Supplementary Table 1 – Buffer recipes.

2X digestion buffer (100mL)

	concentration				
reagent	stock	final	volume		comment
Triton® X-100	10%	2%	20	mL	
Tris-HCl	1M, pH 7.5	100mM	10	mL	
NaCl	5M	300mM	6	mL	
sodium deoxycholate	5%	0.20%	4	mL	
CaCl2	1M	10mM	1	mL	
H_2O			59	mL	
sodium butyrate	1M	10mM			25uL in 2.5mL of 2X buffer
Micrococcal Nuclease	100%	0.04%			1uL in 2.5mL of 2X buffer

2X labeling buffer (200uL)

	concentration				
reagent	stock	final	volume		comment
PEG 400	100%	30%	60	uL	
Quick Ligase buffer	10x	2x	45	uL	
ATP	10mM	1.75mM	35	uL	
EndIt enzyme	100%	8%	16	uL	
Fast Ligase	100%	8%	16	uL	
EGTA	0.5M	40mM	16	uL	
dNTPs	10mM	0.5mM	10	uL	
sodium butyrate	1M	10mM	2	uL	

carrier buffer (1mL)

	concentration				
reagent	stock	final	volume		comment
digestion buffer	2x	0.5x	250	mL	
PEG 400	100%	18%	175	mL	
EGTA	0.5M	20mM	40	mL	
PBS	1x	0.54x	535	mL	

stopping buffer (100mL)

	concentration				
reagent	stock	final	volume		comment
Triton® X-100	10%	1%	10	mL	
EGTA	0.5M	30mM	6	mL	
EDTA	0.5M	30mM	6	mL	
Tris-HCl	1M, pH 7.5	50mM	5	mL	
NaCl	5M	150mM	3	mL	
sodium deoxycholate	5%	0.10%	2	mL	
H_2O			68	mL	

2X ChIP elution buffer (100mL)

	concentration				
reagent	stock	final	volume		comment
NaCl	5M	600mM	12	mL	
SDS	10%	1%	10	mL	
EDTA	0.5M	10mM	2	mL	
Tris-HCl	1M, pH 7.5	10mM	1	mL	
H_2O			75	mL	
					need to incubate at 65 before use

SC-PCR (25uL)

	concentration				
reagent	stock	final	volume		comment
Hercules buffer	5x	1x	5	uL	
SC-PCR primers	50uM each	3uM each	1.5	uL	denature - 95C 3min
DMSO	100%	4%	1	uL	denature - 95C 30sec
dNTPs			0.5	uL	anealing - 55C 10sec
Hercules Taq Enzyme			0.5	uL	elongate - 72C 1min
DNA			17	uL	elongate - 72C 10min

Supplementary Table 2 – A list of 1152 sequences of the barcode adaptors used. See corresponding Excel file.

Supplementary Table 3A – Sequencing performance. The number of reads, alignment rate and distinct rate for every Hiseq run performed. Runs that did not use a full lane are marked with *.

Sequencing Run #	Cell types	Epitopes	# of samples	Million reads	Bases with PF>=Q30	% aligned
1	ES MEF EML	H3K4me3 H3K4me2	23	322	93%	75%
2	ES MEF ES-MEF mix	H3K4me3 H3K4me2	43	250	86%	79%
3*	ES	H3K4me2	13	89	82%	66%
4*	ES	H3K4me2	5	57	84%	77%
5	ES	H3K4me2	37	203	85%	70%

Supplementary Table 3B – Overview of experiments. The number of samples and single cells identified for each cell-type and epitope is presented, as well as the number of cells remaining after preprocessing of the clustering algorithms. The distribution of number of reads per cell is also presented for the cells remaining after filtering.

cell type	Epitope	# of samples	# of cells	single cells per sample	# of cells after QC	# of reads per cell (mean ± STD)
ES	H3K4me3	12	1155	100	314	542 ± 252
ES	H3K4me2	69	9207	133	4643	796 ± 200
MEF	H3K4me3	12	1279	111	376	544 ± 218
MEF	H3K4me2	14	921	68	762	634 ± 291
EML	H3K4me3	2	491	246	193	510 ± 256
mix	H3K4me3	10	1716	171.6	1020	381 ± 275

Supplementary Table 4 – A list of the 91 data-sets used to create the signature library. See corresponding Excel file.

