High-throughput single-nucleus hybrid sequencing reveals genome-transcriptome correlations in cancer

Siran Li[1*], Joan Alexander[1], Jude Kendall[1], Peter Andrews[1], Elizabeth Rose[1], Hope Orjuela[1], Sarah Park[1], Craig Podszus[1], Liam Shanley[1], Rong Ma[2], Nissim Ranade[1], Michael Ronemus[1], Arvind Rishi[3], David L. Donoho[2], Gary L. Goldberg[4], Dan Levy[1], Michael Wigler[1*]

[1] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

[2] Department of Statistics, Stanford University, Stanford, CA, USA

[3] Department of Pathology and Laboratory Medicine, Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, USA.

[4] Department of Obstetrics and Gynecology, Division of Gynecologic Oncology, Zucker School of Medicine at Hofstra/Northwell, Northwell Health, New Hyde Park, NY, USA

* Correspondence: siranli@cshl.edu; wigler@cshl.edu

Abstract

Single-cell genomic analyses can provide information on cellular mutation and tumor heterogeneity, whereas single-cell transcriptomic analyses can distinguish cell types and states. However, the disconnect between genomic and transcriptomic spaces limits our understanding of cancer development. To address this, we developed a novel high-throughput method that simultaneously captures both DNA and RNA from single nuclei and new algorithms for the quantitative clustering and filtering of single-cell data. We applied this hybrid protocol to 65,499 single nuclei extracted from frozen biopsies of five different endometrial cancer patients and separately clustered the genome and expression data. We also analyzed 34,651 and 21,432 nuclei using RNA-only and DNA-only protocols, respectively, from the same samples to verify the clustering. Multiple tumor genome and/or expression clusters were often present within an individual patient, and different tumor clones could project into distinct or shared expression states. Almost all possible genome-transcriptome correlations were observed in the cohort. Stromal clusters were largely shared between patients, but some patients possessed unique stromal components, or mutant stroma with a significant loss of the X chromosome. This study reveals the complex landscape involving genome and transcriptome interactions at single-cell level, and provides new insights into mutant stroma as a potential clinical biomarker.

Main

To enhance our understanding and treatment of cancer, it is important to understand its heterogeneity, its interaction with the host, and the role the host plays in assisting or inhibiting the invasive somatic clone. Single-cell analysis offers one possible route to improved understanding. Integrating genomic and transcriptomic information would enable us to better explore the stromal reaction to neoplasm and the diversity of transcriptional states within tumors, making it possible to understand cooperation and competition between cancer cells and stroma, to view the emergence of malignant from pre-malignant cells, and to discern the forces that drive particular expression states. To achieve this aim and demonstrate its potential, we developed and applied a high-throughput single-cell

analysis of DNA and RNA to five cases of uterine cancer. We show the method yields new information, and provides insights into the nature of the cancer stroma.

Methods for high throughput single-cell DNA[1-3] or RNA[4-11] alone are well established. For investigating both omics, there are methods[12-16] for inferring copy number states from RNA-seq data based on the assumed positive correlation between genome copies and gene expression. While there are low-throughput methods[17-28] for capturing both nucleic acids from single cells, among the limited high-throughput whole-genome and whole-transcriptome techniques[29], none have been applied to tumor biopsies to systematically study the cancer heterogeneity and stromal mutations with the throughput demonstrated in this study.

In this paper, we introduced an innovative high-throughput technology in which both DNA and RNA sequences can be captured from the same cell nucleus. Post-sequencing, the DNA and RNA layers are bioinformatically separated by their mapping properties and subsequently analyzed. We chose to work with nuclei because nuclei from frozen biopsies were a more available and abundant clinical sample source than cells. This multi-omics method is a progression from the single-omic BAG platform[30], in which single cells were encapsulated into individual balls of acrylamide gel, with either DNA or RNA captured by Acrydite primers that were copolymerized into the gel matrix. We applied our hybrid DNA-RNA approach to frozen tissue biopsies of five patients with endometrial cancer. We found sufficient transcriptional complexity in the nuclear RNA to cluster by cell type and state, and sufficient copy number information in the DNA layer to readily distinguish stroma from tumor components by genomic DNA analyses. These clustering patterns were confirmed by published RNA-only and DNA-only protocols. In addition, we developed a novel multinomial algorithm, the “multinomial wheel,” to quantitatively measure the deviation of each single cell from the main clusters. This allowed us to effectively remove cell collisions, and gave us insights into the minority of cells showcasing tumor genomes and stromal expression patterns and those with normal genomes but tumor expression profiles.

Although tumor expression clusters are highly distinct between patients, within a given patient, we often observed multiple tumor expression states. We found that one tumor DNA clone may project into one or more of these distinct expression states, which may or may not be shared by another tumor clone from the same patient. In five patients, we observed virtually every possible projection pattern between genomic and transcriptomic states. Conversely, stromal expression clusters are largely shared between patients, although in different proportions. Two patients exhibited unique stromal components, and all five displayed instances of mutant stroma. In the patient with the worst clinical outcome, almost half of her plasma cells lost one copy of the X chromosome. These observations demonstrated the potential and immediate applicability of this multiomic method in the investigation of cancer evolution, stromal mutations, and the intricate interplay between genome and transcriptome.

Results

The Hybrid Platform

We study DNA and RNA templates from the same individual nuclei using BAG technology, a flexible platform open to design modifications[30]. The BAG platform extends the capabilities of traditional droplet methods[3,4] by polymerizing droplets containing nuclei in the presence of Acrydite-modified primers. The Acrydite-modified primers form a primer-template duplex with the nucleic acid contents of the cell or nucleus. Once polymerized, the “**b**alls of **a**crylamide **g**el” or BAGs are removed from oil and processed in aqueous solution. For DNA-only or RNA-only protocols, the primers that capture nucleic acids are extended using polymerase or reverse transcriptase, resulting in template copies from the same individual nuclei being covalently bound together in the same bag. After processing, we randomly distribute BAGs into 96 wells, affixing one of 96 unique well-barcode sequences to each template on

each BAG in each well. Pooling the BAGs together and randomly splitting them again is defined as a "pool-and-split" process. We repeat this process three times to generate a unique signature of 96^3 or about one million unique **BAG barcodes**. During the first two pool-and-split labeling steps, in addition to the BAG barcodes, we also add eight random bases to each template. When combined with the four bases from the genomic sequences, these twelve random bases form a **varietal tag** (or UMI) that uniquely labels each template molecule.

In this paper, we used BAGs to capture DNA and RNA from the same individual nucleus or cell under mild denaturing conditions. To capture both types of nucleic acids, we used a mixture of Acrydite-modified random T/G and oligo d(T) primers (**Fig. 1a**). The primers are then extended using reverse transcriptase and DNA polymerase under conditions favorable to both reactions (see **Methods** and **Supplementary Protocol**). We apply the pool-and-split labeling process with varietal tags as described above, and then combine all the BAGs into a single PCR reaction, followed by tagmentation to generate a sequencing library.

The targets of our investigation are frozen biopsy samples from five uterine cancer patients. The samples include tumor tissue, which we denote as **Tumor 1** through **Tumor 5**. For three patients (1, 2 and 4), we also have samples from normal adjacent tissue, which we denote as **Normal 1**, **Normal 2**, and **Normal 4**. In addition to our new **hybrid** protocol for simultaneously analyzing both DNA and RNA, we also generated sequencing libraries using traditional BAG sequencing of DNA alone (labeled as **DNA-only**) and RNA alone (labeled as **RNA-only**) on the same samples. The main differences between the three protocols are illustrated in **Supplementary Fig. 1a**. Additionally, we applied the 10x Genomic Chromium v3 single-cell RNA sequencing method on Tumor 1 for comparison purposes.

By design, the paired-end reads in the sequencing library are asymmetric: one end (Read 2) of the read-pair contains information about the template sequence, while the other end (Read 1) contains the BAG barcode, the varietal tag, and some template sequence. We first confirm that the reads have the correct structure with BAG barcodes and NLAIII cutting sites at the correct positions (**Fig. 1a**), and then extract the BAG barcode, varietal tag, and sequence information. We then use HISAT2[31], a sensitive alignment program useful for mapping both DNA and RNA reads, to map the template sequences. On average, 84% of reads have a high-quality map with a mapping quality score equal to 1 or 60 and are only primary mappings. Since sequence amplification occurs before tagmentation, we often obtain reads from different fragmented copies of the same initial template. These reads will have the same BAG barcode, varietal tag, and similar mapping regions, but they will have different sequence information on the opposite end of the read-pair. For this reason, we group reads with the same BAG barcode and varietal tag into **templates**, and templates with the same BAG barcode are grouped into **BAGs**.

We only consider BAGs with a sufficient number of templates, using the "elbow bend" in the cumulative distribution to establish the cutoff. For nuclei from frozen tissue biopsies, on average, we observe 20 reads per template, 8200 templates per BAG, and 3500 BAGs per experiment when using the hybrid protocol. The sequence and its context are used to discern whether templates are derived from DNA or RNA. For each template, we document all its mapping information, including overlap with exons, introns, UTRs, and behavior at splice sites. We established four categories:

1. **Exonic**: If greater than 90% of template bases are within a single gene transcript **and** either >50% of the bases mapped to exons of that gene or the template includes a known splice junction of the target gene.
2. **Intergenic**: If all template bases are intergenic.
3. **Intronic**: If less than 10% of template bases map to exons and not intergenic.
4. **Uncategorized**: does not satisfy 1-3.

We apply these categorization rules to all seven datasets for all protocols. The full distribution of each category per sample is shown in **Supplementary Fig. 1b**. From the DNA-only protocol, we find that 2.5% of templates are Exonic, 52% are Intergenic and 45% are Intronic, aligning closely with the proportions

anticipated from random sampling of the genome. In contrast, the great majority (>80%) RNA-only data templates are mapped to genes, and a sizable proportion are mapped to exons (about 18%) and only 17% are Intergenic. Templates from the hybrid protocol were intermediate to the two: 10% Exonic templates, 30% Intergenic templates, and 55% Intronic templates. The hybrid protocol is composed of a mixture of DNA-only and RNA-only distributions. To divide the templates into RNA or DNA components, we apply a conservative rule: we restrict RNA data analysis to Exonic templates (**RNA layer**) and we restrict DNA data analysis to Intergenic templates (DNA layer) (see Methods for detailed filtering process). The distribution of total unique templates, unique Exonic templates, and unique genes for each sample in the cancer cohort—as well as for a human-mouse mixture experiment—using the hybrid protocol, is presented in **Supplementary Fig. 2**.

Splitting layers comparison

Because our informatics relies on a conservative segregation of templates into molecular layers, we expect some degradation of signal from excluded templates. To measure the extent of this loss, we apply this same layer-splitting method to the hybrid protocol, as well as to the DNA-only and RNA-only datasets. Before conducting separate layer analyses, we found it necessary to introduce additional measures to control bias in the hybrid data, as we observed interference of RNA with DNA profiling. For DNA profiling and clustering, we excluded certain intergenic genomic hotspots that many nuclear RNA sequences map to, which significantly increased the quality of copy number data from the hybrid protocol (**Supplementary Fig. 3**). These hotspots were present in data from both tumor and normal tissues. The copy number data quality from the hybrid protocol was substantially superior to that derived from the RNA-only protocol, with a quantitative measure of the signal-to-noise ratio being discussed in the following section.

