Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity

Longqi Liu[1,2,3,13], Chuanyu Liu[1,2,4,13], Liang Wu[1,2,4,13], Andrés Quintero[5,6,13], Yue Yuan[1,2,4,13], Mingyue Wang[1,2,4], Mengnan Cheng[1,2,4], Liqin Xu[1,2], Guoyi Dong[1,2], Rui Li[1,2,3], Yang Liu[1,2,4], Xiaoyu Wei[1,2,4], Jiangshan Xu[1,2,4], Xiaowei Chen[2], Haorong Lu[2], Dongsheng Chen[1,2], Quanlei Wang[1,2,4], Qing Zhou[1,2], Xinxin Lin[1,2], Guibo Li[1,2], Shiping Liu[1,2], Qi Wang[5,7], Hongru Wang[8], J. Lynn Fink[1], Zhengliang Gao[9], Xin Liu[1,2], Yong Hou[1,2], Shida Zhu[1,2], Huanming Yang[1,10], Yunming Ye[3], Fang Chen[1,2,11], Carl Herrmann[5,6,7], Roland Eils[6,12,*], Zhouchun Shang[1,2,9,*] & Xun Xu[1,2,*]

[1]BGI-Shenzhen, Shenzhen, China. [2]China National GeneBank, BGI-Shenzhen, Shenzhen, China. [3]Harbin Institute of Technology Shenzhen Graduate School, Xili University Town, Shenzhen, China. [4]BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China. [5]Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [6]Health Data Science Unit, Heidelberg University Hospital, Heidelberg, Germany. [7]Institute of Pharmacy and Molecular Biotechnology and Bioquant Center, University of Heidelberg, Heidelberg, Germany. [8]Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing, China. [9]Department of Regenerative Medicine, Tongji University School of Medicine, Shanghai, China. [10]James D. Watson Institute of Genome Sciences, Hangzhou, China. [11]Section of Molecular Disease Biology, Department of Veterinary Disease Biology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [12]Center for Digital Health, Berlin Institute of Health and Charité, Berlin, Germany. [13]These authors contributed equally to this work.

*Correspondence should be addressed to R.E. (roland.eils@bihealth.de), Z.S. (shangzhouchun@genomics.cn) or X.X. (xuxun@genomics.cn)

Abstract

Integrative analysis of multi-omics layers at single cell level is critical for accurate dissection of cell-to-cell variation within certain cell populations. Here we report scCAT-seq, a technique for simultaneously assaying chromatin accessibility and the transcriptome within the same single cell. By applying our integrated approach to multiple cancer cell lines, we discovered genomic loci with coordinated epigenomic and transcriptomic variability. In addition, decomposition of combined single-cell chromatin accessibility and gene expression features by a non-negative matrix factorization (NMF) based method identified signatures reflecting cell type specificity and revealed a profound regulatory relationship between the two layers of omics. We further characterized subpopulations associated with distinct regulatory patterns within patient-derived xenograft models and discovered epigenomic and transcriptomic clues that drive tumor heterogeneity. The ability to obtain

these two layers of omics data will help provide more accurate definitions of "single cell states" and enable the deconvolution of regulatory heterogeneity from complex cell populations.

Main

The rapid proliferation of single cell sequencing technologies has greatly improved our understanding of heterogeneity in terms of genetic, epigenetic and transcriptional regulation within cell populations[1]. We, and others, have developed single-cell whole genome[2], exome[3, 4], methylome[5] and transcriptome[6, 7] technologies and applied these approaches to analyzing the complexity of cell populations in tumorigenesis, developmental process and cellular reprogramming[8]. Meanwhile, single-cell epigenome techniques, including single cell ChIP-seq[9], ATAC-seq[10, 11], DNase-seq[12] and Hi-C[13, 14], have been developed to decipher histone modifications, transcription factor (TF) accessibility landscapes, and 3D chromatin contacts, respectively, in single cells. These techniques provide important information on regulatory heterogeneity by assessing chromatin structure across various cell types.

Recent progress on measuring multi-omics in the same cells has enabled analysis of associations between different layers of regulation on gene expression. Current approaches allow parallel sequencing of the genome and transcriptome, or the methylome and transcriptome, in the same single cell[15-19]. However, it remains challenging to study associations between chromatin accessibility (CA) and gene expression (GE), and a comprehensive investigation of regulatory heterogeneity using multi-layer omics at the single-cell level is still lacking. Here, we present scCAT-seq (**s**ingle-**c**ell **c**hromatin **a**ccessibility and **t**ranscriptome **seq**uencing), a technique that integrates single-cell assays for transposase-accessible chromatin using sequencing (scATAC-seq) and single-cell RNA-seq (scRNA-seq) to measure chromatin accessibility and gene expression simultaneously in single cells. scCAT-seq employs a mild lysis approach and a physical dissociation strategy to separate the nucleus and cytoplasm of each single cell. Thereafter, the supernatant cytoplasm component is subjected to the Smart-seq2 method as described previously[7]. The precipitated nucleus is then subjected to a Tn5 transposase-based and carrier DNA-mediated protocol to capture fragments within accessible regions (**Fig. 1a** and **Supplementary Methods**). Beyond parallel transcriptome and chromatin accessibility in the same single cell, scCAT-seq will be particularly useful for analyzing samples when the amount of input material is limited.

We applied scCAT-seq to the K562 chronic myelogenous leukemia cell line, which has been widely used in the ENCODE project. We sorted single cell and multi-cell samples (e.g., 500 cells) using flow cytometry into wells of 96-well plates. Empty wells were used as negative controls. Samples were then processed using the scCAT-seq protocol. qPCR analysis confirmed the successful capture of single cell nuclei during library preparation (**Supplementary Fig. 1a**). We generated combined CA and GE profiles from a total of 192 samples. Of the 176 single cell profiles, 74 (42.0%) of them passed both CA and GE data quality control criteria (**Supplementary Fig. 1b** and **Supplementary Methods**).

For scCAT-seq-generated chromatin accessibility data, we obtained an average of 1.74 x 10^5 uniquely mapped, usable fragments from single cells (**Supplementary Table 1**). Similar to bulk ATAC-seq[20], the chromatin accessibility fragments show fragment-size periodicity corresponding to integer multiples of nucleosomes (**Supplementary Fig. 1c**) and are strongly enriched on accessible regions (over an average of 30% of fragments) (**Fig. 1b** and **Supplementary Table 1**). We found that about 9% of the fragments were mapped to the mitochondrial genome (**Supplementary Fig. 1d**) which is largely reduced in comparison to standard bulk ATAC-seq studies (typically over 30%)[20]. The aggregate single-cell chromatin accessibility profile closely reproduced the profile generated from 500 cells (Pearson correlation value > 0.9, **Supplementary Fig. 1e**). In comparison to the published scATAC-seq profiles by Buenrostro *et al.*[10], we obtained a higher numbers of usable fragments per single cell (**Supplementary Fig. 1f**) which is likely due to more complete digestion of accessible regions. Although the scCAT-seq data showed lower signal-to-noise ratio (**Supplementary Fig. 1g**), the correlation between single cells increases remarkably (**Supplementary Fig. 1h**), suggesting that scCAT-seq is able to capture the chromatin features more accurately.

For mRNA data generated by scCAT-seq, we obtained an average of 4.6 million reads covering over 8000 genes (GENCODE v19, TPM > 1), which is comparable to published scRNA-seq profiles by Pollen *et al.*[21] (**Supplementary Fig. 1j** and **Supplementary Table 2**). Consistent with published Smart-seq profiles, our mRNA data showed full coverage of the transcript body **(Fig. 1b)**. The aggregate profile was close to the RNA-seq profile obtained from 500 cells (Pearson correlation value > 0.9, **Supplementary Fig. 1i**), suggesting that scCAT-seq is able to accurately quantify gene expression of single cells. The density of chromatin accessibility and mRNA reads of all single cells surrounding a

constitutively accessible region showed that scCAT-seq data recapitulates major features obtained by separately performed bulk ATAC-seq and RNA-seq (**Fig. 1c**).

Gene expression regulation is associated with the structure of the *cis*-regulatory elements (e.g., histone modifications, DNA methylation) and the binding of *trans*-factors (e.g., TFs, epigenetic modifiers)[22]. Therefore, we examined the overall distribution of single-cell CA fragments across different genomic contexts, as well as the expression levels of the putative regulated genes. We observed that the CA fragments were enriched at *cis*-regulatory elements with active histone modifications (e.g., H3K27ac, H3K9ac and H3K4me3), whereas repressive or inaccessible regions (e.g., H3K27me3 and H3K36me3-associated regions) showed lower fragment density (**Fig. 1d**). We also observed inconsistencies between CA and GE. For example, we found low levels of CA fragments on H3K36me3-associated regions but high levels of gene expression. This is not surprising because H3K36me3 is known to be enriched on the active gene body which is occupied by nucleosomes and is inaccessible[22]. Notably, genes with bivalent marks (co-enrichment of H3K4me3 or H3K4me1 and H3K27me3) showed similar level of accessibility as active genes (co-enrichment of H3K4me3 or H3K4me1 and H3K27ac, but lack of H3K27me3), and both of them showed higher levels of accessibility than inactive genes (enrichment of H3K27me3, but not H3K27ac, H3K4me1 and H3K4me3). Conversely, the expression levels of bivalent genes were remarkably lower than active genes and were similar with that of inactive genes. We also investigated the distribution of CA fragments across genomic contexts bound by different TFs and found an overall consistent pattern between chromatin accessibility and gene expression level. Notably, we observed substantial decrease of expression levels of genes associated with binding of EZH2 while the accessibility level showed just a moderate change (**Fig. 1e**). This pattern is similar with that of bivalent genes and is in agreement with the role of EZH2 which, as part of the repressive polycomb complex, catalyzes H3K27me3. Thus, the combined signatures from scCAT-seq well reflect and are useful to assess the transcriptional state of genes within different genomic contexts. This approach will be of high value for many biological applications, for example, studying the heterogeneous transition of bivalent genes during development or cellular reprogramming.

