SYSTEMS BIOLOGY

A universal sequencing read interpreter

Yusuke Kijima[1,2,3], Daniel Evans-Yamamoto[2,4], Hiromi Toyoshima[2], Nozomu Yachie[1,2,5]*

Massively parallel DNA sequencing has led to the rapid growth of highly multiplexed experiments in biology. These experiments produce unique sequencing results that require specific analysis pipelines to decode highly structured reads. However, no versatile framework that interprets sequencing reads to extract their encoded information for downstream biological analysis has been developed. Here, we report INTERSTELLAR (interpretation, scalable transformation, and emulation of large-scale sequencing reads) that decodes data values encoded in theoretically any type of sequencing read and translates them into sequencing reads of another structure of choice. We demonstrated that INTERSTELLAR successfully extracted information from a range of short- and long-read sequencing reads and translated those of single-cell (sc)RNA-seq, scATAC-seq, and spatial transcriptomics to be analyzed by different software tools that have been developed for conceptually the same types of experiments. INTERSTELLAR will greatly facilitate the development of sequencing-based experiments and sharing of data analysis pipelines.

INTRODUCTION

In the last couple of decades, harnessing microarray and high-throughput DNA sequencing, the concept of DNA barcodes has enabled a range of pooled biological screens. Earlier examples include the establishment of the yeast deletion collection, where each strain was constructed to have two unique DNA barcodes at a deletion locus (*1*). The barcoded yeast strains can be pooled and subjected to a single growth competition assay whose individual relative growth changes can be read out by barcode quantities measured by microarray or high-throughput sequencing before and after the competition (*2*). This strategy pioneered the field of chemical genomics to screen drug target genes (*3*, *4*). Soon after, the same concept was also applied to mammalian cell culture–based genome-wide gene knockdown (*5*) and knockout assays (*6*, *7*). In these assays, cells are transduced by a lentiviral library encoding short-hairpin (sh)RNAs or CRISPR-Cas9 guide (g)RNAs. Cell growths conferred by different perturbations can be massively quantified by polymerase chain reaction (PCR) amplification and sequencing of the small shRNA- or gRNA-encoding DNA fragments. Furthermore, experimental systems that produce chimeric fusions of distal genomic regions and those of DNA barcodes associated with different factors have enabled the exploration of chromatin conformations (*8*), protein interactions (*9*–*12*), genetic interactions (*13*), and spatial cellular distribution of single-molecule RNAs (*14*) in large scale. In single-cell and spatial genomics, single-cell identifiers (IDs), spatial IDs, and unique molecular IDs (UMIs) are used to uniquely tag corresponding transcriptomes or genomic DNA fragments, which led to the development of single-cell RNA sequencing (scRNA-seq) (*15*–*18*), scATAC-seq (*19*, *20*), spatial transcriptomics (*21*, *22*), and spatial genomic (*23*) technologies. The above-mentioned methods each enable multiplexing of a number of experiments at once and produce a sequencing library. Sequencing libraries from different assays can also be further multiplexed for a single sequencing run by fusing additional library-specific, unique DNA barcode(s) to each sequencing library DNA. The output DNA molecules of these experiments have a range of complexities, some of which encode multiple information segments whose combinations are sometimes designed to be read out by multiple reads (e.g., paired-end reads and index reads).

However, there have been common issues—most of these sequencing-based experiments have been developed with their own proprietary software tools for specific sequence read structures. While many of such tools have advanced downstream data analysis capabilities, they often cannot be reused even for sequencing reads produced by conceptually the same types of experimental systems. New experimental methods have been repeatedly proposed for conceptually identical analyses with improved performances and different read structures, and data analysis tools that process essentially the same information have been developed for their respective read structures. These reinventions of the wheel have particularly been observed in the scRNA-seq field (*24*). These software tools cannot be exchanged for different scRNA-seq library structures or cross-validated by applying them to the same scRNA-seq dataset. Several efforts have been made to develop flexible software tools that are capable of analyzing different read structures of a certain category of experiments, such as UMI-tools (*25*), zUMIs (*26*), scumi (for UMI-based RNA-seq and scRNA-seq) (*27*), and SnapATAC (for scATAC-seq) (*28*), yet they are not effective for the ongoing development of new experiments that produce unique read structures.

Any sequencing data analyses follow identification of sequence segments in each read (e.g., identification of cell ID, UMI, and cDNA-encoding regions in scRNA-seq reads) and downstream analyses of the extracted sequence segments and values (e.g., mapping to the reference genome and UMI counting of each RNA species in scRNA-seq). Therefore, we propose two solutions to the community: (i) The development of sequencing read interpreters and data analysis tools separately—if a read interpreter only extracts data values encoded in sequencing reads, then its data analysis pipeline should be applicable for sequencing reads of other experiments that produce the same data structures; and (ii) the

[1]School of Biomedical Engineering, Faculty of Applied Science and Faculty of Medicine, The University of British Columbia, Vancouver, BC V6T 1Z3, Canada. [2]Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo 153-8904, Japan. [3]Department of Aquatic Bioscience, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo 113-8657, Japan. [4]Institute for Advanced Biosciences, Keio University, Tsuruoka 997-0035, Japan. [5]Twitter: @yachielab.
*Corresponding author. Email: nozomu.yachie@ubc.ca

development of a read translator—if sequencing reads of a certain format could be translated into another read structure, the existing data analysis pipelines developed for the specific read structure could be used to analyze other read structures. In this study, we have identified that these two ideas can be achieved by a single universal tool, which we have developed and called INTERSTELLAR (interpretation, scalable transformation, and emulation of large-scale sequencing reads).

RESULTS

Overview of INTERSTELLAR

INTERSTELLAR interprets high-throughput sequencing reads and translates them into sequencing reads of another read structure (Fig. 1). The flexible sequence segment identification of "source reads" enables economic development of their downstream data analysis pipelines. A user-defined set of extracted sequence segments can then be converted into "values," according to how the user defined the read interpretation, which can further be used to translate the source reads into "destination reads." This read translation enables current data analysis tools that originally only accept a specific read structure to analyze sequencing reads of another structure.

INTERSTELLAR first identifies sequence segments of reads in FASTQ files according to the user's definition provided in a process configuration file (step 1). The source read structure can be defined flexibly, where multiple segments on each sequencing read are specified by combinations of their lengths, locations, and neighboring sequence motifs using regular expression, followed by identification of valid sequence segments according to their average sequencing quality (Q) scores. Three types of attributes, "combinatorial," "parental," and "local," can be used to associate multiple sequence segments with each other. A combinatorial segment group can be defined to collectively denote a specific information value. A parental segment (or combinatorial parental segment group) can be paired with an independent set of local segments (or combinatorial local segment groups), where sequence-to-value conversion of the local segment(s) is independently performed for its parental segment. For example, cell IDs and UMIs of typical scRNA-seq reads can be defined as parental segments and their local segments, respectively, where the same UMI sequences associated with different cell IDs are interpreted as different objects. Multiple-source read structures can also be defined for a single set of input sequencing reads that are produced by a one-shot sequencing of different libraries.

The segment identification process can be performed independently for fragmented FASTQ files using distributed computing, where each fragmented process yields segmented sequences and count-up information for each unique segment sequence. The sequence count-up information derived from different fragmented processes is then merged to compute an error-corrected sequence for each unique segment sequence (step 2). INTERSTELLAR

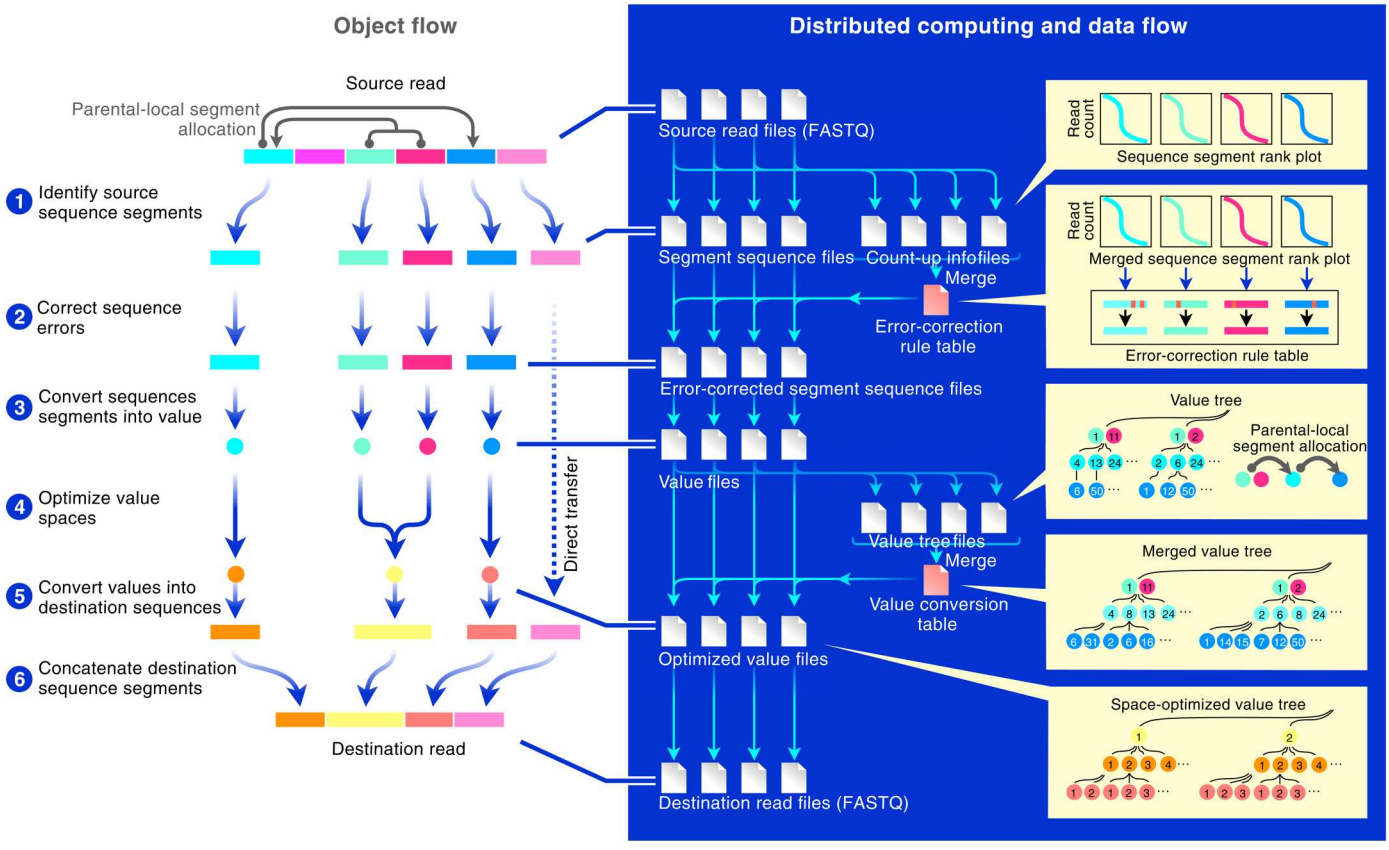

Fig. 1. Overview of INTERSTELLAR. Conceptual diagram representing how INTERSTELLAR (interpretation, scalable transformation, and emulation of large-scale sequencing reads) interprets and translates sequencing reads with its file management and distributed computing strategies.

enables four error correction options: "imputation-to-majority," "mapping-to-allowlist," Bartender (*29*), and a user-developed plugin. In the imputation-to-majority correction, a merged rank-read count curve of each sequence segment is first obtained, and its knee point (the maximum curvature point) is determined. Segment sequences below the knee point are then corrected to their closest similar sequences above the knee point using the Levenshtein distance metric. Similarly, the allowlist mapping uses the Levenshtein distance metric to map input segment sequences to a user-provided allowlist. In these two options, the minor segment sequences (above the Levenshtein distance threshold) are ignored. The barcode sequence correction pipeline Bartender can also be used, where input segment sequences are first grouped into clusters based on the Hamming distance metric, and minor sequences in each cluster are imputed into the top majority sequence. In contrast to the imputation-to-majority strategy, Bartender can potentially rescue valid sequences that are poorly represented in the pool. Alternatively, users can provide a shell script as a user-defined plugin to use a customized error correction method. Once an error correction rule table is generated, it is used to error-correct segment sequences originating from each of the fragmented FASTQ files using distributed computing. The above-mentioned read interpretation process can be applied to any high-throughput sequencing read analysis, and the generated error-corrected segment sequence files enable efficient development of their downstream data analysis pipelines.

