

De novo transformer modeling improves recovery of genetic cell types from sparse single-cell RNA sequencing

Jack M. Craig^{1,2}, M. Erfan Mowlaei^{1,3}, Sayaka Miura⁴,
Xinghua Shi^{1,3}, & Sudhir Kumar^{1,2,3*}

¹Institute for Genomics and Evolutionary Medicine, Temple University

²Department of Biology, Temple University

³Department of Computer & Information Sciences, Temple University

⁴Department of Biology, University of Mississippi

***Correspondence:** s.kumar@temple.edu

Abstract

Single-cell RNA sequencing (scRNA-seq) simultaneously provides gene-expression profiles and genetic variants from individual cells, creating an opportunity to relate cellular phenotypes to their somatic evolutionary histories. However, delineation of genetic type (GTs) from scRNA-seq remains difficult because most variant positions are unobserved in individual cells and the observed base calls contain substantial false-positive and false-negative errors. We evaluated some existing phylogenetic and imputation methods using one simulated dataset and two empirical tumor datasets. We found that extreme sparsity prevented reliable recovery of known or independently inferred GTs when multiple GTs were present. This led us to adapt the STICI transformer architecture to train a separate model de novo on each sparse cell-variant (CV) matrix. These data-specific models predicted millions of missing bases, greatly reducing matrix sparsity. Phylogenetic analyses of the imputed CV matrices showed substantially improved recovery of GTs in both simulated and empirical datasets. In the empirical dataset, transformer-based analysis also suggested finer-scale genetic structure within some previously reported GTs that was not apparent with the existing methods. These results demonstrate that highly sparse scRNA-seq datasets contain substantially more recoverable lineage information than previously appreciated and that de novo transformer modeling provides an effective approach for recovering much of this hidden information. Nevertheless, sequencing errors persisted, limiting reconstruction of cellular lineage structure and leaving significant room for methodological improvement before expression phenotypes can be examined reliably in the context of their cellular evolutionary relationships.

Introduction

Single-cell RNA sequencing (scRNA-seq) has transformed the study of cellular heterogeneity by enabling genome-wide measurement of gene expression in individual cells (Mereu et al., 2020; Moravec et al., 2023). Analyses of these expression profiles have yielded fundamental insights into cellular differentiation during development, tissue organization, and the initiation and progression of diseases such as cancer (Potter, 2018; Lim et al., 2020; Tirosh and Suva, 2024; Boxer et al., 2025). As a result, scRNA-seq has become one of the principal technologies for investigating tumor evolution and cellular diversity.

Beyond measuring gene expression, scRNA-seq also captures expressed genetic variants carried by individual cells. Consequently, every scRNA-seq experiment contains two complementary sources of biological information: gene-expression profiles that describe the functional state of cells and somatic genetic variants that record their evolutionary history. Current analyses routinely exploit the first source by clustering cells into expression-based cell types (ETs) (Hwang et al., 2018; Kiselev et al., 2019; Zhang et al., 2019; Heumos et al., 2023). The second source has the potential to delineate genetically related groups of cells, referred to here as genetic types (GTs), and reconstruct their evolutionary relationships. Joint analysis of ETs and GTs would make it possible to determine whether similar expression states arise through common ancestry, transitions during cellular differentiation, or convergent evolution among genetically distinct cell populations (Patel et al., 2014; Chung et al., 2017; Puram et al., 2017; Kim et al., 2020; Huzar et al., 2022). Such analyses would provide a direct framework for understanding how genetic evolution gives rise to somatic cellular phenotypes during normal development and disease progression.

Current studies often infer GTs using large-scale genomic alterations, particularly copy-number alterations (CNAs), because these events can be detected more reliably than individual nucleotide variants in scRNA-seq data (Hadi et al., 2020; Tarabichi et al., 2021). Consequently, CNA-defined cell clusters have become a practical surrogate for GTs in studies of tumor evolution (Gao et al., 2021). While highly informative, CNAs represent relatively infrequent mutational events and therefore provide only coarse genetic resolution. By contrast, single-nucleotide variants (SNVs) accumulate more frequently. If accurately recovered from scRNA-seq datasets, SNVs would enable substantially finer delineation of GTs and provide a richer record of cellular evolutionary history than CNAs alone.

The expressed SNVs detected in scRNA-seq data can be organized into cell-variant (CV) matrices, in which cells are rows and variant sites are columns (**Fig. 1**). In principle, molecular phylogenetic methods can then be applied to infer evolutionary relationships among cells and delineate GTs. In practice, however, scRNA-seq detects only a small

fraction of the transcripts expressed in each cell, and only a subset of these transcripts contain informative variants (Ziegenhain et al., 2017; Mereu et al., 2020; Chen et al., 2021). Consequently, CV matrices are extremely sparse, often containing missing observations for a vast majority of cell-site combinations even in high-quality datasets (Sparta et al., 2024). The data are further compromised by false-positive and false-negative base calls introduced during sequencing and variant-calling procedures (Svensson et al., 2017; Schnepf et al., 2019; Lu et al., 2021). Thus, extensive missing data and sequencing errors severely limit the recovery of cellular genetic relationships from scRNA-seq datasets (Liu et al., 2023; Moravec et al., 2023; Dou et al., 2024).

Several computational methods have been developed to address missing data and sequencing errors in single-cell sequencing analyses. BEAM uses phylogenetic relationships among evolutionarily proximate cells to iteratively improve base assignments (Miura et al., 2018), PhylinSic employs probabilistic smoothing to infer missing information in scRNA-seq datasets (Liu et al., 2023), and CellPhy jointly models genotype evolution and sequencing error during phylogenetic inference (Kozlov et al., 2022). When sparsity becomes extreme, the initial evolutionary signal required for successful imputation or phylogenetic reconstruction may itself be insufficient, limiting accurate delineation of GTs (see *Results*).

Recent advances in transformer architectures suggest a fundamentally different strategy for recovering genetic information from sparse datasets. Transformer models have demonstrated an ability to identify complex dependencies among sequence positions and accurately predict missing genomic variants from incomplete observations (Devlin et al., 2018; Mowlaei et al., 2025). Their success raises the possibility that they may also recover missing somatic variants from sparse scRNA-seq datasets. This expectation is biologically plausible because somatic mutations are inherited through successive mitotic cell divisions without recombination. Consequently, related cells preserve characteristic patterns of variant co-occurrence that reflect their shared ancestry. Even when only a small fraction of variant positions is directly observed, these lineage-dependent patterns may provide sufficient information for transformer models to infer many of the missing genetic states.

