SOFTWARE TOOL ARTICLE

scRepertoire: An R-based toolkit for single-cell immune receptor analysis [version 1; peer review: 2 approved with reservations]

Nicholas Borcherding1-4, Nicholas L. Bormann5

1Department of Pathology, University of Iowa, Iowa City, IA, USA
2Medical Scientist Training Program, University of Iowa, Iowa City, IA, USA
3Cancer Biology Graduate Program, University of Iowa, Iowa City, IA, USA
4Holden Comprehensive Cancer Center, University of Iowa, Iowa City, IA, USA
5Department of Psychiatry, University of Iowa, Iowa City, IA, USA

Abstract

Single-cell sequencing is an emerging technology in the field of immunology and oncology that allows researchers to couple RNA quantification and other modalities, like immune cell receptor profiling at the level of an individual cell. A number of workflows and software packages have been created to process and analyze single-cell transcriptomic data. These packages allow users to take the vast dimensionality of the data generated in single-cell-based experiments and distill the data into novel insights. Unlike the transcriptomic field, there is a lack of options for software that allow for single-cell immune receptor profiling. Enabling users to easily combine mRNA and immune profiling, scRepertoire was built to process data derived from 10x Genomics Chromium Immune Profiling for both T-cell receptor (TCR) and immunoglobulin (Ig) enrichment workflows and subsequently interacts with the popular Seurat R package. The scRepertoire R package and processed data are open source and available on GitHub and provides in-depth tutorials on the capability of the package.

Keywords

Single-cell RNA sequencing, immune receptor profiling, R, clonotypic analysis

This article is included in the RPackage gateway.
Introduction
The molecular resolution offered by single-cell sequencing (SCS) technologies has led to extensive investigations in the realms of developmental biology, oncology, and immunology. In terms of the latter field, SCS offers the ability to couple the exploration of transcriptomic heterogeneity in immune cells along a disease process with clonality. A number of methods exist for dimensional reduction of mRNA data, reviewed by Chen et al. that have been implemented into R packages to assist in processing and analysis of SCS experiments. However, a gap exists in the processing of V(D)J sequencing, descriptive statistics, clonal comparisons, and repertoire diversity with the current SCS R packages.

With these limitations in mind, scRepertoire was generated (Figure 1). Built using R, scRepertoire is a toolkit to assist in the analysis of immune profiles for both B and T cells, while interacting with the popular Seurat pipeline. scRepertoire also includes processed single-cell mRNA and V(D)J sequencing data of 12,911 tumor-infiltrating and peripheral-blood T cells derived from three renal clear cell carcinoma patients, which is characterized below to demonstrate the capabilities of the package.

Methods
Operation
System requirements for running scRepertoire include the installation of R v3.5.1 and the Seurat R package (v3.1.2). Utilization of scRepertoire is dependent on the total number of single-cells being processed, with a base estimate of 1 Gb of random-access memory and a modern CPU.

Data
The isolation and processing of the 10x-Genomics-based single-cell mRNA and V(D)J Chromium sequencing data for immune cells has previously been described. In addition, T cells were identified using expression values for canonical T cell markers: CD3D, CD4, CD8A, CD8B1 and previous clustering. T cells were isolated and reclustered using the integration method from the Seurat R package (v3.1.2) with 20 principal components and a resolution of 0.5. All code used to generate the figures appearing in the manuscript is available at https://github.com/ncborcherding/scRepertoire.

Implementation
The scRepertoire was built and tested in R v3.5.1. Analysis for scRepertoire was inspired from the bulk immune pro...

