CANCER

scAllele: A versatile tool for the detection and analysis of variants in scRNA-seq

Giovanni Quinones-Valdez[1], Ting Fu[2], Tracey W. Chan[3]†, Xinshu Xiao[1,2,3,4,5,6]*

Single-cell RNA sequencing (scRNA-seq) data contain rich information at the gene, transcript, and nucleotide levels. Most analyses of scRNA-seq have focused on gene expression profiles, and it remains challenging to extract nucleotide variants and isoform-specific information. Here, we present scAllele, an integrative approach that detects single-nucleotide variants, insertions, deletions, and their allelic linkage with splicing patterns in scRNA-seq. We demonstrate that scAllele achieves better performance in identifying nucleotide variants than other commonly used tools. In addition, the read-specific variant calls by scAllele enables allele-specific splicing analysis, a unique feature not afforded by other methods. Applied to a lung cancer scRNA-seq dataset, scAllele identified variants with strong allelic linkage to alternative splicing, some of which are cancer specific and enriched in cancer-relevant pathways. scAllele represents a versatile tool to uncover multilayer information and previously unidentified biological insights from scRNA-seq data.

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) affords a unique glimpse into the transcriptome at the single-cell resolution, revealing great cellular heterogeneity (*1*). Although this type of data harbors rich information of a cell's transcriptome, most studies focused exclusively on gene expression without tackling other important aspects such as single-nucleotide variants (SNVs) (*2*) or allele-specific expression (*3*–*5*). In addition, previous analyses of genetic variants in scRNA-seq data used methods originally designed for bulk RNA-seq or DNA sequencing (DNA-seq) (*6*, *7*) because of the lack of tools specifically tailored for variant calls in scRNA-seq.

Variant calling in RNA poses substantial computational challenges. Typically, variant callers rely on resolving the haplotypes from the next-generation sequencing reads (*8*, *9*). However, this strategy has limited applications at the RNA level where alternative splicing, allele-specific expression, or RNA editing may affect minor allele frequencies of the variants or obscure the true haplotype proportion in the reads. Furthermore, in scRNA-seq, most expressed genes have shallow coverage (*10*, *11*), highlighting the need of accurate variant detection with limited reads. To date, no method exists that explicitly focuses on both SNVs and insertions/deletions (INDELs) in scRNA-seq.

Following their identification, the next major challenge is to link nucleotide variants to their potential molecular function. To this end, RNA-seq data have unique advantages given the afforded multilevel information: gene expression, transcript isoforms, and sequence variations. Using bulk RNA-seq, numerous studies leveraged this strength to uncover allelic bias of genetic variants in gene expression or splicing (*12*, *13*), which, for example, can lead to discovery of functional cis-acting variants that alter splicing (*13*–*16*). Despite their typical low coverage, scRNA-seq data provide similar multilevel information. However, no method exists to leverage these features of scRNA-seq to examine allelic association with alternative RNA processing, such as splicing.

Here, we introduce scAllele, a versatile tool that performs both variant calling and association analysis between variant alleles and alternative splicing using scRNA-seq. As a variant caller, scAllele reliably identifies SNVs and microindels (less than 20 bases) with low coverage. It implements RNA-friendly haplotype filtering by accounting for potential RNA editing sites and allele-specific splicing. Following variant calling, scAllele identifies significant associations between variant alleles and alternative splicing, which provides direct evidence of allele-specific splicing.

Using scRNA-seq data associated with well-characterized genotypes, we show that scAllele outperforms other commonly used methods, especially for microindel identification. We apply scAllele to scRNA-seq data derived from lung cancer samples. Our analysis identifies variants that have significant allelic linkage to splicing isoforms, some of which are enriched in cancer cells and cancer-relevant pathways. Thus, scAllele is an integrative analysis tool that uncovers multilevel information in scRNA-seq.

RESULTS

Algorithm overview

scAllele calls nucleotide variants via local reassembly (Fig. 1A). To scan variants in the entire transcriptome, we split the mapped reads into read clusters (RCs), defined as genomic intervals containing overlapping reads. Reads from each RC are subsequently decomposed into overlapping k-mers and reassembled into a directed de Bruijn graph (dBG). The reference genomic sequence is included in the reassembly to serve as the reference haplotype in the RC. The nodes of the graph represent k-mers derived from the read sequence. Two nodes with k-mers overlapping by k-1 bases are connected with a directed edge. The "bubbles" in the graph represent differences among all sequences including the reads and genome reference sequence.

To identify nucleotide variants, we first traverse the graph with a depth-first search (DFS) to identify nodes marking the beginning

[1]Department of Bioengineering, University of California, Los Angeles, Los Angeles, CA 90095, USA. [2]Molecular, Cellular, and Integrative Physiology Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90095, USA. [3]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90095, USA. [4]Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, CA 90095, USA. [5]Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA. [6]Institute for Quantitative and Computational Biosciences, University of California, Los Angeles, Los Angeles, CA 90095, USA.
*Corresponding author. Email: gxxiao@ucla.edu
†Current address: Calico Life Sciences, 1170 Veterans Blvd, South San Francisco, CA 94080, USA.

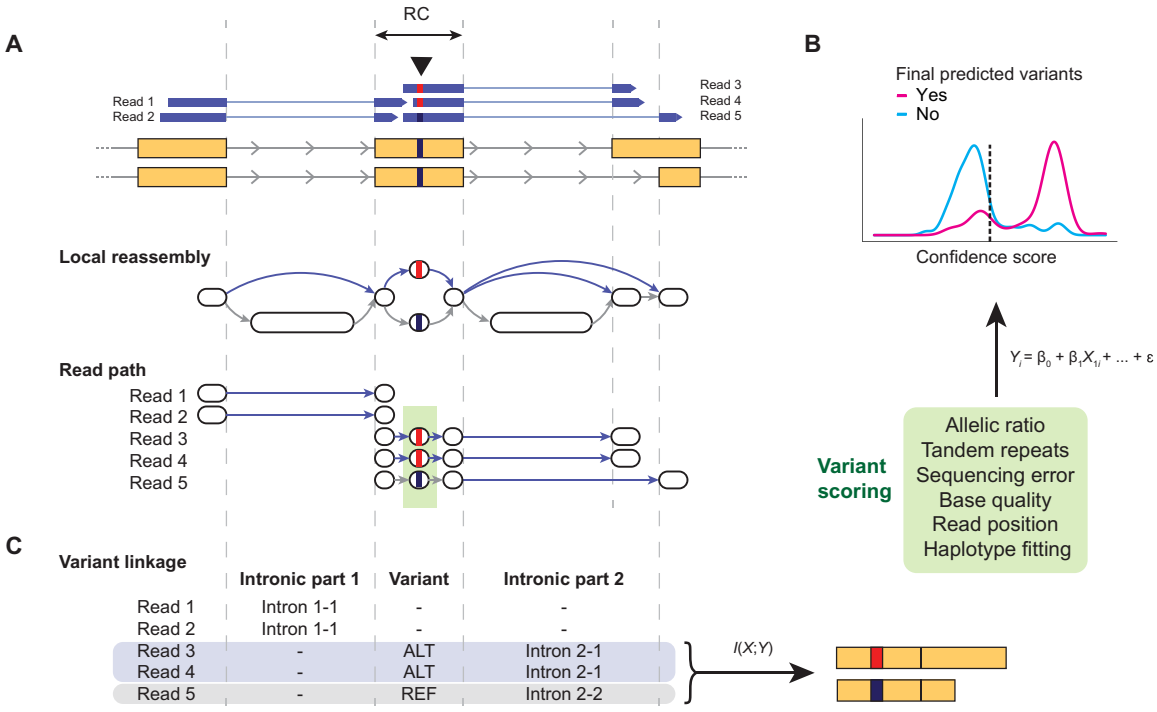

Fig. 1. Algorithm outline. (**A**) Illustration of the main algorithm of scAllele for variant calling. The reads and the reference genomic sequence overlapping an RC are decomposed into k-mers and reassembled into a de Bruijn graph. The graph shown here is a compacted version. The bubbles in the graph indicate a sequence mismatch, i.e., a variant. For each read, scAllele obtains a path for the original read sequence and infers the allele of each variant (including introns). (**B**) Variants (green box in A) identified from the graph are then scored using a GLM. The GLM was trained with different features (green box) to assign a confidence score to the variants. See Materials and Methods for details. (**C**) To identify allele-specific splicing (i.e., variant linkage), scAllele performs a mutual information (MI) calculation between nucleotide variants (SNVs and microindels) and intronic parts (where the alleles are the different overlapping introns) to calculate allelic linkage of splicing isoforms.

and the end of each bubble (source and sink nodes) and their respective pairing (local reassembly; Fig. 1A). Hereafter, we perform a per-read analysis of the graph, where we first obtain the walk in the graph that best matches the read sequence, followed by the identification of variants present in each read. The presence of repeats or low complexity regions greatly complicates the detection of variants because the dBG can be traversed in multiple ways. scAllele overcomes this challenge by performing a Dijkstra-based traversal of the graph with the assumption that the walk with the smallest editing distance best represents the set of variants present in the read. For spliced junction reads, scAllele retains information about splicing and uses the entire read sequence for better assembly. Last, we collect the variants from the RC reads and score them using a generalized linear model (GLM) (Fig. 1B) where the following features are included: read position, base quality, number of neighboring tandem repeats (via a stutter noise model), allelic ratio, sequencing error rate, and RNA-aware haplotype fitting (Materials and Methods). These features capture the genomic and technical contexts of the variants. Because they do not depend on total read coverage at the variant positions, scAllele could potentially handle limited sequencing depth.

