SOFTWARE TOOL ARTICLE

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

Nicholas Borcherding[1-4], Nicholas L. Bormann[5], Gloria Kraus[6]

[1]Department of Pathology, University of Iowa, Iowa City, IA, USA
[2]Medical Scientist Training Program, University of Iowa, Iowa City, IA, USA
[3]Cancer Biology Graduate Program, University of Iowa, Iowa City, IA, USA
[4]Holden Comprehensive Cancer Center, University of Iowa, Iowa City, IA, USA
[5]Department of Psychiatry, University of Iowa, Iowa ity, IA, USA
[6]Faculty of Medicine, Center for Regenerative Therapies Dresden, Technische Universität Dresden, Dresden, Germany

v2 **First published:** 27 Jan 2020, **9**:47
https://doi.org/10.12688/f1000research.22139.1
Latest published: 15 Jun 2020, **9**:47
https://doi.org/10.12688/f1000research.22139.2

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 a number of popular R packages for single-cell expression, such as Seurat. 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.

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 15 Jun 2020	✓ report	✓ report
version 1 27 Jan 2020	? report	? report

1. **Lorenzo Drufuca**, 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

2. **Tim Stuart**, New York Genome Center, New York City, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Nicholas Borcherding (nicholas-borcherding@uiowa.edu)

Author roles: Borcherding N: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Bormann NL**: Methodology, Software; **Kraus G**: Methodology, Software

Competing interests: No competing interests were disclosed.

Grant information: Funding for this project was provided from National Institute of Health F30 fellowship CA206255.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Borcherding N, Bormann NL and Kraus G. **scRepertoire: An R-based toolkit for single-cell immune receptor analysis [version 2; peer review: 2 approved]** F1000Research 2020, **9**:47 https://doi.org/10.12688/f1000research.22139.2

First published: 27 Jan 2020, **9**:47 https://doi.org/10.12688/f1000research.22139.1

REVISED **Amendments from Version 1**

After receiving the very generous reviews, the new version of the manuscript and software reflects our attempts at improving expanding the usability of scRepertoire as a whole. As Drs. Drufuca, Bonnal, and Pagani suggested we have added scRepertoire interaction with a number of R packages and changed the definition of clonotype for B cells. Per the comments made by Dr. Stuart, we have extensively modified the code and accompanying documentation. In addition, we are in the process of submitting the package to Bioconductor. We have added an author, Gloria Kraus, who assisted in the development of the software after the paper was initially submitted. We want to thank both reviews for the suggestions, as well as a number of users, for immensely improving the scRepertoire package.

Any further responses from the reviewers can be found at the end of the article

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[1]. A number of methods exist for dimensional reduction of mRNA data, reviewed by Chen *et al.*[2] 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[3] 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[4–6], as well as SingleCellExperiment and monocle3 class expression objects. 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 patient, which is characterized below to demonstrate the capabilities of the package.

Methods

Operation

System requirements for running scRepertoire[3] include the installation of R v3.5.1 and the 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.