We initially considered applying STICI, a transformer architecture developed for genotype imputation using population reference panels (Mowlaei et al., 2025). However, STICI was trained on germline variants collected from many individuals across diverse populations, whereas scRNA-seq datasets contain somatic variants that are largely unique to an individual tumor and therefore absent from existing reference panels. This distinction prompted us to investigate whether a transformer could instead be trained de novo using only the sparse CV matrix generated from a single scRNA-seq dataset. Such a data-specific model would learn patterns of somatic variant co-occurrence directly from the cells under investigation without relying on external genomic references.

Here, we first evaluate the ability of existing phylogenetic and imputation methods to recover GTs from highly sparse scRNA-seq CV matrices using one simulated dataset and two empirical tumor datasets. We then develop a data-specific STICI (dsSTICI) model, trained de novo for each dataset, and test whether transformer-based recovery of missing bases improves the delineation of GTs. Finally, we examine the limitations revealed by this proof-of-principle study. Our results establish the promise of transformer-based analysis while identifying sequencing-error correction as the principal remaining challenge for reconstructing cellular lineage history from sparse single-cell transcriptomic datasets.

Results

Existing methods fail to recover genetic types from highly sparse CV matrices

We first asked whether existing approaches can reliably delineate GTs from the highly sparse CV matrices generated by scRNA-seq experiments. To address this question, we simulated somatic evolution along a clonal phylogeny containing five genetically distinct clones (GTs), each represented by 1,000 genetically identical cells, with 500 variant positions. The CV matrix was modified to reproduce the characteristics of scRNA-seq data by replacing 95% of bases with the missing symbol (?) and imposing false-positive and false-negative sequencing errors characteristic of scRNA-seq profiles (see *Methods*). A molecular phylogeny was inferred using FastTree because of the large number of cells (Price et al., 2010).

The phylogeny inferred directly from the raw sparse CV matrix failed to recover the known clonal structure. Rather than forming five distinct groups, cells belonging to individual GTs were scattered throughout the phylogeny (**Fig. 2: A1-E1**), making reliable delineation of GTs impossible. This result is consistent with reports from empirical scRNA-seq datasets (Moravec et al., 2023).

To quantify cell clustering corresponding to GTs, we used two complementary measures: the neighborliness index (*NI*) and the distance ratio (*DR*). *NI* is the fraction of cells whose nearest neighbor in the phylogeny belongs to the same GT. An *NI* of 1.0 indicates that every cell is adjacent to a member of the same GT, whereas lower values indicate increasingly poor local clonal clustering. *DR* is the average pairwise patristic distance between cells from different GTs in the inferred phylogeny divided by the average pairwise distance among cells within the same GT. Higher *DR* values indicate stronger separation of GTs and tighter clustering of cells belonging to the same GT (see *Methods*).

In the phylogeny inferred from the raw CV matrix (raw phylogeny), *NI* ranged from 0.34 to 0.90, with an average of 0.63 across GTs. *DR* values were also low and, for some GTs, fell below 1.0, indicating that cells from different GTs were, on average, phylogenetically

closer than members of the same GT. Thus, the raw phylogeny showed quantitatively poor recovery of the known GTs.

We next evaluated BEAM, PhylinSic, and CellPhy. BEAM produced few confident imputations at a Bayesian posterior probability (*BPP*) threshold of 0.7 because its procedure depends on the assumption that the raw phylogeny already contains sufficient evolutionary signal. This assumption is often satisfied for datasets obtained using single-cell DNA sequencing (Miura et al., 2018), in which the characteristic CV matrices have less than 50% sparsity, whereas scRNA-seq datasets can exceed 90% sparsity. After BEAM analysis, CV matrix sparsity increased slightly to 95.1%, and average NI changed by only 0.01. For this reason, cells are dispersed as much as in the raw phylogeny (**Fig. 2: A2-E2**).

PhylinSic yielded an average NI of 0.63 and an average DR of 1.63, little better than the raw phylogeny (**Fig. 2: A4-E4**). It shows slight improvement in clustering GT #1 and #4 cells, but GT #5 cells appear more dispersed. But CellPhy offered noticeable improvement for some GTs (**Fig. 2: A3-E3**). GT #5 (purple) becomes a remarkably cohesive cluster, GT #4 (blue) improves substantially, and GT #3 improves modestly. But GT #1 and #2 become worse. Overall, the average NI was lower than that of the raw phylogeny, whereas the average DR improved modestly (**Fig. 2F** and **2G**, respectively). Overall, none of the tested methods consistently recovered the known GTs of the simulated tumor.

A de novo transformer model recovers information from sparse CV matrices

The poor performance of existing methods prompted us to ask whether information hidden within sparse CV matrices could instead be recovered using transformer-based modeling. As noted earlier, the pretrained STICI model could not be applied directly because it was trained on panels of genomes harboring germline variation in human populations (Mowlaei et al., 2025). Somatic variants are personal in nature and differ among individuals and tumors, so their patterns are not expected to match the population variation used to train STICI.

Therefore, we trained a data-specific STICI model, termed dsSTICI, separately on each CV matrix (see *Methods*). During training, 90% of the non-missing bases in each cell sequence were randomly masked, and the model was trained to predict the masked bases from the remaining observations. In this way, the model learned patterns of somatic variant co-occurrence directly from the cells being analyzed rather than from an external reference panel. Notably, we never used the full CV matrix during dsSTICI training.

For the simulated dataset, model training required 3.3 hours on an NVIDIA H100 GPU. At a transformer base probability (*TBP*) threshold of 0.7, dsSTICI predicted 2,179,113 of the missing bases and changed ten bases present in the input CV matrix. Overall, sparsity

decreased from 95% to 7.8%. The missing nucleotides were filled with the correct base with 94.5% overall accuracy, which included 99.5% accuracy in predicting the original alleles and 78.5% accuracy in predicting variant alleles across all GTs. These results demonstrate that a transformer trained de novo on a single sparse CV matrix can recover a substantial fraction of the matrix's missing genetic information.

