can be precisely sized to contain one individual cell. Individual drops can be filled, steered, split, combined, detected and sorted in microfluidics devices, and thousands of individual drops can be manipulated in less than a minute using only microliters of reagent[18–20].

Here we combined microfluidics, DNA barcoding and next-generation sequencing to acquire low-coverage maps of chromatin state in single cells. We applied the method to profile H3 lysine 4 trimethylation (H3K4me3) and dimethylation (H3K4me2) in mixed populations of mouse embryonic stem (ES) cells, embryonic fibroblasts (MEFs) and hematopoietic progenitors (EML cells), and we show that we can determine the identity of each individual cell and recapitulate high-quality chromatin profiles for each cell state in the mixture. Although the resulting single-cell data are sparse—capturing on the order of 1,000 marked promoters or enhancers per cell—the data are sufficient to identify distinct epigenetic states and to characterize underlying patterns of variability. Within the ES cell population, we detect coherent variations at pluripotency enhancers and Polycomb targets, which

[1]Department of Physics and School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. [2]Epigenomics Program, The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. [4]Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA. [5]Broad Technology Labs, The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [6]Current address: Fraunhofer ICT-IMM, Mainz, Germany. [7]These authors contributed equally to this work. Correspondence should be addressed to B.E.B. (Bernstein.Bradley@mgh.harvard.edu) or D.A.W. (Weitz@seas.harvard.edu).

Received 13 May; accepted 21 September; published online 12 October 2015; doi:10.1038/nbt.3383

appear to reflect a spectrum of differentiation priming, and delineate three subpopulations of cells along this spectrum.

RESULTS

Microfluidics system indexes chromatin from single cells

A fundamental limitation of chromatin mapping technologies relates to the immunoprecipitation (ChIP) step in which an antibody to a modified histone or transcription factor is used to enrich target loci. Low levels of nonspecific antibody binding pull down off-target sites and lead to experimental noise. The issue is exacerbated in small-input experiments, where the amount of on-target epitope may be exceedingly low. Although recent studies have used indexing and amplification procedures to reduce input requirements substantially[21–23], achieving single-cell resolution has remained unattainable.

We reasoned that this limitation might be overcome—at least in part—by labeling chromatin from single cells before immunoprecipitation. Indexed chromatin from multiple cells could then be combined, possibly along with carrier chromatin[24], before immunoprecipitation, thus avoiding the nonspecific noise associated with low input samples. We therefore sought to develop a microfluidics system capable of processing single cells to indexed chromatin fragments (**Fig. 1**, **Supplementary Fig. 1**, **Supplementary Tables 1** and **2** and https://pubs.broadinstitute.org/drop-chip).

We developed a DBM device that captures and processes single cells in ~50-µm-sized aqueous drops (**Figs. 1a** and **2**). As an initial step, we engineered a co-flow drop-maker module in which a suspension of dissociated ES cells is mixed with solution containing weak detergent and micrococcal nuclease (MNase), milliseconds before encapsulation of individual cells in drops (**Fig. 2a** and **Supplementary Video 1**). We confirmed visually that a vast majority of the aqueous drops produced by the module contain either one or zero cells, and confirmed effective cell lysis by fluorescent staining. Under our optimized conditions, MNase preferentially cut accessible linker DNA and efficiently digested the chromatin of single cells within drops (**Fig. 3**). The resulting mix of mono-, di- and trinucleosomes is retained in the same drop as the original cell.

In parallel, we engineered a barcode library consisting of a pool of drops, wherein each drop contains a distinct oligonucleotide adaptor. We designed 1,152 oligonucleotide adaptors each containing a unique 'barcode' sequence, an Illumina-compatible adaptor and restriction sites for selecting 'desired' products (**Fig. 3a**). We then engineered a parallel drop-maker that extracts the oligonucleotides from individual wells in 384-well plates across a pressure gradient into drops, such that each drop contains multiple copies of the same barcode (**Supplementary Fig. 2**). The barcode-containing drops are then combined into a single emulsion (**Supplementary Fig. 2**).

We used a three-point merging device to merge each nucleosome-containing drop with a single barcode-containing drop (**Fig. 2b**). We reinjected a stream of nucleosome-containing drops into one inlet, a stream of barcode-containing drops into a second inlet, and an enzymatic buffer with DNA ligase into a third inlet. The barcode drops (smaller) and the nucleosome drops (larger) pair asymmetrically owing to hydrodynamic forces, and an electric field triggers fusion between one barcode drop, one nucleosome drop and a small aliquot of the enzymatic solution. Barcoded adaptors are ligated to both ends of the nucleosomal DNA fragments, thus indexing the chromatin contents of each drop to their originating cell (**Fig. 2b** and **Supplementary Video 2**).

Although the microfluidics system is designed to yield fusions between one drop containing nucleosomal contents of a single cell and one drop containing a unique barcode, alternate scenarios are possible and must be minimized. First, to mitigate the possibility that one drop might contain more than one cell, we titrated the cell density of the initial suspension such that only 1 in 6 drops contain a cell.

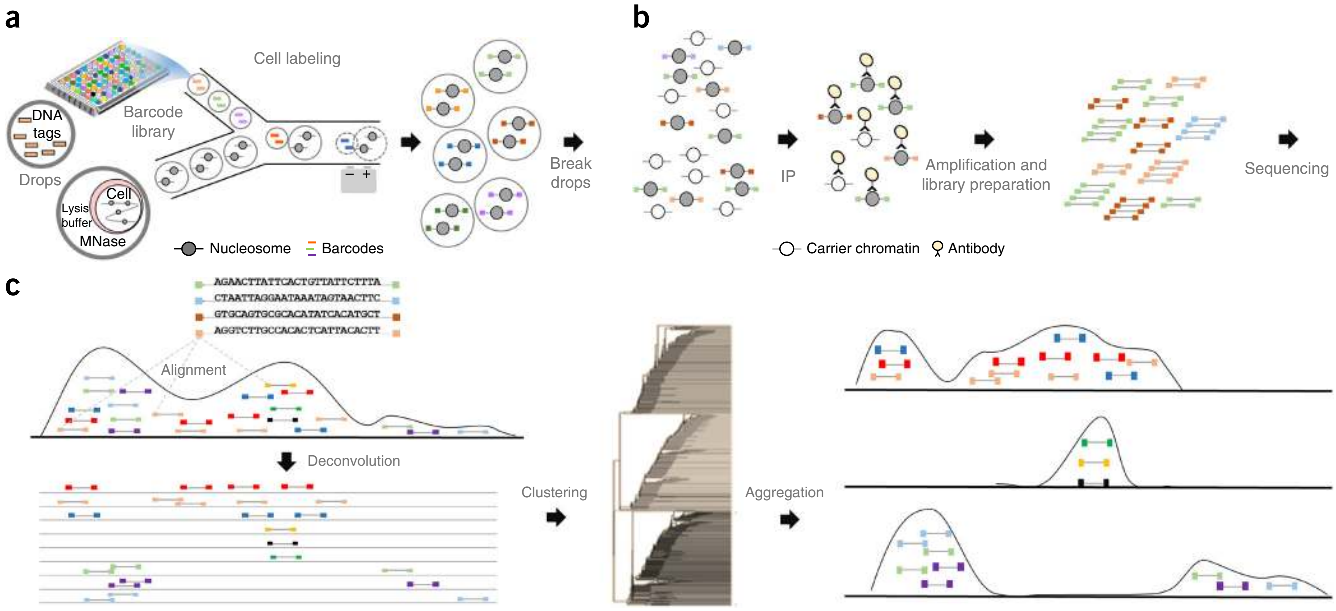

Figure 1 Overview of Drop-ChIP procedure for acquiring single cell chromatin data. (**a**) Microfluidics workflow. A library of drops containing DNA barcodes is prepared by emulsifying DNA suspensions from plates (top left). Cells are encapsulated and lysed in drops and then their chromatin is fragmented (bottom left). Chromatin-bearing drops and barcode drops are merged in a microfluidic device, and DNA barcodes are ligated to the chromatin fragments, thus indexing them to originating cell. (**b**) Combined contents of many drops are immunoprecipitated in the presence of 'carrier' chromatin and the enriched DNA is sequenced. (**c**) Sequencing reads are partitioned by their barcode sequences to yield single cell chromatin profiles (left). An unsupervised algorithm identifies groups of related single cell profiles, which are then aggregated to produce high-quality chromatin profiles for subpopulations (right). See also **Supplementary Figure 1**.

