ScSNViz: a user-friendly toolset for visualization and analysis of Cell-Specific Expressed SNVs

Allen Kim[1], Siera Martinez[1#], Nathan Edwards[2], and Anelia Horvath[1#]

#Correspondence

[1]McCormick Genomics and Proteomics Center, Department of Biochemistry and Molecular Medicine, School of Medicine and Health Sciences, The George Washington University, Washington, DC 20037, USA
[2]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University, Washington, DC 20057, USA

Abstract

Motivation

Understanding genetic variation at the single-cell level is crucial for insights into cellular heterogeneity, clonal evolution, and gene expression regulation, but there is a scarcity of tools for visualizing and analyzing cell-level genetic variants.

Results

We introduce scSNViz, a comprehensive R-based toolset for visualization and analysis of cell-specific expressed Single Nucleotide Variants (sceSNVs) within cell-barcoded single-cell RNA-sequencing (scRNA-seq) data. ScSNViz offers 3D sceSNV visualization capabilities for dimensionally reduced scRNA-seq gene expression data, compatibility with popular scRNA-seq processing tools like Seurat, cell-type classification tools such as SingleR and scType, and trajectory inference computation using Slingshot. Furthermore, scSNViz conducts estimation, summary, and graphical representation of statistical metrics pertaining to sceSNVs distribution and expression across individual cells. It also provides support for the analysis of individual sceSNVs as well as sets comprising multiple expressed sceSNVs of interest.

Availability

ScSNViz is implemented as user-friendly R-scripts, freely available on https://horvathlab.github.io/NGS/scSNViz, supported by help utilities, and requiring no specialized bioinformatics skills for use.

Introduction

Expressed single nucleotide variants (SNVs) fundamentally contribute to transcriptional heterogeneity. Cell-specific expressed SNVs (sceSNVs) encompass DNA-transcribed germline and somatic variants, as well as dynamic and context-dependent RNA-originating variants. These variants, whether originating from DNA or RNA, play significant roles in various cellular processes, impacting protein function, gene expression regulation, transcriptional and translational efficiency, intramolecular communication, and cell signaling. Understanding the distribution and expression of sceSNVs in individual cells can provide insights into cellular characteristics and the functions of these variants. Given the increasing availability of single-cell RNA sequencing (scRNA-seq) datasets and the growing interest in the role of sceSNVs in cellular diversity, there is a need for specialized tools for comprehensive analysis of these variants at the single-cell level.

We introduce scSNViz, a user-friendly toolset designed for quantitative visualization, statistical analysis, and graphical representation of the expression patterns of sceSNVs within individual cells in cell-barcoded scRNA-seq data (e.g.10xGenomics) (Zheng *et al.*, 2017). scSNViz enables quantitative 3D visualization of individual and sets of sceSNVs' expression and cell distribution across diverse cell types and pseudotime trajectories and is compatible with popular tools for scRNA-seq data processing and analysis.

Software Description

ScSNViz is an open-source software developed in R that leverages existing R packages for the processing of single-cell gene and SNV transcriptomics data while introducing novel features for analysis and visualization of sceSNVs. ScSNViz aims to enable users to explore and interpret SNV-related cellular heterogeneity and facilitate the integration of SNV information into existing scRNA-seq analytical workflows.

The workflow for using scSNViz is illustrated in Figure 1a, and the methods are detailed in Supplementary_Methods. ScSNViz requires two inputs: a genes-by-cells gene expression matrix and a list of sceSNVs with cellular barcodes. Users have the option to submit the genes-by-cells matrix as an R object provided that the data has undergone custom-processing, including quality control, scaling, and normalization of gene expression values and is a Seurat object. Alternatively, raw gene expression values generated by popular single-cell RNA sequencing aligners such as Cell Ranger and STARsolo (Kaminow *et al.*, 2021) can be provided. In the latter case, ScSNViz will process the data using default options by Seurat and Slingshot, and, optionally, will classify cell types using scType (Butler *et al.*, 2018; Ianevski *et al.*, 2022; Street *et al.*, 2018). This second option is intended for quick initial exploration of sceSNVs and can be useful for users who may not have extensive experience with scRNA-seq data processing tools.

