EpiCHAOS: a metric to quantify epigenomic heterogeneity in single-cell data

Katherine Kelly[1,2], Michael Scherer[1], Martina Maria Braun[3], Pavlo Lutsik[1,4,*], Christoph Plass[1,*]

[1] Division of Cancer Epigenomics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

[2] Ruprecht Karl University of Heidelberg, Heidelberg, Germany

[3] Computational Biology and Health Genomics, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08003 Barcelona, Spain

[4] Department of Oncology, Catholic University Leuven, Leuven, Belgium

Email addresses:

katherine.kelly@dkfz.de, michael.scherer@dkfz.de, martina.braun@crg.eu, p.lutsik@dkfz.de, c.plass@dkfz.de

***Co-corresponding authors:**

Christoph Plass
c.plass@dkfz.de

Pavlo Lutsik
p.lutsik@dkfz.de

Abstract

Epigenetic heterogeneity is a fundamental property of biological systems, and is recognized as a potential driver of tumor plasticity and therapy resistance. Single-cell epigenomics technologies have been widely employed to study epigenetic variation between – but not within – cellular clusters. We introduce epiCHAOS: a quantitative metric of cell-to-cell epigenetic heterogeneity, applicable to any single-cell epigenomics data type. After validation in synthetic datasets, we applied epiCHAOS to investigate global and region-specific patterns of epigenetic heterogeneity across diverse biological systems. EpiCHAOS provides an excellent approximation of stemness and plasticity in development and malignancy, making it a valuable addition to single-cell cancer epigenomics analyses.

Keywords

Epigenetics, heterogeneity, cancer, single-cell epigenomics, chromatin, DNA methylation, plasticity

Background

Cell-to-cell heterogeneity can be found at multiple levels in all complex biological systems – ranging from that within populations of genetically diverse individuals, to the non-genetic functional heterogeneity between tissues and cell types, and more subtle molecular differences between cells within the same cell type[1]. In recent years this phenomenon of inter-cellular heterogeneity has received particular attention in the study of cancer as a potential driver of tumor progression and therapy resistance [2,3,4].

Although cancer formation begins with the clonal expansion of a single transformed cell, most tumors ultimately acquire extensive heterogeneity, with a single tumor consisting of several populations of cells with diverse phenotypic characteristics [2,5]. This is partly due to the emergence of genetically distinct subclones. However, epigenetic dysregulation is now also recognized as a hallmark of cancer, and it is

40 becoming increasingly evident that profound functional heterogeneity can arise within genetically identical
41 cellular structures due to changes in the epigenetic landscape, including DNA and histone modifications
42 and the larger chromatin architecture [6,7].
43
44 Outside of malignancy, recent studies have highlighted the essential role of cellular heterogeneity in
45 maintaining pluripotency and in shaping differentiation trajectories in the developing organism [8]. Generally,
46 multipotency is accompanied by increased stochastic molecular variation or "noise" which is lost as cells
47 differentiate and commit to a specific cell fate, in which a more stable transcriptional program is acquired
48 [9]. In the context of malignancy, such regulatory heterogeneity or "noise" could similarly act as a driving
49 force towards functional intratumor heterogeneity [9]. Here epigenetic heterogeneity could have potentially
50 detrimental implications. By creating a fitness advantage and improving a tumor's ability to adapt to a
51 variety of intrinsic and extrinsic stresses, heterogeneity increases the chances that part of the tumor will
52 tolerate a range of therapeutic insults, and thus represents a major clinical challenge [5]. Moreover, such
53 heterogeneity can allow the "division of labor" required for a tumor to function efficiently as a system. For
54 example, the functional diversity required to undertake the metastatic cascade – involving degradation of
55 extracellular matrix, invasion into local tissues, survival in transit through the bloodstream and ultimate
56 colonization in a distant organ – is an unthinkable feat for a homogeneous population of cells, but is likely
57 accomplished by a system of phenotypically heterogeneous cells, with each taking on different tasks to the
58 benefit of the whole [10 11].
59
60 An appreciation of intratumor heterogeneity is therefore essential to the study of cancer, however this has
61 historically been overlooked due to the inability of traditional bulk sequencing approaches to unmask
62 intrinsic variations within cell populations. Since the advent of single-cell sequencing technologies, it is
63 now possible to disentangle intratumor and microenvironmental heterogeneity at unprecedented resolution
64 at the genomic, transcriptomic and epigenomic levels [12]. Using technologies such as the Assay for
65 Transposase Accessible Chromatin with sequencing (scATAC-seq), the presence of epigenetically distinct

clusters within tumors – the simplest layer of epigenetic heterogeneity – has been demonstrated across a broad range of cancer types [13,14,15,16]. However, a more complex and underexplored problem is quantifying heterogeneity between cells within a given group or cluster.

A few recent studies have incorporated methods to quantify cell-to-cell heterogeneity at the transcriptional level using distance-based or entropy-based metrics, or to quantify transcriptional "noise" using metrics such as the coefficient of variation [17–19,20]. Others have taken advantage of read-level DNA methylation data to devise metrics of epigenetic heterogeneity that can be applied to bulk datasets [21–25]. However, such strategies have not been extended to single-cell applications and are therefore limited in discerning heterogeneity *between* from heterogeneity *within* cellular clusters. A metric for quantifying epigenetic heterogeneity in single-cell datasets has so far not been established.

To address this we developed epiCHAOS (Epigenetic/Chromatin Heterogeneity Assessment Of Single cells), a distance-based heterogeneity score designed to quantify cell-to-cell epigenetic heterogeneity using single-cell epigenomic data. After systematically validating epiCHAOS in a range of synthetic and real datasets, we demonstrated its ability to capture features of stemness and plasticity in development and malignancy. Additionally, we employed epiCHAOS for investigating region-specific and pathway-specific differences in epigenetic heterogeneity, and demonstrated its applicability to a range of single-cell epigenomic data types.

Results

EpiCHAOS reliably quantifies epigenetic heterogeneity in single-cell epigenomics data

We designed epiCHAOS to assign a heterogeneity score at the level of cell clusters, or other user-defined groups of interest e.g. cell types or treatment conditions. Here, we initially focused on scATAC-seq data as

89 the most commonly used single-cell epigenomics modality. To establish epiCHAOS scores, data from each
90 cluster were extracted as a binarized matrix, representing a peaks-by-cells or tiles-by-cells matrix in the
91 case of scATAC-seq. For each cluster we then computed the distances between all pairs of cells using a
92 count-centered version of the Jaccard distance [26], and then took the mean of all pairwise distances per cluster
93 as its epiCHAOS score (Figure 1A).

94 To assess the performance of epiCHAOS we generated a range of synthetic scATAC-seq datasets in which
95 we controlled the levels of heterogeneity, while maintaining a stable total count (see "Methods" for details).
96 We showed that the epiCHAOS score is highly correlated with the true controlled heterogeneity (Pearson
97 R=0.99) (Figure 1B). Next, we verified that heterogeneity can be detected both in cases of increasing and
98 decreasing counts representing a genome-wide loss/gain of chromatin accessibility. To achieve this we
99 simulated a series of scATAC-seq datasets in which we incrementally increased heterogeneity, while either
100 increasing or decreasing genome-wide chromatin accessibility (see "Methods" for details). We showed that
101 epiCHAOS correctly detects differences in heterogeneity both in cases where counts are added and removed
102 (Figure 1C).
103
104 To confirm that epiCHAOS does not perceive differences in sparsity as differences in heterogeneity, we
105 generated a series of random datasets with varying total number of 1's. We found no correlation between
106 epiCHAOS scores and total number of 1's between datasets (Figure 1D). In real datasets however,
107 differences in coverage may be more complex since they are accompanied by differences in the number of
108 missing values representing false negatives. Thus, we found that higher heterogeneity can be perceived in
109 data with lower total number of fragments. Genome-wide differences in detected ATAC signals can arise
110 from both technical or biological reasons, for example due to the presence of quiescent cells or in cells with
111 copy number alterations (CNAs) (Additional file 1: Figure S1A-D). To exclude any effect of this on
112 epiCHAOS scores we implemented a linear regression-based adjustment for the genome-wide chromatin
113 accessibility across cell clusters, which gave us a count-adjusted heterogeneity score. We showed that this

114 adjusted score is no longer affected by differences in genome-wide chromatin accessibility and is robust to
115 the presence of large-scale deletions and gains in tumor samples (Additional file 1: Figure S1A-D).
116
117 We further validated epiCHAOS using *in silico* mixtures of cell types from the human hematopoietic
118 system. We selected five distinct cell types from a previously published hematopoietic dataset [27] including
119 hematopoietic stem cells (HSCs), monocytes, B-cells, CD8-T cells and plasmacytoid dendritic cells
120 (pDCs). After reducing the peaks matrix to regions which are most differentially accessible between cell
121 types (500 peaks for each cell type), we created mixtures of two to five cell types in all possible
122 combinations, and applied epiCHAOS to each individual cell type and mixture. As expected, epiCHAOS
123 scores were relatively low in individual cell types and increased with the number of cell types in the mixture
124 (Figure 1E). We also confirmed that epiCHAOS was not influenced by technical noise or other potential
125 technical confounders (Additional file 1: Figure S1E-H, Additional file 2: Supplementary Note 1).