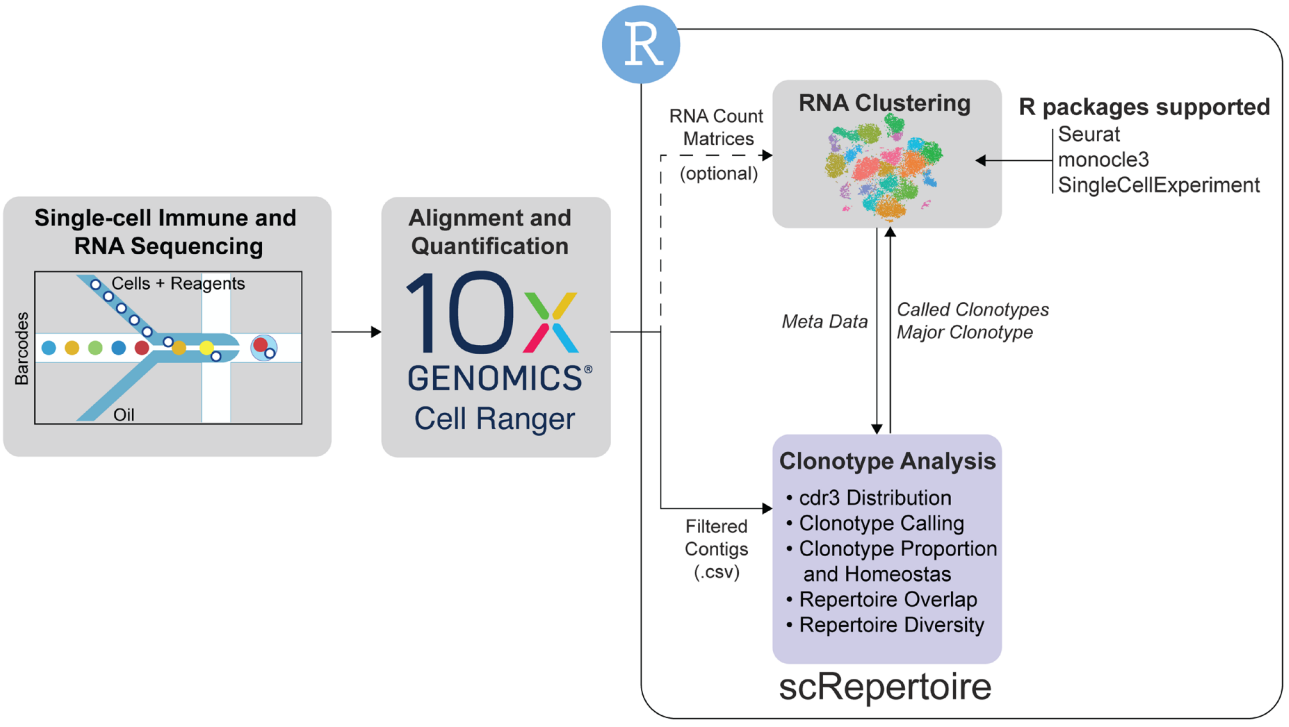

Figure 1. A general workflow for single-cell data analysis involving scRepertoire. The analysis starts with the single-cell immune and mRNA sequencing and Cell Ranger-based alignment with the 10x Genomics pipeline. With the TCR or Ig sequencing, scRepertoire can import the filtered overlapping DNA segments, 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 a number of single-cell expression formats to visualize clonotype data overlaid onto the cell clustering. Likewise, metadata from the expression objects can be imported into scRepertoire to analyze clonotypes by assigned clusters.

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[7,8]. 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[4]. 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 profiling tcR (v2.2.4) R package without derivations in code[9]. 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 vegan (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[3] 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). More advance distribution analysis is also available using the *clonesizeDistribution* function based on recent work using Jensen-Shannon divergence.

Proportional analysis and diversity measures

More in depth analysis of clonal architecture is available. Within the framework of scRepertoire, analysis of clonal homeostasis, 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).

Expression 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 *combineExpression* function in scRepertoire, we can look at the clonotypic frequencies of cells that comprise the