Recovery of missing bases substantially improves genetic cell typing

Simulated dataset

We next determined whether recovering missing bases improved the delineation of GTs. The phylogeny inferred from the dsSTICI-CV matrix showed a marked improvement in the clustering of cells belonging to the same GT (**Fig. 3A-F**). Average *NI* increased from 0.63 to 0.89, with *NI* exceeding 0.70 for every GT. Average *DR* increased to 3.34, revealing large separations between GTs (**Fig. 3G and H**). Thus, dsSTICI substantially improved recovery of simulated GTs.

Ovarian cancer dataset

We applied dsSTICI to an ovarian cancer dataset containing 1,842 cells and 2,970 SNVs (Vázquez-García et al., 2022). Two GTs were detected from CNAs using CopyKAT (Gao et al., 2021), which served as an external benchmark for evaluating GT recovery. The initial CV matrix was 86.0% sparse, and its raw phylogeny had an average *NI* of 0.84 and *DR* of 1.11, with cells from the two CNA-GTs interspersed (**Fig. 4: A1, B1**).

dsSTICI training required approximately one hour on an NVIDIA H100 GPU. At *TBP* $\geq$ 0.7, dsSTICI analysis predicted 3,956,568 missing bases and changed 2,755 bases present in the input matrix, reducing sparsity from 86.0% to 13.8%. The phylogeny inferred from the completed matrix showed greatly improved separation of the two CNA-GTs (**Fig. 4: A4, B4**). Average *NI* increased to 1.0, whereas *DR* increased from 1.11 to 1.79. For the first GT, *DR* more than doubled from 1.12 to 2.28 (**Fig. 4C, D**).

BEAM and CellPhy also recovered the two major CNA-GTs reasonably well (**Fig. 4: A2:B2 and A3:B3**, respectively). Only a few cells were misclassified, resulting in an average *NI* of 0.97 for both, and *DR* values larger than 1 (1.23 for BEAM and 1.19 for CellPhy). However, dsSTICI analysis performed better, producing a higher overall *DR* and more distinct major branches (**Fig. 4: A4-B4**). PhylinSic could not be run on a dataset of this size. Thus, existing approaches can perform well when only two major GTs must be distinguished, although dsSTICI produced a clearer global separation.

Triple-negative breast cancer dataset

We next analyzed a triple-negative breast cancer (TNBC) dataset containing 3,477 cells, 19,576 sites, and six CNA-defined GTs (Gao et al., 2021). The CV matrix was 92% sparse, producing a raw phylogeny with an average *NI* of 0.27 and *DR* of 1.04 (**Fig. 5: A1:F1**). BEAM improved local clustering for several GTs, but the six CNA-GTs were

incompletely separated across the phylogeny, yielding average *NI* and *DR* values of 0.65 and 1.17, respectively (**Fig. 5: A2:F2**). PhylinSic could not handle this dataset, and CellPhy analyses did not complete after several weeks. Therefore, the existing methods that performed reasonably well in the two-GT ovarian dataset were much less effective for this multi-GT problem.

Training dsSTICI required approximately ten hours on an NVIDIA H100 GPU. At *TBP* $\geq$ 0.7, the model predicted 40,255,930 missing bases and changed 7,745 bases in the input CV matrix, reducing sparsity to 33%. The phylogeny inferred from the completed matrix showed marked improvement (**Fig. 5: A3-C3**), with average *NI* increasing from 0.27 to 0.68 and *DR* from 1.04 to 1.46 (**Fig. 5: G, H**). The fourth and sixth GTs showed particularly strong clustering (**Fig. 5: D and F**).

In addition to improving the separation of the six major CNA-GTs, dsSTICI placed some cells of the same CNA-GT into small, cohesive groups, separated by multiple intervening branches, for each GT (**Fig. 5**). These patterns suggest SNV-defined substructure within the broader CNA-GTs and were not apparent in the raw or BEAM phylogenies. Because independent SNV-based clonal assignments are unavailable, these groups should be considered as possibilities rather than confirmed biological lineages.

Therefore, transformer-based recovery of missing sequence information improved genetic cell typing, even in a large, complex tumor dataset.

Sequencing errors remain a major barrier to complete recovery of GTs

Although dsSTICI substantially improved GT recovery, the resulting phylogenies did not perfectly reconstruct the known cellular relationships. In the simulated dataset, not all cells belonging to the same GT formed a single clade, and many cells retained long terminal branches despite the simulated GTs consisting of genetically identical cells (**Fig. 3**). These long terminal branches were caused largely by cell-specific variants. They reflected residual sequencing errors rather than genuine evolutionary divergence.

To investigate this limitation, we compared the distribution of dsSTICI transformer base probabilities for missing bases and simulated sequencing errors (**Fig. 6**). Missing positions were often assigned high *TBP*s, enabling confident imputation of previously unobserved bases. However, nearly all incorrect bases introduced during simulation as sequencing errors also received high *TBP*s. Rather than correcting them, dsSTICI generally propagated them into the completed CV matrix with high confidence.

This behavior follows from the training procedure, as the same sparse and noisy matrix is used both to train and to apply the model. Erroneous observed bases are therefore treated as valid training targets. The present implementation is consequently effective for missing-data imputation but has limited ability to distinguish true somatic variants from sequencing or variant-calling errors. These analyses identify error correction, rather than

imputation alone, as the major remaining challenge for complete recovery of cellular lineage structure.

Discussion

The increasing availability of scRNA-seq data offers substantial promise for studying cellular growth, development, and disease progression. Yet these data have not been effectively harnessed for genetic typing because expressed SNVs are observed only sparsely and are further obscured by sequencing errors. Our analyses show that these limitations do not eliminate the genetic information contained in scRNA-seq datasets. Instead, they make that information difficult to access using existing phylogenetic and imputation methods. By training a transformer de novo on individual CV matrices, we predicted millions of missing bases and substantially improved delineation of GTs in simulated and empirical datasets.

The limited success of existing methods tested here illustrates why recovering GTs from scRNA-seq remains difficult. These approaches depend, in different ways, on the information already present in the observed CV matrix. At the extreme levels of sparsity characteristic of scRNA-seq, the initial phylogenetic signal is often too weak to support accurate reconstruction, smoothing, or iterative improvement. However, their performance also depended on the complexity of the clonal structure. BEAM and CellPhy recovered the two major CNA-defined GTs reasonably well in the ovarian dataset.