We further validated our approach by generating different batches of scCAT-seq profiles from two additional ENCODE cell lines: HeLa-S3 cervix adenocarcinoma and HCT116 colorectal carcinoma cell lines. Clustering of Pearson correlation coefficients for both CA

and GE profiles revealed a strong correlation between single cells from the same cell type (**Supplementary Fig. 2a,b)**, suggesting that scCAT-seq is able to discriminate differences between cell types. A comparison of our datasets with published profiles revealed that the differences across protocols and batches had a substantially smaller effect than difference across cell types (**Supplementary Fig. 2c,d)**.

Next, we explored the dynamic associations between the two omics layers across single cells (**Fig. 2a**). We investigated highly variable regulatory elements and genes, and tested the correlation between accessibility level of single regulatory elements and their putative gene expression. We identified remarkably more positive correlations than negative correlations in all three cell lines (**Fig. 2b**), which is consistent with the known relationship between chromatin accessibility and gene expression in bulk profiles[23]. Interestingly, we found that the regulatory regions that are positively correlated with gene expression are mainly enriched at distal regions (**Fig. 2c**), demonstrating that distal regions (such as enhancers) may be more sensitive in response to the environmental stimulus than proximal regions such as promoters. In addition, the regulatory regions with positive correlation showed significant enrichment of KEGG terms associated with tumorigenesis, such as the WNT signaling pathway[24] in the K562 cell line and the HIF signaling pathway[25] in HCT116 or HeLa-S3 cell lines (**Fig. 2d**). This provides the evidence that the heterogeneous accessibility of regulatory elements accompanies gene expression variation of key cancer associated genes even within presumably homogenous cancer cell lines. This observation is similar to previous reports describing regulatory heterogeneity within *in vitro* cultured pluripotent stem cells[26, 27].

To demonstrate the ability of scCAT-seq to establish more precise correspondences between chromatin accessibility and gene expression in cell populations, we next focused our analysis on a hypothetical cell population merged from the single cells of the three cell types. Here we attempted to derive cellular identity based on regulatory variability. In order to comprehensively extract key features from multi-assay datasets of scCAT-seq, we applied a non-negative matrix factorization (NMF) method, implemented in an R package called Bratwurst[28], to decompose the CA and GE datasets into genomic signatures and extract the corresponding features (regions and genes) (**Fig. 3a** and **Supplementary Methods)**. When applying Bratwurst to CA profiles we obtained a hierarchical decomposition of the cell populations, as shown by Sankey diagram or riverplot (**Fig. 3b**). Using various optimization criteria, the optimal factorization rank was found to be $K_{CA} = 3$

(**Supplementary Fig. 3a** and **Supplementary Methods**), resulting in three signatures (CA1, CA2, CA3) corresponding to the three cell types analyzed in this study (**Fig. 3b**). Applying Bratwurst to GE profiles yielded a similar decomposition (**Fig. 3c**). However, four signatures (GE1, GE2, GE3, GE4, **Fig. 3c** and **Supplementary Fig. 3b**) were identified with the optimal factorization rank $K_{GE} = 4$, suggesting a more complex structure at transcriptome level. Of all seven CA and GE signatures, signature CA1 and GE1 were highly specific to HeLa-S3 cells, while CA3 and GE4 were specific to K562 cells. Interestingly, while HCT116 cells show higher exposure score to CA2, we found different subpopulations within HCT116 that show highly variable exposure scores to GE2 and GE3 (**Fig. 3c**), suggesting transcriptional heterogeneity within the HCT116 cell line. Collectively, our analyses demonstrate that decomposition of chromatin accessibility and gene expression features achieved similar results, but also reflected distinct patterns of regulatory heterogeneity in particular contexts.

To precisely unravel the association between CA and GE, we next applied Bratwurst on the combined CA and GE datasets (CAGE), which produced signatures containing both CA and GE features. In agreement with the NMF analysis using CA or GE separately, with factorization rank $K_{CAGE} = 3$ we found three signatures corresponding to the three cell types (**Supplementary Fig. 4a**), while with $K_{CAGE} = 4$, we again observe additional subpopulations within the HCT116 cell line (**Fig. 3d**). We identified the optimal rank $K_{CAGE} = 3$ (**Supplementary Fig. 3c**), which yielded three signatures, each one with features (genes and accessible regions) that are either specific to one single signature or shared by several of them (**Supplementary Fig. 4b**). To systematically study the relationship between signatures, we calculated the Spearman correlation between exposure scores of single cells to different signatures. Interestingly, we found various correspondences between CA and GE signatures (**Fig. 3e**). For example, signature CA1 was correlated with GE1 (Spearman correlation = 0.41). Signature CA2 was correlated with GE2 and GE3 (Spearman correlations = [0.64, 0.74], respectively), and signature CA3 was correlated with GE4 (Spearman correlation = 0.75). In addition, our analysis showed that the integrated signatures are correlated with both CA and GE signatures (**Fig. 3e**). For example, CAGE1 showed high correlation with CA3 and GE4 (Spearman correlations = [0.81, 0.82], respectively), signature CAGE2 showed high correlation with CA2 and GE2 and GE3 (Spearman correlations = [0.81, 0.92, 0.59], respectively), and signature CAGE3 was highly correlated with CA1 and GE1 (Spearman correlations = [0.60, 0.84]). This indicates that the

integrated signatures are contributed by individual CA and GE signatures with different weights.

We next investigated whether there are regulatory links between accessible regions and genes within the integrated signatures. For each signature, we obtained the signature specific features (genes and accessible regions) and calculated the distances from genes to accessible regions. Interestingly, we found that in all three signatures, the distance of genes to regulatory regions was substantially smaller than randomly selected region and gene sets, revealing a strong regulatory links between these two layers of omics (**Fig. 3f**). This indicates that the integrated signatures are not independent, but are rather interconnected sets of CA or GE features. In addition, functional enrichment analysis showed that the peaks and genes from each signature are significantly enriched at specific regions of the respective cell types (**Supplementary Table 3**). Overall, our analyses clearly demonstrate the potential of scCAT-seq in the accurate definition of cell entities (e.g., subpopulations within complex tissues) through deconvolution of integrated chromatin accessibility and gene expression profiles.

We then assessed whether scCAT-seq could be applied to dissecting intra-tumoral regulatory heterogeneity, which has been shown to correlate with cancer prognoses[29]. We applied scCAT-seq to two lung cancer patient-derived xenograft (PDX) models (**Fig. 4a)**. One is derived from a moderately differentiated squamous cell carcinoma patient (PDX1) and the other one from a large-cell lung carcinoma patient (PDX2), as validated by hematoxylin & eosin staining (**Supplementary Fig. 5a,b** and **Supplementary Methods**). We generated quality-filtered scCAT-seq profiles from FACS-isolated single cells from PDX1 (176 profiles) and PDX2 (167 profiles) tissues, respectively (**Supplementary Fig. 6a,b**). Analyzing the overall correlation between chromatin accessibility and gene expression, we identified remarkably more positive correlations than negative ones (1079 vs. 14 in PDX1, and 2645 vs. 23 in PDX2, **Supplementary Fig. 6c,d**). Compared to K562 cells (see above), this imbalance towards more positive correlations is even more extreme, particularly in PDX2. To assess whether subpopulations represent distinct regulatory characteristics within the PDX models, we applied Bratwurst to CA and GE features. As expected, we found obvious discrimination between the two PDX models when using combined or separate CA and GE profiles (K_{CA} = 2 and K_{GE} = 2, **Fig. 4b** and **Supplementary Fig. 7a-c**). Interestingly, with the increase of factorization rank PDX1-specific signature remained quite stable, whereas the PDX2-specific signature changed dramatically (**Fig. 4b**

and **Supplementary Fig. 7a-c**), revealing prevalent regulatory heterogeneity in particular within PDX2. Applying NMF algorithm to the PDX2 model only we found similar patterns between NMF decompositions when using separate CA and GE profiles (**Supplementary Fig. 7d,e**), suggesting that both epigenome and transcriptome alterations contribute to intra-tumoral variability. We next investigated the link between the two layers of omics by applying NMF on the integrated profiles. We observed a hierarchical structure of heterogeneity with increase of factorization rank (**Fig. 4c** and **Supplementary Fig. 7f**). We obtained three signatures containing both CA and GE features with the optimal rank $K_{CAGE} = 3$ (**Fig. 4c** and **Supplementary Fig. 7g**). We observed that signature CAGE2 divide PDX2 single cells into two subpopulations, while one of the subpopulations could be further distinguished by signatures CAGE1 and CAGE3. Thus, the decomposition yielded three subpopulations within the PDX2 tissue (labeled with subpopulations I, II and III, as shown in **Fig. 4c**). Notably, we found that the exposure score of signature CAGE1 and CAGE3 showed high correlation with signatures GE1 and GE3, respectively, but are poorly correlated with CA signatures, whereas signature CAGE2 showed high correlation with both CA2 and GE2 signatures (Spearman correlations = [0.52, 0.97], respectively, **Fig. 4d**). These findings interestingly suggest that changes of gene expression in signatures CAGE1 and CAGE3 are rather independent of chromatin accessibility, while expression in signature CAGE2, which is specific to subpopulation III, may be driven by variations on chromatin accessibility. In particular, signature CAGE3 shows a very limited number of specific peaks, suggesting that the specific genes of this signature could be driven by promoters which are constitutively open across all cells rather than cell specific regulatory regions. Supporting the idea of signature CAGE2, we found that distances between genes and peaks are generally closer within specific features of signature CAGE2 than the other two signatures (**Fig. 4e**). Interestingly, functional enrichment analysis of genes in signature CAGE2 showed striking enrichment in GO terms related to “cell-cell adhesion” and “GTPase binding", which have been reported to be associated with cancer migration[30] (**Fig. 4f**). This suggests that subpopulations within the PDX tissue may have specific epigenomic and transcriptomic characteristics associated with different cell migratory abilities.