If defined in the process configuration file, then the read translation into destination read structures are next processed for the error-corrected source segment sequences (step 3). Destination read structures can be flexibly specified by using International Union of Pure and Applied Chemistry codes and/or allowlists of destination segment sequences. First, using distributed computing, a segment value file and a value tree are generated from each of the error-corrected segment sequence files, where each unique segment sequence is converted into a numerical value, and parental-local segment allocations of unique values and unique combinatorial value groups are represented in a tree structure. The value tree files originating from the fragmented FASTQ files are then passed to a single computing node to generate a merged value tree. Next, the values in the merged value tree are replaced by new values to minimize the number of numerical value species for each variable in a way that they still uniquely maintain the same tree topology (step 4). Obtaining the value conversion rule table that achieved the optimization of the merged value tree, segment value files are

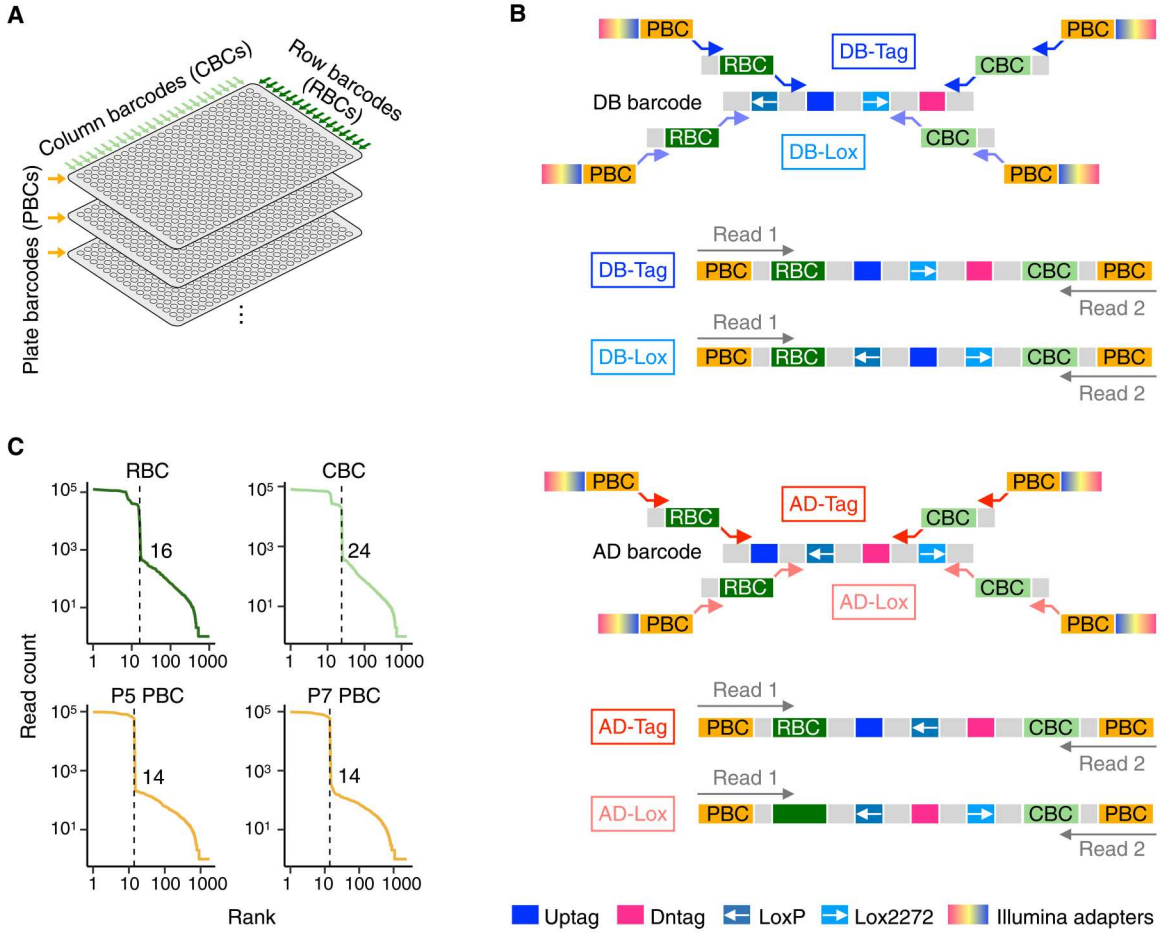

Fig. 2. Interpretation of highly structured RCP-PCR reads. (**A**) The conceptual diagram of row-column-plate polymerase chain reaction (RCP-PCR). (**B**) Two-step PCR amplification and paired-end sequencing of DB and AD barcode cassette libraries. (**C**) Rank-read count plots of row-specific barcodes (RBCs), column-specific barcodes (CBCs), and plate-specific barcodes (PBCs).

separately processed to derive optimized segment value files and then destination FASTQ files by distributed computing, where unique value-to-destination-sequence conversion rules are autonomously generated for the destination read structures (steps 5 and 6). The value space optimization, which takes into account the parental-local segment allocations, is particularly effective when sequence complexities of destination segments are lower (e.g., shorter in length) than those of the corresponding source segments. This process enables a destination segment of less information representativity (versus its source segment) to host all or the maximum possible number of corresponding values represented in the source reads. (When the number of optimized values is over the information representativity of the destination value segment, frequent values are prioritized, and read information associated with any values that are not assigned to a destination segment sequence is ignored.) Throughout the process, the average *Q* scores of source segment sequences are bequeathed from the fragmented FASTQ files through the intermediate segment sequence and value files and given to all letters of corresponding destination segments in the generating FASTQ files. New bases that are not associated with the values inherited from the source reads are all given a *Q* score of 40 in the destination reads. As seen above, the distributed computing process is designed to perform many small conversion tasks in parallel, where the generation of conversion rules that requires monitoring of the entire segment sequence or value space is operated using a single computing node by compressing information from each fragmented task into a small hash table data structure.

Interpretation of highly structured barcode reads

To demonstrate that INTERSTELLAR can be used to analyze highly structured sequencing reads, we first decoded a row-column-plate (RCP)–PCR library generated for massively parallel identification of barcoded plasmid collections used for barcode fusion genetics yeast two-hybrid (BFG-Y2H), an en masse protein interaction technology (*9*). In general, RCP-PCR is developed to identify clonal DNA samples sandwiched by common PCR primer sites that are arrayed into many PCR microwell plates (Fig. 2A). Samples in each microwell plate are first amplified by forward and reverse primers with overhang sequences encoding corresponding plate row-specific barcodes (RBCs) and column-specific barcodes (CBCs), respectively. PCR products are then pooled by plates and subjected to the second round of PCR by primers with overhang sequences encoding sample plate-specific barcodes (PBCs) and Illumina sequencing adapters.

BFG-Y2H uses two types of barcode cassettes, namely, DB-X and AD-Y barcode cassettes (for details, see Materials and Methods). Each type of barcodes encodes site-specific Cre recombination sites, loxP and lox2272, and two barcodes in a different order (Fig. 2B). One of the previously established methods requires subcloning of many barcode cassettes into 384-well plates from a pool of those with degenerated barcode sequences. They are then analyzed by RCP-PCR and high-throughput sequencing for the identification and verification of the isolated barcode cassettes. When using a short-read sequencer whose base calling quality drops along with base calling cycles, two overlapping subregions of the 230–base pair (bp) cassettes can be amplified separately by RCP-PCR for better performance (Fig. 2B).

To demonstrate INTERSTELLAR, DB-X barcode cassettes were subcloned into three and a half 384-well microwell plates (1344 samples), and AD-Y barcode cassettes were subcloned into three 384-well microwell plates (1152 samples). The identification of these barcode cassette samples required a total of 14 plate PCR reactions. The four types of RCP-PCR libraries (two each for DB-X and AD-Y) were mixed 1 M volume each and sequenced. INTERSTELLAR was used to interpret the sequencing reads of the different structures with different barcodes all at once. We first confirmed that the expected numbers of unique RBCs (16 rows), CBCs (24 columns), and PBCs (14 plates) used in this experiment were successfully observed (Fig. 2C). Next, the sequence segment information obtained by INTERSTELLAR was analyzed by another script to identify dominantly representing (or clonal) barcode sequences in each well and to separately interrogate any mutational damages in loxP and lox2272 sequences in each well (fig. S1). For each barcode cassette, two subregion amplification products both contained one of the two barcodes. Last, sample wells with no high-confident agreement for the barcode between the two types of RCP-PCR products were discarded, yielding a total of 287 (18.7%) and 299 (26.0%) high-confident clonal DB-X and AD-Y barcode cassette samples, respectively, whose gain rates were within the range of those expected from the previous study. We also randomly selected 24 independent samples and confirmed by Sanger sequencing that 23 of their sequences were consistent with those identified by RCP-PCR.

Translation of scATAC-seq reads

We next tested a simple translation of high-throughput sequencing reads that did not have a parental-local segment allocation. Using INTERSTELLAR, we emulated a sci-ATAC-seq dataset of *Drosophila* embryogenesis (*30*) in 10x Genomics' Cell Ranger ATAC, originally developed to analyze 10x scATAC-seq libraries. In sci-ATAC-seq, followed by fixation, sample nuclei are split into subpools, where open chromatin regions of nuclei are fragmented by Tn5 transposase with subpool-specific barcodes, yielding barcoded DNA fragments. After combining the barcoded nuclei samples, they are split into subpools again, where unique combinations of Illumina i5 and i7 indexed adapters are concatenated to the fragmented DNA in the nuclei of the corresponding subpool by PCR. Last, PCR products are pooled for a single Illumina sequencing run. This multistep split-and-pool strategy is designed to massively tag open chromatin DNA fragments of single cells with cell-specific combinations of DNA barcodes. Each library was sequenced by a total of four reads: two paired reads to sequence the genomic region and two index reads each to identify a combination of four barcodes (Fig. 3A). On the other hand, 10x scATAC-seq is an emulsion-based method that encapsulates single cells in water-in-oil droplets with unique droplet-specific 16-bp barcodes. In each droplet, open chromatin regions are fragmented by Tn5 and concatenated to the droplet-specific barcodes (i.e., cell IDs), each of whose products is sequenced by a total of three reads (Fig. 3A).

