scGPS introduction

Quan Nguyen and Michael Thompson

2018-12-21

Contents

1. Installation instruction

```
# Prior to installing scGPS you need to install the SummarizedExperiment
# bioconductor package as the following
# source('https://bioconductor.org/biocLite.R') biocLite('SummarizedExperiment')

# To install scGPS from github (Depending on the configuration of the local
# computer or HPC, possible custom C++ compilation may be required - see
# installation trouble-shootings below)
devtools::install_github("IMB-Computational-Genomics-Lab/scGPS")

# for C++ compilation trouble-shooting, manual download and installation can be
# done from github

git clone https://github.com/IMB-Computational-Genomics-Lab/scGPS

# then check in scGPS/src if any of the precompiled (e.g.  those with *.so and
# *.o) files exist and delete them before recompiling

# create a Makevars file in the scGPS/src with one line: PKG_LIBS =
# $(LAPACK_LIBS) $(BLAS_LIBS) $(FLIBS)
```

```
# then with the scGPS as the R working directory, manually recompile scGPS in R
# using devtools to load and install functions
devtools::document()
#load the package to the workspace
devtools::load_all()
```

2. A simple workflow of the scGPS:

The purpose of this workflow is to solve the following task: given a mixed population with known subpopulations, estimate transition scores between these subpopulation

2.1 Create scGPS objects

```
# load mixed population 1 (loaded from sample1 dataset, named it as day2)

devtools::load_all('/Users/quan.nguyen/Documents/Powell_group_MacQuan/AllCodes/scGPS/')

day2 <- sample1
mixedpop1 <- NewscGPS(ExpressionMatrix = day2$dat2_counts, GeneMetadata = day2$dat2geneInfo,
    CellMetadata = day2$dat2_clusters)

# load mixed population 2 (loaded from sample2 dataset, named it as day5)
day5 <- sample2
mixedpop2 <- NewscGPS(ExpressionMatrix = day5$dat5_counts, GeneMetadata = day5$dat5geneInfo,
    CellMetadata = day5$dat5_clusters)
```

2.2 Run prediction

```
# select a subpopulation
c_selectID <- 1
# load gene list (this can be any lists of user selected genes)
genes <- GeneList
genes <- genes$Merged_unique
# load cluster information
cluster_mixedpop1 <- colData(mixedpop1)[,1]
cluster_mixedpop2 <- colData(mixedpop2)[,1]
#run training
LSOLDA_dat <- bootstrap_scGPS(nboots = 2, mixedpop1 = mixedpop1,
    mixedpop2 = mixedpop2, genes = genes, c_selectID  = c_selectID, listData = list(),
     cluster_mixedpop1 = cluster_mixedpop1,
     cluster_mixedpop2 = cluster_mixedpop2)
#> [1] "Total 224 cells as source subpop"
#> [1] "Total 366 cells in remaining subpops"
#> [1] "subsampling 112 cells for training souce subpop"
#> [1] "subsampling 112 cells in remaining subpops for training"
#> [1] "use 107 genes for training model"
#> [1] "use 107 genes 224 cells for testing model"
```

```
#> [1] "rename remaining subpops to 2_3"
#> [1] "there are 112 cells in class 2_3 and 112 cells in class 1"
#> [1] "removing 13 genes with no variance"
#> [1] "standardizing prediction/target dataset"
#> [1] "performming elasticnet model training..."
#> [1] "performning LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>        Df       %Dev    Lambda
#>   [1,]  0 -2.563e-15 3.309e-01
#>   [2,]  1  5.377e-02 3.015e-01
#>   [3,]  1  9.904e-02 2.747e-01
#>   [4,]  2  1.396e-01 2.503e-01
#>   [5,]  2  1.781e-01 2.281e-01
#>   [6,]  2  2.112e-01 2.078e-01
#>   [7,]  2  2.398e-01 1.893e-01
#>   [8,]  2  2.646e-01 1.725e-01
#>   [9,]  2  2.861e-01 1.572e-01
#>  [10,]  3  3.069e-01 1.432e-01
#>  [11,]  3  3.268e-01 1.305e-01
#>  [12,]  4  3.456e-01 1.189e-01
#>  [13,]  5  3.645e-01 1.084e-01
#>  [14,]  5  3.822e-01 9.873e-02
#>  [15,]  6  3.979e-01 8.996e-02
#>  [16,]  6  4.136e-01 8.196e-02
#>  [17,]  6  4.276e-01 7.468e-02
#>  [18,]  9  4.428e-01 6.805e-02
#>  [19,] 10  4.579e-01 6.200e-02
#>  [20,] 11  4.718e-01 5.649e-02
#>  [21,] 11  4.844e-01 5.148e-02
#>  [22,] 12  4.965e-01 4.690e-02
#>  [23,] 12  5.079e-01 4.274e-02
#>  [24,] 13  5.188e-01 3.894e-02
#>  [25,] 15  5.298e-01 3.548e-02
#>  [26,] 16  5.420e-01 3.233e-02
#>  [27,] 17  5.536e-01 2.946e-02
#>  [28,] 18  5.649e-01 2.684e-02
#>  [29,] 21  5.760e-01 2.446e-02
#>  [30,] 22  5.874e-01 2.228e-02
#>  [31,] 24  5.983e-01 2.030e-02
#>  [32,] 26  6.084e-01 1.850e-02
#>  [33,] 29  6.184e-01 1.686e-02
#>  [34,] 32  6.297e-01 1.536e-02
#>  [35,] 33  6.408e-01 1.399e-02
#>  [36,] 34  6.516e-01 1.275e-02
#>  [37,] 34  6.614e-01 1.162e-02
#>  [38,] 37  6.716e-01 1.059e-02
#>  [39,] 38  6.819e-01 9.646e-03
#>  [40,] 43  6.919e-01 8.789e-03
#>  [41,] 46  7.018e-01 8.008e-03
#>  [42,] 48  7.115e-01 7.297e-03
```

```
#>  [43,] 49  7.208e-01 6.648e-03
#>  [44,] 53  7.305e-01 6.058e-03
#>  [45,] 55  7.401e-01 5.520e-03
#>  [46,] 57  7.494e-01 5.029e-03
#>  [47,] 59  7.584e-01 4.582e-03
#>  [48,] 59  7.667e-01 4.175e-03
#>  [49,] 59  7.749e-01 3.804e-03
#>  [50,] 60  7.824e-01 3.466e-03
#>  [51,] 63  7.902e-01 3.159e-03
#>  [52,] 64  7.984e-01 2.878e-03
#>  [53,] 65  8.065e-01 2.622e-03
#>  [54,] 67  8.152e-01 2.389e-03
#>  [55,] 67  8.241e-01 2.177e-03
#>  [56,] 68  8.329e-01 1.984e-03
#>  [57,] 69  8.414e-01 1.807e-03
#>  [58,] 68  8.503e-01 1.647e-03
#>  [59,] 69  8.587e-01 1.501e-03
#>  [60,] 68  8.671e-01 1.367e-03
#>  [61,] 69  8.752e-01 1.246e-03
#>  [62,] 70  8.835e-01 1.135e-03
#>  [63,] 69  8.916e-01 1.034e-03
#>  [64,] 70  8.991e-01 9.424e-04
#>  [65,] 71  9.063e-01 8.587e-04
#>  [66,] 71  9.131e-01 7.824e-04
#>  [67,] 72  9.196e-01 7.129e-04
#>  [68,] 70  9.259e-01 6.495e-04
#>  [69,] 69  9.315e-01 5.918e-04
#>  [70,] 72  9.371e-01 5.393e-04
#>  [71,] 71  9.425e-01 4.914e-04
#>  [72,] 72  9.473e-01 4.477e-04
#>  [73,] 71  9.519e-01 4.079e-04
#>  [74,] 71  9.560e-01 3.717e-04
#>  [75,] 72  9.599e-01 3.387e-04
#>  [76,] 73  9.634e-01 3.086e-04
#>  [77,] 73  9.667e-01 2.812e-04
#>  [78,] 74  9.696e-01 2.562e-04
#>  [79,] 74  9.724e-01 2.334e-04
#>  [80,] 74  9.749e-01 2.127e-04
#>  [81,] 74  9.771e-01 1.938e-04
#>  [82,] 75  9.791e-01 1.766e-04
#>  [83,] 75  9.810e-01 1.609e-04
#>  [84,] 77  9.827e-01 1.466e-04
#>  [85,] 78  9.842e-01 1.336e-04
#>  [86,] 78  9.856e-01 1.217e-04
#>  [87,] 79  9.869e-01 1.109e-04
#>  [88,] 78  9.881e-01 1.010e-04
#>  [89,] 78  9.891e-01 9.207e-05
#>  [90,] 78  9.901e-01 8.389e-05
#>  [91,] 77  9.910e-01 7.644e-05
#>  [92,] 78  9.918e-01 6.965e-05
#>  [93,] 78  9.925e-01 6.346e-05
#>  [94,] 78  9.931e-01 5.782e-05
#>  [95,] 79  9.937e-01 5.269e-05
```

```
#>  [96,] 79  9.943e-01 4.801e-05
#>  [97,] 79  9.948e-01 4.374e-05
#>  [98,] 79  9.952e-01 3.986e-05
#>  [99,] 79  9.956e-01 3.632e-05
#> [100,] 79  9.960e-01 3.309e-05
#> [1] "lambda min is at location 18"
#> [1] "the leave-out cells in the source subpop is 112"
#> [1] "use 112 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 224 cells and 94 genes..."
#> [1] "evaluation accuracy ElasticNet 0.852678571428571"
#> [1] "evaluation accuracy LDA 0.78125"
#> [1] "done bootstrap 1"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 43.8502673796791"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is NaN"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 87.1428571428571"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is NaN"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 19.5488721804511"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is NaN"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 52.5"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is NaN"
#> [1] "Total 224 cells as source subpop"
#> [1] "Total 366 cells in remaining subpops"
#> [1] "subsampling 112 cells for training souce subpop"
#> [1] "subsampling 112 cells in remaining subpops for training"
#> [1] "use 107 genes for training model"
#> [1] "use 107 genes 224 cells for testing model"
#> [1] "rename remaining subpops to 2_3"
#> [1] "there are 112 cells in class 2_3 and 112 cells in class 1"
#> [1] "removing 8 genes with no variance"
#> [1] "standardizing prediction/target dataset"
#> [1] "performming elasticnet model training..."
#> [1] "performming LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
```

```
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>        Df       %Dev    Lambda
#>   [1,]  0 -2.563e-15 3.279e-01
#>   [2,]  1  5.275e-02 2.987e-01
#>   [3,]  1  9.697e-02 2.722e-01
#>   [4,]  1  1.343e-01 2.480e-01
#>   [5,]  2  1.712e-01 2.260e-01
#>   [6,]  2  2.041e-01 2.059e-01
#>   [7,]  2  2.324e-01 1.876e-01
#>   [8,]  2  2.569e-01 1.709e-01
#>   [9,]  2  2.780e-01 1.558e-01
#>  [10,]  2  2.964e-01 1.419e-01
#>  [11,]  2  3.124e-01 1.293e-01
#>  [12,]  2  3.264e-01 1.178e-01
#>  [13,]  2  3.385e-01 1.074e-01
#>  [14,]  3  3.501e-01 9.782e-02
#>  [15,]  4  3.618e-01 8.913e-02
#>  [16,]  5  3.741e-01 8.122e-02
#>  [17,]  5  3.857e-01 7.400e-02
#>  [18,]  6  3.964e-01 6.743e-02
#>  [19,]  9  4.071e-01 6.144e-02
#>  [20,] 11  4.221e-01 5.598e-02
#>  [21,] 11  4.363e-01 5.101e-02
#>  [22,] 13  4.505e-01 4.647e-02
#>  [23,] 15  4.659e-01 4.235e-02
#>  [24,] 17  4.819e-01 3.858e-02
#>  [25,] 19  4.984e-01 3.516e-02
#>  [26,] 22  5.154e-01 3.203e-02
#>  [27,] 24  5.326e-01 2.919e-02
#>  [28,] 24  5.485e-01 2.659e-02
#>  [29,] 25  5.630e-01 2.423e-02
#>  [30,] 28  5.768e-01 2.208e-02
#>  [31,] 31  5.904e-01 2.012e-02
#>  [32,] 33  6.032e-01 1.833e-02
#>  [33,] 35  6.148e-01 1.670e-02
#>  [34,] 37  6.261e-01 1.522e-02
#>  [35,] 40  6.375e-01 1.387e-02
#>  [36,] 44  6.498e-01 1.263e-02
#>  [37,] 45  6.621e-01 1.151e-02
#>  [38,] 47  6.735e-01 1.049e-02
#>  [39,] 50  6.880e-01 9.557e-03
#>  [40,] 54  7.044e-01 8.708e-03
#>  [41,] 56  7.192e-01 7.935e-03
#>  [42,] 57  7.326e-01 7.230e-03
#>  [43,] 57  7.450e-01 6.588e-03
#>  [44,] 58  7.568e-01 6.002e-03
#>  [45,] 59  7.681e-01 5.469e-03
#>  [46,] 60  7.788e-01 4.983e-03
#>  [47,] 62  7.889e-01 4.541e-03
#>  [48,] 65  8.002e-01 4.137e-03
#>  [49,] 66  8.114e-01 3.770e-03
#>  [50,] 66  8.224e-01 3.435e-03
```

```
#>  [51,] 65  8.333e-01 3.130e-03
#>  [52,] 66  8.439e-01 2.852e-03
#>  [53,] 68  8.541e-01 2.598e-03
#>  [54,] 67  8.642e-01 2.367e-03
#>  [55,] 69  8.740e-01 2.157e-03
#>  [56,] 71  8.836e-01 1.966e-03
#>  [57,] 69  8.928e-01 1.791e-03
#>  [58,] 70  9.014e-01 1.632e-03
#>  [59,] 71  9.096e-01 1.487e-03
#>  [60,] 71  9.174e-01 1.355e-03
#>  [61,] 73  9.247e-01 1.234e-03
#>  [62,] 71  9.313e-01 1.125e-03
#>  [63,] 72  9.375e-01 1.025e-03
#>  [64,] 71  9.431e-01 9.338e-04
#>  [65,] 71  9.482e-01 8.508e-04
#>  [66,] 72  9.530e-01 7.752e-04
#>  [67,] 72  9.573e-01 7.064e-04
#>  [68,] 72  9.612e-01 6.436e-04
#>  [69,] 72  9.647e-01 5.864e-04
#>  [70,] 71  9.679e-01 5.343e-04
#>  [71,] 71  9.708e-01 4.869e-04
#>  [72,] 71  9.735e-01 4.436e-04
#>  [73,] 71  9.759e-01 4.042e-04
#>  [74,] 72  9.781e-01 3.683e-04
#>  [75,] 72  9.800e-01 3.356e-04
#>  [76,] 72  9.818e-01 3.058e-04
#>  [77,] 72  9.835e-01 2.786e-04
#>  [78,] 72  9.849e-01 2.539e-04
#>  [79,] 72  9.863e-01 2.313e-04
#>  [80,] 72  9.875e-01 2.108e-04
#>  [81,] 72  9.886e-01 1.920e-04
#>  [82,] 72  9.896e-01 1.750e-04
#>  [83,] 73  9.906e-01 1.594e-04
#>  [84,] 73  9.914e-01 1.453e-04
#>  [85,] 74  9.921e-01 1.324e-04
#>  [86,] 75  9.928e-01 1.206e-04
#>  [87,] 75  9.935e-01 1.099e-04
#>  [88,] 75  9.940e-01 1.001e-04
#>  [89,] 75  9.946e-01 9.123e-05
#>  [90,] 75  9.950e-01 8.313e-05
#>  [91,] 75  9.955e-01 7.574e-05
#>  [92,] 76  9.959e-01 6.901e-05
#>  [93,] 76  9.962e-01 6.288e-05
#>  [94,] 76  9.966e-01 5.730e-05
#>  [95,] 77  9.969e-01 5.221e-05
#>  [96,] 77  9.971e-01 4.757e-05
#>  [97,] 77  9.974e-01 4.334e-05
#>  [98,] 77  9.976e-01 3.949e-05
#>  [99,] 77  9.978e-01 3.598e-05
#> [100,] 77  9.980e-01 3.279e-05
#> [1] "lambda min is at location 22"
#> [1] "the leave-out cells in the source subpop is 112"
#> [1] "use 112 target subpops cells for leave-out test set"
```

```
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 224 cells and 99 genes..."
#> [1] "evaluation accuracy ElasticNet 0.84375"
#> [1] "evaluation accuracy LDA 0.78125"
#> [1] "done bootstrap 2"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 2.13903743315508"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is NaN"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 41.4285714285714"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is NaN"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 1.50375939849624"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is NaN"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 17.5"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is NaN"
```