An important feature of scAllele is the detection of variants at the read level. This feature enables a direct analysis of allelic linkage between the variants and other attributes of the reads. Here, we focus on identifying allelic linkage with alternative splicing via mutual information (MI) (Materials and Methods), same as in our previous work for RNA editing identification (*17*). We consider overlapping introns as "alleles" of the same intronic part (Fig. 1C) and calculate the read coverages of the allele "haplotypes" between introns and nucleotide variants. In this way, we can incorporate splicing isoforms in the MI calculation to identify allele-specific splicing.

The input of scAllele is a bam file (for a single cell) or a list of bam files (each corresponding to one cell). In the latter case, scAllele carries out joint variant calls using all cells provided in the input and produces variant calls for each cell (Materials and Methods). scAllele is a stand-alone tool and only requires bam files to conduct variant calling and linkage detection. However, preprocessing of the bam file is recommended to achieve optimal results (fig. S1).

Evaluation of variant calls in GM12878 and iPSCs

We evaluated the variant-calling function of scAllele using scRNA-seq (Smart-seq2) of GM12878 cells and induced pluripotent stem cells (iPSC) cells from three individuals (*18*). These individuals were carefully genotyped by the Genome in a Bottle (GIAB) (*19*) and 1000 Genomes projects (*20*), thus providing a "ground truth" for method evaluation. We compared the performance of scAllele to those of three other popular variant callers: Freebayes (v.1.3.4), GATK (v.4.2.0.0), and Platypus (v.1.0) (*8*, *9*, *21*). The performance evaluation followed previously published guidelines (*22*) with some modifications to accommodate RNA variants (see Materials and Methods). For GM12878, we used three benchmark datasets: GIAB's list of all genetic variants, GIAB's list of high-confidence genetic variants, and the variant calls based on long-read DNA-seq (Oxford Nanopore) (*23*).

For each dataset and each method, we calculated the true-positive (TP) counts at specific cutoffs of false-positive (FP) counts for

microindels and single-nucleotide polymorphisms (SNPs), respectively (Fig. 2A and fig. S2A). Here, we used TP or FP counts, rather than TP or FP rates, because the ground truth variants that should be captured in the RNA-seq data are unknown (Materials and Methods). For each method, we carried out the analyses for each cell individually (the "single" mode) or via joint variant calls (the "joint" mode). Overall, scAllele achieved the best performance for both microindels and SNPs among all methods, with the joint mode outperforming the single mode. The strength of scAllele in microindel identification is notable as these variants are known for their challenging detection (*24*). For SNPs, most methods achieved highly desirable performance, often exceeding >95% precision (Fig. 2A and fig. S2A) at each FP cutoff, although scAllele still generally outperformed the other methods. Furthermore, scAllele also demonstrated superior performance in capturing microindels or SNPs in "difficult regions" (Fig. 2B and fig. S2B). These difficult regions were defined by GIAB (*19*) as the union of regions with low mappability, high guanine-cytosine (GC) content, low complexity or presence of repeats, and segment duplication among others.

Although the above cells have been analyzed by the GIAB and 1000 Genomes projects, their genotype calls may still miss some TPs. As examples, we experimentally confirmed four microindels categorized as FPs according to the ground truth (Fig. 2C and fig. S3). The four microindels were identified by scAllele and Platypus (two by GATK and three by Freebayes). Thus, the above performance of scAllele (and the other methods) may be a conservative estimation.

One of the hallmarks of scRNA-seq is the limited read coverage per gene. Thus, it is highly desirable to develop variant callers with superior performance at low read coverage. scAllele meets this demand and demonstrates a performance gain relative to the other methods in lowly covered variants (Fig. 2D and fig. S2C). About 90% of the ground truth variants present in the scRNA-seq data were covered by less than five reads in each dataset. Thus, scAllele affords a unique advantage for scRNA-seq data.

Unique to RNA-seq, the allelic read counts of genetic variants reflect their allelic expression levels. Thus, in addition to variant calling, it is necessary to accurately estimate the allelic quantification of each variant. To test the performance of scAllele in this regard, we segregated the ground truth variants into heterozygous and homozygous groups. The heterozygous variants are expected to exhibit an approximately normal distribution in their alternative (ALT) allelic ratios (variant allele read number/total read number), centered around 0.5 (*13*). For homozygous variants, the allelic ratios are expected to be 1. As shown in Fig. 2E and fig. S2D, the results of scAllele largely followed these expectations for both microindels and SNPs. In contrast, other methods resulted in flawed distributions in at least one of the above aspects.

Overall, the above evaluations support the superior performance of scAllele for scRNA-seq variant analysis, especially in handling microindels, an aspect that is much more challenging compared to the most often tackled SNV identification. In addition, the joint variant calling by scAllele showed highly desirable results, which leverages the availability of data from multiple cells without losing the resolution of variant calling at the single-cell level. Thus, joint variant calling is the default mode of scAllele.

Linkage calculation between variants

In addition to variant calling, scAllele enables read-level allelic linkage analysis. This analysis is not possible with other variant callers as read-level information is not extracted. In scAllele, the degree of allelic linkage is quantified as the MI between two types of variants: nucleotide variants and alternatively spliced junctions (Fig. 1). This metric is expected to require a relatively high number of reads harboring both types of variants. To achieve an understanding of the read coverage requirements, we first calculated the MI between pairs of known genetic variants in the GM12878 and iPSC data used in the last section. As expected, the MI of these variant pairs in the RNA is generally high, regardless of read coverage, reflecting the associated DNA haplotypes (Fig. 3A).

As a comparison, we also calculated the MI between pairs of nucleotide variants where at least one variant was not a known genetic variant (Fig. 3B). Because the cell lines have been well genotyped, we assume that all unknown variants observed in the RNA-seq reads are RNA editing sites or sequencing errors. The MI of these variant pairs is expected to be low in general (*17*) unless rare allele-specific RNA editing exists. This expectation of low MI was met at relatively high read coverage ($\geq$10). However, at lower read coverage, the MI is inflated because of the low number of transcripts used for its calculation. Thus, it is necessary to impose a minimum read coverage requirement for MI calculation. In this study, we set this cutoff to be 10 based on the above results. In addition, we required a minimum MI of 0.52 to call significant linkage events, as 95% of the known genetic variant pairs (with $\geq$10 reads) had an MI of 0.52 or greater, and 90% of the unknown variant pairs (with $\geq$10 reads) failed this MI cutoff (Fig. 3C). The read coverage and MI cutoffs can be altered by the user in scAllele.

Although the sequencing depth of a single cell is limited, a typical scRNA-seq dataset includes a large number of cells. Thus, merging data from multiple cells are effective in enhancing the number of testable events (with $\geq$10 reads) for the linkage analysis between genetic variants and splicing. As shown in Fig. 3D, reads from merely five cells allowed 290 of these events to be tested for linkage (with $\geq$10 reads), and a union of 30 cells had 5937 testable events in the GM12878 data. On the basis of this observation, for the linkage calculation, scAllele provides two alternative options by taking as input a single bam file (for single-cell analysis) or a list of bam files (for merged analysis).

scAllele unveils nucleotide variants and allele-specific splicing events in lung cancer cells

Next, we applied scAllele to scRNA-seq data of lung cancer (Smart-seq2) (*25*). We focused on cancer cells and their normal counterparts, epithelial cells, in tumor and matched normal samples of two patients (TH179 and TH238; $n = 574$ cells). We first carried out joint variant calling for the cancer and epithelial cells, respectively, for each patient. An SNV or microindel was retained if it was detected in at least three cells per individual. Furthermore, we compared the presence of the variants in normal epithelial or cancer cells. A variant was defined as cancer-enriched if it was not detected in normal cells or its presence is significantly more frequent in cancer compared to normal cells (corrected $P < 0.1$; Materials and Methods). Otherwise, the variant was labeled as a common variant to cancer and normal cells. As a sanity check, we note that no variant was found to be enriched in normal cells relative to cancer cells.