126 **EpiCHAOS reflects epigenetic heterogeneity associated with developmental plasticity**

127 Previous studies have recognized epigenetic heterogeneity as a feature of uncommitted, multipotent cell
128 states, which decreases in more committed, differentiated cells [9, 28]. To test if epiCHAOS scores align with
129 this expectation we applied our score to scATAC-seq data from a range of developmental contexts. First
130 we tested epiCHAOS in the human hematopoietic system [27,28], where we detected highest epigenetic
131 heterogeneity in hematopoietic stem and progenitor cells (HSCs), Common Lymphoid Progenitors (CLPs),
132 Common Myeloid Progenitors (CMPs), Lympho-myeloid primed progenitor cells (LMPPs), and early
133 erythroid cells. In contrast, more differentiated cells of the myeloid, erythroid and lymphoid lineages
134 showed lower epiCHAOS scores (Figure 2A). Similarly, in data from mouse gastrulation[29], epiCHAOS
135 scores were highest in less differentiated cells including cells of the epiblast, primordial germ cells and in
136 the primitive streak and decreased throughout the formation of distinct meso-, endo- and ectodermal
137 lineages (Figure 2B). In Drosophila embryogenesis, we also observed high epigenetic heterogeneity at the
138 earliest multipotent stages including undifferentiated cells, blastoderm and germ cells (Figure 2C). Among

139 more differentiated tissues, we also detected high heterogeneity scores in the neuronal compartments, likely
140 due to their extraordinary functional heterogeneity [30]. We confirmed using the Descartes scATAC-seq tissue
141 atlas that higher heterogeneity in neural tissues persists outside of an embryonic context [31] (Additional file
142 1: Figure S2).
143
144 To validate these observations we correlated our score with CytoTRACE[32] – a scRNA-based metric
145 designed to capture stemness/plasticity – using scRNAseq data from hematopoietic cells[27]. EpiCHAOS
146 correlated moderately with CytoTRACE scores (Figure 2D), in some cases following a pattern better
147 representative of the differentiation trajectory than CytoTRACE. For example, epiCHAOS detected higher
148 heterogeneity in naive CD8+ and CD4+ T-cells compared to memory T cells. Naive T cells are expected to
149 be more developmentally plastic, but are not detected as such by CytoTRACE. This suggests that
150 epiCHAOS captures some features of cellular plasticity that might not be detectable at the transcriptional
151 level. Similarly in gastrulation and embryogenesis datasets, epiCHAOS correlated with previously
152 annotated metrics of developmental time (Figure 2E-F). This pattern of epigenetic heterogeneity was also
153 in most cases reflected by higher transcriptional heterogeneity in less differentiated cells (Additional file 1:
154 Figure S3).
155
156 Collectively these data suggested that epiCHAOS can provide an accurate approximation of epigenetic
157 heterogeneity indicative of developmental plasticity.

158 **EpiCHAOS correlates with features of plasticity in malignant cells**

159 To investigate epigenetic heterogeneity in malignancy we selected two previously published scATAC-seq
160 datasets from 16 breast[13] and 16 liver[34] cancer patients. We subsetted each dataset for epithelial cells and
161 applied epiCHAOS to the resulting epithelial clusters (Figure 3A, Additional file 1: Figure S4A-B). To
162 understand what features of tumor cells coincide with higher epigenetic heterogeneity, we assigned

163 molecular signature scores to each cluster using the scATAC-seq gene score matrices. Within each tumor
164 type we correlated epiCHAOS against each molecular signature.
165
166 In breast cancer, we noticed that epiCHAOS scores were highly correlated with gene sets relating to
167 epithelial to mesenchymal transition (EMT) – a process which is considered integral to breast cancer
168 plasticity and metastasis [35] (Figure 3B). EMT-related signaling pathways such as TGF-beta and WNT
169 signaling were also correlated with epiCHAOS scores (Additional file 1: Figure S4C). By contrast, gene
170 sets related to estrogen receptor signaling, which are associated with better breast cancer prognosis, were
171 anticorrelated with epiCHAOS scores (Additional file 3: Table S7). Clusters with high epiCHAOS scores
172 also had higher accessibility at genes within various previously described gene expression signatures of the
173 more aggressive invasive and metaplastic breast cancer subtypes (Additional file 1: Figure S4C). Similarly,
174 in liver cancer, EMT-related gene sets as well as previously derived liver cancer prognostic signatures were
175 increased in clusters with high epiCHAOS scores (Figure 3C, Additional file 1: Figure S4C).
176
177 To support a suspected link between epigenetic heterogeneity and plasticity/stemness we applied
178 epiCHAOS to scATAC-seq data from childhood ependymoma comprising multiple differentiated
179 (astrocytes, ependymal cells, neural progenitor cells, and mesenchymal-like cells) as well as
180 undifferentiated tumor cell types [36]. Here, we detected the highest epiCHAOS scores in undifferentiated
181 cell populations, which are known to be enriched in more aggressive ependymomas [37,38] (Figure 3D).
182 Among malignant cell types, the lowest heterogeneity scores were detected in ependymal cells (Figure 3D),
183 which represent the most differentiated ependymoma tumor cells and which are associated with less
184 aggressive disease [37].
185
186 Overall these data support a connection between epigenetic heterogeneity quantified with epiCHAOS and
187 cancer cell plasticity and aggressiveness.

188 **EpiCHAOS reveals elevated levels of epigenetic heterogeneity at PRC2 targeted regions and**
189 **promoters of developmental genes**

190 Beyond comparisons of global epigenetic heterogeneity between cell states, we also wanted to apply
191 epiCHAOS to investigate differences in epigenetic heterogeneity at various classes of genomic loci.
192
193 Focusing on the hematopoietic system, we utilized annotated chromatin/transcription factor binding sites
194 from the Encyclopedia of DNA Elements (ENCODE) database to investigate differences in heterogeneity
195 across different genomic regions. We found that epiCHAOS scores were especially high at Polycomb
196 Repressive Complex 2 (PRC2) targeted regions, as well as at binding sites for CCCTC-binding factor
197 (CTCF) and Cohesin (Figure 4A, Additional file 3: Table S10). Per-region heterogeneity scores tended to
198 be highly correlated between different hematopoietic cell types (Additional file 1: Figure S5A). We
199 validated this finding by comparison to DNA methylation variation in hematopoietic cells as an independent
200 measure of epigenetic heterogeneity. At the DNA methylation level, epigenetic heterogeneity was similarly
201 highest at PRC2 targets, while binding sites for CTCF and Cohesin were also among the most variable
202 regions (Figure 4B). Visualization of peaks-by-cells matrices at examples of regions with high and low
203 epiCHAOS scores reflected the reported differences in heterogeneity (Figure 4C). To understand whether
204 this finding translates to the transcriptional level, we also calculated transcriptional noise for each gene in
205 HSCs based on the coefficient of variation. Consistent with previous reports [39,8], and in line with their
206 epigenetic heterogeneity, we found that transcriptional noise was significantly higher at PRC2 targeted
207 genes compared to non-PRC2 targets (Additional file 1: Figure S6A).
208
209 To further understand the characteristics of regions displaying high epigenetic heterogeneity, we computed
210 epiCHAOS scores at promoter-associated scATAC-seq peaks over a range of gene sets. Gene sets with the
211 highest heterogeneity scores were enriched for developmental processes such as cell fate commitment, cell
212 fate specification and somatic stem cell division (Figure 4D, Additional file 3: Table S11). Per-gene-set

213 heterogeneity scores were correlated between cell types (Additional file 1: Figure S5B), however certain
214 gene sets showed more cell-type specific heterogeneity patterns. For example we noted that genes related
215 to cell fate specification which were among the most heterogeneous gene sets in HSCs, ranked lower in
216 heterogeneity in more differentiated cell types (Additional file 1: Figure S6B). Similarly, promoters of
217 bivalent genes ranked among the most heterogeneous gene sets in HSCs and less so in more differentiated
218 cell types (Figure 4E).
219
220 We wondered whether certain genomic regions contributed to the differences in heterogeneity we had
221 observed between hematopoietic stem/progenitors and more differentiated cell types. Comparing HSCs to
222 monocytes we found that the largest increase in heterogeneity in HSCs was at Enhancer Of Zeste 2 (EZH2)
223 binding sites, followed by BRF2 and KAP1 binding sites (Figure 4F). Comparison of HSCs to differentiated
224 B-cells and CD8 T-cells yielded a similar pattern, suggesting that epigenetic heterogeneity at these key
225 genomic regions might contribute to maintaining pluripotency (Additional file 1: Figure S6C-D).