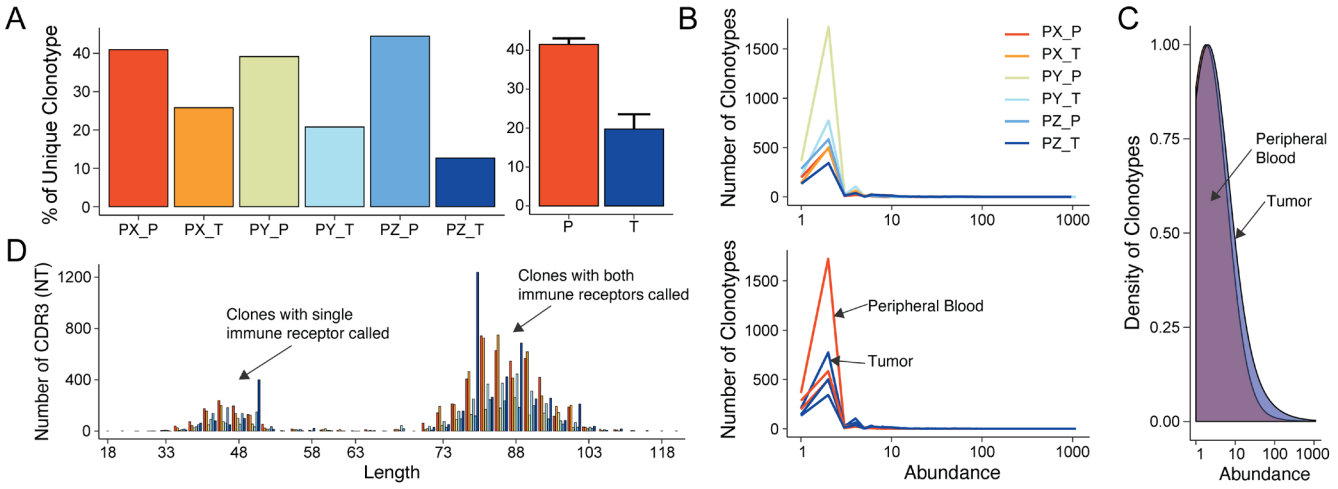

Figure 2. Basic clonotypic analysis functions in scRepertoire. (**A**) Scaled unique clonotypes by total number of TCRs sequenced by patient and type of sample (peripheral, P; tumor, T), using the *quantContig* function. (**B**) Total abundance of clonotypes by sample and type using the *abundanceContig* 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.

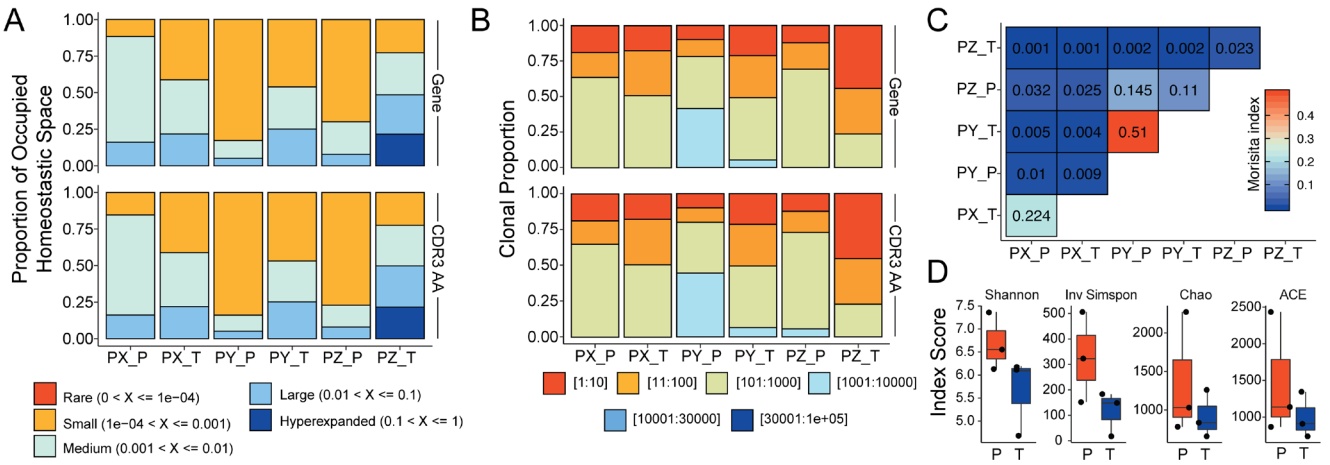

Figure 3. Advanced clonal measures between samples. (**A**) Clonal homeostatic space representations across all six samples using the gene and CDR3 AA sequence for clonotype calling. (**B**) Relative proportional space occupied by specific clonotypes across all six samples using the gene and CDR3 AA 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.