DNA-layer results

Restricting the DNA-only sequence data to the DNA layer results in a 54% reduction in template counts. First, to estimate the effect on the signal-to-noise ratio for single cells between different protocols, we use **Tumor 1** (**Fig. 2a**) which has a 100 MB deletion on chromosome 5 that spans 13 bins for all tumor cells despite different tumor clones. For each single cell, we compare the average template counts of a normal copy 2 region (13 bins from chromosome 10) to the average template counts in the copy 1 region in chromosome 5. The ratio of these averages estimates the signal strength between copy numbers 2 and 1. We compute the mean normalized standard deviations over the bins in these regions to estimate the relative noise. **Supplementary Fig. 4** shows the mean versus standard deviation (SD) for single cells from the DNA layer for DNA-only (green), RNA-only (blue), and the hybrid method (orange). We also show the results of using all the templates from the DNA-only experiments (red). As expected, we have the least noise using all the DNA-only data (SD of 0.148). The signal-noise-ratios (mean/SD) for all-molecules DNA-only data, DNA layers of DNA-only data, hybrid data, and RNA-only data are 13.0, 9.7, 4.3, and 1.0, respectively.

Second, we demonstrate that clustering copy number patterns using DNA-layer molecules or all DNA molecules generates similar DNA clonal information. Using DNA-only data from Tumor 2 as an example (**Supplementary Fig. 5**), we show that Seurat clustering generates the same number of clusters (**Supplementary Fig. 5a,b**) and almost identical copy-number heatmaps (**Supplementary Fig. 5c,d**) using either all molecules or only DNA-layer molecules as bin counts. Both clusterings in Supplementary Fig. 5a and 5b use the same nuclei. The heatmap in **Supplementary Fig. 5e** shows the number of cells in each clone of the two clusterings, and we find that most of the cells belong to the same DNA clones in both clusterings. To quantify this result, we examine all the nuclei pairs in both clusterings and check whether each nuclei pair stays in the same or different DNA clones when clustered using all molecules or

only DNA-layer molecules. We present the result in **Supplementary Fig. 5f** and show that 97.6% of the nuclei pairs are on the diagonal, indicating that the DNA layer preserves the DNA clonal information.

Third, we show the similarity between the DNA layer of the hybrid protocol and the DNA-only protocol. We use tumor clones of Tumor2 as an example. Clusterings of both protocols generate the same number of DNA clones and copy-number patterns (**Supplementary Fig. 6a,b**). To quantify this result, we measure the proximity of every single tumor nucleus to the centroid of each tumor clone. We compute the centroid of each tumor clone by averaging all the nuclei belonging to that clone as determined by Seurat clustering. Given the centroid of each of the four tumor clones and 8 intermediate linear combinations equally spaced between each pair of tumor clone centroids (for a total of 52 possible states), we calculated the most likely state for each nucleus based on multinomial distribution. For each nucleus, the distance between this spot and its Seurat-assigned clone is called its "Distance from home". More details and validation of multinomial wheel analysis will be further discussed in the last section. In **Supplementary Fig. 6c,d**, we plot the most likely state for each tumor nucleus from the hybrid protocol and DNA-only protocol, respectively, with the colors marking its cluster identity. The histograms on the right show the distribution of the distance from home for each nucleus. Compared to the DNA-only protocol where 84.5% of the nuclei are within two units from the centroid of each clone, this number drops to 77.2% for the hybrid protocol. Therefore, we estimate the resolution of the hybrid protocol dropped by about 9% compared to the DNA-only protocol. Furthermore, we calculate the distance of nuclei from the hybrid protocol to the tumor clone centroid determined by the DNA-only protocol, and vice versa. As shown in **Supplementary Fig. 6e,f**, there is a 10.7% reduction of hybrid data to DNA-only centroids compared to hybrid centroids, but there is no reduction of DNA-only data on hybrid protocol centroids, showing the normalized averaged bin counts of each tumor clone for each protocol are similar. Heatmaps of all five cases show clonal similarities between the hybrid protocol and DNA-only protocol (**Supplementary Fig. 7**.)

RNA-layer results

Similar to the comparison between DNA layer and all DNA molecules, we use RNA-only data from Tumor 2 to demonstrate that restricting the analysis to the RNA layer generates similar results to using all of the RNA templates mapped within transcripts (**Supplementary Fig. 8**). UMAP clustering using only the RNA layer templates (**Supplementary Fig. 8a**) or all RNA templates (**Supplementary Fig. 8b**) generates the same number of clusters. Both clusterings were performed using the same group of nuclei. The number of nuclei in each clustering is shown in **Supplementary Fig. 8c**, in which 94.3% of the nuclei remain in the same expression clusters regardless of whether the RNA-layer or all-molecule clustering was performed. We further quantify this result by examining all nuclei pairs to check whether they reside in the same or different clusters under two clusterings **Supplementary Fig. 8d**. We found that 95.2% of the nuclei pairs are on the diagonal, indicating that restricting the analysis to the RNA-layer conserves much of the transcriptional clustering information.

Second, to demonstrate that the expression clustering results were similar between the hybrid and RNA-only protocols, still using Tumor 2 as an example (**Supplementary Fig. 9**), we first clustered the data from each protocol separately (**Supplementary Fig. 9c, 9g**). We then combined the data generated from the two protocols and clustered the merged dataset together, adjusting for technical variations in the protocols (see the **Methods** section). We show nuclei from each protocol in the merged clustering in the panels of **Supplementary Fig. 9a, 9e**. We found that all clusters were populated by nuclei from both protocols (**Supplementary Fig. 9d**). The merged-data clustering generated the same grouping of nuclei as the clustering by either protocol alone. This was illustrated in heatmap matrices of identities (**Supplementary Fig. 9b, 9f**). The columns of the heatmap show clusters from either the hybrid or the RNA protocol alone, while the rows show clusters from the merged data. We found that most of the nuclei fell on the diagonal of the heatmaps, meaning that nuclei of the same type were predominantly grouped together whether clustered alone or in the merged data. For all five cases, the quality of

separations from the hybrid protocols (leftmost panel) and the RNA-only protocols (rightmost panel) was very similar (**Supplementary Fig. 10**).

Finally, we compare the RNA layer of the hybrid protocol to the commonly used 10x Chromium v3 single-cell RNA-seq method (**Supplementary Fig. 11**). We used Tumor 1 as an example since, in addition to diverse stromal components, we observed two distinct tumor expression states (RNA1a and RNA1b) that are independent of the two tumor DNA clones (discussed further in later sections and in **Fig. 2a**). We find that the RNA clustering from 10x Chromium v3 also separates the tumor nuclei into two expression states mainly based on collagen-related gene expressions, and with similar cell proportions to the hybrid protocol (**Supplementary Fig. 11a,b**). To demonstrate the similarities between these two protocols in distinguishing tumor cluster RNA1a from RNA1b, we study the correlation of the fold change between RNA1a and RNA1b of all the genes for the two protocols (**Supplementary Fig. 11c**). We restrict the analysis to genes that were detected in at least 10% of cells in either RNA1a or RNA1b clusters for both protocols. Running an ordinary least squares (OLS) regression of y on x produces a highly significant coefficient of 0.92 (**Supplementary Fig. 11c**). In addition, we plot the ratio of the proportions of cells expressing these genes in RNA1a versus RNA1b for both protocols (**Supplementary Fig. 11d**). Correlation tests under the null hypothesis H_0: $\rho=0$ and the alternative hypothesis H_1: $\rho>0$ confirm that the gene expressions and cell proportions in these two protocols have strong positive correlations with both p-values less than 2.2e-16. The top marker genes that either positively or negatively distinguish RNA1a from RNA1b for both protocols, with p-values from the Wilcoxon Rank Sum test, are listed in **Supplementary Fig. 11e.**

Doublets

We performed a human-mouse nuclei mixture experiment to study the levels of doublets and cross-contamination. Out of 1299 nuclei, after removing templates that mapped to both the mouse genome and human genome, we detected 619 nuclei with the majority (>85%) of the templates mapped to the human genome, and 662 nuclei with the majority (>85%) of the templates mapped to the mouse genome, and there were 18 doubles (1.39%) (**Fig. 1b**). We also show that the level of doublets is consistent between DNA layer and RNA layer, as shown in **Supplementary Fig. 12**, where only 5 out of 1299 nuclei (0.38%) nuclei showed disagreement of identities between two layers. The experiment also presented a low level of cross-contamination. As shown in **Fig. 1c**, the percentage of mouse templates in human nuclei had a median of 0.20%, and the human templates in mouse nuclei had a median percentage of 0.55%. This low level of cross-contamination also is preserved in both the DNA layer and RNA layer (**Supplementary Fig. 12**).

In addition, we performed two mixture experiments using nuclei from tumor biopsies. In each experiment, the frozen material from two patients was mixed prior to preparing single nuclei, sorting, and encapsulating them in BAGs. The source of each nucleus in the mixture experiments could be readily determined by its abundance of germline single nucleotide polymorphisms (SNPs), and collisions/doublets could be readily identified by having SNPs from both sources. One library was a mixture of nuclei from Patient 1 and Patient 5 (**Supplementary Fig. 13a**), and the other was from Patient 2 and Patient 5 (**Supplementary Fig. 13b**). If the SNPs in a BAG were contaminated by at least 15% of SNPs from the other patient, it was considered a collision. Based on this measure, 5.3% (out of 2526 nuclei) and 6.0% (out of 4244 nuclei) of BAGs were judged as collisions in these two experiments, respectively (**Supplementary Fig. 13c, 13d**).

Alluvial diagrams illustrating complex tumor projection patterns

Having validated the clustering patterns from the hybrid platform, we were able to confidently assess the projections of the genomic clusters onto the expression clusters. We first demonstrate this concept through a mixture experiment involving two cell lines: a normal male fibroblast, SKN1, and a

breast cancer cell line, SKBR3. The distributions of the numbers of total unique molecules, Exonic molecules, and detected genes from this experiment are shown in **Fig. 1d-f**. We illustrate the clustering results and heatmaps based on the copy number and gene expression in **Fig. 1g-k**. The alluvial diagram (**Fig. 1i**) shows the projection of the genomic clones into the expression clusters. As expected, we observed a good one-to-one correlation between the genome and transcriptome of each cell type.

After validating the alluvial diagram using cell lines, we used it to illustrate the genome-transcriptome correlations in all five tumor samples (**Fig. 2**). In each panel of Fig. 2, we present the clustering results for both the DNA layer and RNA layer, as well as the copy-number and gene-expression heatmaps to illustrate the features that distinguish each DNA clone or RNA cluster in the five cases. In general, the tumor genome clones projected distinctly from the normal genome clones, although there were exceptions which we term “crossovers” and will discuss in more depth later.

Next, we focused mainly on the tumor genome projection patterns. To classify projections, we used a set of letters and numbers to represent the tumor genome and tumor RNA clusters, respectively. For example, {A:1,2; B:2} indicates tumor clone A projected into RNA clusters 1 and 2, whereas tumor clone B from the same primary tumor tissue projected only into RNA cluster 2. Each of the five tumors had a different projection pattern, and we observed almost all the possible patterns, which are defined as follows: distinct tumor clones could each project into distinct expression clusters (e.g., {A:1; B:2} for Tumor 5), into shared clusters (e.g., {A:1,2; B:1,2} for Tumor 1), or into a combination of distinct and shared clusters (e.g., {A:1,2; B:1} for Tumor 4). Alternatively, multiple DNA clones could project into a single RNA cluster (e.g., {A:1; B:1; C:2; D:2} for Tumor 2), or a single tumor clone could project into two RNA clusters (e.g., {A:1,2} for Tumor 3). We discuss the special aspects of each case in the following paragraphs.