Since *cis*-regulatory elements are comprised of clusters of TF motifs. We assessed whether the regulatory heterogeneity within the PDX2 tissue is driven by TF binding variation by adopting a previously reported analytical method, ChromVAR[31]. This method computes chromatin accessibility deviation scores (z-scores) from a set of peaks that contain a specific TF motif in all single cells. We computed the deviation scores of a total of 386 TF motifs on

peak regions from signature CAGE2 for all PDX2 single cells. We then ranked all the TFs with variable deviation scores. We found that most of the highly variable TFs have been reported to be regulators of tumorigenesis, such as JUNB, NKX2-8, and FOXA1 (**Fig. 4g**). To further investigate possible regulators that are associated with PDX2 heterogeneity, we computed the correlation between TF deviation scores and target gene expression over all single cells in PDX2. Interestingly, we identified many TFs showing positive correlation with gene expression **(Fig. 4h** and **Supplementary Fig. 8)**. The list contains TFs NKX2-8 and FOXA1, which have both been reported to be key regulators of lung cancer[32, 33], suggesting that these TFs might be involved in mediating chromatin accessibility variations to transcriptional heterogeneity in PDX2 tissues. Overall, our results demonstrate that through integrative analysis of the single cell chromatin accessibility and transcriptome data generated with scCAT-seq, we successfully linked epigenomic variability to transcriptional heterogeneity in the context of subpopulation characteristics of cancer models, which would have been difficult to reveal in independent runs of transcriptional and chromatin accessibility sequencing.

Our work shows that scCAT-seq is able to provide high resolution epigenomic and transcriptomic portraits of individual cells. We have shown a positive correlation between chromatin accessibility and gene expression across single cells and identified the key TFs that may be involved in driving the regulatory variability within cell populations. In addition, scCAT-seq enables decomposition of regulatory complexity by extracting and correlating the features from two layers of omics, thus providing the basis for a far more comprehensive study of regulatory mechanisms at the single cell level. Overall, we here demonstrated that scCAT-seq is a highly promising tool for the joint study of multimodal data of single cells, which may pave the way toward a thorough assessment of regulatory heterogeneity in a variety of clinical applications including preimplantation screening and cancer management.

ACKNOWLEDGEMENTS

We thank all members of the Stem Cell and Development Lab (BGI) for their support and Scott Edmunds, Christian Conrad, Kun Ma and Ying Shan for helpful discussion. We also thank Jijun Cheng, Yuan Long and Feifei Zhang from Shanghai LIDE Biotech Co., Ltd. for technical support. This work was supported by the Shenzhen Municipal Government of China Peacock Plan (KQTD20150330171505310), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA16010100), the Shenzhen Municipal Government of China (GJHS20170314152802146), and the Shenzhen Engineering Laboratory for

Innovative Molecular Diagnostics (DRC-SZ (2016) 884). Longqi Liu is funded by the China Postdoctoral Science Foundation (2017M610553).

AUTHOR CONTRIBUTIONS

L.L., C.L., Z.S. and X.X. conceived the idea. C.L. and Y.Yuan designed the scCAT-seq method. C.L. and Y.Yuan performed the majority of the experiments and generated the scCAT-seq data. M.W., M.C., L.X., G.D., Quanlei Wang, J.X., X.C., H.L., G.L., Q.Z., Z.G. and X.Lin assisted with the experiments. L.L. and L.W. designed the data processing pipelines and analyzed the data. A.Q. and C.H. designed and performed the NMF analysis. X.W., D.C., Y.L., R.L., Qi Wang, S.L., Y.H., S.Z., F.C., Y.Ye, X.Liu and H.W. assisted with the data analysis. L.L., C.L., L.W., A.Q. and Y.Yuan prepared the figures. L.L. wrote the manuscript with input from C.L., L.W., A.Q. and Y.Yuan. Z.S., C.H., L.F., R.E. and X.X. revised the manuscript. H.Y., L.L., Z.S., R.E. and X.X. supervised the entire study. All authors read and approved the manuscript for submission.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618-630 (2013).
2. Zong, C., Lu, S., Chapman, A.R. & Xie, X.S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622-1626 (2012).
3. Xu, X. et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886-895 (2012).
4. Hou, Y. et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873-885 (2012).
5. Guo, H. et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23**, 2126-2135 (2013).
6. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377-382 (2009).
7. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096-1098 (2013).
8. Wen, L. & Tang, F. Reconstructing complex tissues from single-cell analyses. *Cell* **157**, 771-773 (2014).
9. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165-1172 (2015).
10. Buenrostro, J.D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490 (2015).
11. Cusanovich, D.A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914 (2015).
12. Jin, W. et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**, 142-146 (2015).

13. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013).
14. Ramani, V. et al. Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263-266 (2017).
15. Dey, S.S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* **33**, 285-289 (2015).
16. Angermueller, C. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229-232 (2016).
17. Cheow, L.F. et al. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nat. Methods* **13**, 833-836 (2016).
18. Hou, Y. et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **26**, 304-319 (2016).
19. Hu, Y. et al. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome. Biol.* **17**, 88 (2016).
20. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213-1218 (2013).
21. Pollen, A.A. et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**, 1053-1058 (2014).
22. Li, B., Carey, M. & Workman, J.L. The role of chromatin during transcription. *Cell* **128**, 707-719 (2007).
23. Boyle, A.P. et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-322 (2008).
24. Polakis, P. Wnt signaling in cancer. *Cold Spring Harb. Perspect. Biol.* **4** (2012).
25. Esteban, M.A. & Maxwell, P.H. HIF, a missing link between metabolism and cancer. *Nature Medicine* **11**, 1047 (2005).
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 Reports* **2**, 881-895 (2014).
27. Nguyen, Q. et al. Single-Cell Transcriptome Sequencing Of 18,787 Human Induced Pluripotent Stem Cells Identifies Differentially Primed Subpopulations. *bioRxiv* (2017).
28. Huebschmann, D. et al. Deciphering programs of transcriptional regulation by combined deconvolution of multiple omics layers. *bioRxiv* (2017).
29. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**, 613-628 (2017).
30. Ridley, A.J. Rho GTPase signalling in cell migration. *Curr Opin Cell Biol* **36**, 103-112 (2015).
31. Schep, A.N., Wu, B., Buenrostro, J.D. & Greenleaf, W.J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975-978 (2017).
32. Hsu, D.S. et al. Characterizing the developmental pathways TTF-1, NKX2-8, and PAX9 in lung cancer. *Proc. Natl. Acad. Sci. U S A* **106**, 5312-5317 (2009).
33. Song, Y., Washington, M.K. & Crawford, H.C. Loss of FOXA1/2 is essential for the epithelial-to-mesenchymal transition in pancreatic cancer. *Cancer Res.* **70**, 2115-2125 (2010).

