Article:
Li, S., Alexander, J., Kendall, J. et al. (19 more authors) (2025) Hybrid BAG-seq: DNA and RNA from the same single nucleus reveals interactions between genomic and transcriptomic landscapes in human tumor samples. Genome Biology, 26 (1). 314. ISSN: 1474-7596

https://doi.org/10.1186/s13059-025-03790-5

Genome Biology

METHODOLOGY

Open Access

Hybrid BAG-seq: DNA and RNA from the same single nucleus reveals interactions between genomic and transcriptomic landscapes in human tumor samples

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], Nissim Ranade[1], Patrick Morris[1], Danielle Stauder[1], Daniel Bradford[1], Zachary Laster[1], Michael Ronemus[1], Arvind Rishi[2], Marina Frimer[3], Rong Ma[4,5,6], David L. Donoho[7], Gary L. Goldberg[1,3], Michael Wigler[1] and Dan Levy[1*]

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

[1] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[2] Department of Pathology and Laboratory Medicine, Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, USA
[3] Department of Obstetrics and Gynecology, Division of Gynecologic Oncology, Zucker School of Medicine at Hofstra/Northwell, New Hyde Park, Northwell Health, NY, USA
[4] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[5] Eric and Wendy Schmidt Center, Broad Institute of MIT and Harvard, Boston, MA, USA
[6] Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA
[7] Department of Statistics, Stanford University, Stanford, CA, USA

Abstract

We introduce hybrid BAG-seq: a high-throughput, multi-omic method that simultaneously captures DNA and RNA from single nuclei. We apply this protocol to 65,499 single nuclei from samples of five uterine cancer patients and validate the clustering using RNA-only and DNA-only protocols from the same tissues. Multiple tumor genome or expression clusters are often present within a patient, with different tumor clones projecting into distinct or shared expression states, demonstrating nearly all possible genome-transcriptome correlations. We also identify mutant stroma with significant X chromosome loss in various cell types and patient-specific stromal subtypes exhibiting aberrant expression patterns.

Background

Single-cell analysis offers an opportunity for insight into cancers and their interactions with the host. From single-cell RNA data, we can obtain valuable information about the cell type and state. This is key in characterizing the various types of normal stroma as well as the diverse expression states within the tumor. From single-cell DNA data, we can determine important lineage information. This enables us to distinguish tumor cells from normal cells and partition tumor cells into subclones. The fusion of these two types of analysis, high-throughput data of RNA and DNA from the same single cells, presents a significant advancement. By assigning both a distinct DNA and RNA identity to each single cell, we may begin to observe the complex interplay between cancer cells and

stroma, the emergence of malignant cells from pre-malignant ones, the driving forces behind specific expression states, and possible patterns of mutation in certain types of host stroma.

Methods for the high-throughput capture of either single-cell DNA [1–3] or RNA [4–11] are well-established and widely utilized in current research. However, when it comes to investigating both omics simultaneously from single cells, existing techniques present certain limitations. There are low-throughput methods capable of capturing both nucleic acids from single cells [12–24], yet these may not be straightforward to scale for high-throughput studies. Additionally, approaches for inferring copy number from high-throughput RNA-alone data have been described [25–29], but these methods rely on the assumption that gene expression and copy number are reliably correlated [30]. The integration of high-throughput genome and transcriptome analysis in a single method has only recently emerged [31–34] and has not yet been applied to human tumor samples to systematically investigate cancer heterogeneity and stromal mutations with the throughput demonstrated in this study. A method proven to be robust and reliable in this context could mark an important leap forward in our understanding of cancer biology at the single-cell level.

In this study, we introduce, characterize, and implement a hybrid high-throughput droplet-based technology that enables the capture of both DNA and RNA templates from the same cell nucleus. This multi-omic technology is an evolution of our BAG platform [35], where single cells are encapsulated into individual balls of acrylamide gel, with nucleic acid templates captured by Acrydite DNA primers and copolymerized into the gel matrix. We use a pool-and-split method to assign unique cell barcodes and varietal tags to each template, similar to what we have previously described [35]. After sequencing, the nucleic acid reads from each cell are partitioned into distinct DNA and RNA layers based on the characteristics of the mapped reads.

We utilized this hybrid BAG approach on frozen tissues obtained from five patients diagnosed with uterine cancer. We found sufficient transcriptional complexity in the nuclear RNA to clearly distinguish cell types. The DNA layer provided sufficient genome copy number information to differentiate stroma from tumor, to identify distinct subclones within the tumor, and to detect mutant stroma. We further confirmed these clustering patterns through comparison with results obtained from validated RNA-only and DNA-only BAG protocols.

Clustering in single-cell analysis presents a unique challenge. Popular methods like tSNE [36] and UMAP [37, 38] provide compelling visualizations; however, underlying algorithmic constraints can force cells to be clustered together. To quantify the likelihood of each cell's cluster affiliation, we enhanced our analysis by modeling clusters with multinomial distributions [39]. Generalizing to a space of pairwise-linear combinations, we identify and remove most doublet collisions—instances of mistaken cellular identity that arise when two distinct cells are assigned the same identity. We also use multinomial distributions to measure inter-cluster distance: two clusters are "far apart" if very few unique templates are required to determine that a random cell from one cluster is unlikely to have originated from the other.

The tumor and host expression clusters from each patient exhibited distinct characteristics. We use the gene count vectors for the clusters to determine differential gene

expression sets that inform labels of cell type and state. Combining expression data from all five patients' samples, we observed that the profiles of the stromal cell types are broadly consistent across all patients, but with some notable differences.

Using high-quality, high-throughput single-cell RNA and DNA data, we explore the complex genomic and expression landscape of five uterine tumors. We observe that cells resembling stromal cell types in their expression profiles are mainly diploid. However, we also identify intriguing instances of subclones within normal cells. In one tumor, we find diploid cells with X chromosome loss accounting for about half the plasma cell component. Upon closer examination of the expression data, we confirmed this DNA-RNA subcluster as the clonal expansion of a single B-cell lineage. Making such observations about stromal clonality requires a multi-omic approach.

Hybrid DNA-RNA data reveal an extensive range of associations between DNA tumor subclones and distinct expression states. Even at this simplest level, these interrelationships span from one-to-one correspondences to more complex many-to-many interactions, and some patterns are strongly suggestive of epigenetic variation. These observations could not be made without a multi-omic approach. Consequently, even deeper analysis was undertaken. We find cases where DNA subclones reveal additional subpopulations within an RNA expression cluster: we separate the RNA expression cluster based on DNA subclones and derive differential gene sets for the subpopulations. These subtle details would also not be observable using RNA-only data.

Overall, our data highlight the potential of multi-omic technologies to better understand cancer evolution, stromal mutations, and the tumor microenvironment. This broad view, coming from single cells alone, can be further honed by using differential gene sets to explore the spatial relationships using multi-probe high-resolution microscopy [40, 41]. Both the broad and detailed views of the inter-connectedness of subpopulations, host and cancer, may enlighten our view of cancer biology and guide future therapeutic strategies.

Results

Experimental design

Samples

We obtained fresh tissue samples of primary tumors from five patients with uterine cancer (Additional file 1: Table S1), and in three cases, distal "normal" endometrium (see Additional file 2: Table S2 for a detailed description). The tumor types surveyed include two carcinosarcomas, a serous carcinoma, an endometrioid adenocarcinoma, and a leiomyosarcoma. Each sample was frozen and pulverized into a powder. From this powder, nuclei were extracted for single-cell DNA, single-cell RNA, and single-cell DNA-RNA ("hybrid") BAG platform. We also used the same source material to perform whole-genome sequencing (WGS). This comprehensive approach ensured that all types of cells were proportionately represented in each method of analysis. To refine our methodology and study doublet collisions, we mixed powders from different patients prior to single-nucleus sorting, as discussed later. Additional file 2: Table S2 provides a comprehensive overview of the datasets utilized in our study, detailing the combination of sample origin (including unique setups like the mixed powder experiment), the associated protocols, and the respective experimental parameters.

The hybrid BAG platform

Our study uses the BAG platform, a versatile method that captures templates from a single-cell entity, either whole cells or nuclei. The BAG platform was built for flexibility, allowing for the reagent customization needed to capture DNA and RNA from the same single cells in a high-throughput manner. As described in Fig. 1A–D, nucleic acid templates (or simply "templates") are captured through primer hybridization to Acrydite-anchored primers embedded into balls of acrylamide gel, shortened to BAGs. This process is followed by primer extension, transcribing the template information of each single cell to primers securely tethered to a single BAG [35]. To establish cell identity, we used a pool-and-split synthesis approach to affix a *BAG tag* to each template, randomly assigning one of a million identities (96^3) to every BAG. During pool-and-split, we also introduced a template tag (also known as varietal tag or UMI) to each template. For the hybrid protocol, we used both oligo-TG primers and oligo-T primers to capture DNA and RNA templates, followed by using DNA polymerase and reverse transcriptase to transcribe the templates onto the anchored primers simultaneously. We then prepared sequencing libraries by tagmentation.

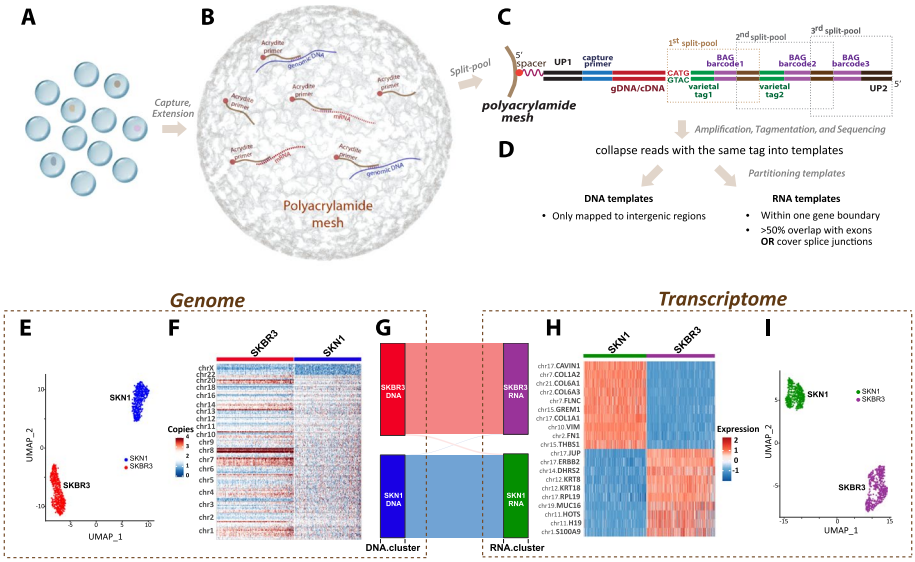

Fig. 1 Overview of hybrid BAG-seq protocol and performance on cell-mixture experiment. **A**–**D** Key steps for hybrid BAG-seq pipeline. **A** *Encapsulation*: Individual cells or nuclei are encapsulated within droplets containing acrylamide and Acrydite-modified primers that are designed to capture both mRNA and genomic DNA. **B** *Polymerization and primer extension*: Gel polymerization is followed by primer hybridization. Acrydite primers are extended by reverse transcriptase and DNA polymerase. **C** *Split-and-pool barcoding*: The double-stranded cDNA and genomic DNA are cleaved with a restriction enzyme. During successive rounds of pool-and-split, BAG-specific barcodes (purple) and template-specific varietal tags (green) are added. This process uniquely tags each molecule while assigning a distinct BAG barcode to templates in the same droplet. **D** *Sequencing and layer assignment*: Post-amplification, the molecules undergo tagmentation and subsequent sequencing. The sequencing reads are analyzed for expected structure, tags are extracted, and reads with identical varietal tags are collapsed into single templates. Templates are then partitioned into either the DNA or RNA layers based on their mapping characteristics. **E**–**I** Performance and genome-transcriptome correlation from a SKN1-SBKR3 mixture single-cell hybrid sequencing experiment: **E** clustering based on DNA copy number; **F** heatmap of DNA copy number variations across chromosomes for SKN1 and SBKR3 cells; **G** correlation between genomic clusters and expression clusters; **H** heatmap of marker genes of expression clusters; and **I** clustering result based on gene-count matrix

Genomic filters

All three protocols—DNA-only, RNA-only, and hybrid—require mapping reads to the genome and then organizing the mapped reads based on their BAG and template tags. First, each read is checked for having the expected structure of barcodes and primer sequences. Second, with the tag and primer sequences removed, the rest of the read is mapped to the reference genome. Reads that share a template tag and a BAG tag, and that co-localize in the genome, are collected into *read aggregates*, and operationally considered a captured template. The average number of reads per captured template for each library is also reported in Additional file 2: Table S2.