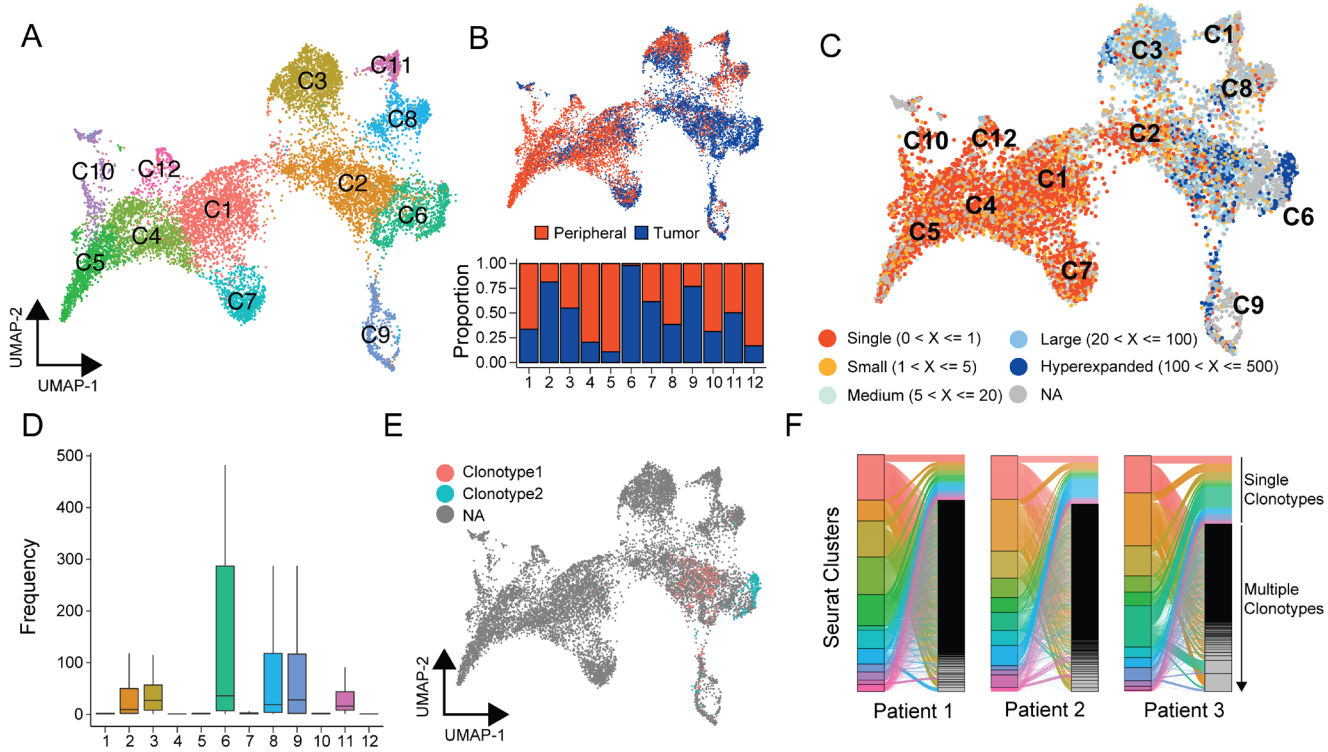

Figure 4. Interaction of scRepertoire with the single-cell expression R packages. (**A**) UMAP projection from Seurat of the ccRCC T cells (n=12,911) into 12 distinct clusters. (**B**) UMAP projection with peripheral blood (red) and tumor (blued) populations highlighted and an accompanying relative proportion composition of each cluster, scaled by the total number of peripheral blood and tumor cells, respectively. (**C**) Using the *combineExpression* function places individual cells into groups by the number of clonotypes, which then can be displayed overlaid with the UMAP projection. (**D**) After combining the clonotype information with the Seurat object, *highlightClonotypes* can be used to specifically highlight the individual clonotypes of interest using the sequence information. (**E**) Interaction of clonotypes between multiple categories can be examined using the *alluvialClonotypes* function.

UMAP-based clusters (Figure 4C). This function also works with the SingleCellExperiment and monocle3 class of expression objects. (Figure 4D). In addition to clonal distribution, we can also use *highlightClonotypes* to set specific sequences of clonotypes to be visualized (Figure 4D), with clonotype 1 referring to the amino acid sequence "CAVNGGSQGNLIF_CSAEREDTDTQYF" and clonotype 2 for the amino acid sequence "NA_CATSATLRVVAEKLFF". Interesting clonotype 2 is restricted to a subcluster of the C6 cluster (Figure 4D). After combining both the clonotype and expression data, interaction between categories, such as cluster label and clonotype frequency can be visualized with the *alluvialClonotypes* function (Figure 4E). This function can also be used to examine the dynamics of single or multiple expanded clonotypes across the categorical variables (Figure 4E). Further, after the attachment of the expression information to a single-cell expression object, the function, *expression2List()* allows users generate analyses based on any categorical variable in the meta data.

Conclusions

scRepertoire[3] 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 platform and provide analysis a number of modalities, including calling clonotypes, clonal space/homeostasis, clonal diversity, and repertoire overlap between samples. Outputs from scRepertoire can combined with dimensional reduction strategies for single-cell RNA quantifications, allowing users to analyze mRNA and immune profiles together. Visualization functions in scRepertoire have a parameter, exportTable, allowing users to examine the quantifications underlying the generation of the graphs. 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

Zenodo: scRepertoire. https://doi.org/10.5281/zenodo.3856827[3].

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

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.3856827[3].