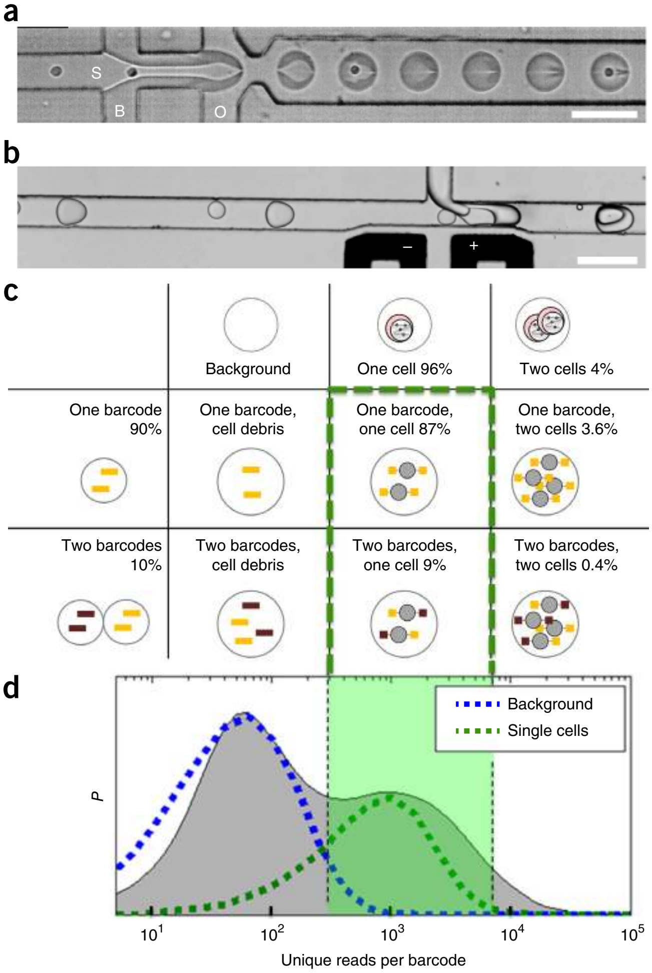

Figure 2 Labeling single-cell chromatin by drop-based microfluidics. (**a**) Micrograph shows an aqueous suspension of cells ('S') co-flowed together with lysis buffer and MNase ('B') as they enter the drop maker junction and disperse in oil ('O'), resulting in the formation of cell-bearing drops (see also **Supplementary Video 1**). (**b**) Micrograph shows cell-bearing drops (~50-µm diameter) and barcode-bearing drops (~30-µm diameter) paired in a microfluidics "three-point merger" device. As adjacent drops flow by the electrodes (+ and –), an induced electric field triggers their coalescence; simultaneously, labeling buffer (B) containing ligase is injected into the merged drops (**Supplementary Video 2**). (**c**) Table depicts estimated frequencies of possible drop fusion outcomes. The number of cells in each drop was measured from **Supplementary Video 1** (see panel **a**). Drops containing cells or cell debris may fuse with one (90%) or two (10%) barcode drops (green frame). Two-barcode fusion events can be detected and corrected *in silico*. Background reads contributed by drops that contain only cell debris are also filtered *in silico*. (**d**) The frequency distribution of barcodes is plotted as a function of the number of reads contributed by each barcode and fitted to a sum of two Poisson distributions, one for the background reads (blue) and one for the single-cells reads (green; see Online Methods). Barcodes in the highlighted range are assumed to originate from single cells and are retained for further analysis. Scale bars, 100 µm.

The remaining empty drops fuse with barcode but their inert contents do not contribute to the eventual sequencing library (**Fig. 2c**). Second, we tuned the system such that each nucleosome-containing drop fuses with either one or two barcodes (**Fig. 2d**), with the understanding that these alternative scenarios can be decoded at the analysis stage (see also Online Methods). Third, we limit each collection to 100 cells paired with barcodes randomly drawn from a library of 1,152 barcodes, ensuring that >95% of barcodes will be unique to a single cell (per Poisson statistics). This conservative approach has little impact on throughput as we multiplex thousands of single cells by collecting multiple samples in parallel and adding a second 'sample' index before sequencing.

Chromatin immunoprecipitation and sequencing

The chromatin fragments generated by the microfluidics platform contain barcode adaptors that index them to originating cells and provide a handle for PCR. We combine indexed chromatin from 100 cells with carrier chromatin from a different organism, perform ChIP and use the enriched DNA to prepare a sequencing library. The barcode adaptors comprise symmetric sequences, such that both ends are available for ligation to nucleosome ends (**Fig. 3a**). Each end contains the same 8-bp barcode (1 out of 1,152 possible sequences) flanked by a universal primer and restriction sites. Adaptor concatemers produced due the large excess of adaptors in the drops (~10^9 copies versus ~10^7 nucleosomal fragments) are eliminated by restriction before amplification (**Fig. 3b**). Symmetrically labeled nucleosomal fragments are amplified by PCR and a second restriction yields an overhang compatible with standard Illumina library preparation. At this stage, we introduce a second 'sample' barcode, enabling us to multiplex thousands of cells in a single sequencing run.

We paired-end sequence these 'Drop-ChIP' samples, reading the 'sample' indexing barcode, the 'single cell' indexing barcode and the intervening genomic DNA. We used HiSeq 2500 (Illumina, USA) for sequencing, with each lane producing on average 320 million reads with high accuracy (88% of reads ≥Q30 (PF)). The typical yield per pool of 100 cells is 7 million aligned reads, of which ~700,000 are unique (**Supplementary Table 3a**). We performed a series of quality controls to ensure homogeneous distribution of barcodes within and across experiments (**Fig. 3d**), to ensure the stability of the barcode library (**Fig. 3e**) and to ensure that barcodes were not mixing or exchanging between drops (**Fig. 3f**). We then filtered the sequencing data to include only reads that contain symmetric barcodes on both sides of the nucleosomal insert (**Fig. 3c**; see also Online Methods) and to exclude highly over-represented barcodes that may have labeled two or more cells (**Fig. 2d**). After filtering, we retain between 500 and 10,000 Drop-ChIP reads per single cell (**Supplementary Table 3b**).

Single-cell profiles deconvolute cell type-specific landscapes

We benchmarked Drop-ChIP in a series of biological settings. We initially focused on three different mouse cell populations: ES cells, MEFs and the hematopoietic line EML. We separately applied suspensions of each cell type to the microfluidic device and labeled individual cells from each population with a different set of barcodes. We then combined labeled chromatin from the three cell types, performed ChIP with H3K4me3 antibody and sequenced the resulting library. We acquired a total 1.1 million (M) uniquely aligned sequencing reads. These reads were distributed on the basis of their barcodes into 868 bins, each corresponding to a single cell.

Visual inspection of single-cell data for 50 individual ES cells and 50 individual MEFs reveals the high quality of the resulting data (**Fig. 4a** and Online Methods). Reads from single cells have a strong tendency to coincide with peaks that are evident in bulk ChIP-seq profiles for the corresponding cell types. The specificity is sufficient that single-cell profiles for ES cells are readily distinguished from single-cell MEF profiles by examination of differentially marked regions (for example, Anxa1 for MEFs; Oct4 and Sox2 for ES cells). Considering all single cells in the H3K4me3 data set, more than 50% of sequencing reads