Supplementary Table 5, Source Data for Figure 5A – The 4 clusters of H3K4me2 ES cells and MEFs. For each cluster, the mean across all cells is given in a separate row for each signature. Table is coded with colors matching the matrix in Figure 5A. See corresponding Excel file.

Supplementary Movie 1 – Cell encapsulation. A slowed down movie of ES cells as they are encapsulated together with digestion buffer in the microfluidic co-flow drop-maker.

Supplementary Movie 2 – Drop labeling. A slowed down movie showing barcode drops (small) and drops containing cellular chromatin (large) merge together with labeling buffer flowing into the T-junction.

Supplementary File 1 – design of microfluidic devices. This compressed folder includes 3 ACAD designs, one for the 96 parallel drop makers, one for the co-flow drop maker and one for the 3 point merger.

Supplementary Note 1 - Quality Control for clustering analysis

We checked that the (H3K4me2-based) ES+MEF clusters that we obtain are not sensitive to arbitrary technical aspects of the analysis. These investigations (detailed below) demonstrate that our results are not driven by some part of the data (robustness under omission of 1/2 the cells), are insensitive to the precise choice of signatures and are the same for two different cell-cell distance metrics. By randomly shuffling data between cells, we show that the correlations between the signatures that drive the clustering disappear.

Independence of clustering on choice of cells:

Of the 5,405 cells that passed QC filters, we randomly selected 50% (2,702 cells) and re-clustered. This was repeated 100 times. Each time, there was a striking correspondence between the new clusters, and the original ES1, ES2, ES3, and MEF clusters of the full-data analysis. For each such bootstrap sample, we calculated the fraction of cells that were assigned to their original cluster. The results (Supplementary Fig. 8B) show that very consistently more than 90% of the cells in any of the clusters are associated with the same cluster even when relying on only half the available data.

Independence of clustering on set of signatures:

To ensure that the clusters found in ES cells are not a consequence of the specific choice of signatures that we have made, we used a different set of 254 signatures (E. Meshorer, personal communication). We used hclust to collect the signatures in 100 groups, and selected a representative signature from each group ('Meshorer signatures'). We then calculated single-cell signature-score vectors for all cells, and proceeded with clustering as before to produce 4 clusters. We also repeated this procedure for the complete set of 314 signatures without any manual curation or filtering. The high degree of agreement between the resulting clusters and the ones obtained with the first set of signatures was used to associate the two sets of clusters, and the new clusters were named suggestively as 'ES1', 'ES2', 'ES3', and 'MEF'. Supplementary Figure 8C shows MDS plots for 1000 randomly sampled cells colored by the original clustering (top), and (using same MDS coordinates) colored according to the new clusters derived from the Meshorer signatures (middle) and all signatures (bottom). The plots show a high degree of similarity between the results of clustering by the three sets of signatures.

Clustering with randomly generated single cells:

It is important to verify that the clusters that we find are not a consequence of the choice of signatures (some evidence for this has already been presented above). This is particularly true because the clusters are driven by correlations between signatures, and the presence of overlapping genomic regions between signatures induces correlations that have nothing to do with true co-variation of signal over different genomic regions across single cells. With a handful of signatures one might address this by computationally eliminating overlaps between them. With 314 signatures, this is impractical. Instead, we clustered the signatures by their overlap-based correlations in an unsupervised manner and then picked one representative from each of the 91 clusters, as described in the Online Methods. Supplementary Figure 8C (top) shows on the Y axis the correlation between signatures based on co-variation across cells in the population, and on the X axis the overlap-based correlations. Each dot

represents a pair of signatures, and the red dots are correlations among the 91 representative signatures from each signature-cluster. There are two notable observations about the figure: a) the overlap-based clustering and representative selection are a good strategy - the overlap correlations between the representative signatures are much smaller than the general range of such correlations among all signatures; b) for all signatures, and for the representative ones in particular, the single-cell based correlations (and anti-correlations) much exceed the overlap-based ones.

