

A wealth of novel cell-specific expressed SNVs from tumor and normal scRNA-seq datasets

Christian Dillard¹, Evgenia Ulianova¹, Prashant NM^{1,2}, Hongyu Liu^{1,3}, Nathan Edwards⁴, and Anelia Horvath^{1#}

#Corresponding author

¹*McCormick Genomics and Proteomics Center, School of Medicine and Health Sciences, The George Washington University, Washington, DC, 20037, USA*

²*Departments of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA*

³*Division of Animal Sciences, University of Missouri, Columbia, MO, 65211, USA*

⁴*Department of Biochemistry and Molecular Medicine, School of Medicine and Health Sciences, The George Washington University, Washington, DC, 20037, USA*

Abstract

We demonstrate a novel variant calling strategy using barcode-stratified alignments on 25 tumor and normal 10XGenomics scRNA-seq datasets (>200,000 cells). Our approach identified 24,528 exonic non-dbSNP single cell expressed (sce)SNVs, a third of which are shared across multiple samples. The novel sceSNVs include unreported somatic and germline variants, as well as RNA-originating variants; some are expressed in up to 17% of the cells, and many are found in known cancer genes. Our findings suggest that there is an unacknowledged repertoire of expressed genetic variants, possibly recurrent and common across samples, in the normal and cancer transcriptome.

Key Words

SNV, SNP, mutation, scRNA-seq, single cell, sceSNV, scExecute, somatic mutation, RNA-editing, post-transcriptional modification

Background

To date, most genetic variation is studied in bulk sequencing datasets, where low (cellular) frequency variants are difficult to distinguish from sequencing errors and other artifacts. Low cellular frequency variants may indicate pre- or early-somatic clonality in cancer and normal tissues or cell-specific RNA post-transcriptional control [1,2]. Furthermore, the ability to detect variants at cell level and knowledge of natural cellular expressed variation are highly compatible with emerging cutting-edge technologies assessing cell-level gene and feature function (Perturb-Seq)[3], characterizing RNA-protein interactions (STAMP), or conducting targeted cell-level genome editing (RADARS) [3–5].

Results and Discussion

Here, we apply a novel strategy that utilizes barcode-stratified alignments and variant calling on 25 tumor and normal publicly accessible scRNA-seq datasets including prostate cancer (pc), cholangiocarcinoma (chlg), neuroblastoma (nb), normal adrenal (na) and normal embryo (ne (>200,000 cells) generated using the 10XGenomics 3'UTR workflow [6–8]. We extracted single cell alignments using scExecute [9], on pooled scRNA-seq alignments generated by STARsolo [10] and on each single cell alignment called variants using GATK [11] and Strelka2 [12] in parallel, retaining for downstream analysis only variants confidently identified by both callers. The pipeline is presented on Figure 1a.

Using this approach, we identified 24,528 exonic non-dbSNP cell-specific expressed single nucleotide variants (sceSNV) from the 25 samples (Table1, and S_Table1), 7,824 of which were observed in more than one sample, and 384 of which are observed in more than half of the datasets (Figure 1b). Of these non-dbSNP sceSNVs, 1,539 (6.3%) are reported in the database of somatic mutations COSMIC [13] and the rest are novel. Some novel sceSNVs were expressed in up to 17% of the cells in a dataset, and many were positioned in known cancer related genes. Cancer genes with multiple novel non-synonymous sceSNVs in more than one sample included *JUN*, *JAK1*, *NFKB1A*, *PIC3R1*, *RAC1* and *RBX1*. The genes with at least 15 novel non-synonymous sceSNVs in more than one sample were *HSP90AA1*, *HSP90AB1*, *ELOB*, *GSTP* and *JUN*. Novel sceSNVs had a higher proportion of missense and stop-codon involving substitutions, and a lower proportion of synonymous variants, as compared to the dbSNP sceSNVs in the same dataset ($p < 0.0001$, chi-squared test, Figure 1c). In addition, the novel sceSNVs had a higher proportion of A>G substitutions (17.7% vs 14.4% in the DbSNP sceSNVs, S_Figure1). For 1272 of the novel sceSNVs loci, we observed changes of the reference nucleotide into two different alternative nucleotides in different cells and samples, and for 120 loci – changes into all three possible alternative nucleotides (See S_Table1).