We translated the sci-ATAC-seq dataset of embryos 6 to 8 hours after egg laying to the read structure of 10x scATAC-seq, performed two-dimensional uniform manifold approximation and projection (UMAP) embeddings of the high-dimensional single-cell genomic accessibility count matrix by Cell Ranger ATAC, and compared it with that from the original sci-ATAC-seq reads obtained using the proprietary data analysis pipeline (Fig. 3B). The cell state

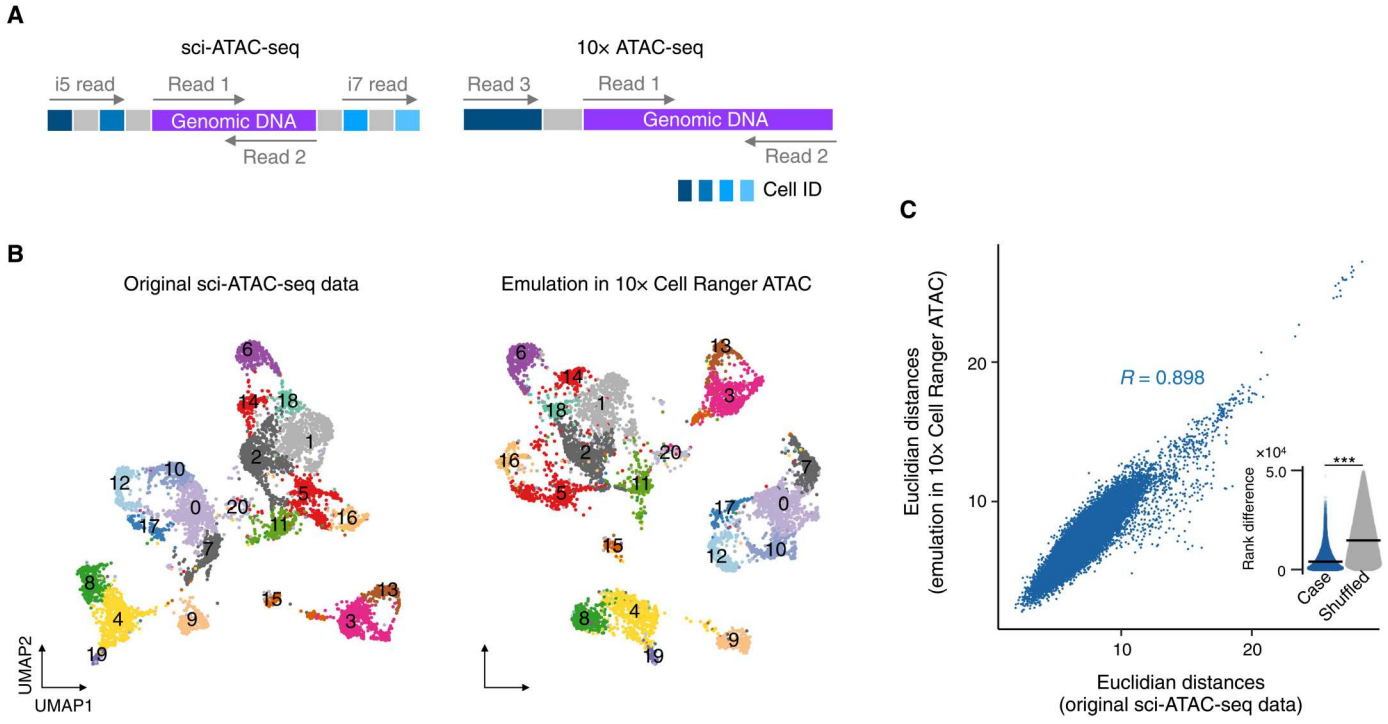

Fig. 3. Translation of scATAC-seq reads. (**A**) Read structures of sci-ATAC-seq and 10x scATAC-seq. ID, identifier. (**B**) Two-dimensional uniform manifold approximation and projection (UMAP) embeddings of sci-ATAC-seq data processed by its original pipeline for *Drosophila* embryo 6 to 8 hours after egg laying and that obtained by Cell Ranger ATAC with the read translation using INTERSTELLAR. Cell state annotations obtained by the original pipeline were applied to both embeddings. (**C**) Correlation in distance of two cells between the high-dimensional genomic accessibility count space of the original sci-ATAC-seq data and that by Cell Ranger ATAC. For each dataset, Euclidean distances in a high-dimensional latent semantic indexing (LSI) space were measured for the same 50,000 randomly sampled cell pairs. The inset sina plot represents rank difference distribution in the Euclidean distance of the same cell pairs before and after translation. The crossbar represents the median. ***$P < 2.2 \times 10^{-16}$ by the two-sided Wilcoxon rank sum test. *R*, correlation coefficient.

clusters identified by the original sci-ATAC-seq pipeline were markedly replicated in the translated dataset analyzed by Cell Ranger ATAC. To assess the data similarity between the original and emulated datasets, we compared Euclidean distances in a high-dimensional latent semantic indexing (LSI) space between randomly selected pairs of cells in the two datasets and found a Pearson's correlation coefficient (*R*) of 0.898 (Fig. 3C). Furthermore, we also measured the rank difference in the Euclidean distance of the same cell pairs in the two datasets and compared it with the random expectation (see Materials and Methods). We demonstrated that the data profiles were significantly preserved after the read translation ($P < 2.2 \times 10^{-16}$). Furthermore, we examined whether the read translation by INTERSTELLAR maintained the ability of the dataset to have its biological information extracted, similar to the original pipeline. The sci-ATAC-seq reads of three embryonic samples of 2 to 4, 6 to 8, and 10 to 12 hours after egg laying were pooled, translated, and analyzed by Cell Ranger ATAC. We confirmed that the analysis successfully recaptured the dynamic diversification of single-cell genomic accessibilities through *Drosophila* embryogenesis and cell state–specific marker gene accessibilities and their genomic distributions (fig. S2, A to E).

Cross-evaluation of different scRNA-seq reads and software tools

Differences in information capacity (or sequence representativity) between source segments and destination segments need to be taken into consideration in some read translations. For example, when the total base pair length of a destination segment(s) is shorter than that of the corresponding source segment(s), the destination segment might not be able to represent all the values observed in the source reads. However, the value space optimization implemented in INTERSTELLAR greatly alleviates this issue by allowing the end user to interpret parental-local segment allocations of the source read structure. For example, UMIs of typical scRNA-seq reads are local to their corresponding transcription products of corresponding single cells that are uniquely encoded in other segments. Some scRNA-seq libraries use shorter UMI segments than others, but in INTERSTELLAR, even the read translations from the latter to the former usually do not have major issues, because the number of unique UMI sequences observed for each of the combinatorial parental segments is practically limited and the value space for the UMIs representing the entire scRNA-seq data can be largely compressed.

To demonstrate that different scRNA-seq read structures can be practically translated to and from each other and analyzed by different software tools of choice that have originally been developed for specific read structures, we obtained four sequencing read datasets of 10x Chromium V3 (mouse heart; 16-bp cell ID and 12-bp UMI) (*31*), Drop-seq (mouse eye; 12-bp cell ID and 8-bp UMI) (*15*), Quartz-Seq2 (mouse stromal vascular fraction; 14-bp cell ID and 8-bp UMI) (*32*), and SPLiT-seq (mouse brain; three combinatorial 8-bp cell ID and 10-bp UMI) (*17*), all of which are representative

state-of-the-art scRNA-seq methods (Fig. 4A). We compared their analysis results obtained using the original software tools with those by 10x Cell Ranger (originally developed for 10x Chromium) and dropseq-tools (originally developed for Drop-seq) with read translation using INTERSTELLAR (Fig. 4B).

While all of the read translations were first performed with the value space optimization where UMI sequences local to the cell ID segments were optimized and translated to destination segment sequences, we also examined another emulation strategy that bequeathed the same UMI sequences to the destination reads for the translations whose destination UMI lengths were equal to or longer than those of the sources (constant sequences were added to the source UMIs to meet the length of the destination UMIs). In these sequencing library emulations, 10x Chromium V3 and Drop-seq read datasets were also self-translated by INTERSTELLAR for 10x Cell Ranger and dropseq-tools, respectively. The transcriptomic profiles of single cells in the translated datasets were compared to those in the original datasets by their correlations in the Euclidean distances of two cells in the high-dimensional transcriptome space and by the rank difference distribution of the Euclidean distances in the two datasets compared to random expectation (Fig. 4C). The "UMI reassignment" (value space optimization) and "UMI bequeathing" strategies both conferred almost identical results. Furthermore, the read translation that required shortening of the UMI lengths with the UMI reassignment strategy also demonstrated similar results to those which did not require UMI shortening, suggesting efficient read translations. A 98.17 and 99.99% of single cells retained the complete source UMI segment values in the emulation of 10x Chromium and SPLiT-seq datasets for dropseq-tools, respectively, whereas the translations without value space optimizations showed markedly poor cell state preservations (Fig. 4D) and Euclidean distance correlations (Fig. 4E), and no single cell retained the complete UMI segment values (Fig. 4F).

While all of the read dataset translations by INTERSTELLAR largely retained the single-cell transcriptome profiles of the original datasets processed by their proprietary tools, the self-emulation of the Drop-seq dataset for dropseq-tools showed a variance in contrast to that of the 10x Chromium V3 for 10x Cell Ranger. This was likely due to the use of the Levenshtein distance–based error correction for interpretation of the cell ID segments in INTERSTELLAR, where 10x Cell Ranger also uses a Levenshtein distance–based error correction, but dropseq-tools (Drop-seq) uses a unique error correction metric trained by its cell ID synthesis errors. [Quartz-Seq pipeline (Quartz-Seq2) and split-seq-pipeline (SPLiT-seq) use Sequence-Levenshtein distance (*33*) and Hamming distance, respectively.] To test this hypothesis, we performed self-emulation of the Drop-seq library by the UMI reassignment strategy with no error correction in the cell ID segments (Fig. 4G) and demonstrated that a much smaller variance was produced when the error correction process was fully performed in dropseq-tools (Fig. 4H). Apart from the variance observed in the self-translation, generally, the other sources of variance between the emulated and original data could be explained by differences in other software-specific downstream data processes that are independent of INTERSTELLAR. For example, the Quartz-Seq pipeline only obtains read count profiles of coding regions, while Cell Ranger and dropseq-tools account for 3′-untranslated regions.

To show the robustness of the read translation with harder tasks, we next translated each of the four scRNA-seq read datasets into a hypothetical read structure that has only 11-bp cell ID and 7-bp UMI segments, both of which are shorter than those of the original datasets. We then translated each of the simulated read datasets back into their source read structures ("round-trip" conversion) and evaluated the information preservation using the corresponding original software tool (fig. S3A). We observed similar cell state distributions before and after the round-trip conversions in all methods (fig. S3, B and C). Notably, the Quartz-Seq data before and after the round-trip conversion showed improved agreement (R = 0.999; fig. S3C) compared to the agreement between the Quartz-Seq data and that emulated in Cell Ranger and dropseq-tools (Fig. 4C), supporting the hypothesis that the differences in the downstream processing pipelines were the sources of variance. In contrast, this was not the case for SPLiT-seq, which may be due to the differences in cell ID error correction methods, cell ID loss from translation into a shorter sequence space, or the combination thereof. Nevertheless, we could not detect any negative effect caused by INTERSTELLAR itself.

To further challenge INTERSTELLAR with translating a more complex read structure with high-order value space optimizations, we performed a round-trip conversion between a pooled library of 10x Chromium scRNA-seq datasets and a hypothetical read structure with multilayered parental-local segment allocations (fig. S4A). While the multiplexed 10x Chromium libraries are sequenced by a combination of four reads, one for cell ID and UMI, one for transcripts, and two others as sample indices, we designed the hypothetical destination read structure to have a total of 17 segment variables in addition to the transcript sequence segment. Analyzing both the original and round-tripped reads using 10x Cell Ranger, we confirmed that INTERSTELLAR could preserve well the single-cell transcriptome profile information (R = 1.000; fig. S4, B and C).

Translation of spatial transcriptomics reads

In the translation of scATAC-seq and scRNA-seq reads, the values extracted from the source segment sequences were assigned to destination segment sequences that were arbitrarily generated or selected from a given allowlist. However, in INTERSTELLAR, the user can also define the translation of source segment sequences to corresponding destination segment sequences by providing a sequence conversion table. To demonstrate the use of this function, we translated spatial transcriptome reads of Slide-seq (*21*) to the read structure of 10x Visium (*22*) and analyzed them by Visium′s proprietary software Space Ranger. The two spatial transcriptomics technologies have been developed with similar conceptual designs. In brief, a tissue sample section is applied to a surface material where reverse transcription (RT) primers with unique positional barcodes are immobilized on distinct locations of the two-dimensional surface. Transcriptomes released from cells of specific positions are captured by proximal, positionally barcoded RT primers and reverse-transcribed such that they are fused to the positional barcodes for pooled high-throughput sequencing. Both read structures are similar to those of scRNA-seq (i.e., UMI used to uniquely identify each RT primer molecule; Fig. 5A). The major difference between the two spatial transcriptomics technologies is in the preparation of positionally barcoded RT primers. In Visium, RT primers with known positional barcode sequences are immobilized on predetermined spots of the two-dimensional surface. In contrast, Slide-