Special aspects of each case

Tumor 1 was a uterine carcinosarcoma with greater than 90% of cells in the tumor tissue having copy number variations (**Fig. 2a**). The copy number heatmap showed two tumor DNA clones, with the primary difference being that clone T1B had a lower copy number in a region on chromosome 13. The alluvial plot showed that cells from each tumor DNA clone projected about equally into both tumor RNA clusters. Cluster RNA1a had a high expression level of fibroblast-specific genes such as fibroblast growth factor receptor genes (*FGFR3*) and collagen genes (*COL9A2*) (**Supplementary Fig. 14a**), consistent with the pathological classification of this tumor as having a sarcomatous component. These fibroblast genes had lower expression in cluster RNA1b. On the other hand, cluster RNA1b had higher expression *of RSPO4*, a key regulator of the Wnt/β-catenin signaling pathway, and *DUSP6*, a negative regulator of the ERK signaling pathway (**Supplementary Fig. 14a**). When we projected the two tumor RNA clusters back into DNA UMAP space, we found that the nuclei from both RNA clusters were randomly distributed in DNA space and unrelated to DNA cluster patterns (**Supplementary Fig. 15a-c**). The two tumor RNA clusters were also not separated by cell cycle, template, or gene counts (**Supplementary Fig. 15d**). The normal nuclei from this patient's primary tumor made up a minor fraction of the total.

Tumor 2 was diagnosed as a uterine serous carcinoma, from which we observed four DNA tumor clones and two tumor RNA clusters (**Fig. 2b**). The two RNA clusters to which these tumor clones projected had many distinguishing gene sets (**Fig. 2b**). Tumor DNA clones T2A and T2B each projected to RNA2a, while T2C and T2D each projected to RNA2b. The major feature shared by T2A and T2B, but not T2C and T2D, was the loss of an entire X chromosome. In fact, one gene that significantly distinguished RNA2a and RNA2b was *XIST*, which had low expression in RNA2a. Compared to tumor 1, the normal genomes of tumor 2 projected to many more distinct RNA clusters. In this example, we observed two normal DNA clusters, one with a single copy of the X chromosome named as DNA clone "Nx". From the alluvial plot in **Fig. 2b**, we found that most of the nuclei in this Nx DNA clone projected to plasma cells as well as T cells. As can be seen in the zoomed image from **Supplementary Fig. 14b**, the projection of “Nx”

into the "Plasma cells" cluster also had a low expression of *XIST*. We quantify this observation by comparing the XIST RNA-layer counts (Exonic templates) between nuclei with two copies and one copy of the X chromosome. Compared to the clone with one copy of the X chromosome, where 95% of the nuclei have 0 XIST RNA counts, there is a significantly higher RNA count (p-value < 1.7e-22, t-test) for the other group of plasma cells where the median XIST RNA count is 3 (**Supplementary Fig. 14b**). As *XIST* RNA is required for X-chromosome inactivation and is only expressed from the inactive X-chromosome[32], this result also verifies that nuclei from the "Nx" DNA clone lost their inactivated X chromosomes instead of not being captured by the hybrid protocol.

Tumor 3 was an endometrial adenocarcinoma, and was the only example in which we did not observe any discernible tumor subclones. Despite this, the tumor projected into two distinct RNA clusters (**Fig. 2c**), which differed most notably in the elevated expression of the estrogen receptor gene *ESR1* in RNA3a (**Supplementary Fig. 14c**). As assessed by immunostaining, about 50% of the tumor cells from the primary tumor expressed the estrogen receptor (**Supplementary Fig. 14c**), which aligned with what we saw from the RNA expression.

Tumor 4 was another uterine carcinosarcoma case. The copy number heatmap showed that the nuclei with copy-number variations were clustered into two DNA clones with significant differences in chromosomes 1, 8, and the X chromosome (**Fig. 2d**). However, for RNA clustering, there was only one major RNA tumor cluster (RNA4a) containing nuclei from both DNA tumor clones. In addition, there were two small RNA clusters, RNA4b and RNA4c, both close to the main cluster RNA4a in RNA space. RNA4c had projections from both tumor DNA clones and had a high expression level of G2-phase marker genes, *MKI67* and *CENPF*. Nuclei in cluster RNA4b were mostly from DNA clone T4A (**Fig. 2d**). Compared to the main tumor RNA cluster RNA4a, RNA4b alone had a high expression level of many actin (*ACTG1*, *ACTB*) and tubulin genes (*TUBA1B*, *TUBA1A*, *TUBB*). RNA4b also highly expressed *EEF2*, an essential factor for protein synthesis, and *GAPDH*, a key enzyme in glycolysis, as shown in **Supplementary Fig. 14d**.

Tumor 5 was a uterine leiomyosarcoma. The biopsy sample from this patient had two sectors, with one being more hemorrhagic than the other. Clustering from each sector was done separately, and overall, the projections were similar. In **Fig. 2e**, we present only the results from Tumor 5-2 (with more nuclei). Tumor 5 had two DNA tumor clones with significant differences in chromosomes 1, 2, 6, 7, 8, 12, and the X chromosome. In RNA clustering, there were two large and one small tumor RNA clusters. The two distinct tumor DNA clones projected mainly to distinct tumor RNA clusters (**Fig. 2e**). Both of the two tumor RNA clusters, RNA5a and RNA5b, had high expression of fibroblast markers, which was concordant with the immunohistochemical analyses showing that the tumor cells were positive for h-caldesmon[33] (**Supplementary Fig. 14e**). RNA5a had higher expression of *TNNT3*, *PLXDC1*, and *MTMR11*, whereas RNA5b had higher expression of *ADAM12*, (involved in skeletal muscle regeneration), *ZFHX4* (related to muscle differentiation), *FN1* (which encodes fibronectin) and collagen genes such as *COL1A1* and *COL6A2*. A small proportion of nuclei from both DNA clones went to the tumor RNA cluster RNA5c, which was cell-cycle related. RNA5c specifically had high expression of typical G2-phase markers *MKI67*, *TOP2A*, *CENPF*, and *CENPE* (**Supplementary Fig. 14e**).

Common stromal and distinct tumor clusters

To obtain a clearer picture of the data, we clustered single-nuclei hybrid data from all patients, separately for RNA and DNA, using the "FindClusters" and "RunUMAP" functions of Seurat (**Fig. 3a** and **Fig. 3b**, respectively). We examined a total of 35,369 nuclei from the sources indicated (**Fig. 3c**). For RNA clustering, we downsampled all nuclei to 400 "Exonic" templates. In addition, and only for the RNA clustering, we created a "garbage" cluster by including 3,500 nuclei from the DNA-only platform treated as hybrid libraries. As expected, the 3,500 nuclei from DNA-only libraries clustered together and were well separated from all other clusters (**Supplementary Fig. 16**). This "garbage" cluster also included 853 low-quality nuclei from the hybrid protocol (**Supplementary Fig. 16**). We removed these nuclei from

further analysis. The aggregated "Exonic" template counts for all the nuclei in each expression cluster are shown in **Supplementary Table 1**.

Tumor-genome clusters were quite distinct from the normal-genome clusters (**Fig. 3b**), and distinct clones within a patient mapped nearby to each other or merged into a single cluster at this resolution of clustering. The intra-tumor clones of patient 2 (T2A, T2B, T2C, and T2D) and patient 5 (T5A and T5B) were still preserved under this resolution. Cluster "N" was distinct from "Nx," the latter being otherwise normal cells with only one copy of the X chromosome.

The projections of nuclei from six sample sources into the combined RNA space (**Fig. 3a**) are shown in **Fig. 3d-i**, where nuclei from a given sample are highlighted either in blue or red, depending on whether they are classified by DNA as tumor or normal genomes. In each panel, the nuclei from other samples are colored in light grey. The projections of the tumor genomes are very distinct between patients, well-separated from each other and the projections of the normal genomes. At this resolution, the intra-tumor RNA expression sub-clusters generally merged together.

By contrast, the normal-genome cell projections from a given patient were quite distinct, and different patients have overlapping normal-genome projections. We labeled these common elements by their distinctive patterns of expression[34] and list the marker genes in **Supplementary Table 2**. The blood components can be further distinguished into finer subtypes, as shown in the zoomed figure of **Fig. 3a**. The counts for these projections of DNA profiles into RNA profiles are shown in **Fig. 4a**. In addition, the hybrid protocol had good consistency between experimental replicates (**Supplementary Fig. 17**).

Exceptions to overlapping normal projections were seen for patient 1, where there was a cluster mainly consisting of normal-genome cells, well-separated from the main stromal clusters. The epithelial-like cluster "EP-T1" from the tumor tissue of patient 1 and the "EP-N1" cluster from an adjacent normal site were distinct from each other or the main epithelial cluster "EP". We believe this distinct “EP-T1” cluster was not due to batch effects, as it had very distinct and plausible gene expression patterns (**Supplementary Fig. 18**), and other stromal-cell projections from this sample mostly overlapped well with other samples.

Loss of the X chromosome in blood elements

The loss of a single X chromosome in some cancer cells (such as in patients 2 and 4)[35], as well as in a small proportion of certain stromal components in cancer patients or the elderly[36], did not surprise us. However, we unexpectedly saw that nearly half of the relatively abundant plasma cells showed losses of the X chromosome in the patient (patient 2) with the worst clinical outcomes (**Supplementary Table 3**). The loss of the X chromosome in somatic lineages was observed in all five endometrial cancer cases. The summary of our data on the projection from normal genomes with and without two copies of the X chromosome is shown in **Table 1**. Another patient (patient 4) with poor clinical outcomes had about 15% loss of the X chromosome in the T cell components, further indicating that this might be a potential biomarker for negative outcomes.

Crossovers and the multinomial wheel

We summarize the projection data for all nuclei from the hybrid protocol (**Fig. 4a**). As represented there and even more visually in the alluvial plots (**Fig. 2**), we saw what we termed "crossovers": nuclei that clustered as tumor or normal genomes, but then clustered in opposition as normal or tumor expression patterns. 357 tumor genomes projected to normal expression clusters, and 401 normal genomes projected to tumor expression clusters. If correct, these crossovers could have profound biological significance (**Discussion**), but it is well-known that in single-cell methods, two cells or nuclei may ‘collide’ and create a merged profile. These collisions could occur at the droplet generation stage, during informatic processing (e.g., barcode collisions), or even biologically by cell

fusion in the host. We developed a filter to eliminate such collisions by making a quantitative measure of the deviation from a cluster.