As shown in Fig. 4A, >140,000 variants were identified in each patient, with most being SNVs. Most SNVs are annotated variants in the Single Nucleotide Polymorphism database (dbSNP) (b151) or Catalogue of Somatic Mutation in Cancer (COSMIC) (human cancer

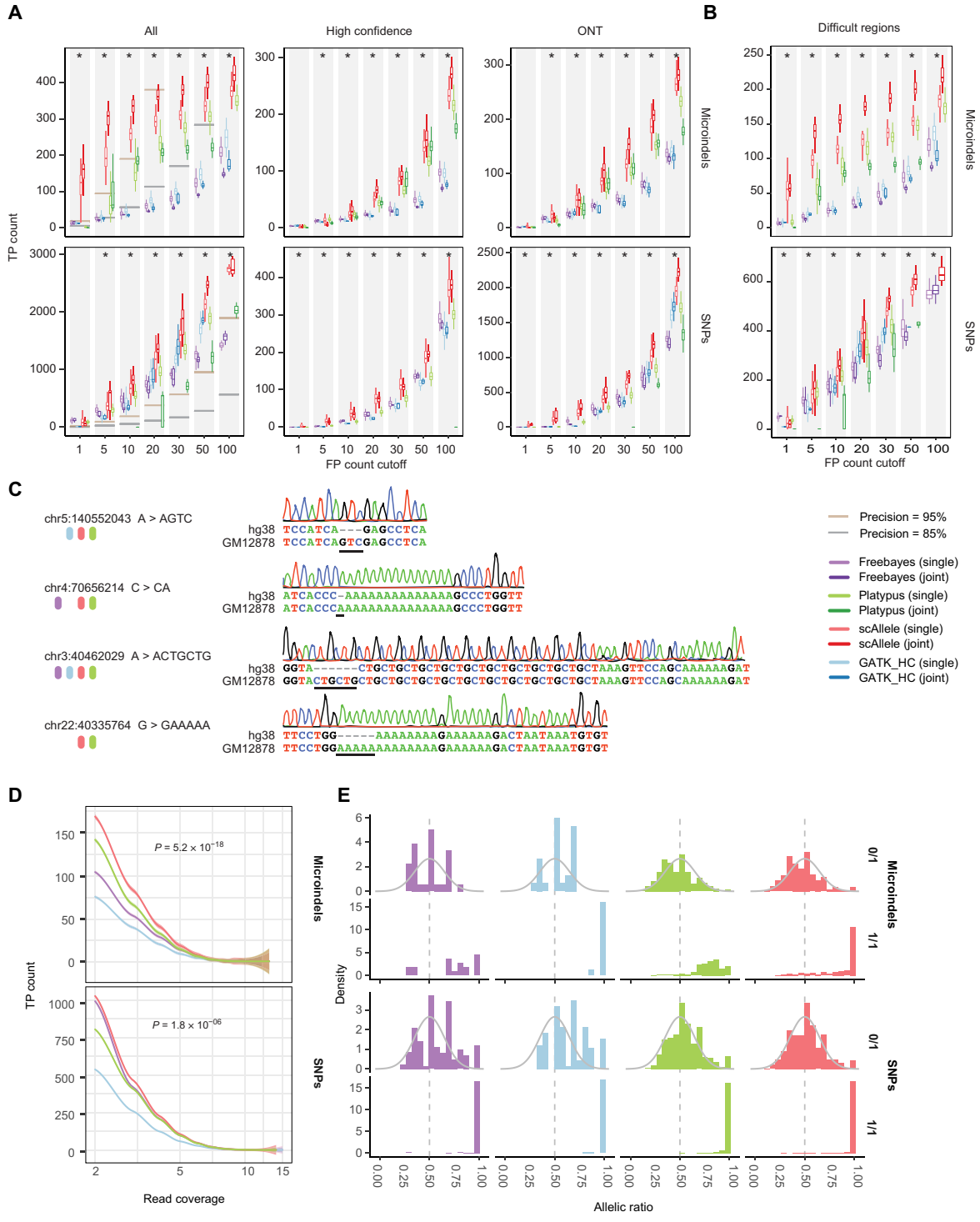

Fig. 2. Performance of scAllele in variant calling of scRNA-seq data of GM12878 cells. (**A**) TP counts of four variant callers were evaluated at different cutoffs of FP count. Performance for microindels (top) and SNPs (bottom) is shown separately. Three sets of ground truth genetic variants were used: all of GIAB-reported small variants (all), GIAB's high-confidence variants (high-confidence), and variants called via Nanopore sequencing by Karst *et al.* [Oxford Nanopore Technologies; (*23*)]. In the panel for "all" variants, the horizontal lines represent the TP count that achieves a precision of 85 or 95% at each FP count cutoff, respectively. Each method was applied to call variants in single cells (single) or jointly across all cells (joint). Asterisks indicate that scAllele (joint or single) had significantly higher TP count than all other methods as determined by a t test ($P < 0.05$). (**B**) Performance in "difficult regions" defined by GIAB. (**C**) Experimental validation of novel microindels via Sanger sequencing. The bars underneath the variant coordinates indicate the methods that detected the variants. (**D**) TP count of four variant callers (at maximum F1 and precision of >0.9) across different read coverages. The curves represent averages of all cells. The *P* value was calculated to compare the TP count (coverage of <5) between scAllele and the second best method (Platypus for microindels and Freebayes for SNPs) via a Wilcoxon rank sum test. (**E**) Allelic ratios of TP variants segregated by their true genotype. 0/1, heterozygous; 1/1, homozygous variant allele. Gray curve is shown for reference purpose only (normal distribution of means = 0.5 and SD = 0.15).

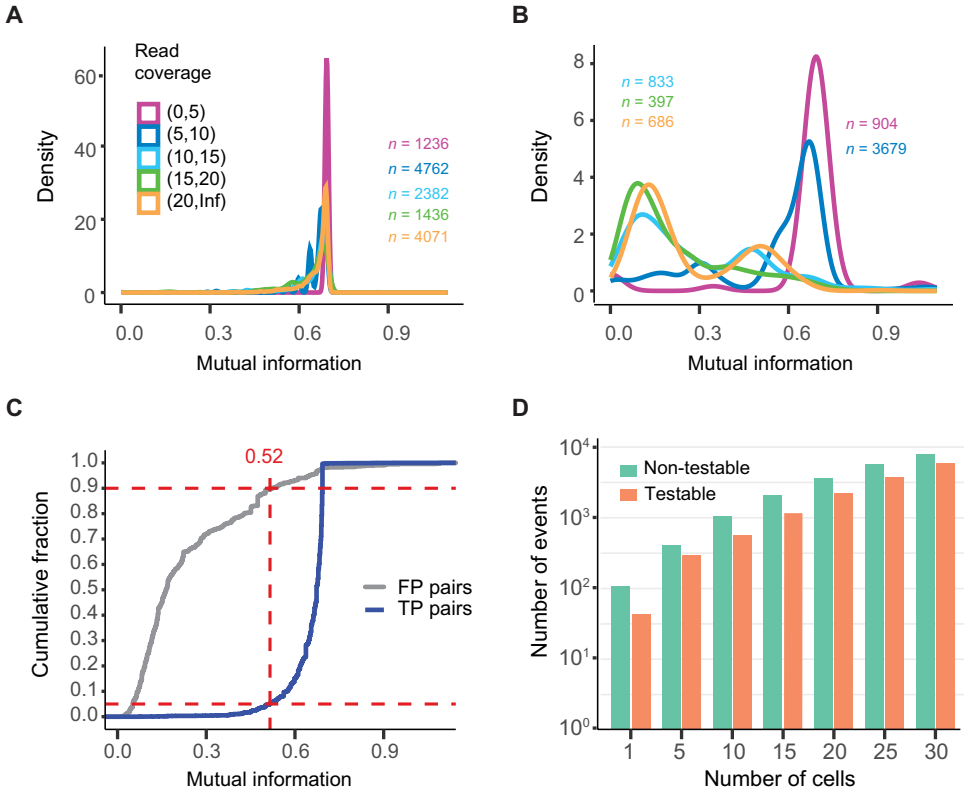

Fig. 3. Linkage calculation between pairs of variants. (**A**) MI distribution (natural log-based) of pairs of true genetic variants (namely, TPs) from the GM12878 and iPSC scRNA-seq segregated by the number of reads covering the pair. Most of these pairs have values close to the theoretical maximum for two alleles [ln(2) = 0.693] regardless of coverage. The number of variant pairs within each read coverage group is shown. (**B**) MI distribution of pairs of variants where at least one is not a known genetic variant (here referred to as FP genetic variants). Similar as in (A), the data were segregated on the basis of read coverage groups (with their corresponding number of variant pairs indicated). (**C**) Cumulative distribution of MI of TP and FP pairs with a minimum read coverage of 10. The MI cutoff of 0.52 was selected as the minimum value for significant linkage between variants. This cutoff rejects 90% FP pairs and 5% TP pairs (dashed lines). (**D**) Number of linkage events between genetic variants and intronic parts with at least 10 total reads (testable) and that with less than 10 reads (nontestable) identified in a one (1) cell or the merged analysis of multiple GM12878 cells.

mutations) database. COSMIC variants constitute a larger fraction among cancer-enriched variants compared to the common variants ($P < 1 \times 10^{-16}$ in both patients, chi-square test). Among the unannotated (i.e., novel) SNVs, some may be novel genetic variants, and others may reflect RNA editing events. A large fraction of the novel SNVs corresponded to A-to-G or C-to-T RNA editing types (28 and 58% in TH179 and 35 and 51% in TH238, respectively). A relatively large fraction of microindels was not annotated in either database, likely reflecting our incomplete knowledge of this type of variants. Cancer-enriched SNVs were more often located in coding exons and 3′ untranslated regions (3′UTRs), less often in introns, compared with common SNVs (Fig. 4B). A similar enrichment in the exonic and 3′UTR regions was observed for cancer-enriched microindels for both patients, relative to the common microindels. These observations support the likely functional importance of these genetic variants.

Furthermore, variants of individual cells can be used to classify cells and detect their individual of origin, an application that is important to experiments where samples from multiple individuals are sequenced together to reduce cost or control for batch effects (*18*). Using scAllele-identified variants and Souporcell (*26*), 99.3% of the cells from the two individuals (TH179 and TH238) were accurately classified, except for four cells, supporting the validity of scAllele's variant calls (fig. S4).

