SCReadCounts: Estimation of cell-level SNVs from scRNA-seq data

Prashant NM[1], Nawaf Alomran[1*], Yu Chen[2], Hongyu Liu[1], Pavlos Bousounis[1], Mercedeh Movassagh[3], Nathan Edwards[2], and Anelia Horvath[1,3#]* **equal contribution, #correspondence*

[1]McCormick Genomics and Proteomics Center, School of Medicine and Health Sciences, The George Washington University, 20037 Washington, DC, USA, [2]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University, Washington, DC 20057, USA,[3] Harvard T.H. Chan School of Public Health, Boston, MA, 02115, USA.

Summary: SCReadCounts is a method for a cell-level estimation of the sequencing read counts bearing a particular nucleotide at genomic positions of interest from barcoded scRNA-seq alignments. SCReadCounts generates an array of outputs, including cell-SNV matrices with the absolute variant-harboring read counts, as well as cell-SNV matrices with expressed Variant Allele Fraction (VAF_{RNA}); we demonstrate its application to estimate cell level expression of somatic mutations and RNA-editing on cancer datasets. SCReadCounts is benchmarked against GATK and Samtools and is freely available as a 64-bit self-contained binary distribution (Linux), along with MacOS and Python installation.

Availability: https://github.com/HorvathLab/NGS/tree/master/SCReadCounts

Supplementary Information: SCReadCounts_Supplementary_Data.zip

Introduction Estimation of single nucleotide variants (SNV) expression from single cell RNA sequencing (scRNA-seq) data is an emerging field with quickly expanding applications, including assessment of allele expression, transcriptional burst kinetics, quantitative loci traits (QTLs), haplotype inference, X-chromosome inactivation, and demultiplexing (Vu *et al.*, 2019; Larsson *et al.*, 2019; Reinius *et al.*, 2016; Van Der Wijst *et al.*, 2018; Hongyu Liu; Prashant *et al.*, 2019, 2020; Edsgärd *et al.*, 2016; Xu *et al.*, 2019; Griffiths *et al.*, 2017; D'Antonio-Chronowska *et al.*, 2019). In cancer, studies on cell-level genetic heterogeneity have been instrumental to trace lineages and resolve subclonal architecture (Vu *et al.*, 2019; Lee *et al.*, 2017; Puram *et al.*, 2017; Venteicher *et al.*, 2017; Müller *et al.*, 2016). Genetically distinct tumor cell populations are shown to exert gene expression (GE) heterogeneity, and to differ in clinical features. However, it is currently challenging to extract genetically distinct cells for downstream analyses (Petti *et al.*, 2019).

To aid these types of studies, we have developed a tool – SCReadCounts - for cell-level estimation of reference and variant read counts (n_{var} and n_{ref}, respectively), from pooled barcoded scRNA-seq alignments. Provided a list of variant sites, SCReadCounts estimates n_{var} and n_{ref}, calculates expressed Variant Allele Fraction ($VAF_{RNA} = n_{var} / (n_{var} + n_{ref})$) and outputs cell-SNV matrices. The cell-SNV matrices can be used as inputs for a wide range of downstream analyses. We demonstrate the application of SCReadCounts to estimate cell level expression of somatic mutations and RNA-editing on cancer datasets from Adrenal Neuroblastoma (Dong *et al.*, 2020). We also exemplify a downstream application to correlate VAF_{RNA} to GE using scReQTL (Liu *et al*., 2020).

Results ScReadCounts requires two inputs: a barcoded scRNA-seq alignment, and a list of genomic positions of interest. In the exemplified workflow (Fig.1a), the barcodes and the Unique Molecular Identifiers (UMIs) are processed using UMItools, and the sequencing reads are aligned to the reference genome (GRCh38) using STAR (v.2.5.7b)(Smith *et al.*, 2017; Dobin *et al.*, 2013). The resulting pooled alignments can be filtered to correct for allele-mapping bias (WASP, Van De Geijn *et al.*, 2015); this filtering utilizes the same list of positions to be used as input for scReadCounts. Examples of positions of interest include SNVs called in the corresponding alignments (i.e. using GATK, see Fig.1a), or user-specified lists of coordinates from external sources, such as sets of somatic mutations (COSMIC) or RNA-editing loci (Auwera Mauricio O. *et al.*, 2002; Tate *et al.*, 2019; Picardi *et al.*, 2017).

SCReadCounts generates three main outputs in a tab-separated value format: (1) a table containing n_{var} and n_{ref} with quality and filtering metrics for each barcode, (2) a cell-SNV matrix with the absolute n_{var} and n_{ref} counts, and, (3) a cell-SNV matrix with the VAF_{RNA} estimated at a user-defined threshold of minimum number of required sequencing reads (minR) (S_Fig.1).