To this end, we developed the "multinomial wheel." The idea behind this is to create a "multinomial state" from each Seurat cluster. There are as many multinomial states as Seurat clusters, one set of multinomial states for genomes and another set for expression. Each multinomial state is represented as a K-length vector summing to one, where K = 300 bins for DNA data and K = 29,637 genes for RNA data. The value of a multinomial vector is the normalized centroid for the cluster. We then computed the deviation of every nucleus from its Seurat-assigned multinomial state as follows. Between every two multinomial states, we created nine new states that were equally spaced linear combinations of the two multinomial vectors. Thus, for N "home" states, we created $9\times N\times(N-1)/2$ new multinomial states, making up the wheel (see **Fig. 4a** for the genome states, and **Supplementary Fig. 19** for expression states). Every home state has $L = 9\times(N-1)$ other states linked directly to it, each from one to ten 'units' away. For each nucleus, we asked which of the L+1 multinomial states (including its home state) would be the most likely to generate its observed template counts. The distance of that nucleus from its Seurat-assigned multinomial state was the number of units to its closest multinomial state. This was defined as the "distance from home" shown in the histograms in **Fig. 4b** and **Supplementary Fig. 19**.

We display each nucleus as a point on the wheel at its closest genome state (**Fig. 4b**, second panel from left). If the distance of the nucleus exceeds 5, the point is colored in red; otherwise, it is colored blue. We show a histogram of the distances (**Fig. 4b**, leftmost panel) and saw that most nuclei were within one-unit distance away from the Seurat-assigned genome multinomial state.

To test the utility of the multinomial wheel to detect collisions, we utilized the two mixture experiments using nuclei from frozen tumor biopsies where the collisions were verified by patient-specific SNPs. In **Fig. 4b** (second panel from right), we show only the nuclei determined as collisions from the first mixture experiment between patients 1 and 5. The BAGs judged by SNPs to be collisions usually reside in the middle of two multinomial states, not close to either one of them (**Fig. 4b**, second panel from right). We showed a histogram of their distances to Seurat-assigned clusters with a peak distance of 4 (**Fig. 4b**, rightmost panel). Most existing doublets detection methods[37-43] for single-cell RNA sequencing start from the individual count vectors of single cells, and then make artificial doublets by adding/averaging random droplet pairs and use these to train the model. However, unlike the existing methods, the multinomial wheel acknowledges the major clusters determined by Seurat UMAP clustering and then measures the deviation of every cell from the centroids of these major clusters. Different from "DoubletDecon"[44] method which is also based on clustering information but decides if a droplet resembles artificial droplets based on a deconvolution algorithm[45], we assume the multinomial of a cluster could be viewed as a k-sided die, with each throw of the die landing on a face with a fixed probability, with probabilities summing to 1. In our context, each face was either a gene (if an expression multinomial) or a genomic bin (if a genomic multinomial), with its probability determined by its relative frequency in the cluster of cells. Each cell in the cluster could be considered as the outcome of N rolls of its multinomial, where N was the count of templates that were observed. This method works for both DNA and RNA space (**Fig. 4b** and **Supplementary Fig. 19**), not only serving as a doublet detector but also providing quantitative measurements of the cells in between different states.

These experiments justified using the multinomial wheel as a filter and removing the genome 'violators' from all BAGs. The violators were nuclei with a distance of ≥2 from their Seurat-assigned home multinomial states. We plotted all violators from the combined DNA analysis in the respective DNA and RNA multinomial wheels (**Supplementary Fig. 19b**). Importantly, the violators of the genome wheel from the mixing experiment were also violators of the expression wheel (**Supplementary Fig. 19c-d**). This filtration was stringent, as we removed about 20% of the BAGs, in excess of our expectation of 5% collisions. We then re-tabulated the projections of nuclei that passed the filter (**Fig. 4c**). In the table of **Fig. 4c**, we highlighted the remaining crossovers in red. Filtration reduced crossovers from 357 to 24 for tumor genomes with normal expression, and from 401 to 24 for normal genomes with tumor

expression. Thus, removing 20% of the BAGs by filtration eliminated greater than 90% of all crossovers. The alluvial plots for each case before and after violator removal are shown in **Supplementary Fig. 20**.

One case of note in **Fig. 4c** was that two nuclei in the adjacent normal tissue of patient 2 (Normal2) showed tumor genomes, and their respective copy-number profiles are exhibited in **Supplementary Fig. 21b**. Using multinomial wheel analyses, we found both nuclei were within 1 unit distance from the same tumor clone T2B, and in RNA space, both were within 1 unit distance from tumor RNA cluster RNA2a. These two nuclei were not crossovers, but they represented tumor subclone infiltration into the adjacent normal tissue for patient 2.

Patients 2 and 5 showed the largest number of residual crossovers, and they also had the largest numbers prior to filtering. We checked all residual crossovers to make sure the Seurat "FindClusters" assignments agreed with UMAP spatial assignments, and that the individual copy number profiles had good quality. We plotted the crossovers from Tumor 5 with normal genomes and tumor expression on the DNA and RNA multinomial wheel in **Supplementary Fig. 21c**, and the crossovers with tumor genomes and normal expressions from Tumor 2 in **Supplementary Fig. 21d**, which gave a clearer picture. We discuss the biological implication of these remaining crossovers in the next section.

Discussion

We sought to develop a high-throughput method for the assessment of both RNA and DNA from individual cells of a population, and to begin to explore its utility in the description of the cellular composition of primary cancer sites. Our experimental design incorporated four elements. First, we chose BAG single-cell technology because of its flexibility and excellent performance for either RNA-only or DNA-only protocols[30]. Second, we collected both RNA and DNA from individual units at the same time in a hybrid protocol because this was simpler than trying to capture the two nucleic acid types sequentially. Third, we chose nuclei over cells because nuclear RNA was sufficient for the classification of cell types[46-48], and isolating intact cells from frozen or fixed biopsy samples is problematic. Finally, we chose to examine nuclei from one target organ, in this case the uterus, so that we could better assess the commonality of the stroma and the diversity of tumor expression[49]. Although this hybrid method, in its current format, has lower genome coverage per cell than the single-nucleus DNA-only or RNA-only protocol[30], it still shows many advantages over plate-based, low-throughput methods that assess both DNA and RNA together (**Supplementary Table 4**). For example, it enables the detection of mutant stroma existing only in a small proportion of most cell types, which is challenging for low-throughput methods. This new method not only enables the analysis of many more cells, but also overcomes some of the challenges imposed by existing techniques, such as high labor intensity, limitation of sample types, and a preference for entire cells over nuclei alone[19,20,50].

At low resolution, we found that the expression clusters of the cancer cells themselves were quite distinct, well separated from each patient and from the normal clusters. At higher resolution, each cancer had more than one expression cluster. The relationship between these tumor expression clusters and the DNA subclonal populations of the cancers was not consistent from patient to patient. In some cases, cells of distinct tumor subpopulations projected to distinct expression clusters; in some cases, the cells of distinct subpopulations projected to the same expression cluster; and in some cases, cells from the same subpopulation "split": they projected to distinct expression clusters. We believe that these split expression patterns are consistent with epigenetic drift rather than being caused by genetic variations. We take note of two special cases: in patient 3, the same cancer population projected to estrogen receptor positive and receptor negative expression clusters; and in patient 1 with uterine carcinosarcoma, each of two cancer subclones projected to high and low collagen expression clusters. In both patients 1 and 3, the distinct cancer expression types were physically interspersed as determined by histopathology. This was consistent with the idea of epigenetic variation, rather than genetic variation, because we would expect the latter to show physical segregation.

Using this multiomic technology, we observed that a significant number of stroma cells lost one copy of chromosome X. Especially in patient 2, almost half of the plasma cells showed loss of one copy of chrX, suggesting extensive clonality in this lineage. These observations raise additional questions: does somatic clonality indicate failure of the immune checkpoint mechanisms? Do these cells hinder or help the tumor penetrate the host? Do the somatic elements travel with the cancer when it metastasizes?

Although aneuploid and diploid lineages generally projected to distinct expression clusters, we initially observed many exceptions that we termed "crossovers." While it would not be surprising to see early tumor lineages without copy number changes begin to express the tumor pattern, it would be surprising to see the entire program expressed so early, before the selection within the host for the predominant tumor clone. Also, if tumor cells can take on the expression pattern of normal cells, they could possibly escape host surveillance or chemotherapy. Such crossovers could therefore be of immense interest, provided they are not artifacts. We therefore refined our methods to minimize possible artifacts. The most likely artifacts are from ‘collisions,’ BAGs that either report multiple nuclei or with coincidental identifiers. Preparing intentionally mixed samples that were distinguishable by SNVs enabled us to determine that collisions occurred in about 5% of the BAG data. After removing possible collisions, the few remaining crossovers merit further future study with larger data sets. Some of these remaining crossovers might be from an earlier cancer lineage or mutated stroma[51-54].

To better understand "crossovers," we needed a tool for quantifying the similarity of a cell to others in its cluster. The widely used clustering program in Seurat was effective at finding clusters, offering a manifest of marker genes that distinguished the clusters, and providing clear graphical displays. However, the current clustering method failed to detect cells intermediate between clusters and was highly dependent on parameter settings. We, therefore, experimented with a simpler mathematical paradigm for clustering, the multinomial distribution and developed the "multinomial wheel" method to filter the collisions. This multinomial wheel algorithm has extensive utility beyond this specific case or method. It can assist in any clustering analyses to provide quantitative measurement of how each single cell fits into each cluster, which would help identify outliers, collisions, and cells in transition from one state to another in either genomic or transcriptomic space.

In summary, we have developed a high-throughput multiomic method that connects genotypes and expression profiles at single-cell resolution. When tumors have copy number changes, it is now possible to distinguish stromal expression patterns from tumor expression patterns. In exploring five uterine tumors, we uncovered all possible patterns of connection between tumor subclones and expression sub-clusters. We saw differences in the proportionate composition by stromal type, and observed clear evidence of genomic variants in stromal subtypes. How these observations relate to cancer biology in general, or to the classification of cancer subtypes and their relation to disease outcome, await more extensive studies. We expect this pilot study opens a window to the complex relationship between genome and transcriptome, and will lead to new insights into cancer biology, new methods for monitoring cancer progression and evaluating clinical prospects, and possibly new treatments.

Methods

Pulverization of frozen biopsy samples in liquid Nitrogen

All patient tissue biopsy samples were pulverized in Liquid Nitrogen (LN2) with a sterile mortar and pestle prior to analysis. Mortar and pestles were submerged in LN2 and cooled to LN2 temperature. The cooled vessels were then partially filled with fresh LN2 and transferred to a basin containing a shallow pool of LN2. The presence of LN2 in both the mortar and basin helped maintain a constant temperature during the pulverization process and prevent sample heating due to friction. The tissue samples were then transferred to the sterile mortar, submerged in LN2, and pulverized until they were mostly a fine, homogeneous powder. Once pulverized, residual tissue material was scraped off the

pestle back into the mortar with a sterile, LN2-cooled disposable spatula. The mortar was then removed from the basin to allow for the LN2 to evaporate out of the mortar. Subsequently, pulverized tissue was immediately collected with a fresh, sterile, LN2-cooled disposable spatula into 2.0 mL DNA LoBind Eppendorf tubes submerged in LN2. Pulverized samples were placed on dry ice with the caps open to allow for temperature equilibration before closing the tubes, and then stored at -80°C until further use. All samples were pulverized with separate sterile mortar and pestles to avoid cross-contamination between tumor and normal adjacent biopsy tissues.