Following variant calling, we carried out linkage analysis to identify allele-specific splicing events in each cell separately. As examples, Fig. 4C shows two significant linkage events. In these cases, the SNPs demonstrated strong allelic linkage with alternative splicing patterns (exon skipping and alternative 5′ splice site, respectively). For the single-cell analysis, across cancer and normal cells, the number of allele-specific splicing events varied greatly, ranging from 0 to 67 events (Fig. 4D), most of which involved SNVs. This number correlated approximately with the number of spliced-junction reads present in each cell (Fig. 4D, insets). On the basis of down-sampling of a few deeply sequenced cells, we observed that 1 million total reads can enable identification of up to nine events per cell (Fig. 4E). In some cells, the number of events plateaued at around 5 million reads. Thus, to afford power for splicing analysis, a relatively large number of scRNA-seq reads is needed per cell. Nonetheless, because scRNA-seq typically involves many cells, the total number of events identified across individual cells can be substantial. In our data, a union of hundreds to thousands of allele-specific splicing events were identified in the cancer or normal cells of each patient (Fig. 4F).

Next, we carried out the allele-specific splicing analysis by merging data of all cancer or normal cells of each patient. As shown in Fig. 4F, the merged analysis uncovered a large number of events, exceeding the total number of events identified across individual cells. Around 42.3 to 49.6% of events from the merged analysis were also identified

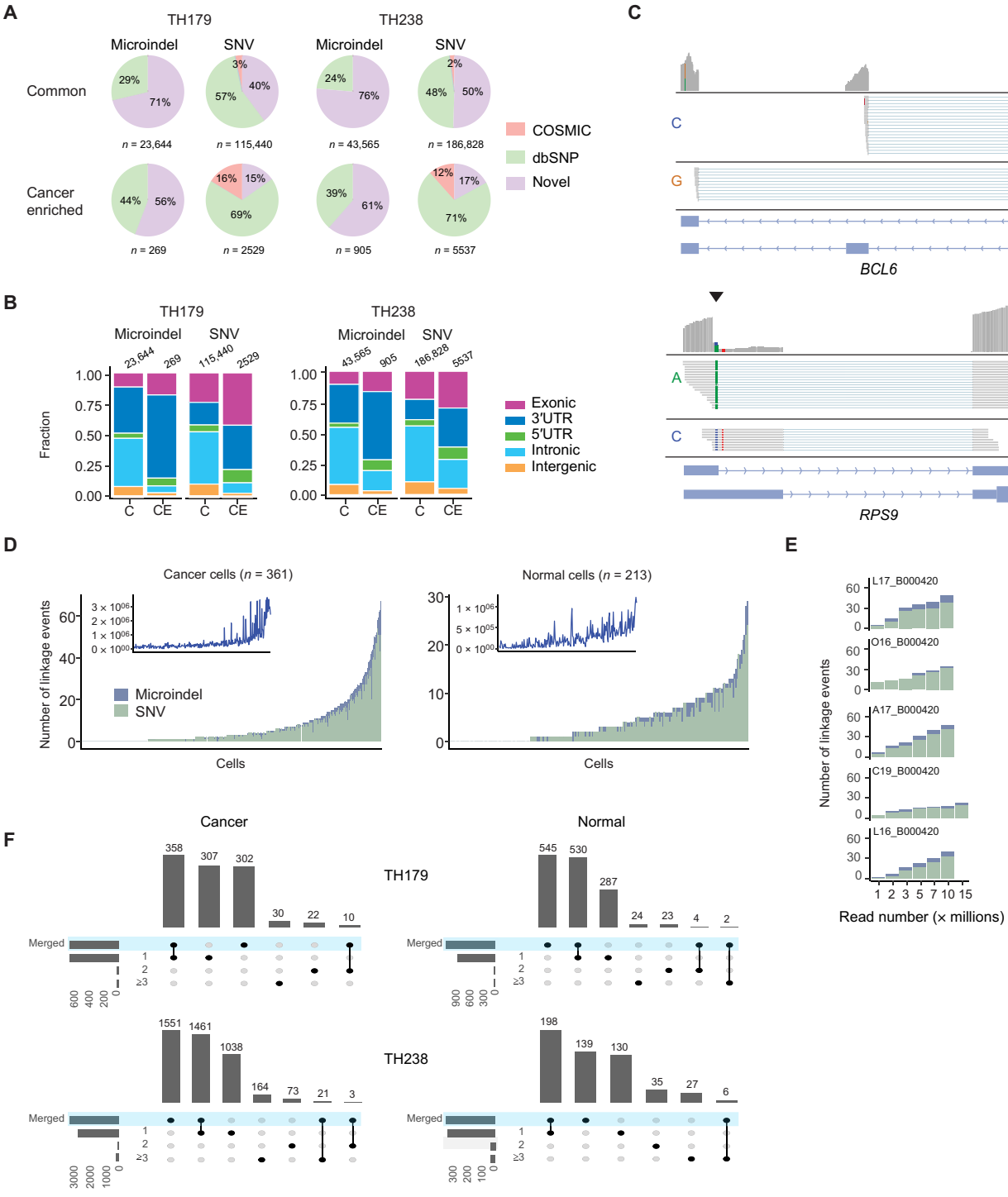

Fig. 4. Genetic variants and allele-specific splicing events detected by scAllele in the lung cancer scRNA-seq dataset. (**A**) Variants identified in each patient (in ≥3 cells). Common denotes variants common to cancer and normal cells. Cancer enriched denotes variants enriched in cancer cells (see main text). Novel denotes variants not annotated in dbSNP or COSMIC. Variants labeled as "COSMIC" may be present in dbSNP as well. (**B**) Distribution of variants in (A) in different types of genomic regions. C, common variants. CE, cancer-enriched variants. (**C**) Integrative Genomics Viewer (IGV) view of two example allele-specific splicing events. The location of the variant is denoted by the black arrowhead. Reads overlapping each variant are segregated according to the allele they harbor (indicated by the nucleotide color and label on the left). (**D**) Number of allele-specific splicing events identified in cancer or normal cells (union of cells from the two patients). The cells are ranked by their total number of events. Events associated with microindels or SNVs are shown in different colors. The insets show the total number of junction reads in the scRNA-seq data of each cell, sorted in the same order as the main panel. (**E**) Number of allele-specific splicing events identified in five deeply sequenced cells at different down-sampled total read coverage. (**F**) Number of allele-specific splicing events identified in the single-cell or merged analysis and their overlaps. The results for the single-cell analysis are shown separately for those identified in 1, 2, or ≥3 cells.

in the single-cell analysis, but each type of analysis had a set of unique events as well. Therefore, whereas the merged analysis achieved enhanced power to discover certain events, other events may only be present in a small number of cells that warrant a single-cell analysis.

Cancer and normal cells exhibit unique and differential allele-specific splicing events

Next, we asked whether cancer and normal cells harbor different allele-specific splicing events. For this analysis, we focused on events identified in the single-cell analysis. Among all these events, 56 were observed in both cancer and normal cells, whereas more events were exclusive to one of the two classes of cells (Fig. 5A). In general, most events were observed in a small number of cells (<5), but there exists a subset of events (17 total) that were present in more than five cells (Fig. 5A).

To identify differential allele-specific splicing events between cancer and normal cells, we focused on two scenarios. In the first scenario, the variants were not present/testable in the normal cells but had significant linkage in the cancer cells, or vice versa (labeled as "cancer-specific" or "normal-specific"; Fig. 5B and table S1). For this scenario, we observed 3005 events that were cancer specific and 986 that were normal specific. Notably, 32 cancer-specific events were observed in at least three cancer cells, whereas only 2 normal-specific events exceeded this level of prevalence (Fig. 5B). Figure 5C shows an example in the gene *IFI44L* (interferon-induced protein 44–like), a type I interferon–stimulated gene with a role in host antiviral response (*27*). The allele-specific splicing event and the associated variant were only observed in cancer cells.