2.3 Summarise results

```
# display the list of result information in the LASOLDA_dat object
names(LSOLDA_dat)
#> [1] "Accuracy"          "ElasticNetGenes"   "Deviance"
#> [4] "ElasticNetFit"     "LDAFit"            "predictor_S1"
#> [7] "ElasticNetPredict" "LDAPredict"
LSOLDA_dat$ElasticNetPredict
#> [[1]]
#> [[1]][[1]]
#> [1] "ElasticNet for subpop1 in target mixedpop2"
#>
#> [[1]][[2]]
#> [1] 43.85027
#>
#> [[1]][[3]]
#> [1] "ElasticNet for subpop2 in target mixedpop2"
#>
#> [[1]][[4]]
```

```
#> [1] 87.14286
#>
#> [[1]][[5]]
#> [1] "ElasticNet for subpop3 in target mixedpop2"
#>
#> [[1]][[6]]
#> [1] 19.54887
#>
#> [[1]][[7]]
#> [1] "ElasticNet for subpop4 in target mixedpop2"
#>
#> [[1]][[8]]
#> [1] 52.5
#>
#>
#> [[2]]
#> [[2]][[1]]
#> [1] "ElasticNet for subpop1 in target mixedpop2"
#>
#> [[2]][[2]]
#> [1] 2.139037
#>
#> [[2]][[3]]
#> [1] "ElasticNet for subpop2 in target mixedpop2"
#>
#> [[2]][[4]]
#> [1] 41.42857
#>
#> [[2]][[5]]
#> [1] "ElasticNet for subpop3 in target mixedpop2"
#>
#> [[2]][[6]]
#> [1] 1.503759
#>
#> [[2]][[7]]
#> [1] "ElasticNet for subpop4 in target mixedpop2"
#>
#> [[2]][[8]]
#> [1] 17.5
LSOLDA_dat$LDAPredict
#> [[1]]
#> [[1]][[1]]
#> [1] "LDA for subpop 1 in target mixedpop2"
#>
#> [[1]][[2]]
#> [1] NaN
#>
#> [[1]][[3]]
#> [1] "LDA for subpop 2 in target mixedpop2"
#>
#> [[1]][[4]]
#> [1] NaN
#>
```

```
#> [[1]][[5]]
#> [1] "LDA for subpop 3 in target mixedpop2"
#>
#> [[1]][[6]]
#> [1] NaN
#>
#> [[1]][[7]]
#> [1] "LDA for subpop 4 in target mixedpop2"
#>
#> [[1]][[8]]
#> [1] NaN
#>
#>
#> [[2]]
#> [[2]][[1]]
#> [1] "LDA for subpop 1 in target mixedpop2"
#>
#> [[2]][[2]]
#> [1] NaN
#>
#> [[2]][[3]]
#> [1] "LDA for subpop 2 in target mixedpop2"
#>
#> [[2]][[4]]
#> [1] NaN
#>
#> [[2]][[5]]
#> [1] "LDA for subpop 3 in target mixedpop2"
#>
#> [[2]][[6]]
#> [1] NaN
#>
#> [[2]][[7]]
#> [1] "LDA for subpop 4 in target mixedpop2"
#>
#> [[2]][[8]]
#> [1] NaN

# summary results LDA
summary_prediction_lda(LSOLDA_dat = LSOLDA_dat, nPredSubpop = 4)
#>    V1  V2                                names
#> 1 NaN NaN LDA for subpop 1 in target mixedpop2
#> 2 NaN NaN LDA for subpop 2 in target mixedpop2
#> 3 NaN NaN LDA for subpop 3 in target mixedpop2
#> 4 NaN NaN LDA for subpop 4 in target mixedpop2

# summary results Lasso to show the percent of cells classified as cells belonging
summary_prediction_lasso(LSOLDA_dat = LSOLDA_dat, nPredSubpop = 4)
#>                 V1                V2
#> 1 43.8502673796791 2.13903743315508
#> 2 87.1428571428571 41.4285714285714
#> 3 19.5488721804511 1.50375939849624
#> 4             52.5             17.5
```

```
#>                                          names
#> 1 ElasticNet for subpop1 in target mixedpop2
#> 2 ElasticNet for subpop2 in target mixedpop2
#> 3 ElasticNet for subpop3 in target mixedpop2
#> 4 ElasticNet for subpop4 in target mixedpop2

# summary accuracy to check the model accuracy in the leave-out test set
summary_accuracy(object = LSOLDA_dat)
#> [1] 78.125 78.125

# summary maximum deviance explained by the model
summary_deviance(object = LSOLDA_dat)
#> $allDeviance
#> [1] "0.4428" "0.4505"
#>
#> $DeviMax
#>                   Dfd          Deviance           DEgenes
#> 1                   0       -2.563e-15    genes_cluster1
#> 2                   1            0.1343    genes_cluster1
#> 3                   2            0.3385    genes_cluster1
#> 4                   3            0.3501    genes_cluster1
#> 5                   4            0.3618    genes_cluster1
#> 6                   5            0.3857    genes_cluster1
#> 7                   6            0.3964    genes_cluster1
#> 8                   9            0.4071    genes_cluster1
#> 9                  11            0.4363    genes_cluster1
#> 10                 13            0.4505    genes_cluster1
#> 11 remaining DEgenes remaining DEgenes remaining DEgenes
#>
#> $LassoGenesMax
#> NULL
```