Next, we estimated the expressed Variant Allele Frequency (VAR_RNA, calculated as the proportion of sequencing reads carrying the alternative nucleotide over all the reads covering the locus) [14], and plotted it over the two-dimensional UMAP projection of the cells clustered and annotated based on gene expression [15,16] using scSNVis [17]. While many sceSNVs did not show a distinct cell-distribution pattern (Figure 1d), some were confined to particular

cell clusters and cell types despite the ubiquitous expression of their harboring gene (Figure 1e).

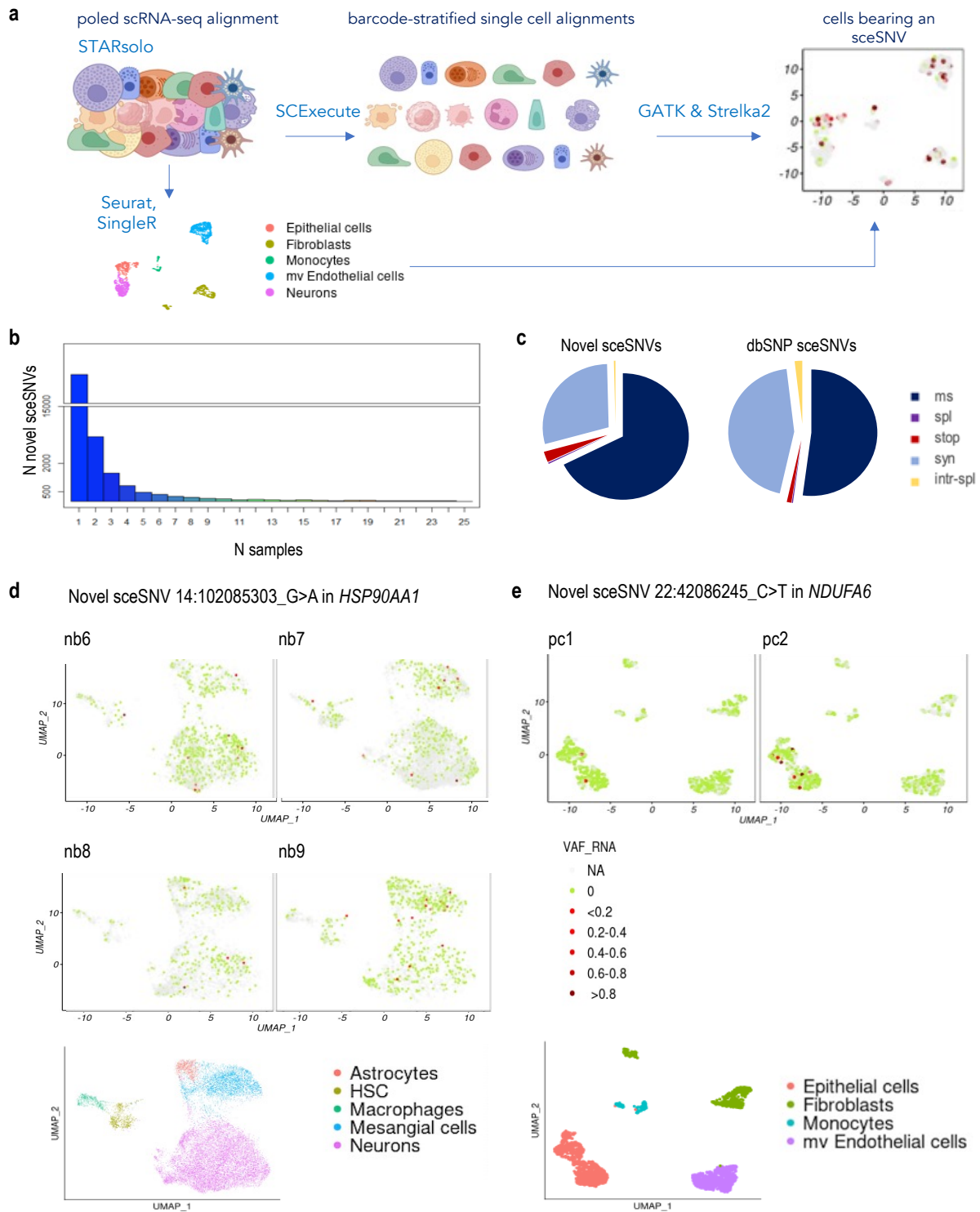


Figure 1. a. Processing pipeline. **b.** Distribution of the number of novel sceSNVs detected in one or more samples. **c.** Distribution of predicted functional annotations between the novel and the previously reported sceSNV; novel sceSNVs are enriched in missense and stop-codon involving substitutions. **d.** top: UMAP projections visualizing the cell distribution and the cellular expressed variant allele frequency (VAF_RNA) of the missense substitution at locus 14:102085303_G>A in the gene *HSP90AA1* across 4 samples from the neuroblastoma dataset. The red color intensity shows the relative expression of the sceSNV in cells with at least 3 sequencing reads covering the sceSNV locus, and the green color indicates that all the reads covering the SNV locus carried the reference nucleotide, consistent with non-zero gene expression. Cells in which the SNV locus is covered by less than 3 reads (corresponding to low or absent gene expression) are shown in grey. Bottom: cell types as classified by SingleR. **e.** top: UMAP projections visualizing the cell distribution of the missense substitution at locus 22:42086245_C>T in the gene *NDUFA6* in the two prostate cancer samples; the expression of the sceSNV is confined to epithelial cells (bottom).

Table 1. SceSNVs in 2 and more cells across 25 tumor and normal samples

Source	SampleID	Code	Chromium version	N cells	Exonic sceSNVs reported in dbSNP	non-dbSNP exonic sceSNVs in 2+ cells	
						Total N exonic sceSNVs	N sceSNVs in COSMIC
prostate cancer	SAMN16086830	pc1	v2	1455	3268	2147	159
	SAMN16086829	pc2	v2	2019	3211	2096	148
cholangio-carcinoma	SAMN13012145	chlg1	v2	3519	6269	4193	314
	SAMN13012146	chlg2	v2	2453	1215	253	32
	SAMN13012147	chlg3	v2	3579	4938	3823	299
	SAMN13012148	chlg4	v2	3769	4438	4022	325
	SAMN13012149	chlg5	v2	2738	5335	4888	316
	SAMN13012150	chlg6	v2	4993	3262	1896	146
neuro-blastoma	SAMN12799275	nb1	v2	4068	3331	2343	230
	SAMN12799274	nb2	v2	5789	2822	1776	135
	SAMN12799273	nb3	v2	6988	2810	2156	151
	SAMN12799272	nb4	v2	2997	2876	2198	189
	SAMN12799270	nb5	v2	6836	3721	3071	226
	SAMN12799269	nb6	v2	6994	3277	2744	180
	SAMN12799266	nb7	v2	12448	2591	1564	144