The second scenario includes variants present and testable for splicing linkage in both cancer and normal cells, but significant linkage was detected with higher prevalence (*P* value of <0.05, Fisher's exact test) in one cell class than the other. For this scenario, we observed 4 events with higher prevalence in cancer cells (cancer-differential; Fig. 5B and table S1) and 13 with higher prevalence in normal cells (normal-differential; Fig. 5B and table S1). Figure 5C shows an example of such an event in the *CTSE* gene, where the C allele of the variant is linked to skipping of the middle exon, whereas the T allele is associated with exon inclusion. This linkage was only observed in normal cells but not cancer cells (despite the presence

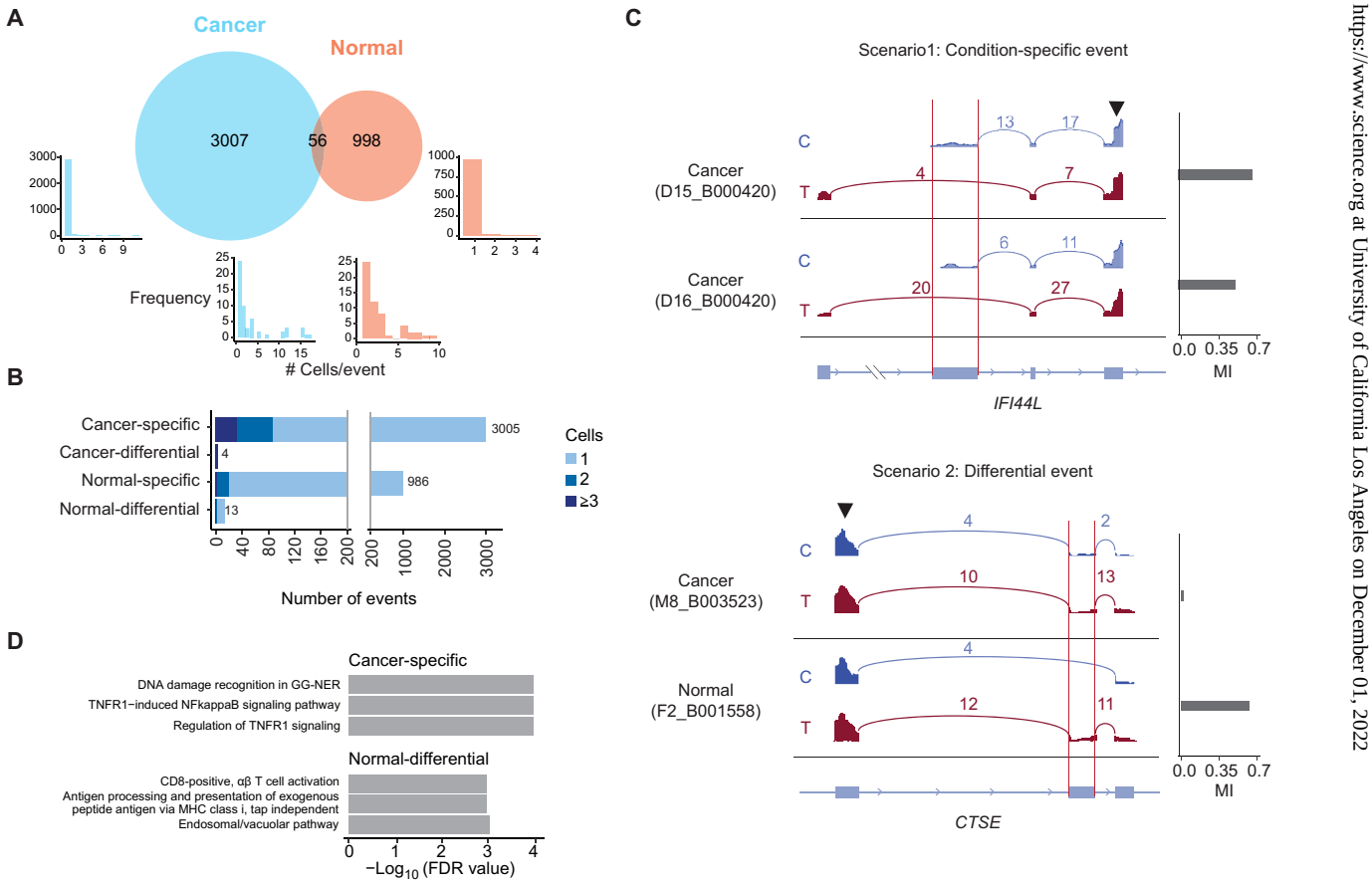

Fig. 5. Comparison of allele-specific splicing events in cancer and normal cells. (**A**) Number of events (identified in single-cell analysis) shared by cancer and normal cells, or exclusive to one of the two classes. Histograms of the number of cells harboring each type of events are also shown for each set and the intersect. (**B**) Number of events categorized as cancer-specific, normal-specific, or differential (see main text for details). The number of cells harboring the events are shown by the shade of the bar. (**C**) Examples of allele-specific splicing events (red vertical lines): scenario 1, a condition-specific event; scenario 2: a differential event. The sashimi plots are split by the allele harbored in the reads (indicated by color and nucleotide label). The read counts are reported for each junction, and the MI between the variant and the splicing event is shown by the bars on the right. Note that only reads harboring the variant (black arrowhead) are shown. (**D**) Top three most significant pathways identified in the functional enrichment analysis of genes harboring allele-specific splicing events in the categories shown in (B). MHC, major histocompatibility complex. GG-NER, Global Genome Nucleotide Excision Repair.

of the variant and adequate read coverage in cancer cells). Notably, the gene *CTSE* encodes for cathepsin E, an aspartic protease with a vital role in protein degradation, bioactive protein generation, and antigen processing and presentation (*28*).

In general, many genes with allele-specific splicing events have cancer relevance. For example, genes with cancer-specific events in Fig. 5B are enriched in inflammatory processes, such as the tumor necrosis factor receptor 1–induced nuclear factor κB signaling pathway (Fig. 5D), which has close relevance to cancer (*29*). Involved in antigen presentation and T cell receptor binding (*30*), two genes from the human leukocyte antigen (HLA) family (*HLA-A* and *HLA-E*) harbored normal-differential events. The four tumor-differential events are located in the gene *SFTPA2*. This gene encodes a pulmonary surfactant–associated protein that lowers the surface tension in the alveoli of mammalian lungs facilitating normal respiration (*31*). Together, these results suggest that allele-specific splicing is involved in many molecular processes relevant to lung cancer.

Run time and memory usage of scAllele

Last, we evaluated the computational performance of scAllele. Figure S5 shows the total number of reads, run time, and RAM usage analyzing the iPSC-derived data. As expected, the run time and memory usage increased with increasing data (number of cells or reads). An analysis of 96 cells needed about 2.5 hours and 14 gigabyte of total memory with 36 cores. When processing a large number of cells, scAllele can be easily adapted to parallelize for chromosome-wise or region-wise analysis.

DISCUSSION

scRNA-seq affords unprecedented views of single-cell transcriptomes. Similar to bulk RNA-seq, scRNA-seq provides information on the single-nucleotide level. However, identification of nucleotide variants in scRNA-seq is challenging because of the limited read coverage per cell. Here, we present scAllele, a versatile tool that not only enables variant calling in single cells but also uncovers allele-specific RNA processing events.

We showed that scAllele outperforms other popular methods in variant calling, especially for microindels, the class of variants that are less well characterized than SNVs. Built upon local reassembly, scAllele refines read alignments and corrects possible misalignments in each read, thus enhancing variant detection accuracy per read. Specific to the nature of RNA-seq data, scAllele handles split reads at the spliced junction by retaining information about splicing and using the entire read sequence for improved assembly. In addition, the Dijkstra-like algorithm enables simplified traversal of the dBG, which makes scAllele more suitable for low complexity and repetitive regions. scAllele uses a GLM model to detect high-confidence variants. The GLM scoring scheme does not depend on total read coverage at the variant position. Instead, it focuses on the context of the variant, such as tandem repeats, base quality of the nearby sequences, overall allelic ratio, and RNA-aware haplotype fitting. As a result, scAllele has an advantage in handling limited sequencing depth and considering RNA-specific features that were rarely considered by other variant calling methods, such as stutter noise in tandem repeats, allele-specific expression, or allele-specific splicing. Furthermore, leveraging the availability of multiple cells in scRNA-seq data, the joint variant calling mode of scAllele yielded superior results. Notably, scAllele's joint-calling mode preserves variant information at the individual cell level and reports the optimal variant call considering both single and joint analyses for each variant. This strategy allows identification of rare variants (such as somatic mutations) that are present in very few cells. These features together confer the superior performance of scAllele.

The read-level variant calling by scAllele enables another advantage, that is, facilitating a detailed view of the allelic bias linked to alternative RNA isoforms. In this work, we focused on allele-specific splicing patterns. A similar approach can be extended to examine other aspects of RNA expression, such as alternative polyadenylation. This type of analysis requires a relatively high read coverage per event, as it simultaneously quantifies alternative alleles of nucleotide variants, alternative RNA isoforms, and their combined linkage patterns. This need of high read coverage is analogous to the fact that alternative isoform analysis in bulk RNA-seq demands deeper read coverage than total gene expression analysis, which is well established. We showed that the number of these events increased with higher scRNA-seq depth, indicating that RNA representation in scRNA-seq was not saturated at lower depth, such as 1 M reads. In addition, to discover splicing events, scAllele is best applied to data generated by scRNA-seq protocols that cover full-length transcripts, such as Smart-seq2. With the continued drop in sequencing cost and innovations in scRNA-seq technologies, we expect to see wide applications of allele-specific and alternative RNA isoform analyses, such as those enabled by scAllele.

We applied scAllele to a lung cancer scRNA-seq dataset (with matched controls). Our analysis identified a large number of nucleotide variants, many of which had enriched presence in cancer cells. Compared to variants common to both normal and cancer cells, cancer-enriched variants were more often cataloged in COSMIC, supporting the validity of the scAllele variant calls. In addition, cancer-enriched variants (both SNVs and microindels) were more often located in coding and 3′UTR regions, which suggests a potential role in altering protein sequences, producing neoantigens, or regulating gene expression. Given the existence of numerous regulatory elements in the 3′UTRs (*32*), genetic variants in these regions may alter many processes, such as mRNA stability, translation, or mRNA localization, which should be investigated in the future. Although microindels are not as abundant as SNVs, they may have critical roles in human diseases (*33*), an area that remains underexplored partly because of the lack of effective methods to identify and analyze these variants. Thus, scAllele fills in a crucial gap in this area.