3. A complete workflow of the scGPS:

The purpose of this workflow is to solve the following task: given an unknown mixed population, find clusters and estimate relationship between clusters

3.1 Identify clusters in a dataset using CORE

(skip this step if clusters are known)

```
# find clustering information in an expresion data using CORE
day5 <- sample2
cellnames <- colnames(day5$dat5_counts)
cluster <-day5$dat5_clusters
cellnames <-data.frame("Cluster"=cluster, "cellBarcodes" = cellnames)
mixedpop2 <-NewscGPS(ExpressionMatrix = day5$dat5_counts, GeneMetadata = day5$dat5geneInfo, CellMetadata

CORE_cluster <- CORE_scGPS(mixedpop2, remove_outlier = c(0), PCA=FALSE)
#> [1] "Performing 1 round of filtering"
#> [1] "Identifying top variable genes"
```

```
#> [1] "Calculating distance matrix"
#> [1] "Performing hierarchical clustering"
#> [1] "Finding clustering information"
#> [1] "No more outliers detected in filtering round 1"
#> [1] "Identifying top variable genes"
#> [1] "Calculating distance matrix"
#> [1] "Performing hierarchical clustering"
#> [1] "Finding clustering information"
#> [1] "writing clustering result for run 1"
#> [1] "writing clustering result for run 2"
#> [1] "writing clustering result for run 3"
#> [1] "writing clustering result for run 4"
#> [1] "writing clustering result for run 5"
#> [1] "writing clustering result for run 6"
#> [1] "writing clustering result for run 7"
#> [1] "writing clustering result for run 8"
#> [1] "writing clustering result for run 9"
#> [1] "writing clustering result for run 10"
#> [1] "writing clustering result for run 11"
#> [1] "writing clustering result for run 12"
#> [1] "writing clustering result for run 13"
#> [1] "writing clustering result for run 14"
#> [1] "writing clustering result for run 15"
#> [1] "writing clustering result for run 16"
#> [1] "writing clustering result for run 17"
#> [1] "writing clustering result for run 18"
#> [1] "writing clustering result for run 19"
#> [1] "writing clustering result for run 20"
#> [1] "writing clustering result for run 21"
#> [1] "writing clustering result for run 22"
#> [1] "writing clustering result for run 23"
#> [1] "writing clustering result for run 24"
#> [1] "writing clustering result for run 25"
#> [1] "writing clustering result for run 26"
#> [1] "writing clustering result for run 27"
#> [1] "writing clustering result for run 28"
#> [1] "writing clustering result for run 29"
#> [1] "writing clustering result for run 30"
#> [1] "writing clustering result for run 31"
#> [1] "writing clustering result for run 32"
#> [1] "writing clustering result for run 33"
#> [1] "writing clustering result for run 34"
#> [1] "writing clustering result for run 35"
#> [1] "writing clustering result for run 36"
#> [1] "writing clustering result for run 37"
#> [1] "writing clustering result for run 38"
#> [1] "writing clustering result for run 39"
#> [1] "writing clustering result for run 40"
#> [1] "Done clustering, moving to stability calculation..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done finding optimal clustering..."
```

3.1 Identify clusters in a dataset using SCORE (Stable Clustering at Optimal REsolution)

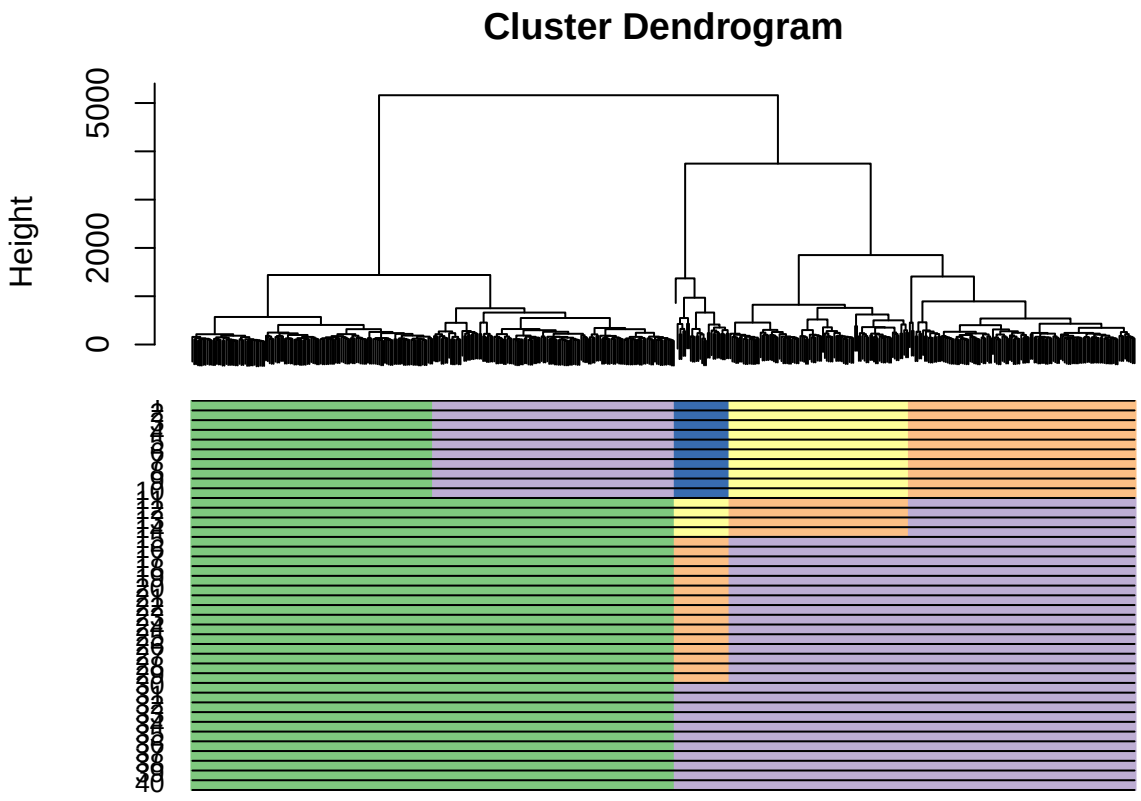
(skip this step if clusters are known) (SCORE aims to get stable subpopulation results, by introducing bagging aggregation and bootstrapping to the CORE algorithm)

```
# find clustering information in an expresion data using SCORE
day5 <- sample2
cellnames <- colnames(day5$dat5_counts)
cluster <-day5$dat5_clusters
cellnames <-data.frame("Cluster"=cluster, "cellBarcodes" = cellnames)
mixedpop2 <-NewscGPS(ExpressionMatrix = day5$dat5_counts, GeneMetadata = day5$dat5geneInfo, CellMetadata

SCORE_test <- CORE_scGPS_bagging(mixedpop2, remove_outlier = c(0), PCA=FALSE,
                           bagging_run = 20, subsample_proportion = .8)
#> [1] "Performing 1 round of filtering"
#> [1] "Identifying top variable genes"
#> [1] "Calculating distance matrix"
#> [1] "Performing hierarchical clustering"
#> [1] "Finding clustering information"
#> [1] "No more outliers detected in filtering round 1"
#> [1] "Identifying top variable genes"
#> [1] "Calculating distance matrix"
#> [1] "Performing hierarchical clustering"
#> [1] "Finding clustering information"
#> [1] "500 cells left after filtering"
#> [1] "Running 20 bagging runs, with 0.8 subsampling..."
#> [1] "Done clustering, moving to stability calculation..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
```

```
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
#> [1] "Done calculating stability..."
#> [1] "Start finding optimal clustering..."
```

