**Chromatin state specific transcriptional regulatory network identification**

**Version** 1.0

**Date** 08/01/2017

**Description** DPM-LGCP stands for Dirichlet Prior mixture for Log Gaussian Cox Process. It is a non-parametric Bayesian clustering technique based on non-homogeneous Poisson process which has been used to identify transcriptional regulatory modules in distinct chromatin states. The algorithm clusters transcription factors sharing similar intensity function (binding patterns). It has broader applications to cases requiring un-supervised clustering of large datasets with prior information on clusters.

**Author** Sharmi Banerjee, Honxiao Zhu, Man Tang, Xiaowei Wu, Wu Feng, David Xie

**System Requirements**

DP-LGCP is independent of operating systems as it is written in R. Basic requirements for running the pipeline include installing R and the following libraries- 'INLA'(INLA 0.0-1468872408, dated 2016-07-18 (14:43:05+0100)), ‘miceadds’, ‘foreach’, ‘doParallel’, ‘igraph’, ‘caTools’, ‘ComplexHeatmap’, ‘pvclust’, ‘MASS’, ‘circlize’, ‘RColorBrewer’, ‘pheatmap’.

**Usage**

Unzip the package. Change the current directory in R to the ‘LGCP\_package' folder containing the code and data organized into subfolders –

* ‘INLA’ contains the R package
* ‘Real’ contains script, data and some results for neural stem cells used in the paper
  + ‘Data’ – peak files for twenty-one transcription factors, diHMM generated domain bed files for thirty domains and mm10 refseq genes.
  + ‘main\_code’ – scripts to generate clusters in chromatin states
  + ‘plotting\_code’ - scripts to generate enrichment maps of TF peaks and TF modules in chromatin states
  + ‘results’ – clustering results (RData files) on chromatin states
* ‘Simulation’ has code and results for simulation data
  + ‘code’ – scripts to simulate transcription factors binding sites with twenty TFs and three clusters and evaluate results by applying clustering algorithm using non-homogeneous Poisson point process
  + ‘data’ – simulated data generated from scripts in ’code’ folder
  + ‘results’ – clustering results
* ‘Neal\_Alg3\_funs’ has the main algorithm clustering functions.

**Setting INLA path**

Extract INLA compressed file and copy INLA folder under LGCP-package to R library located in - /Library/Frameworks/R.framework/Versions/3.3/Resources/library

**Clustering on simulation data**

1. Set working directory in R to the path up to the Simulation directory in LGCP package. E.g. -

setwd('/Users/sharmibanerjee/Documents/Summer2017/LGCP\_package/Simulation/')

1. Set work path in R console. E.g. -

workpath = "/Users/sharmibanerjee/Documents/Summer2017/LGCP\_package/ Simulation/ code/”

1. Run simulation script -

source("script\_simulate\_data.R")

1. Evaluate clustering result -

source("script\_run\_Neal\_alg3.R")

**Simulation case study –**

* Three 1D non-homogeneous Poisson processes with Log-Gaussian intensity were generated. Next, binding site locations of twenty transcription factors were simulated by drawing independent samples from these non-homogeneous Poisson processes. In Figure 1A, the solid curves denote the true intensities.
* Using the simulated binding site locations, we applied the clustering algorithm. The TFs were randomly assigned to any cluster at initializations. The results are shown after ten iterations –
  + initial error rate: 0.394736842105263
  + estimation error rate: 0
  + overall true marginal likelihood: 76.7314595602879
  + overall estimated marginal likelihood: 76.7314595602879
  + true marginal likelihood per cluster: 27.2765131123867, 14.9562841559956, 34.4986622919056
  + estimated marginal likelihood per cluster: 27.2765131123867, 14.9562841559956, 34.4986622919056