For classifying molecules into DNA or RNA origin, templates are assigned to either the RNA or DNA molecular layers based on the composition of their read aggregates. As detailed in the Methods section, templates with largely exonic read aggregates are assigned to the *RNA layer*, whereas templates with strictly intergenic read aggregates are assigned to the *DNA layer*. Any remaining templates are marked as *indeterminate*. Based on our observations, we also found it necessary to introduce additional measures to control bias in the hybrid data. First, we observed interference of RNA with DNA clustering. For DNA profiling and clustering, we excluded certain genomic hotspots, mostly poly-T/A-rich regions, to which many nuclear RNA sequences map. This progressively more stringent criterion significantly increased the quality of copy number data from the hybrid protocol, as shown in Additional file 3: Fig. S1. The bin boundaries were determined empirically to achieve uniform template counts using data integrated from the two normal tissues (Normal 1 and Normal 4) in this study. Second, for clustering analyses, we removed BAGs with fewer than 300 RNA-layer molecules or 600 DNA-layer molecules (discarding approximately 20% of all hybrid BAGs, see Additional file 2: Table S2), and then excluded BAGs where the ratio of DNA to RNA templates fell below one-fifth or exceeds five times the average ratio for that library. These BAGs, which constituted only a very small proportion (0–0.5%) of the nuclei passing minimum-count thresholds (Additional file 3: Fig. S2), could represent poor quality in either the RNA or DNA layer and could result in abnormal clustering results if not removed. We applied these categorization rules to all seven tissue nuclei datasets for all three protocols: hybrid, DNA-only, and RNA-only. The full distribution of each category per sample is shown in Additional file 3: Fig. S3A. Using data from one tumor sample as an example, we demonstrated that the layer categorizations are reasonable. Specifically, DNA clustering identities are very similar (97.6% concordance) whether using all molecules or only DNA-layer molecules as bin counts (Additional file 3: Fig. S4) in the DNA-only protocol, and a similar conclusion (95.2% concordance) was observed for gene expression clustering results in the RNA-only protocol (Additional file 3: Fig. S5), when restricting the analysis to the RNA layer or using all of the RNA templates mapped within transcripts.

We initially applied the molecular-layer concept to a mixture experiment involving two human cell lines: a normal fibroblast, SKN1, and a breast cancer cell line, SKBR3. The distributions of the basic parameters and downsampling curves from this experiment are shown in Additional file 3: Fig. S3B–G. We illustrate the clustering results and heatmaps based on the copy number and gene expression in Fig. 1E–I. The genomic and transcriptomic features from the hybrid protocol successfully recapitulate the published features of these two cell lines [35, 42]. The alluvial diagram (Fig. 1G) 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. Only 1.06% (10 out of 941) of the cells from the DNA cluster of one cell type (either SKN1 or SKBR3) projected to the RNA cluster of the other cell type, probably due to cell doublets.

Clustering RNA and DNA layers

We used the Seurat package [43] for our single-cell sequencing analysis—a tool widely recognized for its utility in gene expression clustering. For the RNA layer, we followed the standard methodology for expression clustering via "RunUMAP" and "FindClusters" functions. We extended the application of Seurat to cluster the DNA layer. We explored a range of DNA bin sizes and ultimately used a total of 300 bins for the DNA clustering and copy number analysis, as it provided a good balance of genomic resolution and average per-bin counts for a high signal-to-noise ratio and copy number reliability. These empirical bins range from 4.7 to 36.5 Mbp, with a median of 9.5 Mbp, and we treated the normalized raw DNA template count within each bin as a "gene input" for the clustering process and heatmap plotting. The segmentation information was only used for cluster-based copy number profile plotting, as shown in Fig. 2 and Additional file 3: Figs. S6–S9, panel B. This approach allowed us to identify shared copy number profiles that we could leverage in a manner similar to gene expression clustering (further detailed in Methods section).

Multinomial wheel

Whether dealing with DNA or RNA data, a common question often arises: how well does a single cell "fit" within its assigned cluster? To investigate this question within the context of a pre-established set of clusters, we first use multinomial probabilities. We take as a given the N clusters that Seurat identifies. For each cluster, we sum the gene (or genomic bin) count data over the cells in that cluster and normalize by the total template count. This results in a probability vector that represents the average gene frequency of the cluster. Utilizing this vector, we can calculate the probability that an observed single-cell count vector arose from each of the N clusters.

However, multinomial probabilities do not translate into a useful metric for deviation from a cluster. Every cell is assigned with almost no ambiguity to one of the major clusters. To properly identify cells that fall between clusters, we incorporate mixed cluster states into a multinomial wheel. For every pair of clusters, A and B with multinomial vectors p_A and p_B, we also consider the mixed state AB_α with multinomial vector $\alpha\ p_A + (1-\alpha)\ p_B$ for α in [0.1, 0.2, …, 0.9]. This results in a total of $9^*(N \text{ choose } 2) + N$ cluster states. By doing this, we can segregate the cells into two categories: core cluster members that stay close to an original cluster, and transitional members that fall between two clusters.

The hybrid platform is comparable to DNA-only and RNA-only platforms

In this section, we compare the hybrid BAGs to the DNA-only and RNA-only BAG platforms. We first focus on comparing the hybrid DNA layer to the DNA-only data, and then the hybrid RNA layer to the RNA-only data. We use the tumor tissue sample from

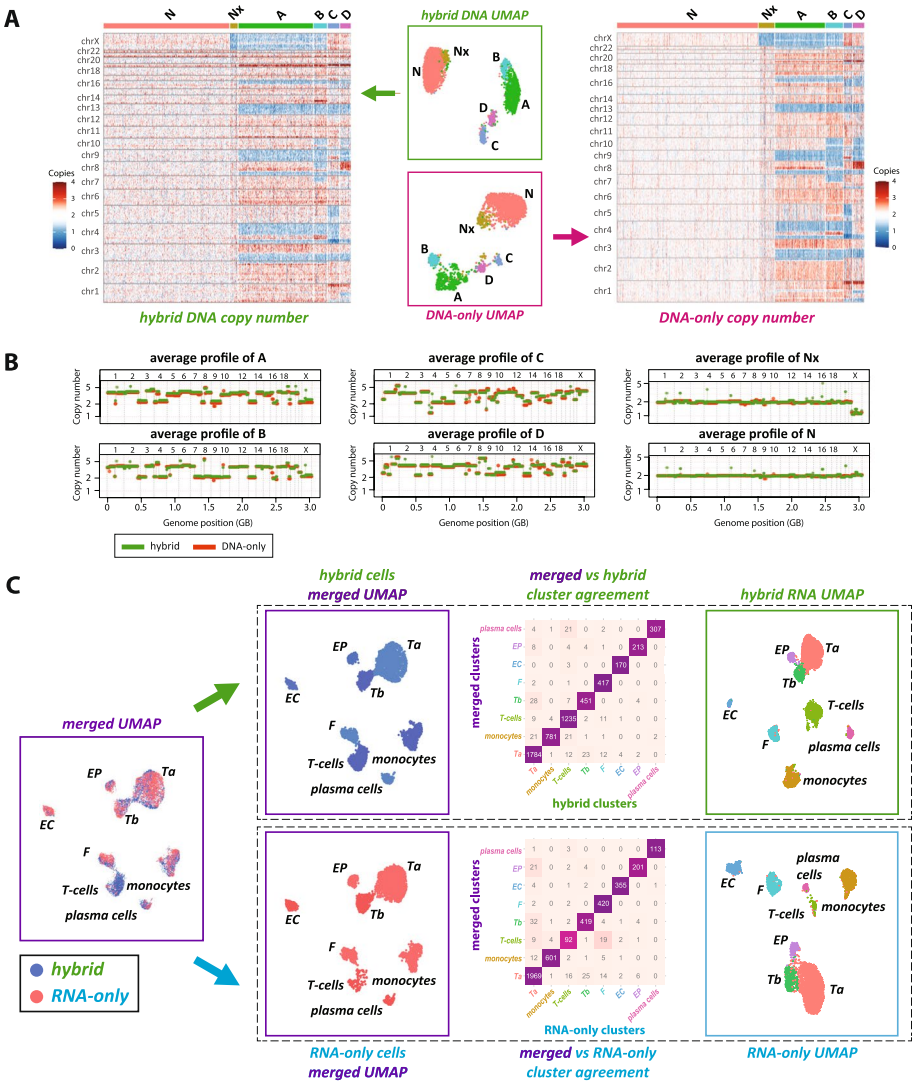

Fig. 2 Comparative clustering analysis using the hybrid BAG-seq protocol versus DNA-only and RNA-only protocols. **A** Single-cell DNA copy number analysis of a tumor sample from patient 2, comparing the hybrid protocol data (green) with the DNA-only protocol data (red). The tumor sample has six distinct copy number profiles: normal diploid cells (*N*), diploid cells with X chromosome loss (Nx), and four aneuploid tumor clones (**A**, **B**, **C**, and **D**). Central plots show UMAP visualizations of single nuclei, color-coded according to cluster identity. Adjacent heatmaps display the binned copy number variations for each single nucleus, arranged by cluster identity on the x-axis against genomic bins on the y-axis, with red indicating amplification and blue indicating deletion. **B** Aggregated copy number profiles derived from summing across all single nuclei within the same cluster, showing high concordance between hybrid (green) and DNA-only (red) datasets. **C** RNA expression analysis of the hybrid data (green) compared to RNA-only data (blue). To the far left, hybrid and RNA-only data are co-clustered into eight expression clusters: two tumor expression clusters (*Ta* and *Tb*) and six somatic cell types—fibroblasts (*F*), epithelial cells (*EP*), endothelial cells (*EC*), T-cells, monocytes, and plasma cells. These merged clusters are then segregated by hybrid or RNA-only origin. To the far right, each protocol's data are independently clustered into the same eight expression clusters. The central heatmaps quantify the agreement between the merged clusters with the respective hybrid and RNA-only clusters. Analogous plots for the other four patients are presented in the supplementary figures

patient 2 as a representative example, with similar comparisons for the other tumor tissue samples provided in Additional file 3: Figs. S6–S9.

DNA layer

Figure 2A compares the DNA layer from the hybrid data (left) with the data from the DNA-only BAG protocol (right). The central scatter plots display the clustering results in UMAP space, with the hybrid data (in the green box) above and the DNA-only data (in the red box) below. Each point represents a single nucleus, color-coded by its DNA copy number cluster. Both methods resolved six clusters, which we manually aligned based on pattern similarity. The N cluster includes cells with a typical diploid profile, whereas the Nx cluster represents a subpopulation of diploid cells with loss of an X chromosome. The remaining clusters—A, B, C, and D—exhibit varied aneuploid copy number profiles. The adjacent heatmaps illustrate the distribution of copy number changes across the genome, with deletions in blue, amplifications in red, and the diploid state in white. Each single cell is represented by a column and the cells are grouped by their DNA cluster.

To verify the congruence of profiles between platforms, we compared the average copy number profiles for each cluster in Fig. 2B, with DNA-only data in red and the DNA layer of the hybrid data in green. To quantify the similarity of clustering results between the two protocols, we used the multinomial wheel approach to measure the proximity of every single tumor nucleus to the centroids of tumor clones determined both by its own protocol (either hybrid or DNA-only) and by the other protocol. As shown in Additional file 3: Fig. S10, projecting DNA-only data to either DNA-only multinomial states or hybrid multinomial states showed no signal reduction (84.5% versus 84.5% nuclei within 2 units to the centroids), and similarly high concordance was observed when projecting hybrid data to either hybrid multinomial states or DNA-only multinomial states (77.2% versus 68.9% nuclei within 2 units to the centroids). Additionally, we examined heterozygous SNPs in both platforms and found similar patterns of loss of heterozygosity (LoH) and allele imbalance. These allele imbalance patterns (Additional file 3: Fig. S11) align with the copy number calls, in that when the copy number is an odd integer, allele imbalance is always observed.