Figure 1. A general workflow for single-cell data analysis involving scRepertoire. The analysis starts with the single-cell immune and mRNA sequencing and Cell Ranger-based alignment with the 10x Genomics pipeline. With the TCR or Ig sequencing, scRepertoire can import the filtered overlapping DNA segments, or contigs. The alignments are filtered by cell type of interest and combined using the individual cell barcodes. Clonotypes can be called using the gene sequence of the immune receptor loci, CDR3 nucleotide sequence or CDR3 amino acid sequence. After clonotype assignment, more extensive clonotypic analysis can be performed at the individual sample level or across all samples. General outputs from scRepertoire can be imported into Seurat objects to visualize clonotype data overlaid onto the cell clustering. Likewise, metadata from the Seurat objects can be imported into scRepertoire to analyze clonotypes by assigned clusters.
Clonotypes can be called using the combination of immune loci genes, a more sensitive approach, or the nucleotide/amino acid sequence of the complementary-determining region 3 (CDR3). In addition to the base functions in R, data processing was performed using the dplyr (v0.8.3) and reshape2 (v1.4.3) R packages. Visualizations are generated using the ggplot2 (v3.2.1) and ggalluvial (v0.11.1) R packages with color pallets derived from the use of colorRamps (v2.3) and RColorBrewer (v1.1.2) R packages. Diversity metrics are calculated using the vega (v2.5-6) R package. Visual outputs of functions are stored as layers of geometric or statistical ggplot layering, allowing users to easily modify presentation.

Results
Clonal analysis
scRepertoire\(^1\) can be used to call clonotypes using the CDR3 amino acid/nucleotide sequences, by gene usage, or by the combination of CDR3 nucleotide sequences and genes. Using the quantContig function, unique clonotypes can be visualized as raw values or scaled to the size of the library for samples or by type (Figure 2A). The total abundance of clonotypes can also be visualized calling abundanceContig (Figure 2B) or relative abundance of clonotypes (Figure 2C). Additionally, the distribution of CDR3 nucleotide or amino acid sequences for clonotypes can be visualized with lengthContig (Figure 2D).

Proportional analysis and diversity measures
More in depth analysis of clonal architecture is available. Within the framework of scRepertoire, analysis of clonal homoeostasis, or the clonal space occupied by clonotypes of specific proportions, can be visualized by clonalHomeostasis function (Figure 3A). Similarly, clonalProportion can be called to look at the proportion of clonal space occupied by specific clonotypes (Figure 3B). Overlap between the samples can be calculated and visualized with clonalOverlap, using either the overlap coefficient or Morisita index methods (Figure 3C). Measured of diversity across samples or groups can be quantified with the clonalDiversity function, demonstrating an overall reduction in clonal diversity in tumor samples (Figure 3D).

Seurat interaction
After the processing and analysis of the TCR repertoire with the base features, the next step is using scRepertoire to interact with the single-cell mRNA data. The expression data for the 12,911 cells built into the package have already been clusters (Figure 4A), with a clear distribution of the clusters into peripheral-blood- versus tumor-predominant (Figure 4B). Using the combineSeurat function in scRepertoire, we can look at the clonotypic frequencies of cells that comprise the UMAP-based clusters (Figure 4C), with notable expansion in the C2, C3, and C6 clusters (Figure 4D). The C7 and C8 clusters also have a relatively high frequency. In addition to clonal distribution, we can also use highlightClonotypes to set specific sequences of clonotypes to be visualized (Figure 4E), with clonotype 1 referring to the amino acid sequence “CAVNGGSGN-LIF_CSAEREDTDTQYF” and clonotype 2 for the amino acid sequence “NA_CATSATLRVV_AEKLFF”. Interesting clonotype 2 is restricted to a subcluster of the C6 cluster (Figure 4E). After combining both the clonotype and expression data, interaction between categories, such as cluster label and clonotype frequency can be visualized with the alluvialGraph function.

Conclusions
scRepertoire\(^1\) is a R-based toolkit for the analysis of single-cell immune receptor profiling. The package is able to take the annotated filtered outputs from the 10x Genomics Cell Ranger

---

**Figure 2.** Basic clonotypic analysis functions in scRepertoire. (A) Scaled unique clonotypes by total number of TCRs sequenced by patient and type of sample (peripheral, P; tumor, T), using the quantContig function. (B) Total abundance of clonotypes by sample and type using the abundanceContig function. (C) Relative abundance of clonotypes using density comparing peripheral blood to tumor samples. (D) CDR3 nucleotide length analysis by sample using the lengthContig function. The bimodal nature of the curve is a function of calling clonotypes for cells with both one and two immune receptors sequenced.
Figure 3. Advanced clonal measures between samples. (A) Clonal homeostatic space representations across all six samples using the gene and CDR3 AA sequence for clonotype calling. (B) Relative proportional space occupied by specific clonotypes across all six samples using the gene and CDR3 AA sequence for clonotype calling. (C) Morisita overlap quantifications for clonotypes across all six samples. (D) Diversity measures based on clonotypes by sample type using Shannon, Inverse Simpson, Chao, and abundance-based coverage estimator (ACE) indices.