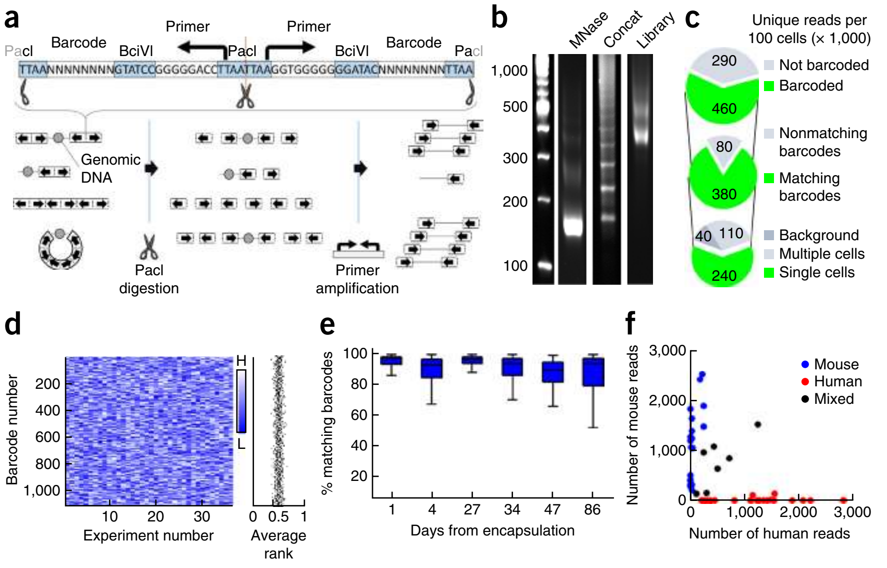

Figure 3 Symmetric barcoding and amplification of chromatin fragments. (**a**) Barcode adapters (top) are 64-bp double-stranded oligonucleotides with universal primers, barcode sequences and restriction sites, whose symmetric design allows ligation on either side. Schematic (bottom left) depicts possible outcomes of ligation in drops, including symmetrically labeled nucleosomes, asymmetrically labeled nucleosomes and adaptor concatemers. Concatemers are removed by digestion of PacI sites formed by adaptor juxtaposition (bottom center), allowing selective PCR amplification of symmetrically adapted chromatin fragments (bottom right). See also **Supplementary Figure 2**. (**b**) Gel electrophoresis for DNA products at successive assay stages. Left lane, DNA ladder; MNase, DNA fragments purified after capture, lysis and MNase digestion of single cells in drops confirm efficient digestion to mononucleosomes (~1 million drops collected); Concat, Illumina library prepared from adaptor-ligated chromatin fragments without PacI digestion reveals overwhelming concatemer bias; Library, Illumina library prepared from adaptor-ligated chromatin fragments digested with PacI, reveals appropriate MNase digestion pattern, shifted by the size of barcode and Illumina adapters. (**c**) Pie charts depict numbers of uniquely aligned sequencing read that satisfy successive filtering criteria (values reflect data from 100 single cells, averaged over 82 trials). We select reads that have barcode sequences on both ends (top) with matching sequence (middle). We then apply a Poisson model to identify barcodes that represent single cells (bottom). (**d**) Heat map depicts homogeneity of barcode selection. Barcodes (rows) are colored according to their relative prevalence (rank order) across 37 experiments (columns). The absence of bias toward particular barcodes (light or dark horizontal stripes) indicates the homogeneity of the barcode library. The mean normalized rank over all barcodes (right) is close to 0.5, consistent with balanced representation. (**e**) Stability of the barcode library emulsion over time. The fraction of reads with matching barcodes on both ends is plotted as a function of time from encapsulation of the barcode library. (**f**) The microfluidics system was applied to barcode a mixed suspension of human and mouse cells. For each barcode, plot depicts the number of reads aligning to the mouse genome (*y* axis) versus the number of reads aligning to the human genome (*x* axis). The data suggest that a vast majority of barcodes is unique to a single cell.

fall within known positive regions, defined from bulk ChIP-seq data (**Fig. 4b**). This proportion is essentially identical to the proportion of reads in bulk ChIP-seq data sets that fall within enriched intervals. The sensitivity of the single-cell profiles is compromised by the low per-cell sequencing coverage. Only ~800 peaks are detected per cell, which corresponds to an overall sensitivity for peak detection of ~5% (**Fig. 4b**). The overall accuracy of the single-cell data is nonetheless supported by the very strong concordance of aggregated data to conventional ChIP-seq measurements (**Supplementary Fig. 3**).

Although the single-cell profiles lack sensitivity for *de novo* peak calling, we reasoned that detection of ~800 true peaks with high specificity might be sufficient to classify or group individual cells with related chromatin landscapes. Indeed, we found that detection of just a few hundred peaks was sufficient to distinguish single-cell MEF profiles from single-cell ES cell profiles with nearly 100% accuracy. For example, single-cell profiles are readily and accurately identified as ES cell or MEF by comparison against conventional ChIP-seq maps (**Fig. 4c**). Moreover, we could apply an unsupervised-clustering approach to distinguish the respective cell states without any prior information about their landscapes. Representing each single cell profile as the number of reads in nonoverlapping 5-kb windows spanning the genome, we calculated the covariance between all pairs of cells. We then used a divisive hierarchical clustering algorithm to cluster the cells based on pairwise distances (DIANA; see Online Methods). This unbiased analysis distinguished three main groups of cells, which are clearly evident in a cluster tree (**Fig. 4d**). Because each cell type had been labeled with a distinct barcode set in this pilot, we were able to evaluate the accuracy of the clustering. We found that >97% of cells in the first cluster were EML cells, >91% of cells in the second cluster were ES cells and >97% of cells in the third cluster were MEFs. Moreover, when we aggregated reads from single-cells in each cluster, the resulting profiles closely matched conventional ChIP-seq data for EML cells, ES cells and MEFs, respectively (**Supplementary Fig. 4**). We note that high-quality single-cell level information was absolutely critical for deconvoluting these populations: when we compromised *in silico* the resolution of our single cell profiles by randomly combining sets of five cells, we were no longer able to distinguish ES cells from MEFs or to deconvolute profiles for the cell types in the mixed population (**Supplementary Fig. 5**).

Finally, we performed an additional Drop-ChIP experiment in which we mixed ES cells and MEFs before their application to the microfluidics device. The DIANA algorithm again effectively resolved single cells of each type based solely on their chromatin profiles, enabling us to produce aggregate profiles for ES cells and MEFs, which closely match conventional ChIP-seq data (**Supplementary Fig. 6**).

Epigenetic states distinguished in a population of ES cells

Transcriptional activity varies between individual cells, even within apparently homogeneous cell types or tissues. Yet how this transcriptional heterogeneity relates to cell-to-cell variability in the underlying gene regulatory elements remains an open question. We therefore examined H3K4me2, a marker of promoters and enhancers. H3K4me2 profiles have been used to survey regulatory element activity genome-wide in a range of cell and tissue types[1,4]. However, the extent to which these landscapes vary across single cells in a population has yet to be determined.

We acquired H3K4me2 data for 4,643 ES cells, cultured in serum with LIF, and 762 MEFs (numbers reflect cells retained after quality controls; see Online Methods). The DIANA algorithm readily clustered these cells into two major groups. Aggregation of reads from cells in the larger group yielded a chromatin profile that closely matched a

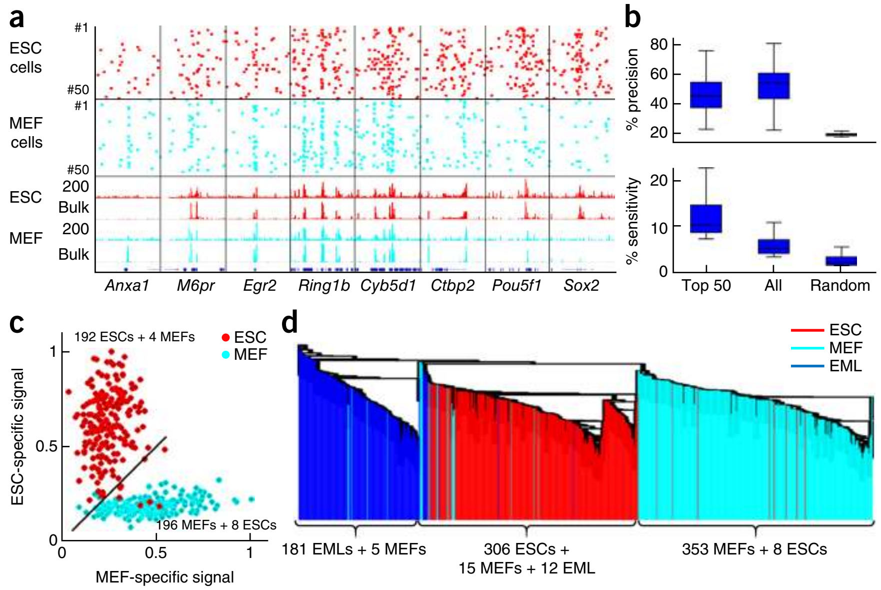

Figure 4 Single-cell H3K4me3 chromatin data inform about subpopulations of known cell types. (**a**) Drop-ChIP data is shown for 50 ES cells (ESCs) and 50 MEFs across representative gene loci. Each row represents data from a single cell. Each column includes reads in 330-kb regions centered on selected genes (*Anxa1*, chr19: 20465000; *M6pr*, chr6: 122269000; *Egr2*: chr10: 67022000; *Ring1b*, chr17: 34262000; *Cyb5d1*, chr11: 69207000; *Ctbp2*, chr7: 140254000; *Pou5f1*, chr17: 35612000; *Sox2*, chr3: 34573000). A high proportion of reads aligns to genomic positions enriched in both bulk ChIP-seq assays ('Bulk') and aggregated chromatin profiles from 200 single-cell assays ('200'), providing evidence that single-cell data are informative. (**b**) The precision (fraction of single-cell reads overlapping known H3K4me3 peaks) and sensitivity (fraction of known H3K4me3 peaks occupied by single-cell reads) are plotted for the top 50 ES cells by sensitivity and for all ES cells in the data set. These data are compared to random profiles simulated by arbitrarily positioning reads. Middle bar marks the median, box covers the 25th–75th percentiles and whiskers cover the 1st–99th percentiles. The average ES cell H3K4me3 profile has a precision of 53% ± 12% and a sensitivity of 7% ± 4%, whereas the average ES cell H3K4me2 profile has a precision of 42% ± 5% and a sensitivity of 3% ± 2% (not shown). (**c**) For 400 single-cell H3K4me3 profiles, scatterplot depicts normalized detection of ES cell–specific intervals versus MEF-specific intervals. In this experiment, ES cells (red) and MEFs (green) were separately barcoded in the microfluidics device, but collectively immunoprecipitated and processed. A naive classification (black line) distinguishes ES cell profiles from MEF profiles with >95% specificity and sensitivity. (**d**) ES cells, MEFs and EML cells were separately barcoded but collectively processed to acquire 883 single-cell profiles (314 ES cells, 376 MEFs, 193 EMLs). These profiles were clustered using an unsupervised divisive hierarchical clustering algorithm (see Online Methods). The hierarchal tree discriminates between cell types with >95% accuracy, indicating that the information content of single-cell profiles is sufficient to accurately group related cells and thereby distinguish cell states within a mixed population. See also **Supplementary Figures 3–6** and Online Methods.

corresponding bulk H3K4me2 profile for ES cells, while aggregation of reads from the smaller group yielded a profile consistent with MEFs.