In the cancer and normal epithelial cells, we identified a large number of allele-specific splicing events. As expected, merging data from all cancer or normal cells yielded more events than analyzing single cells separately. Yet, many events identified in the merged analysis were also detected in individual cells, confirming their presence in multiple cells. Notably, the single-cell analysis uncovered events that were not detected in the merged analysis. This observation likely reflects the existence of events that occur in a small number of cells that were diluted away when data from many cells were merged. Thus, both merged and single-cell analyses should be conducted to obtain a comprehensive view of allele-specific splicing. We further categorized these events on the basis of their relative prevalence in cancer or normal cells. Although most events were observed in a small number of cells, likely because of low read coverage in single cells, this categorization provides an approximate overview of their relative enrichments. Among these events, many have important relevance to cancer, such as those in the *CTSE* and *IFI44L* genes

in Fig. 5C. Our results suggest that scRNA-seq data have useful information to uncover important alternative splicing events, linking genotypes to this molecular phenotype.

In summary, scAllele offers a unique approach to maximize the information extracted from scRNA-seq datasets. With the emergence of scRNA-seq data from a large spectrum of samples, scAllele will lead to a granular view of the genetic landscape of each cell and the potential genetic drivers of gene expression complexity.

MATERIALS AND METHODS

scAllele: Detailed outline

To scan variants in the entire transcriptome, we grouped the sequencing reads into RC. An RC is made up of a group of overlapping read segments. Here, we define read segments as regions of the reads split by the "N" CIGAR elements (i.e., introns). On the basis of this definition, it is expected that RCs likely overlap exons. For each RC, scAllele reassembles the reads and calls variants. The entire sequence of each read was considered for the construction of the dBG not only the segment that overlaps the RC. In this way, information in the flanking introns of each RC is preserved. Multimapped, chimeric, polymerase chain reaction (PCR) duplicates, and low mapping quality reads were removed as well as reads with ≥5 soft-clipped bases or trailing homopolymers ($n \geq 15$). An additional "reference read" was included as part of the RC. This read contains the genomic reference sequence of the entire range spanned by the RC.

In each RC, the reads were decomposed into overlapping k-mers (k-1 overlap), which are the nodes of the dBG. The edges represent consecutive nodes (i.e., two k-mers overlapping by k-1) in the reads. Every edge was labeled with the name of the reads that contained this consecutive pair of k-mers and the position in the read where the k-mers were located.

The graph was then processed by compacting and removing certain nodes. Walks on the graph that contain consecutive nodes of in-degree = 1 and out-degree = 1 can be merged into a single node that contains a sequence length of $k + n - 1$, where n is the number of nodes being merged. In addition, subsequences in the reference read that did not overlap with other reads (which are usually intronic segments) were also compressed. This step greatly simplifies the graph because the intronic regions are generally several thousand bases long, much longer than the average RC. Other nodes were removed from the graph if they did not provide useful information. For example, we defined the actual start and end of the RC as the first and last nodes that originated from the reference read. By definition, these nodes have in-degree = 0 and out-degree = 0, respectively. Additional nodes that complied with the degree requirement but did not originate from the reference read were labeled as alternative starts and ends. These alternative starts and ends also represent differences among read sequences. However, since they do not form a bubble, it is not possible to infer the variant causing this difference.

Subsequently, scAllele inferred the walk on the graph that matched the original sequence of each read. These walks were named "read walks." Because some nodes were removed or merged in the previous step, this walk is not necessarily the same sequence of nodes obtained from the initial read decomposition. As a result, many of the original reads were matched by the same read walk, reducing the number of distinct reads to process.

In the compacted/cleaned dBG, we identified the bubble structures by locating the source nodes, the sink nodes, and the walks connecting them via DFS of the graph. These structures represent variants, and with the DFS, we can identify which specific source node, sink node, and connecting walk correspond to each allele. This information was then used to identify the variants and their alleles present on each read walk. In the case of highly interconnected/cyclic graphs (due to existence of repeats or low complexity regions), this assignment was aided by a Dijkstra-like algorithm, which identifies the most likely set of variants on a read walk by minimizing the editing distance between the read walk sequence and the reference sequence. More specifically, first, all the end-to-end read walks were identified. Then, by calculating the cumulative edit distance at every node and traversing the graph through different walks, we can select the best walk. Note that introns were also considered a type of variants in these intermediate steps and were processed in the same way as the nucleotide variants. However, we did not assign an edit distance to them.

The variants were further processed by normalizing, left aligning, and atomizing. Different features were collected for each variant including read counts for each allele, base qualities, read positions, and count of tandem repeats flanking the variant. An additional feature, namely, haplotype fitting was calculated using the entire set of reads and variants from each RC. These features were then used to score the quality of the variant (see the "Variant scoring" section). At this point, the variant calling step was complete. Because scAllele identifies variants at the read level, this information was stored in memory for subsequent MI analysis, based on which the linkage between nucleotide variants and splicing isoforms was calculated (see the "Linkage analysis" section).

Variant scoring

We trained a GLM using ground truth genetic variants and various features obtained from the main algorithm of scAllele. The features included the variant's ALT allelic ratio (AB), the number of tandem repeats neighboring the variant (TandemRep), sequencing error rate (SER), median base quality in the variant's proximity, read position, and the haplotype fitting, as detailed below.

Low ALT allelic ratio is often indicative of an FP variant, likely due to existence of sequencing errors. We can calculate the probability of observing an allelic ratio (AB) if it is resulted from a sequencing error using the binomial distribution

$$p(\mathrm{AB} \mid \mathrm{seq.error} = f) = \binom{\mathrm{DP}}{\mathrm{AC}} f^{\mathrm{AC}} (1-f)^{\mathrm{DP-AC}}$$

where AC is the ALT allele counts, and DP is the total read count. The value of f is the probability of error, which, in most cases, corresponds to the SER. We used 0.01 for this variable, which is the maximum error rate for the Illumina sequencing platforms (*34*). In tandem repeats, however, the probability of error is expected to increase because of the propensity of PCR slippage. We then defined f as follows

$$f = \begin{cases} \mathrm{SER} + 0.075 \text{ if TandemRep} \geq 5 \text{ and varLength} = 1 \\ \mathrm{SER} + 0.035 \text{ if TandemRep} \geq 5 \text{ and varLength} \geq 2 \\ \mathrm{SER \text{ otherwise}} \end{cases}$$

These values are approximations of the empirical estimation of stutter noise made in lobSTR (*35*). Stutter noise was found to be a function of the variant length (varLength) and the number of tandem repeats (TandemRep). The probability $p(\mathrm{AB} \mid \mathrm{seq.error})$

was used as an additional feature in the GLM and served as an interaction term between AB and TandemRep.

Another feature in the GLM was derived from base quality scores. In case of SNVs, we simply used the base quality at the mismatch position. For microindels, we used the median base quality in the neighboring region of the variant (±7 bases) since the original position of the microindel is, in many cases, ambiguous. In addition, the median read position was used as a variable in the GLM because the 3′ ends of the reads tend to have lower base quality, also considering the fact that the bubble structures in the dBG are less reliable if they only use the ends of the read walks.

Next, scAllele calculated another metric called haplotype fitting. This refers to the ability to cluster the variant alleles into two potential haplotypes based on their colocalization in the reads. We clarify that we do not aim to infer the actual haplotype because RNA-seq data are not ideal for this task. This step simply checks for multiallelic variants and allele combinations that result in more than two haplotypes. For this step, we discarded potential RNA editing sites, and we performed the clustering at the RC level, which, most of the time, matches exonic coordinates. In this way, the haplotype is not confounded by nongenetic variants or allele-specific splicing.

The regression of the GLM was performed using the scikit-learn package (*36*) from Python. The training data consist of genetic variants identified in scRNA-seq data originated from the GM12878 cell line. We used the ground truth from GIAB (*19*) and trained the model on a subset of the dataset used in the "Evaluation of variant calls in GM12878 and iPSCs" section. The training data were not used to derive the performance results in that section. Specifically, we selected five cells to train the GLM and report the results on the remaining 55 cells in Fig. 2. The GLM was trained with label weights to account for the imbalance of labels. We trained a separate model for SNPs, insertions, and deletions respectively. The data were randomly split 25 times into 0.67 and 0.33 proportions for training and testing, respectively. The mean Area Under Receiver-Operating Characteristic curve (AUROC) values in the training dataset (five cells) were 0.87, 0.89, and 0.88 for the three models (SNPs, insertions, and deletions), respectively. The AUROC on the testing dataset (remaining 55 cells) was 0.68, 0.89, and 0.85, respectively. This classifier is set as the default model of scAllele and was used for the analysis of all other datasets presented here (iPSCs and lung cancer data).

Note that the above performance was calculated relative to all variants identifiable by scAllele in the training data because the TP variants specific to RNA-seq are unknown. In RNA-seq, the ground truth variants are, in principle, restricted to those that are transcribed and captured in the reads. Among dbSNP variants, only ~1.4% are located in exons (*37*). The identifiability of an mRNA variant in a specific scRNA-seq dataset depends on the expression level of the mRNA, allele-specific expression status of the variant, and sequencing depth. As a result, when evaluating different methods, one cannot directly determine the ground truth without using a variant calling method, which defeats the purpose of method comparison. Thus, in this work, we used TP count at fixed thresholds of FP counts as performance metrics (Fig. 2 and fig. S2).