Figure 1

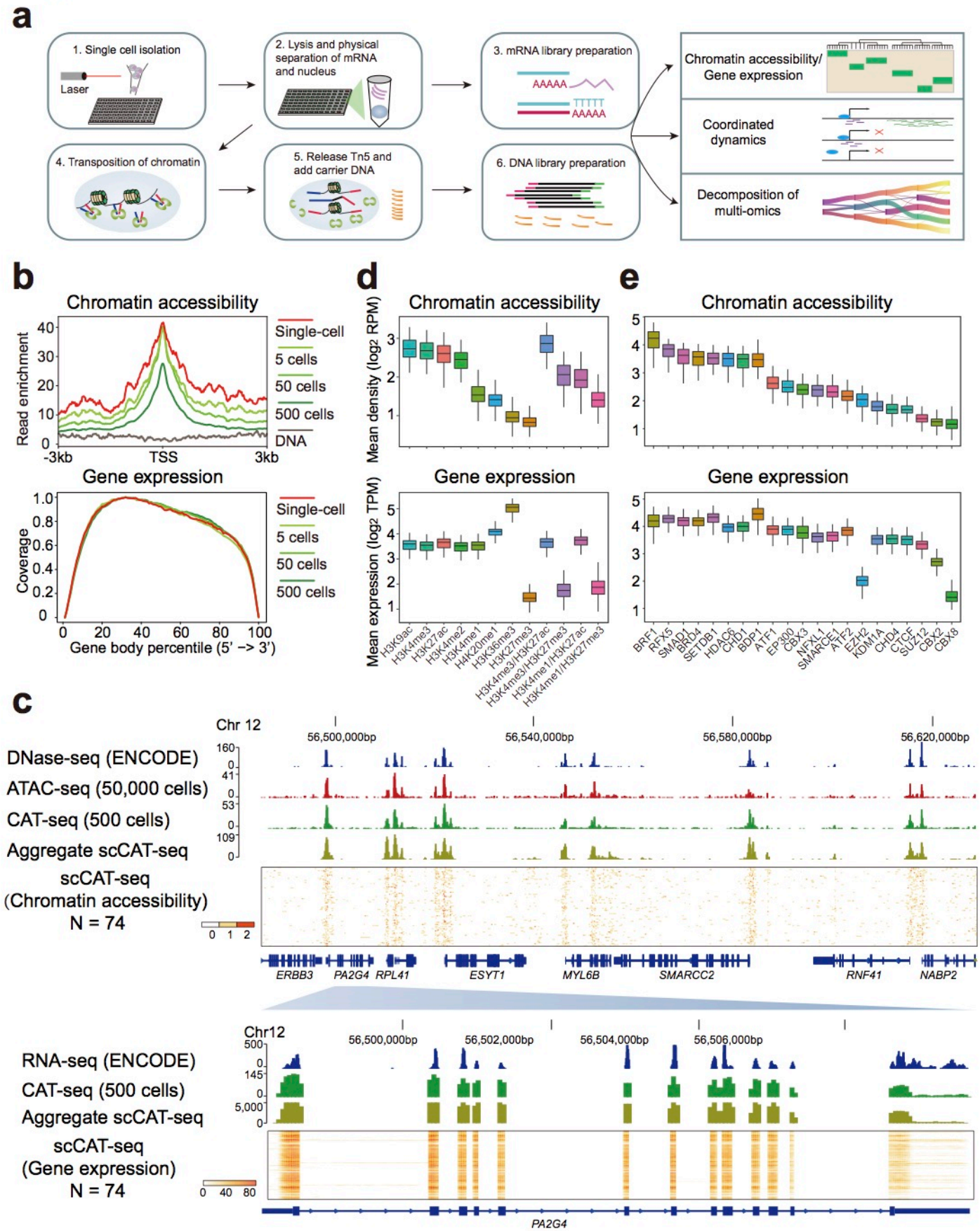

a
1. Single cell isolation
Laser
2. Lysis and physical separation of mRNA and nucleus
3. mRNA library preparation
AAAAA
AAAAA
4. Transposition of chromatin
5. Release Tn5 and add carrier DNA
6. DNA library preparation
Chromatin accessibility/ Gene expression
Coordinated dynamics
Decomposition of multi-omics
b
Chromatin accessibility
Read enrichment
Single-cell
5 cells
50 cells
500 cells
DNA
-3kb
TSS
3kb
Gene expression
Coverage
Gene body percentile (5' -> 3')
d
Chromatin accessibility
Mean density (log2 RPM)
Gene expression
Mean expression (log2 TPM)
H3K9ac
H3K4me3
H3K27ac
H3K4me2
H3K4me1
H4K20me1
H3K36me3
H3K27me3
H3K4me3/H3K27ac
H3K4me3/H3K27me3
H3K4me1/H3K27ac
H3K4me1/H3K27me3
e
Chromatin accessibility
Gene expression
BRF1
RFX5
SMAD1
BRD4
SETDB1
HDAC6
CHD1
BDP1
ATF1
EP300
CBX3
NFXL1
SMARCE1
ATF2
EZH2
KDM1A
CHD4
CTCF
SUZ12
CBX2
CBX8
c
Chr 12
56,500,000bp
56,540,000bp
56,580,000bp
56,620,000bp
DNase-seq (ENCODE)
ATAC-seq (50,000 cells)
CAT-seq (500 cells)
Aggregate scCAT-seq
scCAT-seq
(Chromatin accessibility)
N = 74
ERBB3
PA2G4
RPL41
ESYT1
MYL6B
SMARCC2
RNF41
NABP2
Chr12
56,500,000bp
56,502,000bp
56,504,000bp
56,506,000bp
RNA-seq (ENCODE)
CAT-seq (500 cells)
Aggregate scCAT-seq
scCAT-seq
(Gene expression)
N = 74
PA2G4

Figure 1. scCAT-seq provide accurate genome-wide measure of both chromatin accessibility and gene expression. (**a**) Overview of the scCAT-seq protocol. (**b**) Top panel: chromatin accessibility read enrichment around the transcription start site (TSS). Bottom panel: coverage of mRNA reads along the body of transcripts. Titration series (one single-cell, 5 cells, 50 cells, 500 cells) were marked by the indicated colors. All profiles were generated using the scCAT-seq protocol with the indicated number of cells as input. (**c**) A representative region showing consistent pattern of chromatin accessibility and transcriptome across datasets generated using different number of input cells. The bulk ATAC-seq track was generated using 50,000 K562 cells. The DNase-seq and bulk RNA-seq data of K562 cells were downloaded from ENCODE. The scCAT-seq tracks were chromatin accessibility (upper) and gene expression reads density (bottom) of a total of 74 K562 single cells. (**d**) Top panel: average chromatin accessibility reads density around regions that are enriched by the indicated individual or combined histone modifications. Bottom panel: average expression level of genes associated with regions that are enriched by the indicated individual or combined histone modifications. (**e**) Top panel: average chromatin accessibility reads density within regions that are bound by the indicated transcription factors. Bottom panel: average expression level of genes associated with regions that are bound by the indicated transcription factors.

Figure 2

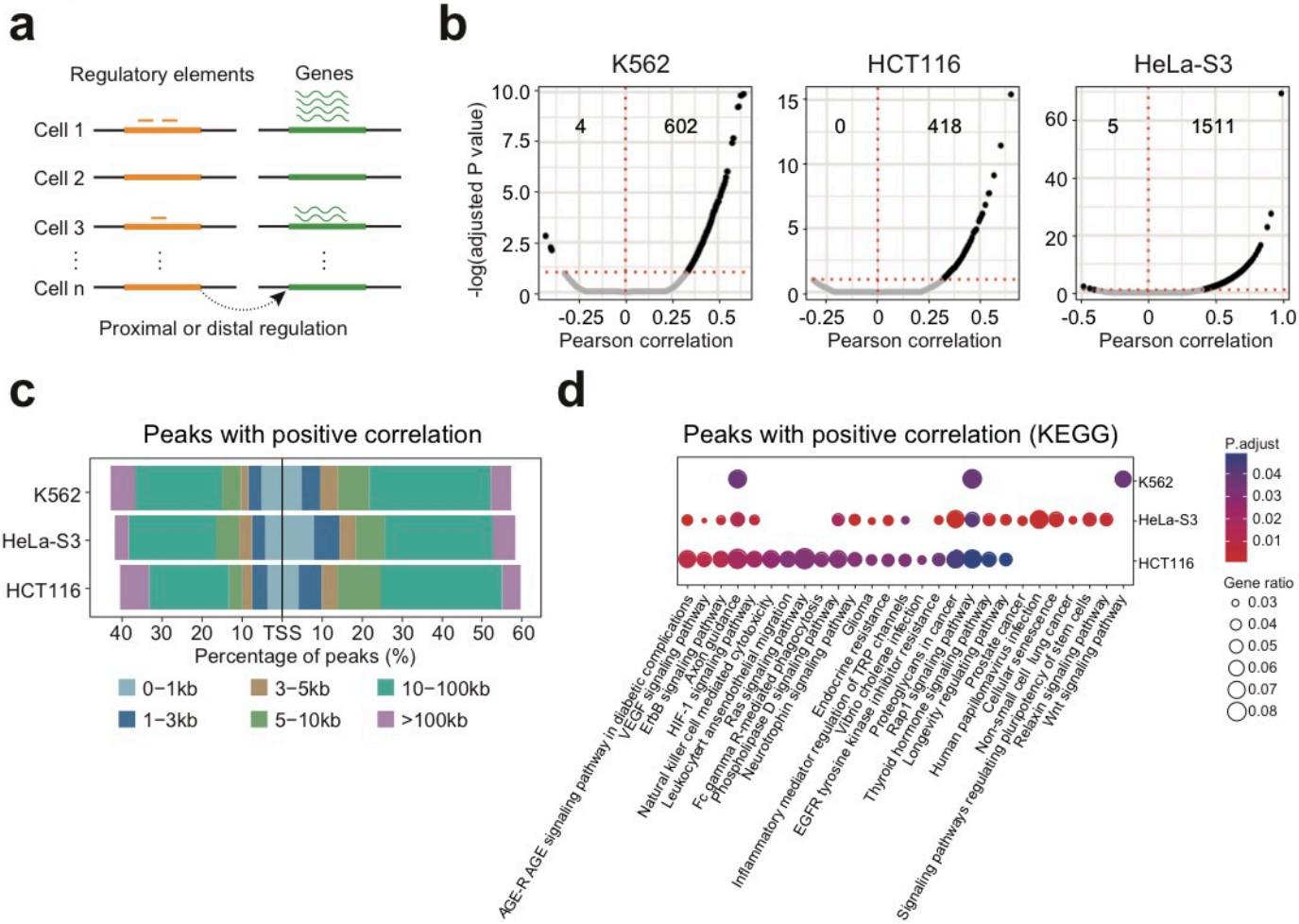

a
Regulatory elements
Genes
Cell 1
Cell 2
Cell 3
Cell n
Proximal or distal regulation
b
K562
HCT116
HeLa-S3
-log(adjusted P value)
4
602
0
418
5
1511
Pearson correlation
c
Peaks with positive correlation
K562
HeLa-S3
HCT116
TSS
Percentage of peaks (%)
0−1kb
1−3kb
3−5kb
5−10kb
10−100kb
>100kb
d
Peaks with positive correlation (KEGG)
AGE-R AGE signaling pathway in diabetic complications
VEGF signaling pathway
ErbB signaling pathway
Axon guidance
HIF-1 signaling pathway
Natural killer cell mediated cytotoxicity
Leukocyte transendothelial migration
Fc gamma R-mediated phagocytosis
Phospholipase D signaling pathway
Neurotrophin signaling pathway
Glioma
Endocrine resistance
Regulation of TRP channels
Vibrio cholerae infection
Kinase inhibitor resistance
Proteoglycans in cancer
Rap1 signaling pathway
Thyroid hormone signaling pathway
Longevity regulating pathway
Prostate cancer
Human papillomavirus infection
Cellular senescence
Non-small cell lung cancer
Signaling pathways regulating pluripotency of stem cells
Relaxin signaling pathway
Wnt signaling pathway
P.adjust
Gene ratio

Figure 2. scCAT-seq uncovers coordinated dynamics between chromatin accessibility and gene expression across single cells. (**a**) A schematic representation showing the correlation analysis between chromatin accessibility and gene expression. (**b**) Correlation analysis between chromatin accessibility of individual elements and the putative gene expression. Shown are Pearson correlation coefficients versus the Benjamini-Hochberg adjusted P-value (transformed by -log10). Significant associations (adjusted P-value <= 0.05) are highlighted in black. The number of positive (+) and negative (-) associations are indicated as well. (**c**) Genomic distribution of regulatory elements that are positively associated with their putative genes. (**d**) KEGG functional enrichment of regulatory elements that are positively correlated with gene expression.

