1   **Spaniel: analysis and interactive sharing of Spatial Transcriptomics data**

2   Rachel Queen[1], Kathleen Cheung[1], Steven Lisgo[2], Jonathan Coxhead[3], Simon Cockell[1]

3   1. Bioinformatics Support Unit, Newcastle University, United Kingdom
4   2. Institute of Genetic Medicine, Newcastle University, Newcastle, United Kingdom
5   3. Genomics Core Facility, Newcastle University, United Kingdom

6

7   **Abstract**

8   Spatial Transcriptomics allows the sequencing of the complete transcriptomes from barcoded

9   regions of intact tissue. The technology has the potential to answer a wide range of biological

10 questions concerning cellular function, but analysis of the data presents a number of challenges

11 which are not met by existing analysis tools. Here we present Spaniel, an R package providing a

12 framework for analysing and sharing Spatial Transcriptomics data.

13 **Main**

14 Whilst technologies such as single cell RNAseq provide tools for dissecting cellular heterogeneity,

15 dissociation of the tissue mean that information about spatial relationship of the cells is lost. Spatial

16 Transcriptomics (ST) [1] is a method for sequencing spatially barcoded transcriptomes from intact

17 tissue sections. This provides a method of linking gene expression data to spatial location of a cell

18 within a tissue cross-section, to further elucidate cellular function. Computational analysis of

19 Spatial Transcriptomics data has a unique set of requirements which are not satisfied by existing

20 analysis tools. Firstly, there is a strong visual element to the analysis and gene expression must be

21 linked back to a histological image of the tissue. Secondly, to answer complex questions there is

22 need for data integration both between different Spatial Transcriptomics datasets and with other

23 datasets such as single cell RNAseq. Finally, a close collaboration between a computational

24 biologist analysing the data and the wet-lab biologist with specialised knowledge of the tissue type

25 is required to interpret the experiments, but as each Spatial Transcriptomics dataset generates a

26 large amount of data and a number of decisions must be taken during analysis, it difficult to share

27 data in a static manner.

28   Numerous tools exist for analysing single cell RNAseq and low input RNAseq experiments [2], but

29   these lack methods for visualising histological data. The ST Viewer tool [3] has been designed for

30   analysing and visualising ST data but as it is a standalone tool it is not suitable for integrative

31   analysis of multiple datasets.

32   Here we introduce Spaniel, an R package which provides a framework for analysing and sharing

33   Spatial Transcriptomics data. Spaniel provides methods for thorough quality control, visualisation

34   and pre-processing of the data. It uses two pre-existing S4 objects designed for single cell analysis,

35   namely Seurat object [4] and SingleCellExperiment object [5], providing strong potential for integration

36   of and with other single cell datasets. Spaniel also provides a Shiny app which allows results to be

37   shared in an interactive manner.

38

39   To test the Spaniel application and produce the figure presented here, we used publicly available

40   sequencing data [1] from the mouse olfactory bulb and the hematoxylin and eosin (H&E), HE_Rep1

41   taken from the Spatial Transcriptomics website (http://www.spatialtranscriptomicsresearch.org/).

42   Each Spatial Transcriptomics experiment involves sectioning a 5-16 µm slice of tissue which is

43   placed on a slide containing a grid of spatially barcoded polyT probes. The probes are positioned in

44   spots on the slide in a 35 * 35 grid, where each spot covers 10 – 100 cells. The sequencing data was

45   generated using paired end sequencing where read 1 contains the spatial barcode and UMI and read

46   2 contains the transcript sequence. Spaniel takes a spatial transcriptomic expression matrix where

47   each row corresponds to a gene and each column corresponds to a spot coordinate. To create the

48   expression data, from the sequencing data, the paired FASTQ files were demultiplexed with a

49   publically available perl script

50   (https://github.com/tallulandrews/scRNASeqPipeline/blob/master/0_custom_undo_demultiplexing.

51   pl) using the spatial barcodes encoded in read 1. Read 2 from successfully demultiplexed pairs were

52   trimmed for quality using Trimmomatic version 0.36 [6]. A reference was created using Ensembl

53   mouse reference genome Release M20 (GRCm38.p6). The trimmed reads were aligned to this

54 reference using STAR version 2.5.3a [7] in single read alignment mode. The number of reads were

55 quantified using HTSEQ version 0.6.1 [8] and a count matrix was created. The H & E images were

56 cropped to region around the edges of the spots and resized to 1000 x 1071 pixels with a resolution

57 of 72dpi using a photo editor.