Figure 4. Interaction of scRepertoire with the Seurat R package. (A) UMAP projection of the ccRCC T cells (n=12,911) into 12 distinct clusters. (B) UMAP projection with peripheral blood (red) and tumor (blue) populations highlighted and an accompanying relative proportion composition of each cluster, scaled by the total number of peripheral blood and tumor cells, respectively. (C) Using the combineSeurat function places individual cells into groups by the number of clonotypes, which then can be displayed overlaid with the UMAP projection. (D) combineSeurat also calculates the frequency of clonotypes, which can be used to more closely examine cluster composition, such as the indicate boxplot. (E) After combing the clonotype information with the Seurat object, highlightClonotypes can be used to specifically highlight the individual clonotypes of interest using the sequence information. (F) Interaction of clonotypes between multiple categories can be examined using the alluvialGraph function.
platform and provide analysis of a number of modalities, including calling clonotypes, clonal space/homeostasis, clonal diversity, and repertoire overlap between samples. Outputs from scRepertoire can be combined with dimensional reduction strategies for single-cell RNA quantifications, allowing users to analyze mRNA and immune profiles together. Under the creative commons v4.0 license, the scRepertoire package is freely available from the GitHub repository and is extensively annotated to assist in implementation and modification.

Data availability

Source data


Folder ‘Data’ contains all data required to run the vignettes described in the Results. This is also available on GitHub.

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software availability

Source code is available from GitHub: https://github.com/ncborcherding/scRepertoire.

Archived source code at the time of publication: https://doi.org/10.5281/zenodo.3612216.

License: Creative Commons Attribution 4.0 International.

References

This paper presents a new R package for the analysis of T-cell and B-cell clonotypes inferred from V(D)J recombination events. The package is designed for the analysis of data generated using the 10x Genomics V(D)J kit that allows clonotypes and transcriptomes to be co-assayed in the same individual cells.

Overall the package appears to fill an unmet need in the analysis of single-cell V(D)J data, and works nicely with the existing Seurat package for single-cell analysis.

While the focus of the paper is on the software, a more detailed explanation of the biology of V(D)J recombination and how this information is detected in single cells would improve the clarity of the manuscript. A description of the major existing challenges and goals of immune clonotype analysis could also be added.

I do not have extensive expertise in the immunology and so my review focuses on the software implementation. I leave discussion of the value and impact of the software to other reviewers. However I will reiterate comments from other reviewers in saying that there should be a distinction made between T and B cell clonotypes, due to the Ig class switching that occurs in B cells.

The paper also states that “The expression data for the 12,911 cells built into the package”. I believe the clonotype information is included in the package in the data directory, but I don’t think the expression data is included. Including large gene expression matrix within the R package is not advisable anyway, and the authors should make it available through other means (NCBI GEO, for example).

**Code review**

I would encourage the authors to submit the package to CRAN or another R repository. This assist in the distribution of the package, and the process of getting the package accepted on CRAN will likely greatly improve the code quality. At a minimum, the authors should try to get the package to pass R CMD check with no notes, warnings, or errors.
Currently the package includes all rendered figures in the vignettes folder. I would recommend removing these files, and also removing the ‘ggsave’ function calls within the vignette to prevent these files being written while compiling the vignette. This will reduce the size of the package.

Several functions in the package use scoping assignment to assign variables in the global environment. This is considered bad practice and should be avoided in all cases.

Extra files such as .DS_Store should be removed from the git repository and from the package. Use the .gitignore and .Rbuildignore files for this.

Avoid importing code within R functions, for example `require(ggplot2)` calls. Instead, document the dependencies using roxygen2, for example `@importFrom ggplot2 ggplot`.

To access data in a Seurat object, I highly recommend using the functions defined in Seurat for this purpose rather than accessing the slots directly. For example, use obj[[1]] to access metadata rather than obj@meta.data and Idents(obj) rather than obj@active.ident

In general the documentation of functions can be greatly improved. Try to include a text description of each function, a detailed description of the parameters, document the returned values, and include an executable example.