RNA layer

We next analyzed the RNA layer of hybrid data compared to the RNA-only protocol. We first combined all the nuclei from both the hybrid and RNA-only platforms and clustered the integrated dataset into 8 clusters as shown in Fig. 2C (leftmost "merged UMAP" plot). While each cluster is labeled with a unique identifier, hybrid nuclei are shown in blue, and RNA-only nuclei in red. We then split the merged dataset by experimental origin with hybrid nuclei shown above and RNA-only nuclei below. The rightmost plots in the panel reflect the clustering of each dataset independently into the same eight identified categories.

We reserve a discussion of the differentially expressed genes for later, but currently label the clusters as monocytes, T-cells, F (fibroblasts), EC (endothelial), EP (epithelial), plasma cells, and two distinct tumor RNA clusters, Ta and Tb. The central agreement matrices, formatted as heatmaps, show the consistency of cell classification across platforms within the merged dataset. The top matrix compares hybrid cluster assignments to merged dataset classifications, while the bottom does the same for

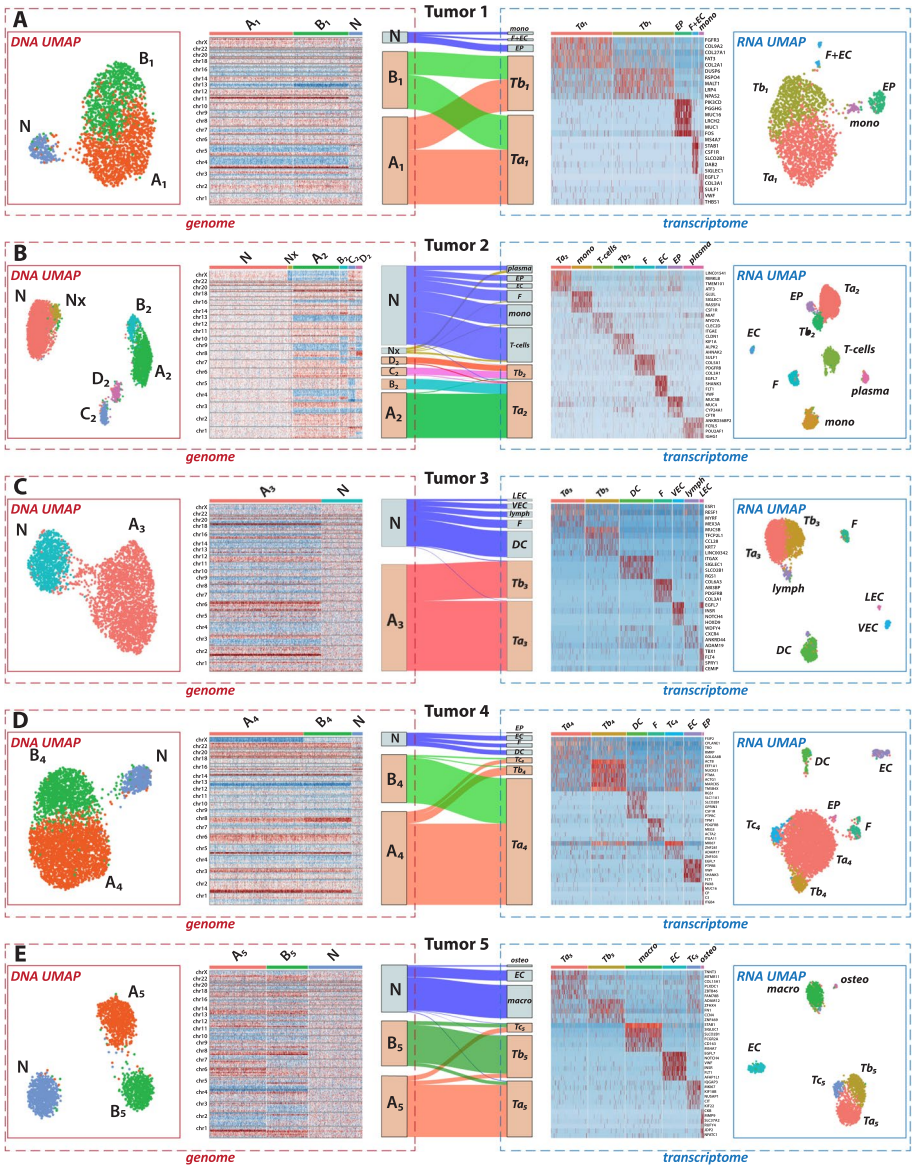

Fig. 3 Connecting genomic and transcriptomic data across five tumor samples using alluvial diagrams. **A**–**E** Summary of the hybrid data analysis of tumor tissue samples from five individual patients, show the various connections between genomic and transcriptomic landscapes. For each patient: The "genome" sections on the left, bordered in red, display the DNA layer information, including UMAP visualizations of single nuclei colored by genomic cluster identities and accompanying heatmaps showing copy number variations across the genome. The "transcriptome" sections, bordered in blue, present the RNA layer data, including UMAP plots of single nuclei colored by expression cluster identities. These are accompanied by heatmaps of marker genes that are distinctly upregulated within specific expression clusters. At the center of each panel, alluvial diagrams connect the DNA and RNA data, linking genomic cluster identities (left) to the RNA expression clusters (right). The thickness of the flow lines represents the proportion of nuclei that belong to a specific genomic cluster (X) and an expression cluster (Y), illustrating the integrative analysis facilitated by the hybrid BAG-seq platform

RNA-only data. For both datasets, a significant proportion (95%) of cells align diagonally, confirming that cluster identities are well preserved across the two platforms.

Profiling genome and transcriptome of tumor samples using hybrid data

For each of the five tumors, we clustered the DNA and RNA layers of the hybrid data, respectively (Fig. 3). The DNA clusters are presented on the far left, accompanied by copy number heatmaps similar to the previous illustrations. On the far right, RNA clusters are displayed, along with a heatmap that illustrates the relative expression levels across sets of differentially expressed genes (blue for low expression, red for high).

Based on the dual DNA and RNA identities assigned to single cells in the hybrid dataset, we quantified the frequency with which cells from a specific DNA cluster appear in a given RNA cluster. This cross-layer association is visualized using alluvial diagrams. For example, in panel A from Tumor Sample 1, the alluvial diagram illustrates that the diploid N cells map exclusively with non-malignant expression profiles (macrophage, F + EC, and EP). Conversely, both tumor clones A_1 and B_1 map to the two tumor expression profiles (Ta_1 and Tb_1).

We now delve into the unique characteristics of each of the five tumor specimens, drawing comprehensive insights from both DNA and RNA layers. Focusing on the tumor genome projection patterns, each of the five tumors displayed a different projection pattern, and we observed almost all the possible 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. We have seen: 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). Furthermore, based on the alluvial plots and the underlying tumor cell identity contingency tables, we calculated the Rand Index and Adjusted Rand Index to quantify the correspondence between DNA and RNA cluster assignments (Additional file 4: Table S3). Higher values of these indices indicate stronger concordance between the two clustering schemes. Tumors 2 and 5, which showed more evident alignment between DNA and RNA clusters, exhibited higher Rand Index and Adjusted Rand Index values compared to Tumors 1, 3, and 4. The latter tumors displayed more mixed projections across modalities, suggesting greater transcriptional plasticity.

With the details of the five cases using the hybrid protocol provided in Additional file 5: Supplementary text, we briefly highlight the important findings regarding these respective cases here.

Tumor 1 was a uterine carcinosarcoma, which is pathologically observed as a biphasic tumor containing both carcinomatous and sarcomatous components. Echoing this in the RNA analyses, cluster Ta_1 showed high expression of fibroblast-specific genes, including *FGFR3*, *COL9A2*, and *COL27A1*, in keeping with the pathological classification of this tumor as having a sarcomatous component, whereas these fibroblast genes had lower expression in cluster Tb_1. This separation between Ta_1 and Tb_1 in gene expression

patterns was consistent across the hybrid, RNA-only BAG protocol, and 10 × Chromium v3 platform (Additional file 3: Fig. S12). We found that the two tumor clones, which mainly differ in copy number of chromosome 13, each project equally to tumor RNA clusters Ta_1 and Tb_1. To explore whether any hidden patterns were missed during clustering, we projected DNA-layer clone identities onto the RNA-layer UMAP space and found that nuclei from both DNA clones are randomly interspersed, showing little correspondence between DNA and RNA clusters (Additional file 3: Fig. S13). Similarly, randomly interspersed patterns were observed when projecting RNA-layer cluster identities onto the DNA-layer UMAP coordinates. Tumor 2 is a uterine serous carcinoma and presents a more complex DNA and RNA landscape than Tumor 1. The tumor sample served as our example in the previous section comparing the BAG platforms (Fig. 2). In contrast to Tumor 1, the DNA-RNA correspondence in Tumor 2 is visually more straightforward: tumor DNA clones A_2 and B_2 mainly project to RNA cluster Ta_2, while clones C_2 and D_2 correspond to Tb_2. Regarding resolution of the RNA-layer clustering of the hybrid protocol, we stopped at the hierarchical level where both the hybrid and RNA-only protocol yielded consistent clustering patterns and featured genes. We did not pursue finer subdivisions of RNA clusters that appeared in only one protocol. Given the hierarchical nature of RNA clustering, we report the top-level, biologically meaningful RNA clusters for correspondence analyses and alluvial plots. This rule applies to all five cases. However, in cases where distinct DNA clones are associated with a shared RNA expression profile, such as A_2 and B_2 with Ta_2, we further dissect the RNA cluster to test whether these DNA subgroups differ in gene expression. To assess how copy number variation affects gene expression, we compared gene expression profiles between the subgroups Ta_2-A_2 and Ta_2-B_2, as well as Tb_2-C_2 versus Tb_2-D_2, as shown in Additional file 3: Fig. S14. The top genes separating these subgroups are not necessarily located in regions with copy number differences that distinguish the two DNA clones.

Tumor 3 is an endometrial adenocarcinoma. It contains a single tumor DNA clone that projects to two distinct RNA clusters—one estrogen receptor (ER) positive and the other ER negative. This observation is consistent with the pathological findings, with the immunohistochemistry (IHC) images showing that ER-positive and ER-negative tumor cells are spatially intermixed within the tissue (Additional file 3: Fig. S15).

Tumor 4 is another uterine carcinosarcoma case. Unlike the other carcinosarcoma case Tumor 1, the tumor cells here globally exhibit sarcomatous-like features, with one RNA cluster, Ta_4, comprising about 90% of the tumor cells. These cells are proportionally derived from two distinct tumor DNA clones, A_4 and B_4, which show substantial differences in copy number. In contrast to this major cluster, a smaller tumor cluster, Tb_4, exhibits upregulation of genes related to cytoskeletal organization, cell motility, protein synthesis, and cellular metabolism, and is primarily derived from clone A_4. In addition, unlike the first three cases, a separate RNA cluster characterized by proliferation markers is clearly distinguishable at this clustering resolution, but disproportionately comes from one tumor DNA clone A_4. As in previous cases, when multiple DNA clones project to a single RNA cluster, we performed differential expression analysis between the DNA-defined subgroups, in this case, Ta_4-A_4 and Ta_4-B_4. While there are significantly differentially expressed genes between these two subgroups (Additional file 3: Fig. S16A), the projections of A_4

and B_4 onto Ta_4 appear randomly intermixed in the RNA UMAP space (Additional file 3: Fig. S16B), similar to what was observed in the tumor RNA clusters of Tumor 1. This suggests limited correspondence between DNA copy number variation and RNA expression, despite the much more pronounced genomic differences between clones A_4 and B_4 compared to the differences between A_1 and B_1.

Tumor 5 is a uterine leiomyosarcoma, with the most intuitively straightforward tumor DNA-RNA correspondence among these five cases, as it exhibits a nearly one-to-one mapping between the two tumor DNA clones and two RNA clusters, despite some cross-projections. Similar to Tumor 4, an RNA cluster representing proliferating cells is distinguishable at this hierarchical level, but unlike Tumor 4, both tumor DNA clones proportionally contribute to this cluster Tc.