Next, we randomly redistributed the reads between the single cells. In the reshuffled data set the number of reads in each single cell was preserved as well as the number of single cells supporting each genomic position. We then recalculated the cell-based correlations between the signatures (Supplementary Fig. 8C (bottom)). One sees that the correlations are greatly reduced in the random set compared to the real data, essentially being as low as the degree of overlap between them allows. This is strong evidence that the structure in the data that drives the ES clusters is an outcome of genuine cell-to-cell co-variation in the data, and not grounded in the signatures themselves.

Controlling for cell cycle:

To examine whether the ES clusters that we identified are related to cells in different cell-cycle stages, we used two sets of cycling genes - those that have higher expression in G1/S and those whose expression is higher at G2/M. We created a signature from each gene set, which includes all gene loci as well as distal regions marked by H3K4me2 (as inferred from aggregate single-cell data) that are within 50kb of a gene in the set. We generated signatures from these sets and then calculated the distributions of signature scores across the different ES clusters. The distributions of G1/S and G2/M signature scores show no significant difference between ES clusters in contrast to other biological relevant signatures such as OCT4, as presented in Supplementary Figure 8E. We thus do not find evidence that the ES cell sub-populations reflect different cell cycle stages or expression.

Supplementary Note 2 - Data comparison with single cell RNA expression

Expression and H3K4me2 enrichment trends correlate across cell populations

For this analysis we used RNA-Seq data from Kumar et al [1], which includes expression (tpm) of 8,885 genes in 171 mouse ES cells. As a first step, we centered and scaled each gene so that its average over the cell population is 0 and standard deviation 1. For each of the genomic signatures (described above), we collated all genes whose promoters overlap the signature. An expression score for a signature was defined as the difference between the mean (normalized) expression of the genes associated with the signature in a given cell, and the mean expression of all 8,885 genes in that cell. We thus computed normalized 'expression' scores for a signature across all single ES cells based on the expression of the associated genes (relative to overall gene expression).

We calculated the signature expression scores for 6 signatures that demonstrated robust patterns of variation across the single-cell chromatin data for ES cells. These include 2 polycomb-related signatures (Ezh2, Ring1B), and 4 pluripotency signatures (Tcf3, Oct4, Sox2, and Nanog). We removed genes that were associated with more than one signature from all but one signature, so that the final gene lists associated with each signature were completely disjoint.

Oct4	Sox2	RING1B	Tcf3	Ezh2	Nanog
1441	392	112	1668	574	1951

We then computed pairwise correlations between the signatures based on either (a) the H3K4me2 signature scores across single cells or (b) the expression scores across single cells. Figure 6B, C displays these correlations as heat maps, and the order of signatures was determined by unsupervised clustering. The pairwise correlations are plotted against one another in Figure 6D.

Clustering RNA-Seq data and agreement with chromatin-based clusters

We used K-means clustering to partition the 171 cells (using kmeans) into two clusters based on the 6 expression signatures. The histograms in Supplementary Figure 10 show the distributions of the signature scores across the two clusters, confirm that the signatures are indeed different between the two clusters, and demonstrate that cluster #1 shows higher expression for the polycomb-related signatures, while cluster #2 is shows higher expression for the pluripotency signatures (analogous to ES3 and ES1, respectively).

Finally, we used the two extreme ES cell clusters, ES1 and ES3, to produce two sets of genes: the first set includes genes with higher H3K4me2 in ES1 (446 genes); the second set includes genes with higher H3K4me2 in ES3 (47 genes). Supplementary Figure 10 shows the mean (normalized) expression of each gene set in each of the two clusters derived from the expression data. The ES1 gene set shows higher mean expression in expression cluster #2. The ES3 gene set shows higher mean expression in cluster #1. Thus, single-cell chromatin profiling and single-cell RNA profiling data for ES cells reveal similar patterns of variability at similar signature gene/element sets. Both approaches also identify two sub-populations, based on coherent differences in pluripotency and Polycomb signatures, with related patterns of enrichment at common gene/promoter sets.