58 Spaniel includes a series of tools to aid the quality control and analysis of Spatial Transcriptomics

59 data and is designed to be used alongside existing tools to aid integration of multiple datasets. It

60 includes functions to create either a Seurat S4 object or SingleCellExperiment S4 object which are

61 designed for single cell experiment analysis and contain slots for both expression data and

62 metadata. The package provides function to import a raw count matrix file and a barcodes text file.

63 The barcodes can be either be barcode file provided by Spatial Transcriptomics giving the

64 coordinates of each probe or adjusted barcodes obtained by pre-processing the image using ST Spot

65 Detector. The Scater R [9] package is used to calculate QC metrics for SingleCellExpeiment objects.

66 Spaniel also provides a function to create a rasterised grob, compatible with ggplot2, from a pre-

67 processed H&E image which is used as the background image for the plots. In this example, the

68 sample is from the mouse olfactory bulb which is a multi-layered structure found in the forebrain of

69 vertebrates (Figure1A).

70 The ST_plot function provides a method to visualise metrics contained within the Seurat object

71 overlaid onto the image of the tissue. This plot can be used initially as a quality control step as well

72 as to visualise different clustering solutions or gene expression of specific genes.

73 Quality Control is a crucial first step in the analysis of this data. By inspecting the number of genes

74 and number of reads per spot, using the ST_plot function, it is possible to pinpoint potential

75 problems with the data that may confound downstream analysis. For example, spots which fall

76 outside of the tissue with high number of counts may be an indicator of overall background RNA

77 contamination. Figure 1B shows the number of genes per spot where spots with at least 2000 genes

78 per spot are coloured in blue and spots with fewer than 2000 genes are coloured in red. The

79 majority of spots above this threshold fall within the tissue area as expected, but a few fall outside

80   the tissue area perhaps indicating contamination. One of the primary aims of Spatial

81   Transcriptomics analysis is the identification of anatomical regions which share similar

82   transcriptomes. As the package is tightly integrated with the Seurat package and

83   SingleCellExperiment object it possible to choose a wide range of clustering methods such as the

84   Seurat FindClusters function to identify regions sharing gene expression patterns. After clustering

85   Seurat can be used to detect genes which are differentially expressed between clusters. The clusters,

86   shown in figure 1C, appear to match the different layers of the olfactory bulb. For example, spots

87   which fall into cluster 0 are detected in the granule cell layer. Through differential gene expression,

88   *Nrgn* was found to be upregulated in cluster 0 compared to the other clusters (Figure 1D). This gene

89   is neuron-specific and is the top gene associated with the granule cell layer of the olfactory bulb in

90   the Allen Mouse Brain Atlas[10].

91

92   Spaniel's Shiny application can be hosted on a local computer or any service running R server (e.g.

93   shinyserver.io) and offers the user 4 plots showing a) the number of genes detected per spot, b) the

94   number of reads detected per spot, c) clustering results, d) the gene expression of a selected gene.

95   The app uses a pre-processed S4 object and grob which are both saved in in RDS format allowing

96   data to be transferred between the computational biologist performing the analysis and other

97   researchers within the group. The gene plot allows users to select any individual gene to plot, whilst

98   the cluster plot enables the visualisation of clustering results at multiple resolutions. Installation

99   instruction and code can be found here: https://github.com/RachelQueen1/Spaniel