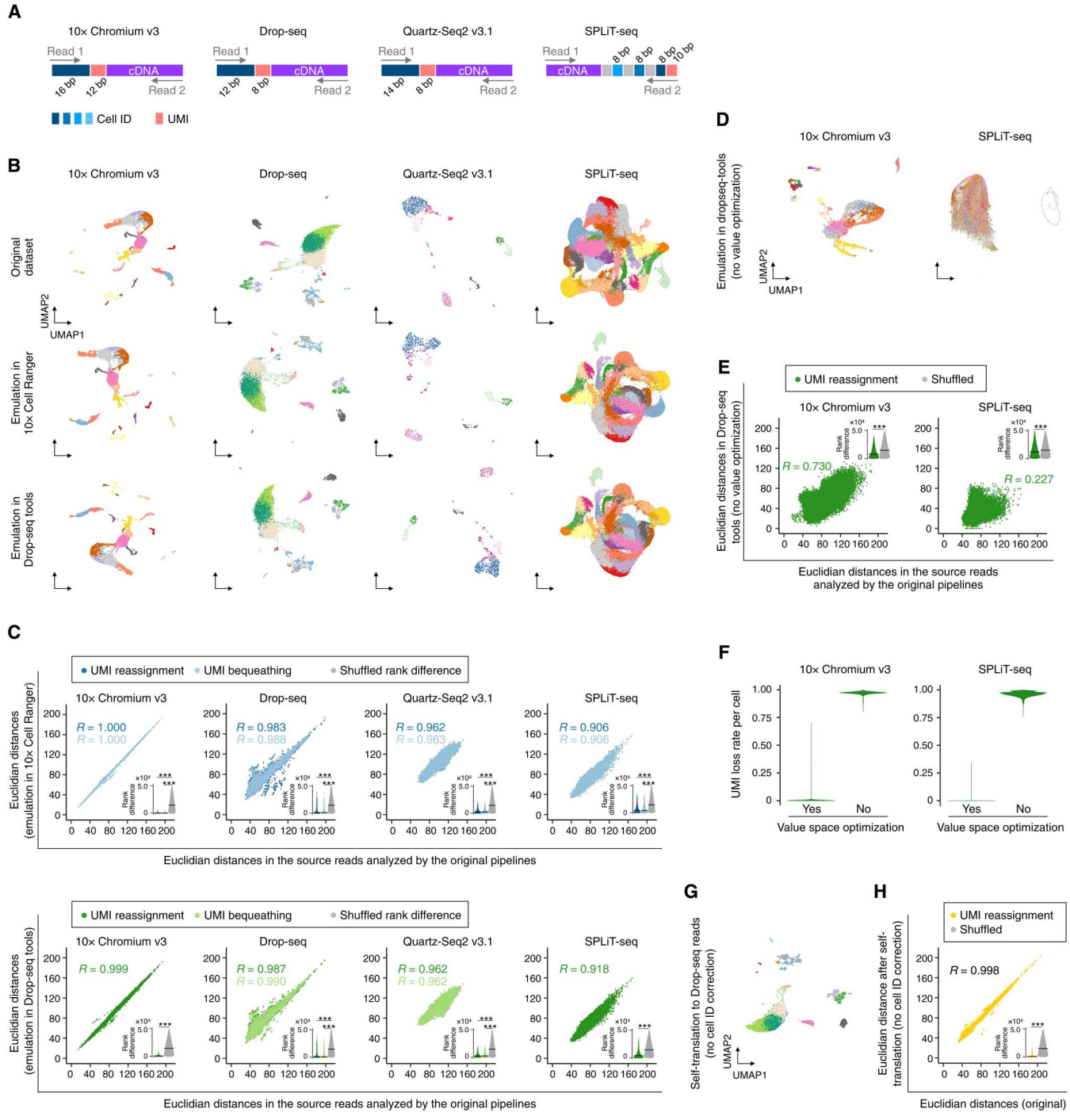

Fig. 4. Cross-evaluation of different scRNA-seq reads and software tools. (**A**) Read structures of different single-cell RNA sequencing (scRNA-seq) methods. bp, base pair. (**B**) Two-dimensional UMAP embeddings of scRNA-seq datasets processed by their original pipelines and those analyzed using 10x Cell Ranger and dropseq-tools by read translation using INTERSTELLAR with the unique molecular ID (UMI) reassignment strategy. Cell state annotations obtained by the original pipelines were applied to the translated results. (**C**) Correlation in distance of two cells between the high-dimensional transcriptome spaces of the original datasets and those translated for Cell Ranger and dropseq-tools with the UMI reassignment and UMI bequeathing strategies. For each dataset, Euclidean distances in the gene expression count matrix were measured for 50,000 randomly sampled cell pairs. The bottom-right inset sina plot of each panel represents rank difference distribution in the Euclidean distance of the same cell pairs before and after translation. The crossbar represents the median. (**D**) Two-dimensional UMAP embeddings of 10x Chromium and SPLiT-seq datasets processed by their original pipelines and those analyzed by dropseq-tools using INTERSTELLAR without value space optimizations. (**E**) Correlation in distance of two cells between the high-dimensional transcriptome spaces of the original datasets and those translated for dropseq-tools without value space optimizations. (**F**) UMI loss rate per cell with and without value space optimizations. (**G**) Two-dimensional UMAP embedding of the Drop-seq dataset self-translated for dropseq-tools with no cell ID error correction. (**H**) Correlation in distance of two cells between the high-dimensional transcriptome spaces of the original and self-translated Drop-seq datasets. ***$P < 2.2 \times 10^{-16}$ by the two-sided Wilcoxon rank sum test.

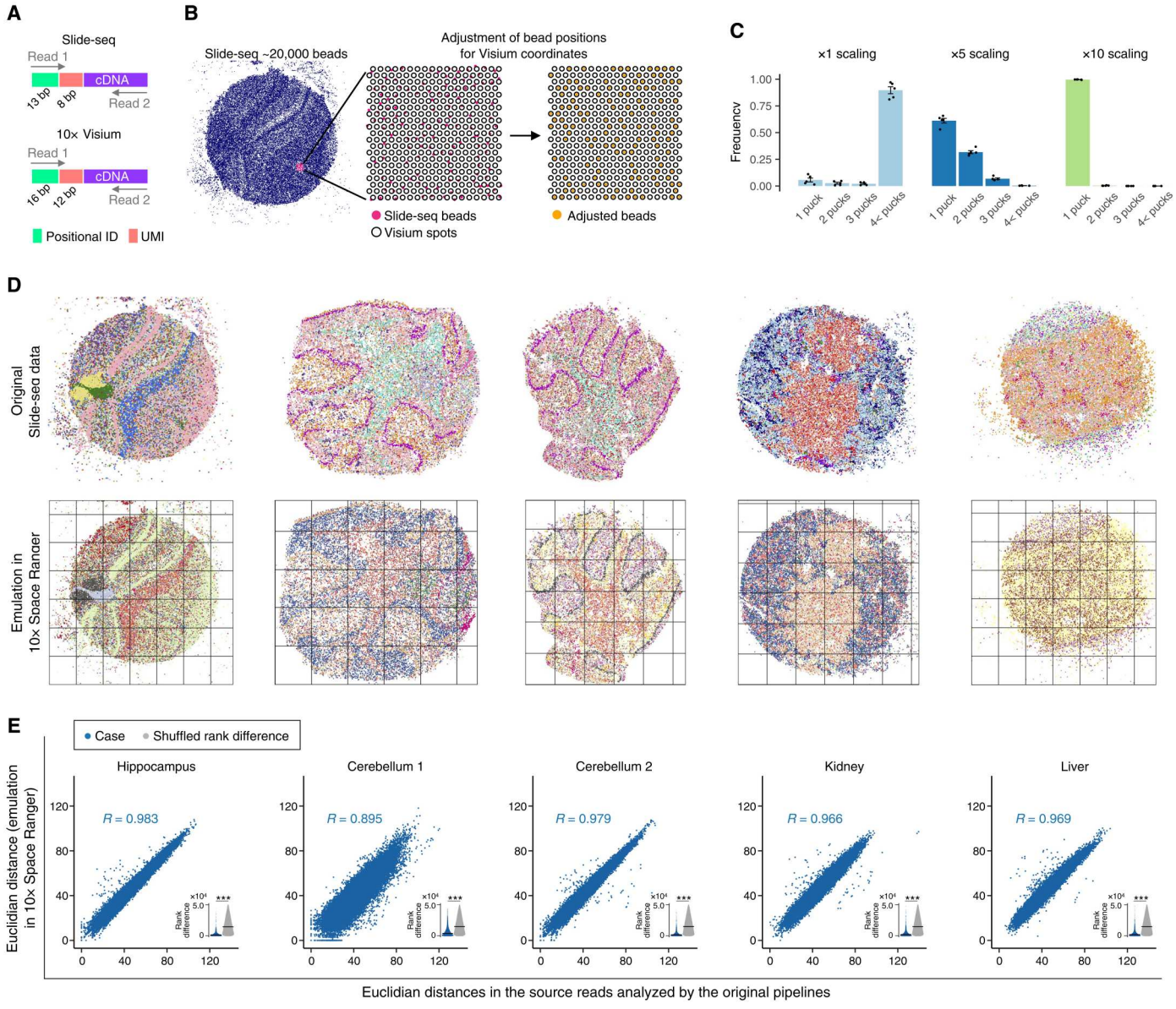

Fig. 5. Translation of spatial transcriptomics reads. (**A**) Read structures of Slide-seq and 10x Visium. (**B**) Strategy to associate Slide-seq positional barcodes to those of multiple 10x Visium slides. Multiple Visium slides are first tiled across an enlarged Slide-seq field with a given scaling factor. Slide-seq positional barcodes are then associated to the closest Visium positional barcodes. (**C**) Relative frequency distributions in number of Slide-seq positional barcodes assigned per Visium positional barcode with scaling factors of ×1, ×5, and ×10. Error bar indicates mean ± SEs. (**D**) Original Slide-seq datasets and those analyzed by 10x Space Ranger with ×10 scaling. Each grid represents a tiled Visium slide. The spatial data points are color coded according to their gene expression profile clusters identified independently in the analysis of each sample. (**E**) Correlation in Euclidean distance of two positional transcriptome profiles (UMI count matrices) between the original Slide-seq datasets and those translated and analyzed using Space Ranger with the read translation. Randomly sampled 50,000 positional barcode pairs with unique correspondences between the original and translated datasets were analyzed for each tissue sample. The inset sina plot represents rank difference distribution in the Euclidean distance of the same cell pairs before and after translation. The crossbar represents the median. ***$P < 2.2 \times 10^{-16}$ by the two-sided Wilcoxon rank sum test.

seq uses uniquely barcoded beads of unidentified barcode sequences, distributes them onto a glass slide surface, and retrospectively identifies the positional barcode sequences by sequencing by ligation before applying sample tissue.

We analyzed five Slide-seq libraries (hippocampus, cerebellum 1, cerebellum 2, kidney, and liver) using Space Ranger. While 10x Visium and Space Ranger are designed to analyze 4992 spatial spots at a time, a Slide-seq slide is usually composed of more than 20,000 barcoded beads. Therefore, we scaled the Slide-seq sample space coordinates and tiled the Visium slides, where each tile was treated as a single Visium experiment, and Slide-seq bead positions were further aligned to the most proximal Visium spots of the corresponding tiles (Fig. 5B). We first sought the best scaling size of the Slide-seq slide that enabled the assignment of single Slide-seq beads to unique Visium spots. Among the three expansion scales of ×1, ×5, and ×10, we found that the ×10 scaling enabled an average of

99.6% Slide-seq positional barcodes across the five samples to find their unique Visium spots (Fig. 5C and table S1). After generating a sequence conversion table for Slide-seq positional barcodes to their corresponding Visium spot barcode of corresponding tiles, reads were translated using INTERSTELLAR and analyzed by Space Ranger. As a result, the spatial gene expression patterns obtained by the read translations were markedly similar to those analyzed by the original pipeline (Fig. 5D). Euclidean distances of two positional pairs in the high-dimensional transcriptome space were highly correlated before and after the read translation (Fig. 5E). The median rank differences of Euclidean distances before and after read translation were all significantly lower than those of random expectation but seemed highly dependent on the sequencing quality. The median rank differences before and after translation for hippocampus, cerebellum 1, cerebellum 2, kidney, and liver were 1479, 3874, 1626, 2054, and 1944, respectively, where their average read counts per positional barcode were 5363, 315, 10,262, 9544, and 15,481, respectively.

