



Detecting somatic mutations in rare clones using single-cell multiomics

Rhys Gillman, Sonam Dukda, Jerome Sadir, et al.

Genome Res. published online August 13, 2026

Access the most recent version at doi:[10.1101/gr.281609.125](https://doi.org/10.1101/gr.281609.125)

P<P	Published online August 13, 2026 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Open Access	Freely available online through the <i>Genome Research</i> Open Access option.
Creative Commons License	This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution-NonCommercial 4.0 International license), as described at http://creativecommons.org/licenses/by-nc/4.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

Detecting somatic mutations in rare clones using single-cell multiomics

Rhys Gillman^{1*}, Sonam Dudka^{1*}, Jerome Samir², Raymond Louie³, Chris Goodnow⁴, Fabio Luciani^{2,5},
Mandeep Singh⁴, Matt A. Field^{1,4,#}

¹ Centre for Tropical Bioinformatics and Molecular Biology, College Science and Engineering, James
Cook University, Cairns, Queensland, 4878, Australia.

² School of Biomedical Sciences, University of New South Wales Sydney, Sydney, New South Wales,
2033, Australia.

³ School of Computer Science and Engineering, University of New South Wales Sydney; Sydney, New
South Wales, 2033, Australia.

⁴ Immunogenomics Lab, Garvan Institute of Medical Research, Darlinghurst, New South Wales ,
2010, Australia.

⁵ Westmead Institute for Medical Research, Westmead, NSW, 2145, Australia

* Contributed equally

Corresponding author

Email: matt.field@jcu.edu.au

Running Title: Identifying Rare Clonal Mutations with SCARCE

Abstract

Somatic mutations are increasingly recognised as drivers of diseases beyond cancer, including
autoimmune disorders. However, identifying rare, cell type-specific causal mutations remains
challenging due to their low frequency within heterogeneous cell populations. Traditional bulk
sequencing approaches lack the resolution to detect rare variants, underscoring the need for novel
methods specifically designed for single-cell data of heterogeneous cell populations.

We present an integrated single-cell multiomics computational framework, SCARCE (Single-Cell
Analysis of Rare Clonal Events), tailored for single-cell DNA sequencing (scDNA-seq) to statistically
prioritise rare somatic mutations within defined cell subpopulations. By comparing variant frequencies
across subpopulations, identified through either variant-based clustering or cell type annotation from

surface marker expression, we identify variants enriched in specific cell populations. Our method applies multiple user-adjustable filters and statistical enrichment tests to distinguish true somatic variants from technical artifacts.

SCARCE successfully identifies rare somatic mutations across three distinct datasets using technologies including MissionBio Tapestry and clonally-amplified whole-genome sequencing of single cells. We demonstrate that SCARCE correctly isolates and identifies true variants in a cell population comprising just 10 of 16,316 cells (0.06% of the total population). Furthermore, in an extensively characterised sample with known causal variants, SCARCE correctly identifies all known pathogenic variants among its top-ranked candidates.

SCARCE offers several advantages over existing tools in the field. By integrating genetic and phenotypic information at single-cell resolution, our approach opens new avenues for understanding the clonal origins of diseases driven by somatic mutations in small cell populations.

Introduction

Somatic mutations are ubiquitous, found in healthy individuals and known to play a key role in cancer initiation and progression. Furthermore, somatic mutations are increasingly recognised as driving a wide range of diseases beyond cancer (Solis-Moruno et al. 2023). Neurological diseases can be driven by somatic mutations, with mutations in genes like *MTOR* and *AKT3* causing focal cortical dysplasia and epilepsy (Lim et al. 2015). Age-related diseases also exhibit somatic mutation signatures, including clonal haematopoiesis of indeterminate potential (CHIP), where acquired mutations in hematopoietic stem cells drive cardiovascular disease and increased mortality risk independent of traditional cancer development (Libby et al. 2019). Additionally, somatic mutations accumulate in various tissues during aging, contributing to organ dysfunction and age-related pathologies such as atherosclerosis, where mutations in vascular cells promote inflammatory responses and plaque formation (Andreassi et al. 2000). Finally, in autoimmune and inflammatory disorders, somatic mutations can create pathogenic self-reactive cell clones that escape normal negative selection, leading to conditions such as autoimmune lymphoproliferative syndrome, where mutations in apoptosis-regulating genes cause aberrant lymphocyte survival and autoimmunity (Holzelova et al. 2004). While somatic mutations in

cancer typically drive uncontrolled proliferation, non-cancer somatic driver mutations often result in more subtle functional alterations affecting cell behaviour, developmental processes or immune regulation.

Bulk sequencing methods such as whole-genome (WGS) or exome sequencing (WES) have been employed successfully to identify germline driver mutations in a wide range of complex diseases (Field 2020; Hamzeh et al. 2021; Field 2022; Dukda et al. 2025) and cancers (Morin et al. 2011; Wilmott et al. 2015). Generally, however, bulk methods have proven unsuitable for detecting rare somatic mutations in mixed cell populations due to their inherent limitations in two critical areas: i) sensitivity and resolution of the technology, ii) the inability to differentiate cell types. Particularly challenging are cases where a somatic mutation occurs in a small subset of cells, such as in clonal haematopoiesis or early-stage cancer. These rare mutations may be present at very low allele frequencies within the heterogeneous population of cells. Since bulk sequencing captures a mixture of DNA from heterogeneous cell populations, the rare mutant alleles can be masked by the more abundant wild-type alleles, making it difficult to detect such mutations. Further, the sensitivity of bulk sequencing is often not high enough to distinguish these mutations from sequencing errors or background noise, leading to high false negative rates (Field et al. 2019). In contrast, techniques like deep sequencing or single-cell sequencing are better suited for detecting rare somatic mutations, as they offer higher sensitivity by sequencing more reads and/or analyzing individual cells, which increases the likelihood of capturing mutations present in a small fraction of the population (Dou et al. 2018).

Early methods to identify rare, cell type-specific somatic mutations focussed on improving the quality of individual reads, thus improving the confidence and sensitivity of rare variant calls. One example employed molecule tagging using a unique molecular identifier (UMI) to input DNA either prior to or simultaneously with amplification of target sequences to acquire consensus reads per molecule. Processing such data required UMI aware variant detection software such as UMI-VarCal (Sater et al. 2020) or DeepSNVMiner (Andrews et al. 2016). Similarly, duplex sequencing methods utilise complementary tags to sequence both strands of double-stranded DNA, identifying amplification and

sequencing errors that don't align between the two strands (Schmitt et al. 2012). However, these tools often fail to scale to large numbers of samples.

More recently, advances in molecular barcoding technology have led to the development of an increasing number of single-cell platforms capable of examining thousands of cells independently. These increasingly multiomic platforms are promising for a wide variety of applications; however, most have primarily focused on RNA-based applications which are known to be problematic for variant detection for a variety of reasons including incomplete transcript representation, isoform complexity and expression of mutated alleles (Lahnemann et al. 2020). Early multiomic platforms such as G&T-Seq enabled single cell DNA-based variant calling, but require sorting of cells into 96 or 384-well plate assays. In this case, cells are often pre-sorted into disease and control groups for somatic variant detection on a small numbers of cells (Singh et al. 2020b; Young et al. 2025). More recently, MissionBio's Tapestry platform offers high-throughput processing of up to tens of thousands of single cells using a microfluidic droplet-based approach, preparing libraries for amplicon DNA sequencing and producing hundreds of thousands of variant calls across thousands of cell lines (Pellegrino et al. 2018).

To differentiate cell types, single-cell protein sequencing of cell surface markers represents a powerful approach for high-resolution cell phenotyping, which has revolutionised our understanding of cellular heterogeneity within complex tissues. This technology, primarily implemented through methods such as Cellular Indexing of Transcriptomes and Epitopes by sequencing (CITE-seq), AbSeq, and related antibody-based approaches, known collectively as antibody-derived tags (ADT), enables simultaneous measurement of dozens to hundreds of surface proteins at single-cell resolution (Stoeckius et al. 2017). MissionBio's Tapestry platform, in addition to high-throughput single cell amplicon sequencing of DNA, also captures surface marker expression using ADT sequencing. While currently largely employed in differentiating immune cell types using combinations of well-known markers, cell surface markers are increasingly being employed in other applications including solid tissue biopsies (Lischetti et al. 2023), greatly broadening its application. When simultaneously combined with single cell DNA-based technologies, this approach enables the comparison of genotypes between specific cell

populations, providing a powerful approach for identifying cell type-specific somatic mutations (Singh et al. 2025).