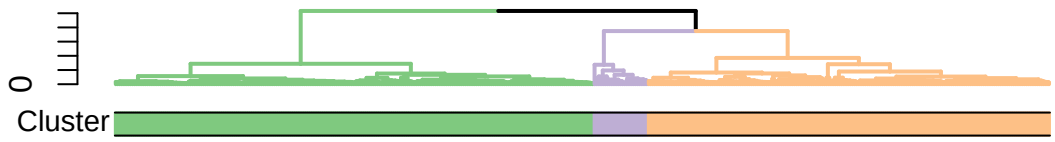
3.2 Visualise all cluster results in all iterations

```
##3.2.1 plot CORE clustering
plot_CORE(CORE_cluster$tree, CORE_cluster$Cluster) #plot all clustering bars
```

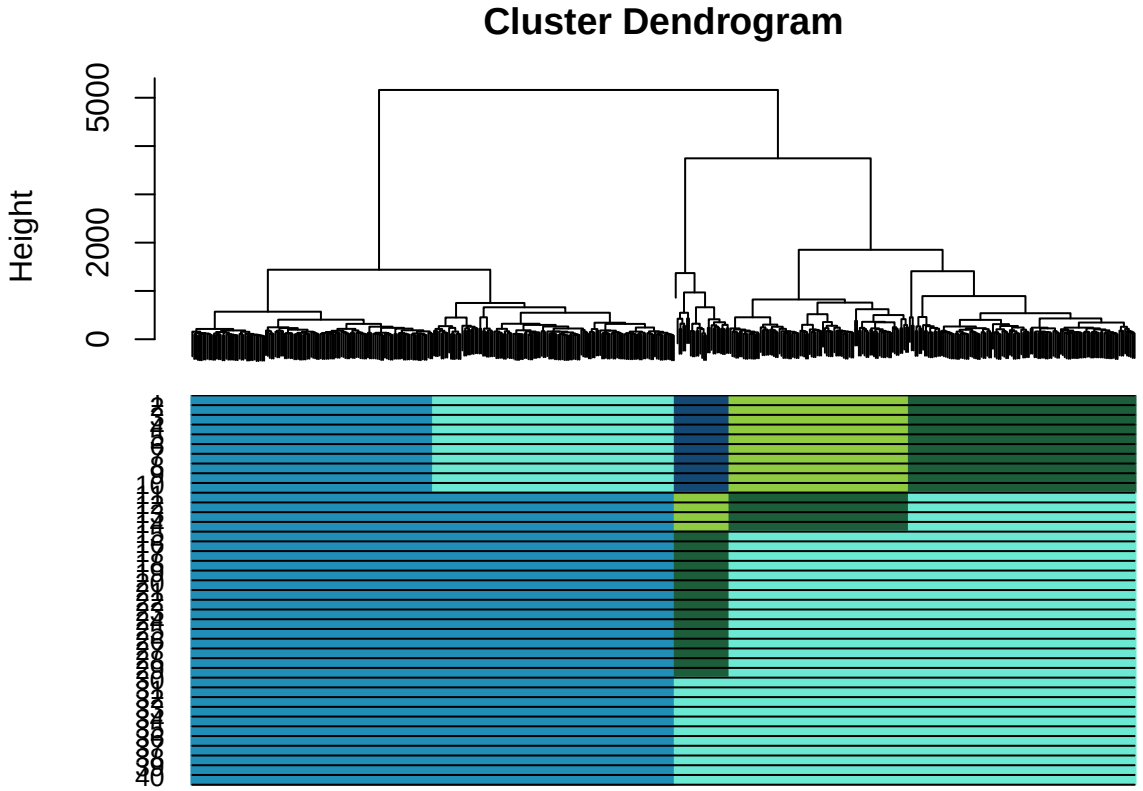



```
#extract optimal index identified by CORE_scGPS
key_height <- CORE_cluster$optimalClust$KeyStats$Height
optimal_res <- CORE_cluster$optimalClust$OptimalRes
optimal_index = which(key_height == optimal_res)
#plot one optimal clustering bar
plot_optimal_CORE(original_tree= CORE_cluster$tree,
                  optimal_cluster = unlist(CORE_cluster$Cluster[optimal_index]), shift = -2000)
```

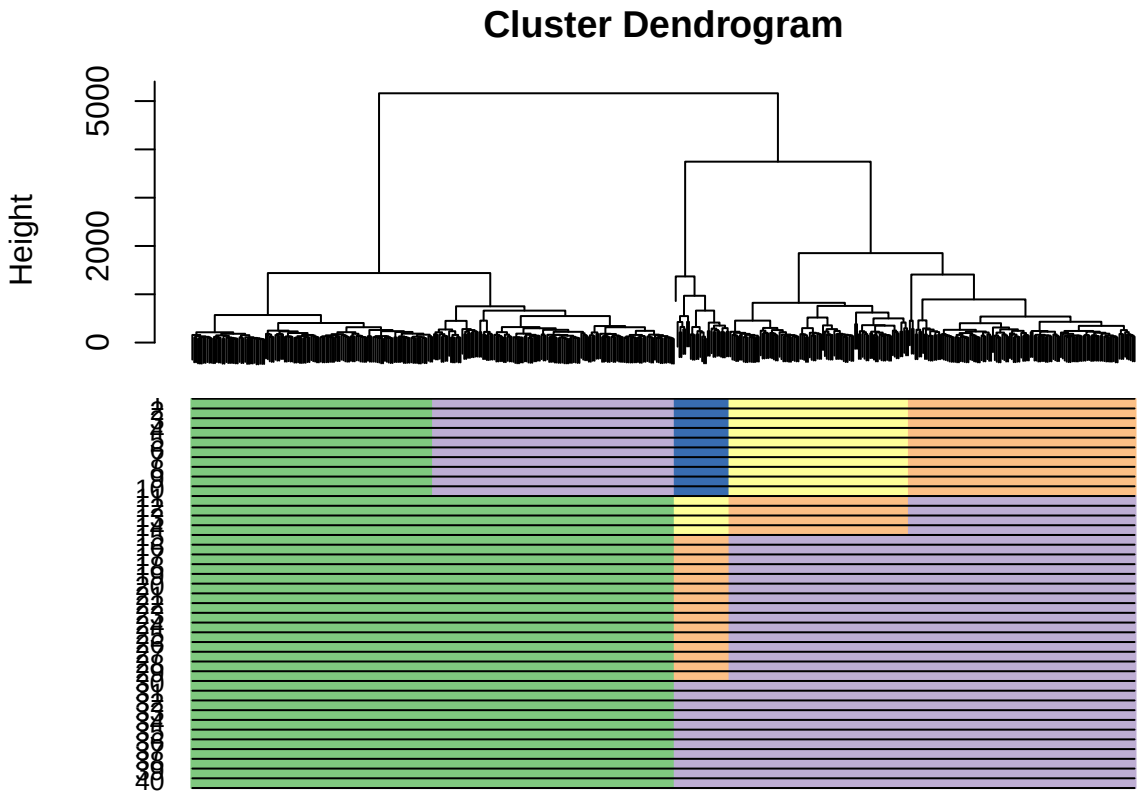
```
#> [1] "Ordering and assigning labels..."
#> [1] 2
#> [1] 128 270  NA
#> [1] 3
#> [1] 128 270 393
#> [1] "Plotting the colored dendrogram now...."
```



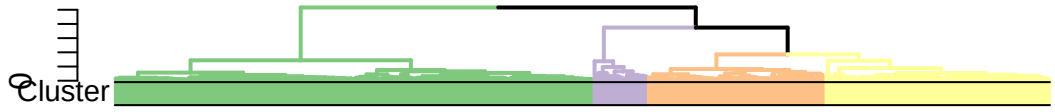

```
#> [1] "Plotting the bar underneath now...."
# you can customise the cluster color bars (provide color_branch values)
plot_CORE(CORE_cluster$tree, CORE_cluster$Cluster, color_branch = c("#208eb7", "#6ce9d3", "#1c5e39", "#8
```




```
##3.2.2 plot SCORE clustering
plot_CORE(SCORE_test$tree, list_clusters = SCORE_test$Cluster)#plot all clustering bars
```



