CellexalVR: A virtual reality platform for the exploration and analysis of single-cell gene expression data

Oscar Legetth[1], Johan Rodhe[1], Joel Pålsson[1], Mattias Wallergård[3], Stefan Lang[1,2], and Shamit Soneji[1,2,*]

[1]Division of Molecular Hematology, BMC, Lund University, Lund Sweden

[2]Lund Stem Cell Center, Lund, Sweden

[3]Department of Design Sciences, Lund university

[*]Corresponding author

Single-cell RNAseq is a powerful tool for the dissection of cell populations at the transcriptome level, and a myriad of techniques are available to project cells on to 3-dimensional space to construct cellular maps that aid the visualisation of heterogeneity and any sub-populations formed. Current visualisation methods for 3-dimensional data on conventional computer displays are poor, and coupled with a lack of intuitive point/cell selection methods often hinders a rapid exploration of finer details contained in the data. Here we present CellexalVR (www.cellexalvr.med.lu.se), a feature-rich, fully interactive, and immersive virtual reality environment for the analysis of single-cell RNAseq experiments that allows researchers to quickly and intuitively gain an understanding of their data.

Single-cell RNAseq (scRNAseq) data can be generated with high-throughput due to the introduction of automated steps regarding library generation such as the Fluidigm C1 and 10X Genomics Chromium platform among others. As a consequence the quantity of scRNAseq data is growing rapidly, and the number of cells per experiment is projected to increase accordingly [8]. The volume of data being generated out-strips the number of bioiformaticians needed to thoroughly analyse it, meaning that some aspects have to be carried out by the ”wet-scientists” performing the experiments. In more general terms, projects such as the Human Cell Atlas [6] will single-cell profile a massive number of cells that will be of general interest to the wider scientific community, and methods will be needed for scientists of varied computational ability to explore it.

The analysis of single-cell RNAseq data is often performed using scripting, primarily using packages for the R/Python languages such as monocle [9] and Seurat [1] among others. A common step in many studies after pre-processing is a dimension-reduction step where cells are arranged in 2/3 dimensional space to visualise

cell-cell relationships and the subgroups they form. A number of methods have been applied/devised such as tSNE [10], diffusion maps [2], ZIFA [5] and pseudo-timing methods such as DDRTree [3] and SPRING [11], all of which result in different projections given they use different underlying methods. The visualisation of 3D projections has thus far been restricted to traditional 2D computer displays using openGL for example. This has several shortcomings, specifically, having limited viewing angles and only being rotatable around the center. The main drawback is the plot is still essentially 2D when viewed on a flat display, therefore, there is no sense of depth and complex projections are harder to comprehend. Also, these plots are not interactive therefore selecting cells for further analyses isn't possible, so to get around this the projections are plotted in 2-dimensions on which the user can gate the cells required, or apply clustering algorithms to capture cell groupings of interest. Clustering is often a supervised process where the user will alter the value of K until a grouping based on what they can see in the data is achieved, leading to what is essentially a time consuming trial-and-error process.

Here we present CellexalVR, a virtual reality (VR) platform developed for use with the HTC Vive that overcomes these blocks. By placing all representations of the data (MDS plots, heatmaps, networks) in VR we have created an immersive environment to explore and analyse scRNAseq data. In VR the MDS plots have visual depth, and can be interacted with intuitively, for example, one can grab each graph and move them to gain any view required as if they were a physical object. If the user defined "play-area" is big enough, one can even walk around them. As multiple MDS plots can be loaded in a single session, they can be cross-compared with ease. For example, cells of interest in a tSNE plot can be traced and connected to their counterparts in a diffusion map allowing the user to visualise directly the differences between the two MDS methods. Another use of this feature would be to determine the effect of preprocessing steps on the outcome when the same MDS method is subsequently applied. CellexalVR also allows sub-populations to be selected directly by passing them through a selection tool from which heatmaps and transcription factor (TF) correlation networks can be generated. These too are also interactive, for example, TF networks can be directly compared to locate common TF-TF pairs.