Despite these advances, there are currently limited options for simultaneous end-to-end processing of these single-cell DNA sequencing and ADT datasets. For example, several options are available for Tapestry data including Optima (Pei et al. 2023) and Mosaic (Zhang et al. 2023) packages, however, these tools are primarily limited to filtering and visualization functions. To identify such mutations, we developed an analysis package, SCARCE (Single-Cell Analysis of Rare Clonal Events), for statistical prioritization of enriched variants in rare cell populations, implemented in R (R Core Team 2026). When ADT data are generated, cell types are annotated in a semi-automated manner, providing the user with various visualizations to decide on final cell type annotations. Variants are annotated using VEP (McLaren et al. 2010) and filters can be applied at three levels: per variant call, within cell types, and across all cells. The workflow prioritises variants enriched within each cell type, making the identification of cell type-specific somatic mutations simple and intuitive.

Results

We designed SCARCE to simplify the analysis of somatic variants in heterogeneous scDNA-seq data. Firstly, subpopulations of cells are identified and annotated. This is achieved either using surface-marker expression to differential biological cell types, or by clustering cells based on their mutation profiles to identify clonal populations, making SCARCE flexible to different sequencing technologies. SCARCE then identifies variants that are enriched in annotated subpopulations of cells by comparing the frequency of each mutation within and outside of the subpopulation (Figure 1).

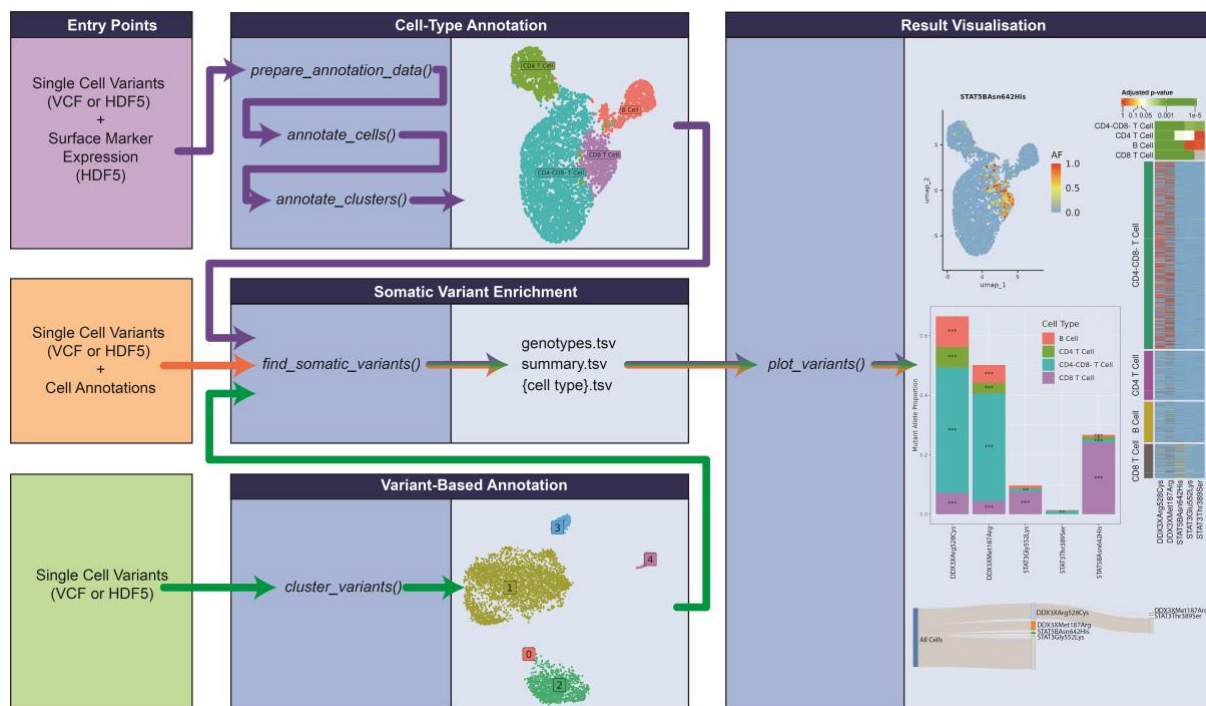


Figure 1: SCARCE Workflow diagram. SCARCE offers three different workflows depending on the input data available. The single-cell variants with surface marker expression pathway (purple path) begins with semi-automated cell type annotation based on ADT surface marker expression to define cell clusters. Users may alternatively provide their own cell type annotations to skip this step (orange path). Alternatively, if no surface marker expression data are available, cells can instead be clustered based on their variant genotypes to identify clusters carrying unique somatic variants (green path). All pathways converge at somatic variant enrichment, which statistically prioritises variants enriched in cell clusters, followed by visualisation of these results.

We first compared SCARCE's somatic variant enrichment with Mission Bio's in-house tools Mosaic and the R package Optima. Each tool was used to analyse Tapestry scDNA-seq data from an example dataset retrieved from the Mission Bio public data repository. We analysed data from 16,316 cells with 114,788 raw variant calls derived from a mixed sample containing two peripheral blood mononuclear cells (PBMCs) samples spiked with 0.06% cells of the KG-1 acute myeloid leukaemia macrophage cell

line. The targeted sequencing panel employed was the Tapestry Single-Cell DNA Myeloid panel. We sought to determine whether the tools could successfully identify variants unique to the KG-1 cells, which were expected to be represented by approximately 10 or fewer cells total. This would serve as a proxy and an extreme test case for the lower limits for somatic variant detection.

Both Optima and Mosaic offer cell-clustering based on variant genotypes, so for comparison we used the *SCARCE::cluster_variants()* function to perform the same task using the variant-only workflow. Cell populations that have acquired somatic mutations are expected to form independent clusters. As expected, SCARCE identified three clusters, including two of approximately equal size, likely corresponding to the two PBMC samples, and a small cluster consisting of only 10 cells (**Figure 2 A**). Neither Optima or Mosaic, which cluster cells based on a much smaller number of variants, appeared to successfully isolate a cluster corresponding with the KG-1 cells (**Figure 2 B-C**).

All three tools support generating heatmaps of filtered variants which can help visualise variant enrichment in a cell population (**Figure 2 D-F**). Notably, however, Optima and Mosaic do not offer statistical methods for variant enrichment or prioritisation. While several variants appeared to be significantly depleted in these small cell clusters, no obviously enriched variants were observed in the output of these tools.

The ability to prioritise variants based on enrichment allows SCARCE to be much less stringent with its initial filtering of variants, allowing the detection of real somatic variants that may be filtered out by other tools. The heatmap produced by *SCARCE::plot_variants()* (**Figure 2 D**) displays a *p*-value for variant enrichment in each cell population based on a Wald test for the mutant allele odds ratio (OR). Additionally, variants are ordered by these *p*-values for prioritisation and may be filtered to display only significantly enriched variants or using thresholds for OR or Wald Z-score. Using SCARCE, we generated a heatmap of the significantly ($p < 0.05$) enriched variants predicted to be somatic in cluster 0. Notably, none of these 6 variants passed the stringent filtering applied by Optima and Mosaic, and were therefore not included in their heatmaps (**Figure 2 E-F**).