We next considered whether the single-cell clustering patterns might reveal additional substructure among the ES cells, potentially reflecting subpopulations with distinct regulatory states. The existence of such subpopulations is supported by prior studies that examined gene reporters and transcriptional signatures at single-cell resolution[25–29]. However, when we used DIANA to cluster individual ES cells based on their H3K4me2 data, we found that the results were highly sensitive to algorithm parameters and technical attributes, such as mean single-cell coverage.

We therefore implemented an alternative approach based on the assumption that functionally related genomic elements, which tend to vary coherently across cell types, also vary coherently across individual ES cells. We reasoned that our sensitivity to detect subsets of ES cells with distinct regulatory patterns would be increased by considering such element sets or 'signatures', which would have higher signal-to-noise ratio in our data than individual elements. This strategy is analogous to signature-based methods that have been successfully applied in the analysis of single-cell RNA-seq, DNA methylation and chromatin accessibility data[10,13,28] and in the interpretation of cancer mutations[30,31]. To identify signature sets relevant to chromatin states, we collected 314 publicly available ChIP-seq profiles for histone modifications, transcription factors and chromatin regulators. We collated target (enriched) regions for each profile and then clustered the profiles based on the overlaps between these target regions. We thereby identified 91 representative signatures, each composed of a set of genomic elements with shared chromatin states (for example, H3K9me3 in ES cells), transcription factor binding (for example, Oct4 targets), and/or chromatin regulator occupancy (for example, p300 targets) (**Supplementary Fig.** 7 and **Supplementary Table 4**).

For each individual ES cell (or MEF), we calculated the number of reads overlapping each signature, thereby creating a matrix of 5,405 single cells by 91 signatures. Agglomerative hierarchical clustering of the signature matrix distinguished several prominent groups of cells with correlated chromatin landscapes (**Fig. 5a**; see Online Methods). The major division segregated all MEFs from ES cells, which were distributed across several clusters. To visualize the relationship between cells, we derived multidimensional scaling (MDS) plots from the signature matrix (**Fig. 5b**). The MEFs show a relatively tight distribution, suggestive of more concordant H3K4me2 landscapes. By contrast, the individual ES cells cover a much larger region within the MDS plot, segregating into three loose groups (**Fig. 5b**). The tighter distribution among individual MEFs may relate to observations that such lineage-committed cells adopt a relatively constrained chromatin state. By contrast, ES cell chromatin is notable for its accessible and plastic state[32].

Several lines of evidence support the robustness and validity of the signature-based clustering (**Supplementary Note 1**). First, the most prominent division accurately distinguishes ES cells from MEFs (98% of ES cells are correctly classified as ES cells; 95% of MEFs are correctly classified as MEFs). Second, the signature-based clusters are independent of read coverage (**Supplementary Fig. 8a**). Third, the signature-based clusters are robust with respect to the removal of subsets of single cells. When we repeatedly simulated the clustering after randomly removing 50% of the cells, only a fraction of cells at the edges of MDS clusters switched their assignments (**Fig. 5c** and **Supplementary Fig. 8b**). By contrast, when reads were randomly reassigned between cells, the correlation structure driving the clustering was lost (**Supplementary Fig. 8c**). We also tested our sensitivity to detect small subpopulations by removing cells from one of the clusters *in silico*. We found that sensitivity depended on the frequency of the subpopulation and the total number of sampled cells, such that

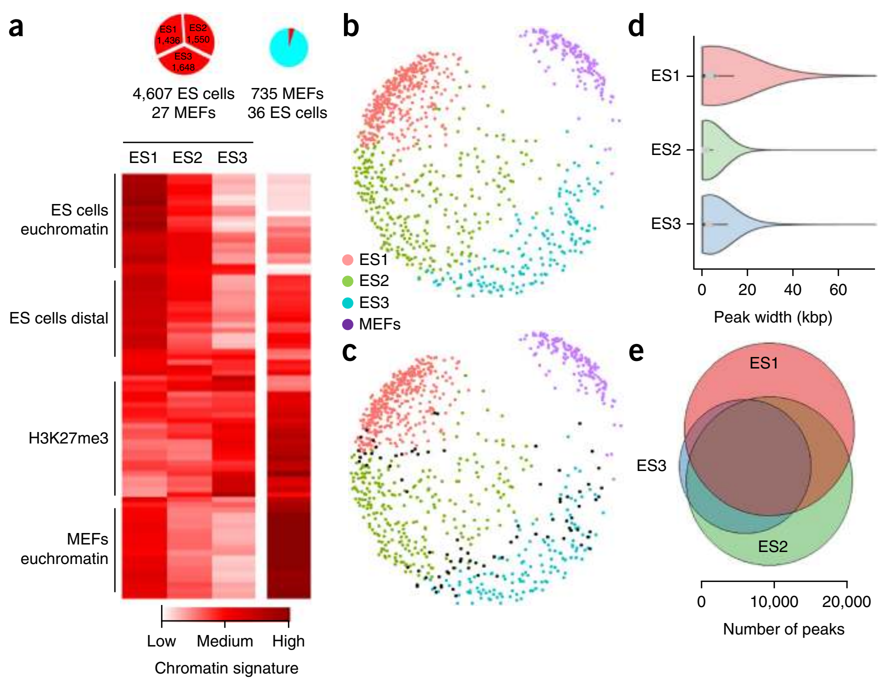

Figure 5 A spectrum of ES cell subpopulations with variable chromatin signatures for pluripotency and priming. (**a**) Single-cell H3K4me2 data for 4,643 ES cells and 762 MEFs were subjected to agglomerative hierarchical clustering based on their scores in 91 signature sets of genomic regions (see Online Methods). Pie chart at left depicts the proportions of individual ES cells that cluster into each of three clusters (1,436 cells in ES1, 1,550 cells in ES2 and 1,648 cells in ES3), and pie chart at right depicts the relative numbers of ES cells and MEFs that cluster into a fourth group, which corresponds to MEFs. Heat map (below) depicts the mean signature scores (rows) for each cluster (columns). (**b**) Multidimensional scaling (MDS) plot compares the chromatin landscapes of single ES cells and MEFs (colored dots). The distance between any two dots (cells) approximates the distance between their 91-dimensional signature vectors. The plot shows 1,000 single cells (randomly sampled from the 5,405 cells with H3K4me2 data), colored on the basis of their cluster association. Tight co-localization of the MEF cluster and, to a lesser degree, the ES1 cluster suggests that the corresponding landscapes are relatively more homogeneous. In contrast, the ES2 and ES3 clusters are more broadly distributed and may reflect a gradient of single cell states. (**c**) MDS plot as in **b**, but with cells that frequently switched clusters in bootstrapping tests on varying subsets of cells indicated in black (see Online Methods). These unstable cells are exclusively located on the borders between clusters. (**d**) Violin plots show the distribution of peak widths for peaks called from aggregate ES1, ES2 or ES3 profiles (see Online Methods). (**e**) Venn diagram depicts the relative numbers and overlaps of peaks called from aggregate ES1, ES2 or ES3 profiles. The ES1 cluster is notable for higher pluripotency-signature scores, larger numbers of peaks and tighter internal concordance. In contrast, the ES3 cluster has higher activity over Polycomb signatures and increased heterogeneity, potentially reflecting a mixture of primed states. See also **Supplementary Figures 7** and **8**, **Supplementary Note 1** and the online source data for this figure.

detection of rarer subsets requires analysis of larger numbers of cells (for example, detecting a subpopulation present at 5% requires the analysis of 1,000 cells in total; **Supplementary Fig. 9**).

To test the dependence of the clusters on the set of signatures used, we repeated the agglomerative hierarchical clustering using (i) all 314 signatures without any filtration or (ii) a distinct collection of signatures from a recently established resource of functional genomic data sets (E. Meshorer, Hebrew University, personal communication). In both cases, we again distinguished a tight cluster of MEFs, as well as three groups of ES cells that closely correspond to the groups derived using the original 91 signatures (**Supplementary Fig. 8d**). Finally, to exclude the possibility that the ES cell clusters reflect different cell cycle signatures, we tested, but found no evidence, for differential activity of cell cycle–related genes (**Supplementary Fig. 8e**).