Combined clustering of all patient samples

Integrative clustering analysis and cluster phylogeny

In our previous sections, we examined one patient at a time, while in this section, we aim to understand how patient expression profiles interrelate and potentially achieve finer resolution in our clustering analysis. To gain such a global picture of the hybrid data, we performed cluster analysis on the combined data from all patients, separately for RNA layer and DNA layer, as demonstrated in Fig. 4. For RNA layer clustering, we lowered the threshold for RNA template count from 400 to 300. This analysis includes hybrid data from the five tumors (Tumor i) presented in the previous section, as well as normal tissue samples from patients 1, 2, and 4 (Normal i). We then explore which expression clusters map to which patient and within each patient to which genome cluster, diploid copy number (flat), diploid copy number with one chromosome X (X-loss), or complex aneuploid genome (CN+). In this way, we reveal diverse and common cellular and functional repertoires. These results are summarized in Table 1.

Specifically, our gene expression clustering procedure incorporates several novel features, which we detail below for clarity. First, we include 3500 "empty" cells derived from the RNA layer of DNA-only BAGs. These empty cells were clustered together in Fig. 4A. This cluster also attracts hybrid BAGs with severe RNA template depletion. Overall, 2.4% of hybrid cells coalesce into the empty state.

As before, we use Seurat's FindClusters and UMAP functions to cluster and display 40,149 nuclei, which uses the UMAP coordinates to visualize the single cells, as shown in Fig. 4A. After the initial clustering, we further resolved the diploid cells to delineate subpopulations, described in the next section. Specifically, we chose four sub-regions of the initial UMAP (blood elements, fibroblasts, endothelial cells, epithelial cells) for further subclustering. Given the single-cell pool's diverse cell type composition—ranging from tumor, epithelial, endothelial, myeloid, etc. —an iterative clustering approach is a reasonable strategy. For the tumor subclusters, we adopted the case-specific cluster information defined in the previous section.

We use UMAP coordinates to obtain a planar representation of the cells, and our method of multinomial analysis to color the cells by expression type. The aggregate of templates in each subcluster forms a gene probability vector, wherein the total number of templates mapping to a gene is normalized by the total number of templates mapping to *any* gene, resulting in a probability distribution. This probability vector determines a

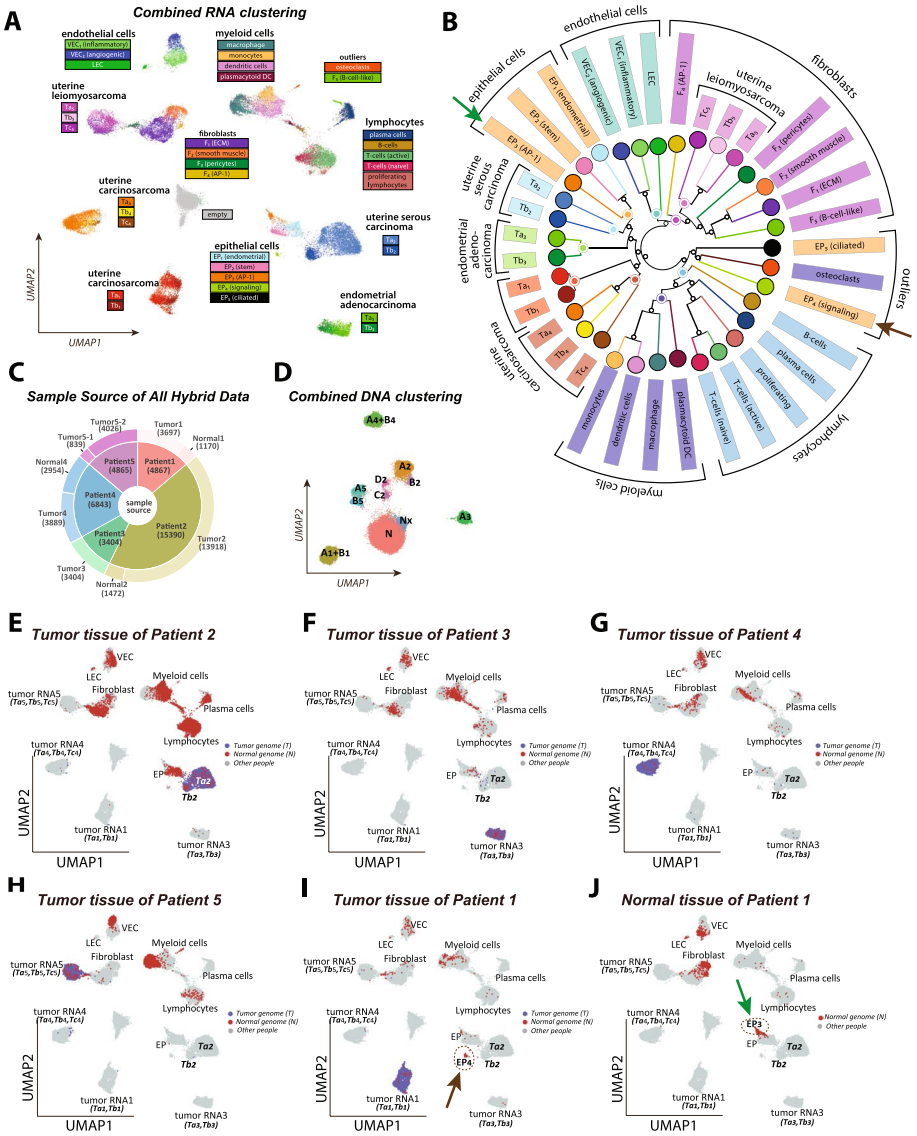

Fig. 4 Integrated cluster analysis and unique cluster identification using aggregate tumor and normal tissue data from all patients. **A** UMAP scatter plot showing the RNA layer data from both tumor and normal tissue samples across all patients, totaling 40,149 nuclei. Stromal clusters are defined through iterative subclustering, while tumor subclusters are defined from individual tumor analyses as presented in Fig. 3. An empty cluster (gray) consists of nuclei mainly from the RNA layer of seven DNA-only experiments. Each point is colored according to the cluster with the highest likelihood as determined by a multinomial model. **B** The neighbor-joining tree illustrates the relationships among the stromal and tumor subtypes. The tree is computed from inter-cluster distances based on multinomial distributions. **C** The source of nuclei from the hybrid protocol used in the combined analyses. **D** Combined DNA clustering after removing nuclei clustered to the "empty" state of the RNA clustering in **A**. **E**–**J** The tumor-genome (blue) and normal-genome (red) nuclei projections on the RNA UMAP space for six tissue samples. The tumor-genome or normal-genome information is determined by the combined DNA analysis in **C**. Unique stromal components specific to certain tissues are circled in dashed lines and indicated by arrows in **I** and **J**, and marked in **B**

multinomial distribution for each cluster that in turn, assigns a cluster probability for each single cell based on its gene counts. In Fig. 4A, each point's color reflects the cluster most likely to have generated the gene counts in that cell.

Table 1 Co-clustering counts across all samples. This table presents the distribution of nuclei for each patient (P1 to P5) and tissue sample (Tumor or Normal) categorized by DNA cluster: diploid (flat, white), diploid with one X chromosome (X-loss, blue), and aneuploid (CN+, red). It details the total count of nuclei within each category, along with sub-counts for each expression cluster

			Stromal cells																						Tumor									
			blood elements										endometrial					fibroblast				endothelial			tumor 1		tumor 2		tumor 3		tumor 4			tumor 5		
			myeloid cells					lymphocytes														lymph	vascular													
	sample name	total count	monocytes	dendritic cells	macrophage	plasmacytoid DC	osteoclasts	T-cells (naive)	T-cells (active)	proliferating lymphocytes	plasma cells	B-cells	EP1 (endometrial)	EP2 (stem)	EP3 (AP-1)	EP4 (signaling)	EP5 (ciliated)	F1 (ECM)	F2 (smooth muscle)	F3 (pericytes)	F4 (AP-1)	F5 (B-cell like)	LEC	VEC1 (inflammatory)	VEC2 (angiogenic)	Ta1	Tb1	Ta2	Tb2	Ta3	Tb3	Ta4	Tb4	Tc4	Tb5	Tc5
P1	Normal 1 (flat)	1080		8				2	1	1	1		2	8	306		3	47	358	7	99	3	16	213	5										
	Tumor 1 (flat)	168	8	8	19					1			3	3	9	86		6	2	4			1	12	2	3	1								
	Normal 1 (X-loss)	58						1	1						2			4	35		6		1	8											
	Tumor 1 (X-loss)	3							1															1		1									
	Normal 1 (CN+)	32												1	2			1	18				1	9											
	Tumor 1 (CN+)	1770			3								2		1	3		2								1090	669								
P2	Normal 2 (flat)	1418		20	8	3		13	23	1	1	1	94	469	5		7	405	180	9	2	4	13	154	1										
	Tumor 2 (flat)	2939	347	342	10	38		251	825	67	193	30	196	20	1			364	19	9	2	10	10	113	33			43	14						
	Normal 2 (X-loss)	34		1				2	1				1	9				9	6			1		4											
	Tumor 2 (X-loss)	229	11	17				19	42	3	89		4					18	2			2		5				14	3						
	Normal 2 (CN+)	20		1									1	3				6	4			1		1				2	1					
	Tumor 2 (CN+)	2382	12	9	1			3	47	2	11		8	2	2			15	3	2		1		6	2			1767	487						
P3	Tumor 3 (flat)	928	28	400	13	60	7	16	17	18	19	11	9	4			1	149		2		2	37	69	21					11	34				
	Tumor 3 (X-loss)	14		6		1					1							2						3						1					
	Tumor 3 (CN+)	2462	1	11	1	1		1	3				4											2	2				2	1518	916				
P4	Normal 4 (flat)	2775	8	93	17	36		86	77	17	521	35	162	8	19		5	731	327	50	2	111	46	415	9										
	Tumor 4 (flat)	358		64	14	5		4	3	28	8		2	2				53	33	20		1	8	67	36							7	3		
	Normal 4 (X-loss)	127		2	1	1		20	13	1	25	1	5					21	18	1		6	5	7											
	Tumor 4 (X-loss)	27		3	1				1	2	1							5	2					4	3							2	3		
	Normal 4 (CN+)	52				1		3	1	2	10							18	9	1		2		4	1										
	Tumor 4 (CN+)	3594		2	3				2	2	3		1	4				6	8				1	7	1					1	2	3045	272	140	
P5	Tumor 5 (flat)	637	24	43	311	19	21	2	3	15								7		4		2	2	9	153									9	6	7
	Tumor 5 (X-loss)	20		1	8		1											1		1				1	6									1	
	Tumor 5 (CN+)	1244		6	18		1		1											5		3			18									622	467	103

We then use the multinomial distributions to establish a distance metric between any two subclusters, each with its own multinomial distribution. The asymmetric disparity between two multinomial distributions, A and B, is determined by simulation, in which we determine the number of unique RNA-layer templates of a cell simulated from distribution A that are needed to preferentially assign the cell to A rather than to B, for some given confidence threshold. We then symmetrize the disparity measure, combining A to B and B to A. Clusters with distinct profiles will require fewer templates to establish separation, while those with similar profiles require more templates for differentiation. We invert this measure to establish a pairwise distance (Additional file 3: Fig. S17) and use neighbor-joining to build a phylogeny over the expression clusters (Fig. 4B). We rank the genes that best separate two sets of expression clusters, A and B, using binomial methods, rather than the FindMarkers function within Seurat. We use the ratio of total templates in A and B to determine the null expectation of the ratio for any given gene. We then apply a binomial test to the observed counts for each gene in A and B. The marker genes distinguishing tumor expressional subclusters for each patient can be found in Additional file 6: Table S4, while those for stroma clusters across all five patients are listed in Additional file 7: Table S5.