226 **EpiCHAOS is applicable to single-cell epigenomics data from any modality**

227 After demonstrating epiCHAOS' capability in scATAC-seq data, we wanted to test its application to other
228 epigenomics layers. For this we first applied epiCHAOS to single-cell nucleosome, methylation and
229 transcription sequencing (scNMT-seq) DNA methylation data from mouse gastrulation[40]. Here we observed
230 highest epigenetic heterogeneity in the epiblast compared to more differentiated germ layers (Figure 5A),
231 as well as a moderate correlation between promoter-wide DNA methylation-based and ATAC-based
232 epiCHAOS scores from the same cells (Figure 5B).
233
234 Next we utilized single-cell Targeted Analysis of the Methylome (scTAM-seq) [41] data from mouse
235 hematopoiesis [42]. In line with our previous observations at the level of chromatin accessibility (Figure 2A),
236 we found that DNA methylation-based heterogeneity was increased in more primitive hematopoietic
237 progenitor cells such as HSCs and early MPPs compared to more differentiated progenitors such as GMPs

238 and pre-B cells (Figure 5C). Notably, we found a continuous decrease of epiCHAOS scores along the
239 myeloid differentiation trajectory (HSC, MPP3, MPP4, GMP), while erythroid progenitors had elevated
240 levels of heterogeneity in line with the scATAC-seq data shown in Figure 2A.
241
242 Finally, applying epiCHAOS to H3K27me3 single-cell chromatin immunoprecipitation sequencing
243 (scChIP-seq) data from a previous study of breast cancer therapy resistance[43], we detected higher epigenetic
244 heterogeneity in resistant compared to sensitive cells, supporting the idea that intratumor epigenetic
245 heterogeneity may serve as a driver of treatment resistance (Figure 5D).
246
247 We conclude that epiCHAOS is applicable to all single-cell epigenomics data types and not restricted to
248 scATAC-seq data. This proves its utility across a wide range of epigenomic studies and enables the
249 comparison of heterogeneity in multi-modal datasets.

250 Discussion

251 Single-cell sequencing technologies offer unparalleled possibilities for studying epigenomic heterogeneity
252 in development and disease. However, few studies have tackled the question of how cell-to-cell
253 heterogeneity can be quantified at the level of cell populations, beyond the presence of distinct cell states.
254 For instance, with current metrics one is unable to rank clusters or groups of cells by their heterogeneity
255 levels. Here we introduce epiCHAOS: a quantitative metric for intercellular epigenetic heterogeneity
256 computed at the single-cell level. We demonstrate how epiCHAOS can be used to provide unprecedented
257 insights into the patterns of epigenomic variation across a wide range of applications including development
258 and malignancy.
259
260 We have demonstrated the functionality of our method by applying epiCHAOS to various scATAC-seq
261 datasets to investigate both global and region-specific heterogeneity. Our findings across a range of

262 developmental contexts align with the accepted notion that less differentiated and more developmentally
263 plastic cell types have higher heterogeneity compared to more differentiated and functionally specialized
264 ones – a phenomenon that has never been demonstrated in single-cell epigenomics data [9,28]. Moreover, we
265 provide evidence to support epigenetic heterogeneity as a hallmark of plasticity in cancer, paving the way
266 for more detailed pan-cancer analyses in this domain.
267
268 Our findings regarding the preferential heterogeneity of specific genomic regions are also in agreement
269 with previous studies of transcriptional noise and DNA methylation-based heterogeneity. For example Kar
270 *et al.* have previously shown that PRC2 targets display higher cell-to-cell variation in gene expression, with
271 a low transcriptional burst frequency giving rise to oscillations between on and off states over time [39]. The
272 idea that genes with bivalent chromatin states exhibit higher transcriptional noise has also been previously
273 shown by Faure *et al.* [44] and Kumar *et al.*[8] Similarly, Feinberg & Irizarry have demonstrated that DNA
274 methylation stochasticity is increased at developmental genes [45]. Here we demonstrate that these patterns
275 of variability can be observed at the epigenetic level between single cells, and that they appear not only to
276 be the most heterogeneous regions, but also the regions that are most preferentially heterogeneous in stem
277 compared to more differentiated cells. It is notable that PRC2 targeted regions are also known to exhibit
278 increased epigenetic variation across multiple cancers, compared to normal tissues [46]. This supports the idea
279 that epigenetic heterogeneity at these regions might be a driving force in tumor evolution and raises
280 interesting questions about the frequent disruption of PRC2 complex components in cancer [47].
281
282 Understanding how epigenetic heterogeneity is shaped throughout tumor evolution e.g. initiation,
283 progression, remission and relapse, and deciphering its role in the progression to metastasis and therapy
284 response could have important translational implications. EpiCHAOS should therefore be especially
285 valuable to cancer researchers, particularly those interested in plasticity and stemness, where the role of
286 epigenetic heterogeneity remains underexplored. Outside this domain, we expect that our metric should be
287 broadly applicable for studies of heterogeneity for example in developmental biology, aging and immunity.

Conclusions

EpiCHAOS provides a quantitative metric of cell-to-cell epigenetic heterogeneity, which should complement single-cell epigenomics studies of cancer and development. We have shown that epiCHAOS offers an excellent approximation of stemness and plasticity in various developmental contexts, as well as in cancer. Comparison of epiCHAOS scores at different genomic regions highlighted increased heterogeneity of polycomb targets and developmental genes. EpiCHAOS is applicable to a variety of single-cell epigenomics data types including measurements of chromatin accessibility, DNA methylation and histone modifications.

Materials and Methods

Calculating epiCHAOS scores

Heterogeneity scores were calculated for each given cluster or otherwise defined group of cells using a binarized matrix of scATAC or other single-cell epigenomics data. First pairwise distances were calculated between all cells within the cluster using a chance-centered version of the Jaccard index which controls for differences in the relative number of ones and zeros [26]. Afterwards, the mean of all pairwise distances per cluster was computed. To remove any further effect of sparsity a linear regression model of the raw heterogeneity scores was fit against the total number of detected accessible peaks, averaged across cells in the respective cluster, and the residuals of this model were taken as the adjusted scores. Finally, the scores were transformed to an interval of 0-1, and subtracted from 1 to convert the similarity metric to a distance metric.

ScATAC-seq data from the Hep-1 liver cancer cell line [48] was used to test whether epiCHAOS scores correlate with measures of technical noise. Cells were stratified into bins (20 bins of 100 cells each) based

310 on various quality control metrics: FRIP scores, TSS enrichment scores and nucleosome ratio, which were
311 calculated using ArchR [49].
312
313 DNA CNAs were inferred in the Hep-1 cell line using epiAneuFinder [50] with a windowSize=100,000 and
314 minFrags=20,000. To investigate the influence of CNAs on epiCHAOS scores, the most prominent
315 examples of large subclonal copy number gains (gain on chromosome 5) and deletions (deletion on
316 chromosome 13) were selected by visual inspection, and cells were stratified based on the presence or
317 absence of each alteration. EpiCHAOS scores were calculated across peaks in the affected chromosome
318 and compared between cells with diploid or alternative states. To correct for CNAs when applying
319 epiCHAOS to cancer datasets, a per-chromosome count-corrected epiCHAOS score was derived, where
320 epiCHAOS scores were calculated per chromosome, implementing a linear regression-based adjustment
321 for the total coverage on that chromosome, and then the average of per-chromosome scores was taken as
322 the global epiCHAOS scores.
323
324 Unless otherwise specified, epiCHAOS was calculated using the entire peaks-by-cells matrix. ENCODE
325 Transcription Factor Binding Site (TFBS) regions from the LOLA core database[51] were used for
326 comparisons of heterogeneity at different genomic regions, for which scATAC peaks matrices were
327 subsetted to obtain peaks overlapping with each genomic region. Similarly for comparisons across gene
328 sets, data were subsetted for peaks overlapping with promoters of each gene set using the GO:BP gene sets
329 from MsigDB[52].

330 **Generating synthetic datasets with controlled heterogeneity**

331 To test the performance of epiCHAOS, synthetic datasets were generated *in silico* in a way that simulates
332 the structure of binarized scATAC-seq peak matrices. First a series of 100 synthetic datasets with controlled
333 heterogeneity was created, in which each dataset has an equal total count. To do this a random binary matrix
334 was created, which would represent the first dataset in the series. In each subsequent dataset, homogeneity

335 was incrementally introduced by removing a set number of 1's from selected n rows, and adding them to a
336 different selected n rows, in such a way that a constant number of 1's is maintained, while heterogeneity
337 decreases.

338 To test situations where the genome-wide chromatin accessibility is increasing or decreasing, binarized data
339 from an example scATAC-seq dataset were perturbed to create datasets of increasing heterogeneity with
340 either addition or removal of 1's. Specifically, 10, 20, 30, 40 and 50% of 1's were selected at random and
341 replaced by 0's, and corresponding numbers of 0's were selected at random and replaced by 1's.

342 To test that epiCHAOS is not influenced by differences in sparsity, a series of 100 random binary datasets
343 was generated with each dataset having equal dimensions and incrementally increasing total number of 1's.
344 Their epiCHAOS scores were then computed and tested for a correlation with their total count.

345 As an additional validation approach semi-synthetic scATAC-seq datasets were created by mixing data
346 from distinct cell types. Using scATAC-seq data from human bone marrow [27] five cell types were selected;
347 HSCs, Monocytes, CD8-CM T cells, B-cells and plasmacytoid DCs. The top 500 differentially accessible
348 peaks for each cell type were identified and used to create *in silico* mixtures of two to five cell types in all
349 possible combinations.