The sample cohort

We studied samples from five patients (patient 1 – patient 5). The samples were from biopsies of their uterine cancers (Tumor 1 – Tumor 5), and in three patients also from adjacent normal endometrial sites (Normal 1, Normal 2, and Normal 4). For most samples (Normal 1, Tumor 1, Tumor 2, Tumor 3, Normal 4, Tumor 4, Tumor 5), we sequenced single nuclei of the same sample on each of three platforms: DNA-only, RNA-only, and the hybrid protocol. We performed comparison analyses and showed the validity of the hybrid protocol mainly using the above five trio data sets.

Hybrid BAG generation

We dissolved the pulverized tissue in ice-cold NST detergent buffer[47] and stained with DAPI. We performed single-nuclei sorting using DAPI-H vs. DAPI-A single-nuclei gate on a FACSAria II SORP cell sorter to remove debris and clumps. We confirmed (data not shown) that single-nuclei sorting based on ploidy would not be able to distinguish cancer cells from normal cells because the hypodiploid peak of cancer cells often overlaps with the diploid peak of normal cells[47]. Single nuclei were loaded into the microfluidic device described in detail in a previous publication[30]. Nuclei were encapsulated into droplets with an average diameter of 120 microns. For the capture of nucleic acids, we used 5' Acrydite oligonucleotides. All the Acrydite-modified oligonucleotides became covalently co-polymerized into the gel ball matrix. They also all contained, at their 5' end, a universal PCR primer (UP1) for subsequent amplification. For RNA-only protocol, we used oligo-dT; for DNA-only protocol, we used random T/G primers, and followed their respective published protocols[30]. To capture both RNA and DNA together in the new hybrid protocol, we used both Acrydite primer designs, but we altered the protocol in two important ways.

The first critical change was an incubation step at 85°C for 5 minutes instead of 95°C for 12 minutes for DNA denaturation in the DNA-only protocol. Otherwise, we observed significant destruction of the RNA.

The second critical change took place after the BAGs were formed. The RNA and genomic DNA trapped in the BAGs were used as templates to make covalently bound copies, and in the new hybrid protocol, both reverse transcriptase and DNA polymerase were used. Template-switch-oligos were also introduced in the hybrid protocol so that the cDNA products which were covalently linked to the BAG matrix ended with a double-stranded region. This double-stranded DNA region included an NLA-III cleavage site. Subsequently, DNA polymerase (Klenow) was added to extend the captured genomic DNA from primers, forming a copy that was also covalently linked to the BAGs. Some, perhaps most, of the cDNA-mRNA sequence was further partially converted to double-stranded cDNA. BAGs were pooled and the covalently captured DNA and cDNA were cleaved with NLAIII leaving a sticky end used for subsequent extensions.

BAG barcodes and varietal tags were added to the 3' ends of the covalently captured nucleic acids in split-and-pool reactions. The BAG barcodes were present on both the genomic-DNA and RNA copies. The varietal tags were used for counting. The first BAG barcode and varietal tag were added by ligation extension (described in detail in the supplementary experimental protocol), leaving a common 3' sequence identical across all the molecules and BAGs. The second BAG barcode and varietal tag were

added by hybridization extension of the common 3' sequence, along with a second common sequence adapter for the third split-and-pool step. The third barcode was added by a split PCR, using the first universal PCR primer (UP1) and the second common sequence adapter as part of the PCR primer sequences.

These amplified products were pooled and converted by tagmentation into paired-end Illumina sequencing libraries. One end of the reads contained BAG barcode and varietal tag, as well as genomic or transcriptomic sequence information. The other end from random tagmentation was mostly genomic or transcriptomic sequence information.

Initial data processing

Sequencing libraries were sequenced in paired-end 150 bp format using an Illumina NovaSeq 6000. Briefly, each processing step is described in more detail in the immediately following sections. We first checked the structure of each read pair in the fastq files. For the good read pairs with the correct structure as shown in **Fig. 1a**, we extracted the BAG barcode, varietal tag, and genomic sequences from both reads. We then mapped the genomic (including transcriptomic) sequence to the reference genome with gene transcript information. Finally, we combined the mapping information from all reads belonging to each varietal tag for each BAG barcode. In the end, we obtained a template data table with each row containing the information of an original template/molecule. In the following section, we explain each processing step from the fastq file to the template table in detail.

Step 1 – Check sequence structure

First, we filtered out reads from the fastq files where either Read 1 or Read 2 were less than 100 bases. Second, we examined if the sequences from the expected BAG barcode positions exactly matched one of the 96×96×96 barcodes, and if the "CATG" cutting site was in the expected location, allowing for one base mismatch. We removed read pairs that did not satisfy these requirements. Third, from Read 1 which started with barcodes and varietal tags, we trimmed away the first 80 bases containing the BAG barcode, varietal tag, and adapter sequences, and also checked if the reverse complementary sequence of the universal primer ("CCAAACACACCCAA") or oligo-dT ("AAAAAAAAAAAAAA") was present. If present, it meant we had reached the end of the template, so these primer-related sequences were trimmed off for downstream mapping. Similarly, for Read 2, the tagmentation end, we checked and removed the adapter sequence ("GAGCGGACTCTGCG") from the first split-and-pool if it existed. After trimming, we required both Read 1 and Read 2 to be at least 30 bases long. All the bases from Read 1 and Read 2 after trimming were then used for paired-end mapping (Step 3).

Step 2 – Extract BAG barcode and varietal tags

If a read pair passed Step 1, we extracted the BAG barcode and varietal tag information from the first 80 bases of Read 1, and this information was appended to the read ID. The 17 base BAG barcodes came from three cycles of the split-and-pool procedure, of which five bases came from the 1st-split, six bases came from the 2nd-split, and six bases came from the 3rd-split. There were 96 different barcodes for each split, so there were altogether 96×96×96 (≈ 1 million) varieties. The 12 base varietal tag came from both the split-and-pool primers and the genomic sequence. Out of these twelve bases, four bases came from the 1st-split, four bases came from the 2nd-split, and four bases came from the genomic sequence that was two bases away from the "CATG" cutting site. These twelve bases provide 4^{12} (≈ 16 million) varieties for each BAG.

Step 3 – Map to the human genome

After steps 1 and 2 above read pairs were mapped to the UCSC hg19 human genome using HISAT2 version 2.1.0[31]. The reference genome we used included the primary chromosomes and

unlocalized and unplaced contigs. Alternate haplotypes were not included in the genome index. HISAT2 can take a file with known splice sites to use for alignment. This file was generated using a gtf formatted file extracted from the NCBI refSeq gene annotation table from the UCSC genome browser and the HISAT2 program, hisat2_extract_splice_sites.py. The bam files were then sorted and indexes using samtools. In subsequent data analysis steps we designate by mapped reads the reads that HISAT2 marks as being part of a proper pair and a primary mapping having a read mapping quality score greater than zero.

Step 4 – Combine read information with original template information

We grouped the mapped reads based on their BAG barcodes. For the reads with the same BAG barcode, we sorted the varietal tags by the number of reads associated with each tag in descending order. We performed a "rollup" algorithm on the sorted varietal tags, and discarded varietal tags within a Hamming distance of one from a more abundant varietal tag having at least ten times more reads. We assumed the eliminated varietal tags originated from the tags with more abundant reads but contained sequencing or PCR errors.

Using the varietal tags from the above "rollup" step, we aggregated the mapped segments for all the reads with the same varietal tag. We checked the total coverage of each varietal tag against all exons and transcript boundaries from the NCBI refSeq gene annotation file downloaded from the UCSC genome browser, and wrote out one line per varietal tag with all the useful information into a "template table". Each line of the template table contains the following information: BAG barcode, varietal tag, chromosome, start mapping position, end mapping position, start and end mapping position for each fragment if there was more than one continuous fragments, total bases covered by this template, number of reads, number of genes, gene list, bases overlapping with the transcript of the best-matched genes, bases overlapping with exons, number of splice junctions, number of unspliced sites, bases overlapping with the coding regions, 5'UTR, and 3'UTR of the gene. The downstream data analyses were mainly based on the information from this table. The best-mapped gene was deemed to be the gene from the annotated transcript file having the highest overlap to the transcript. If more than one transcript had the same overlap then best was determined by overlap to exons, then overlap to coding sequence, then the number of splice junctions, then the fewest unspliced sites. If more than one gene tied for all these criteria, then all genes are listed in the template table.

Template processing

Sequence classification

Starting from the template data table described above, each initial molecule was classified as one of the four categories: "Exonic"," Intronic", "Intergenic", and "Uncategorized". This process was applied uniformly regardless of the protocol types (RNA-only, DNA-only, or hybrid). We classified a template as an "Exonic" template if over 90% of its bases were mapped within one gene. Furthermore, we refined the "Exonic" classification only if 50% or more covered bases from this template were exonic, or if 20% or more covered bases were exonic and at least one splicing event was observed (RNA layer). If all the bases from a template were mapped to intergenic regions, we classified it as "Intergenic". If a template was not classified as "Intergenic", but less than 10% of its covered bases were exonic and no splicing events were observed, this template was classified as "Intronic". Only a small proportion of templates failed to be classified into the above three categories, and these templates were classified as "Uncategorized".

For expression clustering, we only used "Exonic" templates assigned to a single gene regardless of protocols. For copy number clustering, we tested four versions of template choices on all the libraries, and presented the comparative results on the two normal tissue samples (Normal 1 and Normal 4) in **Supplementary Fig. 3**, which we will discuss in the next section.

Copy number plot varieties

We demonstrated four progressively improved versions of copy number estimation, named "all_molecules", "no_exon", "no_gene", and "no_gene.avoid50closeTN". The method "all_molecules" simply used all molecules for each retained nucleus for copy number as the name would imply. The method "no_exon" used molecules both classified as "Intronic" and "Intergenic" in the previous paragraph. The method "no_gene" only used "Intergenic" templates with no bases covering a transcript.

The method "no_gene.avoid50closeTN" (DNA layer) only retained the "Intergenic" molecules from the "no_gene" method that were at least 50 bases distant from RNA hotspots. We defined an RNA hotspot as the genomic region between two "Intergenic" templates that were within 50 bases of each other in RNA-only libraries from all tumor and normal samples in the cohort. RNA hotspots were expected to be some combination of actual unannotated transcripts and regions of DNA that were prone to being copied by reverse transcriptase. As these hotspot sequences distorted copy number profiles in normal and tumor biopsy specimens, we eliminated certain intergenic regions when determining copy number profiles for the hybrid protocol.

Empirical bin boundary generation for copy number

Separately for each of the four copy number molecule selection methods above, we used the genomic positions of all molecules from normal DNA samples to determine empirical bin boundaries for 300 bins with approximately equal molecule counts per bin. Excluding any molecules mapping to chromosome Y, we assigned to each chromosome 1-22 and X a number of bins in proportion to its fraction of total molecule counts. Within each chromosome, bin boundaries were assigned greedily from the start of the chromosome so that all but the final bin contained at least the same number of molecules that was equal to the total counts of molecules (or referred to as templates) divided by the number of bins for that chromosome. The observed count of molecules per bin was recorded as a normalization factor for later use during per-sample copy number estimation. This normalization factor could vary by up to 30% between chromosomes because a small number of bins (300) can only be imperfectly allocated by chromosomal molecule counts.