Figure 3

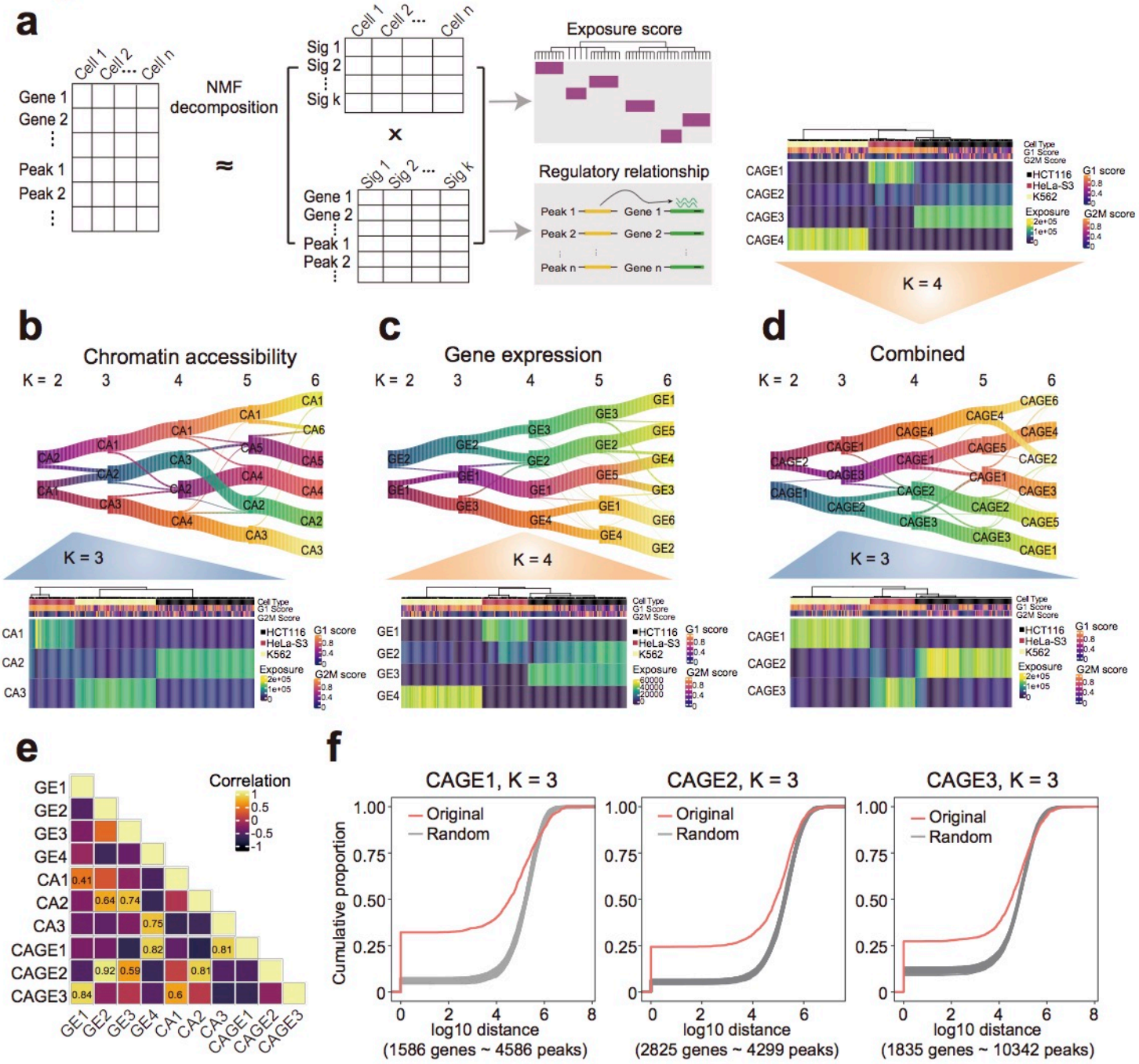

a
Cell 1
Cell 2
Cell n
Gene 1
Gene 2
Peak 1
Peak 2
NMF decomposition
≈
Sig 1
Sig 2
Sig k
x
Exposure score
Regulatory relationship
Peak 1
Peak 2
Peak n
Gene 1
Gene 2
Gene n
Cell Type
G1 Score
G2M Score
HCT116
HeLa-S3
K562
Exposure
G1 score
G2M score
K = 4
b
Chromatin accessibility
K = 2
3
4
5
6
K = 3
c
Gene expression
K = 4
d
Combined
K = 3
e
Correlation
f
CAGE1, K = 3
CAGE2, K = 3
CAGE3, K = 3
Original
Random
Cumulative proportion
log10 distance
(1586 genes ~ 4586 peaks)
(2825 genes ~ 4299 peaks)
(1835 genes ~ 10342 peaks)

Figure 3. Decomposition of combined chromatin accessibility and gene expression features reveals a profound relationship between the two omics layers. (**a**) Schematic representation showing the framework of applying NMF algorithm to the decomposition of scCAT-seq data. (**b**-**d**) Results of the NMF applied to the CA (b), GE(c) or combined datasets (d) respectively. Top panel: riverplot showing tree-like representation of the signatures extracted from dataset profiles of the three cell lines at different factorization ranks. Nodes represent the signatures, and the edge strength indicates the similarity between signatures linked by the edges. The factorization rank increases from left to right. Bottom panel: heatmap showing the exposure scores of all cells in the three cell types to each signature with the indicated K values. (**e**) Heatmap showing Spearman correlation coefficients between signatures from NMF decomposition of separate (CA or GE) and combined (CAGE) scCAT-seq profiles. (**f**) Distance between signature-specific peaks and genes within each combined signature (red line), comparing with distance between randomly selected peak-gene pairs (gray line).

Figure 4

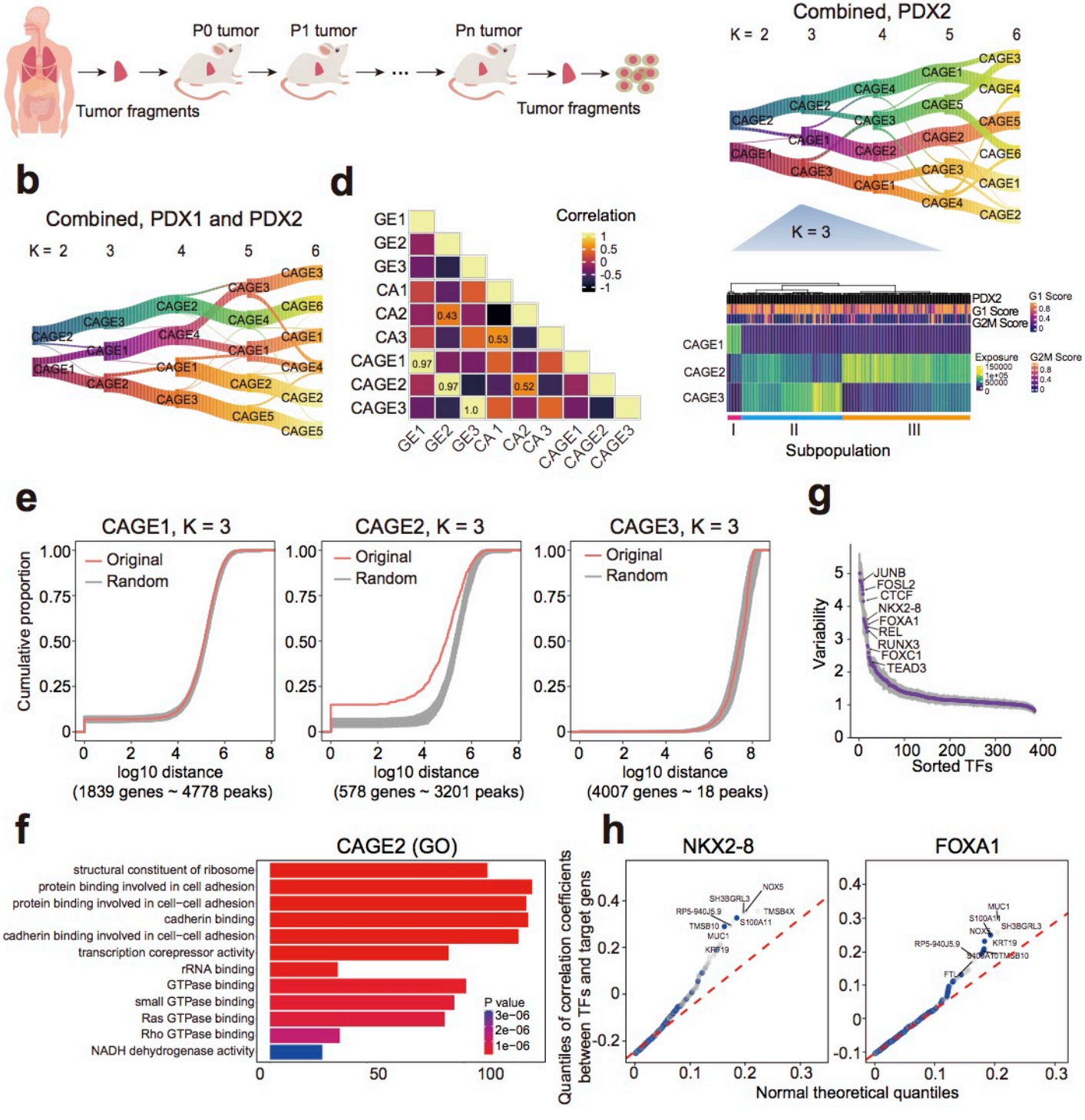
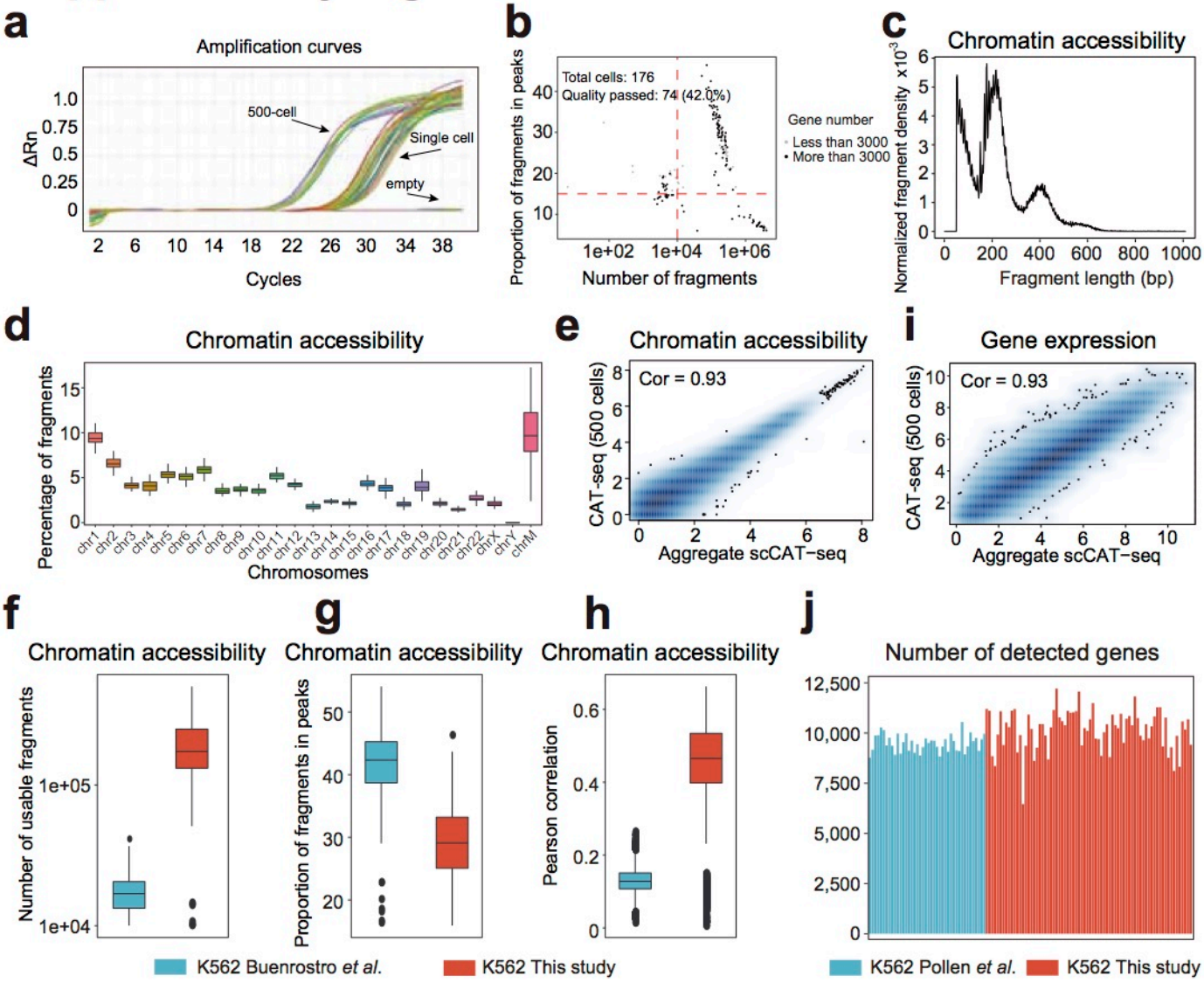