In contrast, BEAM was much less successful in separating six GTs in the TNBC dataset, and CellPhy and PhylinSic were computationally infeasible at that scale. Thus, the difficulty is not simply the presence of missing data; it increases when sparse observations must resolve many genetically related lineages. The poor performance of these methods, therefore, does not imply that the data lack lineage information; rather, it indicates that the information is distributed across sparse observations in a form that conventional approaches cannot readily exploit.

The success of dsSTICI is biologically understandable because somatic cells divide by mitosis without recombination. Mutations are therefore inherited together across generations of cells, preserving characteristic patterns of variant co-occurrence. Successive mutations also generate nested combinations of variants during clonal diversification. A transformer can consider dependencies among many sites simultaneously, allowing it to use these distributed patterns to predict missing bases. The marked reduction in sparsity and improved clustering of known or independently inferred GTs suggest that the model captured biologically meaningful patterns of shared somatic variation. Therefore, dsSTICI can leverage multi-site contexts that substantially improve its predictions, compared with methods that do not account for dependence among sites.

and treat each position in the CV matrix as independent information during statistical analysis; see also (Kumar et al., 2026).

However, improved base completion did not yield perfect delineation of GTs because the model rarely corrected erroneous observed bases, which often received high transformer base probabilities, thereby producing long cell-specific branches. Because training was performed on the same noisy matrix whose missing elements were subsequently imputed, observed errors were treated as correct targets and could be memorized by the model. Such memorization is a known issue with noisy input data (Pompanon et al., 2005; Harutyunyan et al., 2020). Thus, the present results distinguish two related but different problems: imputing bases that were not observed and correcting bases that were observed incorrectly. dsSTICI was effective primarily for the former even with a minuscule fraction of available bases in the CV matrix.

Importantly, however, while CNA-defined GTs provide a useful independent benchmark in tests of empirical datasets, they are not exact ground truth for SNV-defined GTs. First, CNA inference is not without error (Gao et al., 2021; Song et al., 2025), and GTs based on CNAs represent a coarser view of evolution, whereas SNVs can provide substructure within a CNA-GT. This distinction is especially relevant in the TNBC analysis. The dsSTICI phylogeny contained several small, cohesive groups of cells from the same CNA-GTs, separated by substantial internal branches, whereas a comparable structure was not evident in the raw or BEAM phylogenies. Such groups may represent finer SNV-defined subGTs nested within broader CNA-GTs. Their phylogenetic separation makes them more suggestive of coherent lineage divergence than isolated cell-specific errors, although matched DNA sequencing or other independent genetic evidence will be required for validation. Accordingly, agreement with CNA-GTs measures recovery of major genetic groups but may underestimate the finer evolutionary structure accessible through SNVs.

Even with these limitations, the biological implications are substantial. Reliable GT assignment would allow expression-based cell states to be interpreted in their evolutionary context. Investigators could determine whether similar transcriptional programs arose through common descent or independently in genetically distinct lineages, identify state transitions accompanying clonal expansion, and examine the origins of treatment resistance or metastatic dissemination using the same cells (Jun et al., 2023; Liu et al., 2023). In this sense, scRNA-seq contains two coupled records of cellular biology: a functional record in gene expression and an evolutionary record in somatic variation. We find that transformer modeling shows significant potential to recover the latter from highly incomplete data.

Conclusions

The present study establishes a proof of principle but also defines the principal obstacles that remain. Recovering missing bases alone is insufficient for complete genetic cell typing when erroneous observed bases are retained with high confidence. Future methods, particularly advanced transformer architectures, will therefore need to be developed to distinguish uncertainty due to missing observations from that due to sequencing and variant-calling errors. Information such as read depth, allele balance, base quality, and mapping quality may help determine how strongly an observed call should contribute to training and inference. Methods will also need to be tested across datasets containing different numbers and arrangements of GTs, because the contrast between the two-GT and six-GT analyses indicates that performance can decline as lineage complexity increases.

In summary, highly sparse scRNA-seq datasets retain substantially more recoverable genetic information than previously appreciated. De novo transformer modeling provides an effective first step toward accessing this information and improving genetic cell typing. It may also expose candidate genetic structure below the resolution of CNA-based classifications. Nevertheless, substantial room for improvement remains, particularly in sequencing-error correction, validation of candidate subGTs, and reliable analysis of complex multi-GT tumors.

Methods

Simulated data

We simulated a sparse, error-prone CV matrix by adapting an existing simulated dataset from a previous study that benchmarked methods for inferring cellular phylogeny (Miura et al., 2023). This dataset consisted of five GTs, each comprising 1,000 cells, with each cell characterized by 500 SNVs. The original simulation incorporated a 1% false-positive rate and a 20% false-negative rate (Miura et al., 2023). To mimic the sparsity observed in empirical scRNA-seq data, we randomly replaced bases with missing values until overall sparsity reached 95%. Because the original dataset encoded only the presence or absence of mutations, we randomly assigned nucleotide states to convert it into sequence data.

Empirical data

We obtained empirical scRNA-seq data from prior studies of metastatic ovarian cancer (OC) (Vázquez-García et al., 2022) and triple-negative breast cancer (TNBC) (Gao et al., 2021). For both datasets, raw sequencing reads (FASTQ files) were processed into gene-expression matrices using the Cell Ranger Count pipeline from 10x Genomics

(Genomics, 2019). Position-sorted read alignments (BAM files) were also generated for downstream SNV calling.

For gene-expression analysis, ambient RNA contamination, doublets or multiplets, and low-quality cells with elevated mitochondrial RNA content were removed using SoupX (Young and Behjati, 2020), DoubletFinder (McGinnis et al., 2019), and miQC (Hippen et al., 2021), respectively. Cell-type annotation was performed using SingleR (Aran et al., 2019), with annotations from the original study for the OC dataset and from the Human Breast Cell Atlas for the TNBC dataset (Kumar et al., 2023). CNAs were inferred using CopyKAT (Gao et al., 2021). Ploidy status was estimated for each cell, and only cells inferred by CopyKAT to be aneuploid were retained for CNA-based GT annotation.

For SNV calling, ovarian cancer cells were used in the OC dataset, and luminal epithelial and basal mammary cells were used in the TNBC dataset. SNVs were identified using cellSNP-lite (Huang and Huang, 2021) with default settings, including a minimum minor-allele frequency of 0.1 and minimum read depth of 100. SNVs with frequencies greater than 0.5% were retained for phylogenetic inference. After processing, the OC dataset contained 1,842 cells and 2,970 SNVs, whereas the TNBC dataset contained 3,477 cells and 19,576 SNVs.