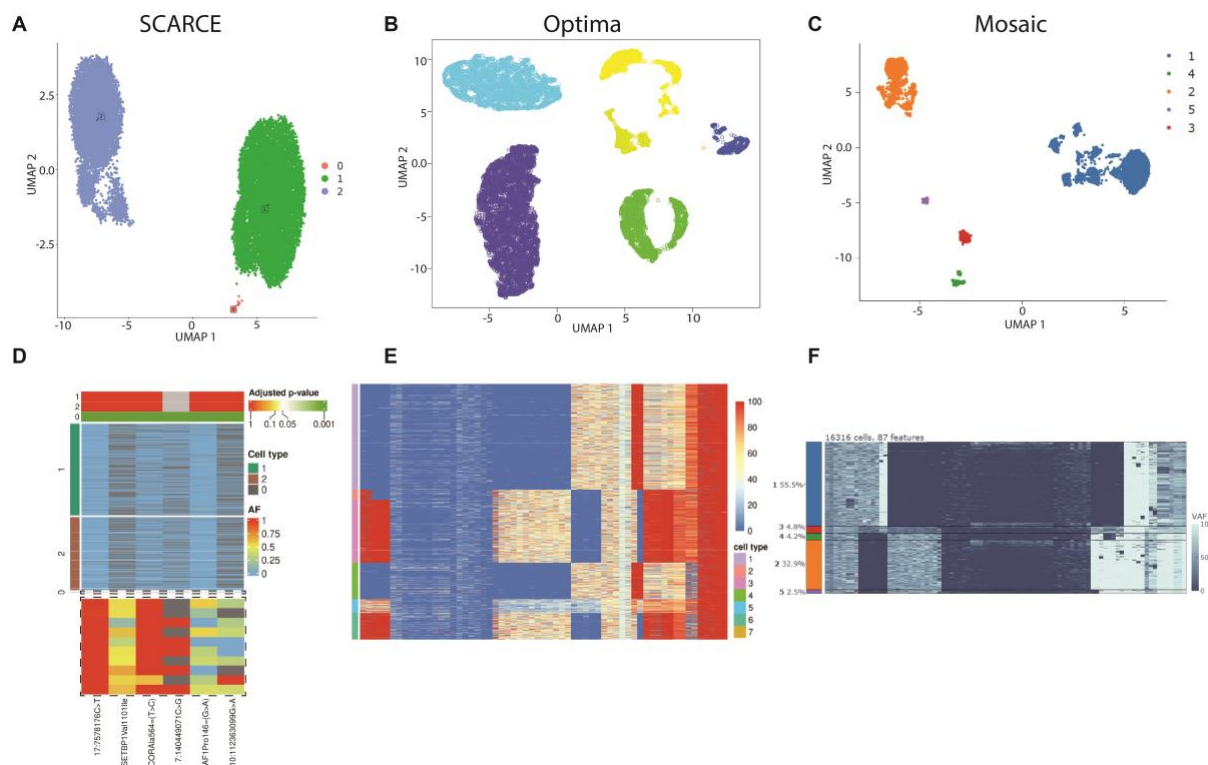


Figure 2: Somatic variant detection comparison of SCARCE with Optima and Mosaic.

Tapestri sequencing data from 16,316 cells were analysed from a sample composed of a 50:50 mix of PBMCs with a 0.05% spike-in of the KG-1 cell line. Variant-based cell clustering was performed by SCARCE (A), Optima (B) and Mosaic (C). Heatmaps were also generated to visualise variant enrichment in the cell type clusters using each tool, respectively (D-F). For SCARCE, variants are ordered by p -value (Wald p , odds ratio (OR)) and filtered to $p < 0.05$, $OR > 5$. An expanded view of the top variants identified by SCARCE showing just the cells for cluster 0 is also shown. For Optima (E) and Mosaic (F), genomic location labels were omitted due to illegibility.

To determine whether the variants identified by SCARCE were true positives, we retrieved a list of known variants in the KG-1 cell line from Depmap (<https://depmap.org/portal/>). Due to the targeted sequencing used in the Tapestri platform, only a single variant overlapped with a known variant in the

cell line: a non-coding splice-donor variant in *TP53* (build hg19, Chr17:7578176C>T). This variant was identified by SCARCE and ranked as the most highly enriched variant. Given that this variant was identified in 10/10 (100%) cells in cluster 0 and in only 30/15,734 (0.2%) of the other clusters, we confirmed cluster 0 as the KG-1 cell line. Additionally, all 6 variants frequently had high genotype quality within cluster 0, and relatively lower quality outside of this cluster, reinforcing that variant calls outside of this cluster were likely false positives (Supplementary Figure S1).

Additionally, SCARCE offers several additional visualisations to help users identify somatic variants. **Supplementary Figure S2 A** shows a bar plot of mutant cell proportions per cell-type with significant enrichment highlighted. The distribution of called genotypes or allele frequencies can also be visualised on a UMAP (**Supplementary Figure S2 B**). Finally, SCARCE can produce two plots that assist in the identification of variants shared between multiple cells, an UpSet plot (**Supplementary Figure S2 C**) and a Sankey plot (**Supplementary Figure S2 D**). These can be particularly useful for identifying clonal driver mutations and clonal lineages.

To assess whether this sensitivity remains in real data, we applied SCARCE to Tapestri data from a patient (PR02-4) with pancreatic ductal adenocarcinoma (PDAC) available from a previous publication (Zhang et al. 2023). The original paper identified three potential somatic variants (*KRAS*^{Gly12Asp}, *KRAS*^{Gly12Val}, *TP53*^{Cys242Tyr}) using bulk whole-exome sequencing (WES) followed by Tapestri sequencing. Using SCARCE, two major cell populations were identified containing 3400 and 209 cells, respectively (**Figure 3 A**). While all three variants were found in the data, *KRAS*^{Gly12Asp} and *TP53*^{Cys242Tyr} were the only significantly enriched variants identified (**Figure 3 B**).

To test whether the clonal population would still be identified and whether the variants would remain significantly enriched in smaller clonal populations, we randomly subsampled cells from the smaller clonal population to 100, 50, 25 and 10 cells, repeating this over 10 iterations while rerunning SCARCE (Figure 3C). In all cases, except for iteration 8 of the 10 cell subsampling, the mutant cell population was identified and both variants remained significantly enriched (**Figure 3 D**).

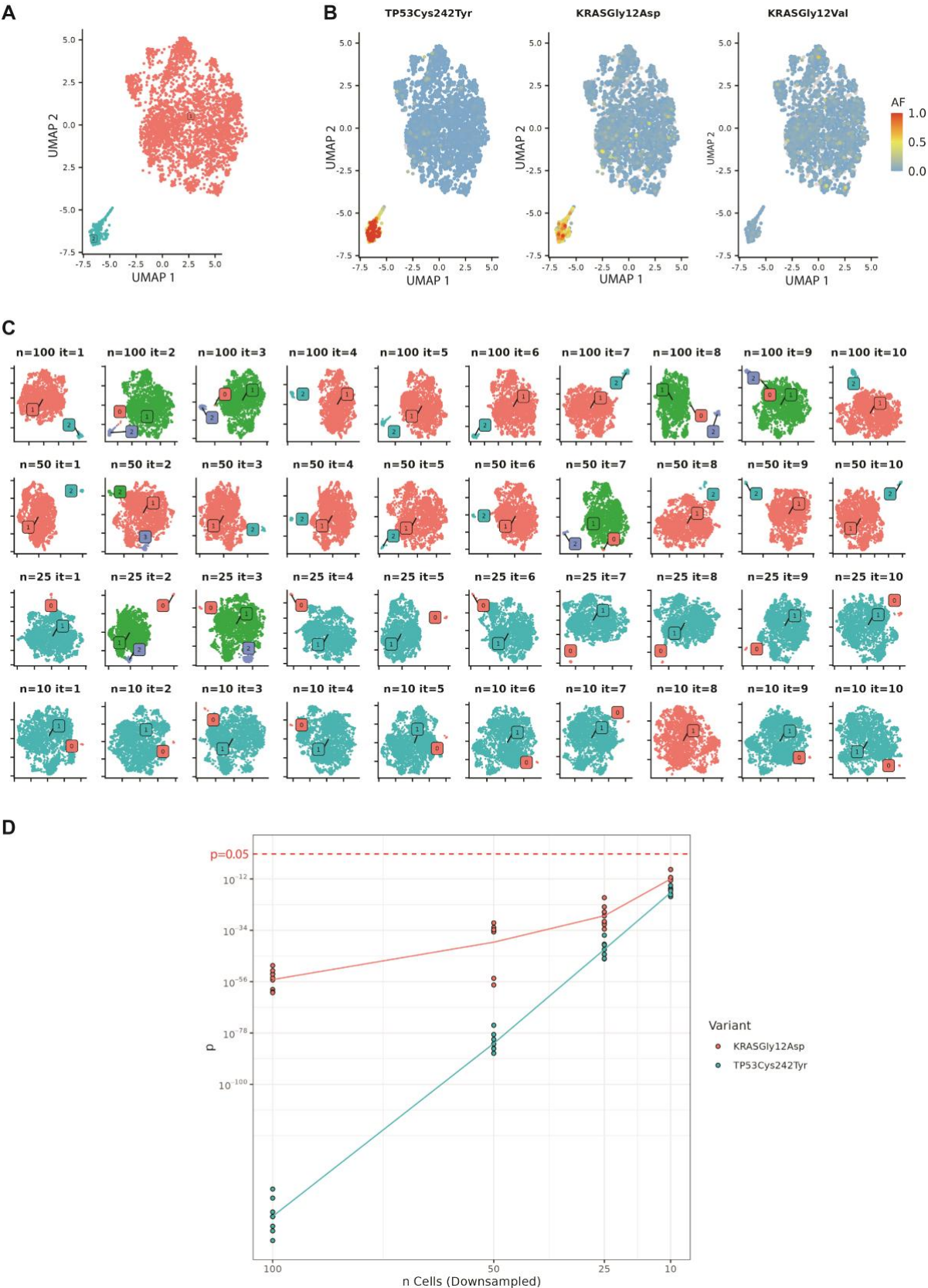


Figure 3: Somatic variant enrichment in pancreatic ductal adenocarcinoma (PDAC) using SCARCE. Tapestry scDNA-seq data from patient PR02-4 (Zhang et al. 2023) was