350 **scATAC-seq data processing and analysis**

351 Publicly available scATAC-seq datasets for human hematopoiesis [27], mouse gastrulation [29], drosophila
352 embryogenesis [33], breast cancer [13], liver cancer, [34] ependymoma [36] and liver cancer cell lines [48] were
353 downloaded from the respective publications. For analyses in developmental datasets and in ependymoma
354 processed counts matrices were used as provided by the authors, where cell types were previously
355 annotated. For breast and liver cancer datasets fragments files were downloaded, processed and analyzed
356 using ArchR[49]. Cells with TSS enrichment scores < 4 or number of fragments < 1000 were removed, and
357 doublets were filtered out using default parameters. Iterative LSI was performed followed by clustering

358 using the Seurat method. Gene score matrices were generated using ArchR and used for subsetting cancer
359 datasets for epithelial cells based on epithelial cellular adhesion molecule (EPCAM) scores. After
360 reclustering epithelial cells, peak calling was performed using macs2[53]. To assign gene set/pathway scores
361 to each cluster, gene set annotations were obtained from MSigDB using the msigdbr R package[52]. Gene
362 scores were first averaged across all cells within each cluster, and then the mean score of all genes within
363 a given gene set was computed to assign gene set scores per cluster.

364 **Differential heterogeneity analysis**

365 Differential heterogeneity analyses were performed for each region using a permutation approach, whereby
366 the difference in epiCHAOS scores between two selected cell types were compared with that between pairs
367 of 1000 randomly computed groups sampled from the same pool of cells. P-values were calculated as the
368 quantile of the distribution of sampled permutations for which the difference in heterogeneity scores was
369 greater than that between the two test groups.

370 **DNA methylation variation**

371 Data from Adelman et al. 2019 [54] was used to calculate DNA methylation variability by computing variance
372 per CpG site in HSC-enriched lineage-negative (Lin− CD34+ CD38−) samples across the eight male
373 donors. To address the issue of data sparsity, a maximum quantile threshold of 0.005 was established for
374 missing values per site. Any sites that surpassed this threshold were removed. For each ENCODE TFBS,
375 the average of variances was calculated for all CpG sites overlapping with the respective regions.

376 **scRNA-seq analyses**

377 Publicly available scRNAseq datasets for human hematopoiesis[27], mouse gastrulation [29] and drosophila
378 embryogenesis[33] were downloaded from the respective publications and analyzed using Seurat. Cell-to-cell
379 transcriptional heterogeneity was calculated by computing pairwise Euclidean distances according to the
380 methods of Hinohara *et al.* [19]. Developmental potential was calculated per cell using CytoTRACE [32] and

assigned as a mean per cell type for downstream analyses. Transcriptional noise per gene was estimated using the coefficient of variation as previously described [55]. A list of PRC2-target genes used for comparison of transcriptional noise was obtained from Ben-Porath *et al.* [56]

scChIP-seq analysis

scChIP-seq counts matrices representing 50kb non-overlapping bins of H3K27me3 from human breast cancer patient-derived xenograft cells that were sensitive or resistant to Capecitibine (HBCx-95 and HBCx-95-CapaR) were downloaded from GSE117309 and processed as described by Grosselin et al. [43]. Cells having a total number of unique reads in the upper percentile were removed as outliers, and genomic regions not represented in at least 1% of all cells were filtered out. Data corresponding to non-standard chromosomes and the Y chromosome were excluded. Cells with a total number of unique reads < 1600 were removed. Counts matrices were binarized and cells from each condition were subsampled to select five groups of 100 cells each for epiCHAOS calculation.

scNMT-seq and scTAM-seq analysis

scNMT-seq DNA methylation and ATAC data from mouse gastrulation [40], summarized across promoters, gene bodies and CpG islands, were accessed using the SingleCellMultiModal R package. scTAM-seq data from mouse hematopoiesis were obtained from Scherer et al.[42], downloaded from https://figshare.com/ndownloader/files/42479346, and analyzed using Seurat.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All datasets analyzed in this article are publicly available and can be accessed through their associated publications as mentioned in the "Methods" section. R functions for calculating epiCHAOS scores and additional analysis scripts used throughout the article are deposited on GitHub at CompEpigen/epiCHAOS (github.com).

Competing interests

The authors have no competing interests.

Funding

The work was funded by the German Research Foundation, DFG, SFB1074 subproject B11N (CP), the Carreras Foundation (CP), DKFZ Dr. Rurainski Postdoctoral Fellowship (MS) and the Helmholtz International Graduate School (KK).

Authors' contributions

KK developed the method, devised and performed the majority of analyses. CP proposed the project idea and supervised the work. PL co-supervised the work and proposed strategies for assessing differential heterogeneity and accounting for copy number alterations. MS contributed to the overall design and planning of the manuscript, proposed the comparison with DNA methylation-based heterogeneity and the application of the method to single-cell DNA methylation data. MB performed the calculation of DNA methylation variation. KK wrote the manuscript and CP, MS and PL all discussed and contributed to its final version.

Acknowledgements

We would like to thank Karsten Rippe, Isabelle Seufert, Sebastian Waszak, Charles Dussiau, Moritz Mall, Manu Saraswat, and members of the division of Cancer Epigenomics at the DKFZ for insightful discussions about the method. We are also thankful to Thomas Hielscher for statistical recommendations, and Dieter Weichenhan for comments on the manuscript.

Figure Legends

Figure 1. EpiCHAOS reliably quantifies epigenetic heterogeneity in single-cell epigenomics data

A. Schematic describing epiCHAOS calculation. Using single-cell epigenomics data in binarized matrix form, epiCHAOS scores are assigned per cluster by computing the mean of all pairwise cell-to-cell distances using a chance-centered Jaccard index followed by regression-based adjustment for sequencing depth. μ=mean per cluster. **B.** Scatter plot illustrating correlation between epiCHAOS scores (epiCHAOS) and controlled heterogeneity across 100 synthetic datasets. Pearson correlation coefficient shown. **C.** Barplots illustrate increasing heterogeneity after perturbation of scATAC-seq data from sorted monocytes by either randomly adding or randomly removing 10-50% of 1's. **D.** Scatterplot illustrating the absence of correlation between epiCHAOS scores (epiCHAOS) and total coverage across a series of 100 randomly generated datasets ordered by increasing total counts. Pearson correlation coefficient shown. **E.** Validation of epiCHAOS using *in silico* mixtures of hematopoietic cell types. UMAP embedding illustrates scATAC profiles from five selected cell types of human bone marrow [27]. After selecting 500 top differentially accessible peaks for each cell type, *in silico* mixtures of two to five cell types in all possible combinations were created. Boxplots show the relationship between epiCHAOS scores (epiCHAOS) and number of cell types (y-axis) after *in silico* mixing.

444 **Figure 2. EpiCHAOS reflects epigenetic heterogeneity associated with developmental**
445 **plasticity**

446 **A-C.** Barplots (left) showing epiCHAOS scores (epiCHAOS) computed in scATAC-seq data from **A.**
447 human hematopoiesis [27], **B.** mouse gastrulation [29] and **C.** drosophila embryogenesis [33]. EpiCHAOS scores
448 were computed per-cell type as annotated in the original publications. Plots are ordered by epiCHAOS
449 scores and progenitor cells and undifferentiated tissue types are highlighted in blue. UMAP embeddings
450 (right) illustrating epiCHAOS scores in the same datasets as in barplots. UMAPS are colored by epiCHAOS
451 scores computed per annotated cell/tissue type. **D-F.** Scatter plots correlating epiCHAOS scores
452 (epiCHAOS) with developmental time as defined by (D) CytoTRACE score, averaged across cells (human
453 hematopoietic system), (E) developmental time in days at sample collection, averaged across cells (mouse
454 gastrulation), or (F) predicted developmental time in hours (Drosophila embryogenesis). Linear regression
455 lines and R-squared values are displayed.

456 **Figure 3. EpiCHAOS correlates with features of plasticity in malignant cells**

457 **A.** Schematic describing breast and liver cancer datasets used for epiCHAOS calculation. **B-C.** Dot plots
458 illustrate ordered Pearson coefficients after correlation of per-cluster epiCHAOS scores against gene set
459 scores for all Hallmark Gene Ontology biological processes in breast (B) and liver cancer (C). Top 5
460 correlations in each dataset are highlighted and labeled below. **D.** UMAP embedding (left) of scATAC-seq
461 data from five primary and two metastatic childhood ependymoma tumors [36]. Cells are colored by
462 epiCHAOS scores (epiCHAOS) computed for each cell type. Undifferentiated cells are highlighted in red.
463 Bar plot (right) ordered by epiCHAOS scores (epiCHAOS) for all malignant and non-malignant cell types
464 annotated in ependymoma tumors. Malignant cell types are highlighted in blue.

465 **Figure 4. EpiCHAOS reveals elevated levels of epigenetic heterogeneity at PRC2 targeted**
466 **regions and promoters of developmental genes**

467 **A.** Ordered dot plot showing epiCHAOS scores (epiCHAOS) computed across region sets for each
468 ENCODE chromatin factor binding site in hematopoietic stem cells, ordered by epiCHAOS scores. Top 20
469 region sets (cell type and binding site) are labeled. **B.** Scatter plot comparing epiCHAOS scores
470 (epiCHAOS) with DNA methylation variation at each transcription factor binding site as in (A). PRC2
471 targets (binding sites for EZH2/SUZ12, red), CTCF targets (orange) and Cohesin binding sites (binding
472 sites for RAD21/SMC3, blue) are highlighted. X-axis represents the average of per-CpG methylation
473 variances across 10 individuals at all CpGs overlapping with the respective region set. Values are scaled to
474 a 0-1 range. **C.** Heatmaps show examples of high-epiCHAOS (EZH2 binding sites) and low-epiCHAOS
475 (FOS binding sites) regions in the peaks matrix of scATAC-seq data from HSCs. **D.** Ordered dot plot
476 showing epiCHAOS scores (epiCHAOS) computed across promoters for each gene ontology biological
477 process in hematopoietic stem cells, ordered by epiCHAOS scores. **E.** Bar plot comparing epiCHAOS ranks
478 for the set of bivalent genes across different hematopoietic cell types. The higher the rank indicates that the
479 selected gene set has higher epiCHAOS scores compared to other gene sets in that cell type. Ranks were -
480 log(10) transformed for display. **F.** Volcano plot illustrates differential heterogeneity between
481 hematopoietic stem cells and monocytes. Differential heterogeneity was tested for each ENCODE TFBS
482 (binding sites measured in K562 cells). For each TFBS, the -log10(p-value) obtained by permutation test is
483 displayed on the y-axis, and the difference in epiCHAOS scores between the two cell types is displayed on
484 the x-axis.