Existing Methods

We analyzed simulated and empirical datasets with BEAM (Miura et al., 2018b) using the BEAMfast.py implementation and default parameters. PhylinSic (Liu et al., 2023) was run on default parameters. CellPhy (Kozlov et al., 2022) was run using the “fast” method for the simulated dataset, as “full” and “search” methods were computationally infeasible. All three methods could be run on the simulated data. CellPhy could be run on the OC dataset, but neither PhylinSic nor CellPhy was completed for the TNBC data due to the dataset size.

Training and applying dsSTICI

The STICI architecture was originally designed for germline genomic imputation (Mowlaei et al., 2025), which we adapted for CV matrix completion. The original STICI model was trained using a complete reference panel of genome sequences from the 1000 Genomes Project. For scRNA-seq CV matrices, training and inference were instead performed separately for each sparse dataset to build a data-specific model, dsSTICI, which required a few alterations.

During training, 90% of the non-missing bases in each cell sequence were randomly masked, and the model was trained to predict the masked bases. For example, if a CV matrix was 95% sparse, 90% of the remaining 5% observed bases were masked during a training instance, leaving 0.5% of all positions available as input. We used categorical

cross-entropy and Kullback-Leibler divergence loss functions and ignored missing sites during training. We did not use the Mach-Rsq loss function employed in STICI because it was designed to improve genotype-dosage prediction and was not relevant to scRNA-seq base prediction (Mowlaei et al., 2025). The model achieved 93.4% accuracy in training. After training, dsSTICI produced a distribution of transfer base probabilities (TBP) for every site in each sequence. A base was assigned when TBP was at least 0.7. This procedure imputed missing bases and, in some cases, changed bases present in the input CV matrix.

Phylogenetic inference and assessment of clonal grouping

We used FastTree for phylogenetic analysis of simulated, empirical, and dsSTICI-completed CV matrices because it can analyze datasets containing thousands of cells (Price et al., 2010). We used a GTR model (-gtr) for full-base-resolution alignments in the OC and simulated datasets and the standard nucleotide option (-nt) for the binary TNBC alignment.

We quantified clonal grouping using two metrics. For the simulated dataset, true clonal identities were known. For empirical datasets, CNA-defined GT assignments inferred by CopyKAT were used as an external benchmark (Gao et al., 2021).

The neighborliness index (NI) is the fraction of cells assigned to a given GT whose nearest topological neighbor is also assigned to that GT. When the nearest neighbor is a clade containing multiple cells, the clonal assignment for that clade is taken to be the most common identity among its members, with ties broken at random. NI ranges from 0.0 to 1.0, with higher values indicating stronger local clustering of cells from the same GT.

The distance ratio (DR) is the average pairwise patristic distance between cells from different GTs divided by the average pairwise distance among cells from the same GT. Greater distances between GTs and shorter distances within GTs therefore produce higher DR values, indicating stronger recovery of clonal structure.

	SNV_1	SNV_2	SNV_3	SNV_4	SNV_5
Cell_1	?	?	?	?	?
Cell_2	?	?	?	A	T
Cell_3	T	?	?	?	?
Cell_4	?	?	?	?	?
Cell_5	?	G	?	?	?

Figure 1: A cell-variant (CV) matrix for a scRNA-seq dataset.

A scRNA-seq typically produces highly sparse CV matrices where some cells have base calls at a site (shaded boxes), whereas most positions are unobserved (“?”).

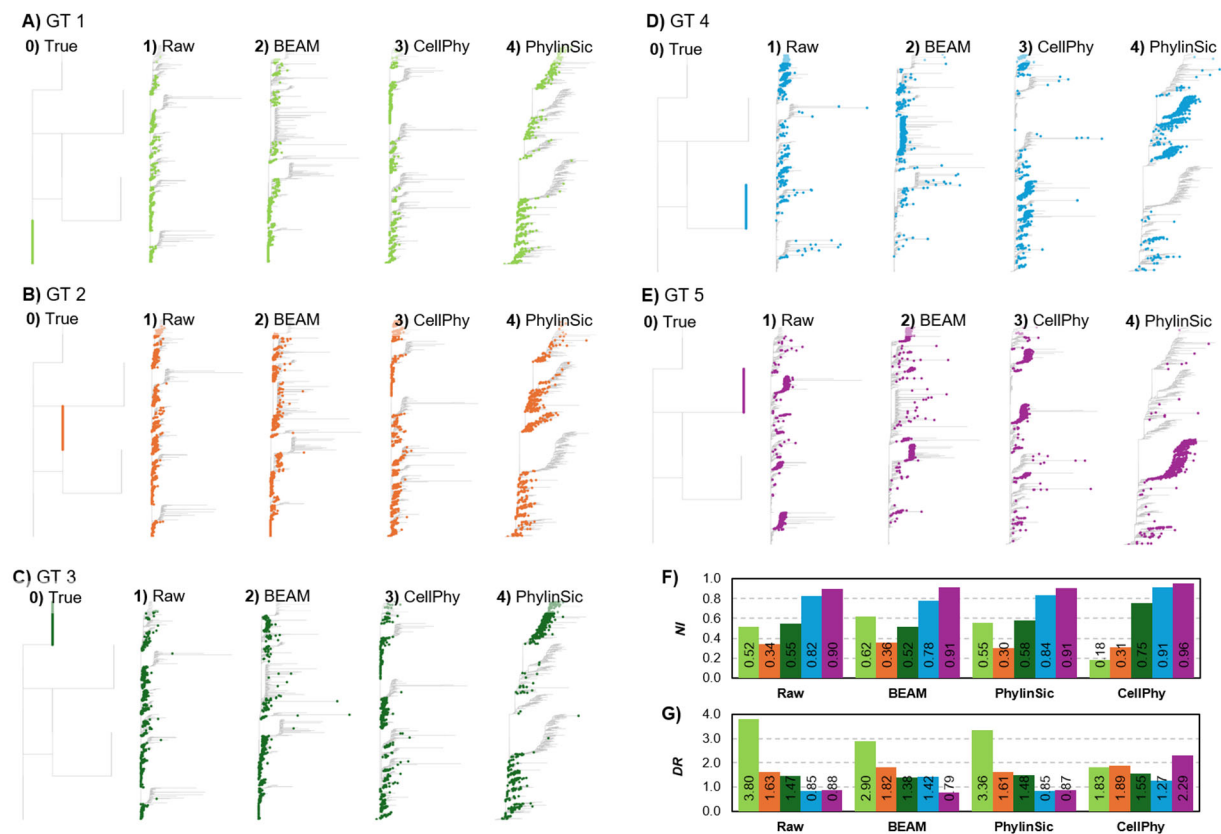


Figure 2: Distribution of cells associated with each color-coded GT (GT) in a simulated dataset after analysis by existing methods.