	SAMN12799264	nb8	v2	16554	1970	1229	99
	SAMN12799263	nb9	v2	4273	2314	1616	123
	SAMN15453063	nb10	v3	12441	3613	490	49
	SAMN15453064	nb11	v3	7582	3528	282	24
normal fetal adrenal	SAMN12799261	fa1	v2	9135	1151	385	27
	SAMN12799259	fa2	v2	4910	1251	430	27
	SAMN12799258	fa3	v2	26638	2944	2680	152
	SAMN12799257	fa4	v2	22283	2619	2369	158
normal embryo	SAMN15453062	ne1	v3	19414	5801	962	59
	SAMN15453069	ne2	v3	14192	3498	309	23
Statistics				sum	sum	sum	sum
				208067	82053	49922	3735

Variant calling from individual scRNA-seq alignments is a new and unexplored approach; therefore, to minimize false positives among the novel sceSNVs, we performed stringent quality filtering and examination of the sceSNV confidence at several levels. First, we used for our analyses the intersection of the highest quality calls in at least two cells per dataset by two callers widely used for RNA variant detection, GATK and Strelka2 (S_Methods) [18]. In parallel, for all novel sceSNV positions we estimated the variant read counts across all cells in each dataset using a method for cell-level tabulation of the sequencing read counts bearing reference and variant alleles from barcoded scRNA-seq alignments. [14]. SCReadCounts was fully concordant with the variant call results, identifying variant reads in all cells where sceSNV was called. Third, for 500 arbitrarily selected sceSNVs, we visually reviewed the local alignment using the Integrated Genomics Viewer (IGV) [19]. This analysis showed that for up to 2% of the calls per sample the originally called variant participates in a more complex alteration, involving two or more consecutive nucleotides. These calls were removed across the entire dataset by assessment the presence of more than one call within any 25 consecutive bases in the same cells. The remaining loci showed high quality alignment and confident variant presence (examples on S_Figure2, S_Figure3 and S_Figure4). Fourth, for the same set of 500 sceSNVs we examined the between-cell occurrence and distribution. Many sceSNVs show preferential presence in particular cell-clusters, suggesting a relationship between the sceSNV and cluster-specific gene expression (Figure 1e and S_Figure5) and the

related molecular context of cell-specific sceSNV. Fifth, we restricted our analysis to exonic regions, comprising the best-known reference variational context; we additionally filtered out difficult to assess repeated genomic regions and removed ribosomal genes due to their significant number of homologous genes and pseudogenes. We note that while these stringent measures are likely to increase false negatives, in this pilot approach we prioritize high-quality, high confidence sceSNV calls over a comprehensive search for sceSNVs.