Coherent variations at pluripotency elements and bivalent promoters

We considered the biological significance of the three ES cell subpopulations defined from the single cell data, which we termed ES1, ES2 and ES3. First, we examined the distribution of signature scores across these subpopulations. We observed notable differences in the H3K4me2 signal distributions over pluripotency-related signatures, such as Oct4 or Sox2 targets[33]. Cells in the ES1 group tend to have the highest signal over pluripotency signatures, ES2 cells intermediate signals and ES3 cells the lowest signals over these elements (**Fig. 6a**). These differences in target element activity may relate to the heterogeneous expression of pluripotency factors, previously documented in ES cell populations[27,29]. We observed the opposite pattern for a signature composed of targets of FoxA2, with progressively higher signals in ES2 and ES3. FoxA2 is an endodermal transcription factor whose regulatory targets are dynamically activated during early ES cell differentiation[34]. Although FoxA2 expression is rarely evident in undifferentiated ES cells, this signature may reflect a degree of lineage priming associated with very low expression of factors involved in early specification.

The respective subpopulations also vary in terms of their signal distributions over Polycomb and CoREST targets. Polycomb targets correspond to bivalent domains, which are inactive but poised in pluripotent cells[35,36]. CoREST is a potent repressor that silences neural-related genes in ES cells. H3K4me2 signals over Polycomb and CoREST signatures are lowest in ES1, consistent with a pure pluripotent state, but progressively increase in the ES2 and ES3 populations (**Fig. 6a**). In fact, the Polycomb signatures correlate inversely with pluripotency signatures across all single ES cells in the data set (**Fig. 6b**). Thus, the latter populations show reduced chromatin activity at pluripotency targets and increased activity at sites that are normally inactive in pluripotent cells.

We also generated aggregate H3K4me2 profiles for the ES1, ES2 and ES3 subpopulations by combining reads from the cells in each cluster (see Online Methods). Comparison of these profiles confirmed differences over elements in the various signatures, most notably pluripotency and Polycomb targets. We also observed global differences between the landscapes. H3K4me2 peaks in the ES3 profile are present in fewer numbers and are narrower than in ES1 and ES2 (**Fig. 5d**,**e**). In addition to their global accessibility[32], pluripotent cells have relatively larger numbers of elements marked by distal chromatin signatures[37]. The reduced H3K4me2 peaks in ES3 may thus be an additional reflection of a primed chromatin state. Alternatively or in addition, the changes in ES3 may reflect a spectrum of subthreshold priming events associated within alternative early fates;

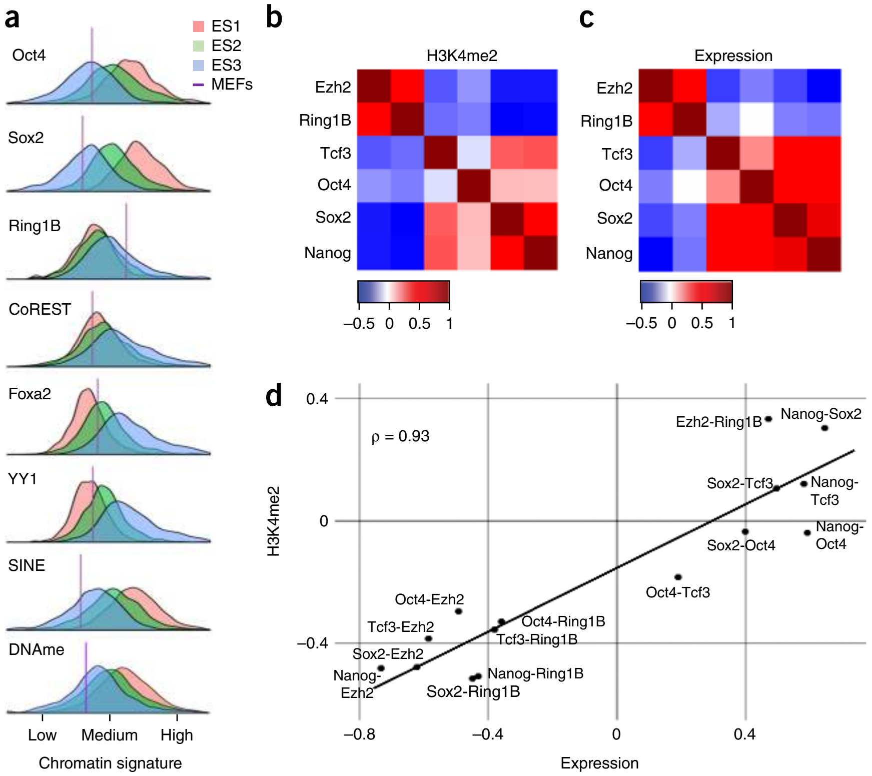

Figure 6 Orthogonal single-cell assays corroborate ES cell subpopulations and cell-to-cell variability in regulatory programs. (**a**) The distribution of single-cell scores for eight dominant signatures is plotted for ES1, ES2 and ES3. Vertical lines depict the mean score of each signature in MEFs. DNAme signature consists of 10,000 regions identified by Kelsey *et al.*[28] as most variable in their methylation status in ES cells. (**b**) Heat map depicts positive and negative correlations between six selected signatures, based on co-variation of H3K4me2 across single ES cells. (**c**) Heat map depicts positive and negative correlations between six selected signatures, based on co-variation of expression across single ES cells (See **Supplementary Note 2**). (**d**) Scatterplot depicts correlations between the indicated signature pairs across single ES cells, as determined from H3K4me2 or RNA expression data. Best-fit line and Pearson correlation are also indicated. Thus, orthogonal single-cell techniques lead to similar conclusions regarding ES cell subpopulations and underlying patterns of variability in pluripotency and Polycomb signatures, suggestive of a continuum from pluripotent to primed states. See also **Supplementary Figure 10**.

such a scenario might explain the relatively higher variance of the ES3 single cell profiles (**Fig. 5b**). Together, our findings suggest that the respective subpopulations reflect a continuum of ES cell states with varying degrees of pluripotency- or priming-related chromatin features.

Earlier studies have documented variability in pluripotency factor expression and DNA methylation levels across individual ES cells, which may in part reflect naive and primed subpopulations[27,38,39]. Our findings suggest that this cell-to-cell variability is accompanied by widespread alterations at pluripotency-associated regulatory elements, lineage-specific genes and Polycomb targets. Yet lineage-specific genes and Polycomb targets show sparse expression in ES cells[27,40], suggesting that the chromatin alterations may occur with relative independence from downstream transcriptional changes. However, a recently published single-cell RNA sequencing study for ES cells[41] reported that the expression of pluripotency genes and Polycomb targets is variable across individual ES cells—a conclusion that directly parallels our findings. Indeed, when we directly analyzed the single-cell RNA data, we found that the composite expression of pluripotency-related genes anticorrelates with the composite expression of Polycomb-target genes across single cells, again consistent with our chromatin findings (**Fig. 6c,d**). Furthermore, clustering of the single cell RNA profiles, based on these signature gene sets, distinguished two ES cell subpopulations with features of ES1 and ES3, respectively (**Supplementary Fig. 10** and **Supplementary Note 2**). This concordance between single-cell chromatin and RNA profiling supports our technological approach and biological findings.

DISCUSSION

Access to genomic information is controlled by cell type–specific chromatin structures. Chromatin maps provide a systematic means to identify regulatory sequences and track their activity across cellular states[1]. However, current methods yield averaged 'ensemble' profiles that are insensitive to internal heterogeneity. This is a major limitation given that cellular heterogeneity is inherent to most, if not all, tissues, cell types and models.

Here we sought to overcome this limitation by combining drop-based microfluidics with genomic barcoding to establish a platform for profiling chromatin at single-cell resolution. Although our method was able to detect cell-cell variations, this first attempt has limitations that will need to be addressed through further innovations. The coverage per cell will need to be increased by improved ligation efficiency, more efficient amplification and/or alternative barcoding methods. It may also be valuable to replace MNase digestion with other fragmentation strategies, thus expanding the strategy's applicability beyond chromatin marks. Similarly, the use of barcoded beads could substantially increase the number of cells per sample and improve the efficiency of our method[18,19].