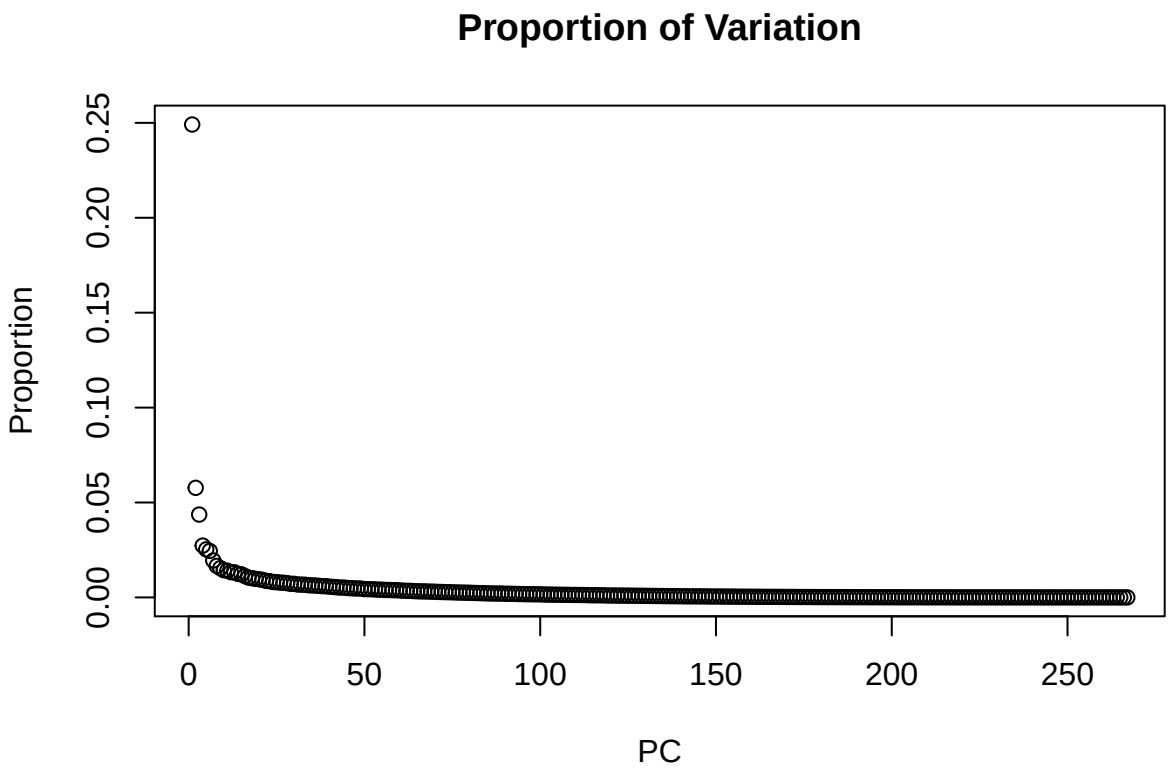

```
#plot one stable optimal clustering bar
plot_optimal_CORE(original_tree= SCORE_test$tree,
                  optimal_cluster = unlist(SCORE_test$Cluster[SCORE_test$optimal_index]), shift = -100)
#> [1] "Ordering and assigning labels..."
#> [1] 2
#> [1] 128 270  NA  NA
#> [1] 3
#> [1] 128 270 333  NA
#> [1] 4
#> [1] 128 270 333 440
#> [1] "Plotting the colored dendrogram now...."
```



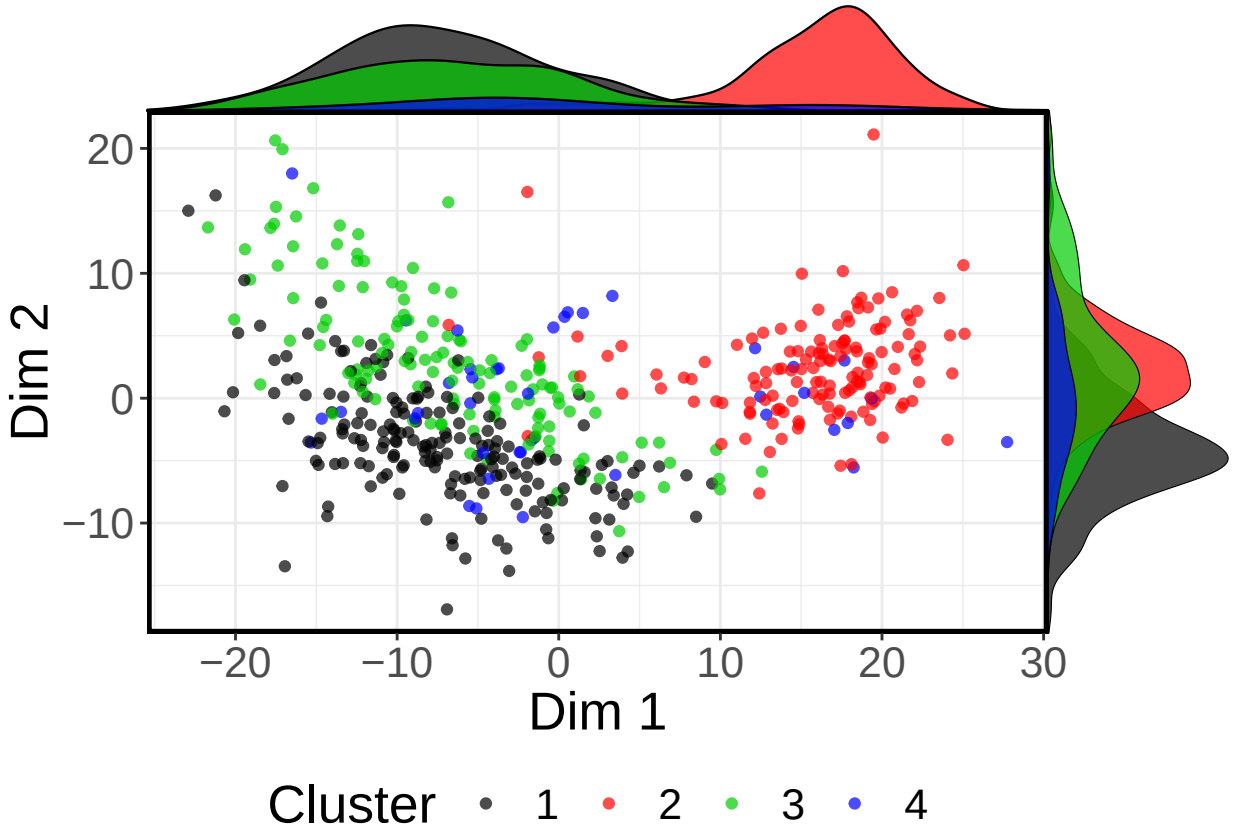

```
#> [1] "Plotting the bar underneath now...."
```

3.4 Compare clustering results with other dimensional reduction methods (e.g., CIDR)

```
library(cidr)
t <- CIDR_scGPS(expression.matrix=assay(mixedpop2))
#> [1] "building cidr object..."
#> [1] "determine dropout candidates..."
#> [1] "determine the imputation weighting threshold..."
#> [1] "computes the _CIDR_ dissimilarity matrix..."
#> [1] "PCA plot with proportion of variance explained..."
```



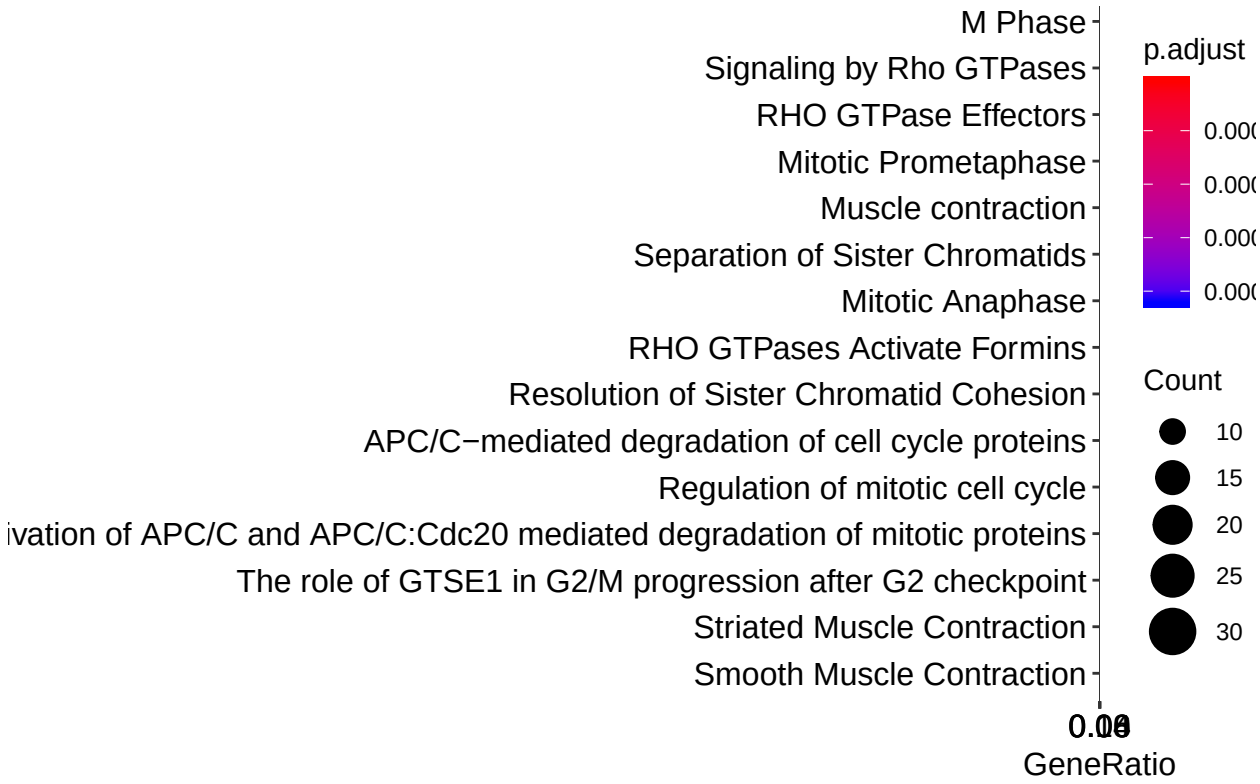

```
#> [1] "find the number of PC..."
#> [1] "perform clustering..."
p2 <-plotReduced_scGPS(t, color_fac = factor(colData(mixedpop2)[,1]),palletes =1:length(unique(colData(m
p2
```