Copy number estimation

For each copy number variant separately, each selected molecule incremented a bin based on the established bin boundaries for that method. Each bin count was then divided by the per-bin normalization factor, and the result was multiplied by 2 divided by the median value over all bins. Assuming a mostly diploid sample, this process resulted in a copy number profile for the sample that was centered at a value of 2. Circular binary segmentation (DNAcopy version 1.50.1)[55] in R was then performed on the copy number profile using parameters alpha=0.02, nperm=1000, undo.SD=0.5 and min.width=2. For each profile, we also computed a quantity we call 'terrain' which was the sum of the absolute value of adjacent bin copy number differences. To produce copy number input for the "CreateSeuratObject" function of Seurat, the per-bin normalization factor was applied to each raw bin count for each cell, and a second per-cell normalization factor was then used so that each cell's total normalized count was set equal to its total unnormalized count.

RNA clustering

The RNA clustering was performed using Seurat package (version 3.1.5), and using the standard Seurat clustering pipeline[56]. The gene names were also appended with the chromosome information to distinguish any ambiguous locations. We removed the ribosomal protein genes for clustering. For comparing expression clustering between the hybrid protocol and RNA-alone protocol, we normalized the gene-template matrix by cell, and removed the PCA components that most significantly

distinguished protocol differences. We normally used at least 15 PCA components for clustering. This approach gave us similar clustering results as the "IntegrateData" function in Seurat v4. For the combined RNA clustering of all the hybrid data, we downsampled the gene matrix to 400 Exonic templates per nucleus, and included nuclei with more than 300 Exonic templates for clustering. In the clustering process, we only used genes that showed up in at least 30 nuclei, and nuclei with at least 150 genes; we used the top 5,000 variable gene features for PCA analysis and used the first 50 PCA components for subsequent UMAP and FindCluster functions.

Copy number clustering

Similar to RNA clustering, we used "RunUMAP" and "FindClusters" functions of Seurat to cluster nuclei based on copy number. For each library, we had a bin-counts matrix, similar to the gene matrix for RNA clustering. There were 300 rows in the matrix, representing 300 genomic bins. Each column represented a nucleus. Each element of the 2D matrix represented the tag counts of the corresponding bin in the corresponding nucleus. We first normalized the matrix by columns: for each nucleus, we divided each bin count by the mean of 300 bins and then multiplied by 2. We not only used these 300 normalized single bin counts for clustering; additionally, we also included the median normalized bin counts of every two and three adjacent bins, as long as these adjacent bins were within the same chromosome. The reason for this step was that copy number segmentation usually requires similar amplification or deletion patterns in at least two contiguous bins. By doing this, we appended another 277 rows from the two adjacent bins and 254 rows from the three adjacent bins onto the original 300-row normalized bin-count matrix.

We performed clustering using the new matrix with 831 rows. We used a workflow similar to that for RNA clustering, but we did not use "NormalizedData" function since the matrix had already been normalized. For “FindVariableFeatures” function, we used the top 500 features by inputting “selection.method = “vst”, nfeatures = 500”.

Copy number heatmap

The single-nucleus copy-number heatmap was plotted using Seurat "DoHeatmap" function. Each row represented the median normalized counts of two adjacent bins, except for the first bin of each chromosome, in which we used the normalized count of that single bin. The total of 300 rows were sorted in genomic order, with chromosome Y eliminated.

Multinomial Wheel

To build a multinomial wheel in DNA space, we first computed a multinomial vector to represent each Seurat cluster. Each multinomial vector had 300 elements, representing 300 genomic bins. Each element was the total bin counts from all the nuclei in that cluster. We normalized each vector to sum to one, serving as the multinomial probability vector representing that cluster. Next, we computed the linear combination of multinomial probability vectors of every two Seurat clusters, and created 9 equally spaced sampling states $C_{1,2,\ldots,9} = pA + (1-p)B$, for p = (0.1, 0.2,…, 0.9), where A and B are the two original states. We then assigned the nucleus to the state with the highest likelihood. In R language, we used the "dmultinom" function to compute multinomial probabilities.

We applied a similar idea to create the RNA multinomial wheel. Different from the DNA multinomial vector where each element was a genomic bin, in RNA space, each element represented one of the 29,637 genes. We computed the sum of gene counts for each Seurat cluster $V_{1,2,\ldots,n}$ (n is the number of Seurat clusters, and V_i is a 29,637-element vector, $i = 1,2,\ldots,\mathrm{n}$), but unlike DNA, there were many elements still being zero which could not be used as a multinomial probability vector. We solved the problem by adding a small value to each element that was proportional to the total expression level of every gene, so that each vector V_i^* does not contain zero elements. For each gene element j, we did

the following transformation: $V_i^*[j] = V_i[j] + \left(0.05 \times (\sum_j V_i[j])\right) \times (\sum_i V_i[j]) \div (\sum_{i,j} V_i[j])$. We then normalized each vector V_i^* to obtain the multinomial probability vector for cluster i.

Supplementary Materials

Supplementary Fig. 1 to Supplementary Fig. 21
Supplementary Table 1 to Supplementary Table 4
Supplementary Protocol

Acknowledgments

We thank P. Moody for cell sorting, Q. Gao for histology assistance, A. Runnels and E. Ghiban for Illumina sequencing assistance, C. Danyko, A. Stepansky, and D. Stauder for technical assistance, D. Tuveson for helpful discussion, A. Kapedani, M. A. Green, and S. Chin from the Clinical Research Team in the Department of Obstetrics and Gynecology of Long Island Jewish Medical Center for clinical information assistance. Additionally, we extend our gratitude to S. S. Fox, M. Heywood, K. Quinn, B. M. Weil, J. Jacob from Biobanking/Anatomic Pathology Team in the Northwell Health Biospecimen Repository (NHBR) in Northwell Health Cancer Institute for sample transfer and pathological information assistance. This work was supported by a grant from the Simons Foundation, Life Sciences Founders Directed Giving-Research (award number 519054 to M.W.); The Breast Cancer Research Foundation (BCRF, to M.W.); and by support from the Cold Spring Harbor Laboratory and Northwell Health Affiliation (awarded to M.W.).

Author contributions

S.L. and M.W. conceived the idea and designed the study; S.L., J.K., P.A., and M.W. developed bioinformatic analysis programs; S.L. and M.W. developed the experimental protocol; S.L., J.A., E.R, H.O., S.P., and L.S performed single-cell BAG-seq experiments; J.A. and C.P. assisted in sample transfer and sample preparation; A.R. performed histology analyses; G.L.G provided samples and clinical guidance; S.L., J.K., P.A., R.M., N.R., M.R., D.L.D., D.L., and M.W. performed informatics analyses and results interpretation; S.L., D.L., and M.W. wrote the manuscript with input from all coauthors.

Declaration of interests

The authors declare that they have no competing interests.

Data availability

Illumina sequencing data for all the single-nucleus libraries are available at NCBI Sequencing Read Archive (SRA) with accession code (PRJNA773107).

Code availability

Code is available through https://github.com/siranli01/DNA_RNA.

References

1. Lan, F., Demaree, B., Ahmed, N. & Abate, A.R. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nature biotechnology* **35**, 640-646 (2017).
2. Vitak, S.A. *et al.* Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature methods* **14**, 302-308 (2017).

3. Andor, N. *et al.* Joint single cell DNA-Seq and RNA-Seq of cancer reveals subclonal signatures of genomic instability and gene expression. *Biorxiv*, 445932 (2018).
4. Macosko, E.Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214 (2015).
5. Klein, A.M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).
6. Gierahn, T.M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature methods* **14**, 395-398 (2017).
7. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661-667 (2017).
8. Rosenberg, A.B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176-182 (2018).
9. Habib, N. *et al.* Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925-928 (2016).
10. Gao, R. *et al.* Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. *Nature communications* **8**, 1-12 (2017).
11. Zheng, G.X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 1-12 (2017).
12. Gao, R. *et al.* Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nature biotechnology* **39**, 599-608 (2021).
13. Fan, J. *et al.* Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome research* **28**, 1217-1227 (2018).
14. Elyanow, R., Zeira, R., Land, M. & Raphael, B.J. STARCH: Copy number and clone inference from spatial transcriptomics data. *Physical Biology* **18**, 035001 (2021).
15. Patel, A.P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).
16. Erickson, A. *et al.* Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature* **608**, 360-367 (2022).
17. Dey, S.S., Kester, L., Spanjaard, B., Bienko, M. & Van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nature biotechnology* **33**, 285-289 (2015).
18. Macaulay, I.C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature methods* **12**, 519-522 (2015).
19. Han, K.Y. *et al.* SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome research* **28**, 75-87 (2018).
20. Hou, Y. *et al.* Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell research* **26**, 304-319 (2016).
21. Bian, S. *et al.* Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science* **362**, 1060-1063 (2018).
22. Han, L. *et al.* Co-detection and sequencing of genes and transcripts from the same single cells facilitated by a microfluidics platform. *Scientific reports* **4**, 1-9 (2014).
23. Van Strijp, D. *et al.* Complete sequence-based pathway analysis by differential on-chip DNA and RNA extraction from a single cell. *Scientific reports* **7**, 1-9 (2017).
24. Kong, S.L. *et al.* Concurrent single-cell RNA and targeted DNA sequencing on an automated platform for comeasurement of genomic and transcriptomic signatures. *Clinical chemistry* **65**, 272-281 (2019).
25. Cheow, L.F. *et al.* Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nature methods* **13**, 833-836 (2016).
26. Rodriguez-Meira, A. *et al.* Unravelling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel RNA sequencing. *Molecular cell* **73**, 1292-1305. e8 (2019).

27. Li, W., Calder, R.B., Mar, J.C. & Vijg, J. Single-cell transcriptogenomics reveals transcriptional exclusion of ENU-mutated alleles. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **772**, 55-62 (2015).
28. Yu, L. *et al.* scONE-seq: A single-cell multi-omics method enables simultaneous dissection of phenotype and genotype heterogeneity from frozen tumors. *Science Advances* **9**, eabp8901 (2023).
29. Yin, Y. *et al.* High-throughput single-cell sequencing with linear amplification. *Molecular cell* **76**, 676-690. e10 (2019).
30. Li, S. *et al.* Copolymerization of single-cell nucleic acids into balls of acrylamide gel. *Genome research* **30**, 49-61 (2020).
31. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**, 907-915 (2019).
32. Panning, B., Dausman, J. & Jaenisch, R. X chromosome inactivation is mediated by Xist RNA stabilization. *Cell* **90**, 907-916 (1997).
33. Watanabe, K., Kusakabe, T., Hoshi, N., Saito, A. & Suzuki, T. h-Caldesmon in leiomyosarcoma and tumors with smooth muscle cell-like differentiation: its specific expression in the smooth muscle cell tumor. *Human pathology* **30**, 392-396 (1999).
34. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).
35. Spatz, A., Borg, C. & Feunteun, J. X-chromosome genetics and human cancer. *Nature Reviews Cancer* **4**, 617-629 (2004).
36. Zhou, Y. *et al.* Single-cell multiomics sequencing reveals prevalent genomic alterations in tumor stromal cells of human colorectal cancer. *Cancer cell* **38**, 818-828. e5 (2020).
37. Xi, N.M. & Li, J.J. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. *Cell systems* **12**, 176-194. e6 (2021).
38. Wolock, S.L., Lopez, R. & Klein, A.M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems* **8**, 281-291. e9 (2019).
39. Lun, A.T., McCarthy, D.J. & Marioni, J.C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5**(2016).
40. Bais, A.S. & Kostka, D. scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics* **36**, 1150-1158 (2020).
41. Gayoso, A., Shor, J., Carr, A.J., Sharma, R. & Pe'er, D. DoubletDetection (Version v2. 4). *Zenodo, DOI* **10**(2018).
42. McGinnis, C.S., Murrow, L.M. & Gartner, Z.J. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell systems* **8**, 329-337. e4 (2019).
43. Bernstein, N.J. *et al.* Solo: doublet identification in single-cell RNA-Seq via semi-supervised deep learning. *Cell systems* **11**, 95-101. e5 (2020).
44. DePasquale, E.A. *et al.* DoubletDecon: deconvoluting doublets from single-cell RNA-sequencing data. *Cell reports* **29**, 1718-1727. e8 (2019).
45. Gong, T. & Szustakowski, J.D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29**, 1083-1085 (2013).
46. Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nature medicine* **26**, 792-802 (2020).
47. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90 (2011).
48. Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature biotechnology* **38**, 737-746 (2020).
49. Puram, S.V. *et al.* Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611-1624. e24 (2017).