Recent studies have used polyadenylated cellular RNA barcodes for scRNA-seq to obtain single-cell transcriptomic information together with cell clone barcodes [e.g., CellTagging (*34*) and LARRY (*35*)], cell lineage barcodes [e.g., scGESTALT (*36*)], and genetic perturbation reagent information [e.g., Perturb-seq (*37*) and CROP-seq (*38*)]. To demonstrate that INTERSTELLAR can analyze such multimodal sequencing read datasets, we also translated the recently published sci-Space (*39*) reads for 10x Cell Ranger. The sci-Space read structure is composed of read 1, always encoding cell ID and UMI segments; and read 2, encoding cDNA sequence, spatially deposited positional ID, section ID, or slide ID (Fig. 6A), where a combination of the three IDs paired with the same cell ID defines a predefined two-dimensional coordinate among multiple sci-Space slides for that single cell. The cell states identified from the translated reads using Cell Ranger and their spatial positions were markedly similar to those of the original reads analyzed by the proprietary tool (Fig. 6B).

Interpretation of long-read sequencing reads

Long-read sequencing (*40*) has also thrived in the development of new experiments, such as full plasmid sequencing, currently replacing Sanger sequencing, and the construction of a barcoded deep mutational scanning (DMS) library (*41*, *42*). To demonstrate INTERSTELLAR for the interpretation of long-read sequencing reads, we applied INTERSTELLAR to analyze a PacBio sequencing dataset that was previously prepared to identify a barcoded DMS library of *MSH2* variants (*42*). In the original work, 242 variants of 2928-bp *MSH2* gene were synthesized and cloned together with 13-bp random DNA barcodes, and yeast strains with the variant expressing vectors were pooled to assess growth effect, where the strain enrichment was measured by high-quality Illumina sequencing only targeting the short barcode regions. To identify barcode sequences corresponding to each variant, AssemblyByPacBio (ABP) (*43*) was used to align the reads to the reference sequence to identify both variants and barcodes (Fig. 7A). Last, PacRAT (*44*) was used to identify and correct barcodes that are uniquely assigned to variants in the allowlist.

We replaced the segment sequence identification process by ABP with INTERSTELLAR, where 20-bp upstream and downstream sequences of the barcode and *MSH2*-coding region were simply identified in the reads using regular expression to extract their respective inner regions (Fig. 7B). The read count distribution of unique barcode species identified both by ABP and INTERSTELLAR were markedly higher than those identified by one of them (Fig. 7C). However, INTERSTELLAR identified substantially more barcodes that are almost inclusive of those identified by ABP (Fig. 7D, top half). The barcode species identified by both pipelines were linked to more variants in the allowlist and more enriched for the expected length of 2928 bp than the others. When the same analysis was limited to barcode species identified by either pipeline with a barcode count of two or more, the agreement between ABP- and INTERSTELLAR-identified barcodes and the enrichment of barcodes linked to variants in the allowlist were both elevated as expected (Fig. 7D, bottom half). The DMS assay part adopts this barcode-variant cross-reference table and measures only barcode abundances in a screening condition using high-quality short-read sequencing to assess their associated variant effects. Because barcodes in the table that are not shown in the screening are ignored, false-positive barcodes identified from the long-read sequencing would not be a problem. Accordingly, we suggest that the use of INTERSTELLAR in the generation of the cross-reference table could only increase the sensitivity of the DMS assay without decreasing the assay sensitivity.

DISCUSSION

The structure of any sequencing library is always designed by stipulating information to be encoded in the DNA sequence with positions of sequence segments or by sectioning them using constant marker sequences (otherwise the library cannot be analyzed after sequencing). After any assay followed by sequencing of the library, sequence segments are extracted and error-corrected for downstream analyses. INTERSTELLAR has full capacity to decode any of these reads with the flexible regular expression system and the parental-local associations of values encoded in the sequence segments.

We performed read translations and data analyses using different software tools for scATAC-seq, scRNA-seq, and spatial transcriptomics reads and compared the results with those from the original reads analyzed by the original proprietary software tools. Although the overall results were markedly similar between the original and emulated results, there were different levels of the variances observed. The differences in the results can be explained by three potential sources: (i) the read interpretation process, (ii) the destination segment assignment process, and (iii) differences in the value analysis processes between different software tools, where INTERSTELLAR is responsible for the first two. From the scRNA-seq read translation demonstrations, the error correction step of the read interpretation process was suggested to be a potential major source of the variance seen, in which the error correction of the read interpretation was likely to make the error correction steps implemented in different software tools ineffective (i.e., overriding of the error correction strategy by INTERSTELLAR). Although the Levenshtein distance metric is the default for the non–allowlist-based error correction of INTERSTELLAR, and this is practically not an issue for most sequencing read data analyses, it can be replaced with Bartender or a user-developed plugin. The destination segment sequence assignment process is the only potential source of the loss of information encoded in the source reads when the information capacity (or representativity) of a destination

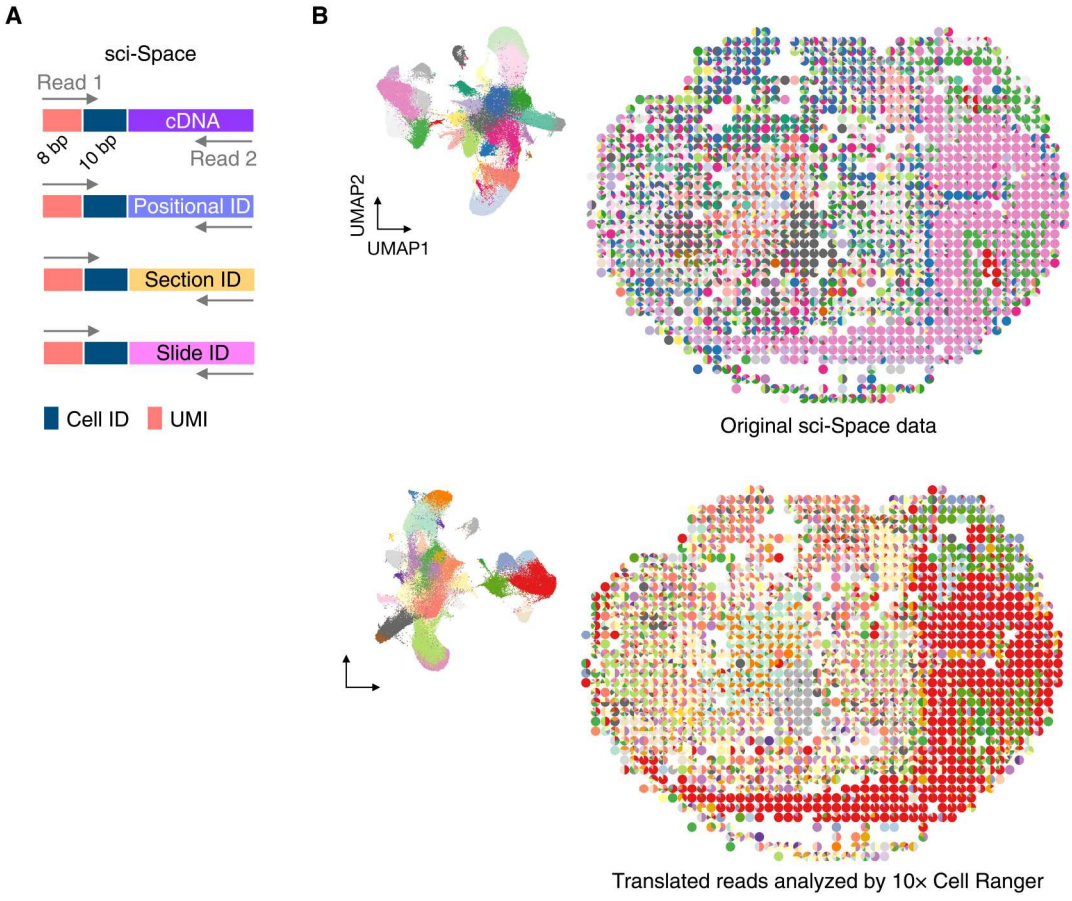

Fig. 6. Translation of multimodal scRNA-seq reads. (**A**) Read structures of sci-Space. (**B**) Two-dimensional UMAP embeddings of cells and the spatial distributions of cell states for the original sci-Space data (top) and the translated data analyzed by 10x Cell Ranger (bottom). The cell state clusters are color coded according to their gene expression profile clusters identified independently in each analysis.

segment is less than that of the corresponding source segment. To address this issue, we implemented theoretically the best value space optimization strategy that uses the user-defined information of parental-local segment allocations and successfully demonstrated that the information loss could be minimal for the read translations with a reduction in sequence representativity.

In the last couple of decades, beyond the (epi)genomic and transcriptomic analyses of clinical samples and various species, applications of massively parallel short-read sequencing technologies have enabled the development of wide-ranging biological assays, and the field continues to expand rapidly. While it has been a practice to develop and combine proprietary sequencing read interpreters and data analysis pipelines with the development of new sequencing-based assays, we propose a shift to a next form, where the community uses a common sequencing read interpretation and translation platform, such as INTERSTELLAR, and develops only the data analysis parts and shares them separately for the best utilization of data processing resources.

MATERIALS AND METHODS

Datasets

scATAC-seq, scRNA-seq, spatial transcriptomics, and *MSH2* DMS datasets obtained for this study are listed in table S2. For the sci-ATAC-seq dataset, we found only FASTQ files where cell IDs are recorded in the header line of each entry. Therefore, we generated new FASTQ files such that each sequencing read entry encoded the cell ID with per base *Q* scores of 30. Similarly, for the Slide-seq datasets, we extracted positional barcodes, UMIs, and cDNA sequences from the available BAM files and regenerated FASTQ files by setting the per base *Q* scores of positional barcodes and UMIs to 37. *Q* scores for the cDNA sequences were inherited from the BAM files.

Preparation of barcoded plasmids for RCP-PCR

In BFG-Y2H (*9*), the structures of DB-X and AD-Y barcode cassettes are 5′-loxP′-U1-Uptag-U2-lox2272-D1-Dntag-D2-3′ and 5′-U1-Uptag-U2-loxP′-D1-Dntag-D2-lox2272-3′, respectively, where U1, U2, D1, and D2 are common PCR amplification handles specific to DB-X barcodes or AD-Y barcodes, Uptag and Dntag are unique IDs assigned to a specific protein X- or Y-encoding gene fused to the respective Y2H domains (DB or AD), and loxP′ (reverse complement of loxP) and lox2272 are site-specific Cre recombination sites (Fig. 2B). In this study, DB-X and AD-Y barcode