The single-cell chromatin data are sparse, with only about 1,000 peaks detected in each individual cell due to low coverage. Nonetheless, specificity is high, with ~50% of reads aligning to known positive sites. The accuracy and information content can be appreciated through visualization of the single-cell tracks (**Fig. 4a**) and by comparing aggregate data for as few as 50 cells to conventional profiles. Regardless, the primary goal of our single-cell study is to find patterns of cell-to-cell variation across a population, rather than to examine an individual given cell. Several lines of evidence establish the capacity of our assay to acquire such information. First, the data from each single cell contains ample information to decipher its cell identity based on comparisons to known landscapes. Moreover, an unbiased clustering procedure applied to Drop-ChIP data generated for a mixed population of cells could effectively distinguish the 'cell type' of each single-cell profile with nearly 100% accuracy. Finally, aggregate profiles derived for each unbiased cluster closely matched conventional profiles for the respective substituents of the mixed population. Although this approach has been successful, we note that its success relies on the existence of a coherent chromatin state in a sufficient number of sampled cells. Power to distinguish such subpopulations thus benefits from sampling large numbers of cells and from the high throughput of microfluidics systems.

We used the method to investigate the cell-to-cell variability of different types of regulatory elements. We profiled H3K4me2, a marker of promoters and enhancers, in thousands of individual ES cells. We then asked whether coherent variations in the single cell chromatin data might unveil subpopulations with distinct epigenetic states. To maximize our sensitivity in regard to distinguishing closely related cell states, we implemented a clustering procedure based on 'signature' sets of elements. In this way, we were able to delineate three subpopulations (ES1, ES2 and ES3) whose identity is robust to permutations. The subpopulations are distinguished by their signals over loci bound by pluripotency- or differentiation-associated transcription factors or targeted by epigenetic repressors, including Ezh2, Ring1B and REST. Specifically, the ES1 population sustains high pluripotency factor activity and robust silencing over Polycomb and CoREST targets, and it may thus be analogous to 'naive' ES cells[39]. By contrast, the ES3 population exhibits signs of differentiation priming, including increased chromatin activity over enhancers implicated in early endodermal lineages and subtle derepression of Polycomb targets. This population also appears relatively heterogeneous, with lower concordance between individual cells potentially reflecting alternate priming states. Remarkably similar patterns of cell-to-cell variability are evident in single-cell RNA expression data generated for an analogous ES cell population[41]. Here again, pluripotency factors and Polycomb targets are seen to vary coherently across individual cells, with positive and negative correlations among gene and regulator sets showing striking parallels to their corresponding patterns of chromatin activity (**Fig. 6b–d**). We suggest that integration of single-cell chromatin and single-cell expression data may allow more precise coupling of regulatory elements with target genes and deeper understanding of their functional dynamics and relationships.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. GEO: GSE70253.

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

ACKNOWLEDGMENTS

We thank A. Regev, N. Yosef, E. Shema, I. Tirosh, H. Zhang, S. Gillespie and J. Xing for their valuable comments and critiques of this work. We also thank G. Kelsey for sharing single-cell data for comparisons. This research was supported by funds from Howard Hughes Medical Institute, the National Human Genome Research Institute's Centers of Excellence in Genome Sciences (P50HG006193), ENCODE Project (U54HG006991), the National Heart, Lung, and Blood Institute (U01HL100395), the National Science Foundation (DMR-1310266), the Harvard Materials Research Science and Engineering Center (DMR-1420570) and the Defense Advanced Research Projects Agency (HR0011-11-C-0093).

AUTHOR CONTRIBUTION

All authors designed experiments and approved the final manuscript. A.R. and O.R. performed experiments. A.R., O.R. and N.S. performed computational analyses. A.R., O.R. and R.A.S. developed experimental protocols. A.R., O.R. and N.S. developed analytical methods and tools. A.R., O.R., A.G., B.E.B. and D.A.W. conceived and designed the study. B.E.B., N.S., A.R., O.R. and D.A.W. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Rivera, C.M. & Ren, B. Mapping human epigenomes. *Cell* **155**, 39–55 (2013).
2. Baylin, S.B. & Jones, P.A. A decade of exploring the cancer epigenome–biological and translational implications. *Nat. Rev. Cancer* **11**, 726–734 (2011).
3. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
4. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
5. Shalek, A.K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
6. Kalisky, T. & Quake, S.R. Single-cell genomics. *Nat. Methods* **8**, 311–314 (2011).
7. Munsky, B., Neuert, G. & van Oudenaarden, A. Using gene expression noise to understand gene regulation. *Science* **336**, 183–187 (2012).
8. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
9. Brown, C.R., Mao, C., Falkovskaia, E., Jurica, M.S. & Boeger, H. Linking stochastic fluctuations in chromatin structure and gene expression. *PLoS Biol.* **11**, e1001621 (2013).
10. Cusanovich, D.A. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
11. Murphy, P.J. *et al.* Single-molecule analysis of combinatorial epigenomic states in normal and tumor cells. *Proc. Natl. Acad. Sci. USA* **110**, 7772–7777 (2013).
12. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
13. Patel, A.P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
14. Xu, X. *et al.* Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886–895 (2012).
15. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
16. Sackmann, E.K., Fulton, A.L. & Beebe, D.J. The present and future role of microfluidics in biomedical research. *Nature* **507**, 181–189 (2014).
17. Guo, M.T., Rotem, A., Heyman, J.A. & Weitz, D.A. Droplet microfluidics for high-throughput biological assays. *Lab Chip* **12**, 2146–2155 (2012).
18. Klein, A.M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
19. Macosko, E.Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
20. Rotem, A. *et al.* High-throughput single-cell labeling (Hi-SCL) for RNA-Seq using drop-based microfluidics. *PLoS ONE* **10**, e0116328 (2015).
21. Adli, M., Zhu, J. & Bernstein, B.E. Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. *Nat. Methods* **7**, 615–618 (2010).
22. Wu, A.R. *et al.* Automated microfluidic chromatin immunoprecipitation from 2,000 cells. *Lab Chip* **9**, 1365–1370 (2009).
23. Lara-Astiaso, D. *et al.* Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
24. O'Neill, L.P., VerMilyea, M.D. & Turner, B.M. Epigenetic characterization of the early embryo with a chromatin immunoprecipitation protocol applicable to small cell populations. *Nat. Genet.* **38**, 835–841 (2006).
25. Hackett, J.A. & Surani, M.A. Regulatory principles of pluripotency: from the ground state up. *Cell Stem Cell* **15**, 416–430 (2014).
26. Hough, S.R. *et al.* Single-cell gene expression profiles define self-renewing, pluripotent, and lineage primed states of human pluripotent stem cells. *Stem Cell Rep.* **2**, 881–895 (2014).
27. Singer, Z.S. *et al.* Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Mol. Cell* **55**, 319–331 (2014).
28. Smallwood, S.A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).
29. Chambers, I. *et al.* Nanog safeguards pluripotency and mediates germline development. *Nature* **450**, 1230–1234 (2007).
30. Ben-Porath, I. *et al.* An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nat. Genet.* **40**, 499–507 (2008).
31. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
32. Meshorer, E. & Misteli, T. Chromatin in pluripotent embryonic stem cells and differentiation. *Nat. Rev. Mol. Cell Biol.* **7**, 540–546 (2006).
33. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
34. Li, Z. *et al.* Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell* **151**, 1608–1616 (2012).
35. Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.* **8**, 532–538 (2006).
36. Bernstein, B.E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
37. Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
38. Farlik, M. *et al.* Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep.* **10**, 1386–1397 (2015).
39. Nichols, J. & Smith, A. Naive and primed pluripotent states. *Cell Stem Cell* **4**, 487–492 (2009).
40. Ku, M. *et al.* Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* **4**, e1000242 (2008).
41. Kumar, R.M. *et al.* Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* **516**, 56–61 (2014).

ONLINE METHODS

The procedures for Drop-ChIP are explained in **Supplementary Figure 1** and in a dedicated web portal with an interactive flow chart (https://pubs.broadinstitute.org/drop-chip).

Buffers. Recipes for all buffers are described in **Supplementary Table 1**.

Reagents. For the inert carrier oil we use HFE-7500 (3M, USA) with 1% w/w of a block co-polymer surfactant of perfluorinated polyethers (PFPE) and polyethylene glycol (PEG) (008-FluoroSurfactant, Ran Biotechnologies, USA). To separate the emulsion we use a commercially available demulsifier (1*H*,1*H*,2*H*,2*H*-perfluoro-1-octanol, CAS # 647-42-7, Sigma-Aldrich, USA). Antibodies for Immuno Precipitation were purchased from Millipore (H3K4me3: 07-473, H3K4me2: 07-030).

Microfluidic devices. We fabricate polydimethylsiloxane (PDMS) devices using photolithography and coat them with fluorophilic Aquapel (Rider, MA, USA) to prevent wetting of drops on the channel walls. Electrodes are fabricated on chip using low melting temperature solder[42]. The designs used to fabricate the devices are available in ACAD format (**Supplementary Design Files**). We use OEM syringe pumps (KD Scientific, MA, USA) to drive the fluidics and a fast camera (HiSpec1, Fastec Imaging,USA) to image encapsulation and drop fusion.