Last, we sought to define a "quality score" for the scAllele variant call. First, we consider the log likelihood of the GLM regression as a regression score

$$\log\left(\frac{p(\text{Variant} = \text{True})}{p(\text{Variant} = \text{False})}\right) = \text{GLM}(\text{feature1}, \text{feature2}, \ldots)$$

Meanwhile, the quality score (QUAL) of a standard VCF file format (specified by VCFtools, v.4.2) is a Phred-scaled form

$$\text{QUAL} = -10 \times \log_{10}(p(\text{Variant} = \text{False}))$$

Thus, a GLM regression score of zero corresponds to QUAL = 3.01, which represents equal probability of a variant being true or false. On the basis of the benchmark evaluation, we observed that the AUROC score was usually maximized at regression scores between 1 and 2 ($10.4 \leq \text{QUAL} \leq 20.0$), whereas the F1 score was maximized at the lowest QUAL (3.01). Thus, in scAllele, the default score format is QUAL with a cutoff of 3.01. However, the user can choose to use regression scores to define quality of variant calls, and the score cutoff can also be defined by the user.

For joint variant calling, scAllele first calls and scores variants in each individual cell. A joint QUAL score is then calculated by combining features from all cells, e.g., mean base quality across cells, allelic ratio across cells, etc. The final QUAL score used for each variant is set to be the highest score among all cell-level scores and the joint score. In this way, variants that are cell specific (e.g., somatic mutations) still maintain their cell-level scores as the joint score is likely low given the low allelic frequencies of these variants across the cell population. In contrast, germline variants, present in most cells, likely receive higher joint scores than cell-level scores because of, for example, less stutter error effect due to higher read counts and higher mean base quality.

Linkage analysis

scAllele detects variants at the read level allowing for allelic linkage detection. For every RC, all the reads that overlap a variant position were collected with their corresponding allele recorded (REF or ALT). Reads from different RCs were pooled together after scanning an entire chromosome. In paired-end data, an RC may not contain both mates of the pair. Thus, by merging reads from different RCs, we can increase the number of potential linkages.

For every pair of variants that were less than 100 kb apart, scAllele retrieved the reads that overlapped both variants. Using these reads, one array per variant was constructed containing the allele information of the variant. We used Python's scikit-learn package (*36*) to calculate the MI between these two arrays, same as the linkage calculation between nucleotide variants and splicing isoforms (details below).

For the calculation of linkage between nucleotide variants and splicing isoforms, scAllele first grouped overlapping introns into an "intronic part" (Fig. 1A). These overlapping introns were considered as alleles of the intronic part. Thus, in this calculation, an intronic part is analogous to a nucleotide variant. In this way, the MI between an intronic part and a nucleotide variant [$I(v_1, v_2)$] can be calculated as follows

$$I(v_1, v_2) = \sum_{a_i \in N} \sum_{a_j \in N} p(a_i, a_j) \times \log\left(\frac{p(a_i, a_j)}{p(a_i)\,p(a_j)}\right)$$

where a_i and a_j represent the alleles of variants 1 and 2, respectively, and N represents the collection of all possible alleles. The probability of observing a_i, a_j or (a_i, a_j) was calculated using the maximum likelihood method.

As described in the "Linkage calculation between variants" section in Results, we required a minimum of 10 common reads for a

pair of variants. In addition, a minimum MI of 0.52 was required to define a significant linkage.

scRNA-seq processing and mapping

All datasets used in this study were generated via the Smart-seq2 protocol, which allows for full transcript coverage, making it ideal for variant identification and alternative splicing analysis. Raw scRNA-seq fastq files from the GM12878 cell line were retrieved from ENCODE (accession: ENCSR000AIZ, two biological replicates). These replicates were deeply sequenced (about 30 million reads). Thus, we down-sampled each replicate into 30 alignment files with roughly 1 million reads each. The goal was to resemble a shallowly sequenced sample to test our method on low-coverage data.

The scRNA-seq data from iPSCs, corresponding to individuals NA19098, NA19101, and NA19239, were obtained from National Center for Biotechnology Information (NCBI; accession number: GSE77288) (http://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77288). The lung cancer (*25*) dataset was also downloaded from NCBI (BioProject PRJNA591860) (https://ncbi.nlm.nih.gov/bioproject/28889).

Raw reads from all samples were preprocessed using fastqc (v.0.11.7) (*38*) to check for adapter content and overrepresented sequences. If present, these sequences were removed using cutadapt (v.1.9) (*39*). The 3′ end of reads with low base quality was also trimmed using sickle (v.1.33) (*40*). The reads were aligned using two-pass STAR alignment (v.2.7.0c) (*41*). Last, we marked PCR duplicates using the tool MarkDuplicates from Picard Tools (v.2.25.2) (fig. S1) (*42*).

Evaluation of multiple variant callers

Variants were called by three other tools: Platypus, GATK-HaplotypeCaller, and Freebayes. The recommended preprocessing steps and parameters were used. For Platypus, the variant call was performed after processing the alignment file with Oppossum. For GATK, the variant calls were carried out following the "best practices" steps for RNA variant calling. For Freebayes, the variant call was performed with default parameters. Joint variant calls for GATK was performed by running individual cells in the Genomic Variant Call Format (GVCF) mode followed by the variant consolidation and combine GVCFs step. Freebayes and Platypus perform joint calls as default when the user inputs multiple bam files. For all methods, the predicted variants were then filtered to remove those with ALT allele read count of <2 or with A-to-G, C-to-T mismatches as they may represent RNA editing sites. Variants with labels indicating low quality were also removed.

We evaluated the performance of the variant callers using the vcfeval function from rtg tools (v.3.12) (*43*) according to the benchmarking standards reported previously (*22*), with the following parameters

```
RTG vcfeval – T 1 – b $TRUTH_VCF – c $QUERY_VCF – o $OUTPUT –
t $REF_SDF – f 'QUAL' ––bed – regions = $RC_BED ––all – records –
–decompose ––ref – overlap ––sample ALT ––output – mode = 'annotate'
```

The variable RC_BED is a bed file containing all the genomic regions covered by at least one read. We used this file to reduce the running time of the software. The option "--sample ALT" was used to skip genotype matching and is more appropriate for RNA data. For the evaluation of variant calling in difficult regions, we used the union of bed files containing difficult regions for variant calling from GIAB, which merges regions of low mappability, high GC content, segment duplication, low complexity, functional regions, and other difficult regions.

Detection of cancer-enriched variants and annotation

We applied scAllele (joint variant calling) to the lung cancer dataset from Maynard *et al.* (*25*) and selected cancer and normal epithelial cells corresponding to two individuals (TH238 and TH179), where both cancer and normal tissue biopsies were obtained. Then, we focused on variants present in at least three cells of an individual. We calculated the prevalence of each variant across cells. Using the hypergeometric test, we evaluated the enrichment of each variant in cancer cells compared to normal. The *P* value obtained was then corrected for multiple testing using the Benjamini-Hochberg (BH) method. We defined a given variant as cancer enriched if the BH-corrected *P* value is ≤ 0.1 or if the variant is not present in any normal cell.

We further overlapped the variants with the COSMIC (cancer.sanger.ac.uk) (*44*) and, subsequently, dbSNP (b151) (*45*) databases. From the COSMIC annotation, we only selected variants that were confirmed to be somatic and were found in lung tissue. A variant was labeled "novel" if it is not present in either database.

Clustering by genotype

We obtained the allele counts for each variant in each cell of the two individuals in the lung cancer dataset from the VCF files produced by scAllele (the AC and RC fields). With these variant calls, we ran Souporcell (*26*) to cluster the cells of the two patients (with parameter $k = 2$).

Detection of differential linkage events

To detect differential linkage (i.e., allele-specific splicing), we selected common linkage events (same nucleotide variant and same introns) between the cells of the same individual. We then classified them into four categories explained by the two proposed scenarios. For scenario 1, we selected linkage events that were present in cancer cells, but not in normal cells, or vice versa. These events were grouped into the categories cancer-specific and normal-specific, respectively. For scenario 2, we selected linkage events that were present in both types of cells but significantly more prevalent in one compared to the other. These events were grouped into the categories "cancer-differential" and "normal-differential." To detect differential prevalence, we used the Fisher's Exact test for the number of cells with the linkage event and the number of cells that were testable for linkage in each group of cells.

Functional enrichment analysis

We performed functional enrichment analysis for genes containing the differential or condition-specific events identified in the lung cancer dataset. For this analysis, we used Cytoscape (*46*) network analysis, which uses the STRING network database (*31*). We selected terms from the following molecular pathway databases: Reactome Pathways, Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways, WikiPathways, and Gene Ontology Biological Process. Cytoscape uses a collection of methods to perform enrichment analysis based on the overrepresentation of genes of a defined molecular pathway in the query gene list. This overrepresentation is typically calculated using a hypergeometric test. The top three most

significant terms (as determined by the false discovery rate value) from these databases are reported in Fig. 5D.