Acknowledgements

We would like to thank Davide Angeletti of the University of Gothenburg and Jae Seung Moon of Stanford University for extensive beta testing and suggestions for improvement.

References

1. Papalexi E, Satija R: **Single-cell RNA sequencing to explore immune cell heterogeneity.** *Nat Rev Immunol.* 2018; **18**(1): 35–45.
PubMed Abstract | Publisher Full Text

2. Chen G, Ning B, Shi T: **Single-Cell RNA-Seq Technologies and Related Computational Data Analysis.** *Front Genet.* 2019; **10**: 317.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Borcherding N, Bormann NL: **scRepertoire (Version 1.2.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.3856827

4. Stuart T, Butler A, Hoffman P, *et al.*: **Comprehensive Integration of Single-Cell Data.** *Cell.* 2019; **177**(7): 1888–1902.e21.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Macosko EZ, Basu A, Satija R, *et al.*: **Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.** *Cell.* 2015; **161**(5): 1202–14.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Butler A, Hoffman P, Smibert P, *et al.*: **Integrating single-cell transcriptomic data across different conditions, technologies, and species.** *Nat Biotechnol.* 2018; **36**(5): 411–420.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Borcherding N, Ahmed KK, Voigt AP, *et al.*: **Transcriptional heterogeneity in cancer-associated regulatory T cells is predictive of survival.** *bioRxiv.* 2018; 478628.
Publisher Full Text

8. Vishwakarma A, Bocherding N, Chimenti MS, *et al.*: **Mapping the Immune Landscape of Clear Cell Renal Cell Carcinoma by Single-Cell RNA-seq.** *bioRxiv.* 2019; 824482.
Publisher Full Text

9. Nazarov VI, Pogorelyy MV, Komech EA, *et al.*: **tcR: an R package for T cell receptor repertoire advanced data analysis.** *BMC Bioinformatics.* 2015; **16**: 175.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 03 August 2020

https://doi.org/10.5256/f1000research.27282.r64820

✔ **Raoul Jean Pierre Bonnal**
National Institute of Molecular Genetics, Milan, Italy
Lorenzo Drufuca
School of Medicine and Surgery, Università degli Studi di Milano Bicocca, Milan, Italy
Massimiliano Pagani
National Institute of Molecular Genetics , Milan, Italy

I can confirm that the authors addressed all the issues and the paper is now suitable for indexing.

Competing Interests: No competing interests were disclosed.

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.

Reviewer Report 09 July 2020

https://doi.org/10.5256/f1000research.27282.r64821

✔ **Tim Stuart**
New York Genome Center, New York City, NY, USA

The authors have now addressed all my original comments.

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.

Version 1

Reviewer Report 09 March 2020

https://doi.org/10.5256/f1000research.24415.r60873

?

Tim Stuart
New York Genome Center, New York City, NY, USA

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[[]] 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.

Author Response 09 Jun 2020

Nicholas Borcherding, University of Iowa, Iowa City, USA

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.

The authors agree that a more extensive explanation of VDJ biology would be a benefit, however, due to the limitations of less than 1000 words and 10 citations, we are limited by the structure of the F1000Research. We have added citations to the readme file to help users.

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).

The seurat object with the package is available in the zenodo repository and linked in the github repository. The raw data has been deposited in GEO and more information can be found in the two preprints cited.

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.

Thank you for the suggestions, the authors believe this point helped the structure of the package immensely. We have used the devtools check function and as of the version in the github repository and zenodo at resubmission, there is no warnings, errors, or notes. We have begun the process of submitting the package to Bioconductor and have passed the build phase of the process.

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.

Thank you for the suggestion, this has been completed.

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.

Scoping assignments in these instances are to give users the options of exporting the tables or matrices that form the visualization. We have modified the activity of the exportTable parameter to return a data frame used if TRUE or visualization if FALSE (default).

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.

Thank you for the suggestion, we have removed .DS_Store.

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

Thank you for the suggestion, we have made this change.

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[[]] to access metadata rather than obj@meta.data and Idents(obj) rather than obj@active.ident

Thanks for the suggestion, we have changed this in the new version of the R package.

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.

Both reviewers have made this excellent suggestion, the new version of scRepertoire has more comprehensive documentation, including descriptions and working examples.

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.

Thank you for the suggestion, we have modified this for all functions.

Replace code like `class(df)[1] == "Seurat"` with `inherits(x = df, what = "Seurat")`
We have made this change and also replace all class() evaluations.

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.

