This is the documentation for code to run Bayesian cluster analysis on single-molecule microscopy (SMLM) data sets. This document assumes it is being used on a Linux platform.

1. **Installing R**

The code runs in the R programming language. To install, download R from: <http://www.r-project.org/>

We then recommend the use of R-Studio which provides a graphical user interface for R. This is also freely available at:

<http://www.rstudio.com/>

The first time the code is run, two additional R plugins are required and can be installed using the tools-> install packages menu. The two packages required are “splancs” and “igraph”.

1. **Data format.**

We have included an example data set, however, should you wish to test the code with your own data it needs to be in a specific format. – the same as our example

1. Create a folder in the same folder as the three R files with the name of the experiment, for example “Test”
2. Within that folder place a text file called config.txt. This specifies the parameters of the model against which cluster proposal will be scored. The supplied config file contains our recommended settings.
   * xlim and ylim specify the range of the data i.e. the minimum and maximum x and y values in the data set. Our provided data example has values from 0-3000 but for real data this might be for example 23000-26000. We have found analysing 3000x3000nm regions is good, but you can analyse larger or smaller areas.
   * histbins and histvalues specify a prior on the sizes (radii) of clusters as described in our manuscript.
   * Alpha is the prior on the Dirichlet concentration parameter as described in our manuscript
   * pbackground is the prior on the probability of a point being in the background, again as described in our paper
3. In the same folder add sequentially numbered sub folders named 1,2,3 etc. Our code will batch process all such folders. Each one contains one region which will be analysed.
   * Each region is specified using a text file called data.txt.
   * This is a comma separated file with carriage return
   * The first line must be x,y,sd,ID
   * The data is then in 4 comma separated columns, one line per molecule. X and y are the coordinates in nm. sd is the localisation precision of each molecule. If this is unknown, then set to zero and the code should assume it is precise. The 4th column is the cluster to which the molecule belongs (for simulated data). Molecules not in clusters are labelled sequentially.
   * If using real data and the ID is not known, leave this column out and delete the ID heading
4. **Once the data is in place you can run the cluster analysis**
   * Open RStudio
   * Set the working directory to the folder containing the three R files by typing for example setwd(“~/Myfolder/”)
   * Open the file named run.R
   * In line 2, type the name of the folder you want to process, for example foldername=”Test”
   * Click “Source” to run the code
   * The code should then run. Estimate 30mins per data set (depending on the size and number of molecules)
   * After the code has finished, it will have created a folder in each 1,2,3 sub-folder called “labels” with all the cluster proposals from different combinations of r and T
   * It will also have created a matrix called r\_vs\_thresh.txt containing the scores of each combination. Plotting these, for example using contourf in Matlab gives the r-T plots we show in our paper
5. Post-processing

* In RStudio you can then open postprocessing.R and click Source to run it. This will generate:
  + Histograms of the number of clusters per region and histograms of the cluster radii for the whole batch process that was performed
  + Additionally, within each data folder it will generate a map of the best scored cluster proposals (+ the “real” answer if simulated data was used with an ID column). It will also generate a text file called “summary.txt” with the key cluster descriptors for that region