Cell cultures. Mouse embryonic stem cells from a male mouse embryo (v6.5, NBP1-41162, Novus, USA) were cultured on mitotically inactivated mouse embryonic fibroblasts (MEFs, Globalstem, USA). ES cells were maintained in medium containing Knockout DMEM (Gibco, USA), 15% Fetal Bovine serum, 1% Pen/Strep (Gibco, USA), 1% Non-essential amino acids (Gibco, USA), 1% Glutamax (Gibco, USA), 0.01% LIF (ESG1107, Millipore, USA) and 0.0004% beta-mercaptoethanol. Mouse embryonic fibroblasts (Globalstem, USA) were cultured in the same medium but without LIF. EML (CRL-11691, ATCC, USA) were grown in Iscove's modified Dulbecco's medium (IMDM) with 4 mM L-glutamine adjusted to contain 1.5 g/L sodium bicarbonate containing 200 ng/mL mouse stem cell factor, 1% Pen/Strep (Gibco, USA) and 20% Fetal Bovine serum. Human K562 cells were grown in DMEM (Gibco, USA), 20% Fetal Bovine serum, 1% Glutamax (Gibco, USA) and 1% Pen/Strep (Gibco, USA). Cell lines were tested for mycoplasma contamination and ES cells authenticated by measuring Oct4 levels, characteristic morphology and chromatin state.

Preparation of unlabeled chromatin. About 100M K562 cells were suspended in 1 mL of 1× digestion buffer. The suspension is incubated at 4 °C for 10 min to lyse the cells, after which MNase is activated by incubating at 37 °C for 15 min and inactivated by adding 40 µL of 0.5 M EGTA (final concentration of 20 mM). Next, we centrifuged the lysate for 5 min at max speed, separate the chromatin supernatant and mix it with 1 mL of 2× stopping buffer.

Barcode and primer design. The design of barcode adapters is shown in **Supplementary Figure 2a**. A sequence of 5 guanine nucleotides on each side of the barcode is not complementary and forms a loop. These loops were designed to prevent the formation of hairpins or stem-loops that inhibit priming during amplification of labels. The 1,152 barcode sequences are listed in **Supplementary Table 2**. To prime the barcoded genomic DNA, we use the following SC-PCR primer sequences:

TAAGGTGGGGGGGATAC 59.6(Tm)
TAAGGTCCCCCGGATAC 59.6(Tm)

Barcode library generation. Barcodes were commercially synthesized (IDT, USA) and suspended in 10 mM Tris at a concentration of 500 µM in 384-well plates. We use a 96 parallel drop-maker microfluidic chip with aqueous inlets for each drop-maker that precisely fit one quarter of a 384-well plate and that are immersed in 96 different wells, each containing a unique barcode. Oil with surfactant is distributed to all drop-makers via a common inlet that is connected to a pressurized (9 psi) oil reservoir. The plate and the microfluidic parallel device are placed in a pressure chamber while a common outlet for all 96 barcode drop-makers is located outside the pressure chamber. Upon pressurizing the chamber (6 psi), each of the 96 barcode solutions is forced through its own drop-maker, thereby forming an emulsion of ~35 µm diameter drops where every drop contains about 1 billion copies of one of the 96 barcodes. The process is repeated until all barcodes are encapsulated. Before use, the emulsion is pooled in a single tube and mechanically mixed by rolling the tube for 5 min.

Cell encapsulation. Cells were suspended at a concentration of 5 M/mL in PBS and loaded in a syringe together with a magnetic stirrer bar stirred by a motorized magnet located externally to prevent sedimentation of the cell suspension. The suspension of cells is co-flowed at a 1:1 ratio with 2× digestion buffer containing both a detergent for cell lysis and Micrococcal Nuclease (MNase). MNase is an endonuclease that digests single-stranded nucleic acids, but is also active against double-stranded DNA and under optimized conditions will preferentially digest the open DNA at the inter-nucleosomal regions, resulting in the fragmentation of chromatin into primarily mono-nucleosomes. The two aqueous phasescell suspension and buffer—meet immediately before passing through the microfluidic drop making junction so that they only mix inside the ~50-µm-diameter drops containing them (**Supplementary Video 1**). After encapsulation, drops were incubated at 4 °C for 10 min for lysis and then at 37 °C for 15 min for MNase digestion.

Barcode-cell drop fusion. Drops containing native chromatin from single-cells and drops containing barcodes are reinjected into a custom 3-point merger microfluidic device. The third inlet in the 3-point merging chip is fed with 2× labeling buffer, optimized for both end repair of dsDNA and blunt end ligation in the same solution. A high voltage amplifier (2210, TREK, USA) which supplies a 100 V square A/C wave at a frequency of 25 kHz is used to drive the device electrodes which induce an electric field that electro-coalesces the 3 phases (cell drops, barcode drops and labeling buffer). After merging, all fused drops are collected in a single tube preloaded with a bed of carrier drops that protect the sample drops from evaporating or wetting the tube walls. The carrier drops are 70 µm in diameter, similar to the size of the fused drops, and contain a carrier buffer optimized to match the mixed buffers in the fused drops, thereby minimizing the osmotic forces acting on the sample drops. To simplify the distribution of samples into wells downstream, we use 2 mL of carrier drops for every 10,000 cells collected. After collection, the mixed emulsion is incubated at room temperature for 2 h to allow ligation.

Extracting samples from fused drops. The 2 mL of emulsion containing fused drops and carrier drops are distributed in aliquots of 20 µL into wells containing 20 µL of 1% surfactant oil. This ensures that each well contains a sample of about 100 labeled cells. Each well is then topped with 50 µL stopping buffer that stops the ligation reaction and 25 µL of unlabeled chromatin from ~1M K562 cells. The unlabeled chromatin acts as a buffer, minimizing nonspecific binding during ChIP and protecting the minute amounts of labeled chromatin from being lost during liquid handling. To separate the emulsion, 10 µL of demulsifier is added to each well and the plate is centrifuged at 1,000 r.p.m. for 30 s. The aqueous phase in each well, containing labeled chromatin from ~100 cells, separates above the oil phase and is transferred to a new well for ChIP.

ChIP. Each sample of ~100 cells was incubated at 4 °C overnight with 1–3 µL of antibodies (see reagents). The complexes were precipitated with 20 µL of protein-A coated magnetic beads (10008D, Life Technologies, USA) in a total volume of ~125 µL per sample. Beads were washed sequentially twice with low-salt immune complex wash, twice with high-salt immune complex wash, once with LiCl immune complex wash, and twice with TE (10 mM Tris-HCl). Wash volumes are 100 µL per sample, except for the last wash, where the immunoprecipitated chromatin remains bound to the beads in 21.5 µL of TE per sample for downstream reactions and is eluted later in the library preparation.

Library preparation. To minimize the abundance of barcode adaptors concatemers, we add 1 µL of PacI restriction enzyme (R0547L, NEB, USA) and 2.5 µl of NEB Buffer 1 to each sample of 100 cells in 21.5 µL of TE and incubate at 37 °C for 2 h and then at 65 °C for 20 min. This is done immediately after ChIP washing steps and while the chromatin is still bound to the ChIP beads. PacI digest

doi:10.1038/nbt.3383

in between bound concatemers and in the middle of each adaptor to form 30 bp DNA fragments that can be easily filtered out using simple size selection (see **Fig. 3a** and **Supplementary Fig. 2a**). Next, we elute the chromatin by adding 25 µL of 2× elution buffer, digest RNA contaminates by adding 3 µL of RNase (11119915001, Roche Diagnostics, USA) and incubate at 37 °C for 20 min and remove the nucleosomes by adding 3 µL of Proteinase K (P8102S, NEB, USA) and incubating at 37 °C for 2 h and deactivating at 65 °C for 30 min. We purify the DNA using 1.5× AMPure XP beads (A63880, Beckman Coulter, USA) and follow with 14 rounds of Single-Cell-PCR (SC-PCR, **Supplementary Table 1**) to amplify the labeled DNA and with another purification using 1.1× AMPure XP beads. To reduce unspecific Illumina adaptor ligation we first dephosphorylate all 5′ ends by adding 1 µL pf Antarctic Phosphatase (M0289L, NEB, USA) and 2.5 µL of Antarctic Phosphatase Buffer in a total volume of 25 µL including the DNA and incubating at 37C for 30 min. We then purify the DNA using 1.1× AMPure XP beads, add 1 µL of BciVi enzyme (R0596L, NEB, USA) and 2.5 µl of NEB Buffer 4 in a total volume of 25 µL including the DNA and incubate at 37 °C for 1 h. This will specifically cleave the labeled DNA, leaving an A overhang at the 5′ end of all DNA fragments with single cell adapters. To ligate Illumina adapters, we purify DNA using 1.1× AMPure XP beads, reduce the sample volume to 4 µL via evaporation, add 0.5 µL Quick Ligase (M2200L, NEB, USA), 6 µL of 2× Quick Ligation Reaction Buffer and 1.5 µL Illumina adapters diluted 1:150 and incubate at room temperature for 15 min. Before amplifying the Illumina adapters we apply PacI again to digest concatemers that may have formed during the ligation step. For this, we first purify DNA using 0.7× AMPure XP beads and then use the same concentrations and incubation times as the first application of PacI. Finally, we purify DNA using 0.7× AMPure XP beads and amplify the Illumina adapters by adding 12.5 µL of PCR Mix (PfuUltra II Hotstart PCR Master Mix, 600850, Agilent Technologies, USA) and 0.5 µL of Illumina Primers at 25 µM in a total volume of 25 µL including the DNA and thermocycling (initial denaturation at 95 °C for 3 min, 14 rounds of 95 °C for 30 s, 55 °C for 30 s and 72 °C for 1 min, and final extension at 72 °C for 10 min). The amplified sample is purified one last time using 0.7× AMPure XP beads and then the DNA content is measured and the sample is sequenced.