processed through SCARCE using default parameters. Two cell populations were identified using variant-based clustering (A). *KRAS*^{Gly12Asp} and *TP53*^{Cys242Tyr} were the only significantly ($p < 0.05$) enriched variants identified. *KRAS*^{Gly12Val} is also shown as it was identified in the previous publication (B). Cells from population “2” were randomly subsampled to 100, 50, 25 and 10 cells over 10 iterations (C). Benjamini–Yekutieli adjusted p -values for *KRAS*^{Gly12Asp} and *TP53*^{Cys242Tyr} enrichment were analysed for each iteration and random subsampling (D).

Next, we used SCARCE, Optima, and Mosaic to identify driver somatic variants in lymphocytes from a patient with refractory coeliac disease (RCD) using Tapestry sequencing data for 4,858 cells. Importantly, this sample was previously characterised extensively using multiple sequencing technologies to identify five likely driver mutations: *DDX3X*^{p.M187R}, *DDX3X*^{p.R528C}, *SOCS3*^{p.V84I}, *STAT5B*^{p.N642H}, *STAT3*^{p.E552K} (Singh et al. 2025). All three tools returned four of the five driver mutations in their filtered results except for *SOCS3*^{p.V84I}, which was filtered by all three tools due to low depth reads and poor quality alignment scores. **Figures 4 B-C** show the heatmaps produced by Optima and Mosaic, respectively. In the cell clusters produced, only two of these driver variants show obvious sub-population enrichment, namely *DDX3X*^{p.R528C} and *DDX3X*^{p.M187R}. In contrast, the somatic variant prioritisation performed by SCARCE identifies all four of these variants that passed filtering with high significance, and two additional potential somatic variants are identified: *STAT3*^{Ser614Arg} and *STAT3*^{Thr389Ser} (**Figure 4 A**). Although rarer, these two *STAT3* variants strongly enriched within cluster 2 (*STAT3*^{Ser614Arg} OR=47.5, $p=1.05 \times 10^{-22}$; *STAT3*^{Thr389Ser} OR=34.3, $p=2.53 \times 10^{-13}$) and are annotated as “probably damaging” and “possibly damaging”, respectively by PolyPhen, making them likely true variants. The UMAP visualises the enrichment of these variants, and the Sankey plot (**Figure 4 E**) again shows limited co-occurrence, in agreement with the previous findings (Singh et al. 2025).

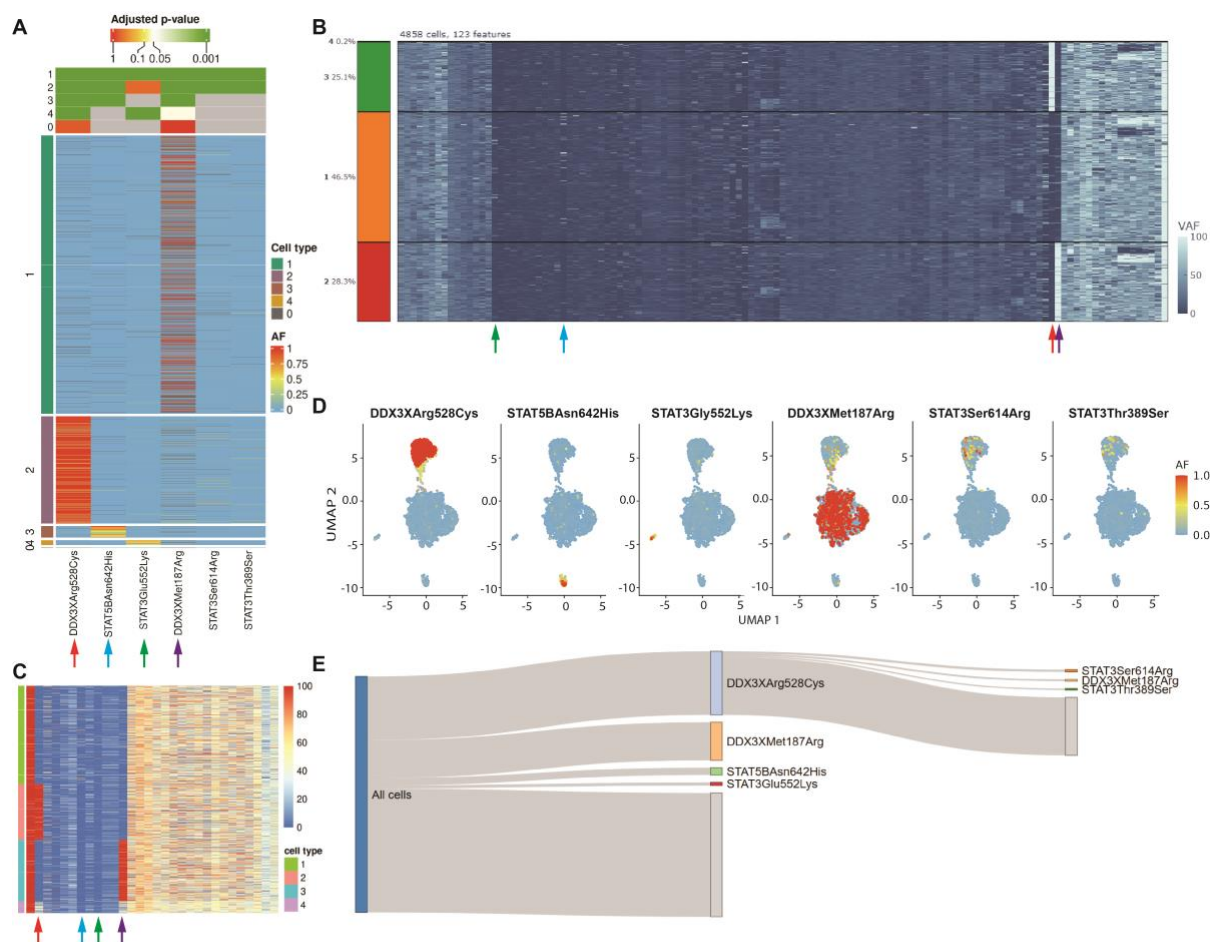


Figure 4: Variant-based clustering and somatic cell type analysis in patient with refractory coeliac disease (RCD) using SCARCE, Optima and Mosaic. A) Significant variants ($p < 0.05$, $OR > 5$) identified using SCARCE. Variant clustering and visualisation was also performed using Optima (B) and Mosaic (C). Genomic location labels were omitted due to illegibility, though common variants amongst technologies are indicated by colour-coded arrows. SCARCE visualisations further show localisation of allele frequency (AF) of enriched variants in cell sub-populations on UMAP (D) and limited co-occurrence (E).

A key feature of SCARCE is the ability to utilise ADT data from Tapestry sequencing to cluster and annotate cell populations, allowing users to identify somatic variant enrichment in a known cell type. SCARCE assists users in cell-type annotation by performing semi-automated annotation using the

SCINA R package (Zhang et al. 2019) and producing plots which help the user decide on a final assignment for each cluster. Examples of these plots are shown in **Supplementary Figure S3**.

SCARCE was next run using ADT clustering for the same RCD patient analysed previously (**Figure 5**). Importantly, this workflow identifies the same known driver mutations as before, but reveals their cell-type enrichment. For example, both *DDX3X* driver mutations were found to be highly enriched in CD4-CD8- T cells, while *STAT5B*^{p.N642H} and *STAT3*^{p.E552K} were enriched in CD8+ T cells. These findings were, again, in agreement with the previous publication (Singh et al. 2025).