50. Zachariadis, V., Cheng, H., Andrews, N. & Enge, M. A highly scalable method for joint whole-genome sequencing and gene-expression profiling of single cells. *Molecular Cell* **80**, 541-553. e5 (2020).
51. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome research* **20**, 68-80 (2010).
52. Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312-317 (2019).
53. Yizhak, K. *et al.* RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364**, eaaw0726 (2019).
54. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532-537 (2019).
55. Olshen, A.B., Venkatraman, E., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572 (2004).
56. Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902. e21 (2019).

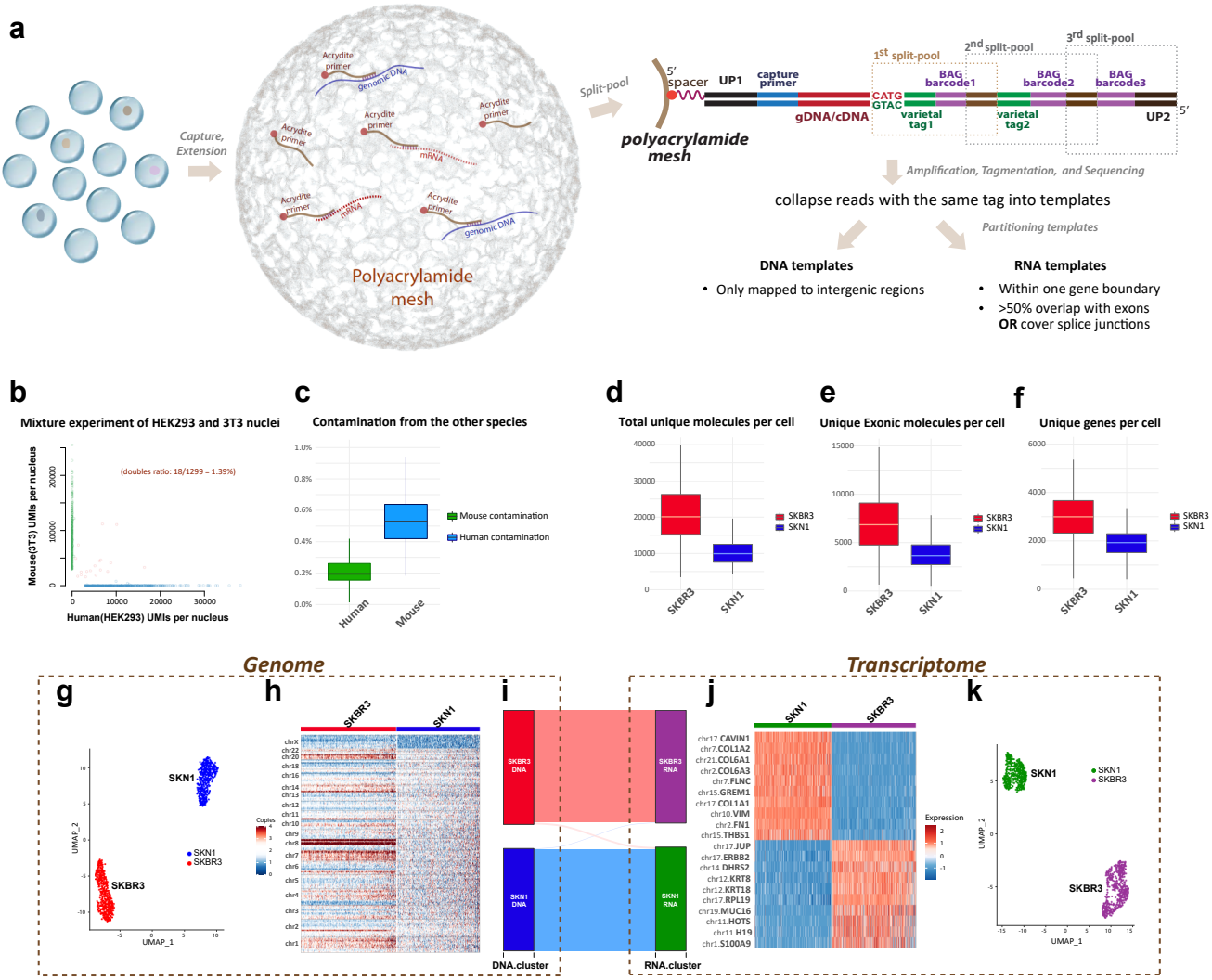


Fig. 1. Overview and basic performance of the single-nucleus/cell hybrid sequencing protocol to simultaneously capture and analyze the genome and transcriptome. a, Workflow showing the major steps of the single-nucleus/cell hybrid protocol. **b-c**, Doublets ratio (1.39% out of 1299 nuclei) (b) and cross-contamination level (0.2% mouse templates in human nuclei and 0.5% human templates in mouse nuclei) (c) from the other species of a HEK293-3T3 nuclei mixture experiment. **d-k**, Performance and genome-transcriptome correlation from a SKN1-SBKR3 mixture single-cell hybrid sequencing experiment; total unique molecules (d), exonic templates (e), and gene counts per cell (f); clustering based on DNA copy number (g) and copy number heatmap (h); clustering based on gene-count matrix (k) and heatmap of marker genes (j), and the correlation between two genomic clusters and two expression clusters (i).

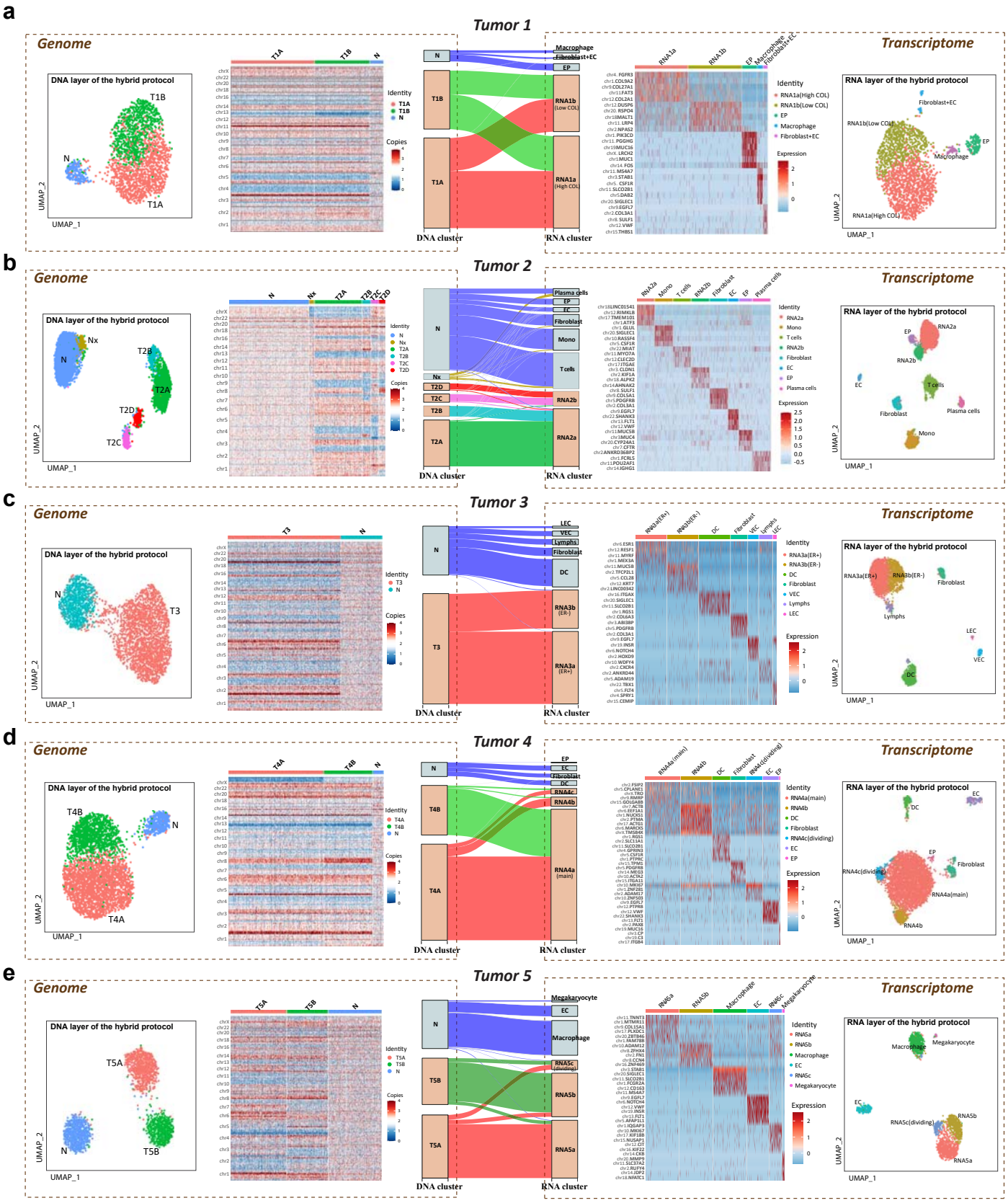


Fig. 2. Alluvial diagrams showing the genome-transcriptome correlations of the five tumor samples.
a-e, For each one of the five tumor samples Tumor1-5, we show the genomic clustering (leftmost panel) and copy number heatmap (second panel from the left), expression clustering (rightmost panel) and marker gene heatmap (second panel from the right), and the alluvial plots connecting the genome clones with RNA expression clusters (middle panel).