Sequencing. We use Illumina HiSeq to sequence 2 × 60 bp paired end reads. The first 11 sequencing cycles are dark to prevent low complexity failure when reading the non-variable regions of the barcode adaptor.

Filtering single-cell reads. Barcodes are expected at the first 8 bp of the first read and bp #12–19 of the second read. Half of PacI recognition site "TTAA" will follow the barcode sequence, and the rest of the read is genomic. Since barcodes are symmetric, both ends may be sequenced, so several combinations for read #1 and read #2 are possible, all representing the same fragment, as shown in **Supplementary Figure 2b**. Reads with barcode sequences not matching any of the 1,152 barcodes were discarded. Remaining reads were aligned to mm9 genome using Bowtie2 (ref. 43) in paired end mode, trimming the first 23 bp on each 5′ end and discarding multi-mapped reads and reads that are longer than 1 kb (syntax: "bowtie2 -X 1000 -- trim5 23 -x mm9 -1 [read#1.fastq] -2 [read#2.fastq] --S [output.sam]"). Of the remaining distinct reads, only those reads with matching barcodes on both ends were saved, with the following exception: if two (and only two) barcodes happen to mutually label 10% or more of reads associated with either of the two barcodes, then those barcodes are treated as identical and all reads labeled by either or both barcodes are considered to have matching barcodes on both ends. This exception handles cases where two barcode drops fuse with one cell drop. Finally, to determine those barcodes that are associated with single cells, the numbers of reads per barcode were analyzed based on Poisson statistics (**Supplementary Note 3**). The reads associated with the chosen barcodes, along with their barcode of origin, were used in downstream analysis.

Visualizing and assessing precision and sensitivity of single-cell chromatin profiles. To visualize the information content attainable by Drop-ChIP (**Fig. 4a**), we selected 100 single-cell H3K4me3 profiles (50 ES cells and 50 MEFs). These examples were selected based on high read coverage over target regions. The reads from each single-cell profile were plotted across representative regions. Although these best case examples better illustrate the accuracy of the profiles, visualization of essentially any subset of single cells recapitulates similar enrichment over target regions. We calculated the precision of each single-cell profile from the fraction of reads overlapping known peaks, and sensitivity from the fraction of known peaks overlapping single-cell reads (peaks defined from corresponding bulk profiles).

Supervised classification of single-cell tracks into ES and MEF cell types. For 400 H3K4me3 tracks (200 ES cells and 200 MEFs), we calculated the fraction of reads overlapping with peaks specific to either ES cells or MEFs (based on bulk H3K4me3 profiles). We plotted the ES cell score of each single-cell vs. its MEF specific score, with both scores normalized to a maximum of 1. A simple comparison between the two scores correctly classifies cells with >95% accuracy (**Fig. 4c**).

Clustering ES cells, MEFs and EMLs based on H3K4me3 single-cell profiles. We counted reads intersecting with 5-kb genomic bins to produce a vector v of ~500,000 values for each of the cells. Next we binarized the data to reduce any possible bias that might originate from over-represented bins (for example, repetitive regions):

$$v = (v_1, v_2, v_3, \ldots)$$

$$b = (b_1, b_2, b_3, \ldots)$$

$$b_i = \begin{cases} 1 & v_i > 0 \\ 0 & v_i = 0 \end{cases}$$

To reduce noise we filtered out low coverage cells and non-informative bins by selecting only single cells that occupy at least 250 bins, and restricting the set of bins to only those that were occupied by at least 2% but no more than 50% of the single cells.

We divided each binary vector by the total number of nonzero bins to control for cell-coverage variability, and calculated pair-wise covariances:

$$\tilde{b} = \frac{b}{\sum_i b_i}$$

$$C_{\alpha\beta} = \begin{cases} cov(\tilde{b}_\alpha, \tilde{b}_\beta) & \alpha \neq \beta \\ 0 & \alpha = \beta \end{cases}$$

where α and β are indices for individual cells.

Finally, we used a divisive clustering algorithm to cluster the columns of C by applying the function "diana" from the "cluster" R package.

Peak calling. We use Scripture[44] with a segmentation length of 1,000–5,000 bp to identify enriched regions in chromatin profiles.

Chromatin signatures collection and analysis. To build our signature library, we first collected 314 available ChIP-seq data sets from GEO and ENCODE, called peaks for each data set using Scripture, and defined the signature as the set of all 5-kb genomic bins overlapping the peaks of a data set. Pearson correlations ρ_{ij} between signatures correspond to the degree of overlap of genomic regions between them, and we used the distance function $d_{ij} = 1 - \rho_{ij}$ to cluster the signatures by applying the R function hclust (using the complete linkage method). Finally, we set a threshold that cut the dendrogram into 91 biologically meaningful clusters each consisting of highly overlapping maps and manually chose a representative signature from each cluster, taking into account quality of data and biological relevance. The correlation between the 91 signatures is shown in **Supplementary Figure 7** and the signature names and their public sources are listed in **Supplementary Table 4**.

Clustering H3K4me2 using chromatin signatures scores. To cluster H3K4me2 single-cell profiles, we first calculated the coverage, or score of cells in each of the chromatin signatures: we binned the reads of each single cell in 5-kb genomic bins and then calculated the number of bins that overlapped with each signature profile to produce a matrix of 10,128 cells (9,207

doi:10.1038/nbt.3383

ES cells and 921 MEFs) over 91 signatures. We used two specific signatures, the H3K4me2 signature score of ES cells and MEFs, to filter out single-cell profiles with a low ChIP signal. For ES cells and MEFs separately, we compared the single-cell scores for the respective H4K3me2 signature to a distribution of signature scores obtained by randomly choosing reads from input ChIP-seq bulk experiments of the same cell type (Whole Cell Extract, WCE). We filtered out cells with H3K4me2 signature scores that are lower than the 95% percentile of the H3K4me2 signature score of WCE virtual single cells. 7,327 cells (6,432 ES cells and 895 MEFs) satisfied this criterion and were retained for the next step (these were also retained for unsupervised clustering using DIANA, which classified the two cell types at >95% purity). We normalized each cell for coverage and standardized (subtracted the mean and divided by standard deviation) the distribution of each signature variable over the remaining cells. We applied two distance metrics, Euclidean and Manhattan, to create two pairwise distance matrices and then separately applied the R agglomerative hierarchical clustering method hclust (using the complete linkage method) on each of the matrices. We found 4 to be the minimal number of clusters required to separate the ES cells and MEFs. Clustering using the two metrics agreed on 84% of the cells. To make downstream results less dependent on the choice of metric, we decided to keep only those cells on which both metrics agreed. As a final step of cleaning up potentially noisy data, we noticed that when we partitioned the data to 5 clusters, 3 large (>1,400 cells) ES clusters are formed, one clear MEF cluster, and an additional smaller, somewhat more mixed cluster (360 cells, 26 of which are MEFs), and we have discarded the cells in the last cluster remaining with 4,643 ES cells and 762 MEFs. All subsequent analyses of population heterogeneity in H3K4me2 (**Figs. 5** and **6**) use these 5,405 cells.

Multidimensional scaling (MDS) plots. For these plots (**Fig. 5b**,**c** and **Supplementary Fig. 8d**), we used ρ_{ij}, the Pearson correlation between signature-scores vectors of single cells, for the distance function: $d_{ij} = 1 - \rho_{ij}$. The MDS was calculated from a matrix of these distances using the isoMDS function in the MASS R package[45], which implements Kruskal's non-metric multidimensional scaling.

Analysis code. Analysis and plots were performed using Matlab, R and ggplot.

42. Mazutis, L. *et al.* Single-cell analysis and sorting using droplet-based microfluidics. *Nat. Protoc.* **8**, 870–891 (2013).
43. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
44. Guttman, M. *et al. Ab initio* reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).
45. Venables, W.N. *Modern Applied Statistics with S* 4th edn. (Springer, New York, 2002).

doi:10.1038/nbt.3383