On the DNA front, from a total of 35,369 nuclei from the sources indicated in Fig. 4C studied using the hybrid protocol, after removing the 2.4% (859 out of 35,369) low-quality nuclei that were clustered into the empty state in the gene expression clustering result, we clustered the DNA-layer of all the remaining nuclei based on genomic bin counts. Similar to the RNA clustering result, in the DNA space in Fig. 4D, tumor-genome clusters were quite distinct from the normal-genome clusters, and distinct clones within a patient mapped nearby to each other or merged into a single cluster at this resolution of clustering.

Common and unique features in stromal subpopulations

Combining RNA and DNA clustering information, the projections of nuclei from six sample sources into the combined RNA space (Fig. 4A) are shown in Fig. 4E–J, 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 gray. The projections of the tumor genomes are very distinct between patients, well separated from each other and the projections of the normal genomes.

Although the normal-genome stromal clusters are mostly shared between patients, the distribution and abundance of different subtypes vary significantly by patient and sample. For example, there are substantial differences in immune cell proportions among the five patients, as illustrated in Fig. 4E–J with more quantitative data in Table 1, suggesting variation in tumor microenvironment and immune response. In addition, there are several gene expression clusters that only belong to certain tissues and certain patients, implying influences from tumor microenvironments and personal genomes. For example, the epithelial-like cluster EP_4 (Fig. 4I, indicated by the brown arrow) from the tumor tissue of patient 1 is significantly different from the EP_3 cluster (Fig. 4J, indicated by the green arrow) from the distal normal uterine site in the same patient, and they are both distinct from the main epithelial cluster EP_1. These two clusters are also pointed to by arrows in the hierarchical wheel in Fig. 4B, far away from each other. We believe these distinct clusters are not due to batch effects or template counts, not only because they have distinct and biologically plausible gene expression patterns (Additional file 7: Table S5), but also because these distinct clusters overlapped well in experimental replicates and showed the same patterns in RNA-only datasets (Additional file 2: Table S2). These unique stromal subpopulations will be discussed in detail in the following section.

From a hierarchical and phylogenetic perspective, the multinomial tree (Fig. 4B) provides a more quantitative view of these diverse stromal and tumor RNA clusters. The branches of the tree largely preserve cell-type categories. The blood elements share a common branch (blue labels) with a myeloid-derived sub-branch (dark blue) distinct from lymphocytes (light blue); a branch of epithelial cells (orange); fibroblasts (purple) and endothelial cells (green). Most of the subclusters fall on their expected branch. The exception includes a single sub-branch containing two epithelial subclusters, EP_4 and EP_5, and osteoclasts, a myeloid cell type. In general, clusters that are close together in the UMAP (Fig. 4A) share a common branch in the tree (Fig. 4B). The major exception is cluster F_5, B-cell-like fibroblasts, which are near the B-cells in the UMAP but nearer to the fibroblasts in the tree.

The tumor clones from each patient occupy distinct sub-branches in the tree. The uterine leiomyosarcoma (purple), a muscle-derived tumor, has expression subclusters on the fibroblast branch of the tree. The other four tumors share a deep branch with the epithelial cells. One sub-branch contains the two uterine carcinosarcomas (red and dark red). Nearer the epithelial cells in the tree, are the endometrial adenocarcinoma (green) and nearer still the uterine serous carcinoma (blue). The branch lengths provide a relative measure of similarity, showing that Ta_1 and Tb_1 are highly similar as are Ta_3 and Tb_3. In contrast, the subclones of Ta_2 and Tb_2 are far apart.

Multinomial wheel and crossovers

In the previous section, we observed that the majority of nuclei exhibit concordant DNA and RNA profiles: diploid DNA with stromal RNA expression (flat, *N*) or complex DNA

patterns with tumor RNA expression (CN+, T). These concordant nuclei constitute the expected biological behavior; however, there are a subset of cells that do not match this pattern. Additional file 8: Table S6 shows the counts for each patient of cells that are diploid or complex (flat or CN+) and map to normal clusters or tumor clusters (*N* or T). Across all patients and tissue samples, 1–5% of nuclei have flat copy number profiles and tumor expression patterns, or a combination of copy number variations and stromal expression patterns. Although accounting for a small proportion of the total dataset, these crossover nuclei may represent an interesting population. Alternatively, they may be the result of unresolved doublet collisions [44]. Additional file 9: Table S7 summarizes the counts and calculates the proportion of concordant and crossover nuclei per patient.

To determine if crossover nuclei are a unique biological state or collision artifacts, we employed the multinomial wheel to differentiate mixed states. Integrating DNA-layer data across all hybrid-protocol experiments, we constructed a multinomial wheel with tumor clusters, diploid cells and diploid cells with X-loss as the individual spokes of the wheel (Additional file 3: Fig. S18A). Some tumor DNA clusters are so similar (A_1 and B_1, A_4 and B_4) that we collapse each of those pairs into a single node (A_1/B_1 and A_4/B_4). For each pair of the 11 vertices, we created 9 equally spaced sampling states, resulting in a multinomial wheel with (11 choose 2) = 55 spokes and 9*(11 choose 2) = 495 intermediate states.

From the DNA counts of each nucleus, we evaluate the probability that those observations are derived from each of the (495 + 11) = 506 multinomial distributions in the wheel. We assign each nucleus to the node with the highest posterior probability. Notably, more than 60% of nuclei align with a "pure" cluster on the multinomial wheel by residing on a vertex, and 85% are situated within two units' distance from a pure cluster (Additional file 3: Fig. S18A). Informed by two mixture experiments, detailed in the Methods and Additional file 3: Fig. S18B–E, we apply a collision filter, marking for removal nuclei that are 3 or more units from a vertex and fall between a normal DNA cluster (N, Nx) and a tumor DNA cluster (A_i, B_i, etc.). Additional file 7: Table S5 and Additional file 8: Table S6 (filtered nuclei) show the counts after applying the collision filter.

If crossovers are the result of cryptic collisions, the collision filter will disproportionately reduce the frequency of crossovers compared to the concordant nuclei between the DNA and RNA layers. Indeed, our observations align with these expectations: while the collision filter removed 14% of concordant nuclei, it removed a substantial 71% of crossover nuclei (Additional file 8: Table S6). Removing collisions does not alter the results presented in the previous section. A version of Table 1 restricted to nuclei that pass the collision filter can be found in Additional file 10: Table S8.

The multinomial wheel can be used for much more than detecting doublets. It quantitatively measures how much a single cell deviates from the centroids of major clusters in DNA or RNA space and can provide insights into genomic heterogeneity, particularly for cells transitioning between states. It does not rely on PCA or UMAP coordinates, and by using only cluster centroids as reference points, it offers an interpretable, coordinate-free metric for cell-wise distance that aids interpretation of nonlinear, reduced-dimensional embeddings such as UMAP. For example, in Additional file 3: Fig. S19, we extracted tumor cells from three cases in which only two major tumor clones were present and

projected them onto a one-dimensional space based on their best multinomial probability. The proportion of cells classified as "in transition" by the one-dimensional multinomial wheel agrees with the intuitive spatial distribution observed in UMAP space and is significantly higher than expected from doublet rates, suggesting the presence of intermediate or transitioning cells that share features of both clones.

Loss of X chromosome in stromal lineage

In this final analysis, we revisit the occurrence of X chromosome loss in diploid cells, particularly noted in the plasma cell population from patient 2's tumor sample. X chromosome loss, while reported before in various tissues associated with cancer [45, 46] or aging [47, 48], presented an unusual prevalence in our dataset among the stromal cell types in the tissue. For example, about half of the plasma cells in patient 2 exhibited X chromosome loss, a finding that merited further investigation into its biological implications.

To determine whether the X chromosome loss was the result of a clonal event or occurred independently across different cells, we performed a haplotype analysis: If all affected cells lost the same X chromosome, it would suggest a common ancestry; whereas if they lost different copies of the X, it would point to independent, convergent events. The tumor genome from patient 2 has two subclones with X-loss (A_2, B_2), and we used these two subclones to phase the SNPs on the X chromosome. Upon aggregating SNPs within each nucleus, we found that the Plasma-Nx nuclei share a common haplotype, suggesting a clonal nature, whereas the T-cell-Nx population shows both haplotypes (Fig. 5A). Given the robust coverage of the Plasma-N and Plasma-Nx subclusters, we were able to conduct a differential analysis on their aggregate gene expression data. As shown in Fig. 5B, there are four genes with significantly different expressions: *XIST* and *TSIX* are under-expressed in the X-loss population, affirming the X-loss observed in the DNA layer. The other two genes are *IGHG1* and *IGHG2*, hinting at a potential compositional difference between these plasma cell populations. As *XIST* RNA is required for X chromosome inactivation and is only expressed from the inactive X chromosome [49–51], these results also verify that nuclei from the "Nx" DNA clone lost their inactivated X chromosomes.

Considering the distinct IGH expression in the plasma subclusters, we further investigated the clonality of the *Plasma-Nx* population through VDJ recombination analysis. Using MiXCR [52] to count the unique VDJ recombination patterns, our analysis strongly suggests that the Plasma-Nx cells are primarily derived from a singular B-cell lineage expansion, in contrast to the Plasma-N cells, which originated from a diverse array of lineages (Fig. 5C).

To corroborate our findings and further visualize the loss of the X chromosome, we applied RNAScope technology on the same tumor tissue from patient 2. The spatial transcriptomics images (Fig. 5D–I) provide a vivid illustration of the XIST expression patterns. DAPI staining (blue) marks the nuclei, while XIST (red) and IGHG (green) show specific gene expressions. In Fig. 5D, half of the cells display no XIST signal, indicating X-loss, likely from tumor A_2 or B_2 clones with deletions in the X chromosome. Figure 5E and F show IGHG, the plasma cell marker, and XIST expression within two different regions, showing two opposite phenomena. While most of the plasma cells in

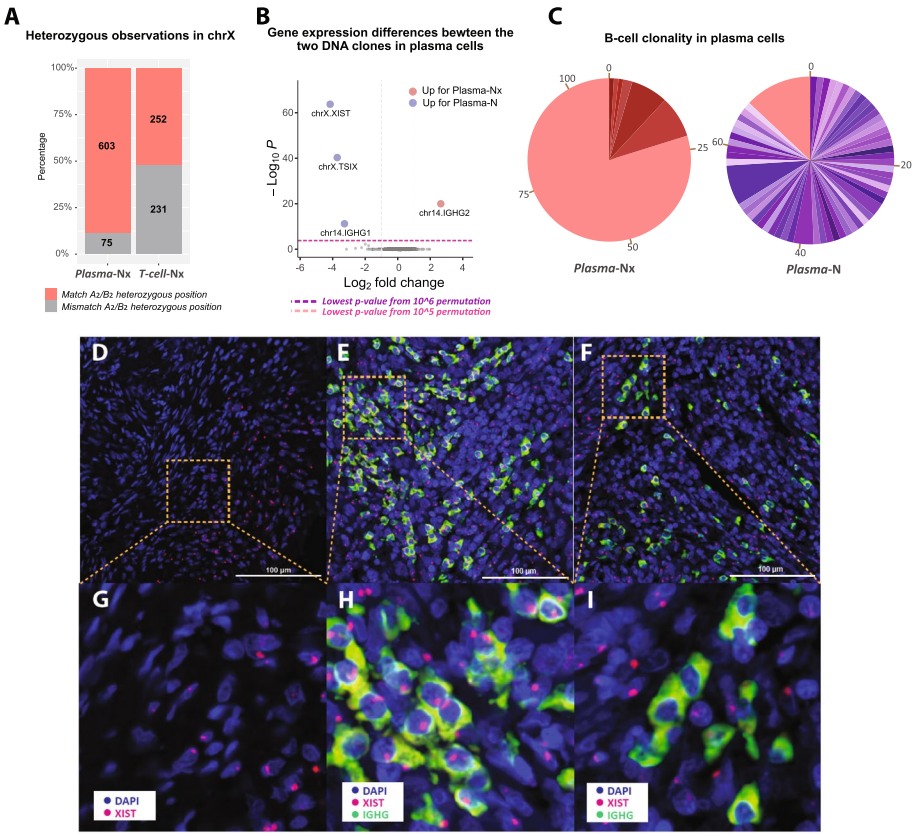

Fig. 5 Analysis of chromosome X loss in somatic cells of the primary tumor 2. **A** Bar plot showing the ratio of X-haplotype observations in the X-loss populations of plasma (Plasma-Nx) and T-cell (T-cell-Nx) nuclei from patient 2. Tumor subclones A2 and B2 with only one copy of the chromosome X are used to phase the X chromosome SNPs in the Plasma-Nx and T-cell-Nx populations as belonging to one haplotype (red, match A2/B2) or the other (gray, mismatch A2/B2). T-cell-Nx nuclei exhibit a balanced distribution of SNVs from both haplotypes, while Plasma-Nx nuclei show a pronounced bias toward the A2/B2 haplotype. **B** A volcano plot shows the genes with statistically significant expression differences between diploid plasma cells (Plasma-N) and plasma cells with one X chromosome (Plasma-Nx). **C** Pie charts showing VDJ recombination results for Plasma-N and Plasma-Nx nuclei. Each color represents a unique B-cell clone identified by its CDR3 sequence, with the size indicating the clone's prevalence. The Plasma-N chart shows a diverse clonal makeup with few dominant clones, while the Plasma-Nx chart shows a clear dominance of one lineage (red), constituting 80% of the population and matching the signature of the primary clone in Plasma-N. **D**–**I** Spatial transcriptomics images illustrating XIST expression in the same tumor tissue of patient 2. **D** RNAscope images displaying the expression of XIST (red) along with DAPI staining of nuclei, showing half of the cells with no XIST signal detection. **E**, **F** IGHG (green, plasma cell marker) and XIST (red) expression along with DAPI staining of nuclei. **E** shows a region consisting of normal plasma cells with XIST gene expression, while **F** shows another region of plasma cells with no XIST red dots detected. **G**–**I** are zoomed-in images of **D**–**F**, respectively