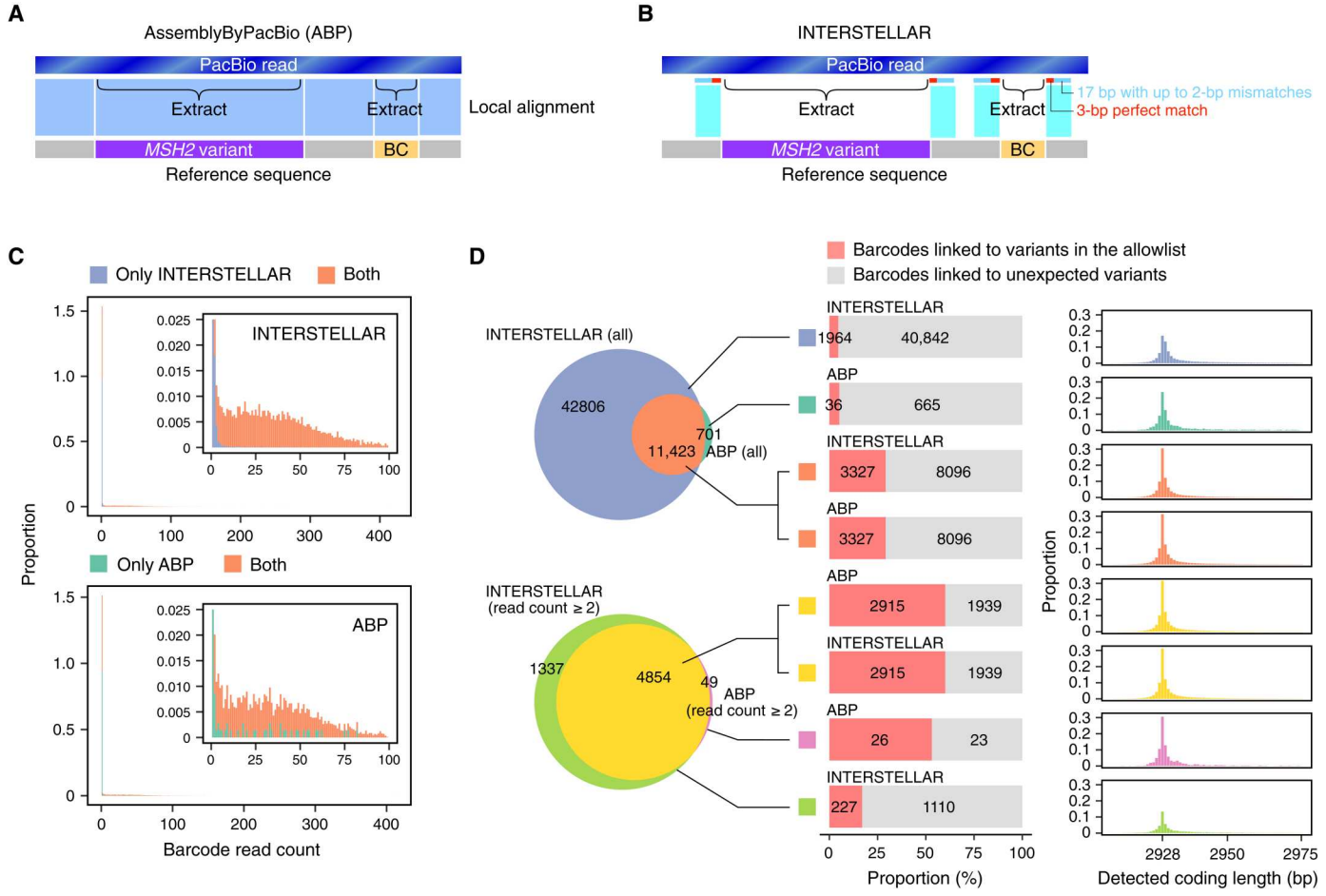

Fig. 7. Interpretation of long-read sequencing reads. (**A**) Read segmentation strategy by AssemblyByPacBio (ABP). (**B**) Read segmentation by INTERSTELLAR. In the ABP workflow, the sequencing reads are first aligned to the reference sequence. The *MSH2* variants and barcodes are then extracted on the basis of their positions aligned to the reference. When INTERSTELLAR was used, we extracted coding variant and barcode segments by simply identifying their 20-bp upstream and downstream sequences with fuzzy matching (3-bp perfect match for the inner edge and up to two mismatches for the remaining 17-bp region). (**C**) Read count distribution of barcodes identified by INTERSTELLAR (top) and ABP (bottom). (**D**) Left: Venn diagrams for barcode species detected by the two workflows. Top diagram: With no read count threshold for identified barcode species. Bottom diagram: With a read count threshold of two or more. Middle: Proportion of barcodes whose *MSH2* variants detected by each corresponding tool were involved in the allowlist. Right: Length distribution of coding variant segments identified by each corresponding tool.

plasmids were respectively constructed from pDN0510 and pDN0509 (*45*) by assembling BFG-Y2H DB-X and AD-Y barcode fragment pools (*9*) by three-fragment Gibson DNA assembly (*46*) as follows: DB-X Uptag and Dntag fragment pools were amplified using the random barcode templates DB-BC-UP and DB-BC-DN, with the primer pairs DB-BC-UP_F/DB-BC-UP_R and DB-BC-DN_F/DB-BC-DN_R, respectively. Similarly, AD-Y Uptag and Dntag fragment pools were amplified using the random barcode templates AD-BC-UP and AD-BC-DN, with the primer pairs AD-BC-UP_F/AD-BC-UP_R and AD-BC-DN_F/AD-BC-DN_R, respectively. Each PCR was performed in a 35-μl volume, including 3.6 μl of 10 pM template, 0.7 μl each of 10 μM primers, 0.2 μl of Phusion DNA polymerase, 7 μl of 5× Phusion HF detergent-free buffer (Thermo Fisher Scientific, F520L), and 0.28 μl of 25 mM deoxynucleoside triphosphates (dNTPs) with the following thermal cycler conditions: 98°C for 30 s, five cycles of 98°C for 10 s, 65°C for 10 s, and 72°C for 10 s, and then 24 cycles of 98°C for 10 s and 72°C for 10 s, followed by 72°C for 5 min for the final extension. The pDN0509 and pDN0510 backbones were linearized by PI-Psp I [New England Biolabs (NEB)] following the manufacturer's instruction. Each Gibson DNA assembly of the Uptag pool, Dntag pool, and linearized backbone was performed in a total of 20-μl volume with 25 fmol of each of the backbone and barcode fragments. The reaction was incubated at 50°C for 1 hour, and 1 μl was used to transform 50 μl of One Shot ccdB Survival 2 T1^R Competent Cells (Thermo Fisher Scientific, A16460) according to the manufacturer's instructions. The transformation samples were spread on 245 mm × 245 mm square LB + ampicillin plates and incubated overnight at 37°C for colony isolation. Single colonies were picked by QPix 450 robot (Molecular Devices) and arrayed into 384-well plates with liquid LB + ampicillin media. Oligonucleotides used in this protocol are listed in table S3.

RCP-PCR

To identify the clonal barcodes with high-quality base calling with a short paired-end sequencing, two different RCP-PCRs are

performed against the same samples. In brief, DB-Tag RCP-PCR and DB-Lox RCP-PCR respectively identify 5′-U1-Uptag-U2-lox2272-D1-Dntag-D2-3′ and 5′-loxP′-U1-Uptag-U2-lox2272-D1-3′ regions of the same sample wells, and AD-Tag RCP-PCR and AD-Lox RCP-PCR, respectively, identify 5′-U1-Uptag-U2-loxP′-D1-Dntag-D2-3′ and 5′-U2-loxP′-D1-Dntag-D2-lox2272-3′ regions of the same sample wells (Fig. 2B). The first RC-PCRs for DB-Tag, DB-Lox, AD-Tag, and AD-Lox fragments were performed with the primer sets encoding row or column IDs (table S3). Each RC-PCR was performed in 10 μl of volume, including 4 μl of 16-fold diluted overnight culture as templates, 1 μl each of 2 μM primers, 0.2 μl of Phusion DNA polymerase, 2 μl of 5× Phusion HF buffer (NEB), and 0.08 μl of 25 mM dNTPs. The thermal cycler conditions for DB- and AD-Tag RC-PCRs were: 95°C for 3 min, 30 cycles of 95°C for 10 s, 63°C for 10 s, and 72°C for 15 s, and then 72°C for 5 min for the final extension. The conditions for DB- and AD-Lox RC-PCRs were: 95°C for 3 min, 30 cycles of 95°C for 10 s, 66°C for 10 s, and 72°C for 15 s, and then 72°C for 5 min for the final extension. (Note that 4× 96-well PCRs were performed for each 384-well template sample plate for better sample handling.) The RC-PCR products were pooled by plates and purified using FastGene PCR purification kit (Nippon Genetics) and subjected to plate PCR (P-PCR) using custom indexed primers for Illumina library preparation listed in table S3. Each P-PCR was performed in a 40-μl volume including 2× Phusion High-Fidelity PCR Master Mix (NEB), 1 μl each of 10 μM forward and reverse plate primers, and 1 ng of size selected RC-PCR product with the following thermal cycler condition: 98°C for 30 s; 15 cycles of 98°C for 10 s, 60°C for 10 s, and 72°C for 1 min; and then 72°C for 5 min for the final extension. The P-PCR products were pooled and quantified by quantitative PCR using the KAPA Illumina Library Quantification Kit (Kapa Biosystems) and sequenced by MiSeq (Illumina, 2× 250-bp paired-end sequencing).

Interpretation of RCP-PCR reads

Using INTERSTELLAR, we identified Uptag, Dntag, loxP, and/or lox2272 segments of DB-Tag, DB-Lox, AD-Tag, and AD-Lox reads with RBCs, CBCs, and PBCs. We discarded any segment sequences whose average *Q* scores were below 20 or whose minimum per base *Q* scores were below 10. P5 PBCs, P7 PBCs, RBCs, and CBCs were error-corrected using the allowlists with a maximum Levenshtein distance threshold of 1. The process configuration file of INTERSTELLAR is available at https://github.com/yachielab/Interstellar/blob/main/config/fig2_RCPPCR/rcppcr.conf or https://doi.org/10.5281/zenodo.7250991.

Identification of high-quality clonal barcode cassettes

To determine sample wells containing clonal barcode cassettes, we analyzed the RCP-PCR data interpreted by INTERSTELLAR with the following criteria. For each well, we first determined the most dominant Uptag and Dntag in the Tag RCP-PCR result. If the occupancies of the most frequent tags were both above 50%, then the Uptag and Dntag were regarded as clonal in the well. The validities of loxP and lox2272 were separately evaluated using the reads from the Lox RCP-PCR, with the criterion of 70% or more reads encoding the correct sequences. Because DB- and AD-Lox RCP-PCR reads, respectively, encode Uptags and Dntags, we also determined the most dominant Uptag or Dntag from the Lox RCP-PCR reads of each well with the same criterion used to call the clonal Uptags and Dntags from the Tag RCP-PCR reads. Last, wells with single dominant Uptag and Dntag pairs, valid loxP and lox2272 sequences, and no conflict in Uptag or Dntag between Tag and Lox RCP-PCRs was called to contain high-quality clonal barcode cassettes. The Python script for this process is available at https://github.com/yachielab/Interstellar/blob/main/utils/analyze_rcppcr.py or https://doi.org/10.5281/zenodo.7250991. For validation, we randomly selected 24 wells of predicted clonal barcode cassettes, cultured the corresponding bacterial samples overnight in LB + ampicillin media at 37°C, extracted plasmids using FastGene plasmid kit (Nippon Genetics), and performed Sanger sequencing with Term-Rvs primer (table S3).

Translation of sci-ATAC-seq reads

Using INTERSTELLAR, we identified four combinatorial cell IDs and a genomic DNA region of each sci-ATAC-seq read combination and translated them into the read structure of 10x scATAC-seq. We discarded any read containing genomic DNA segments whose average *Q* scores were below 20 or whose minimum per base *Q* scores were below 5. Cell IDs in the source reads were error-corrected with the maximum Levenshtein distance threshold of 1 using a position-specific cell ID allowlist. Each combination of cell IDs was translated into a 16-bp barcode selected from the cell ID allowlist of 10x Chromium scATAC-seq. The process configuration file of INTERSTELLAR is available at https://github.com/yachielab/Interstellar/blob/main/config/fig3_sciATAC/sciATAC_to_10xATAC.conf or https://doi.org/10.5281/zenodo.7250991.

scATAC-seq data analysis

We first analyzed the sci-ATAC-seq reads of *Drosophila* single cells for 6 to 8 hours after egg laying (GSE101581) by using 10x Cell Ranger ATAC (version 1.2.0) with read translation using INTERSTELLAR. FlyBase version R6.25 and Ensemble BDGP6.95 were used as a reference genome and for genomic annotation, respectively, to obtain a genomic accessibility count matrix. For comparison, we obtained the original raw genomic accessibility count matrix [2–kilo–base pair (kbp) bins across the genome; GSE101581] produced in the original study. Following the workflow used in the original study, cells with the lowest 10% read counts were discarded, resulting in 7092 cells. Furthermore, the genomic accessibility count matrix was limited to the top 20,000 2-kbp bins of frequently mapped reads across cells for the subsequent steps. After obtaining the genomic accessibility count matrix from the translated 10x scATAC-seq reads using Cell Ranger ATAC, the following analyses were limited to the 7092 cells observed in both matrices. Both genomic accessibility count matrices were processed by Signac version 1.0.0 (*47*). For each matrix, accessibility count normalization was performed by RunTFIDF(), and the normalized matrix was processed by RunSVD() for low-dimensional data projections by singular value decomposition (SVD). After identifying the top 30 LSI components, LSIs correlated with single-cell read depth with Pearson's *R* of more than 0.5 were removed (LSI components 1 and 4 and components 1 and 5 were removed from the original and translated datasets, respectively). The remaining 28 LSI components were used for UMAP embedding of the data using RunUMAP() to identify cell clusters using *k*-nearest neighbor (kNN) clustering by FindNeighbors() and FindClusters() with a resolution parameter of 1. We also analyzed the sci-ATAC-seq

reads of *Drosophila* single cells for all of the available developmental stages in the same study (2 to 4, 6 to 8, and 10 to 12 hours after egg laying) by using Cell Ranger ATAC with read translation. The data analyses by Cell Ranger ATAC were first independently performed for three stage-specific samples—each sample with the same reference genome and genomic annotation. We aggregated the genomic accessibility count matrices from all samples into a single matrix using Cell Ranger ATAC and analyzed it by Signac. Low-quality cells were discarded according to the instruction of Signac; the nucleosome signal scores and transcription start site (TSS) enrichment scores of cells were computed by NucleosomeSignal() and TSSEnrichment(), respectively, and cells with nucleosome signal scores of <2, TSS enrichment score of >2, and % reads mapped to identified accessibility peaks of >40 were retained. Furthermore, identified accessibility peaks with 200 to 100,000 mapped reads across retained single cells were used to construct the high-quality genomic accessibility count matrix, followed by read count normalization and low-dimensional data projection as described above.