Performance We compared the SCReadCounts estimations with the analogous modules of the mpileup utility of Samtools (Li *et al.*, 2009) and the haplotype caller of GATK. SCReadCounts default options generate nearly identical values to mpileup and GATK (Fig.1b and S_Fig.2). SCReadCounts uses, by default, a very simple read-filtering criteria, but it can also be readily configured to achieve scenario-specific, mpileup-consistent, or GATK-consistent results, with optional explicit output of the number of reads discarded by each filtering rule. In regard to efficiency, on our system (2x14 cores CPUs with 1.5TB RAM compute node) processing of a file containing ~5000 cells, ~150mln reads, and ~80K SNVs, requires approximately 4h for the estimation of n_{var} and n_{ref}, and up to 2 minutes for the generation of the cell-SNV matrices. The later enables the users to quickly generate VAF_{RNA} matrices at various minR.

Applications We first tested the ability of SCReadCounts to assess the expression of known somatic mutations in the neuroblastoma scRNA-seq dataset. To do that we extracted from COSMIC 404,693 single nucleotide substitutions located in Cancer Census Genes and not overlapping with known germline SNV loci (S_Table 1). SCReadCounts estimated detectable expression of a number of COSMIC mutations in a low to moderate proportion of the individual cells. An example is COSV67805199 in the gene *RHOH*, with a variable VAF_{RNA} (minR = 5) across a number of cells from sample SRR10156295 (Fig.1c).

Next, we sought to assess if SCReadCounts can detect cell-specific RNA-editing levels. We used the same set of neuroblastoma samples, this time assessing 101,713 single nucleotide editing events from the REDI database (S_Table_2) (Picardi *et al.*, 2017). At minR = 5, SCReadCounts identified multiple loci with variable levels of editing, some with apparent clustering in the two-dimensional space of certain cell-types (cell-types are classified using Seurat and SingleR (Hafemeister and Satija, 2019; D. *et al.*, 2019) An example of RNA-editing in position 14:100846310_A>G) located in the cancer-implicated lincRNA *MEG3* is shown on Fig. 1d. Cells with edited *MEG3* were seen predominantly in neurons, where they showed clear positional clustering. The levels of editing showed positive correlation (albeit with small effect size) with the expression of the harboring *MEG3* (cis-scReQTL, Fig1.e left,), and higher size effect correlations with other genes (trans-scReQTLs), for example the Neuronal Vesicle Trafficking Associated *NSG1* (Fig.1e, right).

Considerations When applying SCReadCounts, the following considerations are in place. First, as mentioned earlier, modeling sequencing errors in the UMI is essential. Second, estimation of n_{var}