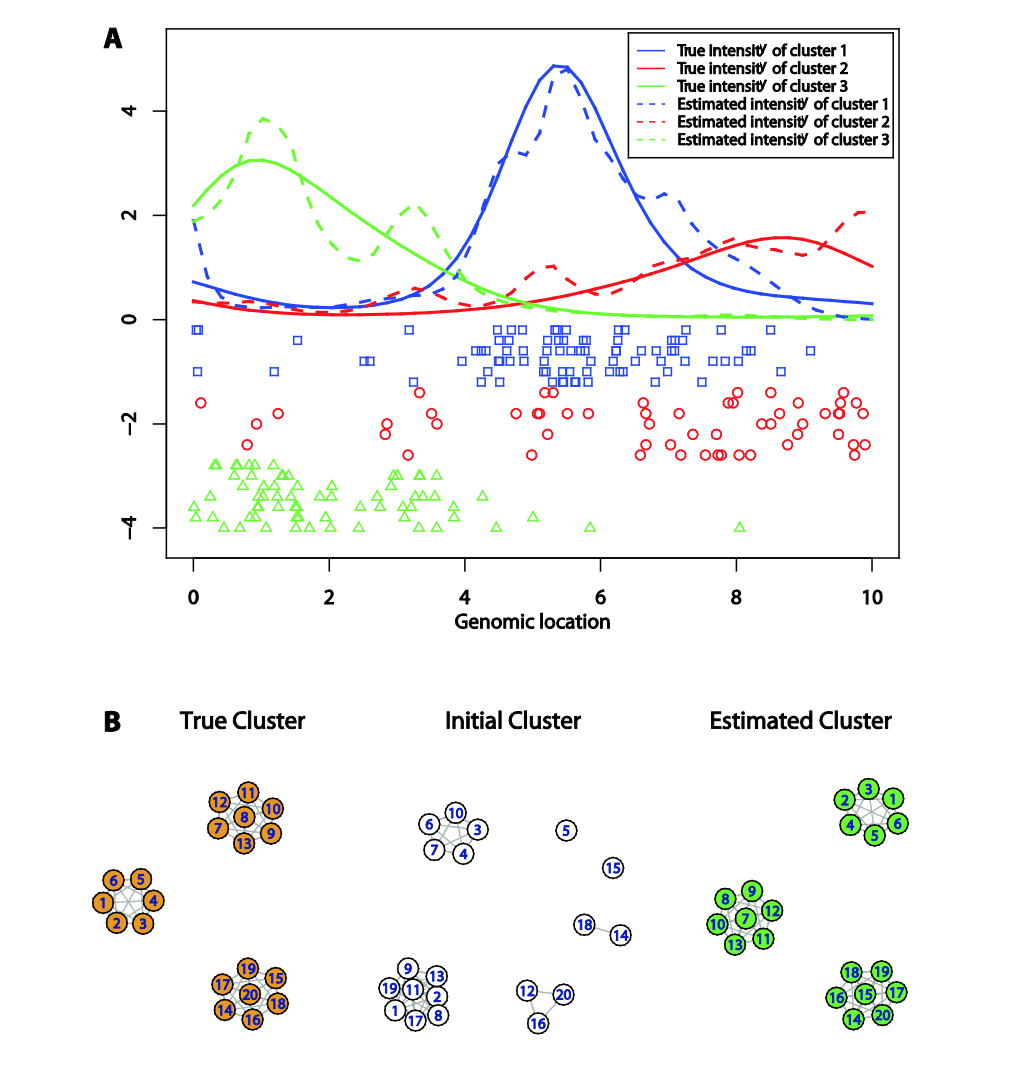


Figure 1: Clustering results obtained on simulated data. (A). Estimated and True intensity curves for three clusters along with the binding site locations. (B) True, Initial and Estimated clusters.

**Clustering on real data**

1. Set working directory in R to the path up to the Real directory in LGCP package. E.g. -

setwd('/Users/sharmibanerjee/Documents/Summer2017/LGCP\_package/Real/ main\_code/')

1. Set work path in R console. E.g. -

workpath = "/Users/sharmibanerjee/Documents/Summer2017/LGCP\_package/”

1. Run the following scripts –

source("script\_generate\_TF\_peaks\_in\_diHMM\_states.R")

source("script\_get\_clusters\_for\_each\_diHMM\_state.R")

1. Files generated
   1. Clustering results will be stored in the ‘results’ folder under ‘Real’ with the file names *diHMM\_celltype\_cluster\_domainname.RData*
   2. TF peaks in each chromatin state will be generated with names *celltype\_TFpeaks\_diHMM\_chromatinstate.RData*

**Plotting enrichment heat-maps**

1. Set working directory in R to the path up to the Real directory in LGCP package. E.g. -

setwd('/Users/sharmibanerjee/Documents/Summer2017/LGCP\_package/Real/ plotting\_code/')

1. Set work path in R console. E.g. -

workpath = “/Users/sharmibanerjee/Documents/Summer2017/LGCP\_package/”

1. Run the following scripts to generate heat map in Figure 2(a) in main paper

source("TF\_peak\_enrichment\_in \_states.R")