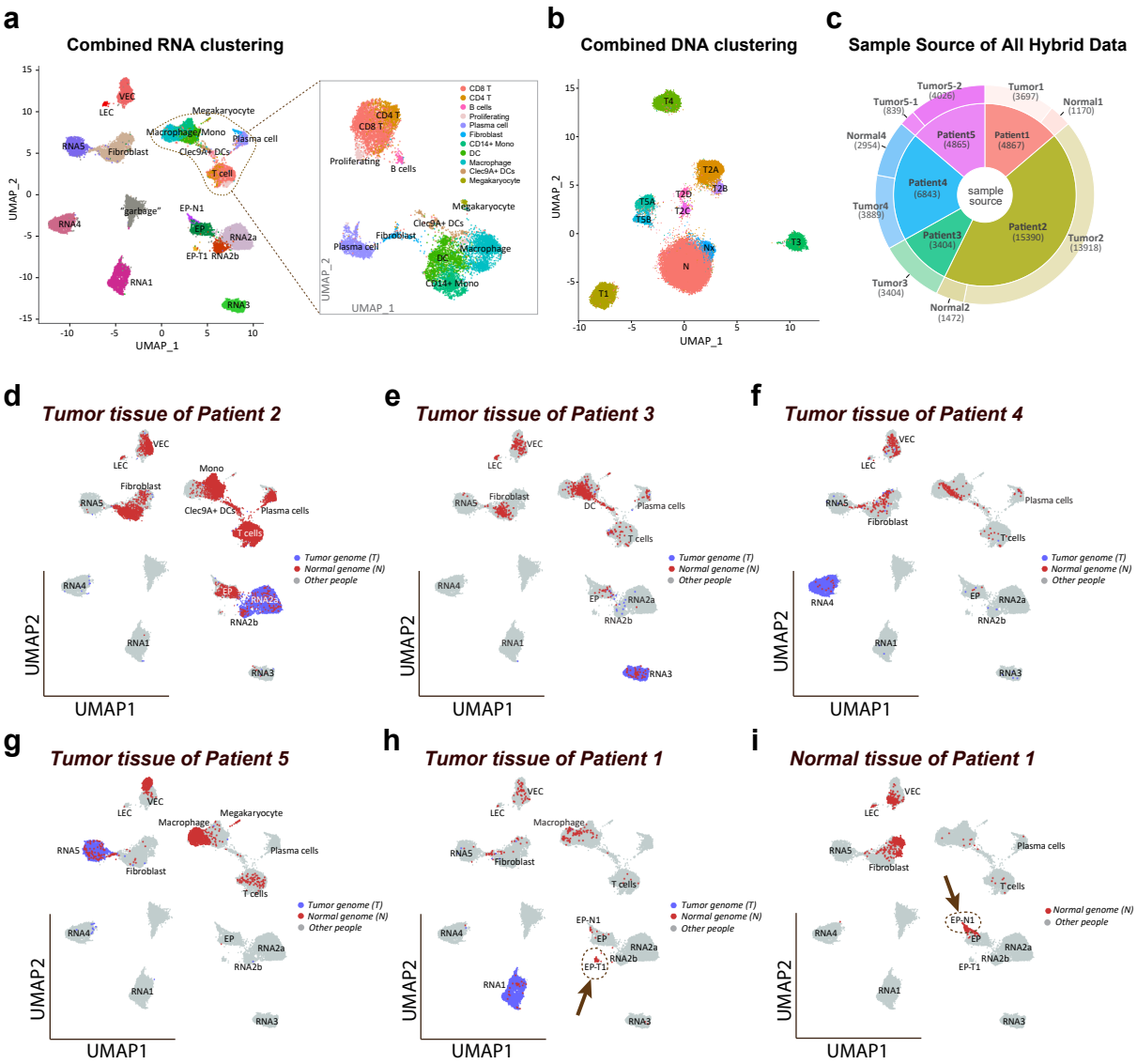


Fig. 3. RNA and DNA clustering using a combination of all nuclei from the hybrid protocol, distinguishing largely common stromal and distinct tumor clusters.

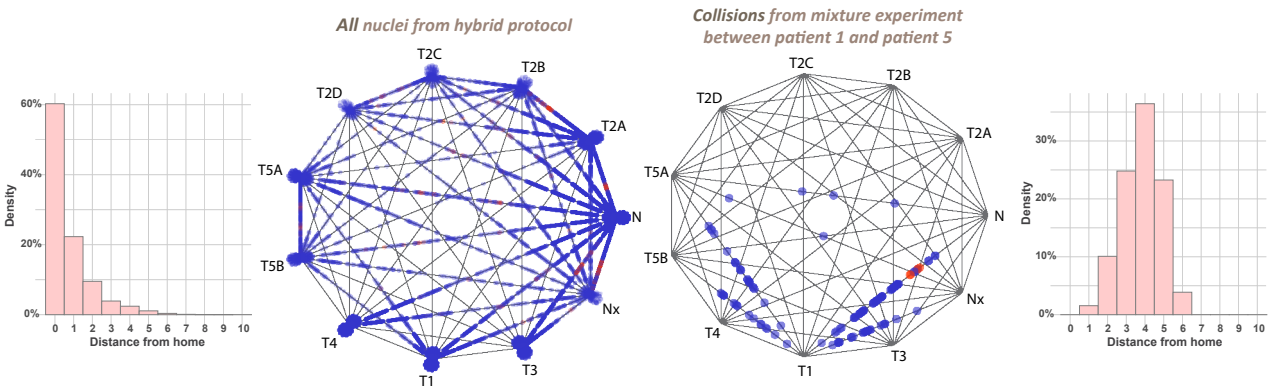
a, Combined RNA clustering using all nuclei from the hybrid protocols and 3,500 nuclei from seven DNA-only libraries. A zoomed plot in the same figure, circled by dash line, presents the subtypes of white blood cells when clustered separately. **b**, Combined DNA clustering after removing nuclei clustered to the "garbage" RNA state in (a). **c**, The sample sources of nuclei from the hybrid protocol in the combined analyses. **d-i**, The projections of tumor-genome (blue) and normal-genome (red) nuclei into the RNA UMAP space for six biopsy samples. The tumor-genome or normal-genome information is determined by the combined DNA analysis in (b); **d**, Tumor2; **e**, Tumor3; **f**, Tumor4; **g**, Tumor5; **h**, Tumor1; **i**, Normal1; in **h-i**, Unique stromal components are circled in dashed lines and indicated by arrows.

a

	RNA1	RNA2 (RNA2a+RNA2b)	RNA3	RNA4	RNA5	Macrophage/Mono	T cells	Clec9A+ DCs	Plasma cells	Fibroblast	EC (VEC+LEC)	EP	EP-N1	EP-T1	Megakaryocyte
Normal1_N		1		1		8	7	1	1	560	240	13	306		
Normal1_T															
Tumor1_N	25	3	3	2	8	75	6	1		23	46	8	2	118	
Tumor1_T	3358			1	3	1				2				3	
Normal2_N	1	2	2	1		32	38	4	2	625	173	572			
Normal2_T		3				2				9	1	2			
Tumor2_N	2	163	10	3	15	1929	3474	143	638	1072	575	606	1	2	2
Tumor2_T	1	5017	7	10	1	40	81	4	19	38	16	24	1		
Tumor3_N		3	41		4	483	42	46	19	151	130	16			7
Tumor3_T		8	2415			12	5	1	1			2			
Normal4_N		3		1	5	136	221	62	656	1140	484	189	5		
Normal4_T															
Tumor4_N		1		13	5	104	16	6	10	109	118	3			
Tumor4_T	1	3	4	3437	1	2			2	13	8	2			
Tumor5_N					83	1609	89	2		24	441	1			90
Tumor5_T		2		15	2393	27	1	1		17	19				1

Tumori_N means nuclei from the tumor tissue of patient i (Tumori) with normal genome (_N).
Normalj_T means nuclei from the adjacent normal tissue of patient j (Normalj) with tumor genome (_T).

b



c

	RNA1	RNA2 (RNA2a+RNA2b)	RNA3	RNA4	RNA5	Macrophage/Mono	T cells	Clec9A+ DCs	Plasma cells	Fibroblast	EC (VEC+LEC)	EP	EP-N1	EP-T1	Megakaryocyte
Normal1_N				1		7	8	1	1	489	226	10	270		
Normal1_T															
Tumor1_N						61	5	1		14	33	5	2	88	
Tumor1_T	3009									2					
Normal2_N			1			30	31	4	2	497	142	494			
Normal2_T		2													
Tumor2_N		9	2		1	1531	2745	115	486	825	456	495	1		2
Tumor2_T		4399		2		2	6	1		2	3	4			
Tumor3_N			2		1	448	39	45	17	137	120	7			7
Tumor3_T			1882												
Normal4_N				1		120	207	57	612	1047	444	153	4		
Normal4_T															
Tumor4_N						84	10	4	5	83	104	2			
Tumor4_T	1		1	3053						1					
Tumor5_N					6	1333	76	2		16	328	1			63
Tumor5_T				3	2096					3					

Fig. 4. Multinomial wheel analyses quantifying the deviation of each cell to the major clusters and removing most of the cross-overs.

a, The projection of normal-genome (_N) or tumor-genome (_T) nuclei from each biopsy sample into RNA clusters based on clustering results in panels (a) and (b) of Fig. 3. **b**, We show the most-likely position for each nucleus on the DNA multinomial wheel based on the multinomial analysis (second panel from the left), and display in histogram the distance between the multinomial-wheel assignment and the Seurat assignment for all nuclei (first panel from the left). In addition, in the second panel from the right, we only show the nuclei with mixed identities (collisions) from the first mixture experiment of Tumor 1 and Tumor 5 on the same DNA multinomial wheel, and display in histogram the distance between the multinomial-wheel assignment and the Seurat assignment for all collisions (first panel from the right). **c**, The projection table in (a) after removing the suspected collisions which are between 2 and 10 units away from the major clusters .

The projection of normal-genome nuclei with 2 or 1 copy of chrX onto stroma cell types

	Macrophage /Mono	T cells	Clec9A+ DC	Plasma BC	Fibroblast	VEC	LEC	EP	EP-N1	EP-T1	Megakaryocyte
Normal1_N_2chrX	5	4	1	1	241	109	7	5	182		
Normal1_N_1chrX					**17** (6.6%)	**4**	**1**				
Tumor1_N_2chrX	39	2	1		6	17	2	1	1	40	
Tumor1_N_1chrX										1	
Normal2_N_2chrX	23	21	2	2	385	103	10	392			
Normal2_N_1chrX	**1**	**2**			**5**	**1**		**6**			
Tumor2_N_2chrX	1119	2130	88	221	613	319	34	406	1		2
Tumor2_N_1chrX	**23** (2.0%)	**104** (4.7%)	**4**	**149 (40.3%)**	**16** (2.5%)	**4**		**2**			
Tumor3_N_2chrX	212	29	31	10	76	35	17	3			2
Tumor3_N_1chrX	**2**		**1**	**1**	**1**						
Normal4_N_2chrX	61	112	36	286	508	222	16	90	2		
Normal4_N_1chrX		**20 (15.2%)**		**15** (5.0%)	**13** (2.5%)	**3**	**1**	**1**			
Tumor4_N_2chrX	41	4	3	3	36	50	2	2			
Tumor4_N_1chrX	**1**	**1**			**3**	**1**					
Tumor5_N_2chrX	897	56	2		12	209	2	1			40
Tumor5_N_1chrX	**11** (1.2%)	**2**				**1**					

Normali_N_2chrX means the adjacent normal tissue from patient i (Normali) with normal genome (_N) with both copies of chrX (_2chrX).
Tumorj_N_1chrX means the tumor tissue from patient j (Tumorj) with otherwise normal genome (_N) but only one copy of chrX (_1chrX).
Nuclei with tumor DNA genome are not shown in this table.

Table 1. The projection of normal-genome cells with two or one copy of chrX into stromal expression clusters.