The identified novel sceSNVs include previously unreported somatic and germline variants, as well as RNA-originating variation. Without cell-level matched DNA, the origin of sceSNVs is difficult to assign. However, for some sceSNVs, likely origin can be inferred from their cellular allele expression (VAF_RNA) and distribution across cells and samples. For example, while the above-described filtering cannot completely exclude germline DNA variants, they are unlikely to constitute a large proportion of novel sceSNVs because they are not reported in dbSNP and are observed in only a modest proportion of the cells per sample. On the other hand, low cellular frequency sceSNVs are consistent with both somatic DNA sceSNVs and sceSNVs resulting from post-transcriptional RNA-modifications such as RNA-editing. Somatic sceSNVs observed in multiple samples are likely to be reported in COSMIC, suggesting that the non-COSMIC novel sceSNVs are enriched in post-transcriptional RNA-modifications. While our set of novel sceSNVs do not contain previously reported RNA-editing loci and we exclude repetitive genome regions, known to contain the highest frequency of editing events, low cellular frequency editing of exonic positions is possible [20,21]. Another possibility for some non-COSMIC somatic sceSNVs is that they exist only in the period between their incidence and cell death, for example, if they impair critical mechanisms for cell-survival or replication. Such a scenario would prevent replication of cells bearing the sceSNV and result in low cellular frequency, and consequently challenge detection by bulk sequencing techniques. In regard to VAF_RNA, RNA-editing sceSNVs are likely to have higher VAF_RNA variance and attain any value between 0 and 1, whereas for many somatic mutations in biallelically expressed genes VAF_RNA is expected to be closer to 0.5. Finally, sceSNVs resulting from random transcription errors are unlikely to have high prevalence among our set as they are expected to be seen in a single molecule (i.e., represented by only one read per cell), which calls are excluded by our stringent quality filtering. Of note, regardless of DNA- or RNA-origin, the sceSNVs represent part of the functional cellular transcriptome, can exert effects on the proteins sequence and function, and can increase molecular variation of the cell at the multi-omics level.

Finally, to assess potential links between the novel sceSNVs and cell-specific gene expression, for a subset of sceSNVs we performed differential expression analysis between sceSNV-bearing cells and the rest of the cells in the dataset using Deseq2 [22]. Notably, for many sceSNV, we observed gene expression differences concordant across different samples, and consistent with current knowledge. For example, DE analysis of cells with and without the novel stop codon substitution 1:45511398_C>T in *PRDX1* identified 5 and 7 significantly deregulated genes in the two prostate cancer samples respectively (padj < 0.2), four of which are shared between the two samples and deregulated in the same direction (S_Table2). Three of these four genes – *TAGLN*, *ACTA2* and *MYL9* - participate in a well-known network (S_Figure6). The observation of shared and concurrently deregulated genes in cells bearing the same sceSNVs provides additional evidence for these sceSNVs as true positives and suggests causative, mechanistic and functional implications. Possible scenarios

include sceSNVs regulating the expression of one or a set of genes, or RNA-modification events taking place in cells with similar gene expression, both potentially more frequent for sceSNVs co-clustered in cells of similar types.

Conclusions

Here, we explore for the first-time expressed genetic variation at cell-level. Our findings suggest that there is an unappreciated repertoire of cell-level expressed genetic variation, possibly recurrent and common across samples, that participates in transcriptome function and dynamics in both cancer and normal cells. While the DNA- or RNA-origin of these variants is currently difficult to confidently determine, their appearance and, for some, relationship to certain gene-sets and cell types, suggests novel mechanisms and function for expressed genetic variation. We also demonstrate an assessment strategy for cellular and functional context by studying deregulated genes in the cells bearing specific sceSNVs. Furthermore, correlation between their VAF_{RNA} and the expression of harboring or other genes using scReQTL [23] may also provide needed biological context. The analysis of sceSNVs from scRNA-seq data is crucial to support emerging methods that use cell-level introduction and tracking of RNA-variants for manipulating cellular behavior and temporal deconvolution of cellular events [3–5]. Interpretation of these new methods' results will require prior knowledge of naturally occurring cell-level genetic and transcript variation, which we explore in this work. The study of expressed cell-specific variants in scRNA-seq data, as demonstrated here, has the potential to link expressed variation to tissue evolution and cell fate and has a role in the successful implementation of emerging single-cell biotechnologies.