a
P0 tumor
P1 tumor
Pn tumor
Tumor fragments
Tumor fragments
b
Combined, PDX1 and PDX2
K = 2 3 4 5 6
c
Combined, PDX2
K = 2 3 4 5 6
K = 3
PDX2
G1 Score
G2M Score
Exposure
CAGE1
CAGE2
CAGE3
I II III
Subpopulation
d
Correlation
GE1 GE2 GE3 CA1 CA2 CA3 CAGE1 CAGE2 CAGE3
e
CAGE1, K = 3
CAGE2, K = 3
CAGE3, K = 3
Original
Random
Cumulative proportion
log10 distance
(1839 genes ~ 4778 peaks)
(578 genes ~ 3201 peaks)
(4007 genes ~ 18 peaks)
g
Variability
Sorted TFs
JUNB
FOSL2
CTCF
NKX2-8
FOXA1
REL
RUNX3
FOXC1
TEAD3
f
CAGE2 (GO)
structural constituent of ribosome
protein binding involved in cell adhesion
protein binding involved in cell−cell adhesion
cadherin binding
cadherin binding involved in cell−cell adhesion
transcription corepressor activity
rRNA binding
GTPase binding
small GTPase binding
Ras GTPase binding
Rho GTPase binding
NADH dehydrogenase activity
P value
h
NKX2-8
FOXA1
Quantiles of correlation coefficients between TFs and target gens
Normal theoretical quantiles

Figure 4. Dissection of regulatory heterogeneity within PDX models. (**a**) Schematic representation of the experiments. Patient tumor tissues were propagated by xenograft transplantation in humanized immunocompromised NOG mice. PDX cells were dissociated and digested into single cells and then subjected to scCAT-seq workflow. (**b**) Riverplot showing tree-like representation of the signatures extracted from combined scCAT-seq profiles (CA and GE) of PDX1 and PDX2 single cells at different factorization ranks. (**c**) Top panel: riverplot showing the signatures extracted from combined scCAT-seq profiles of the PDX2 single cells at different factorization ranks. Bottom panel: heatmap showing cell exposure to the signatures extracted with K = 3 in PDX2 single cells. (**d**) Heatmap showing Spearman correlation coefficients between signatures from NMF decomposition of separate (CA or GE) and combined (CAGE) scCAT-seq profiles of PDX2. (**e**) Distance between peaks and genes within each combined signature (red line), comparing with distance between randomly selected peak-gene pairs (grey line). (**f**) GO functional enrichment of genes and peaks in combined signature 2 (CAGE2). (**g**) Observed cell-to-cell variability within sets of genomic regions associated with transcription factor motifs. (**h**) qqplot showing the excess of positive correlations between transcription factor motif deviation score (computed by ChromVAR) and their putative gene expression, comparing with a randomised dataset. Shown are two examples of motifs including NKX2-8 (left) and FOXA1 (right). Each dot represents a gene, and the blue dots indicate that the motif is present in the promoter of the gene. The names of the most significant genes are indicated as well.

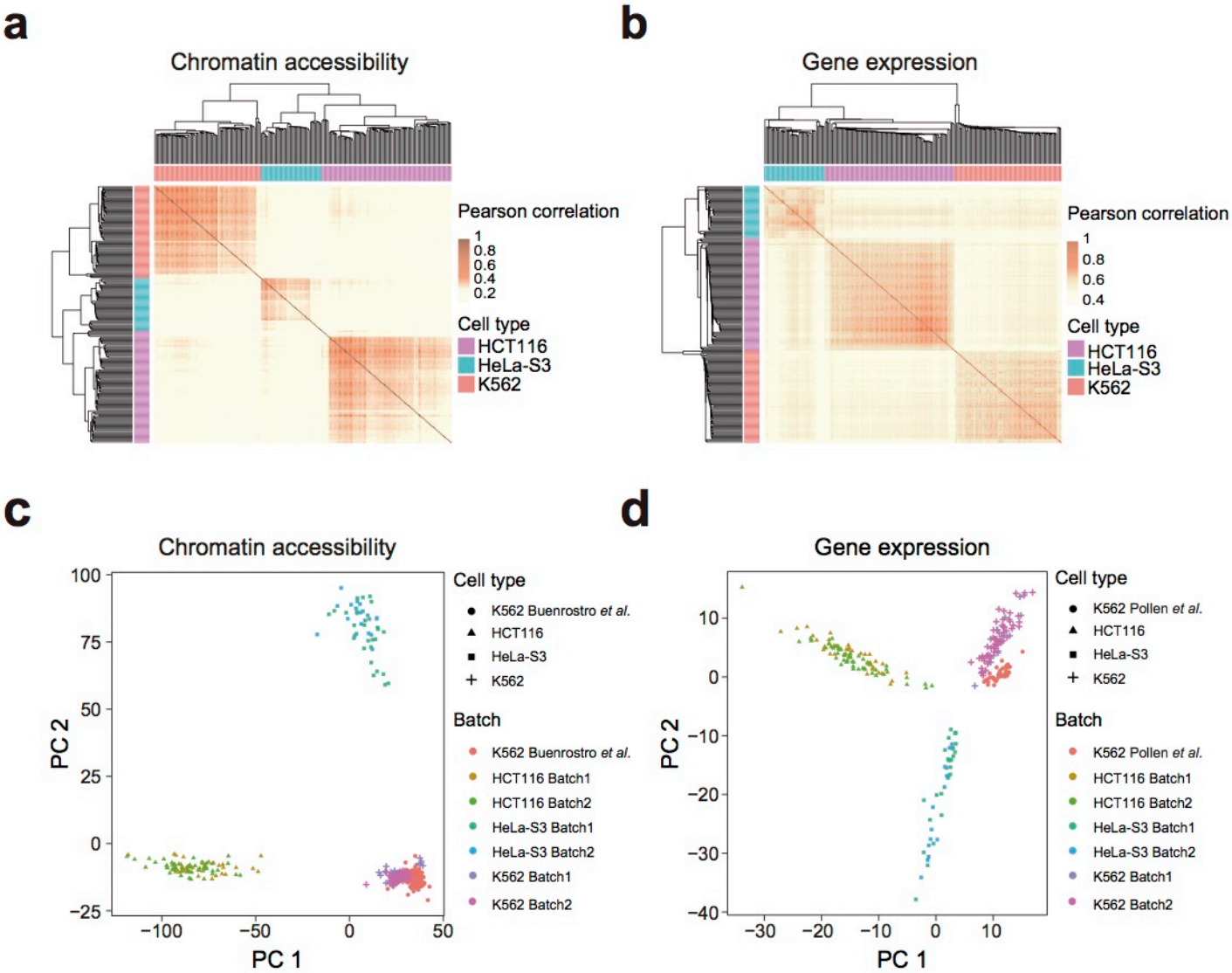
Supplementary Figure 1



Supplementary Figure 1

Quality metrics of scCAT-seq data. (**a**) qPCR amplification curve using materials in the bottom of wells after the separation step of the scCAT-seq protocol. Wells containing 0, 1 and 500 cells were analyzed. After the separation step the materials were amplified for 8 cycles using primers targeting the Tn5 adaptor. The PCR product was then purified and amplified by qPCR using primers targeting an accessible region in the human genome. (**b**) K562 scCAT-seq profiles were quality-filtered according to the number of total fragments, proportion of fragments within reference peaks and detected gene numbers. (**c**) DNA fragment size distribution of chromatin accessibility fragments from single cells display characteristic nucleosome-associated periodicity. (**d**) Percentage of the single cell chromatin accessibility reads mapped to each chromosome and the mitochondrial genome. (**e**) Correlation between aggregate single-cell chromatin accessibility profiles and bulk ATAC-seq profiles. (**f**-**h**) Comparison of number of usable chromatin accessibility fragments (f), proportion of fragments within accessible regions (g) and Pearson correlation coefficients (h) between scCAT-seq and published scATAC-seq profiles. (**i**) Correlation between aggregate gene expression profiles from single cells and gene expression profiles from 500 cells. (**j**) Comparison of the number of detected genes between scCAT-seq and published scRNA-seq profiles.