3.5 Find gene markers and annotate clusters

```
#load gene list (this can be any lists of user-selected genes)
genes <-GeneList
genes <-genes$Merged_unique

#the gene list can also be objectively identified by differential expression analysis
#cluster information is requied for findMarkers_scGPS. Here, we use CORE results.

#colData(mixedpop2)[,1] <- unlist(SCORE_test$Cluster[SCORE_test$optimal_index])

suppressMessages(library(locfit))
suppressMessages(library(DESeq))

DEgenes <- findMarkers_scGPS(expression_matrix=assay(mixedpop2), cluster = colData(mixedpop2)[,1],
                             selected_cluster=unique(colData(mixedpop2)[,1]))
#> [1] "Start estimate dispersions for cluster 1..."
#> [1] "Done estimate dispersions. Start nbinom test for cluster 1..."
#> [1] "Done nbinom test for cluster 1 ..."
#> [1] "Adjust foldchange by subtracting basemean to 1..."
#> [1] "Start estimate dispersions for cluster 2..."
#> [1] "Done estimate dispersions. Start nbinom test for cluster 2..."
#> [1] "Done nbinom test for cluster 2 ..."
#> [1] "Adjust foldchange by subtracting basemean to 1..."
#> [1] "Start estimate dispersions for cluster 3..."
#> [1] "Done estimate dispersions. Start nbinom test for cluster 3..."
```

```
#> [1] "Done nbinom test for cluster 3 ..."
#> [1] "Adjust foldchange by subtracting basemean to 1..."
#> [1] "Start estimate dispersions for cluster 4..."
#> [1] "Done estimate dispersions. Start nbinom test for cluster 4..."
#> [1] "Done nbinom test for cluster 4 ..."
#> [1] "Adjust foldchange by subtracting basemean to 1..."

#the output contains dataframes for each cluster.
#the data frame contains all genes, sorted by p-values
names(DEgenes)
#> [1] "DE_Subpop1vsRemaining" "DE_Subpop2vsRemaining" "DE_Subpop3vsRemaining"
#> [4] "DE_Subpop4vsRemaining"

#you can annotate the identified clusters
DEgeneList_1vsOthers <- DEgenes$DE_Subpop1vsRemaining$id

#users need to check the format of the gene input to make sure they are consistent to
#the gene names in the expression matrix

#the following command saves the file "PathwayEnrichment.xlsx" to the working dir
#use 500 top DE genes
suppressMessages(library(DOSE))
suppressMessages(library(ReactomePA))
suppressMessages(library(clusterProfiler))
enrichment_test <- annotate_scGPS(DEgeneList_1vsOthers[1:500], pvalueCutoff=0.05, gene_symbol=TRUE)
#> [1] "Original gene number in geneList"
#> [1] 500
#> [1] "Number of genes successfully converted"
#> [1] 487

#the enrichment outputs can be displayed by running
dotplot(enrichment_test, showCategory=15)
```



4. Relationship between clusters within one sample or between two samples

The purpose of this workflow is to solve the following task: given one or two unknown mixed population(s) and clusters in each mixed population, estimate and visualise relationship between clusters