485 **Figure 5. EpiCHAOS is applicable to single-cell epigenomics data from any modality**

486 **A.** Barplots comparing epiCHAOS scores (epiCHAOS) across different lineages of mouse gastrulation
487 using scNMT-seq DNA methylation data from Argelaguet et al. [40] Methylation data summarized across
488 promoters, gene-bodies or CpG islands were used for epiCHAOS computation.. **B.** Scatter plot comparing

489 promoter-wide epiCHAOS scores across different gastrulation lineages using single-cell DNA methylation
490 [epiCHAOS (DNAm)] and ATAC-seq [epiCHAOS (ATAC)] data from the same cells as in A. **C.** UMAP
491 embedding generated from scTAM-seq DNA methylation data from Scherer et al.[42]. Hematopoietic
492 progenitor state and epiCHAOS scores are annotated. GMP: granulocyte monocyte progenitor, EryP:
493 Erythroid progenitor, MEP: myeloid/erythroid progenitor, MPP: multipotent progenitor, HSC:
494 hematopoietic stem cell, MkP: megakaryocyte progenitor. **D.** EpiCHAOS scores calculated using scCHIP-
495 seq data for H3K27me3 from Grosselin et al.[40,43]. Cells were derived from a PDX breast cancer model,
496 separated based on sensitivity or resistance to Capecitabine. Ten subsamples of 100 cells each were taken
497 per condition. Boxplot shows comparison of epiCHAOS scores (epiCHAOS) between sensitive and
498 resistant cells. Wilcoxon p-value < 0.005.

499 # References

500 1. Gough, A. *et al.* Biologically Relevant Heterogeneity: Metrics and Practical Insights. *SLAS Discov* **22**,
501 213–237 (2017).
502 2. Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328–
503 337 (2013).
504 3. Hofmann, W. K., Trumpp, A. & Müller-Tidow, C. Therapy resistance mechanisms in hematological
505 malignancies. *International journal of cancer* **152**, (2023).
506 4. Marusyk, A., Janiszewska, M. & Polyak, K. Intratumor heterogeneity: the Rosetta stone of therapy
507 resistance. *Cancer Cell* **37**, 471 (2020).
508 5. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev.*
509 *Clin. Oncol.* **15**, 81–94 (2018).
510 6. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).
511 7. Nuno, K. *et al.* Convergent epigenetic evolution drives relapse in acute myeloid leukemia. *Elife* **13**,
512 (2024).

513 8. Kumar, R. M. *et al.* Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* **516**,
514 56–61 (2014).
515 9. Pujadas, E. & Feinberg, A. P. Regulated noise in the epigenetic landscape of development and disease.
516 *Cell* **148**, 1123–1131 (2012).
517 10. Chapman, A. *et al.* Heterogeneous tumor subpopulations cooperate to drive invasion. *Cell Rep.* **8**, 688–
518 695 (2014).
519 11. Lambert, A. W., Pattabiraman, D. R. & Weinberg, R. A. Emerging Biological Principles of Metastasis.
520 *Cell* **168**, 670–691 (2017).
521 12. Baslan, T. & Hicks, J. Unravelling biology and shifting paradigms in cancer with single-cell
522 sequencing. *Nat. Rev. Cancer* **17**, 557–569 (2017).
523 13. Kumegawa, K. *et al.* GRHL2 motif is associated with intratumor heterogeneity of cis-regulatory
524 elements in luminal breast cancer. *NPJ Breast Cancer* **8**, 70 (2022).
525 14. Cejas, P. *et al.* Subtype heterogeneity and epigenetic convergence in neuroendocrine prostate cancer.
526 *Nat. Commun.* **12**, 5775 (2021).
527 15. Zhu, Q. *et al.* Single cell multi-omics reveal intra-cell-line heterogeneity across human cancer cell
528 lines. *Nat. Commun.* **14**, 8170 (2023).
529 16. Guilhamon, P. *et al.* Single-cell chromatin accessibility profiling of glioblastoma identifies an invasive
530 cancer stem cell population associated with lower survival. *Elife* **10**, (2021).
531 17. Wu, F. *et al.* Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-
532 small cell lung cancer. *Nat. Commun.* **12**, 2540 (2021).
533 18. Ma, L. *et al.* Tumor Cell Biodiversity Drives Microenvironmental Reprogramming in Liver Cancer.
534 *Cancer Cell* **36**, 418–430.e6 (2019).
535 19. Hinohara, K. *et al.* KDM5 Histone Demethylase Activity Links Cellular Transcriptomic Heterogeneity
536 to Therapeutic Resistance. *Cancer Cell* **35**, 330–332 (2019).
537 20. Teschendorff, A. E. & Enver, T. Single-cell entropy for accurate estimation of differentiation potency
538 from a cell's transcriptome. *Nat. Commun.* **8**, 15599 (2017).

539 21. Landau, D. A. *et al.* Locally disordered methylation forms the basis of intratumor methylome variation
540 in chronic lymphocytic leukemia. *Cancer Cell* **26**, 813–825 (2014).
541 22. Guo, S. *et al.* Identification of methylation haplotype blocks aids in deconvolution of heterogeneous
542 tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat. Genet.* **49**, 635–642
543 (2017).
544 23. Landan, G. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated
545 regions in normal and cancerous tissues. *Nat. Genet.* **44**, 1207–1214 (2012).
546 24. Xie, H. *et al.* Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic*
547 *Acids Res.* **39**, 4099–4108 (2011).
548 25. Scherer, M. *et al.* Quantitative comparison of within-sample heterogeneity scores for DNA
549 methylation data. *Nucleic Acids Res.* **48**, e46 (2020).
550 26. Chung, N. C., Miasojedow, B., Startek, M. & Gambin, A. Jaccard/Tanimoto similarity test and
551 estimation methods for biological presence-absence data. *BMC Bioinformatics* **20**, 644 (2019).
552 27. Granja, J. M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype
553 acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).
554 28. Mohammed, H. *et al.* Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions
555 during Mouse Early Gastrulation. *Cell Rep.* **20**, 1215–1228 (2017).
556 29. Argelaguet, R. *et al.* Decoding gene regulation in the mouse embryo using single-cell multi-omics.
557 *bioRxiv* 2022.06.15.496239 (2022) doi:10.1101/2022.06.15.496239.
558 30. Yao, Z. *et al.* A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain.
559 *Nature* **624**, 317–332 (2023).
560 31. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370**, (2020).
561 32. Gulati, G. S. *et al.* Single-cell transcriptional diversity is a hallmark of developmental potential.
562 *Science* **367**, 405–411 (2020).
563 33. Calderon, D. *et al.* The continuum of embryonic development at single-cell resolution. *Science* **377**,
564 eabn5800 (2022).

565 34. Craig, A. J. *et al.* Genome-wide profiling of transcription factor activity in primary liver cancer using
566 single-cell ATAC sequencing. *Cell Rep.* **42**, 113446 (2023).
567 35. Lüönd, F. *et al.* Distinct contributions of partial and full EMT to breast cancer malignancy. *Dev. Cell*
568 **56**, 3203–3221.e11 (2021).
569 36. Aubin, R. G. *et al.* Pro-inflammatory cytokines mediate the epithelial-to-mesenchymal-like transition
570 of pediatric posterior fossa ependymoma. *Nat. Commun.* **13**, 3936 (2022).
571 37. Gojo, J. *et al.* Single-Cell RNA-Seq Reveals Cellular Hierarchies and Impaired Developmental
572 Trajectories in Pediatric Ependymoma. *Cancer Cell* **38**, 44–59.e9 (2020).
573 38. Gillen, A. E. *et al.* Single-Cell RNA Sequencing of Childhood Ependymoma Reveals Neoplastic Cell
574 Subpopulations That Impact Molecular Classification and Etiology. *Cell Rep.* **32**, 108023 (2020).
575 39. Kar, G. *et al.* Flipping between Polycomb repressed and active transcriptional states introduces noise
576 in gene expression. *Nat. Commun.* **8**, 36 (2017).
577 40. Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**,
578 487–491 (2019).
579 41. Bianchi, A. *et al.* scTAM-seq enables targeted high-confidence analysis of DNA methylation in single
580 cells. *Genome Biol.* **23**, 229 (2022).
581 42. Scherer, M. *et al.* Somatic epimutations enable single-cell lineage tracing in native hematopoiesis
582 across the murine and human lifespan. *bioRxiv* 2024.04.01.587514 (2024)
583 doi:10.1101/2024.04.01.587514.
584 43. Grosselin, K. *et al.* High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states
585 in breast cancer. *Nat. Genet.* **51**, 1060–1066 (2019).
586 44. Faure, A. J., Schmiedel, J. M. & Lehner, B. Systematic Analysis of the Determinants of Gene
587 Expression Noise in Embryonic Stem Cells. *Cell Syst* **5**, 471–484.e4 (2017).
588 45. Feinberg, A. P. & Irizarry, R. A. Evolution in health and medicine Sackler colloquium: Stochastic
589 epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc.*
590 *Natl. Acad. Sci. U. S. A.* **107 Suppl 1**, 1757–1764 (2010).