Comparison of the original and translated high-dimensional datasets

To compare two high-dimensional count matrices obtained by applying different data processing methods to the same scATAC-seq, scRNA-seq, or spatial transcriptome read dataset, we adopted the following metric. First, 50,000 pairs of high-dimensional data points (e.g., transcriptome profiles of single cells or spatial positions) were randomly sampled, and their Euclidean distances in the two datasets were compared. Furthermore, to quantitatively evaluate the similarity of the two datasets in a nonparametric manner, we defined the rank difference $\Delta R_{i,j}$ of the same high-dimensional data pairs (i, j) between the two datasets as follows and evaluated their distribution compared to that obtained from two data pairs each independently sampled from the two datasets

$$\Delta_{i,j} = \mid R(i,j) - R(i',j') \mid$$

where i and j are randomly sampled high-dimensional data points in a set $\ominus$ ($i,j \in \ominus$), i' and j' are the corresponding data points in a projected (translated) set $\ominus'$ ($i',j' \in \ominus'$), and $R(x,y)$ represents the rank of Euclidean distance of x and y ($x,y \in X$).

Translation of scRNA-seq reads

10x Chromium V3 scRNA-seq, Drop-seq, Quartz-Seq2 v3.1, and SPLiT-seq reads were analyzed by INTERSTELLAR to identify their cell ID(s), UMI, and cDNA sequence segments. We discarded sequencing reads with any segment whose average Q score was below 20 or whose minimum per base Q score was below 5. The cell IDs of Drop-seq reads were corrected by the imputation-to-majority correction with the maximum Levenshtein distance threshold of 1. For 10x Chromium V3, Quartz-Seq2 and SPLiT-seq reads, the cell IDs were corrected using allowlists with the maximum Levenshtein distance threshold of 1. In the read translation for Cell Ranger, the cell ID values and UMI values were assigned to sequence segments selected from the whitelist of 10x Chromium V3 and 12-bp random sequence segments, respectively. In the read translation for dropseq-tools, the cell ID and UMI values were assigned to 12- and 8-bp random sequence segments, respectively. We listed the runtime information of the read translation from each technology into 10x Chromium read structure in table S4. For the UMI bequeathing strategy, the source UMI sequences were elongated by A nucleotides to adjust the UMI lengths if necessary. For the round-trip conversion of the four scRNA-seq dataset, we translated the original read datasets into a hypothetical read structure of a smaller information capacity with 11-bp cell ID and 7-bp UMI and translated them back to the original read structure. The process configuration files of INTERSTELLAR are available at https://github.com/yachielab/Interstellar/tree/main/config or https://doi.org/10.5281/zenodo.7250991.

scRNA-seq data analysis

We translated scRNA-seq read datasets of 10x Chromium V3 scRNA-seq, Drop-seq, Quartz-Seq2 v3.1, and SPLiT-seq and analyzed them by 10x Cell Ranger (version 3.0.2) and dropseq-tools (version 2.3.0). For comparison, we also analyzed the original read datasets by their proprietary software tools [i.e., 10x Cell Ranger, dropseq-tools, Quartz-Seq pipeline, and split-seq-pipeline (commit c3923ea), respectively]. The mouse reference genome GRCm38 was commonly used throughout these analyses. In the analysis of both the original SPLiT-seq read dataset and those translated and analyzed by the other two software tools, cDNA segments mapped to intronic regions were also accounted for to estimate gene expression. Gene expression count matrices obtained from the original software tools were processed to filter out low-quality cells with the following criteria using Seurat version 3.2.0 (*48*). For those from 10x Cell Ranger, dropseq-tools, and Quartz-Seq pipeline, we removed cells whose numbers of detected genes were ≤200 or ≥2500 or whose UMI proportion from mitochondrial genes was ≥20%. For the original gene expression matrix of SPLiT-seq obtained by split-seq-pipeline, we removed cells whose numbers of detected genes were ≤250 or ≥2500 or UMI proportion from mitochondrial genes was ≥5%, considering the expectation of low mitochondrial reads in the single-nucleus RNA-seq. Last, cells commonly observed in the gene expression matrices obtained by the original software tool, 10x Cell Ranger and dropseq-tools, were retained, yielding 5003, 11,334, 1048, and 185,722 cells for ones sourced from the original 10x Chromium V3, Drop-seq, Quartz-Seq2, and SPLiT-seq datasets, respectively. Using Seurat, all of the filtered gene expression matrices were then processed by NormalizeData() with a scale factor of 10,000 and ScaleData(), followed by the extraction of the top 5000 highly variable genes by FindVariableFeatures() for principal components analysis (PCA) by RunPCA(). Using the top 20 principal components, we carried out two-dimensional UMAP embedding and kNN clustering of each dataset by RunUMAP(), FindNeighbors(), and FindClusters() with a resolution parameter of 0.6.

Highly complex round-trip conversion

We simulated a sequencing read pool of a highly complex hypothetical read structure from a total of four 10x Chromium V3 scRNA-seq read datasets (two heart and two neural samples) obtained from the 10x Genomics website. Because these datasets had already been demultiplexed, we provided simulated Illumina i5 and i7 index reads in addition to each paired-end read entry for cell ID plus UMI and cDNA. The i5 indices were provided to discriminate individual FASTQ datasets, and the i7 indices were provided to indicate the sample type (heart or neural), in which they could serve as parental sequence segments of their associated i5 indices. The sequencing read datasets were pooled and interpreted by

INTERSTELLAR as described above. The interpreted reads were then translated into a destination read structure with value space optimization, where 15-bp i7 index, 10-bp i5 index, 16-bp cell ID, and 12-bp UMI segment sequences were translated into arbitrarily designed three 2-bp units, four 3-bp units, five 4-bp units (each unit was limited to a selection from ten 4-bp sequences in an allowlist), and five 2-bp units, respectively. Last, the simulated reads were translated back to the read structure of 10x Chromium V3. The original and round-trip reads were both analyzed by 10x Cell Ranger. The gene expression matrices were derived by the same criteria described above. The process configuration file of INTERSTELLAR is available at https://github.com/yachielab/Interstellar/tree/main/config/figS4_10X_roundtrip or https://doi.org/10.5281/zenodo.7250991.

Translation of spatial transcriptomics reads

Using INTERSTELLAR, we analyzed Slide-seq reads and identified their positional barcodes, UMIs, and cDNA fragments. We discarded sequencing reads with any segment whose average *Q* score was below 20 or minimum per base *Q* score was below 5 and reads whose positional barcodes were not found in the allowlist with the perfect match. After obtaining a sequence conversion table between positional barcodes of a Slide-seq slide and those of multiple 10x Genomics Visium slide tiles, reads were grouped by destination Visium tile. For each Visium tile group, we translated the reads into the Visium read structure using INTERSTELLAR with the UMI bequeathing strategy, where the segment length is adjusted by adding A nucleotides. The process configuration file of INTERSTELLAR is available at https://github.com/yachielab/Interstellar/blob/main/config/fig5_Slide-seq/Slide_to_10xVisium.conf or https://doi.org/10.5281/zenodo.7250991. The resulting FASTQ files of Visium tiles were independently processed by 10x Genomics Space Ranger (version 1.0.0) with the options "--slide = V19L01-041 --area = C1" using a fake slide image (https://github.com/yachielab/Interstellar/blob/main/utils/fake_spaceranger_box.jpeg or https://doi.org/10.5281/zenodo.7250991) such that whole Visium spots were recognized to be covered by a tissue sample and processed. For each Slide-seq tissue sample, the Space Ranger results of multiple tiles were merged and analyzed by Seurat version 3.2.0 to obtain a single gene expression count matrix of spatial positions, with the same protocol applied for the scRNA-seq data analyses above, except that the top 3000 highly variable genes were used for PCA and the top 30 principal components were used for UMAP embedding and kNN clustering.

Translation and analysis of sci-Space data

Using INTERSTELLAR, we analyzed sci-Space reads of slide IDs 7 to 14 listed in table S2, identified their cell ID and UMI segments, and translated the reads into a single pair of FASTQ files for 10x Cell Ranger with the UMI bequeathing strategy. The process configuration file is available at https://github.com/yachielab/Interstellar/blob/main/config/fig6_sci-Space/sciSpace.conf or https://doi.org/10.5281/zenodo.7250991. After analyzing the translated reads by Cell Ranger, we processed the expression matrices obtained from the original pipeline (https://ftp.ncbi.nlm.nih.gov/geo/series/GSE166nnn/GSE166692/suppl/GSE166692_sciSpace_count_matrix.mtx.gz) and Cell Ranger with the same criteria used for the scRNA-seq analysis above. From each dataset, we independently performed kNN clustering and obtained cell state labels. Cell state occupancies in each spot were plotted as pie charts using an R package scatterpie v0.1.7 (https://github.com/GuangchuangYu/scatterpie) based on the coordinate information of cells from the original study (https://ftp.ncbi.nlm.nih.gov/geo/series/GSE166nnn/GSE166692/suppl/GSE166692_sciSpace_cell_metadata.tsv.gz).

Interpretation of *MSH2* DMS reads

The *MSH2* DMS long-read sequencing reads were independently processed by ABP and INTERSTELLAR. When ABP is used, the reads were first aligned to the reference sequence (FASTA file) using BWA-MEM (*49*) with options "-M -L 80", and the alignment results output in a BAM file were sorted using SAMtools (*50*). We then used extractBarcodeInsertPairs.py with options "-l 13 -p 6558 --start 3377 --end 6306" to extract the coding variant and barcode regions with their quality scores, followed by the extraction of barcode-variant combinations using unifyAssignment.py. The Python scripts were obtained from https://github.com/shendurelab/AssemblyByPacBio/ (commit 0cb2d1d). In parallel, using INTERSTELLAR, we extracted coding variant and barcode segments by identifying their 20-bp upstream and downstream sequences with fuzzy matching (3-bp perfect match for the inner edge and up to two mismatches for the remaining 17-bp region). The process configuration file of INTERSTELLAR is available at https://github.com/yachielab/Interstellar/blob/main/config/fig7_MSH2_DMS/MSH2.conf or https://doi.org/10.5281/zenodo.7250991. We generated a sequence-quality score table for extracted barcodes following the format produced by extractBarcodeInsertPairs.py and used unifyAssignment.py to obtain a barcode-variant table. The barcode-variant tables generated by ABP and INTERSTELLAR were subjected to PacRAT (*44*) to correct the variant sequences for each barcode.

Statistical analysis

The Euclidean distance correlations in high-dimensional data space between the original and translated results were all measured by Pearson's correlation. The statistical tests to compare the rank difference distributions to random expectations were performed by the two-sided Wilcoxon rank sum test.