100
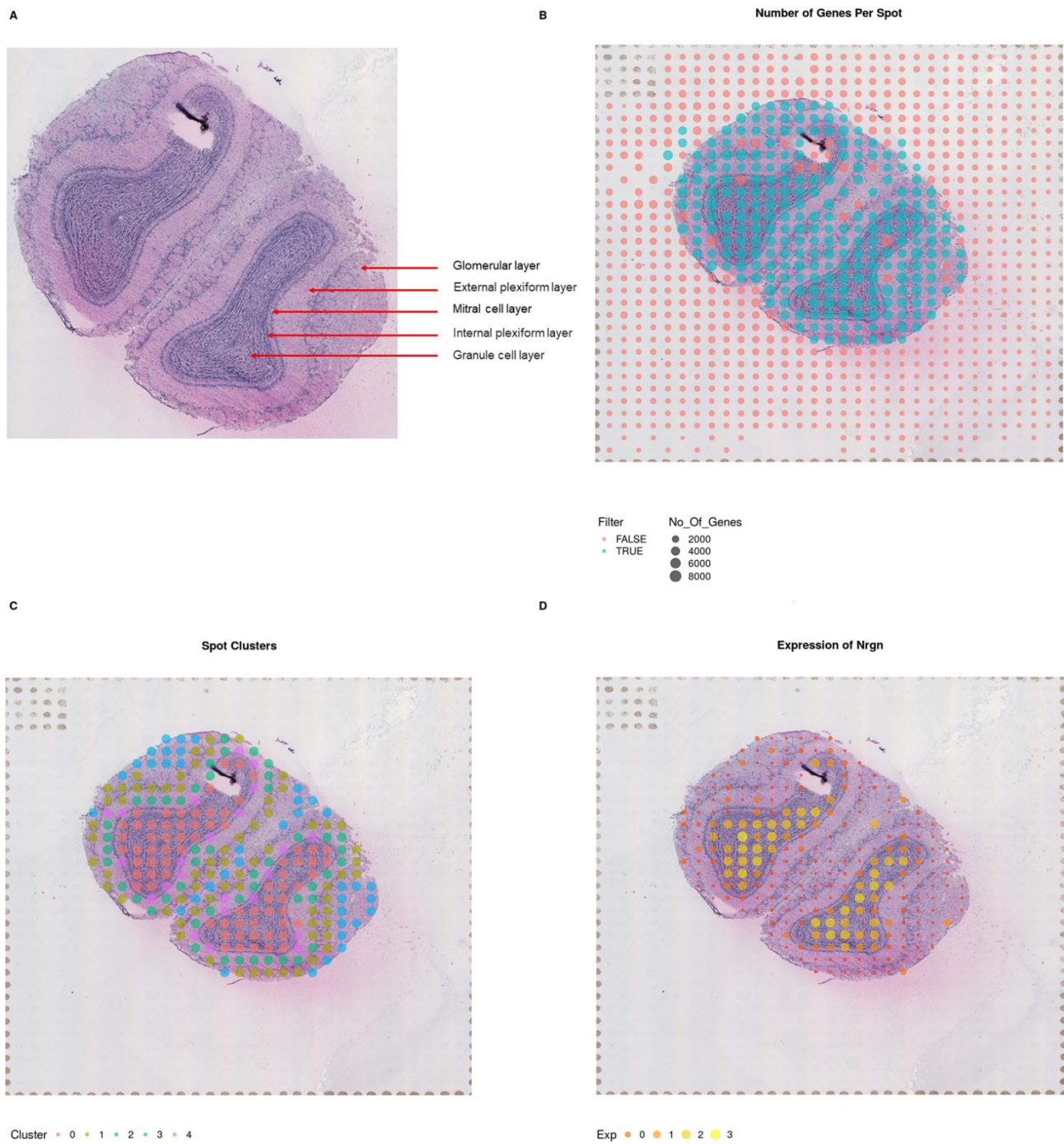
101

102 Figure 1: A) H&E image of mouse olfactory bulb showing the different layers. Example plots using

103 Spaniel showing B) quality control metrics such as the number of genes detected per spot, C) the

104 results of Seurat's FindClusters clustering analysis, D) the expression of the *Nrgn* gene.

105

106

107

108 **References**

109 1.   Patrik L. Ståhl et al. Visualization and analysis of gene expression in tissue sections by
110 spatial transcriptomics. *Science* **353**, 78-82 (2016).
111 2.   Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape
112   with the scRNA-tools database. *PLoS Comput Biol* **14**, e1006245 (2018).
113 3.   Fernandez Navarro, J., Lundeberg, J. & Stahl, P.L. ST viewer: a tool for analysis and
114   visualization of spatial transcriptomics datasets. *Bioinformatics* **35**, 1058-1060 (2019).
115 4.   Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell
116   transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*
117   **36**, 411-420 (2018).
118 5.   Lun A, R.D. SingleCellExperiment: S4 Classes for Single Cell Data. . (2019).
119 6.   Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
120   sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
121 7.   Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21
122   (2013).
123 8.   Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with high-
124   throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).
125 9.   McCarthy, D.J., Campbell, K.R., Lun, A.T. & Wills, Q.F. Scater: pre-processing, quality
126   control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*
127   **33**, 1179-1186 (2017).
128 10.   Lein, E.S. et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**,
129   168-176 (2007).
130