4.1 Start the scGPS prediction to find relationship between clusters

```
#select a subpopulation, and input gene list
c_selectID <- 1
#note make sure the format for genes input here is the same to the format for genes in the mixedpop1 an
genes = DEgenes$DE_Subpop1vsRemaining$id[1:500]

#run the test bootstrap with nboots = 2 runs

cluster_mixedpop1 <- colData(mixedpop1)[,1]
cluster_mixedpop2 <- colData(mixedpop2)[,1]

sink("temp")
LSOLDA_dat <- bootstrap_scGPS(nboots = 2, mixedpop1 = mixedpop1,
    mixedpop2 = mixedpop2, genes = genes, c_selectID  = c_selectID,
     listData = list(),
     cluster_mixedpop1 = cluster_mixedpop1,
     cluster_mixedpop2 = cluster_mixedpop2)
#> [1] "Total 224 cells as source subpop"
```

```
#> [1] "Total 366 cells in remaining subpops"
#> [1] "subsampling 112 cells for training souce subpop"
#> [1] "subsampling 112 cells in remaining subpops for training"
#> [1] "use 482 genes for training model"
#> [1] "use 482 genes 224 cells for testing model"
#> [1] "rename remaining subpops to 2_3"
#> [1] "there are 112 cells in class 2_3 and 112 cells in class 1"
#> [1] "removing 5 genes with no variance"
#> [1] "standardizing prediction/target dataset"
#> [1] "performming elasticnet model training..."
#> [1] "performming LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>        Df       %Dev   Lambda
#>   [1,]  0 -2.563e-15 0.310200
#>   [2,]  1  2.469e-02 0.296100
#>   [3,]  1  4.734e-02 0.282600
#>   [4,]  1  6.824e-02 0.269800
#>   [5,]  2  9.416e-02 0.257500
#>   [6,]  2  1.185e-01 0.245800
#>   [7,]  2  1.414e-01 0.234600
#>   [8,]  3  1.632e-01 0.224000
#>   [9,]  4  1.865e-01 0.213800
#>  [10,]  5  2.092e-01 0.204100
#>  [11,]  7  2.326e-01 0.194800
#>  [12,]  7  2.549e-01 0.186000
#>  [13,]  7  2.758e-01 0.177500
#>  [14,]  7  2.956e-01 0.169400
#>  [15,]  8  3.143e-01 0.161700
#>  [16,]  8  3.323e-01 0.154400
#>  [17,]  8  3.492e-01 0.147400
#>  [18,]  9  3.658e-01 0.140700
#>  [19,]  9  3.819e-01 0.134300
#>  [20,] 10  3.972e-01 0.128200
#>  [21,] 10  4.117e-01 0.122300
#>  [22,] 10  4.256e-01 0.116800
#>  [23,] 10  4.389e-01 0.111500
#>  [24,] 11  4.515e-01 0.106400
#>  [25,] 12  4.643e-01 0.101600
#>  [26,] 12  4.767e-01 0.096950
#>  [27,] 12  4.885e-01 0.092550
#>  [28,] 12  4.998e-01 0.088340
#>  [29,] 14  5.111e-01 0.084330
#>  [30,] 14  5.225e-01 0.080490
#>  [31,] 16  5.340e-01 0.076830
#>  [32,] 18  5.459e-01 0.073340
#>  [33,] 22  5.586e-01 0.070010
#>  [34,] 23  5.716e-01 0.066830
#>  [35,] 25  5.840e-01 0.063790
#>  [36,] 25  5.962e-01 0.060890
#>  [37,] 24  6.078e-01 0.058120
```

```
#>  [38,] 24  6.189e-01 0.055480
#>  [39,] 26  6.296e-01 0.052960
#>  [40,] 29  6.403e-01 0.050550
#>  [41,] 29  6.506e-01 0.048250
#>  [42,] 30  6.604e-01 0.046060
#>  [43,] 31  6.699e-01 0.043970
#>  [44,] 31  6.791e-01 0.041970
#>  [45,] 34  6.882e-01 0.040060
#>  [46,] 35  6.973e-01 0.038240
#>  [47,] 37  7.063e-01 0.036500
#>  [48,] 37  7.152e-01 0.034840
#>  [49,] 37  7.237e-01 0.033260
#>  [50,] 38  7.319e-01 0.031750
#>  [51,] 39  7.402e-01 0.030310
#>  [52,] 39  7.482e-01 0.028930
#>  [53,] 40  7.559e-01 0.027610
#>  [54,] 41  7.636e-01 0.026360
#>  [55,] 42  7.712e-01 0.025160
#>  [56,] 42  7.785e-01 0.024020
#>  [57,] 45  7.858e-01 0.022920
#>  [58,] 47  7.933e-01 0.021880
#>  [59,] 49  8.008e-01 0.020890
#>  [60,] 53  8.084e-01 0.019940
#>  [61,] 53  8.161e-01 0.019030
#>  [62,] 54  8.234e-01 0.018170
#>  [63,] 56  8.306e-01 0.017340
#>  [64,] 58  8.375e-01 0.016550
#>  [65,] 58  8.443e-01 0.015800
#>  [66,] 59  8.510e-01 0.015080
#>  [67,] 58  8.573e-01 0.014400
#>  [68,] 59  8.635e-01 0.013740
#>  [69,] 59  8.694e-01 0.013120
#>  [70,] 61  8.750e-01 0.012520
#>  [71,] 63  8.805e-01 0.011950
#>  [72,] 64  8.858e-01 0.011410
#>  [73,] 68  8.909e-01 0.010890
#>  [74,] 70  8.958e-01 0.010400
#>  [75,] 70  9.005e-01 0.009924
#>  [76,] 72  9.050e-01 0.009473
#>  [77,] 74  9.093e-01 0.009042
#>  [78,] 76  9.134e-01 0.008631
#>  [79,] 76  9.173e-01 0.008239
#>  [80,] 77  9.211e-01 0.007864
#>  [81,] 77  9.246e-01 0.007507
#>  [82,] 79  9.280e-01 0.007166
#>  [83,] 80  9.313e-01 0.006840
#>  [84,] 80  9.344e-01 0.006529
#>  [85,] 82  9.374e-01 0.006232
#>  [86,] 82  9.403e-01 0.005949
#>  [87,] 81  9.430e-01 0.005679
#>  [88,] 81  9.456e-01 0.005421
#>  [89,] 81  9.481e-01 0.005174
#>  [90,] 80  9.504e-01 0.004939
```

```
#>  [91,] 81  9.527e-01 0.004715
#>  [92,] 81  9.548e-01 0.004500
#>  [93,] 80  9.569e-01 0.004296
#>  [94,] 81  9.589e-01 0.004100
#>  [95,] 81  9.607e-01 0.003914
#>  [96,] 82  9.625e-01 0.003736
#>  [97,] 83  9.642e-01 0.003566
#>  [98,] 84  9.659e-01 0.003404
#>  [99,] 84  9.674e-01 0.003250
#> [100,] 84  9.689e-01 0.003102
#> [1] "lambda min is at location 47"
#> [1] "the leave-out cells in the source subpop is 112"
#> [1] "use 112 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 224 cells and 477 genes..."
#> [1] "evaluation accuracy ElasticNet 0.852678571428571"
#> [1] "evaluation accuracy LDA 0.848214285714286"
#> [1] "done bootstrap 1"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 73.7967914438503"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 38.5026737967914"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 66.4285714285714"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 82.8571428571429"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 30.0751879699248"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 35.3383458646617"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 57.5"
```

```
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 55"
#> [1] "Total 224 cells as source subpop"
#> [1] "Total 366 cells in remaining subpops"
#> [1] "subsampling 112 cells for training souce subpop"
#> [1] "subsampling 112 cells in remaining subpops for training"
#> [1] "use 482 genes for training model"
#> [1] "use 482 genes 224 cells for testing model"
#> [1] "rename remaining subpops to 2_3"
#> [1] "there are 112 cells in class 2_3 and 112 cells in class 1"
#> [1] "removing 7 genes with no variance"
#> [1] "standardizing prediction/target dataset"
#> [1] "performning elasticnet model training..."
#> [1] "performning LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>        Df       %Dev   Lambda
#>   [1,]  0 -2.563e-15 0.292900
#>   [2,]  1  2.202e-02 0.279600
#>   [3,]  1  4.223e-02 0.266900
#>   [4,]  1  6.089e-02 0.254800
#>   [5,]  1  7.822e-02 0.243200
#>   [6,]  3  1.007e-01 0.232100
#>   [7,]  5  1.242e-01 0.221600
#>   [8,]  6  1.492e-01 0.211500
#>   [9,]  7  1.747e-01 0.201900
#>  [10,]  7  1.994e-01 0.192700
#>  [11,]  7  2.224e-01 0.184000
#>  [12,]  7  2.438e-01 0.175600
#>  [13,]  8  2.648e-01 0.167600
#>  [14,]  8  2.845e-01 0.160000
#>  [15,]  8  3.029e-01 0.152700
#>  [16,]  9  3.205e-01 0.145800
#>  [17,]  9  3.371e-01 0.139200
#>  [18,]  9  3.527e-01 0.132800
#>  [19,] 10  3.678e-01 0.126800
#>  [20,] 12  3.839e-01 0.121000
#>  [21,] 12  4.005e-01 0.115500
#>  [22,] 12  4.162e-01 0.110300
#>  [23,] 14  4.322e-01 0.105300
#>  [24,] 16  4.479e-01 0.100500
#>  [25,] 21  4.647e-01 0.095920
#>  [26,] 22  4.830e-01 0.091560
#>  [27,] 23  5.007e-01 0.087400
#>  [28,] 23  5.174e-01 0.083430
#>  [29,] 23  5.333e-01 0.079630
#>  [30,] 24  5.488e-01 0.076010
#>  [31,] 24  5.637e-01 0.072560
#>  [32,] 24  5.779e-01 0.069260
#>  [33,] 24  5.917e-01 0.066110
```

```
#>  [34,] 24  6.048e-01 0.063110
#>  [35,] 25  6.174e-01 0.060240
#>  [36,] 27  6.296e-01 0.057500
#>  [37,] 27  6.416e-01 0.054890
#>  [38,] 27  6.530e-01 0.052390
#>  [39,] 27  6.639e-01 0.050010
#>  [40,] 27  6.744e-01 0.047740
#>  [41,] 28  6.844e-01 0.045570
#>  [42,] 28  6.942e-01 0.043500
#>  [43,] 32  7.038e-01 0.041520
#>  [44,] 33  7.136e-01 0.039630
#>  [45,] 33  7.233e-01 0.037830
#>  [46,] 33  7.325e-01 0.036110
#>  [47,] 33  7.414e-01 0.034470
#>  [48,] 34  7.501e-01 0.032900
#>  [49,] 34  7.586e-01 0.031410
#>  [50,] 34  7.669e-01 0.029980
#>  [51,] 36  7.751e-01 0.028620
#>  [52,] 36  7.831e-01 0.027320
#>  [53,] 36  7.909e-01 0.026080
#>  [54,] 40  7.985e-01 0.024890
#>  [55,] 41  8.063e-01 0.023760
#>  [56,] 41  8.138e-01 0.022680
#>  [57,] 41  8.211e-01 0.021650
#>  [58,] 41  8.280e-01 0.020670
#>  [59,] 42  8.347e-01 0.019730
#>  [60,] 42  8.412e-01 0.018830
#>  [61,] 42  8.474e-01 0.017970
#>  [62,] 42  8.534e-01 0.017160
#>  [63,] 42  8.590e-01 0.016380
#>  [64,] 44  8.646e-01 0.015630
#>  [65,] 43  8.700e-01 0.014920
#>  [66,] 44  8.752e-01 0.014240
#>  [67,] 44  8.803e-01 0.013600
#>  [68,] 46  8.852e-01 0.012980
#>  [69,] 47  8.900e-01 0.012390
#>  [70,] 51  8.946e-01 0.011830
#>  [71,] 52  8.992e-01 0.011290
#>  [72,] 54  9.037e-01 0.010770
#>  [73,] 54  9.080e-01 0.010290
#>  [74,] 55  9.121e-01 0.009818
#>  [75,] 57  9.161e-01 0.009371
#>  [76,] 58  9.199e-01 0.008945
#>  [77,] 58  9.235e-01 0.008539
#>  [78,] 59  9.270e-01 0.008151
#>  [79,] 59  9.303e-01 0.007780
#>  [80,] 59  9.335e-01 0.007427
#>  [81,] 59  9.365e-01 0.007089
#>  [82,] 60  9.394e-01 0.006767
#>  [83,] 60  9.421e-01 0.006459
#>  [84,] 61  9.448e-01 0.006166
#>  [85,] 62  9.473e-01 0.005885
#>  [86,] 62  9.497e-01 0.005618
```

```
#>  [87,] 62  9.520e-01 0.005363
#>  [88,] 62  9.542e-01 0.005119
#>  [89,] 62  9.563e-01 0.004886
#>  [90,] 62  9.583e-01 0.004664
#>  [91,] 63  9.602e-01 0.004452
#>  [92,] 63  9.620e-01 0.004250
#>  [93,] 63  9.637e-01 0.004057
#>  [94,] 63  9.654e-01 0.003872
#>  [95,] 64  9.670e-01 0.003696
#>  [96,] 64  9.685e-01 0.003528
#>  [97,] 64  9.699e-01 0.003368
#>  [98,] 66  9.713e-01 0.003215
#>  [99,] 65  9.726e-01 0.003069
#> [100,] 66  9.738e-01 0.002929
#> [1] "lambda min is at location 81"
#> [1] "the leave-out cells in the source subpop is 112"
#> [1] "use 112 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 224 cells and 475 genes..."
#> [1] "evaluation accuracy ElasticNet 0.875"
#> [1] "evaluation accuracy LDA 0.861607142857143"
#> [1] "done bootstrap 2"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 20.855614973262"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 27.27272727273"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 88.5714285714286"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 75"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 9.02255639097744"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 43.609022556391"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 42.5"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 57.5"
```

```
sink()
```