Excellent suggestion and this has been fixed.

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

Thank you for the suggestion, part of the change in the "call" to cloneCall variable in the functions also including creating a function to deal with the user selection. These lines were consistently repeated across multiple functions, but now have been reduced. We have also extensively modularized the code and added these to utils.R .

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.

This has been completed.

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

Agreed, although redundant, this was a request from an early beta tester, we plan to keep the function in.

Competing Interests: None

Reviewer Report 03 March 2020

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

?

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[1] and Arnaout[1] or Startrac package by Zhang and colleagues[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[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

1. Kaplinsky J, Arnaout R: Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nature Communications*. 2016; **7** (1). Publisher Full Text
2. Zhang L, Yu X, Zheng L, Zhang Y, et al.: Lineage tracking reveals dynamic relationships of T cells in colorectal cancer.*Nature*. **564** (7735): 268-272 PubMed Abstract | Publisher Full Text
3. Koch H, Starenki D, Cooper SJ, Myers RM, et al.: powerTCR: A model-based approach to comparative analysis of the clone size distribution of the T cell receptor repertoire.*PLoS Comput Biol*. **14** (11): e1006571 PubMed Abstract | Publisher Full Text

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.

Author Response 09 Jun 2020

Nicholas Borcherding, University of Iowa, Iowa City, USA

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.

Thank you for your input. To clarify, indeed a gap exists. The vast majority of packages available for V(D)J single-cell analysis is python-based reconstructions of TCR/BCR sequences (see, scTCRseq , TRAPeS and TraCeR). The tcR package that the reviewer indicates is in fact deprecated and only functional for bulk sequencing, there is no barcoded-based method. However, the authors of tcR have a newer R package, immunarch, that is freely available and produced by the company immunomind. This newer package offers single-cell support, but the workflow within the package does not allow for the combination of both TCR/BCR chains, and will call clonotypes with a single chain. In addition, as the reviewers keenly mention in the accompanying comments, single-cell technologies have unique pitfalls that need to be addressed as opposed to just creating wrapper functions. These concepts have been added to a newer version of the manuscript.

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()).

We have added extensive documentation to the package.

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 and Arnaout or Startrac package by Zhang and colleagues. Interaction with such packages would greatly increase scRepertoire analytical effectiveness.

We have added the function StartracDiversity() to the development version of the package to allow for the migr, exp, and tans indices to be calculated from the Seurat object after the addition of the clonotype information. This function is detailed in the vignette due to limited word count of this article.

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 and could be easily incorporated into scRepertoire adding considerable power to it (though assessment of required numerosity should be introduced).

We have added the function clonesizeDistribution() to allow users to examine the hierarchical clustering based on clone size distributions of samples to the scRepertoire package along with a comprehensive description and covered the usage in the vignette accompanying the package. Due to the length limitation for the manuscript, we only reference the function in the manuscript with the appropriate citation.

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.

We agree with the need to add thresholding for clonotype assignment for the BCR in the context of somatic hypermutation. We have separated the combineContig() function into combineTCR() and combineBCR() functions, the latter organizes the data similarly, but defines nucleotide sequences by the normalized Hamming Distance, with sequences greater than >0.85 assigned to the same group. In the strict definition of the clonotypes in the combineBCR() the vgene is also added to both the light and heavy chains. We have kept the ability to visualize and analyze clonotypes at the amino acid, nucleotide sequence and gene level as well, to all users more choices.

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.

When appropriate, the visualization functions have the exportTable variable that allows a user to get the table or matrix that enables the visualization. Although this feature is already present, we will add discussion to the manuscript and accompanying vignette.

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.

If the authors understand the suggestion, the reviewers are asking for additional filtering options for combineTCR() and combineBCR(), two separate functions created from the reviewers earlier points. The function has the ability to remove barcodes with NA values or with multiple chains. More specifically to the combineTCR() function we have added removeMulti (removes all barcodes with >= 3 chains) and filterMulti (selects the 2 highest expressing chains). In terms of the chain filtering and visualizations, we are more than happy to add additional filtering and visualization tools for users. If the reviewer could clarify more on the how they envisioned the chain composition visualization, we would appreciate more insight.

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.

We have added the variable NA into the new combineBCR() and combineTCR(), which functions by removing cells without clonotype information. If we understand the reviewer correctly, the differential joining function would also need to address the seurat object as a whole instead of just merging the meta.data.

Competing Interests: None