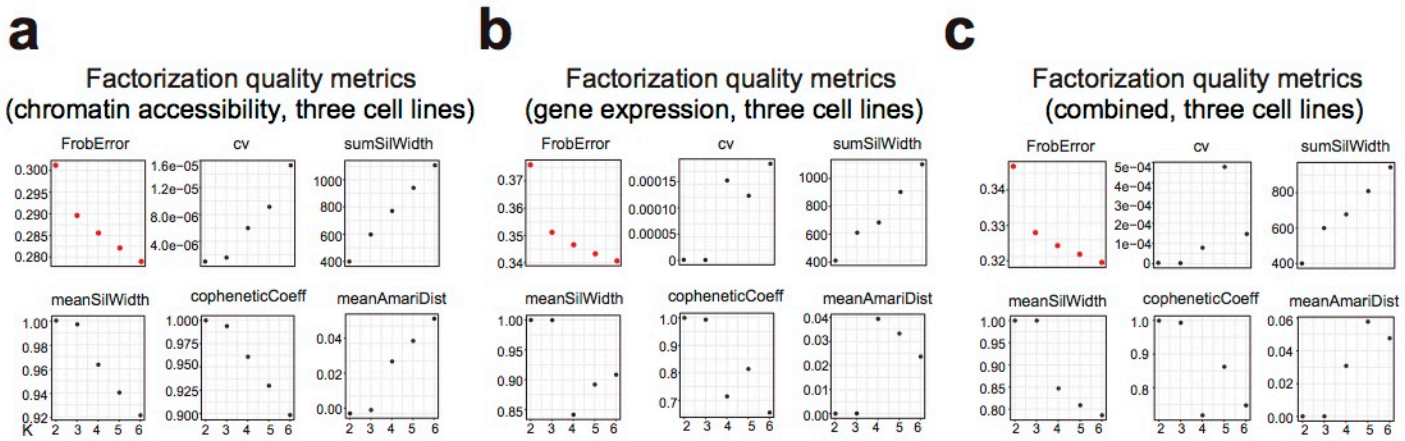
Supplementary Figure 2



Supplementary Figure 2

Correlation analysis across scCAT-seq profiles. (**a**-**b**) Heatmap showing the correlation between scCAT-seq-generated chromatin accessibility (a) and gene expression (b) profiles. (**c**) Principle component analysis of different batches of scCAT-seq-generated chromatin accessibility data and published datasets. (**d**) Principle component analysis of different batches of scCAT-seq-generated gene expression data and published datasets.

Supplementary Figure 3

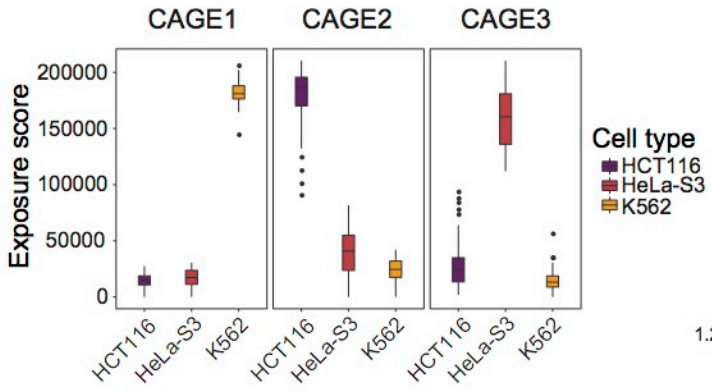


Supplementary Figure 3

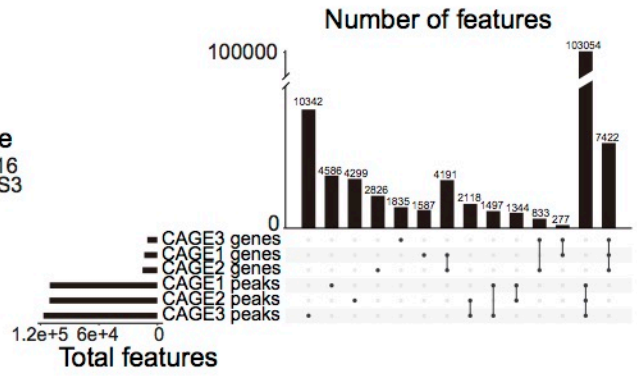
Quality metrics of CA and GE dataset NMF with different factorization ranks. Shown are Frobenius errors, cophenetic coefficients and mean Amari distance for CA (a), GE (b) and combined signature (c) factorizations. The optimal factorization rank was selected as the K that best satisfies the quality metrics criteria: minimize the Frobenius error and the mean Amari distance, while maximizing the cophenetic correlation coefficient.

Supplementary Figure 4

a



b



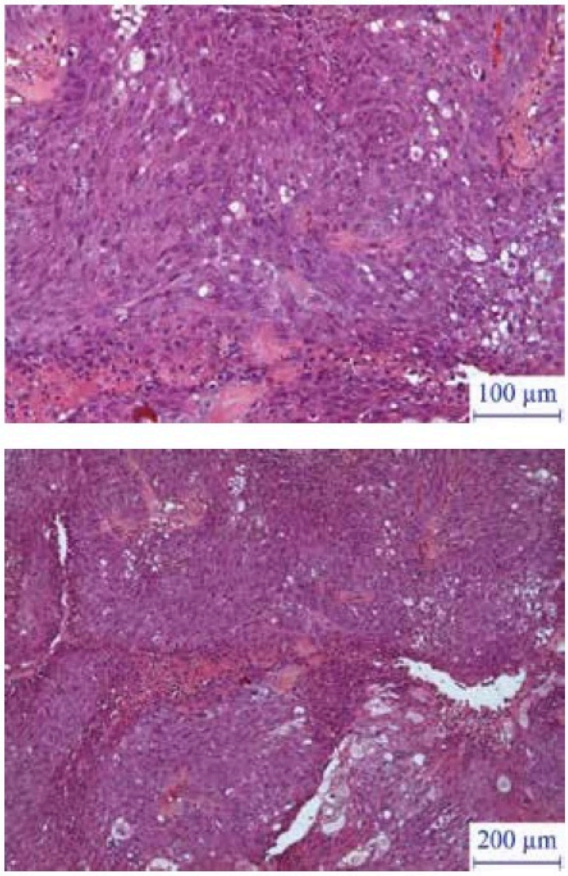
Supplementary Figure 4

Analysis of the combined signatures from NMF analysis of scCAT-seq profiles of the three cell types. (**a**) Boxplots showing exposure scores of each signature in the three cell types. (**b**) Number of gene or peak features that are either specific to one signature or shared by some of the signatures.