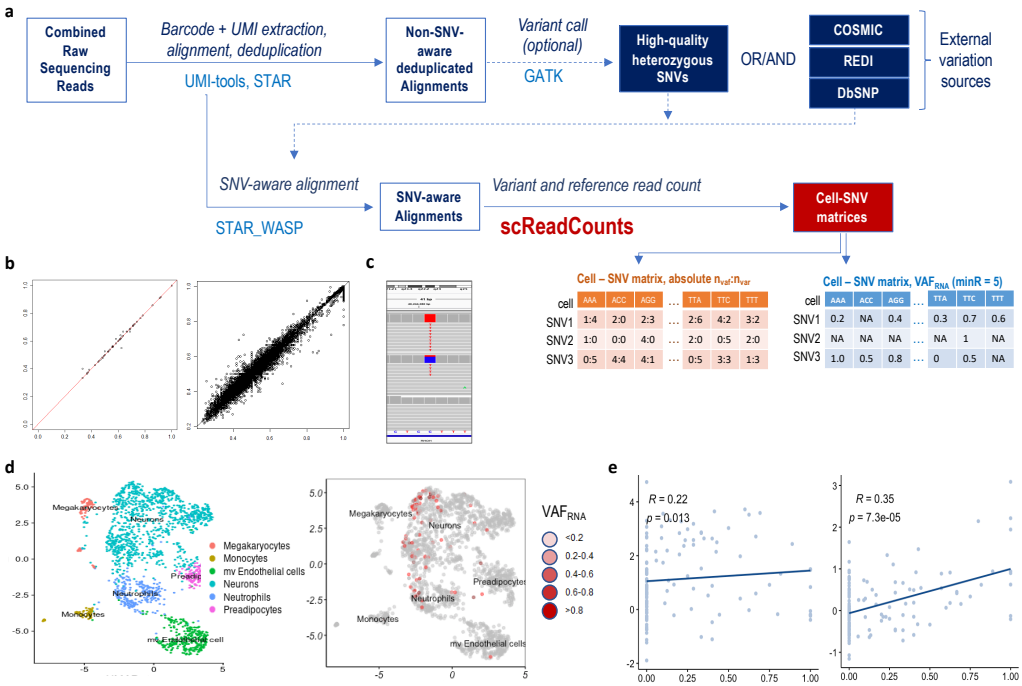

Figure 1. a. ScReadCounts workflow using publicly available tools. **b.** Concordance of read counts estimations (VAF_{RNA}) between SCReadCounts (y-axis) and mpileup (x-axis) from an individual cell alignment (left), and a pooled alignment (right); sample SRR10156295. **c**. IGV visualization of variable VAF_{RNA} of the somatic mutation COSV67805199 in the gene *RHOH* in three individual cells of sample SRR10156295. **d**. Two-dimensional UMAP clusters showing cells classified by type (left) and visualizing RNA-editing levels in the gene *MEG3*, where the intensity of the color corresponds to the proportion of edited reads (sample SRR10156295). **e**. Left: Cis-scReQTL correlation between edited levels (x-axis) and GE (y-axis) of *MEG3.* Right: trans-scReQTL between editing levels (x-axis) in MEG3 and the *NSG1* GE (y-axis); the sample is SRR10156295).

is sensitive to mapping, therefore WASP-correction of the alignments is recommended. Third, when estimating VAF_{RNA}, the selection of minimal required number of reads is important; our results show that for the most of the current scRNA-seq datasets, minR=5 provides a reasonable balance between randomness of sampling (high minR) and inclusivity (low minR) (Prashant *et al.*, 2020).

In summary, we believe that SCReadCounts supplies a fast and efficient solution for estimation of scRNA-seq genetic variance.

Funding: This work was supported by MGPC, The George Washington University; [MGPC_PG2019 to AH].

Conflict of Interest: None declared.

References

Auwera Mauricio O.,G.A.V. der C. *et al.* (2002) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinforma.*

D.,A. *et al.* (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.*

D'Antonio-Chronowska,A. *et al.* (2019) Association of Human iPSC Gene Signatures and X Chromosome Dosage with Two Distinct Cardiac Differentiation Trajectories. *Stem Cell Reports*.

Dobin,A. *et al.* (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*.

Dong,R. *et al.* (2020) Single-Cell Characterization of Malignant Phenotypes and Developmental Trajectories of Adrenal Neuroblastoma. *Cancer Cell*.

Edsgärd,D. *et al.* (2016) Scphaser: Haplotype inference using single-cell RNA-seq data. *Bioinformatics*.

Van De Geijn,B. *et al.* (2015) WASP: Allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*.

Griffiths,J.A. *et al.* (2017) Mosaic autosomal aneuploidies are detectable from single-cell RNAseq data. *BMC Genomics*.

Hafemeister,C. and Satija,R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*

Hongyu Liu, et al. scReQTL: an approach to correlate SNVs to gene expression from individual scRNA-seq datasets.

Larsson,A.J.M. *et al.* (2019) Genomic encoding of transcriptional burst kinetics. *Nature*.

Lee,J.K. *et al.* (2017) Spatiotemporal genomic architecture informs precision oncology in glioblastoma. *Nat. Genet.*

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*.

Müller,S. *et al.* (2016) Single-cell sequencing maps gene expression to mutational phylogenies in PDGF - and EGF - driven gliomas . *Mol. Syst. Biol.*

Petti,A.A. *et al.* (2019) A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat. Commun.*

Picardi,E. *et al.* (2017) REDIportal: A comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.*

Prashant,N.M. *et al.* (2019) Estimating allele-specific expression of SNVs from 10x Genomics Single-Cell RNA-Sequencing Data. 1–12.

Prashant,N.M. *et al.* (2020) Estimating the allele-specific expression of snvs from 10× genomics single-cell rna-sequencing data. *Genes (Basel)*.

Puram,S. V. *et al.* (2017) Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell*.

Reinius,B. *et al.* (2016) Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat. Genet.*

Smith,T. *et al.* (2017) UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.*

Tate,J.G. *et al.* (2019) COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.*

Venteicher,A.S. *et al.* (2017) Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science (80-.).*

Vu,T.N. *et al.* (2019) Cell-level somatic mutation detection from single-cell RNA sequencing. *Bioinformatics*.

Van Der Wijst,M.G.P. *et al.* (2018) Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.*

Xu,J. *et al.* (2019) Genotype-free demultiplexing of pooled single-cell RNA-seq. *Genome Biol.*