CellexalVR comprises of two components. The first is the VR interface which has been implemented in Unity (`https://unity3d.com/`), and an R package cellexalvrR (`https://github.com/sonejilab/cellexalvrR`) that performs two functions. The first is to undertake back-end calculations during a CellexalVR session, and second is to provide easy-to-use functions to export scRNAseq data from an R session into a set of input project files that CellexalVR can read. The main advantage of compartmentalising CellexalVR is so bioinformaticians can alter the R package to modify/add methods without needing knowledge of C# which is the language the VR interface is coded in. At a minimum CellexalVR should be provided with the gene expression data (highly variable genes only are recommended), and one set of MDS coordinates. CellexalVR will also import cell surface marker intensities captured during index sorting, and categorical metadata for cells and genes. Detailed documentation and instructional videos are provided on the project website.

Figure 1 shows the current highlights of CellexalVR using data from mouse hematopoietic stem and progenitor cells [4]. The MDS plot(s) are automatically loaded in when a new CellexalVR session is initiated, and when more than one is present they are all loaded simultaneously (Fig 1a). These can be coloured according to the expression of a selected gene (here, Gata1 has been chosen) (Fig 1b) using a keyboard in the virtual environment (FigS1). After the projections have been coloured by a gene, the user has the option to calculate and display the top 10 correlated and anti-correlated genes which is done by clicking the chevron next to the gene name, and the subsequent lists appear to the left (Fig 1b middle). Each of these can be selected to recolour the graph according to the expression of that gene. Cells of interest are captured by passing them through a coloured plane that extends from the action controller (Fig 1c bottom and FigS2) when the selection tool is activate from the menu. As cells are passed through the user experiences haptic feedback as the cells are selected and coloured, and the corresponding cells in the other MDS plots are coloured the same, giving instant feedback on where these cells reside in other projections. A new group is initiated by a left/right-click on the action controller touchpad that brings up a new colour. Once the desired groups have been defined and saved the user can produce a heatmap of differentially expressed genes (Fig 1d). The heatmap can be resized and rearranged, but importantly the order the cells appear in the heatmap is defined by the order in which they are passed through the selection tool, particularly useful when selecting through pseudotime projections as the heatmap will preserve this pseudotime information. Another option is to construct partial correlation networks of transcription factors (Fig 1e). These networks can be viewed in 2D/3D, and can be cross-compared to see which TF-TF pairs are in common between the different networks generated from each population. Clicking on a gene name will recolour the MDS plots by the expression of the gene selected. A third option is to trace the selected cells to their counterparts in other MDS plots. Fig 1f shows a group of cells that have been selected in the left DDRTree projection and then traced to the tSNE plot on the right. These seemingly similar cells in fact split into two further groups when both projections are considered together, and these two groups can be captured by passing the selection tool over cubes placed on the connecting lines. All of these functions are triggered using one touch operations on the controller mounted menu system (Fig 1a, bottom left and FigS3). Figures generated during a session can be exported as images, and selections made during the session are saved in the R object (see methods) as they are created.

CellexalVR will also import and display metadata for cells and genes. For example, cells can be coloured by their type (FigS3) and any other characteristics which the user has assigned, for example, cell-cycle phase. Index sorting data is another important source of information that CellexalVR can visualise, therefore cells can be coloured according to the expression of cell surface markers used during sorting (FigS4). As the input matrices are generic it means CellexalVR will also take data from CITEseq [7]. In general, as pre-processing of data is done prior to import, CellexalVR will take data from any scRNAseq technology. Currently the limit for the number of cells that can be displayed in total across all MDS plots is approximately 15,000. Beyond that users may experience lag, but efforts are under way to increase this number.

In order to expedite the learning process we have extensive documentation and video tutorials on the project website, and our test users became proficient in around 30 minutes regardless of age, video game or VR experience. While CellexalVR does not remove the need for a bioinformatican, it allows non-bioinformaticians to interact and analyse single-cell expression data in a fast and intuitive manner previously not possible. With the amount of scRNAseq data set to increase rapidly alternative methods are going to be required to navigate it, and we have shown that VR is an attractive solution.

Availability and system requirements.

We are currently inviting individuals to test the current version of CellexalVR by contacting us via the form at `https://www.cellexalvr.med.lu.se/download.html`. Users will need a gaming-class computer with a high-end graphics card (for example an NVIDIA GTX1080) running Windows 10, and an HTC Vive.

Acknowledgements.