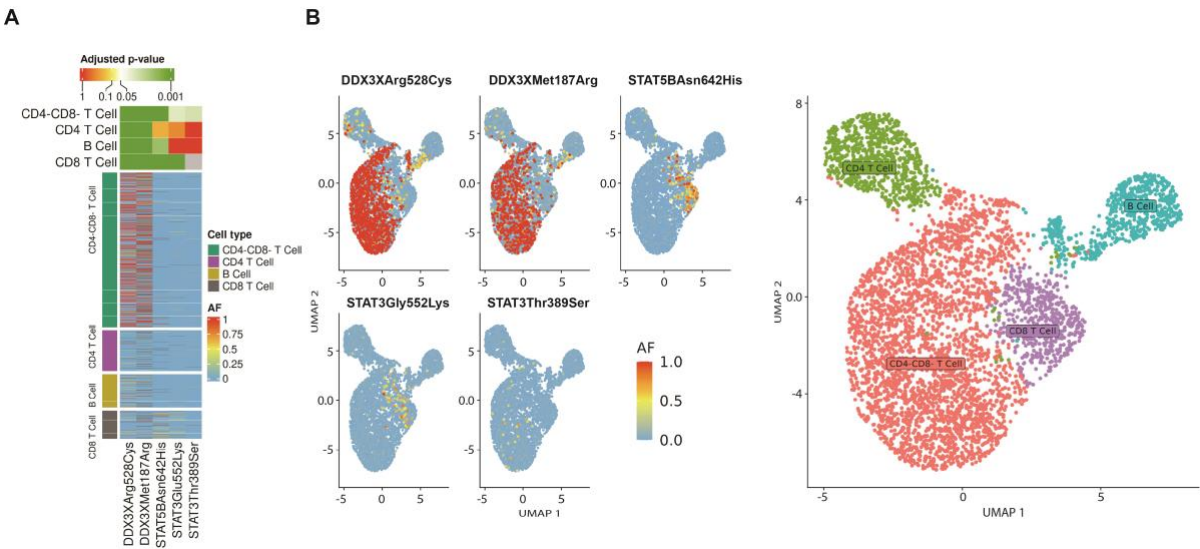


Figure 5: SCARCE variant enrichment analysis using ADT surface marker-based clustering. A) Heatmap of significant ($p < 0.05$, OR > 5) somatic variants identified by SCARCE. B) UMAP plot indicating allele frequency (AF) and localisation of significant variants.

Finally, while SCARCE was designed to utilise the powerful multiomic data capabilities offered by Tapestry or similar technologies, it can support any single-cell DNA sequencing technology that generates a multi-sample VCF where each ‘sample’ references a clone. To demonstrate this functionality, we used SCARCE to identify enriched variants in a patient with cryoglobulinaemic vasculitis who underwent clonally-amplified WGS performed on two aberrant cell populations (“blue branch” $n=16$, “orange branch” $n=16$, control $n=2$) (Singh et al. 2020a). Notably, these branches are

known to differ by the presence of a *KLHL6*^{Leu90Pro} variant, present only in the “blue branch” cells. Due to the vast amount of data available in WGS as opposed to the targeted sequencing performed by Tapestry, we limited this analysis to only Chromosome 3 for this demonstration. **Figures 6 A-B** show the results of this analysis. As expected, due to multiple hypothesis corrections across such a large number of variants and a small sample size in this study, no variants reached statistical significance. However, in such cases, users can instead focus on effect size. To do so, SCARCE offers the ability to filter results by Wald Z-score and OR instead, with the heatmap customisable to display these features. Doing so identifies *KLHL6*^{Leu90Pro} among the top priority variants, along with several other previously unreported variants in this sample. Notably, the Sankey plot (**Figure 6C**) also highlights co-occurrence of many variants.

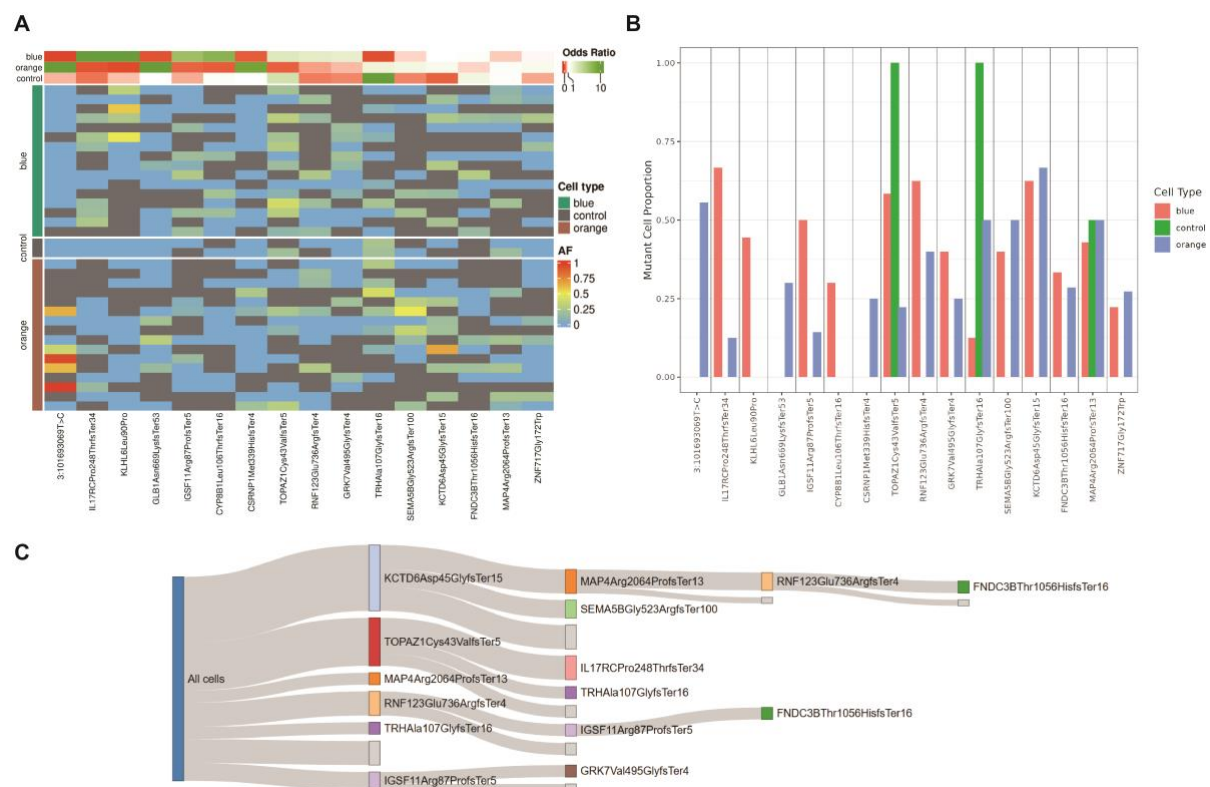


Figure 6: SCARCE somatic variant prioritisation in clonally-amplified WGS sequencing. Data consists of WGS performed on 34 cells from known clonal branches (blue n=16, orange n=16, control n=2). This analysis was restricted to Chromosome 3 for demonstration purposes. SCARCE was used to generate a heatmap (A) of the top priority variants ordered by OR, as

well as a bar plot showing mutant allele proportions per cell annotation (B) and a Sankey plot showing co-occurrence of variants (minimum set size =2) (C).

Discussion

Here we present a simple and intuitive R package, SCARCE, to statistically prioritise and visualise somatic variants enriched in subpopulations detectable in scDNA-Seq data. Somatic mutations occurring in small clonally expanded cell populations are now understood to be key drivers in both cancer and autoimmune disease (Singh et al. 2020a; Singh et al. 2025), and can be extremely difficult to detect using traditional bioinformatics pipelines, reinforcing the need for new tools to be developed.

Recent years have seen rapid expansion of scDNA-seq library preparation technologies. For example, SCARCE works natively with HDF5 data from Tapestry, a technology which uses a microfluidic approach to capture single cells and perform targeted DNA sequencing of thousands of cells paired with ADT assays for surface marker expression (Pellegrino et al. 2018). However, SCARCE can be applied to any technology that produces multi-sample VCF-style files with one ‘sample’ column per cell. In this paper we additionally analysed clonally amplified data from isolated single cells using multiple displacement amplification (MDA), a precursor to NANOSeq, used by BioSkryb’s ResolveOME. ResolveOME employs whole-genome primary-template directed amplification (PTA) and sequencing with paired transcriptome sequencing of up to 384 isolated clones (Marks et al. 2023). This approach, therefore, sacrifices cell throughput but offers much greater sequencing capacity. Finally, we have tested on G&Tseq, an earlier technology which also offers paired genomic and transcriptomic sequencing of isolated single cells using MDA (Macaulay et al. 2015). Although these methods lack Tapestry’s ADT-based surface marker annotations, this can be recovered through flow cytometry-based isolation and annotation of individual cells prior to library preparation.