Supplementary Note 3 - Poisson based statistics for choosing cell-representing barcodes

To distinguish barcodes representing single cells from barcodes representing multiple cells or from barcodes representing empty drops with background reads we use a Poisson model to represent the distribution of the number of reads per barcode. Under the assumption that each drop is associated randomly with any of the 1152 barcodes with a probability *1/1152* and that the association of each drop is independent of the other drops, we can use the Poisson distribution to describe the probability of finding, in a collection of n drops, x drops all carrying the same barcode:

n – total number of drops

$$P(x) = \mathrm{P}_{\lambda=\frac{n}{1152}}(x)$$

where

$$\mathrm{P}_{\lambda}(x) = \frac{\lambda^x}{x!}e^{-\lambda}.$$

The number of DNA fragments labeled within each drop is modeled using the gamma distribution, conveniently providing a non-negative distribution of values with a given mean μ and coefficient of variation *c*. Thus, the number of reads X associated with each of the 1152 unique barcodes distributes as the convolution of the two distributions:

X – the number of reads associated with each of the 1152 unique barcodes

$$N_{\mu,c,n}(X) = P_{\lambda}(1)\mathrm{G}_{\mu,c}(X) + P_{\lambda}(2)\sum_{k_1=0}^{X}\mathrm{G}_{\mu,c}(k_1)\mathrm{G}_{\mu,c}(X-k_1) + \cdots$$

With

$$\lambda = \frac{n}{1152},$$

$$\mathrm{G}_{k,\theta}(X) = \frac{\mathrm{X}^{k-1}}{\Gamma(\mathrm{k})\theta^k}e^{-\frac{x}{\theta}}$$

is the gamma distribution, and the conversion $\mu, c \Leftrightarrow k, \theta$ is given by $\mu = k\theta$ and

$$c = \frac{1}{\sqrt{k}}.$$

We fit the observed distribution of reads per barcode by a combined distribution of two populations of drops: n_e cell-free drops, with an average of $\mu_e \pm c$ % background reads per drop (Fig. 2D, blue) and n_s drops containing single-cells, with an average of $\mu_s \pm c$ % reads per cell (Fig. 2D, green). We optimize the values of n_e, n_s, μ_e, μ_s and *c* to fit the experimentally measured distribution of number of reads per barcode. We repeat this process for each experiment (see Supplementary Table 3B) and find that on average the coefficient of variation is $c = 1 \pm 0.2$, the signal to noise ratio, which is the ratio between the number of reads per single-cell drop to the number of reads per empty drop is

$SNR = \frac{\mu_s}{\mu_e} = 43 \pm 23.$

To isolate barcodes representing single-cells we choose a threshold number of reads X_{min} such that $1 - \sum_{X=1}^{X_{min}} N_{\mu_e,c,n_e}(\mathrm{X}) = 10^{-2}$

meaning that the probability of a barcode with X_{min} transcripts or more to be associated with cell-free drops is less than 1%; we also choose a threshold number of reads X_{max} such that $1 - \sum_{X=1}^{X_{max}} N_{\mu_s,c,n_s}(\mathrm{X}) = 10^{-2}$

meaning that the probability of a barcode with X_{max} transcripts or more to be associated with single-cell drops is less than 1%. Barcodes with X reads falling within this range $X_{min}<X< X_{max}$ are chosen as those barcodes representing single cells. The number of single-cell drops as a fraction of the number of barcodes is

$\frac{n_s}{\#\ of\ barcodes} = 14\% \pm 3\%.$

The number of such barcodes and the number of reads they contribute is summarized for all experiments in Supplementary Table 3B.

1. Kumar, R.M. et al. Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* **516**, 56-61 (2014).