Fig. 5E contain the XIST signal, implying the presence of the inactivated X chromosome; the majority of the plasma cells in Fig. 5F lack the XIST signal, suggesting that these cells have lost the inactivated X chromosome. Both the plasma cells with and without XIST signals were found spatially localized in the tumor microenvironment.

We next extended our X-loss analysis to the other tissue samples. For patient 1, we noted a 5% X-loss in the diploid component of the normal distal sample, predominantly in the smooth muscle (F_2) expression cluster. However, establishing clonality in this context was challenging because the tumor subclones in this patient retained both X chromosomes, making it difficult to phase the lost X haplotype for validation. In patient 4, we

identified significant X chromosome loss in the blood component of the normal distal sample, particularly in naïve and active T-cell populations. Additionally, a smaller but notable percentage of fibroblast ECM and smooth muscle cells (F_1 and F_2) also exhibited X-loss. In contrast, patients 3 and 5 displayed minimal X chromosome loss, accounting for 0.5% of the diploid population.

Discussion

In our analysis of the hybrid data, we observed a small but significant population of cells that appear to violate their identities: tumor cells with normal expression and diploid cells with tumor expression. The former suggests a sort of mimicry in which the cancer cells appear like normal stroma, while the latter may point to a pre-aneuploid tumor precursor state or stromal cells dramatically altered by the tumor microenvironment. However, the application of a collision detection method significantly reduced these crossover populations. This reduction has tempered our initial belief in the prevalence of mimicry or progenitor states as observed in the hybrid dataset. The existence of such cells cannot be completely dismissed, and further refinements to the method, larger datasets, and additional data types will be required to determine their frequency and biological significance.

The current version of hybrid BAG-seq yields relatively low DNA-layer template counts, lower than those obtained from the DNA-only protocol. This is likely due, in part, to suboptimal primer denaturation and hybridization time and temperature conditions, which have not yet been fully optimized and could be improved in the future to increase yield. In addition, we applied very conservative filtering criteria for DNA-layer molecules to ensure copy number accuracy-restricting analysis to intergenic templates and excluding hotspot regions. Furthermore, the majority of captured templates, especially for nuclei samples, actually fall within intronic regions, which we did not use for either DNA-layer or RNA-layer analyses in this work, but they are worth further investigation to develop better algorithms for assignment to the RNA or DNA layer. Regarding RNA yield, we used NST buffer for nuclei isolation but did not test alternative methods that may better preserve RNA integrity and potentially further enhance recovery for different tissue types [53–56].

Most tumors exhibit detectable copy number differences [57]. Because of the instability of aneuploid genomes during replication, we can often find unique copy number changes that trace a tumor's lineage. As we have observed in this study, even stromal cells may carry detectable copy number changes that mark their lineal descent. Unfortunately, some tumors do not have copy number changes and many bifurcations in the tumor phylogeny go unremarked by copy number events, and the vast majority of phylogenic branches of the stroma do not carry a unique, detectable copy difference. Therefore, to enhance our ability to identify unique subclones in both the tumor and stroma, we are working on smaller genomic changes such as single nucleotide variants (SNVs) [58–61], small insertions and deletions (indels), and microsatellite length variations (MSLVs). Both our previous single-omic DNA/RNA BAG-seq and the current hybrid BAG-seq are based on primer hybridization and extension. This foundation enables a natural extension of BAG-seq as a target-specific capture platform for enriching tumor or germline variants, whether located on chromosomal or mitochondrial sequences.

These developments promise to provide detailed insights into tumor phylogeny and could positively affect the DNA capture yield.

Although the DNA analyses in this study were mostly based on coverage-derived copy number information, we also performed loss-of-heterozygosity (LoH) analyses using germline SNPs from matched blood whole-genome sequencing. These SNP-based analyses were used to validate cluster differences inferred from copy number variation and also to provide additional phylogenetic information. In the same vein, this DNA-RNA paired dataset also presents an opportunity to explore allele-specific expression (ASE) in the RNA layer, which could potentially help resolve subclones with similar copy number profiles but different allelic gene expression patterns. A transcriptome-wide ASE analysis is for future studies, but we have included a proof-of-concept example by analyzing RNA-layer data from nuclei of the A_2 and B_2 DNA clones, both projecting to the same RNA cluster Ta_2. By examining all genes on chromosome 10 with sufficient coverage, we observed significant and consistent allele ratio differences between the A_2 and B_2 aggregates (Additional file 3: Fig. S20), echoing the LoH patterns identified in the DNA-only data (Additional file 3: Fig. S11) and supporting the potential of RNA-layer allele analysis to further refine subclonal architecture and DNA-to-RNA cluster projections.

Building on its capacity to detect genomic mutations, future versions of the BAG platform could also be adapted to incorporate epigenetic features. For example, transposons modified with Acrydite could be incorporated via transposase to access open chromatin regions, which can be covalently linked to BAGs. This linkage of original genomic DNA molecules to BAGs would also allow for parallel genomic variation and epigenetic DNA methylation analyses.

The RNAscope assay on X chromosome loss in this study briefly touches on the potential of using spatial transcriptomics to image stromal mutations, which are difficult to validate or infer using single-omic sequencing techniques. In the spatial context, hybrid BAG-seq holds even greater promise. With the advancement of spatial platforms like Xenium, Visium, and MERSCOPE that can resolve hundreds of genes across millions of cells, hybrid BAGs can guide the selection of informative genes that best represent tumor subclones, somatic variations, and even stromal mutations, thereby empowering spatial transcriptomics assays for genomic lineage tracing and mutation-informed mapping.

Conclusions

RNA-only data cannot definitively distinguish a stromal expression state from that of the tumor or tumor precursors. Identifying which expression patterns belong to stromal cells and which belong to tumor cells becomes particularly challenging when stromal populations exhibit atypical expression or when tumor cells mimic normal cells. Moreover, there is little hope of connecting tumor genomic changes marked by DNA clonality with changes in expression states marked by RNA clusters.

To close this gap in our understanding requires integrating single-cell DNA and RNA together from the same nuclei, ideally with thousands of nuclei per tissue sample. We achieved this by leveraging our BAG (balls of acrylamide gel) platform. The BAG platform uses a microfluidics oil-emulsion technique to co-encapsulate single nuclei with Acrydite-modified primers and acrylamide monomers into a droplet. The primers

hybridize to the cellular nucleic acids, which are subsequently copolymerized to convert the droplet into a permeable gel matrix or BAG. Depending on primer design, hybridization conditions, and polymerases, BAGs can capture either DNA or RNA alone, as we have previously shown.

In this paper, we introduce a new method that captures DNA and RNA together in hybrid BAGs. We demonstrate that hybrid BAGs provide sequence data that are consistent with those obtained from DNA-only or RNA-only BAGs. The hybrid data are sufficient to establish the same genomic and expression identities identified by the other two methods. In the five patients we examined, we observed comparable and consistent clusters (Fig. 2, Additional file 3: Figs. S6–S9). This includes even subtle features, such as a single 43 MB deletion that distinguished two tumor subclones from patient 1, and loss of the X chromosome in some diploid cells.

Having established the effectiveness of the hybrid method for each modality, we then explored the interplay between the DNA and RNA layers. We found that the combination of expression and genomic data from hybrid BAGs clarified the subpopulation composition of the primary site of tumors. We used several standard tools from the Seurat package, including cell type clustering, subset selection, and UMAP visualizations. We also developed additional modules: (1) using Seurat to cluster the DNA copy number data; (2) alluvial plots to visualize the mappings of genomic clusters into expression clusters; (3) multinomial methods to determine cell cluster membership and to filter collisions; (4) establishing distance metrics relating clusters; and (5) cluster back propagation, which uses information from the DNA-layer to divide RNA clusters, and the reverse.

By examining the alluvial plots in Fig. 3, we note varied patterns of mapping between the tumor DNA clones and the RNA expression clusters. In patient 1, both A_1 and B_1 map equally to the same two expression clusters and in similar proportions, strongly suggesting that the phenotypic flexibility of the tumor was preserved through the single-cell bottleneck event that generated the B_1 population. The tumor of patient 4 presents a more clinically advanced uterine carcinosarcoma. Its two tumor clonal populations both map to all three expression states, again suggesting that this diversity was preserved during tumor evolution. However, their proportional contributions to each expression state differ significantly. In contrast, patient 2 and patient 5 present a different story, with tumor cells demonstrating a marked dependency between genomic identity and RNA expression.

In cases where different tumor DNA clones project to the same RNA cluster, we use a process unique to hybrid data that we call back-propagation. In back-propagation, we use the DNA-layer identities to look for additional substructures in the RNA-layer expression states (and vice versa) and then test those groups for significant differences in expression. Subclustering obtained in this way usually does not emerge when clustering RNA-layer alone, or it does not appear at the same hierarchical level as other major tumor clusters, but the genes detected through this approach may have important implications for tumor evolution and could serve as lineage-specific expression markers.

In the combined data from all samples, we identified 23 distinct stromal expression clusters. The majority of these clusters (21 of 23) include nuclei from more than one patient, suggesting that many stromal cell types have consistent expression patterns

across individuals. In contrast, every tumor cluster occurs in one and only one patient with little in common between the tumor clusters from different patients.

By analyzing highly expressed genes, we accurately determined the likely cell type or state for most of these clusters. Our analysis revealed two stromal expression clusters unique to patient 1: a fibroblast population in the distal normal tissue with activation of the AP-1 pathway, and an epithelial population in the tumor tissue with unique makers for signaling and response to stimuli. This epithelial population was distinct in both the UMAP and the multinomial tree. Without the DNA-layer from the hybrid data, we could not have confidently concluded that this unique cluster derives from diploid cells.

While most of the diploid cells are copy number 2 everywhere (labeled as N), in every patient we identified a subset of diploid nuclei that exhibited the loss of an X chromosome (labeled Nx). While this phenomenon has been previously observed in both tumors and stromal populations, the hybrid BAGs provide new insights into their role in tumor biology. In patient 2, we identified an X-loss event in a significant proportion of the plasma cell and T-cell populations. Using the RNA data and the DNA data together, we determined that each plasma cell lost the same X chromosome, whereas the T-cells lost either one or the other equally. This strongly suggested that the plasma cell population may have a common origin. Comparing the expression profiles of *plasma-N* and *plasma-Nx*, we observed reduced expression of *XIST* and *TSIX*, in the Nx population, further confirming that the X-loss observed in the DNA layer is not a sequencing artifact but a genuine genomic event. Furthermore, the differential expression analysis showed differences between the *IGH1* and *IGH2* genes. This led us to explore the VDJ recombination region, providing decisive evidence of a clonal origin for this mutant plasma cell population.