For each GT (A-E, color coded), the corresponding cells are highlighted on the phylogeny inferred **0)** from the true simulated CV matrix, with no sparsity or error added; **1)** from the CV matrix with sparsity and error added; **2)** after BEAM analysis; **3)** after CellPhy analysis; and **4)** after PhylinSic analysis. **F)** NI scores **G)** DR scores.

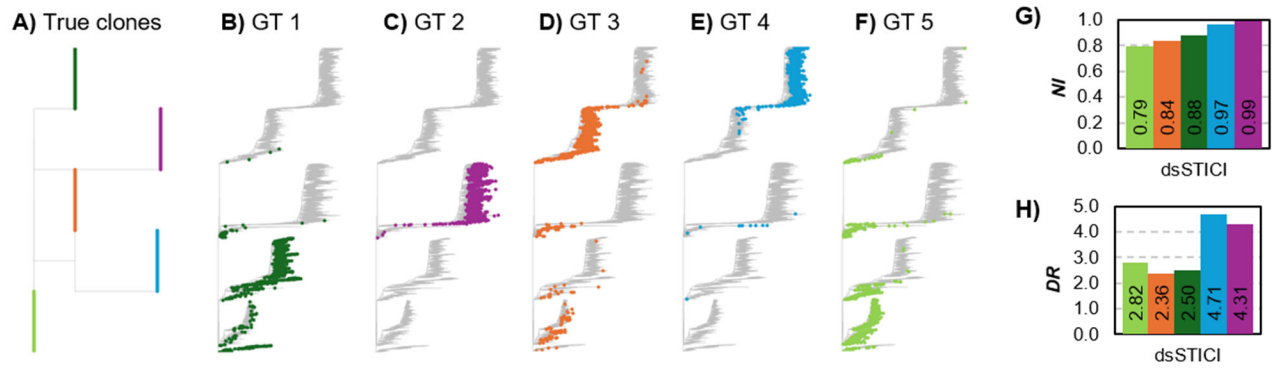


Figure 3: Recovery of simulated GTs after dsSTICI analysis.

A) Phylogeny inferred from the CV matrix after dsSTICI imputations, with cells colored by GT. **B)** NI scores, **C)** DR scores.

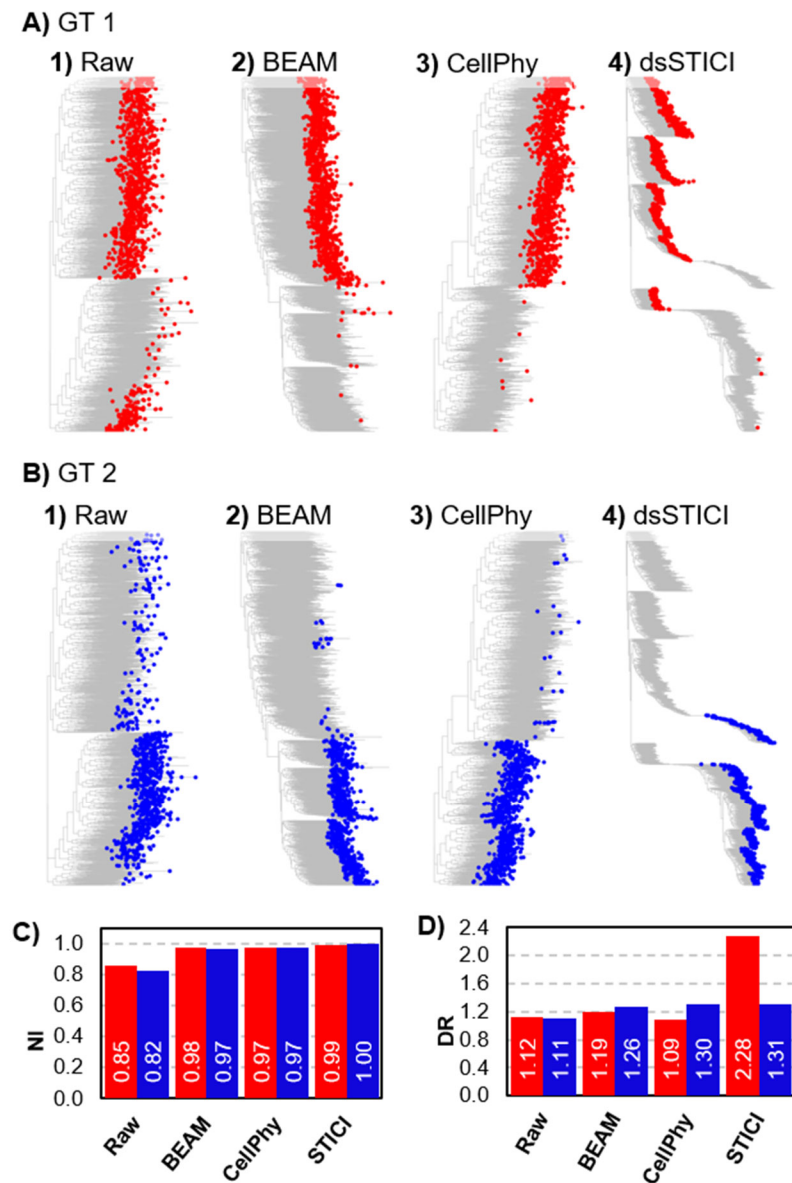


Figure 4: Recovery of two CNA-defined GTs in the empirical ovarian cancer dataset.

For each GT (**A** and **B**, color coded), the corresponding cells are highlighted on the phylogeny inferred **1)** from the raw scRNA-seq dataset; **2)** after BEAM analysis; **3)** after CellPhy analysis; **4)** after dsSTICI analysis. **C)** NI scores. **D)** DR scores.