591 46. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat.*
592 *Genet.* **43**, 768–775 (2011).
593 47. Parreno, V., Martinez, A.-M. & Cavalli, G. Mechanisms of Polycomb group protein function in cancer.
594 *Cell Res.* **32**, 231–253 (2022).
595 48. Wang, S. *et al.* Single-cell multiomics reveals heterogeneous cell states linked to metastatic potential
596 in liver cancer cell lines. *iScience* **25**, 103857 (2022).
597 49. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin
598 accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
599 50. Ramakrishnan, A. *et al.* epiAneufinder identifies copy number alterations from single-cell ATAC-seq
600 data. *Nat. Commun.* **14**, 5846 (2023).
601 51. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory
602 elements in R and Bioconductor. *Bioinformatics* **32**, (2016).
603 52. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell*
604 *systems* **1**, 417 (2015).
605 53. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
606 54. Adelman, E. R. *et al.* Aging Human Hematopoietic Stem Cells Manifest Profound Epigenetic
607 Reprogramming of Enhancers That May Predispose to Leukemia. *Cancer Discov.* **9**, 1080–1101
608 (2019).
609 55. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics.
610 *Nat. Methods* **11**, 637–640 (2014).
611 56. Ben-Porath, I. *et al.* An embryonic stem cell-like gene expression signature in poorly differentiated
612 aggressive human tumors. *Nat. Genet.* **40**, 499–507 (2008).

Figure 1

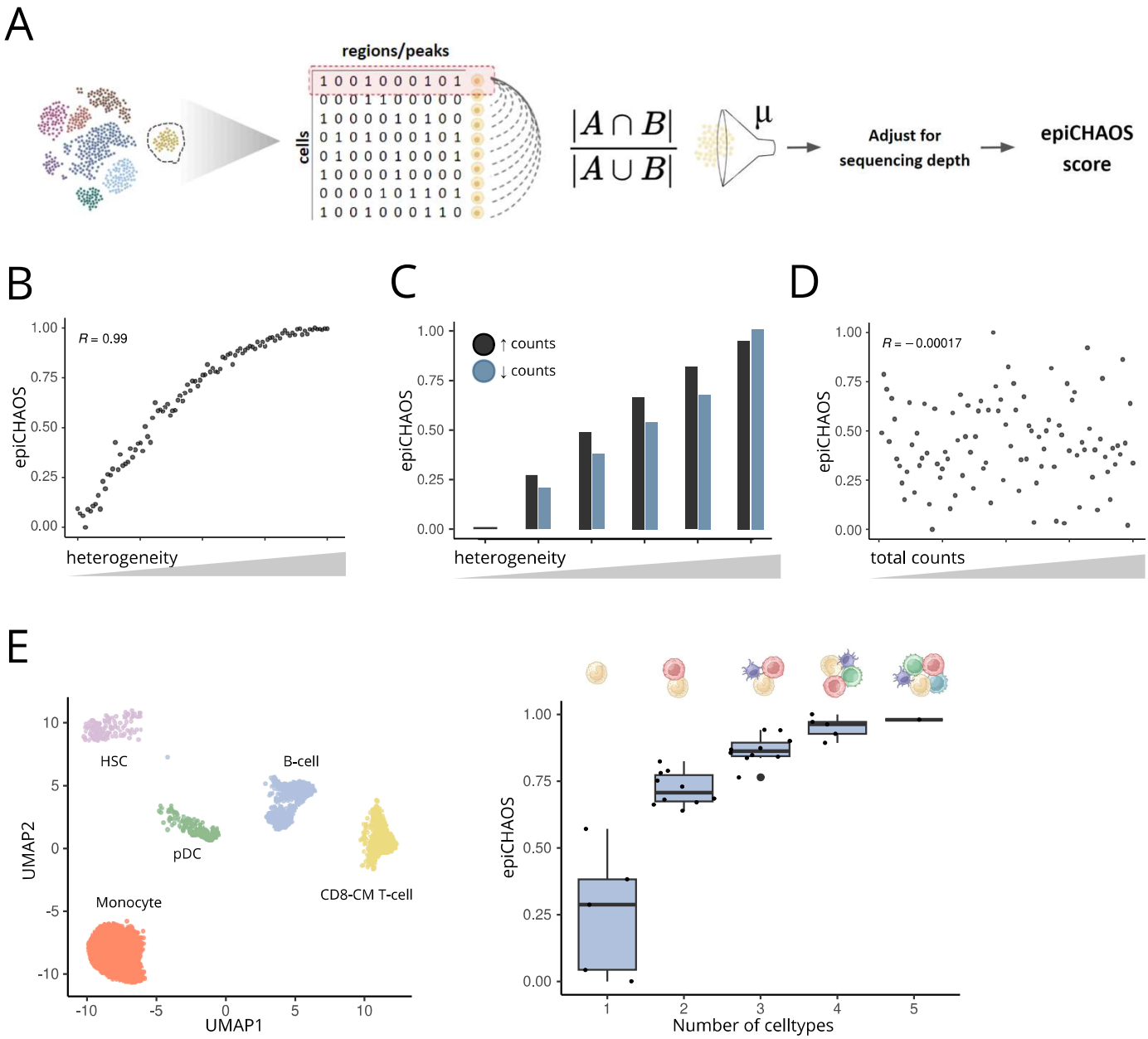

A
regions/peaks
cells
$\frac{|A \cap B|}{|A \cup B|}$
μ
Adjust for sequencing depth
epiCHAOS score
B
R = 0.99
epiCHAOS
heterogeneity
C
↑ counts
↓ counts
epiCHAOS
heterogeneity
D
R = −0.00017
epiCHAOS
total counts
E
HSC
B-cell
pDC
Monocyte
CD8-CM T-cell
UMAP2
UMAP1
epiCHAOS
Number of celltypes