The second input consists of a list of sceSNVs with their respective cellular barcodes in a tab-delimited format. For quantitative visualization of sceSNV expression, it is recommended that the list of sceSNVs be processed through SCReadCounts (Prashant *et al.*, 2021) or a similar tool. SCReadCounts tabulates the variant and reference read counts (N_VAR and N_REF, respectively) for each sceSNV across all cells, including cases where the genomic position of interest is exclusively covered by reference reads. This capability is crucial because it (1) distinguishes between a negative signal (no expression of the variant nucleotide in the locus) and no signal (no expression of the SNV-harboring gene), and (2) facilitates quantitative visualization of sceSNV expression. Using the provided N_VAR and N_REF counts, scSNViz calculates the expressed Variant Allele Frequency (VAF_RNA = N_VAR / (N_VAR + N_REF), where VAF_RNA $\in$ [0,1]), converts N_VAR, N_REF and VAF_RNA into color gradients, and overlays these gradients onto dimensionally reduced representations of the cells.

In cases where the sceSNVs are provided as the second input but have not been processed with SCReadCounts for N_VAR and N_REF estimation, scSNViz will generate an image distinguishing cells with the sceSNVs from those without. While this utility is not quantitative and provides less information regarding the expression of the SNV and its harboring gene across the cells (Figure_S1), which can still be valuable in various contexts, for example identifying cells with somatic mutations.

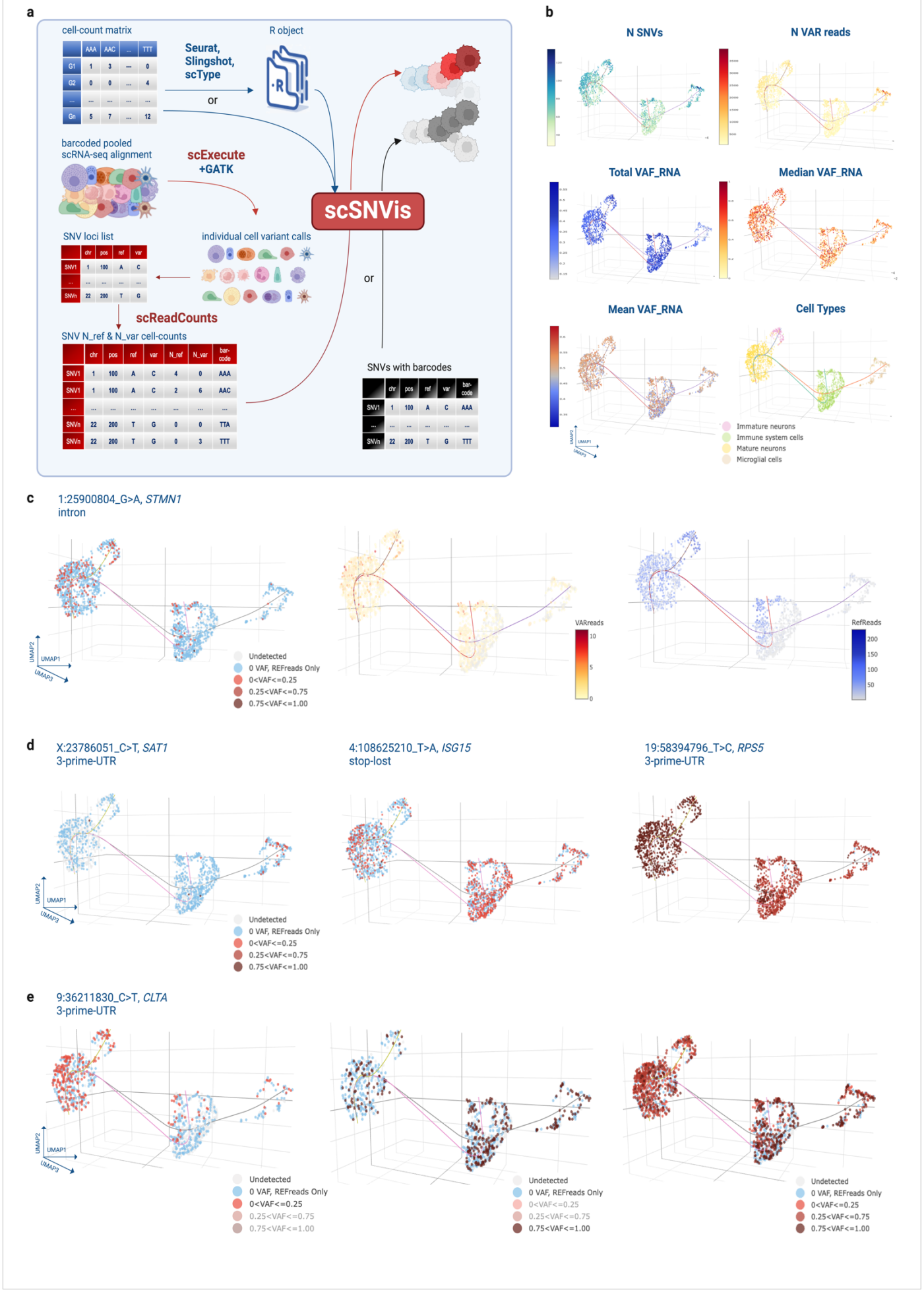
a
cell-count matrix
R object
Seurat, Slingshot, scType
or
barcoded pooled scRNA-seq alignment
scExecute +GATK
scSNVis
SNV loci list
individual cell variant calls
scReadCounts
SNV N_ref & N_var cell-counts
or
SNVs with barcodes
b
N SNVs
N VAR reads
Total VAF_RNA
Median VAF_RNA
Mean VAF_RNA
Cell Types
Immature neurons
Immune system cells
Mature neurons
Microglial cells
c
1:25900804_G>A, STMN1
intron
Undetected
0 VAF, REFreads Only
0<VAF<=0.25
0.25<VAF<=0.75
0.75<VAF<=1.00
VARreads
RefReads
d
X:23786051_C>T, SAT1
3-prime-UTR
4:108625210_T>A, ISG15
stop-lost
19:58394796_T>C, RPS5
3-prime-UTR
e
9:36211830_C>T, CLTA
3-prime-UTR