Methods

Pulverization of frozen tissue samples in liquid nitrogen

All patient tissue 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 tissues.

The sample cohort

We studied samples from five patients (patient 1–patient 5). The samples were from their uterine cancers (Tumor 1–Tumor 5), and in three patients also from an adjacent normal endometrial site (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

A detailed step-by-step bench protocol is archived at protocols.io [62], and in this section we highlight some critical steps in hybrid BAG generation. We dissolved the pulverized tissue in ice-cold NST detergent buffer [42] and stained it 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 [42]. Single nuclei were loaded into the microfluidic device described in detail in a previous publication [35]. 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 copolymerized 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 [35]. 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 min instead of 95 °C for 12 min 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 introduced by ligation extension, as described in the step-by-step hybrid BAG-seq experimental protocol [62], leaving a common 3′ sequence identical across all 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. 1C, 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 $96 \times 96 \times 96$ 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 ("AAAAAA AAAAAAAA") 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 $96 \times 96 \times 96$ ($\approx$ 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} ($\approx$ 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 [63]. 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 indexed using samtools. In subsequent data analysis steps, we designate as mapped reads those that HISAT2 identifies as part of a proper pair, with a primary mapping and a 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, which we will discuss in the next section.

Refinement of copy number estimation methods

We demonstrated four progressively refined versions of copy number estimation, named "all_molecules," "no_exon," "no_gene," and "no_gene.avoid50closeTN" (Additional file 3: Fig. S1). The "all_molecules" method simply used all molecules from each retained nucleus for copy number estimation, as the name implies. The "no_exon" method used only molecules classified as "Intronic" or "Intergenic," as described in the previous paragraph. The "no_gene" method further restricted the data to only "Intergenic" templates, excluding any molecules with bases overlapping annotated transcripts. For each sample, we first downsampled the DNA-only reads using a per-cell binomial draw so that the total number of DNA molecules matched those of the hybrid library. Cells with fewer molecules than the hybrid depth were left unchanged.

When comparing copy number data derived from intergenic templates in the hybrid protocol to those in the RNA-only protocol, we noticed that certain high-count genomic bins in the RNA-only data also appeared in the hybrid protocol. To investigate this, we aggregated all intergenic templates from seven RNA-only BAG-seq libraries (from two normal and five tumor tissues), sorted them by genomic position, and measured the distances between consecutive templates. Histograms of these distances revealed that the majority ($>77\%$) of adjacent intergenic RNA-only molecules were within 50 bp (Additional file 3: Fig. S21). Therefore, we defined RNA hotspots as genomic stretches where all consecutive inter-molecular distances are ≤ 50 bp after sorting by genomic position. We later found that these hotspot regions often contain poly-A (or poly-T) genomic DNA sequences. Since these hotspots distorted copy number profiles in both normal and tumor specimens, we eliminated them from downstream analysis. This filtering step

improved copy number quality, as demonstrated using normal tissue samples (Additional file 3: Fig. S1).

The final method, "no_gene.avoid50closeTN," retained only the "Intergenic" templates from the "no_gene" method that were located at least 50 bases away from RNA hotspots. We observed that each successive filtering criterion progressively improved copy number quality. To quantify this improvement, we introduced a "terrain" metric to measure coverage uniformity across the genome. It is defined as the sum of absolute differences between adjacent bins; lower terrain values indicate smoother and more uniform copy number profiles. As shown in Additional file 3: Fig. S1, the DNA-only libraries consistently exhibit lower terrain scores than hybrid libraries, but the difference diminishes as more stringent filters are applied—from "all_molecules," to "no_exon," to "no_gene," and finally to "no_gene.avoid50closeTN." This final criterion, "no_gene.avoid50closeTN," was used to define the DNA-layer molecules of the hybrid BAG-seq protocol and was applied in all downstream analyses throughout this study.

Empirical binning strategy for copy number analysis

Separately for each of the four copy number molecule selection methods described above, we used the genomic positions of all molecules from two normal-tissue DNA samples to determine empirical bin boundaries for 300 genomic bins, each containing approximately equal molecule counts. Excluding any molecules mapped to chromosome Y, we assigned to chromosomes 1–22 and X a number of bins proportional to their fraction of the total molecule count. Within each chromosome, bin boundaries were assigned greedily from the chromosome start, such that all but the final bin contained at least the number of molecules equal to the chromosome-specific total molecule count divided by its assigned number of bins. The observed count of molecules per bin was recorded as a normalization factor for later use during per-sample copy number estimation. This factor accounts for residual bin-to-bin differences resulting from unequal molecule distributions and rounding of bin assignments per chromosome. Because a small total number of bins (300) were used, perfect proportional allocation was not possible, and normalization factors could vary by up to 30% between chromosomes.

Additional file 11: Table S9 provides the details of the 300 empirical bins used for hybrid protocol DNA-layer copy number analysis. This table includes the start and end genomic positions of each bin, bin length, GC content, and the actual number of intergenic molecules from the two normal samples ("n_good") used to compute per-bin normalization weights in single-cell copy number analysis. We used the following formula to transform the raw per-cell bin count vector into a normalized bin count vector, without changing the sum of total bin counts for each cell. Applying this transformation does not affect clustering results but makes the bin counts of normal cells more uniform and improves the visual appearance of the copy-number heatmaps. Let us define:

$\boldsymbol{raw_{cb}}$: raw count of molecules in cell $\boldsymbol{c}$, bin $\boldsymbol{b}$;

$\boldsymbol{n_b}$: number of observed intergenic molecules ("n_good" from Additional file 11: Table S9) in bin $\boldsymbol{b}$;

$\mathrm{Tc} = \sum_b \frac{\boldsymbol{raw_{cb}}}{n_b}$: total normalized weight for cell $\boldsymbol{c}$;

$\mathrm{Nc} = \sum_b raw_{cb}$: total raw molecules in cell $\boldsymbol{c}$;

Then, the normalized value is $\boldsymbol{norm_{cb}} = \left(\frac{\boldsymbol{raw_{cb}}}{\boldsymbol{n_b}}\right) \bullet \frac{\boldsymbol{N_c}}{\boldsymbol{T_c}}$.

Segmentation and visualization of copy number profiles

Copy number segmentation was used only to generate visual copy number profiles for population-based plotting, such as in Fig. 2B, Additional file 3: Figs. S6–S9 (panel B), and Additional file 3: Fig. S1. We aggregated nuclei from a cluster and used the empirical bin boundary information specific to the platform and filtering criteria described above. Circular binary segmentation (CBS) was then performed using the DNAcopy R package (version 1.50.1) [64]. Prior to segmentation, we applied LOESS-based GC-content normalization to the binned template count ratios. Segmentation was run using the function cbs.segment.uber01.0_3k (included in our GitHub codebase indicated below), with parameters alpha = 0.1, nperm = 1000, undo.SD = 0.25, and min.width = 2. The segmented copy number values were scaled by a multiplier selected to minimize the total squared difference between the scaled values and their nearest integers, thereby aligning the segmentation output with approximate integer copy number for improved interpretability.

RNA clustering

The RNA clustering was performed using Seurat package (version 3.1.5) following the standard Seurat clustering pipeline [43]. 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-only protocol, we normalized the gene-template matrix by cell and excluded the PCA components that most significantly distinguished protocol differences. We typically 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 5000 variable gene features for PCA analysis and used the first 50 PCA components for subsequent UMAP and FindCluster functions. The detailed Seurat parameters and R code have been uploaded to GitHub and can be found at: https://github.com/siranli01/DNA_RNA.

RNAScope imaging analysis

RNAScope Multiplex Fluorescent Reagent Kit v2 (Advanced Cell Diagnostics) was employed to visualize RNA transcripts within tissue sections. Formalin-fixed, paraffin-embedded (FFPE) tissue slides from Tumor 2 were prepared according to the manufacturer's instructions. Following deparaffinization and rehydration, sections underwent hydrogen peroxide treatment and target retrieval. Protease treatment was then applied to facilitate probe penetration.

The following RNAScope probes were used for target detection: IGHG-pool (Cat No. 481901), which targets IGHG (1–4) with 11–19 ZZ pairs, and human XIST (Cat No. 311231-C2), which targets the XIST RNA transcript. Signal amplification was performed using the RNAScope Multiplex Fluorescent Reagent Kit v2. The TSA Vivid Dyes were used for fluorescent signal development: IGHG was visualized using TSA Vivid

Fluorophore 520 dye (Cat No. 323271), and XIST was visualized using TSA Vivid Fluorophore 570 dye (Cat No. 323272). Amplification steps followed the standard protocol provided by Advanced Cell Diagnostics, ensuring optimal signal-to-noise ratio.

Imaging was conducted on a Nikon Ti spinning-disk confocal microscope equipped with a YOKOGAWA spinning-disk system and controlled by Nikon Elements software AR 5.42.04. Fluorophores were excited with the following laser lines: DAPI at 405 nm, IGHG (TSA Vivid Fluorophore 520) at 488 nm, and XIST (TSA Vivid Fluorophore 570) at 561 nm. Fluorescent signals for DAPI, IGHG, and XIST were pseudo-colored for visualization.

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, but 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 distributions

For a cluster X, we sum the RNA template counts over each gene for all cells in the cluster. If a gene contains zero counts over the population of cells, we assign a value of ½, to avoid zero probabilities when comparing to a cell that containing a gene unobserved in cluster X. We normalize by the count vector by its total to obtain a probability distribution over the set of genes. We compute this probability vector for each stromal expression cluster as determined by the Seurat iterative clustering of the stromal cell types, for each tumor cluster as determined in the individual tumor RNA clusters, and for an

"empty" cluster composed of the RNA-layer from 500 DNA-only BAGs. The expression multinomial distributions are used for two purposes:

1. Cluster assignment. We assign each nucleus to the distribution with the maximum likelihood of generating its observed counts, assuming a uniform prior on the space of clusters. These identities determine the color of the points in Fig. 4A and the counts in Table 1.
2. Inter-cluster dissimilarity. We compute a pairwise dissimilarity between clusters using the multinomial distributions induced by the cluster average. For any two clusters X and Y, and template count t between 1 and 300, we compute by simulation (100,000 samples per data point) the posterior probability that a single nucleus generated from the multinomial distribution on X and a count of t templates comes from X, conditioned on the prior probability of originating from X or Y with equal probability. For each value of t, we compute the AUROC (equivalent to the average posterior probability). To define a measure of how likely X is correctly identified against Y, we compute M(X, Y), the smallest value of t for which the AUROC exceeds 0.999 (Additional file 3: Fig. S17). To convert this matrix into a pairwise dissimilarity between clusters, we compute a similarity measure S(X, Y) = ½ * [M(X, Y) + M(Y, X)] and invert this to obtain a dissimilarity measure D(X, Y) = 300 − S(X, Y). We then apply the neighbor-joining algorithm to the dissimilarity measure to obtain a tree (Fig. 4B).

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 $p = (0.1, 0.2, \ldots, 0.9)$, 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,$i = 1,2,\ldots,n$), 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 Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03790-5.

Additional file 1: Table S1 Patients overview.

Additional file 2: Table S2 BAG library parameters.

Additional file 3: Figures S1–S21.

Additional file 4: Table S3 Contingency matrices.

Additional file 5: Supplementary text.

Additional file 6: Table S4 Tumor subcluster expression.

Additional file 7: Table S5 Stromal marker genes.

SAdditional file 8: Table S6 Crossover analysis.

Additional file 9: Table S7 Crossover summary, effects of filtering.

Additional file 10: Table S8 Co-clustering all (filtered for collisions).

Additional file 11: Table S9 Empirical bin boundaries for hybrid protocol.

Acknowledgements

We thank P. Moody for cell sorting; J. Preall, C. Regan, and E. Zhang in CSHL single-cell core facility for assistance; Q. Gao for histology assistance; A. Runnels from NYGC and E. Ghiban from CSHL for Illumina sequencing assistance; M. Yao, S. Kleeman, D. Fearon, T. Janowitz, J. Boyd, and D. Tuveson from CSHL and N. Chiorazzi from Northwell for helpful discussion; and A. Kapedani, M. A. Green, and S. Chin from the Clinical Research Team in the Department of Obstetrics and Gynecology of Northwell 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, and 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.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions

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

Funding

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.).