Figure 2

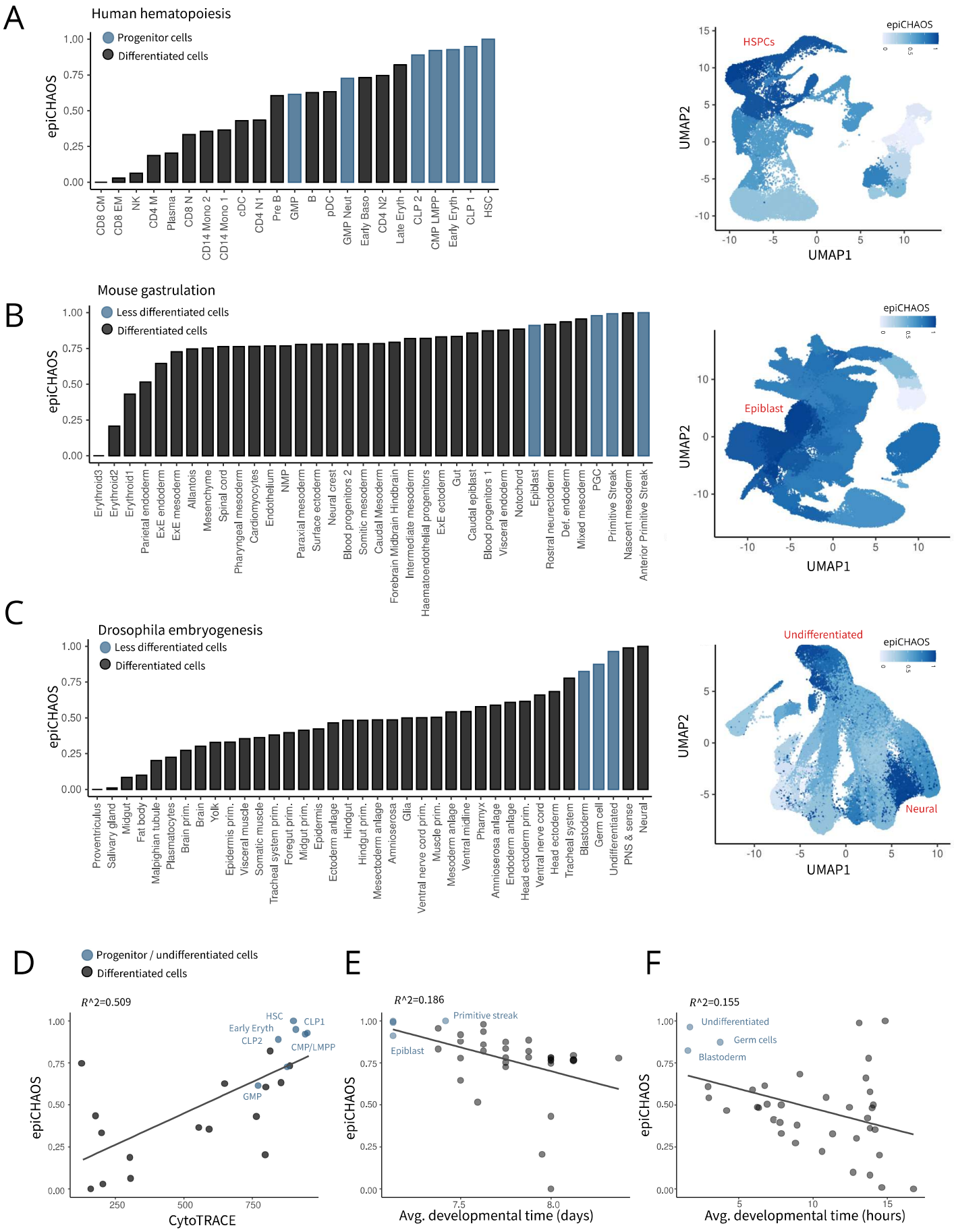

A
Human hematopoiesis
Progenitor cells
Differentiated cells
epiCHAOS
CD8 CM
CD8 EM
NK
CD4 M
Plasma
CD8 N
CD14 Mono 2
CD14 Mono 1
cDC
CD4 N1
Pre B
GMP
B
pDC
GMP Neut
Early Baso
CD4 N2
Late Eryth
CLP 2
GMP LMPP
Early Eryth
CLP 1
HSC
HSPCs
UMAP2
UMAP1
B
Mouse gastrulation
Less differentiated cells
Differentiated cells
Erythroid3
Erythroid2
Erythroid1
Parietal endoderm
ExE endoderm
ExE mesoderm
Allantois
Mesenchyme
Spinal cord
Pharyngeal mesoderm
Cardiomyocytes
Endothelium
NMP
Paraxial mesoderm
Surface ectoderm
Neural crest
Blood progenitors 2
Somitic mesoderm
Caudal Mesoderm
Forebrain Midbrain Hindbrain
Intermediate mesoderm
Haematoendothelial progenitors
ExE ectoderm
Gut
Caudal epiblast
Blood progenitors 1
Visceral endoderm
Notochord
Epiblast
Rostral neurectoderm
Def. endoderm
Mixed mesoderm
PGC
Primitive Streak
Nascent mesoderm
Anterior Primitive Streak
Epiblast
C
Drosophila embryogenesis
Less differentiated cells
Differentiated cells
Proventriculus
Salivary gland
Midgut
Fat body
Malpighian tubule
Plasmatocytes
Brain prim.
Brain
Yolk
Epidermis prim.
Visceral muscle
Somatic muscle
Tracheal system prim.
Foregut prim.
Midgut prim.
Epidermis
Ectoderm anlage
Hindgut
Hindgut prim.
Mesectoderm anlage
Amnioserosa
Glia
Ventral nerve cord prim.
Muscle prim.
Mesoderm anlage
Ventral midline
Pharnyx
Amnioserosa anlage
Endoderm anlage
Head ectoderm prim.
Ventral nerve cord
Head ectoderm
Tracheal system
Blastoderm
Germ cell
Undifferentiated
PNS & sense
Neural
Undifferentiated
Neural
D
Progenitor / undifferentiated cells
Differentiated cells
R^2=0.509
HSC
CLP1
Early Eryth
CLP2
CMP/LMPP
GMP
CytoTRACE
E
R^2=0.186
Primitive streak
Epiblast
Avg. developmental time (days)
F
R^2=0.155
Undifferentiated
Germ cells
Blastoderm
Avg. developmental time (hours)

Figure 3

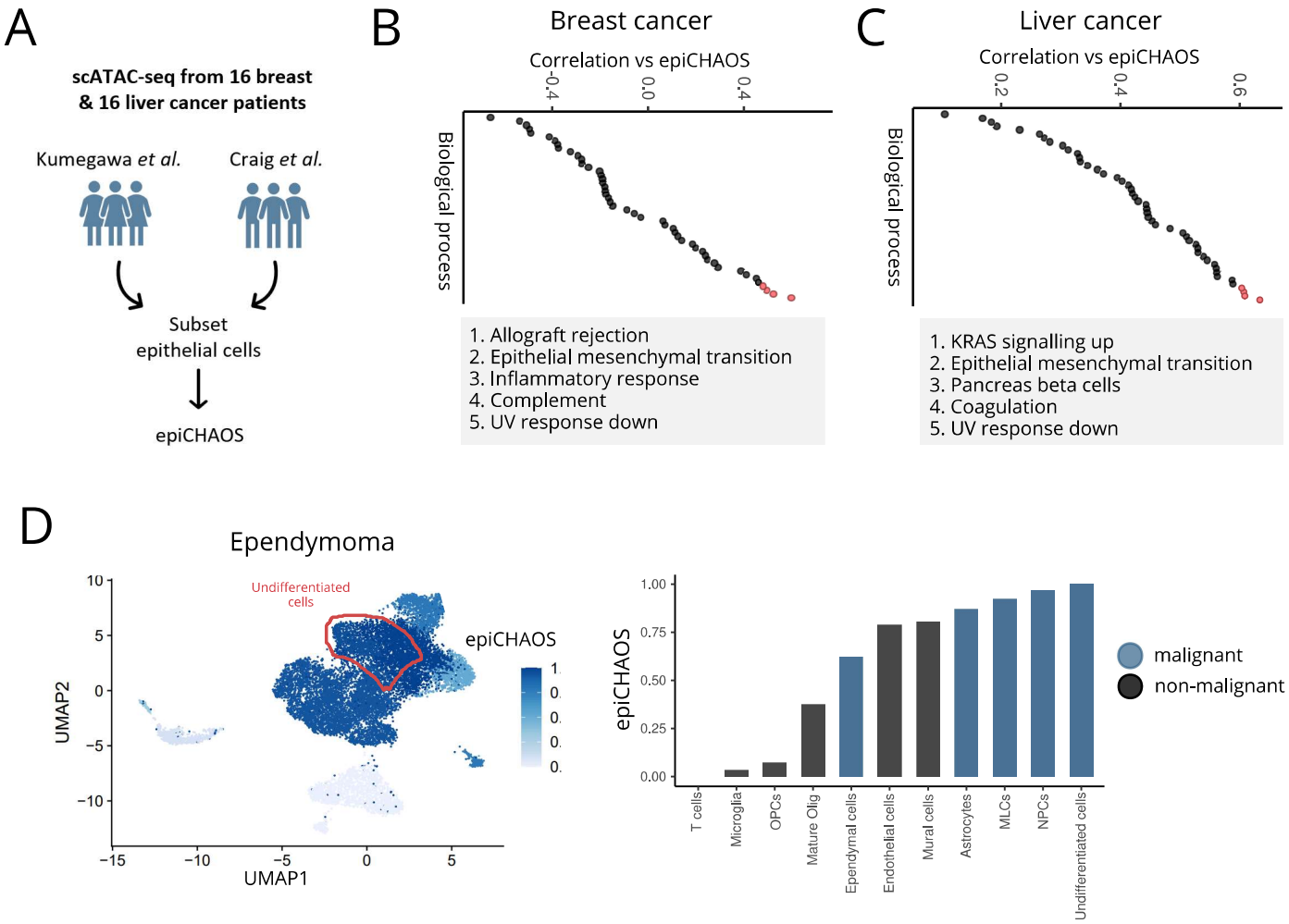

A
scATAC-seq from 16 breast & 16 liver cancer patients
Kumegawa et al.
Craig et al.
Subset epithelial cells
epiCHAOS
B
Breast cancer
Correlation vs epiCHAOS
-0.4
0.0
0.4
Biological process
1. Allograft rejection
2. Epithelial mesenchymal transition
3. Inflammatory response
4. Complement
5. UV response down
C
Liver cancer
Correlation vs epiCHAOS
0.2
0.4
0.6
Biological process
1. KRAS signalling up
2. Epithelial mesenchymal transition
3. Pancreas beta cells
4. Coagulation
5. UV response down
D
Ependymoma
Undifferentiated cells
epiCHAOS
UMAP2
UMAP1
epiCHAOS
1.00
0.75
0.50
0.25
0.00
T cells
Microglia
OPCs
Mature Olig
Ependymal cells
Endothelial cells
Mural cells
Astrocytes
MLCs
NPCs
Undifferentiated cells
malignant
non-malignant