Methods

The methods used to perform this study are described in detail in S_Methods.

References

1. Wijewardhane N, Dressler L, Ciccarelli FD. Normal Somatic Mutations in Cancer Transformation. *Cancer Cell*. 2021;39:125–9.
2. Picardi E, Horner DS, Pesole G. Single-cell transcriptomics reveals specific RNA editing signatures in the human brain. *RNA*. 2017;
3. Replogle JM, Norman TM, Xu A, Hussmann JA, Chen J, Cogan JZ, et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat Biotechnol*. 2020;
4. Brannan KW, Chaim IA, Marina RJ, Yee BA, Kofman ER, Lorenz DA, et al. Robust single-cell discovery of RNA targets of RNA-binding proteins and ribosomes. *Nat Methods*. 2021;
5. Jiang K, Koob J, Chen XD, Krajewski RN, Zhang Y, Villiger L, et al. Programmable eukaryotic protein expression with RNA sensors. *bioRxiv*. 2022;2022.01.26.477951.
6. Ma X, Guo J, Liu K, Chen L, Liu D, Dong S, et al. Identification of a distinct luminal subgroup diagnosing and stratifying early stage prostate cancer by tissue-based single-cell

- RNA sequencing. *Mol Cancer*. 2020;
7. Zhang M, Yang H, Wan L, Wang Z, Wang H, Ge C, et al. Single-cell transcriptomic architecture and intercellular crosstalk of human intrahepatic cholangiocarcinoma. *J Hepatol*. 2020;73:1118–30.
 8. Dong R, Yang R, Zhan Y, Lai H-D, Ye C-J, Yao X-Y, et al. Single-Cell Characterization of Malignant Phenotypes and Developmental Trajectories of Adrenal Neuroblastoma. *Cancer Cell*. 2020;
 9. Edwards N, Dillard C, Prashant NM, Liu H, Yang M, Ulianova E, et al. SCEXecute: cell barcode-stratified analyses of scRNA-seq data. *bioRxiv* [Internet]. 2022;2022.03.27.485988. Available from: <http://biorxiv.org/content/early/2022/03/28/2022.03.27.485988.abstract>
 10. Kaminow B, Yunusov D, Dobin A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv*. 2021;
 11. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr Protoc Bioinforma*. 2013;
 12. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods*. 2018;
 13. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. 2019;
 14. Prashant NM, Alomran N, Chen Y, Liu H, Bousounis P, Movassagh M, et al. SCReadCounts: estimation of cell-level SNVs expression from scRNA-seq data. *BMC Genomics* [Internet]. 2021 [cited 2021 Sep 23];22:689. Available from: <https://bmcbgenomics.biomedcentral.com/articles/10.1186/s12864-021-07974-8>
 15. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*. 2019;
 16. D. A, A.P. L, L. L, E. W, V. F, A. H, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol*. 2019;
 17. Hongyu Liu, Prashant NM, Nawaf Alomran, Pavlos Bousounis, Mercedeh Movassagh, Nathan Edwards and AH. scSNVvis: cell-level visualization of expressed SNVs from scRNA-seq data.
 18. Prashant N, Liu H, Dillard C, Ibeawuchi H, Alsaedy T, Chan KH, et al. Improved SNV discovery from barcode-stratified scRNA-seq alignments. *Genes (Basel)*. 2021;12.
 19. Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. Variant review with the integrative genomics viewer. *Cancer Res*. 2017.
 20. Picardi E, D'Erchia AM, Giudice C Lo, Pesole G. REDportal: A comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res*. 2017;
 21. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, et al. Widespread RNA and DNA sequence differences in the human transcriptome. *Science (80-)*. 2011;
 22. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;
 23. Liu H, Prashant NM, Spurr LF, Bousounis P, Alomran N, Ibeawuchi H, et al. scReQTL: an approach to correlate SNVs to gene expression from individual scRNA-seq datasets. *BMC Genomics* [Internet]. 2021;22:40. Available from: <https://doi.org/10.1186/s12864-020-07334-y>