Despite the increasing number of single-cell DNA-based technologies, pipelines for somatic variant calling remain tailored to bulk sequencing approaches. Common tools like Mutect2 (McKenna et al.

2010), Strelka2 (Kim et al. 2018), and VarScan (Koboldt et al. 2012) can, in theory, be applied to single-cell data by considering each cell as a sample. However, these tools expect input of tumour and normal samples, while for single-cell data it is typically unclear which cells harbour pathogenic somatic mutations. SCARCE instead applies a 1 vs all approach to compare mutant allele counts between a specified cell population and all other cell types iteratively, statistically testing for enrichment in a cell population while prioritising deleterious variants based on Ensembl VEP annotations.

In this publication we compared SCARCE with MissionBio's specific software Optima and Mosaic, which are intended for filtering and visualisation rather than statistical enrichment. A major challenge of single cell DNA sequencing such a large number of cells is the extremely high number of variants that are found, the vast majority of which are likely to be false positives. Because they lack a prioritisation approach, Optima and Mosaic rely on stringent filtering to reduce, in this example, ~100,000 variant calls down to ~20-50 for clustering and visualisation purposes. SCARCE instead applies less-stringent filters to filter out only technical noise, followed by statistical enrichment to focus on probable somatic variants, an approach which reduces the risk of removing true positives. As demonstrated, SCARCE identifies cell-type enriched variants that are filtered out by other tools.

Another key feature of SCARCE is the ability to cluster cells using either ADT- or variant-based information. Comparing the two approaches on the same samples, we found that variant-based clustering is often better at isolating true rogue cell populations, as demonstrated by comparing **Figures 4C and 5C**. While variant-based clustering is also offered by Optima and Mosaic, our use of a greater number of prioritised variant features led to an improved ability to isolate populations, evident in **Figure 2**. On the other hand, ADT-based clustering allows SCARCE to accurately annotate cell types, giving users biological context for enriched somatic variants. For example, **Figure 5B** clearly and intuitively demonstrates biological cell-type specific enrichment of variants.

Single-cell sequencing data can quickly become extremely computationally expensive to analyse, depending on the number of cells analysed. We tracked the peak memory usage by each tool when analysing the two Tapestry datasets presented, and found that peak memory utilisation by SCARCE was

far lower than Optima when analysing large datasets, thanks primarily to our use of chunked loading, though memory use was still much greater than Mosaic (**Supplementary Figure S4**). This means that moderately sized datasets (<5000 cells) can reasonably be run on a local machine without the need for high performance computing infrastructure.

Despite the advantages offered by SCARCE, we note several limitations which we aim to address in the future. For example, in its current implementation, SCARCE does not consider any form of structural variants, including copy number alterations, in its variant prioritisation. Additionally, completely automated cell-type annotation is not offered, requiring users to make informed annotation decisions based on the intuitive visualisations produced by the software. Finally, although SCARCE is designed to be memory efficient by implementing chunked-loading of variant data (which can be user-controlled), the software has been optimised for targeted sequencing panels rather than WGS. This means that analysing WGS data may incur higher than expected memory requirements and runtimes, an issue we plan to address in future releases.

Given the rapid expansion of scDNA-seq technologies, SCARCE's development will continue in the future. In particular, there are several ways that cell clustering could be improved to better isolate cell populations of interest. For example, we have demonstrated that variant-based and ADT-based clustering both offer key advantages and have value independently for grouping cells. However, we expect that a method which combines both may be of significant value, potentially giving both better resolution for clustering and increased biological insight into enrichment results. Furthermore, while some technologies such as G&Tseq and ResolveOME provide paired genomic and transcriptomic data, SCARCE does not currently support cell clustering based on transcriptome data. Finally, in the specific context of autoimmune disease, where immunological clones can be typed based on repertoire sequencing of T cell and B cell receptors, we suspect that the ability to annotate clusters based on repertoire sequencing could offer improved resolution for identifying somatic variants in aberrant, self-reactive clones.

In summary, SCARCE offers users a powerful but simple way to prioritise somatic variants in heterogeneous scDNA-seq data. As scDNA-seq technologies advance, we aim to continue development of SCARCE to assist in the development of diagnostics and treatment options for patients with somatic mutation-driven diseases, including cancer and autoimmune disease.

Methods

Workflow overview:

SCARCE allows two primary workflows and several optional steps dependent on data availability and user-preference (**Figure 1**). The ADT-clustering workflow is designed for experiments that include single-cell variant calls as well as surface marker expression from ADT assays. This workflow begins with semi-automated cell type annotation, which users can skip if utilising their own cell annotations, followed by enrichment of cell type-specific mutations and visualisation. In the variant-only workflow, SCARCE performs cell clustering based on present variants to identify rogue subpopulations of cells. These subpopulations are then annotated and investigated for variant enrichment as per the standard workflow.

Input requirements:

For the ADT-clustering workflow, SCARCE expects surface marker expression in the HDF5 file format, as supplied by Tapestry sequencing, or this can be skipped by providing user-defined cell annotations. Variant data can be provided in the same HDF5 object, or as a multi-sample VCF file with one column per cell. For the variant-only workflow, only the variant information is required, supplied as either an HDF5 object or multi-sample VCF.

Data Preparation:

Surface marker expression data are prepared using the `SCARCE::prepare_annotation_data()` function. ADT data is extracted from the HDF5 object and loaded into a Seurat (v5) (Hao et al. 2024) object using `Seurat::CreateSeuratObject()`. The ADT matrix is normalised with the `Seurat::NormalizeData(normalization.method="CLR")` function and scaled using

Seurat::ScaleData(). Surface markers are filtered to a top user-defined percentile (default = 75%). Finally, PCA and UMAP dimensionality reduction are performed, with user-defined distance metrics (default = “euclidean”), and cell clusters are identified using *Seurat::FindClusters()* at user-defined resolutions (default = 0.2, 0.35, 0.5, 0.8, 1, 1.2).

Custom Cell Annotation:

SCARCE performs semi-automated cell annotation with *SCARCE::annotate_cells()* by first performing an automated per-cell annotation using SCINA (Zhang et al. 2019) and providing a series of plots to help the user ultimately decide on a manual per-cluster annotation (**Supplementary Materials, Supplementary Figure S3**). By default, the CellMarker2.0 (Hu et al. 2023) database provides a list of cell-surface markers associated with each cell-type, but these may be overwritten with user-defined markers. SCINA is run for a maximum of 100 iterations and assigns a single cell type and probability score to each cell. Cells with an assignment probability below a user-defined threshold (default 0.5) are annotated as “unknown”. Plots are produced informing the user of the surface marker expression and the SCINA cell type assignment proportions for each cluster. Finally, the user annotates cell clusters using *SCARCE::annotate_clusters()*.

Variant Processing and Filtering:

Variant data are processed and filtered using *SCARCE::find_somatic_variants()*. Variant formatting, regardless of HDF5 or VCF input, is first normalised to a format that supports Ensembl Variant-Effect Predictor (VEP) input, and multi-allelic sites are expanded to become biallelic. Incorrectly reported variants that cannot be normalised are removed.

Variants are further filtered in three stages:

1. Call-level filters are used to filter variant calls for individual cells. In these cases, the numeric genotype (NGT) call is changed to 3.
 - Depth < min_DP_per_call (default = 10)
 - Genotype Quality < min_GQ_per_call (default = 30)

2. Variant-level filters are used to filter variant calls across all cells

- Alternate allele count < min_varcount_total (default = 30)
- Alternate allele proportion < min_varportion_total (default = 0)
- Alternate allele proportion > max_varportion_total (default = 0.7)
- Total number of cells with non-missing data < min_datacount_total (default = 100)
- Total proportion of cells with non-missing data < min_dataportion_total (default = 0.25)
- Minimum mean genotype quality < min_quality_mean_total (default = 0.3)
- Minimum mean allele frequency < min_allelefreq_mean_total (default = 0)
- Minimum mean allele frequency in cells classified as mutant < min_allelefreq_mean_mutants (default = 0.1)

3. Cell type-specific filters are used to filter variants in the cell type-specific output files

- Alternate allele count in cell type < min_varcount_celltype (default = 1)
- Alternate allele proportion in cell type < min_varportion_celltype (default = 0)
- Cell type cells with non-missing data < min_datacount_celltype (default = 1)
- Proportion of total mutated cells belonging to cell type < min_varportion_in_celltype (default = 0)

To help identify candidate driver mutations from the large number of spurious variant calls, variant annotations are generated using VEP (McLaren et al. 2016) including variant types, gene symbols, protein alterations, population allele frequency, and pathogenicity predictions using SIFT (Sim et al. 2012) and PolyPhen (Adzhubei et al. 2010). The `--flag_pick_allele` flag is used to ensure that a single annotation is returned per allele.