Figure 1. **a**. scSNViz workflow. scSNViz accepts either raw or processed gene-cell expression values as its first input, and a list of sceSNVs with cellular barcodes as its second input. It is recommended that the list of sceSNVs be processed through SCReadCounts, facilitating quantitative visualization based on the number of reference and variant read counts, expressed as color gradients. SCReadCounts-processed data also enables the distinction between SNV loci with solely reference read counts, solely variant read counts, and no read counts, thus allowing for the differentiation between cells with monoallelic reference expression, monoallelic variant expression. The software tools developed by our team and focused on variant analysis from single cells are shown in red. **b**. Sets of sceSNVs: Overall statistical metrics, accompanied by cell type classification by scType (offered as an option in scSNViz). To enhance visualization, users can select from multiple customizable color designs, including mono- and bi-chromatic gradients. The visualization is exemplified on the publicly accessible sample SAMN12799270, isolated from neuroblastoma tumor tissue. **c**. Individual sceSNVs: VAF_RNA, N_VAR, N_REF, are visualized for each individual sceSNV from the submitted list. For a comprehensive understanding of sceSNV expression, exploring the three metrics simultaneously is recommended. **d**. Multiple distinct and characteristic patterns of allelic expression were observed in our study. Examples include exclusive expression of the sceSNV X:23786051_C>T in SAT1 within a subpopulation of cells from the microglia cluster (left), preferential expression of sceSNV 4:108625210_T>A in ISG15 in the immune cells (middle), and preferential monoallelic expression of sceSNV 19:58394796_T>C in RPS5 in neurons, with biallelic expression in the remaining cells. **e**. scSNViz offers a clickable option to selectively visualize only cells with certain VAF_RNA, N_VAR, and N_REF values: cells with low VAF_RNA for sceSNV 9:36211830 C>A in *CLTA* (below 0.25, left), cells with high VAF_RNA (above 0.75, middle), compared to an image visualizing all cells (right). In this example an interesting observation is the neurons showing lower allelic expression of 9:36211830 C>A across most cells of the cluster, whereas in the immune cluster over half cells show high allelic expression with the rest only expressing the reference allele.