Simons Foundation,519054,Breast Cancer Research Foundation

Data availability

Illumina sequencing data for all the single-nucleus libraries are available at NCBI Sequencing Read Archive (SRA) with accession code (PRJNA773107) [65].

An end-to-end pipeline for converting BAG paired-end FASTQ files into gene-resolved read tables suitable for repertoire analysis and quality control is available at GitHub (https://github.com/levycshl/bag-pipe) [66] and archived on Zenodo (https://doi.org/10.5281/zenodo.17048209) [67] under an MIT License.

R scripts for downstream clustering, multinomial wheel analyses, and related analyses are available at GitHub (https://github.com/siranli01/DNA_RNA) [68] and archived on Zenodo (https://doi.org/10.5281/zenodo.17051405) [69] under an MIT License.

Declarations

Ethics approval and consent to participate

Women presenting to the Department of Obstetrics and Gynecology at the Long Island Jewish Medical Center (LIJMC) at Northwell Health with a clinical suspicion of ovarian or endometrial cancer (type I or type II) were prospectively enrolled between October 2018 and March 2021 as part of an ongoing study of endometrial and ovarian cancers. The hybrid BAG-seq method was applied to blood and tissue biopsies collected from five patients (P1–P5) who underwent total hysterectomy and bilateral salpingo-oophorectomy at LIJMC. All patients provided individual informed consent. The study (TAP#1805) was approved by the Northwell Health Biospecimen Repository Committee on Tissue Governance and Protocol Support. Demographic and clinical data related to uterine cancer diagnoses were collected and maintained in the Department of Obstetrics and Gynecology database at Northwell Health. All procedures were performed in accordance with the Declaration of Helsinki.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 14 September 2024 Accepted: 14 September 2025
Published online: 27 September 2025

References

1. Lan F, Demaree B, Ahmed N, Abate AR. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. Nat Biotechnol. 2017;35(7):640–6.
2. Vitak SA, Torkenczy KA, Rosenkrantz JL, Fields AJ, Christiansen L, Wong MH, et al. Sequencing thousands of single-cell genomes with combinatorial indexing. Nat Methods. 2017;14(3):302–8.
3. Andor N, Lau BT, Catalanotti C, Sathe A, Kubit M, Chen J, et al. Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines reveals rules of in vitro evolution. NAR genomics and bioinformatics. 2020;2(2):lqaa016.
4. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161(5):1202–14.
5. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187–201.
6. Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods. 2017;14(4):395–8.
7. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017;357(6352):661–7.
8. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018;360(6385):176–82.
9. Habib N, Li Y, Heidenreich M, Swiech L, Avraham-Davidi I, Trombetta JJ, et al. Div-seq: single-nucleus RNA-seq reveals dynamics of rare adult newborn neurons. Science. 2016;353(6302):925–8.
10. Gao R, Kim C, Sei E, Foukakis T, Crosetto N, Chan L-K, et al. Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. Nat Commun. 2017;8(1):1–12.
11. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8(1):1–12.
12. Dey SS, Kester L, Spanjaard B, Bienko M, Van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. Nat Biotechnol. 2015;33(3):285–9.
13. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015;12(6):519–22.
14. Han KY, Kim K-T, Joung J-G, Son D-S, Kim YJ, Jo A, et al. SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. Genome Res. 2018;28(1):75–87.
15. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. 2016;26(3):304–19.
16. Bian S, Hou Y, Zhou X, Li X, Yong J, Wang Y, et al. Single-cell multiomics sequencing and analyses of human colorectal cancer. Science. 2018;362(6418):1060–3.
17. Han L, Zi X, Garmire LX, Wu Y, Weissman SM, Pan X, et al. Co-detection and sequencing of genes and transcripts from the same single cells facilitated by a microfluidics platform. Sci Rep. 2014;4(1):1–9.
18. Van Strijp D, Vulders R, Larsen N, Schira J, Baerlocher L, Van Driel M, et al. Complete sequence-based pathway analysis by differential on-chip DNA and RNA extraction from a single cell. Sci Rep. 2017;7(1):1–9.
19. Kong SL, Li H, Tai JA, Courtois ET, Poh HM, Lau DP, et al. Concurrent single-cell RNA and targeted DNA sequencing on an automated platform for comeasurement of genomic and transcriptomic signatures. Clin Chem. 2019;65(2):272–81.
20. Cheow LF, Courtois ET, Tan Y, Viswanathan R, Xing Q, Tan RZ, et al. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. Nat Methods. 2016;13(10):833–6.
21. Rodriguez-Meira A, Buck G, Clark S-A, Povinelli BJ, Alcolea V, Louka E, et al. Unravelling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel RNA sequencing. Molecular cell. 2019;73(6):1292–305. e8.
22. Li W, Calder RB, Mar JC, Vijg J. Single-cell transcriptogenomics reveals transcriptional exclusion of ENU-mutated alleles. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2015;772:55–62.
23. Yu L, Wang X, Mu Q, Tam SST, Loi DSC, Chan AK, et al. Scone-seq: a single-cell multi-omics method enables simultaneous dissection of phenotype and genotype heterogeneity from frozen tumors. Sci Adv. 2023;9(1):eabp8901.
24. Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet. 2023;24(8):494–515.
25. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol. 2021;39(5):599–608.
26. Fan J, Lee H-O, Lee S, Ryu D-e, Lee S, Xue C, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome Res. 2018;28(8):1217–27.
27. Elyanow R, Zeira R, Land M, Raphael BJ. STARCH: copy number and clone inference from spatial transcriptomics data. Phys Biol. 2021;18(3):035001.

28. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):1396–401.
29. Erickson A, He M, Berglund E, Marklund M, Mirzazadeh R, Schultz N, et al. Spatially resolved clonal copy number alterations in benign and malignant tissue. Nature. 2022;608(7922):360–7.
30. Ronemus M, Bradford D, Laster Z, Li S. Exploring genome-transcriptome correlations in cancer. Biochem Soc Trans. 2025;53(01):113–20.
31. Yin Y, Jiang Y, Lam K-WG, Berletch JB, Disteche CM, Noble WS, et al. High-throughput single-cell sequencing with linear amplification. Molecular cell. 2019;76(4):676–90. e10.
32. Olsen TR, Talla P, Sagatelian RK, Furnari J, Bruce JN, Canoll P, et al. Scalable co-sequencing of RNA and DNA from individual nuclei. Nat Methods. 2025;22(3):477–87.
33. Otoničar J, Lazareva O, Mallm J-P, Simovic-Lorenz M, Philippos G, Sant P, et al. HIPSD&r-seq enables scalable genomic copy number and transcriptome profiling. Genome Biol. 2024;25(1):1–22.
34. Li Y, Huang Z, Xu L, Fan Y, Ping J, Li G, et al. UDA-seq: universal droplet microfluidics-based combinatorial indexing for massive-scale multimodal single-cell sequencing. Nat Methods. 2025;22(6):1199–212.
35. Li S, Kendall J, Park S, Wang Z, Alexander J, Moffitt A, et al. Copolymerization of single-cell nucleic acids into balls of acrylamide gel. Genome Res. 2020;30(1):49–61.
36. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(Nov):2579–605.
37. McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018.
38. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IW, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38–44.
39. Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model. Genome Biol. 2019;20:1–16.
40. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015;348(6233):aaa6090.
41. Janesick A, Shelansky R, Gottscho AD, Wagner F, Williams SR, Rouault M, et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. Nat Commun. 2023;14(1):8353.
42. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472(7341):90–4.
43. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck III WM, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–902. e21.
44. Xi NM, Li JJ. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. Cell systems. 2021;12(2):176–94. e6.
45. Spatz A, Borg C, Feunteun J. X-chromosome genetics and human cancer. Nat Rev Cancer. 2004;4(8):617–29.
46. Zhou Y, Bian S, Zhou X, Cui Y, Wang W, Wen L, et al. Single-cell multiomics sequencing reveals prevalent genomic alterations in tumor stromal cells of human colorectal cancer. Cancer cell. 2020;38(6):818–28. e5.
47. Machiela MJ, Zhou W, Karlins E, Sampson JN, Freedman ND, Yang Q, et al. Female chromosome X mosaicism is age-related and preferentially affects the inactivated X chromosome. Nat Commun. 2016;7(1):11843.
48. Liu A, Genovese G, Zhao Y, Pirinen M, Zekavat SM, Kentistou KA, et al. Genetic drivers and cellular selection of female mosaic X chromosome loss. Nature. 2024;631(8019):134–41.
49. Panning B, Dausman J, Jaenisch R. X chromosome inactivation is mediated by Xist RNA stabilization. Cell. 1997;90(5):907–16.
50. Weakley SM, Wang H, Yao Q, Chen C. Expression and function of a large non-coding RNA gene XIST in human cancer. World J Surg. 2011;35:1751–6.
51. Huang K-C, Rao PH, Lau CC, Heard E, Ng S-K, Brown C, et al. Relationship of XIST expression and responses of ovarian cancer to chemotherapy. Mol Cancer Ther. 2002;1(10):769–76.
52. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods. 2015;12(5):380–1.
53. Segovia C, Desrosiers V, Khadangi F, Robitaille K, Armero VS, D'Astous M, et al. A versatile and efficient method to isolate nuclei from low-input cryopreserved tissues for single-nuclei transcriptomics. Sci Rep. 2025;15(1):5581.
54. Nadelmann ER, Gorham JM, Reichart D, Delaughter DM, Wakimoto H, Lindberg EL, et al. Isolation of nuclei from mammalian cells and tissues for single-nucleus molecular profiling. Curr Protoc. 2021;1(5):e132.
55. Slyper M, Porter CB, Ashenberg O, Waldman J, Drokhlyansky E, Wakiro I, et al. A single-cell and single-nucleus RNA-seq toolbox for fresh and frozen human tumors. Nat Med. 2020;26(5):792–802.
56. Masilionis I, Chaudhary O, Urben BM, Chaligne R. Nuclei extraction for 10x Genomics Single Cell Multiome ATAC+ Gene Expression from frozen tissue using Singulator™ 100 or 200 (S2 Genomics) V2. 0. 2023.
57. Krasnitz A, Kendall J, Alexander J, Levy D, Wigler M. Early detection of cancer in blood using single-cell analysis: a proposal. Trends Mol Med. 2017;23(7):594–603.
58. Nam AS, Kim K-T, Chaligne R, Izzo F, Ang C, Taylor J, et al. Somatic mutations and cell identity linked by genotyping of transcriptomes. Nature. 2019;571(7765):355–60.
59. Penter L, Borji M, Nagler A, Lyu H, Lu WS, Cieri N, et al. Integrative genotyping of cancer and immune phenotypes by long-read sequencing. Nat Commun. 2024;15(1):32.
60. Lareau CA, Ludwig LS, Muus C, Gohil SH, Zhao T, Chiang Z, et al. Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. Nat Biotechnol. 2021;39(4):451–61.
61. Lindenhofer D, Bauman JR, Hawkins JA, Fitzgerald D, Yildiz U, Marttinen JM, et al. Functional phenotyping of genomic variants using multiomic scDNA-scRNA-seq. bioRxiv. 2024:2024.05. 31.596895.
62. Bradford D, Li S. Hybrid BAG-Seq protocol for DNA-RNA co-assay. protocols.io. 2025.
63. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15.
64. Olshen AB, Venkatraman E, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5(4):557–72.

65. Li S. PRJNA773107. Datasets. Sequence Read Archive (SRA). 2021. https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA773107.
66. Levy D. levycshl/bag-pipe. GitHub. 2025. https://github.com/levycshl/bag-pipe.
67. Levy D. levycshl/bag-pipe. 2025. Zenodo. https://doi.org/10.5281/zenodo.17048209.
68. Li S. siranli01/DNA_RNA. GitHub. 2025. https://github.com/siranli01/DNA_RNA.
69. Li S, Hybrid BAG. R scripts Zenodo. 2025. https://doi.org/10.5281/zenodo.17051405.

Publisher's Note