We thank Rasmus Olofzon, Kristian Berg, Daniel Hellstrom, Daniel Cheveyo, Arvid Carlman, and Christopher Nilsson from the LTH, Lund University for their work on the prototype. Steve Taylor at the CBRG, Oxford University and members of the Lund Stem Cell Centre for testing and feedback.

Funding.

O.L, J.R and J.P are funded by the Knut and Alice Wallenberg Foundation and Hemato-Linnè (Vetenskapsrådet), M.W is funded by H2020, eSSENCE, and the Pufendorf Inst, Lund University. S.L and S.S are funded by StemTherapy which is funded by the Swedish Government.

Author contributions.

O.L, J.R, and J.P developed CellexalVR and the project website. S.L and S.S developed cellexalvrR. M.W provided technical assistance. S.S conceived the study wrote the paper.

References

[1] A. Butler and R. Satija. Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. *bioRxiv*, page 164889, 2017.

[2] L. Haghverdi, F. Buettner, and F. J. Theis. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998, 2015.

[3] Q. Mao, L. Wang, S. Goodison, and Y. Sun. Dimensionality Reduction Via Graph Structure Learning. In *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[4] S. Nestorowa, F. K. Hamey, B. Pijuan Sala, E. Diamanti, M. Shepherd, E. Laurenti, N. K. Wilson, D. G. Kent, and B. Göttgens. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, 128(8):e20–e31, 2016.

[5] E. Pierson and C. Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1), 2015.

[6] A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Göttgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundeberg, P. Majumder, J. C. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Phillipakis, C. P. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T. N. Schumacher, A. Shalek, E. Shapiro, P. Sharma, J. W. Shin, O. Stegle, M. Stratton, M. J. Stubbington, F. J. Theis, M. Uhlen, A. Van Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, and N. Yosef. The human cell atlas. *eLife*, 6, 2017.

[7] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017.

[8] V. Svensson, R. Vento-Tormo, and S. A. Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604, 2018.

[9] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, 2014.

[10] L. Van Der Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[11] C. Weinreb, S. Wolock, and A. M. Klein. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, 2017.

Methods

CellexalVR is built using Unity 3D (https://unity3d.com/), an engine focusing mainly on video games. In CellexalVR, Unity handles many things such as rendering the frames that are displayed on the computer's monitor and in the user's headset, collecting input from the keyboard and the controllers, forwarding events that trigger certain actions and handling all physics simulation. Unity comes with an editor which is the primary development environment that CellexalVR was created within. CellexalVR uses SteamVR (https://steamcommunity.com/steamvr) and OpenVR (https://github.com/ValveSoftware/openvr) for communication with the hardware and VRTK (https://github.com/thestonefox/VRTK) for basic interaction logic.

Inside the Unity engine, everything the user can see is represented with what Unity calls gameobjects. Each gameobject can have a number of components attached to it. For example, gameobjects may use the rigidbody

component to interact with the physics simulation, or custom written scripts that define different behaviours. These components are easily accessible from a programming point of view, and come with a large array of customisable options to fine-tune their behaviour. Gameobjects may also have one or more child gameobjects that will follow the parent if it moves allowing for hierarchies of gameobjects to be defined. For example, when a dataset is loaded in CellexalVR, 3D projections of the data is shown by displaying a small sphere for each cell. Each sphere is its own gameobject that holds a graphpoint script. The graphpoint script contain some information about the point in relation to the 3D projection it is part of. All graphpoint gameobjects have one of the graph gameobjects as their parent. The graph gameobjects have rigidbody components attached to them, allowing them to be moved around, and all graphpoints that are part of the graph will move as well.

File preparation and back-end calculations are implemented in R using the cellexalvrR package (https://github.com/sonejilab/cellexalvrR). CellexalVR is provided as a Windows executable, and the source will be available at https://github.com/sonejilab/cellexalvr.

Data formats. CellexalVR requires multiple, correctly formatted files. This process is simplified by using the R package cellexalvrR that will generate the input files, R object, and SQLite database required. At a minimum CellexalVR should be provided with:

- ***A matrix of gene expression data (C cells x G genes)***. This is processed by cellexalvrR to an SQLite3 database that CellexalVR queries when needed.
- ***At least one set of MDS coordinates placing the cells in 3D space (C cells x 3)***. This can be from any MDS methods the user deems suitable and CellexalVR will accept more than one MDS table. These are exported as 3 column text files (*.mds) with the Cell ID in the first column. cellexalvrR will also calculate coarse convex hulls using the *ashape* R package for each MDS projection and write that to tab-delimited text files (*.hull).