Supplementary Figure 5

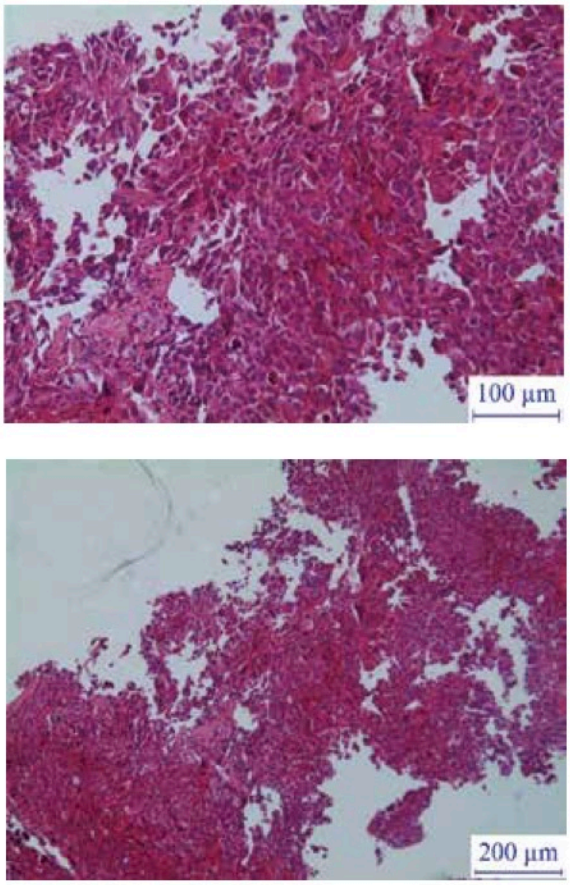
a

PDX1



b

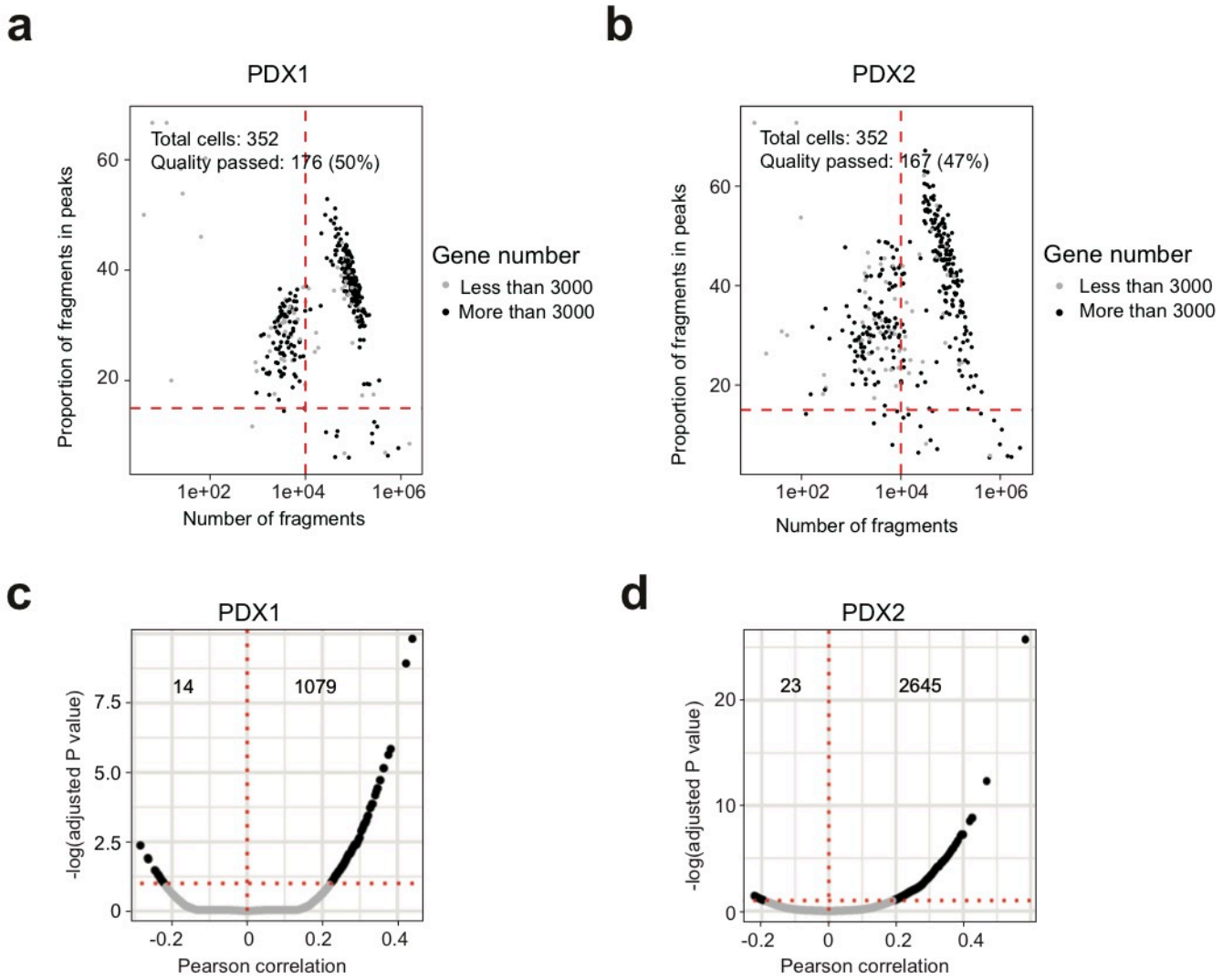
PDX2



Supplementary Figure 5

Hematoxylin and eosin staining images of PDX tissues derived from (a) a moderately differentiated squamous cell carcinoma patient (PDX1) and (b) a large-cell lung carcinoma patient (PDX2).

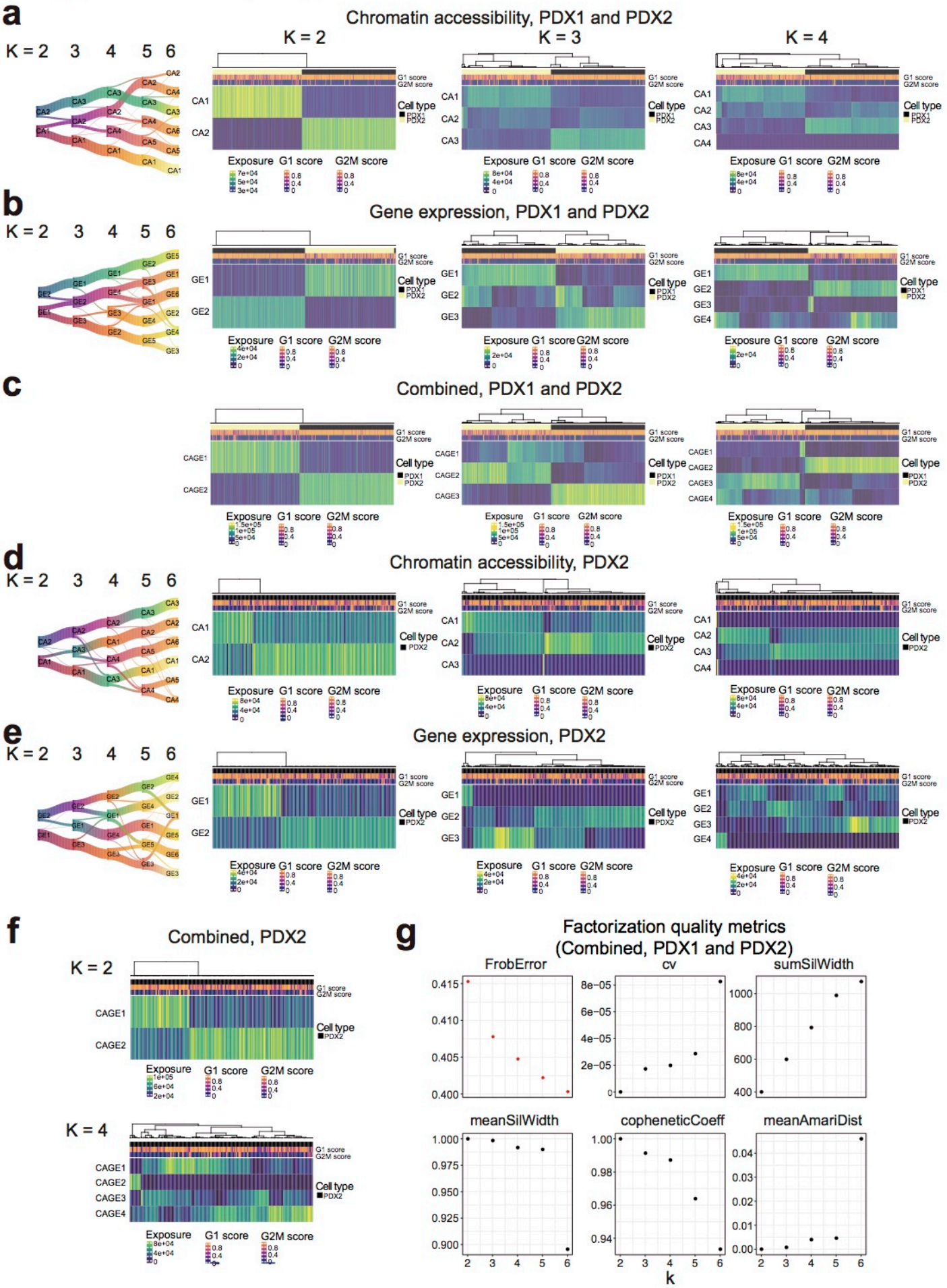
Supplementary Figure 6



Supplementary Figure 6

Quality control and association analysis for PDX profiles. (**a**) PDX1 and PDX2 scCAT-seq profiles were quality-filtered according to the number of total reads, reads within peaks and detected gene number. (**b**) Correlation analysis between chromatin accessibility of individual elements and the putative gene expression using profiles of the two PDX models. Shown are Pearson correlation coefficients versus the Benjamini-Hochberg adjusted P-value (transformed by -log10). Significant associations (P <= 0.05) are highlighted in black. The number of positive (+) and negative (−) associations are indicated as well.

Supplementary Figure 7

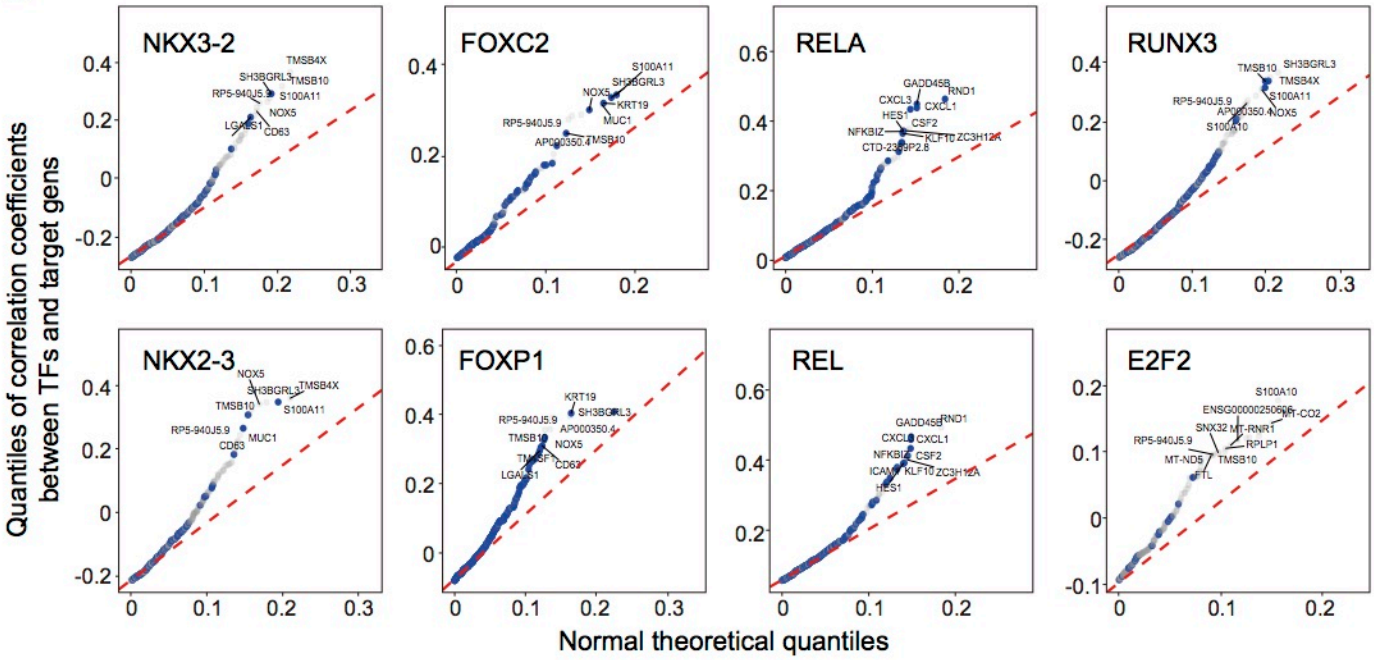


Supplementary Figure 7

NMF decomposition analysis in PDX models. (**a**-**c**) Left panel: riverplot showing the signatures extracted from CA (a), GE (b) and combined (c) profiles of the two PDX models at different factorization ranks. Right panel: heatmap showing the clustering of exposure scores of each signature from CA (a), GE (b) and combined (c) decomposition with K = [2,3,4] in all single cells of two PDX models. (**d**-**e**) Left panel: riverplot showing the signatures extracted from CA (d) and GE (e) profiles of the PDX2 model at different factorization ranks. Right panel: heatmap showing the clustering of exposure scores of each CA (a) and GE (b) signature with K = [2,3,4] in all single cells of the PDX2 model. (**f**) Heatmap showing the clustering of exposure scores of each combined signature with K = [2,4] in all single cells of the PDX2 model. (**g**) Quality metrics of combined PDX2 dataset NMF with different factorization ranks.

Supplementary Figure 8

a



Supplementary Figure 8

qqplot showing the excess of expected positive correlations between transcription factor motif deviation score (z-score computed by ChromVAR) and their putative gene expression compared with a randomized dataset. Shown are examples of motifs. Each dot represents a gene; the blue dots indicate that the motif is present in the promoter of the gene. The names of the most significant genes are indicated as well.