Based on VEP-annotations, a `PRIORITY` flag is assigned to variants with a maximum population allele frequency (`--max_af`) less than a threshold (default = 0.02) and their VEP consequence assigned as: `missense_variant`, `stop_gained`, `stop_lost`, `start_lost`, `frameshift_variant`, `inframe_deletion/insertion` or `protein_altering_variant`.

439

440 *Variant-Based Cell Clustering:*

441 Cells may be clustered and a Seurat object generated using the *SCARCE::cluster_variants()* function,
 442 similar to the method employed by Optima (Pei et al. 2023) and Mosaic (Zhang et al. 2023). This
 443 function first searches for the top *n_features* (default = 100) to use for clustering. Variants are first
 444 filtered to those with an assigned genotype in > 80% of cells. The Bernoulli variance (*var_ngt*) of the
 445 population alternate allele frequency (p_{alt}) is calculated as follows:

$$446 \quad p_{alt} = \frac{\text{Total Alternate Allele Count}}{\text{Total Data Count}}$$

$$448 \quad \text{var_ngt} = p_{alt} (1 - p_{alt})$$

447

449 This down-weights both very rare and very high frequency (likely germline) variants. To retain
 450 informative very-rare events, we additionally select a set of rare variants (<2% frequency) with high
 451 evidence. We first filter rare variants to those with mean genotype quality > 30, then take the variants
 452 with the highest evidence score defined as:

$$453 \quad \text{evidence_score} = \ln(\text{Total Alternate Allele Count} + 1) \times \ln(\text{Mean Depth} + 1) \\ 454 \quad \times \frac{\text{Mean Genotype Quality}}{30}$$

455 The final selection of *n_features* receives 90% from the top Bernoulli variance rankings and 10% from
 456 the rare variant rankings.

457 Using an allele frequency matrix of the selected variant features (v) for all cells, we generate a cosine
 458 similarity matrix for all cells. For two cells, i and j , the cosine similarity (cos) and distance (D) are
 459 calculated using their allele frequency vectors across the selected variants (x_v) only for the variants in
 460 which an allele frequency is available in both cells ($v \in \Omega_{ij}$):

$$\cos(i, j) = \frac{\sum_{v \in \Omega_{ij}} x_{vi} x_{vj}}{\sqrt{\sum_{v \in \Omega_{ij}} x_{vi}^2} \sqrt{\sum_{v \in \Omega_{ij}} x_{vj}^2}}$$

$$D(i, j) = 1 - \cos(i, j).$$

In cases where there is too small an overlap of data availability between cells ($\Omega_{ij} < 10$), cosine distance is set to 1. This distance matrix is then embedded into 2D space using `uwot::umap(n_neighbors = 30, min_dist = 0.3, metric = "euclidean")` and density-based clustering is performed using DBSCAN.

Mutant Cell Annotation:

`SCARCE::annotate_mutants()` is used to annotate mutant or diseased cell populations. This function can take a vector of variants of interest as input along with a Seurat object, and annotate cells carrying these variants as mutants, either taken as the intersection (those carrying all input variants) or union (those carrying any of the input variants) of cells. After doing so, `SCARCE::find_somatic_variants()` can be used to attempt to identify clonal variants present across this cell population.

Cell Type Variant Enrichment:

`SCARCE::find_somatic_variants(run_cell_type_enrichment=TRUE)` reports several statistics to help identify variants enriched in a cell type as specified by the provided cell type annotations. These annotations may refer to actual cell types as identified by surface protein expression using `SCARCE::annotate_clusters()`, variant clusters identified using `SCARCE::cluster_variants()`, or mutant cells identified using `SCARCE::annotate_mutants()`. Statistics are calculated by comparing each cell type of interest against all others in a 1 vs all approach, which is performed for all cell types (default) or only a user-defined selection.

For a given cell type, C , we collect the cell type-specific barcodes, B_C , and the barcodes for all other cell types $B_{\bar{C}}$, and compute:

- a : The number of cell type cells carrying a mutant allele ($NGT_{B_C} \in \{1, 2\}$)
- b : The number of cell type cells homozygous for the reference allele ($NGT_{B_C} = 0$)

- c : The number of “other” cells carrying a mutant allele ($NGT_{B\bar{c}} \in \{1,2\}$)
- d : The number of “other” cells homozygous for the reference allele ($NGT_{B\bar{c}} = 0$)

This yields a 2×2 confusion matrix per variant:

	Alt	Ref
C	a	b
$\bar{C}$	c	d

If any cell count is equal to 0, a Haldane-Anscombe correction is performed by adding 0.5 to each cell.

The OR comparing C and $\bar{C}$ is:

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

The standard error of the log odds-ratio is calculated as:

$$SE(\log(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

This is then used to calculate a Wald Z-score for the $\log(OR)$:

$$Z = \frac{\log(OR)}{SE(\log(OR))}$$

The Wald Z-score is finally used to calculate a p -value on a standard normal distribution. We use the Wald test because it directly operates on the log odds ratio, which provides both an effect size and an associated standard error, enabling straightforward calculation of confidence intervals and Z-scores. For any case where $\min(a, b, c, d) < 5$, a two-sided Fisher’s exact test is used instead from `stats::fisher.test`.

The false discovery rate is controlled by adjusting p -values of all passing variants (`FILTER=""`) using the Benjamini–Yekutieli procedure (Benjamini and Yekutieli 2001).

Benjamini–Yekutieli procedure requires that null p -values are uniform or super-uniform. To assess this, we used the `SCARCE::generate_qq_plot()` function to generate quantile-quantile (QQ) plots from our results based on presumed null variants as described in the **Supplementary Materials**. The QQ plots

showing a super-uniform p -value distributions corresponding with **Figures 3 and 4** are shown in **Supplementary Figure S5**.

Reporting:

In total, four key report files are generated by the *SCARCE::find_somatic_variants()* function and utilised by other downstream functions.

- Summary file (summary*.tsv): Contains one row per variant with VEP annotations, and variant-level summary information. Separate files are generated for all variants (summary.tsv), pass-only variants (summary_pass_only.tsv), and passing priority variants (summary_pass_only_priority).
- Cell type variant enrichment files (*cell_type*.tsv): Contains all the same information as the respective summary files with additional cell type-specific summary information and statistics for variant enrichment. Separate files are generated for all variants (summary.tsv), pass-only variants (summary_pass_only.tsv), and passing priority variants (summary_pass_only_priority).
- Genotype files: Contains all the same information as the summary.tsv file, with an additional column for each cell encoding per-cell allele frequency (AF), depth (DP), genotype quality (GQ), and numeric genotype (NGT) formatted as AF;DP;GQ;NGT.
- Variant counts per cell type: Contains all the same information as the summary.tsv file, with additional columns for each cell type containing the number of cells carrying a mutant allele and the total number of cells with data available.

Visualisation:

All plots are generated using the *SCARCE::plot_variants()* function. This function generates summary plots for a small number of variants of interest, which can be acquired by either retrieving the top N variants from the variant enrichment statistics, or by searching for a gene or specific variants of interest using the *search* and *search_col* arguments. The plots generated by this function include:

- Stacked bar plot: Plots the mutant allele proportions for each cell type
- Upset plot: Shows overlap and uniqueness of variants present between different cell types

- Allele frequency heatmap: A heatmap of the mutant allele frequency per cell clustered into their cell types
- UMAP plots: UMAP plots coloured by NGT and allele frequency
- Sankey plot: A Sankey plot showing the subclustering of variants in order from highest frequency to lowest

Datasets:

Tapestri sequencing data were retrieved from the Mission Bio data portal (portal.missionbio.com) for “patient 1” for AML single-cell measurable residual disease (scMRD) in HDF5 format.

The original data used in this publication, including the Tapestri data and the clonally-amplified WGS data were acquired as previously described (available under ENA: PRJEB31688 and NCBI PRJNA1088187) (Singh et al. 2020a; Singh et al. 2025). These data were processed using hg19 at the time of analysis, and as such the original publications present hg19 coordinate. For this reason, and because all discussed variants are in coding regions that do not vary between the assemblies, we did not convert these coordinates to GRCh38, though SCARCE can support either assembly.