In addition to these, further optional files can be imported. These are:

- ***Surface marker intensities (C cells x S surface markers)***. These are recorded when cells have been index sorted. If using CITEseq, these expression values go into this table.
- ***Cell type information (C cells x T types)***. These allow the user to label each cell as being of a certain type, which can then be displayed in the CellexalVR session. Cells are marked as belonging to a class with a "1", or "0" otherwise.
- ***Metadata for cells (C cells x M meta)***. Further labels for cells, for example cell-cycle stage.
- ***Metadata for genes (G genes x M meta)***. For example, marking genes if they belong to a particular category such as transcription factors, epigenetic factors, or code surface proteins.

To export the necessary file from R, a cellexalvr S4 object needs to be created first using the supplied functions . A typical export process would look as follows:

```
library(cellexalvrR)
#The following lines load the example data from Nesterowa et al (see references).
#Data can be downloaded from cellexalvr.med.lu.se
load("log2data.RData") #expression data (matrix)
load("facs.RData") #surface marker expression (matrix)
load("cell.ids.RData") #cell IDs (matrix)
load("diff.proj.RData") #diffusion map projection coordinates (matrix)
load("ddr.proj.RData") #DDRTree projection coordinates (matrix)
load("tsne.proj.RData") #tSNE projection coordinates (matrix)

log2data[1:10,1:10] #displays the first 10 rows and columns of the expression matrix

#The next 4 lines show how the matricies should look
head(facs)
head(cell.ids)

head(diff.proj)
head(ddr.proj)

#The 3 sets of MDS coordinates are put into a single list
proj.list <- list(diffusion=diff.proj,DDRtree=ddr.proj,tSNE=tsne.proj)
names(proj.list)

#Create a cellexalvr object setting the specie to mouse
cellvr <- MakeCellexaVRObj(log2data,mds.list=proj.list,specie="mouse",
cell.meta=cell.ids,facs.data=facs)

#Output the files to a selected folder.
export2cellexalvr(cellvr,"CellexalOut/")
```

For the 10X Genomics Chromium platform we have functions to convert a Seurat object to a cellexalvr object prior to export, and other functions to help data formatting. See the package manual for further details.

In-session calculations are also performed by cellexalvrR. Genes (anti-)correlated to a gene of interest are calculated using a Spearman rank coefficient. For heatmap generation the user has the option of defining

differentially expressed genes between selected groups using one of i) ANOVA of a fitted linear model, ii) EdgeR, and iii) MAST. Clustering is performed using hierarchical clustering for the top 250 differentially expressed genes as the default, but this is user definable in the configuration file (Fig S5). TF-TF partial correlation networks from within a selected group are calculated using the GeneNet package using a ggmcutoff =0.8. All heatmaps and networks are rendered in the CellexalVR UI. TFs are defined as those in the AnimalTFDB database (urlhttp://bioinfo.life.hust.edu.cn/AnimalTFDB/).

Hardware. CellexalVR was developed for HTC Vive on a gaming class workstation comprising an Intel i7 processor, 16Gb RAM, 1Tb SSD, and an NVIDIA GTX1080 graphics card. An HTC Vive Pro is recommended since the increased resolution greatly enhances the experience, and reading becomes easier.