Provided with the gene-expression values and a set of sceSNVs with cellular barcodes (with or without scReadCounts processing), scSNViz generates the following outputs. First, scSNViz computes overall statistical metrics for the set of sceSNVs, including the number of sceSNVs per cell (N_SNVs), the total count of the N_VAR and N_REF count alleles per cell, as well as the total, median, and mean VAF_RNA (Figure 1b, the set of sceSNVs used is shown in S_Table1). These metrics are calculated across all the sceSNVs in the input list and are designed to offer insights into the overall expressed nucleotide variance within the system.

Specifically, N_SNVs can indicate cell clusters with a high number of expressed variants, associated, for instance, with increased somatic mutagenesis in cancer cells, or with elevated RNA-originating variance related to processes such as RNA editing or transcriptional infidelity. Likewise, N_VAR can pinpoint cell clusters exhibiting elevated expression of variant alleles. This metric can be further analyzed in conjunction with the presence or absence of N_REF to differentiate clusters with high gene expression from those with high variant allele expression. Total VAF_RNA is calculated by dividing the sum of N_VAR counts across the sceSNVs (loci) by the total reads covering the sceSNV loci (N_VAR + N_REF), providing an assessment of SNV expression magnitude across cell populations. The median and mean VAF_RNA offer insights into the central tendency and variability of SNV expression levels within the cell population, respectively, and can be informative when studying processes such as transcriptional bursting, cellular heterogeneity, and regulatory mechanisms influencing gene expression. Like N_SNVs, the remaining metrics can indicate both DNA-transcribed variant expression and RNA-originating variant activity. scSNViz visualizes the overall statistical metrics through individual cell projections in 3D space (see Figure 1b) and histograms delineating metric distribution patterns within bins (Figure_S2), with the latter supporting a summary-level exploration of cellular heterogeneity.

In addition to overall statistical metrics, scSNViz visualizes N_VAR, N_REF, and VAF_RNA for each sceSNV in the submitted list. These three metrics offer distinct aspects of the sceSNV expression, and, considered together, provide comprehensive insights into sceSNV expression dynamics (Figure 1b and Figure 1c). While VAF_RNA indicates the relative expression of the variant allele, it is essential to consider it alongside the absolute numbers of N_VAR and N_REF. Together, these metrics can distinguish between DNA mono- or biallelic events and RNA-originating events that may not adhere to mono- or bi-allelic

patterns. Applying scSNViz to over 40 scRNA-seq datasets from publicly available normal and tumor datasets, we observed that the majority of the individual sceSNVs exhibit highly specific and informative expression patterns. For instance, some sceSNVs exhibit strictly preferential expression in a particular cell cluster or subcluster, while others display expression across most cells in the sample, albeit with varying relative expression of the variant allele in different cell clusters (Figure 1d and Figure_S3). While the first pattern can be consistent with somatic variance or cell-specific RNA-editing, the second may indicate allele-specific regulation of gene expression. In both cases, examining simultaneously VAF_RNA, N_VAR, and N_REF can help distinguish if the level of variant allele expression is simply linked to the expression of the harboring gene or is independent. The latter scenario suggests potential biological functionality of the sceSNV of interest.

Furthermore, scSNViz offers customizable plotting parameters, including colors, legends, labels, axes, cell sizes, and cell borders. It also implements a clickable option to selectively visualize only cells with certain VAF_RNA, N_VAR, and N_REF values (Figure 1e). Additionally, scSNViz provides flexibility in dimensionality reduction methods, supporting UMAP, tSNE, or PCA (Figure_S4).

Moreover, scSNViz provides the option to apply thresholds on parameters to explore specific subsets of sceSNVs based on their expression level or their presence in a certain number of cells. For instance, users can choose to exclude sceSNVs expressed in fewer or more than a specified number of cells (e.g., 20 cells) or opt not to visualize cells expressing sceSNVs covered by low number of sequencing reads (e.g., less than 3 reads). This functionality aids in user quality control and highlights SNVs with specific characteristics. For example, SNV calls in loci covered by only one or two variant reads (N_VAR) are more likely to result from technical factors. Similarly, sceSNVs expressed in only a small number of cells may lack informative power due to insufficient statistical analysis of features of the cells expressing the sceSNV. At the same time, sceSNVs expressed with high N_VAR in a particular cell cluster, while the rest of the clusters mostly exhibit the reference allele, are likely to indicate a biological relationship between the sceSNV and the specific cells.