It is generally not advisable to overwrite functions in base R or other packages with variable names, for example the ‘call’ variable in ‘clonalDiversity’ overwrites the base R ‘call’ function.

Replace code like `class(df)[1] == “Seurat”` with `inherits(x = df, what = “Seurat”)`.

In plotting functions such as clonalOverlap, consider returning the ggplot object rather than printing the object. For example, replace `suppressWarnings(print(plot))` with `return(plot)`. This will allow users to modify the plot that is generated.

Some code sections are duplicated, for example L91:104 and L131:144 in seuratFunctions.R. Consider putting duplicated code into functions.

Imported functions should be added to the namespace. Documenting the imports using roxygen2 (as has been done for parameters and exports) will take care of this.

The highlightClonotypes function is a bit redundant with existing functions in Seurat, ie DimPlot function with the cells.highlight parameter.

Is the rationale for developing the new software tool clearly explained?
Partly

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes
Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?  
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?  
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Single-cell genomics, bioinformatics, epigenomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 03 March 2020

https://doi.org/10.5256/f1000research.24415.r59209

© 2020 Pagani M et al. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Lorenzo Drufuca  
School of Medicine and Surgery, Università degli Studi di Milano Bicocca, Milan, Italy

Raoul Jean Pierre Bonnal  
National Institute of Molecular Genetics, Milan, Italy

Massimiliano Pagani  
National Institute of Molecular Genetics, Milan, Italy

The paper describes a new package (scRepertoire) of functions for the analysis of clonality in single cell experiments in the R analytical environment.

While it is true, as it is claimed, that ‘a gap exists in the processing of V(D)J sequencing, descriptive statistics, clonal comparisons, and repertoire diversity with the current SCS R packages’, nonetheless few packages already exist for the analysis of clonality in bulk samples (such as the inspiring package tcR). From the text it is not immediately clear why there is the need for a brand new package rather than a simpler ‘wrapper’ to bridge the gap between single cell based data and existing analytical packages.

Although package functionalities are clearly illustrated in the paper, software documentation could be improved: behaviour of some functions is not immediately clear without studying the code directly (e.g. the way multi-chains clonotypes are treated by combineContigs()).

Similar to its inspiring package (tcR) and leveraging on the vegan package, scRepertoire implements different measures for diversity. Nonetheless, it has been suggested that conventional diversity measures fail to describe properly clonotype distributions and several complementary methods have been proposed
like the Recon package by Kaplinsky\textsuperscript{1} and Arnaout\textsuperscript{1} or Startrac package by Zhang and colleagues\textsuperscript{2}. Interaction with such packages would greatly increase scRepertoire analytical effectiveness.

Similarly, the function abundanceContig(), especially in the unscaled form, is of little use per se. Interestingly, recent methods have been proposed to for the comparative analysis of clone size distributions\textsuperscript{3} and could be easily incorporated into scRepertoire adding considerable power to it (though assessment of required numerosity should be introduced).

The package claims to be designed both for TCR and BCR analysis but definition of clonality in B cells is slightly different than in T cells due to isotype switch and somatic hypermutation phenomena following activation. Therefore clonotype identity between two cells should be defined differently between BCR and TCR analysis.

Concerning overlap, beside the nice representation as a heatmap, it could be useful to have the chance to output the matrix itself rather than the plot only.

10x vdj methods occasionally fails to reconstruct complete clonotypes or it reconstructs putatively aberrant clonotypes (clonotypes with multiple beta chains). Currently scRepertoire does not allow to filter for specific chain compositions but such feature would be worth adding, together with a graphical visualization of relative frequencies of chain composition across clonotypes.

Paired 10x gene expression profile and vdj scoped analysis are not guaranteed to reconstruct the information for the exact same pool of barcodes, thus the combineSeurat() function could be improved by allowing to specify whether an inner or Seurat-sided joining is to be performed and ensuring that the joining is performed correctly.

References

Is the rationale for developing the new software tool clearly explained?
Partly

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes
Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Integrative Biology, Cancer Immunology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com