Code availability

The scAllele software is available at Zenodo (https://doi.org/10.5281/zenodo.6558451) and in our github repository (https://github.com/gxiaolab/scAllele/).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abn6398

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. E. Papalexi, R. Satija, Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18**, 35–45 (2018).
2. H. Zafar, Y. Wang, L. Nakhleh, N. Navin, K. Chen, Monovar: Single-nucleotide variant detection in single cells. *Nat. Methods* **13**, 505–507 (2016).
3. J. K. Kim, A. A. Kolodziejczyk, T. Illicic, S. A. Teichmann, J. C. Marioni, Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.* **6**, 1–9 (2015).
4. Y. Jiang, N. R. Zhang, M. Li, SCALE: Modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol.* **18**, 1–15 (2017).
5. K. Choi, N. Raghupathy, G. A. Churchill, A Bayesian mixture model for the analysis of allelic expression in single cells. *Nat. Commun.* **10**, 1–11 (2019).
6. P. M. Schnepp, M. Chen, E. T. Keller, X. Zhou, SNV identification from single-cell RNA sequencing data. *Hum. Mol. Genet.* **28**, 3569–3583 (2019).
7. F. Liu, Y. Zhang, L. Zhang, Z. Li, Q. Fang, R. Gao, Z. Zhang, Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome Biol.* **20**, 242 (2019).
8. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907v2 (2012).
9. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
10. T. M. Yamawaki, D. R. Lu, D. C. Ellwanger, D. Bhatt, P. Manzanillo, V. Arias, H. Zhou, O. K. Yoon, O. Homann, S. Wang, C. M. Li, Systematic comparison of high-throughput single-cell RNA-seq methods for immune cell profiling. *BMC Genomics* **22**, 66 (2021).
11. M. J. Zhang, V. Ntranos, D. Tse, Determining sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun.* **11**, 1–11 (2020).
12. J. Fan, J. Hu, C. Xue, H. Zhang, K. Susztak, M. P. Reilly, R. Xiao, M. Li, ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLOS Genet.* **16**, e1008786 (2020).
13. G. Li, J. H. Bahn, J. H. Lee, G. Peng, Z. Chen, S. F. Nelson, X. Xiao, Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res.* **40**, e104 (2012).
14. K. Amoah, Y. H. E. Hsiao, J. H. Bahn, Y. Sun, C. Burghard, B. X. Tan, E. W. Yang, X. Xiao, Allele-specific alternative splicing and its functional genetic variants in human tissues. *Genome Res.* **31**, 359–371 (2021).
15. Y. H. E. Hsiao, J. H. Bahn, Y. Yang, X. Lin, S. Tran, E. W. Yang, G. Quinones-Valdez, X. Xiao, RNA editing in nascent RNA affects pre-mRNA splicing. *Genome Res.* **28**, 812–823 (2018).
16. Y.-H. E. Hsiao, J. H. Bahn, X. Lin, T.-M. Chan, R. Wang, X. Xiao, Alternative splicing modulated by genetic variants demonstrates accelerated evolution regulated by highly conserved proteins. *Genome Res.* **26**, 440–450 (2016).
17. Q. Zhang, X. Xiao, Genome sequence-independent identification of RNA editing sites. *Nat. Methods* **12**, 347–350 (2015).
18. P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, Y. Gilad, Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 1–15 (2017).
19. J. M. Zook, J. McDaniel, N. D. Olson, J. Wagner, H. Parikh, H. Heaton, S. A. Irvine, L. Trigg, R. Truty, C. Y. McLean, F. M. De La Vega, C. Xiao, S. Sherry, M. Salit, An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* **37**, 561–566 (2019).
20. L. Clarke, S. Fairley, X. Zheng-Bradley, I. Streeter, E. Perry, E. Lowy, A.-M. Tassé, P. Flicek, The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Res.* **45**, D854–D859 (2017).
21. A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg, A. O. M. Wilkie, G. McVean, G. Lunter, Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
22. P. Krusche, L. Trigg, P. C. Boutros, C. E. Mason, F. M. De La Vega, B. L. Moore, M. Gonzalez-Porta, M. A. Eberle, Z. Tezak, S. Lababidi, R. Truty, G. Asimenos, B. Funke, M. Fleharty, B. A. Chapman, M. Salit, J. M. Zook, Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
23. S. M. Karst, R. M. Ziels, R. H. Kirkegaard, E. A. Sørensen, D. McDonald, Q. Zhu, R. Knight, M. Albertsen, High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat. Methods* **18**, 165–169 (2021).
24. Z. Sun, A. Bhagwate, N. Prodduturi, P. Yang, J. P. A. Kocher, Indel detection from RNA-seq data: Tool evaluation and strategies for accurate detection of actionable mutations. *Brief. Bioinform.* **18**, 973–983 (2017).
25. A. Maynard, C. E. McCoach, J. K. Rotow, L. Harris, F. Haderk, D. L. Kerr, E. A. Yu, E. L. Schenk, W. Tan, A. Zee, M. Tan, P. Gui, T. Lea, W. Wu, A. Urisman, K. Jones, R. Sit, P. K. Kolli, E. Seeley, Y. Gesthalter, D. D. Le, K. A. Yamauchi, D. M. Naeger, S. Bandyopadhyay, K. Shah, L. Cech, N. J. Thomas, A. Gupta, M. Gonzalez, H. Do, L. Tan, B. Bacaltos, R. Gomez-Sjoberg, M. Gubens, T. Jahan, J. R. Kratz, D. Jablons, N. Neff, R. C. Doebele, J. Weissman, C. M. Blakely, S. Darmanis, T. G. Bivona, Therapy-induced evolution of human lung cancer revealed by single-cell RNA sequencing. *Cell* **182**, 1232–1251.e22 (2020).
26. H. Heaton, A. M. Talman, A. Knights, M. Imaz, D. J. Gaffney, R. Durbin, M. Hemberg, M. K. N. Lawniczak, Souporcell: Robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020).
27. M. L. DeDiego, L. Martinez-Sobrido, D. J. Topham, Novel functions of IFI44L as a feedback regulator of host antiviral responses. *J. Virol.* **93**, e01159-19 (2019).
28. K. Yamamoto, T. Kawakubo, A. Yasukochi, T. Tsukuba, Emerging roles of cathepsin E in host defense mechanisms. *Biochim. Biophys. Acta* **1824**, 105–112 (2012).
29. K. Gong, G. Guo, N. Beckley, Y. Zhang, X. Yang, M. Sharma, A. A. Habib, Tumor necrosis factor in lung cancer: Complex roles in biology and resistance to treatment. *Neoplasia* **23**, 189–196 (2021).
30. A. L. Ackerman, P. Cresswell, Cellular mechanisms governing cross-presentation of exogenous antigens. *Nat. Immunol.* **5**, 678–684 (2004).
31. D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, C. Von Mering, STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
32. C. Mayr, Regulation by 3′-untranslated regions. *Annu. Rev. Genet.* **51**, 171–194 (2017).
33. J. Chen, J. Guo, Comparative assessments of indel annotations in healthy and cancer genomes with next-generation sequencing data. *BMC Med. Genomics* **13**, 1–11 (2020).
34. F. Pfeiffer, C. Gröber, M. Blank, K. Händler, M. Beyer, J. L. Schultze, G. Mayer, Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **8**, 1–14 (2018).
35. M. Gymrek, D. Golan, S. Rosset, Y. Erlich, lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.* **22**, 1154–1162 (2012).
36. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
37. K. Wang, M. Li, H. Hakonarson, ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
38. A. Simons, FastQC: A quality control tool for high throughput sequence data (2010).
39. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J.* **17**, 10–12 (2011).
40. N. Joshi, J. Fass, Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (2011); https://github.com/najoshi/sickle.
41. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
42. Broad Institute, Picard toolkit (2019).
43. J. G. Cleary, R. Braithwaite, K. Gaastra, B. S. Hilbush, S. Inglis, S. A. Irvine, A. Jackson, R. Littin, M. Rathod, D. Ware, J. M. Zook, L. Trigg, F. M. De La Vega, Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *bioRxiv* , 023754 (2015).
44. J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, S. A. Forbes, COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
45. S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, K. Sirotkin, dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
46. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

Acknowledgments: We thank members of the Xiao laboratory for helpful discussions and comments on this work. We appreciate the helpful discussions with S. Mangul. **Funding:** This work was supported in part by grants from the National Institutes of Health (U01HG009417 and R01AG056476 to X.X.) and the Jonsson Comprehensive Cancer Center at UCLA. **Author contributions:** G.Q.-V. and X.X. conceived the study, analyzed the data, and wrote the manuscript. T.F. performed the experiments. T.W.C. contributed to data analysis. All authors read and edited the manuscript. **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. Variant calls and linkage events from the GM12878 and iPSCs for individuals NA19098, NA19101, and NA19239, and the lung cancer cells are available at Zenodo (https://doi.org/10.5281/zenodo.6558593) and in our github repository (https://github.com/gxiaolab/scAllele/tree/main/data).

Submitted 10 December 2021
Accepted 19 July 2022
Published 2 September 2022
10.1126/sciadv.abn6398

ScienceAdvances

scAllele: A versatile tool for the detection and analysis of variants in scRNA-seq

Giovanni Quinones-ValdezTing FuTracey W. ChanXinshu Xiao

Sci. Adv., 8 (35), eabn6398. • DOI: 10.1126/sciadv.abn6398

Science Advances (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.