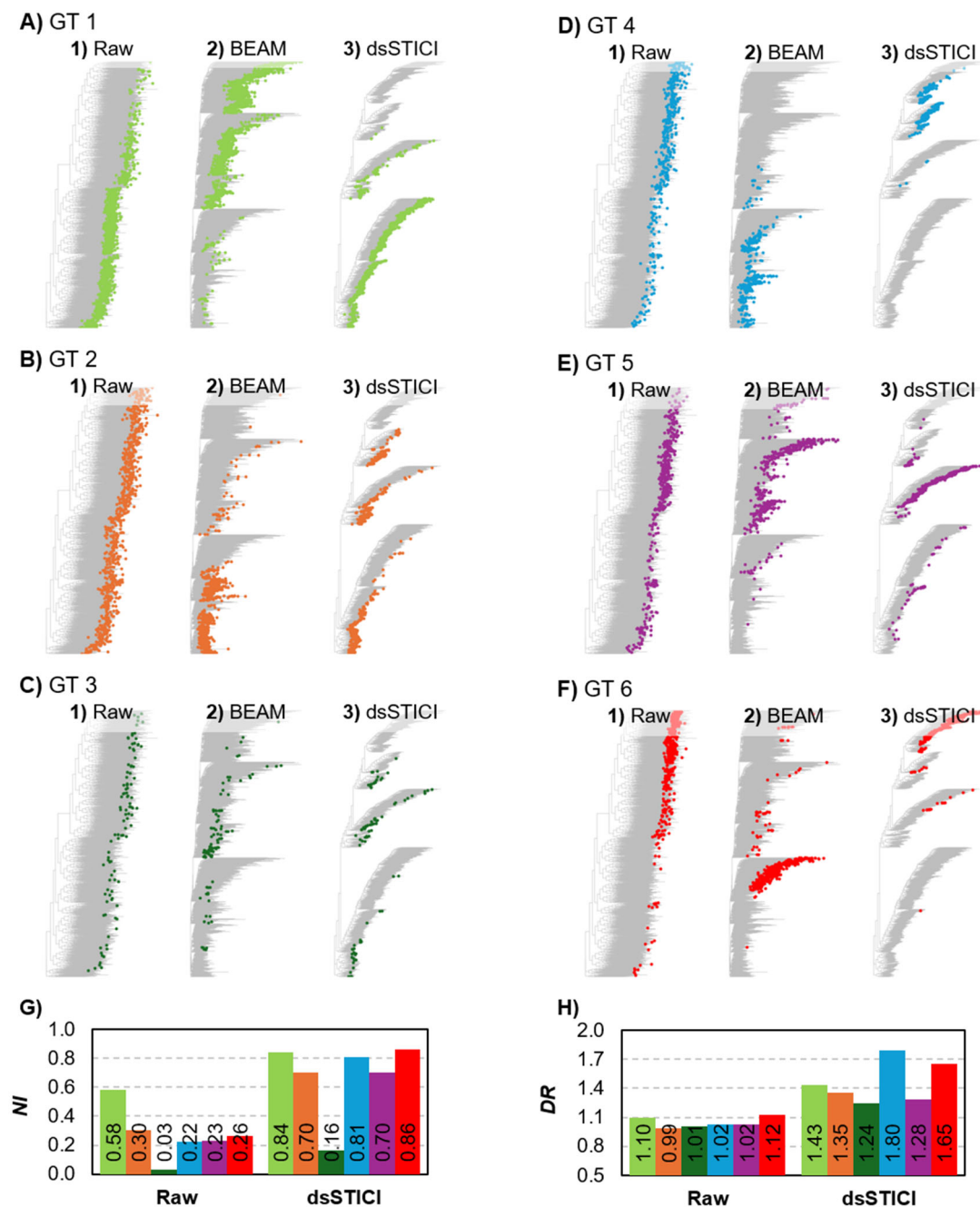


Figure 5: Recovery of six CNA-defined GTs in the empirical triple-negative breast cancer dataset.

For each GT (A-F, color coded), the corresponding cells are highlighted on the phylogeny inferred **1)** from the raw scRNA-seq dataset; **2)** after BEAM analysis; **3)** after dsSTICI analysis. **G)** NI scores. **H)** DR scores.

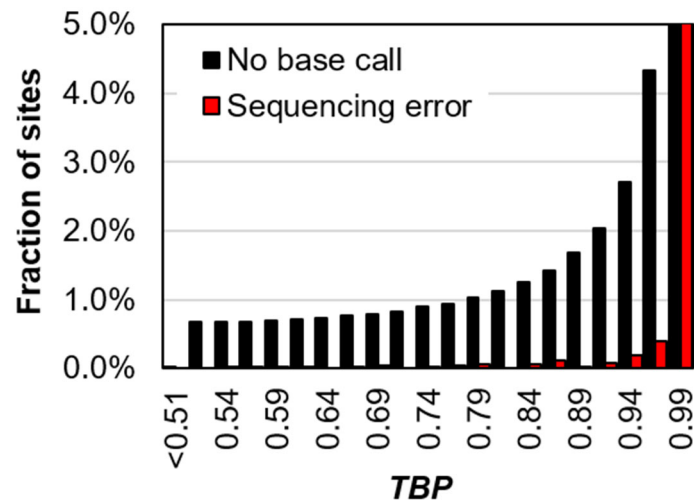


Figure 6. Distribution of transformer base probabilities.

Bars show the fraction of sites in each probability interval for missing bases (black) and simulated sequencing errors (red). The axis is truncated for clarity.