Availability and Implementation

scSNViz is distributed as a freely available R toolkit (https://horvathlab.github.io/NGS/scSNViz), which does not require specialized bioinformatics skills for use. scSNViz delivers robust performance across various computing environments. Typically, its execution takes minutes, whether on high-performance computing (HPC) environments or local machines, including laptops. To assist users in fine-tuning customizable parameters, scSNViz includes a comprehensive help utility that provides detailed information and guidance on usage options, along with access to an example dataset.

Discussion

scSNViz is intended for researchers investigating cell-level DNA and/or RNA-originating nucleotide variance across various biological contexts, such as genotype-phenotype linkage studies, exploration of clonal dynamics within heterogeneous cell populations, and analysis of RNA-originating single-cell genetic variation. ScSNViz offers a user-friendly interface designed for custom visualization of expressed genetic variance derived from scRNA-seq data for both individual sceSNVs and sets of sceSNVs. The visualization of both sets and individual sceSNVs offers distinct information value, providing insights into overall expression patterns and specific allele behaviors, respectively. The integration of these two visualization sources enables users to achieve a comprehensive understanding of the expressed variance landscape at the single-cell level.

The selection of sets of sceSNVs is important, as each set can offer specific insights into different levels of transcriptional heterogeneity, such as the expression of biallelic germline alleles, the evolution of somatic mutations, or the dynamic of the RNA-originating variance. These sets can encompass any genomic positions of interest, including known genetic variation loci like somatic mutational hotspots or RNA-editing sites, for which variant calling may not be necessary. Additionally, sets may consist of variants called or inferred within the dataset itself. In our workflow, we employ cell-level variant calling facilitated by our tool

for cell-level stratification of processes, SCExecute (Edwards *et al.*, 2023), in combination with GATK and Strelka (McKenna *et al.*, 2010; Kim *et al.*, 2018) (See Supplementary Methods).

ScSNViz is versatile for a range of applications, from detailed investigations into allele dynamics of specific mutations and genes to analyzing allele-related biological processes like chromosome X-silencing, transcriptional bursting, and imprinting, as well as clonal lineage dynamics. Moreover, by facilitating the study of relationships between variant and gene expression, scSNViz contributes to understanding sequence-dependent regulation of gene expression.

Given the current understudied nature of sceSNVs from cell-barcoded scRNA-seq data, we anticipate that scSNViz will drive substantial advancements in data generation, analysis, and interpretation of expressed nucleotide variation. Of note, scSNViz is directly applicable to long read scRNA-seq data, including the emerging datasets produced by Nanopore and PacBio which are expected to significantly expand the number and types of mutations identifiable from scRNA-seq data.

In conclusion, we believe that scSNViz offers a rapid and effective solution for assessing cell-level expressed nucleotide variance. Its user-friendly interface, diverse summarization and visualization capabilities, and wide range of customizable features enhance its utility across several biological contexts.

Authors Contribution

AH developed the concept and drafted the manuscript, AK implemented the software, with the assistance of SM. NE tested and optimized the software. AH devised and supervised the study. All authors have read and approved the final manuscript.

Funding

This work has been supported by NCI award R21CA271066 to Anelia Horvath and by McCormick Genomic and Proteomic Center at George Washington University.

References

Butler,A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*, **36**.

Edwards,N. *et al.* (2023) SCExecute: custom cell barcode-stratified analyses of scRNA-seq data. *Bioinformatics*, **39**.

Ianevski,A. *et al.* (2022) Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun*, **13**.

Kaminow,B. *et al.* (2021) STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv*.

Kim,S. *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods*, **15**.

McKenna,A. *et al.* (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, **20**.

Prashant,N.M. *et al.* (2021) SCReadCounts: estimation of cell-level SNVs expression from scRNA-seq data. *BMC Genomics*, **22**.

Street,K. *et al.* (2018) Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, **19**.

Zheng,G.X.Y. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat Commun*, **8**.