Figure 4

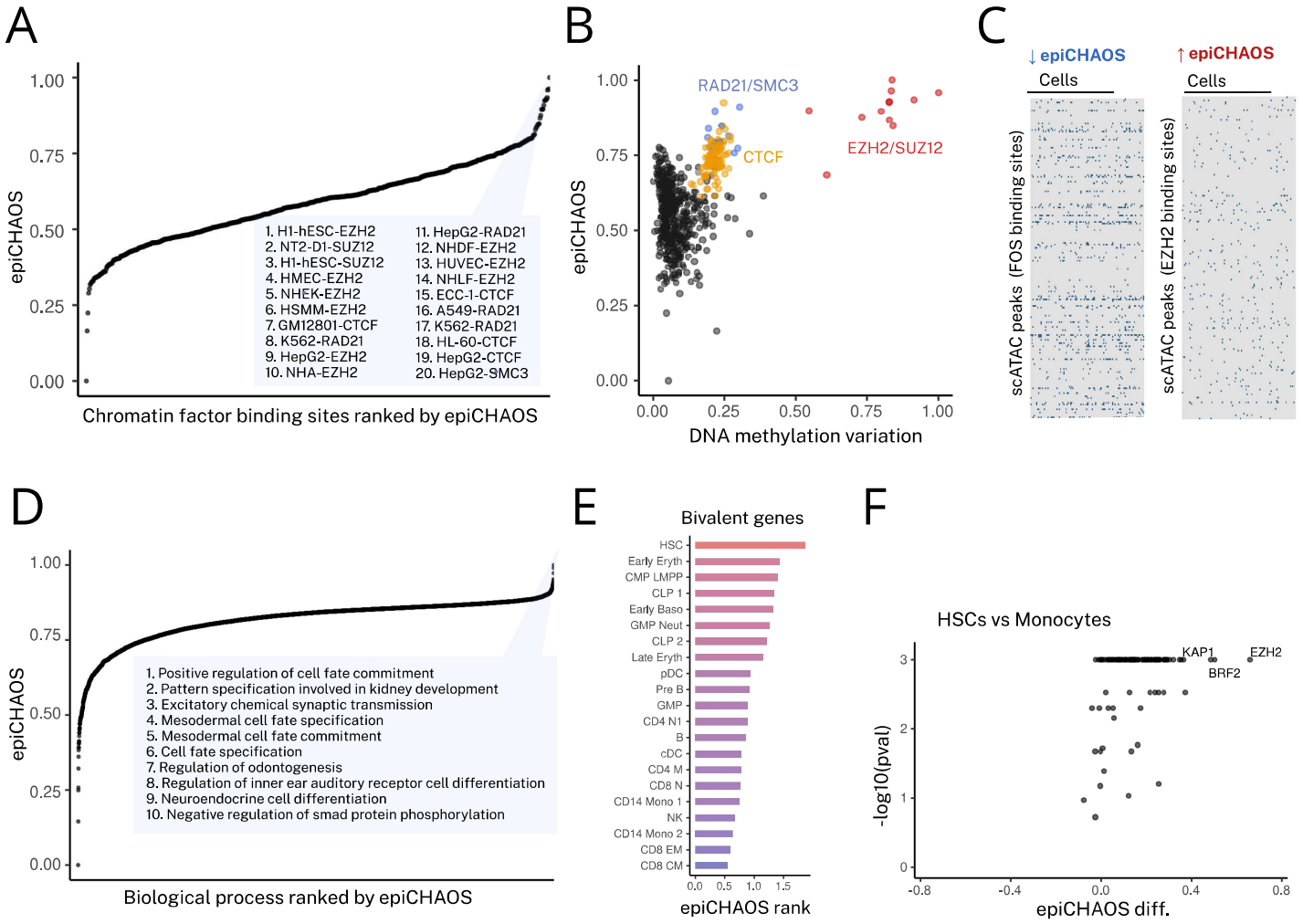
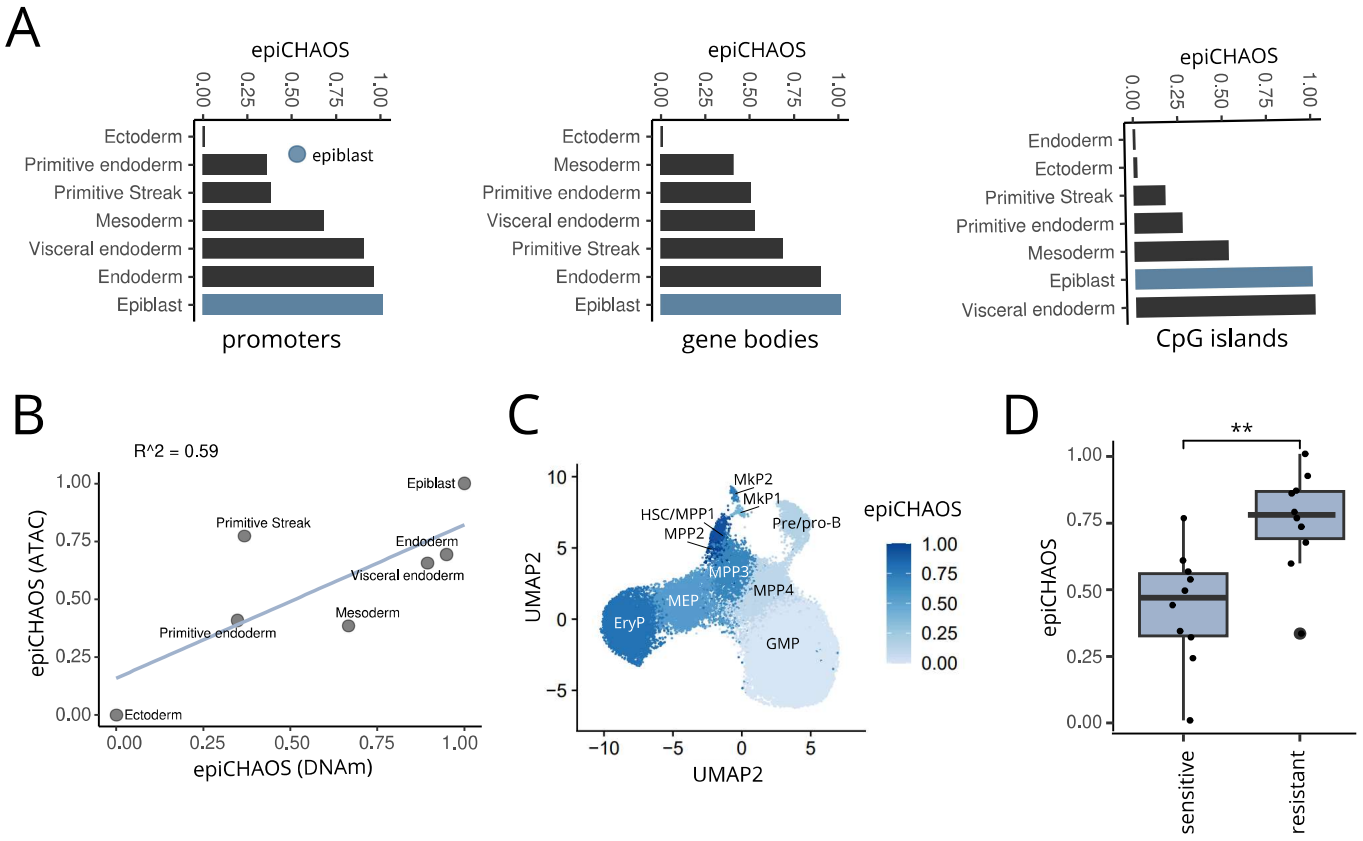

A
epiCHAOS
1.00
0.75
0.50
0.25
0.00
1. H1-hESC-EZH2
2. NT2-D1-SUZ12
3. H1-hESC-SUZ12
4. HMEC-EZH2
5. NHEK-EZH2
6. HSMM-EZH2
7. GM12801-CTCF
8. K562-RAD21
9. HepG2-EZH2
10. NHA-EZH2
11. HepG2-RAD21
12. NHDF-EZH2
13. HUVEC-EZH2
14. NHLF-EZH2
15. ECC-1-CTCF
16. A549-RAD21
17. K562-RAD21
18. HL-60-CTCF
19. HepG2-CTCF
20. HepG2-SMC3
Chromatin factor binding sites ranked by epiCHAOS
B
epiCHAOS
RAD21/SMC3
CTCF
EZH2/SUZ12
0.00
0.25
0.50
0.75
1.00
DNA methylation variation
C
↓ epiCHAOS
↑ epiCHAOS
Cells
Cells
scATAC peaks (FOS binding sites)
scATAC peaks (EZH2 binding sites)
D
epiCHAOS
1.00
0.75
0.50
0.25
0.00
1. Positive regulation of cell fate commitment
2. Pattern specification involved in kidney development
3. Excitatory chemical synaptic transmission
4. Mesodermal cell fate specification
5. Mesodermal cell fate commitment
6. Cell fate specification
7. Regulation of odontogenesis
8. Regulation of inner ear auditory receptor cell differentiation
9. Neuroendocrine cell differentiation
10. Negative regulation of smad protein phosphorylation
Biological process ranked by epiCHAOS
E
Bivalent genes
HSC
Early Eryth
CMP LMPP
CLP 1
Early Baso
GMP Neut
CLP 2
Late Eryth
pDC
Pre B
GMP
CD4 N1
B
cDC
CD4 M
CD8 N
CD14 Mono 1
NK
CD14 Mono 2
CD8 EM
CD8 CM
0.0
0.5
1.0
1.5
epiCHAOS rank
F
HSCs vs Monocytes
KAP1
EZH2
BRF2
-log10(pval)
3
2
1
0
-0.8
-0.4
0.0
0.4
0.8
epiCHAOS diff.

Figure 5


A
epiCHAOS
0.00
0.25
0.50
0.75
1.00
Ectoderm
Primitive endoderm
Primitive Streak
Mesoderm
Visceral endoderm
Endoderm
Epiblast
epiblast
promoters
gene bodies
CpG islands
B
R^2 = 0.59
epiCHAOS (ATAC)
epiCHAOS (DNAm)
C
UMAP2
MkP2
MkP1
HSC/MPP1
MPP2
Pre/pro-B
MPP3
MPP4
MEP
EryP
GMP
D
**
sensitive
resistant