Supplementary Materials

This PDF file includes:
Figs. S1 to S4

Other Supplementary Material for this manuscript includes the following:
Tables S1 to S4

REFERENCES AND NOTES

1. E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, A. M. Chu, C. Connelly, K. Davis, F. Dietrich, S. W. Dow, M. El Bakkoury, F. Foury, S. H. Friend, E. Gentalen, G. Giaever, J. H. Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, D. J. Lockhart, A. Lucau-Danila, M. Lussier, N. M'Rabet, P. Menard, M. Mittmann, C. Pai, C. Rebischung, J. L. Revuelta, L. Riles, C. J. Roberts, P. Ross-MacDonald, B. Scherens, M. Snyder, S. Sookhai-Mahadeo, R. K. Storms, S. Véronneau, M. Voet, G. Volckaert, T. R. Ward, R. Wysocki, G. S. Yen, K. Yu, K. Zimmermann, P. Philippsen, M. Johnston, R. W. Davis, Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).

2. A. M. Smith, L. E. Heisler, J. Mellor, F. Kaper, M. J. Thompson, M. Chee, F. P. Roth, G. Giaever, C. Nislow, Quantitative phenotyping via deep barcode sequencing. *Genome Res.* **19**, 1836–1842 (2009).
3. M. E. Hillenmeyer, E. Fung, J. Wildenhain, S. E. Pierce, S. Hoon, W. Lee, M. Proctor, R. P. S. Onge, M. Tyers, D. Koller, R. B. Altman, R. W. Davis, C. Nislow, G. Giaever, The chemical genomic portrait of yeast: Uncovering a phenotype for all genes. *Science* **320**, 362–365 (2008).
4. T. Roemer, J. Davies, G. Giaever, C. Nislow, Bugs, drugs and chemical genomics. *Nat. Chem. Biol.* **8**, 46–56 (2011).
5. K. Berns, E. M. Hijmans, J. Mullenders, T. R. Brummelkamp, A. Velds, M. Heimerikx, R. M. Kerkhoven, M. Madiredjo, W. Nijkamp, B. Weigelt, R. Agami, W. Ge, G. Cavet, P. S. Linsley, R. L. Beijersbergen, R. Bernards, A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431–437 (2004).
6. O. Shalem, N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. Mikkelson, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench, F. Zhang, Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
7. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
8. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
9. N. Yachie, E. Petsalaki, J. C. Mellor, J. Weile, Y. Jacob, M. Verby, S. B. Ozturk, S. Li, A. G. Cote, R. Mosca, J. J. Knapp, M. Ko, A. Yu, M. Gebbia, N. Sahni, S. Yi, T. Tyagi, D. Sheykhkarimli, J. F. Roth, C. Wong, L. Musa, J. Snider, Y. C. Liu, H. Yu, P. Braun, I. Stagljar, T. Hao, M. A. Calderwood, L. Pelletier, P. Aloy, D. E. Hill, M. Vidal, F. P. Roth, Pooled-matrix protein interaction screens using barcode fusion genetics. *Mol. Syst. Biol.* **12**, 863 (2016).
10. S. A. Trigg, R. M. Garza, A. MacWilliams, J. R. Nery, A. Bartlett, R. Castanon, A. Goubil, J. Feeney, R. O'Malley, S. C. Huang, Z. Z. Zhang, M. Galli, J. R. Ecker, CrY2H-seq: A massively multiplexed assay for deep-coverage interactome mapping. *Nat. Methods* **14**, 819–825 (2017).
11. U. Schlecht, Z. Liu, J. R. Blundell, R. P. S. Onge, S. F. Levy, A scalable double-barcode sequencing platform for characterization of dynamic protein-protein interactions. *Nat. Commun.* **8**, 15586 (2017).
12. J. S. Yang, M. Garriga-Canut, N. Link, C. Carolis, K. Broadbent, V. Beltran-Sastre, L. Serrano, S. P. Maurer, rec-YnH enables simultaneous many-by-many detection of direct protein-protein and protein-RNA interactions. *Nat. Commun.* **9**, 3747 (2018).
13. J. J. Díaz-Mejía, A. Celaj, J. C. Mellor, A. Coté, A. Balint, B. Ho, P. Bansal, F. Shaeri, M. Gebbia, J. Weile, M. Verby, A. Karkhanina, Y. Zhang, C. Wong, J. Rich, D. Prendergast, G. Gupta, S. Öztürk, D. Durocher, G. W. Brown, F. P. Roth, Mapping DNA damage-dependent genetic interactions in yeast via party mating and barcode fusion genetics. *Mol. Syst. Biol.* **14**, e7985 (2018).
14. J. A. Weinstein, A. Regev, F. Zhang, DNA microscopy: Optics-free spatio-genetic Imaging by a stand-alone chemical reaction. *Cell* **178**, 229–241.e16 (2019).
15. E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, S. A. McCarroll, Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
16. A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, M. W. Kirschner, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
17. A. B. Rosenberg, C. M. Roco, R. A. Muscat, A. Kuchina, P. Sample, Z. Yao, L. T. Graybuck, D. J. Peeler, S. Mukherjee, W. Chen, S. H. Pun, D. L. Sellers, B. Tasic, G. Seelig, Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
18. S. Picelli, Å. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, R. Sandberg, Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
19. D. A. Cusanovich, R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, J. Shendure, Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
20. J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, W. J. Greenleaf, Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
21. S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, E. Z. Macosko, Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
22. P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, Å. Borg, F. Pontén, P. I. Costea, P. Sahlén, J. Mulder, O. Bergmann, J. Lundeberg, J. Frisén, Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
23. T. Zhao, Z. D. Chiang, J. W. Morriss, L. M. LaFave, E. M. Murray, I. Del Priore, K. Meli, C. A. Lareau, N. M. Nadaf, J. Li, A. S. Earl, E. Z. Macosko, T. Jacks, J. D. Buenrostro, F. Chen, Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature* **601**, 85–91 (2022).
24. V. Svensson, R. Vento-Tormo, S. A. Teichmann, Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
25. T. Smith, A. Heger, I. Sudbery, UMI-tools: Modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
26. S. Parekh, C. Ziegenhain, B. Vieth, W. Enard, I. Hellmann, zUMIs–A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7**, giy059 (2018).
27. J. Ding, X. Adiconis, S. K. Simmons, M. S. Kowalczyk, C. C. Hession, N. D. Marjanovic, T. K. Hughes, M. H. Wadsworth, T. Burks, L. T. Nguyen, J. Y. H. Kwon, B. Barak, W. Ge, A. J. Kedaigle, S. Carroll, S. Li, N. Hacohen, O. Rozenblatt-Rosen, A. K. Shalek, A.-C. Villani, A. Regev, J. Z. Levin, Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* , 1–10 (2020).
28. R. Fang, S. Preissl, Y. Li, X. Hou, J. Lucero, X. Wang, A. Motamedi, A. K. Shiau, X. Zhou, F. Xie, E. A. Mukamel, K. Zhang, Y. Zhang, M. M. Behrens, J. R. Ecker, B. Ren, Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).
29. L. Zhao, Z. Liu, S. F. Levy, S. Wu, Bartender: A fast and accurate clustering algorithm to count barcode reads. *Bioinformatics* **34**, 739–747 (2018).
30. D. A. Cusanovich, J. P. Reddington, D. A. Garfield, R. M. Daza, D. Aghamirzaie, R. Marco-Ferreres, H. A. Pliner, L. Christiansen, X. Qiu, F. J. Steemers, C. Trapnell, J. Shendure, E. E. M. Furlong, The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**, 538–542 (2018).
31. G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, J. H. Bielas, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
32. Y. Sasagawa, H. Danno, H. Takada, M. Ebisawa, K. Tanaka, T. Hayashi, A. Kurisaki, I. Nikaido, Quartz-Seq2: A high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* **19**, 29 (2018).
33. T. Buschmann, L. V. Bystrykh, Levenshtein error-correcting barcodes for multiplexed DNA sequencing. *BMC Bioinformatics* **14**, 272 (2013).
34. B. A. Biddy, W. Kong, K. Kamimoto, C. Guo, S. E. Waye, T. Sun, S. A. Morris, Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564**, 219–224 (2018).
35. C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo, A. M. Klein, Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, eaaw3381 (2020).
36. B. Raj, D. E. Wagner, A. McKenna, S. Pandey, A. M. Klein, J. Shendure, J. A. Gagnon, A. F. Schier, Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
37. A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, A. Regev, Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866.e17 (2016).
38. P. Datlinger, A. F. Rendeiro, C. Schmidl, T. Krausgruber, P. Traxler, J. Klughammer, L. C. Schuster, A. Kuchler, D. Alpar, C. Bock, Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
39. S. R. Srivatsan, M. C. Regier, E. Barkan, J. M. Franks, J. S. Packer, P. Grosjean, M. Duran, S. Saxton, J. J. Ladd, M. Spielmann, C. Lois, P. D. Lampe, J. Shendure, K. R. Stevens, C. Trapnell, Embryo-scale, single-cell spatial transcriptomics. *Science* **373**, 111–117 (2021).
40. G. A. Logsdon, M. R. Vollger, E. E. Eichler, Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
41. D. M. Fowler, S. Fields, Deep mutational scanning: A new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
42. A. R. Ollodart, C. C. Yeh, A. W. Miller, B. H. Shirts, A. S. Gordon, M. J. Dunham, Multiplexing mutation rate assessment: Determining pathogenicity of Msh2 variants in *Saccharomyces cerevisiae*. *Genetics* **218**, iyab058 (2021).
43. K. A. Matreyek, L. M. Starita, J. J. Stephany, B. Martin, M. A. Chiasson, V. E. Gray, M. Kircher, A. Khechaduri, J. N. Dines, R. J. Hause, S. Bhatia, W. E. Evans, M. V. Relling, W. Yang,

J. Shendure, D. M. Fowler, Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
44. C. C. Yeh, C. J. Amorosi, S. Showman, M. J. Dunham, PacRAT: A program to improve barcode-variant mapping from PacBio long reads using multiple sequence alignment. *Bioinformatics* **38**, 2927–2929 (2022).
45. A. Marchant, A. F. Cisneros, A. K. Dubé, I. Gagnon-Arsenault, D. Ascencio, H. Jain, S. Aubé, C. Eberlein, D. Evans-Yamamoto, N. Yachie, C. R. Landry, The role of structural pleiotropy and regulatory evolution in the retention of heteromers of paralogs. *eLife* **8**, e46754 (2019).
46. D. G. Gibson, L. Young, R. Y. Chuang, J. C. Venter, C. A. Hutchison, H. O. Smith, Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
47. T. Stuart, A. Srivastava, S. Madad, C. A. Lareau, R. Satija, Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
48. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
49. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
50. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, G. P. D. P. Subgroup, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

Acknowledgments: We thank the members of the Yachie laboratory at the University of British Columbia and the University of Tokyo for useful comments and discussions throughout the course of this study, especially S. King for proofreading the manuscript. We appreciate M. Dunham, C. Yeh, C. Amorosi, and A. Ollodart for assisting the PacBio read analyses. Sequencing read analysis was performed using the SHIROKANE Supercomputer at the University of Tokyo, Human Genome Center. **Funding:** This study is supported by Canada Research Chair Program by the Canadian Institutes for Health Research (to N.Y.), Genome British Columbia Pilot Innovation Fund PIF003 (to N.Y.), CIHR Coronavirus Variants Rapid Response Network (CoVaRR-Net) (to N.Y.), Japan Agency for Medical Research and Development (to N.Y.), Japan Society for the Promotion of Science DC2 Fellowship (to Y.K.), and Japan Society for the Promotion of Science DC1 Fellowship (to D.E.-Y.). **Author contributions:** Conceptualization: Y.K. and N.Y. Methodology: Y.K. and N.Y. Software: Y.K. and H.T. Investigation: Y.K. and D.E.-Y. Visualization: Y.K. Writing (original draft): Y.K. Writing (review and editing): Y.K. and N.Y. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. INTERSTELLAR is available and will be updated at https://github.com/yachielab/INTERSTELLAR. The current version of INTERSTELLAR is deposited to https://doi.org/10.5281/zenodo.7250991. All the codes used in this study are available at https://doi.org/10.5281/zenodo.7250991. Test codes are executable at https://colab.research.google.com/drive/1nuqPK_zQSXFXHu-9gZR5w9EfsQhH6ltl?usp=sharing. The accession numbers of the public datasets used in this study are listed in table S2. The RCP-PCR data generated in this study have been submitted to the NCBI BioProject database (www.ncbi.nlm.nih.gov/bioproject) with the accession number PRJNA767068.

Submitted 1 June 2022
Accepted 30 November 2022
Published 4 January 2023
10.1126/sciadv.add2793