Code Availability:

The SCARCE R package is available as Supplemental Code and from <https://github.com/FieldLabFNQOomics/SCARCE>. In addition, the code to generate the figures in this manuscript are available in the Supplemental Material ‘Source Code’.

Competing Interest Statement:

The authors declare no competing interests.

ACKNOWLEDGMENTS

We thank the National Computational Infrastructure (NCI) for their computational support. The work was supported by National Health and Medical Research Council Fellowship APP5121190.

SD, MS, CG and MAF conceived the project. SD, MAF, RL and FL performed preliminary analysis. RG developed the SCARCE package. All authors contributed to the manuscript, with most writing by RG, SD and MAF.

References

- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nature methods* **7**: 248-249.
- Andreassi MG, Botto N, Colombo MG, Biagini A, Clerico A. 2000. Genetic instability and atherosclerosis: can somatic mutations account for the development of cardiovascular diseases? *Environ Mol Mutagen* **35**: 265-269.
- Andrews TD, Jeelall Y, Talaulikar D, Goodnow CC, Field MA. 2016. DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations. *PeerJ* **4**: e2074.
- Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**: 1165-1188, 1124.
- Dou Y, Gold HD, Luquette LJ, Park PJ. 2018. Detecting Somatic Mutations in Normal Cells. *Trends in genetics : TIG* **34**: 545-557.
- Dukda S, Kumar M, Calcino A, Schmitz U, Field MA. 2025. Increasing pathogenic germline variant diagnosis rates in precision medicine: current best practices and future opportunities. *Human genomics* **19**: 97.
- Field MA. 2020. Detecting pathogenic variants in autoimmune diseases using high-throughput sequencing. *Immunol Cell Biol* doi:10.1111/imcb.12372.
- Field MA. 2022. Bioinformatic Challenges Detecting Genetic Variation in Precision Medicine Programs. *Frontiers in Medicine* **9**.
- Field MA, Burgio G, Chuah A, Al Shekaili J, Hassan B, Al Sukaiti N, Foote SJ, Cook MC, Andrews TD. 2019. Recurrent miscalling of missense variation from short-read genome sequence data. *BMC Genomics* **20**: 546.
- Hamzeh AR, Andrews TD, Field MA. 2021. Detecting Causal Variants in Mendelian Disorders Using Whole-Genome Sequencing. *Methods in molecular biology (Clifton, NJ)* **2243**: 1-25.
- Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C et al. 2024. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology* **42**: 293-304.
- Holzelova E, Vonarbourg C, Stolzenberg MC, Arkwright PD, Selz F, Prieur AM, Blanche S, Bartunkova J, Vilmer E, Fischer A et al. 2004. Autoimmune lymphoproliferative syndrome with somatic Fas mutations. *N Engl J Med* **351**: 1409-1418.
- Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, Chen J, Jiang W, Yang K, Ou Q et al. 2023. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic acids research* **51**: D870-D876.
- Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Kallberg M, Chen X, Kim Y, Beyter D, Krusche P et al. 2018. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**: 591-594.

- Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**: 568-576.
- Lahnemann D, Koster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A et al. 2020. Eleven grand challenges in single-cell data science. *Genome biology* **21**: 31.
- Libby P, Sidlow R, Lin AE, Gupta D, Jones LW, Moslehi J, Zeiher A, Jaiswal S, Schulz C, Blankstein R et al. 2019. Clonal Hematopoiesis: Crossroads of Aging, Cardiovascular Disease, and Cancer: JACC Review Topic of the Week. *J Am Coll Cardiol* **74**: 567-577.
- Lim JS, Kim WI, Kang HC, Kim SH, Park AH, Park EK, Cho YW, Kim S, Kim HM, Kim JA et al. 2015. Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. *Nat Med* **21**: 395-400.
- Lischetti U, Tastanova A, Singer F, Grob L, Carrara M, Cheng PF, Martinez Gomez JM, Sella F, Haunerding V, Beisel C et al. 2023. Dynamic thresholding and tissue dissociation optimization for CITE-seq identifies differential surface protein abundance in metastatic melanoma. *Commun Biol* **6**: 830.
- Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM et al. 2015. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature methods* **12**: 519-522.
- Marks JR, Zawistowski JS, Salas-González I, Tate TA, Morozova TV, Blackinton JG, Arvapalli DM, Velivela SD, Harton GL, Gawad C et al. 2023. Unifying comprehensive genomics and transcriptomics in individual cells to illuminate oncogenic and drug resistance mechanisms. *bioRxiv* doi:10.1101/2022.04.29.489440: 2022.2004.2029.489440.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**: 1297-1303.
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome biology* **17**: 122.
- McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)* **26**: 2069-2070.
- Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, Corbett RD, Johnson NA, Severson TM, Chiu R, Field M et al. 2011. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**: 298-303.
- Pei D, Griffard R, Yellapu NK, Nissen E, Koestler DC. 2023. optima: an open-source R package for the Tapestry platform for integrative single cell multiomics data analysis. *Bioinformatics* **39**.
- Pellegrino M, Sciambi A, Treusch S, Durruthy-Durruthy R, Gokhale K, Jacob J, Chen TX, Geis JA, Oldham W, Matthews J et al. 2018. High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. *Genome Res* **28**: 1345-1352.
- R Core Team. 2026. R: A Language and Environment for Statistical Computing. doi:10.32614/R.manuals.

- Sater V, Viailly PJ, Lecroq T, Prieur-Gaston E, Bohers E, Viennot M, Ruminy P, Dauchel H, Vera P, Jardin F. 2020. UMI-VarCal: a new UMI-based variant caller that efficiently improves low-frequency variant detection in paired-end sequencing NGS libraries. *Bioinformatics* **36**: 2718-2724.
- Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012. Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 14508-14513.
- Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research* **40**: W452-457.
- Singh M, Jackson KJL, Wang JJ, Schofield P, Field MA, Koppstein D, Peters TJ, Burnett DL, Rizzetto S, Nevoltris D et al. 2020a. Lymphoma Driver Mutations in the Pathogenic Evolution of an Iconic Human Autoantibody. *Cell* **180**: 878-894 e819.
- Singh M, Jackson KJL, Wang JJ, Schofield P, Field MA, Koppstein D, Peters TJ, Burnett DL, Rizzetto S, Nevoltris D et al. 2020b. Lymphoma Driver Mutations in the Pathogenic Evolution of an Iconic Human Autoantibody. *Cell* doi:10.1016/j.cell.2020.01.029.
- Singh M, Louie RHY, Samir J, Field MA, Milthorpe C, Adikari T, Mackie J, Roper E, Faulks M, Jackson KJL et al. 2025. Expanded T cell clones with lymphoma driver somatic mutations accumulate in refractory celiac disease. *Sci Transl Med* **17**: eadp6812.
- Solis-Moruno M, Battle-Maso L, Bonet N, Arostegui JJ, Casals F. 2023. Somatic genetic variation in healthy tissue and non-cancer diseases. *Eur J Hum Genet* **31**: 48-54.
- Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**: 865-868.
- Wilmott JS, Field MA, Johansson PA, Kakavand H, Shang P, De Paoli-Iseppi R, Vilain RE, Pupo GM, Tembe V, Jakrot V et al. 2015. Tumour procurement, DNA extraction, coverage analysis and optimisation of mutation-detection algorithms for human melanoma genomes. *Pathology* doi:10.1097/PAT.0000000000000324.
- Young C, Singh M, Jackson KJL, Field MA, Peters TJ, Angioletti-Uberti S, Frenkel D, Ravishankar S, Gupta M, Wang JJ et al. 2025. A triad of somatic mutagenesis converges in self-reactive B cells to cause a virus-induced autoimmune disease. *Immunity* doi:10.1016/j.immuni.2024.12.011.
- Zhang H, Karnoub ER, Umeda S, Chaligné R, Masilionis I, McIntyre CA, Sashittal P, Hayashi A, Zucker A, Mullen K et al. 2023. Application of high-throughput single-nucleus DNA sequencing in pancreatic cancer. *Nat Commun* **14**: 749.
- Zhang Z, Luo D, Zhong X, Choi JH, Ma Y, Wang S, Mahrt E, Guo W, Stawiski EW, Modrusan Z et al. 2019. SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. *Genes (Basel)* **10**: 531.