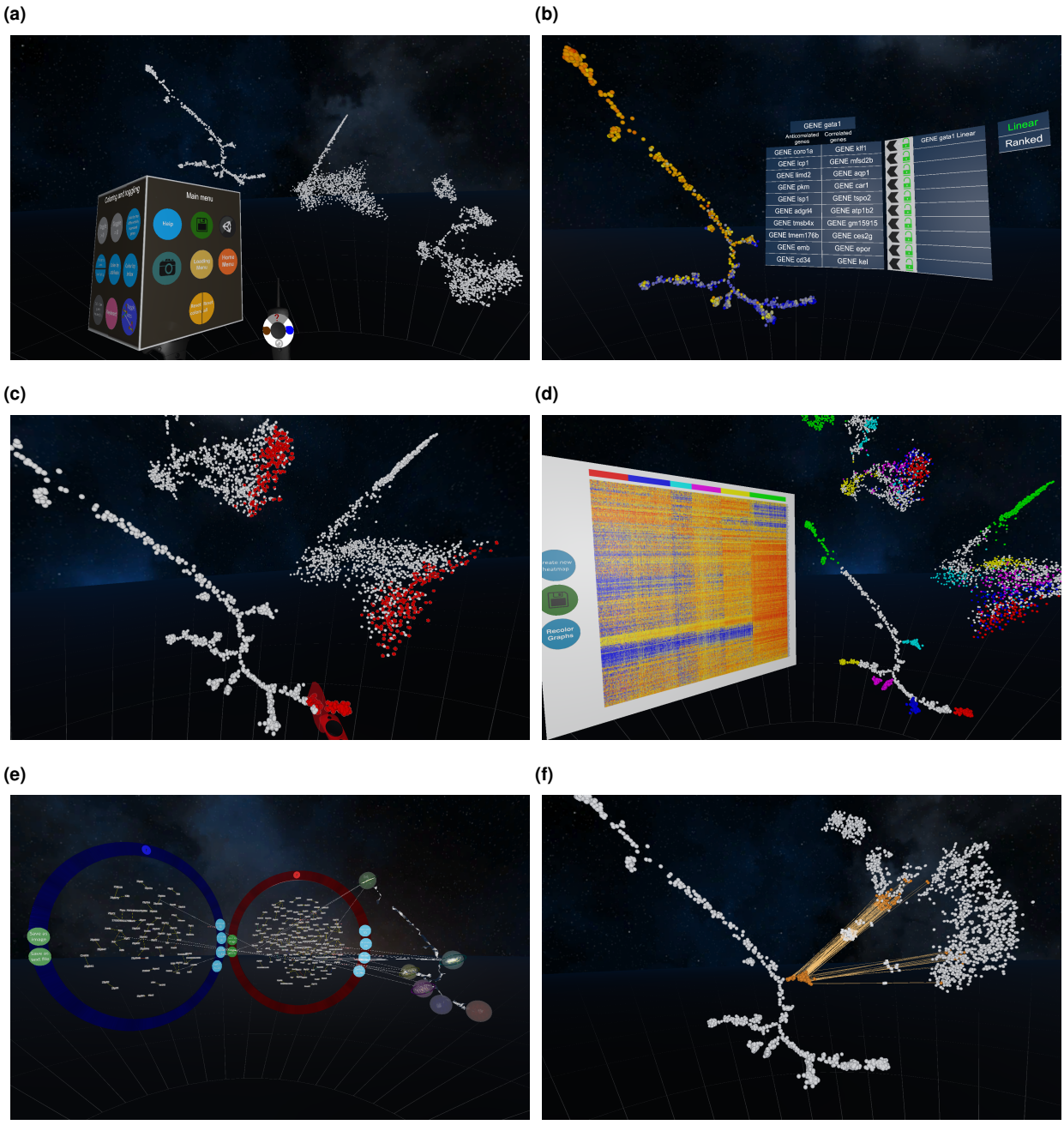

Figure 1: **(a)** Three sets of dimension reduced data (DDRTree, diffusion map, tSNE) from mouse hematopoietic stem and progenitor cells loaded into CellexalVR with the controllers in the foreground. **(b)** Cells coloured by the expression of Gata1, and the top 10 Correlated and anti-correlated genes are calculated by pressing the black chevron. Selecting a gene name with the laser will then re-colour each cell according to the expression of that gene. **(c)** The freehand selection tool. Passing cells through the paddle selects them into a group, and the corresponding cell in the other MDS plots are also highlighted simultaneously. Clicking the action-controller touchpad changes the colour of the selection tool to initiate a new group. **(d)** Differentially expressed genes between free-hand selected groups displayed as a heatmap. Coloured bars at the top of the heatmap show where the cells came from in the original selection. **(e)** Two transcription factor networks that have been generated from two of selected groups of cells. TF-TF pairs in common between the graphs are highlighted by connectors giving a visual measure of network similarity. **(f)** Cells selected from the DDRTree plot (left) are tracked to their counterpart cells in the tSNE plot (right), highlighting the fact they fall into two further potential groups.