References

- Boxer, E., Feigin, N., Tschernichovsky, R., Darnell, N. G., Greenwald, A. R., Hoefflin, R., et al. (2025). Emerging clinical applications of single-cell RNA sequencing in oncology. *Nat. Rev. Clin. Oncol.* 22, 315–326.
- Chen, W., Zhao, Y., Chen, X., Yang, Z., Xu, X., Bi, Y., et al. (2021). A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples. *Nat. Biotechnol.* 39, 1103–1114.
- Chung, W., Eum, H. H., Lee, H.-O., Lee, K.-M., Lee, H.-B., Kim, K.-T., et al. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* 8, 15081.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional Transformers for language understanding. *arXiv [cs.CL]*. doi: [10.48550/arXiv.1810.04805](https://doi.org/10.48550/arXiv.1810.04805)
- Dou, J., Tan, Y., Kock, K. H., Wang, J., Cheng, X., Tan, L. M., et al. (2024). Single-nucleotide variant calling in single-cell sequencing data with Monopogen. *Nat Biotechnol* 42, 803–812.
- Gao, R., Bai, S., Henderson, Y. C., Lin, Y., Schalck, A., Yan, Y., et al. (2021). Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.* 39, 599–608.
- Genomics, 10x (2019). *Technical Note – Resolving Cell Types as a Function of Read Depth and Cell Number*, Document Number CG000148.
- Hadi, K., Yao, X., Behr, J. M., Deshpande, A., Xanthopoulos, C., Tian, H., et al. (2020). Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell* 183, 197–210.e32.
- Harutyunyan, H., Reing, K., Ver Steeg, G., and Galstyan, A. (2020). Improving generalization by controlling label-noise information in neural network weights., in *International Conference on Machine Learning*, (PMLR), 4071–4081.
- Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., et al. (2023). Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* 24, 550–572.
- Hippen, A. A., Falco, M. M., Weber, L. M., Erkan, E. P., Zhang, K., Doherty, J. A., et al. (2021). miQC: An adaptive probabilistic framework for quality control of single-cell RNA-sequencing data. *PLoS Comput. Biol.* 17, e1009290.
- Huang, X., and Huang, Y. (2021). Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics* 37, 4569–4571.

- Huzar, J., Kim, H., Kumar, S., and Miura, S. (2022). MOCA for integrated analysis of gene expression and genetic variation in single cells. *Front. Genet.* 13, 831040.
- Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* 50, 1–14.
- Jun, S.-H., Toosi, H., Mold, J., Engblom, C., Chen, X., O'Flanagan, C., et al. (2023). Reconstructing clonal tree for phylo-phenotypic characterization of cancer using single-cell transcriptomics. *Nat. Commun.* 14, 982.
- Kim, N., Kim, H. K., Lee, K., Hong, Y., Cho, J. H., Choi, J. W., et al. (2020). Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.* 11, 2285.
- Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* 20, 273–282.
- Kozlov, A., Alves, J. M., Stamatakis, A., and Posada, D. (2022). CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data. *Genome Biol.* 23, 1–30.
- Kumar, S., Alibutud, R., Allard, J., Mowlaei, M. E., and Shi, X. (2026). Transforming molecular evolutionary analyses with protein foundation models. *Nature Methods*, (in review).
- Kumar, T., Nee, K., Wei, R., He, S., Nguyen, Q. H., Bai, S., et al. (2023). A spatially resolved single-cell genomic atlas of the adult human breast. *Nature* 620, 181–191.
- Lim, B., Lin, Y., and Navin, N. (2020). Advancing cancer research and medicine with single-cell genomics. *Cancer Cell* 37, 456–470.
- Liu, X., Griffiths, J. I., Bishara, I., Liu, J., Bild, A. H., and Chang, J. T. (2023). Phylogenetic inference from single-cell RNA-seq data. *Sci. Rep.* 13, 12854.
- Lu, T., Park, S., Zhu, J., Wang, Y., Zhan, X., Wang, X., et al. (2021). Overcoming expressional drop-outs in lineage reconstruction from single-cell RNA-sequencing data. *Cell Rep.* 34, 108589.
- McGinnis, C. S., Murrow, L. M., and Gartner, Z. J. (2019). DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* 8, 329–337.e4.
- Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D. J., Álvarez-Varela, A., et al. (2020). Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* 38, 747–755.
- Miura, S., Dolker, T., Sanderford, M., and Kumar, S. (2023). Improving cellular phylogenies through the integrated use of mutation order and optimality principles.

Comput Struct Biotechnol J 21, 3894–3903.

- Miura, S., Huuki, L. A., Buturla, T., Vu, T., Gomez, K., and Kumar, S. (2018). Computational enhancement of single-cell sequences for inferring tumor evolution. *Bioinformatics* 34, i917–i926.
- Moravec, J. C., Lanfear, R., Spector, D. L., Diermeier, S. D., and Gavryushkin, A. (2023). Testing for Phylogenetic Signal in Single-Cell RNA-Seq Data. *J. Comput. Biol.* 30, 518–537.
- Mowlaei, M. E., Li, C., Jamialahmadi, O., Dias, R., Chen, J., Jamialahmadi, B., et al. (2025). STICI: Split-Transformer with integrated convolutions for genotype imputation. *Nat. Commun.* 16, 1218.
- Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 1396–1401.
- Pompanon, F., Bonin, A., Bellemain, E., and Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nat. Rev. Genet.* 6, 847–859.
- Potter, S. S. (2018). Single-cell RNA sequencing for the study of development, physiology and disease. *Nat. Rev. Nephrol.* 14, 479–492.
- Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2: approximately maximum-likelihood trees for large alignments. *PLoS One* 5, e9490.
- Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 171, 1611–1624.e24.
- Schnepp, P. M., Chen, M., Keller, E. T., and Zhou, X. (2019). SNV identification from single-cell RNA sequencing data. *Hum. Mol. Genet.* 28, 3569–3583.
- Song, M., Ma, S., Wang, G., Wang, Y., Yang, Z., Xie, B., et al. (2025). Benchmarking copy number aberrations inference tools using single-cell multi-omics datasets. *Brief. Bioinform.* 26, bbaf076.
- Sparta, B., Hamilton, T., Natesan, G., Aragonés, S. D., and Deeds, E. J. (2024). Binomial models uncover biological variation during feature selection of droplet-based single-cell RNA sequencing. *PLoS Comput. Biol.* 20, e1012386.
- Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., et al. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14, 381–387.
- Tarabichi, M., Salcedo, A., Deshwar, A. G., Ni Leathlobhair, M., Wintersinger, J., Wedge, D. C., et al. (2021). A practical guide to cancer subclonal reconstruction

from DNA sequencing. *Nat. Methods* 18, 144–155.

Tirosh, I., and Suva, M. L. (2024). Cancer cell states: Lessons from ten years of single-cell RNA-sequencing of human tumors. *Cancer Cell* 42, 1497–1506.

Vázquez-García, I., Uhlig, F., Ceglia, N., Lim, J. L. P., Wu, M., Mohibullah, N., et al. (2022). Ovarian cancer mutational processes drive site-specific immune evasion. *Nature* 612, 778–786.

Young, M. D., and Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* 9, giaa151.

Zhang, A. W., O’Flanagan, C., Chavez, E. A., Lim, J. L. P., Ceglia, N., McPherson, A., et al. (2019). Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat. Methods* 16, 1007–1015.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., et al. (2017). Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* 65, 631–643.e4.