Cluster analysis protocol in the hepato-pancreato-biliary progenitor compartment of '*confetti'* mouse embryos

Anca Margineanu[1], David Willnow[2,3], Francesca M. Spagnoli[2]

[1]Max Delbrück Center for Molecular Medicine Berlin; [2]Centre for Stem Cells and Regenerative Medicine, King's College London; [3]Berlin Institute of Health
Anca.Margineanu@mdc-berlin.de; david.willnow@kcl.ac.uk; francesca.spagnoli@kcl.ac.uk

The brainbow or confetti animal models expressing cells uniquely labelled with fluorescent proteins via random genetic recombinations enable researchers to trace the cell fate by lineage analysis, as well as to analyse the distribution of individual clones generated by cellular proliferation [1].

Approaches to identify the clones and to count the cell numbers vary from manually defining the clone limits and counting the cells [2] to combining manual count and measurement of surface coverage with more complex mathematical models to explain the clone evolution [3] or to superimpose the confetti clone with another genetic marker [4] (for a review, see Roy et al. [5]). Other authors developed dedicated software to identify the clones and to perform automatic segmentation, further using the results in Monte Carlo simulations to predict spatial distributions [6].

We describe a protocol to identify cellular clones in the pancreas of '*confetti'* mouse embryos at an unique time point following induction using a doxycycline-activated transgenic system [7]. Images of dorsal and ventral pancreas in cleared mouse embryos were taken in two-photon microscopy using simultaneous three wavelengths excitation. After spectrally unmixing the detection channels of the corresponding fluorescent proteins, the 3D cellular coordinates have been automatically determined and further used to assess the numbers of cells per clone via cluster analysis. We implemented the analysis in the R statistical software using a hybrid algorithm described by Husson et al. [8]. Starting from the biological data estimating the number of cellular divisions from the day of induction until the collection of the samples, we tested 5 possible distributions of cells in clusters. To select the most probably distribution or to decide if additional possibilities must be explored, we used statistical parameters generated by the algorithm in combination with validation indices obtained by running our data via a cluster validation algorithm in R [9]; the results of the validation algorithm were further refined using a mathematical analysis of five selected indices.

Clustering algorithms are part of the exploratory data analysis and, in the absence of the ground truth, a protocol to find the most probable distribution of cells per cluster is necessary. We believe that the steps described here are applicable to other biological data where clone identification is necessary.

1. Livet et al., *Nature*, 2007, **450**:56.
2. Ritsma et al., *Nature*, 2014, **507**:362.
3. Snippert et al., *Cell*, 2010, **143**: 134.
4. Calzolari et al., *Nat. Neurosci.*, 2015, **18**: 490.
5. Roy et al., *Stem Cells*, 2014, **32**:3046.
6. Ghigo et al., *J. Exp. Med.* 2013, **210**: 1657.
7. Willnow et al., 2020, bioRxiv preprint doi: https://doi.org/10.1101/2020.08.06.240176.
8. Husson et al., https://www.agrocampus-ouest.fr/math/.
9. Charrad et al., *J. Stat. Software*, 2014: **61**, 1.