4.2 Display summary results for the prediction

```
#get the number of rows for the summary matrix
row_cluster <-length(unique(colData(mixedpop2)[,1]))

#summary results LDA to to show the percent of cells classified as cells belonging by LDA classifier
summary_prediction_lda(LSOLDA_dat=LSOLDA_dat, nPredSubpop = row_cluster )
#>                 V1               V2                                  names
#> 1 38.5026737967914 27.2727272727273 LDA for subpop 1 in target mixedpop2
#> 2 82.8571428571429               75 LDA for subpop 2 in target mixedpop2
#> 3 35.3383458646617  43.609022556391 LDA for subpop 3 in target mixedpop2
#> 4               55             57.5 LDA for subpop 4 in target mixedpop2

#summary results Lasso to show the percent of cells classified as cells belonging by Lasso classifier
summary_prediction_lasso(LSOLDA_dat=LSOLDA_dat, nPredSubpop = row_cluster)
#>                 V1               V2
#> 1 73.7967914438503  20.855614973262
#> 2 66.4285714285714 88.5714285714286
#> 3 30.0751879699248 9.02255639097744
#> 4             57.5             42.5
#>                                    names
#> 1 ElasticNet for subpop1 in target mixedpop2
#> 2 ElasticNet for subpop2 in target mixedpop2
#> 3 ElasticNet for subpop3 in target mixedpop2
#> 4 ElasticNet for subpop4 in target mixedpop2

# summary maximum deviance explained by the model during the model training
summary_deviance(object = LSOLDA_dat)
#> $allDeviance
#> [1] "0.7063" "0.9365"
#>
#> $DeviMax
#>                  Dfd          Deviance           DEgenes
#> 1                  0       -2.563e-15    genes_cluster1
#> 2                  1           0.07822    genes_cluster1
#> 3                  3            0.1007    genes_cluster1
#> 4                  5            0.1242    genes_cluster1
#> 5                  6            0.1492    genes_cluster1
#> 6                  7            0.2438    genes_cluster1
#> 7                  8            0.3029    genes_cluster1
#> 8                  9            0.3527    genes_cluster1
#> 9                 10            0.3678    genes_cluster1
#> 10                12            0.4162    genes_cluster1
#> 11                14            0.4322    genes_cluster1
#> 12                16            0.4479    genes_cluster1
#> 13                21            0.4647    genes_cluster1
#> 14                22             0.483    genes_cluster1
#> 15                23            0.5333    genes_cluster1
#> 16                24            0.6048    genes_cluster1
#> 17                25            0.6174    genes_cluster1
```

```
#> 18                27            0.6744    genes_cluster1
#> 19                28            0.6942    genes_cluster1
#> 20                32            0.7038    genes_cluster1
#> 21                33            0.7414    genes_cluster1
#> 22                34            0.7669    genes_cluster1
#> 23                36            0.7909    genes_cluster1
#> 24                40            0.7985    genes_cluster1
#> 25                41             0.828    genes_cluster1
#> 26                42             0.859    genes_cluster1
#> 27                43              0.87    genes_cluster1
#> 28                44            0.8803    genes_cluster1
#> 29                46            0.8852    genes_cluster1
#> 30                47              0.89    genes_cluster1
#> 31                51            0.8946    genes_cluster1
#> 32                52            0.8992    genes_cluster1
#> 33                54             0.908    genes_cluster1
#> 34                55            0.9121    genes_cluster1
#> 35                57            0.9161    genes_cluster1
#> 36                58            0.9235    genes_cluster1
#> 37                59            0.9365    genes_cluster1
#> 38 remaining DEgenes remaining DEgenes remaining DEgenes
#>
#> $LassoGenesMax
#> NULL

# summary accuracy to check the model accuracy in the leave-out test set
summary_accuracy(object = LSOLDA_dat)
#> [1] 84.82143 86.16071
```

4.3 Plot the relationship between clusters in one sample

Here we look at one example use case to find relationship between clusters within one sample or between two sample

```
#run prediction for 3 clusters
cluster_mixedpop1 <- colData(mixedpop1)[,1]
cluster_mixedpop2 <- colData(mixedpop2)[,1]
#cluster_mixedpop2 <- as.numeric(as.vector(colData(mixedpop2)[,1]))

c_selectID <- 1
genes = DEgenes$DE_Subpop1vsRemaining$id[1:200] #top 200 gene markers distinguishing cluster 1

LSOLDA_dat1 <- bootstrap_scGPS(nboots = 2, mixedpop1 = mixedpop2, mixedpop2 = mixedpop2, genes=genes, c
#> [1] "Total 187 cells as source subpop"
#> [1] "Total 313 cells in remaining subpops"
#> [1] "subsampling 94 cells for training souce subpop"
#> [1] "subsampling 94 cells in remaining subpops for training"
#> [1] "use 201 genes for training model"
#> [1] "use 201 genes 188 cells for testing model"
#> [1] "rename remaining subpops to 2_3_4"
#> [1] "there are 94 cells in class 2_3_4 and 94 cells in class 1"
#> [1] "standardizing prediction/target dataset"
#> [1] "performming elasticnet model training..."
```

```
#> [1] "performming LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>        Df      %Dev   Lambda
#>   [1,]  0 3.364e-15 0.294500
#>   [2,]  1 2.226e-02 0.281200
#>   [3,]  1 4.265e-02 0.268400
#>   [4,]  1 6.142e-02 0.256200
#>   [5,]  1 7.876e-02 0.244500
#>   [6,]  1 9.485e-02 0.233400
#>   [7,]  1 1.098e-01 0.222800
#>   [8,]  1 1.238e-01 0.212700
#>   [9,]  2 1.414e-01 0.203000
#>  [10,]  3 1.593e-01 0.193800
#>  [11,]  4 1.785e-01 0.185000
#>  [12,]  4 1.970e-01 0.176600
#>  [13,]  5 2.152e-01 0.168500
#>  [14,]  5 2.324e-01 0.160900
#>  [15,]  5 2.487e-01 0.153600
#>  [16,]  5 2.640e-01 0.146600
#>  [17,]  6 2.791e-01 0.139900
#>  [18,]  7 2.939e-01 0.133600
#>  [19,]  7 3.090e-01 0.127500
#>  [20,]  8 3.237e-01 0.121700
#>  [21,]  9 3.379e-01 0.116200
#>  [22,]  9 3.521e-01 0.110900
#>  [23,]  8 3.656e-01 0.105900
#>  [24,]  9 3.784e-01 0.101000
#>  [25,]  9 3.906e-01 0.096450
#>  [26,] 10 4.029e-01 0.092070
#>  [27,] 10 4.155e-01 0.087880
#>  [28,] 11 4.275e-01 0.083890
#>  [29,] 11 4.397e-01 0.080070
#>  [30,] 11 4.512e-01 0.076440
#>  [31,] 12 4.623e-01 0.072960
#>  [32,] 13 4.733e-01 0.069640
#>  [33,] 14 4.842e-01 0.066480
#>  [34,] 14 4.949e-01 0.063460
#>  [35,] 16 5.055e-01 0.060570
#>  [36,] 17 5.164e-01 0.057820
#>  [37,] 19 5.270e-01 0.055190
#>  [38,] 19 5.373e-01 0.052680
#>  [39,] 21 5.486e-01 0.050290
#>  [40,] 21 5.601e-01 0.048000
#>  [41,] 22 5.712e-01 0.045820
#>  [42,] 22 5.821e-01 0.043740
#>  [43,] 23 5.927e-01 0.041750
#>  [44,] 23 6.032e-01 0.039850
#>  [45,] 26 6.142e-01 0.038040
#>  [46,] 27 6.253e-01 0.036310
#>  [47,] 27 6.361e-01 0.034660
```

```
#>  [48,] 30 6.469e-01 0.033090
#>  [49,] 31 6.575e-01 0.031580
#>  [50,] 31 6.676e-01 0.030150
#>  [51,] 32 6.774e-01 0.028780
#>  [52,] 33 6.868e-01 0.027470
#>  [53,] 34 6.962e-01 0.026220
#>  [54,] 34 7.053e-01 0.025030
#>  [55,] 34 7.142e-01 0.023890
#>  [56,] 34 7.226e-01 0.022810
#>  [57,] 38 7.310e-01 0.021770
#>  [58,] 38 7.396e-01 0.020780
#>  [59,] 39 7.478e-01 0.019840
#>  [60,] 40 7.561e-01 0.018930
#>  [61,] 42 7.646e-01 0.018070
#>  [62,] 42 7.731e-01 0.017250
#>  [63,] 42 7.811e-01 0.016470
#>  [64,] 42 7.888e-01 0.015720
#>  [65,] 43 7.963e-01 0.015000
#>  [66,] 45 8.035e-01 0.014320
#>  [67,] 47 8.106e-01 0.013670
#>  [68,] 47 8.176e-01 0.013050
#>  [69,] 47 8.243e-01 0.012460
#>  [70,] 47 8.308e-01 0.011890
#>  [71,] 47 8.370e-01 0.011350
#>  [72,] 48 8.430e-01 0.010830
#>  [73,] 48 8.489e-01 0.010340
#>  [74,] 49 8.545e-01 0.009872
#>  [75,] 50 8.599e-01 0.009423
#>  [76,] 51 8.651e-01 0.008995
#>  [77,] 52 8.703e-01 0.008586
#>  [78,] 52 8.753e-01 0.008196
#>  [79,] 53 8.802e-01 0.007823
#>  [80,] 53 8.849e-01 0.007468
#>  [81,] 53 8.895e-01 0.007128
#>  [82,] 53 8.939e-01 0.006804
#>  [83,] 53 8.981e-01 0.006495
#>  [84,] 53 9.022e-01 0.006200
#>  [85,] 54 9.061e-01 0.005918
#>  [86,] 55 9.099e-01 0.005649
#>  [87,] 55 9.136e-01 0.005392
#>  [88,] 55 9.171e-01 0.005147
#>  [89,] 55 9.205e-01 0.004913
#>  [90,] 55 9.238e-01 0.004690
#>  [91,] 55 9.270e-01 0.004477
#>  [92,] 55 9.300e-01 0.004273
#>  [93,] 55 9.329e-01 0.004079
#>  [94,] 55 9.358e-01 0.003894
#>  [95,] 57 9.385e-01 0.003717
#>  [96,] 58 9.412e-01 0.003548
#>  [97,] 58 9.438e-01 0.003387
#>  [98,] 58 9.462e-01 0.003233
#>  [99,] 58 9.486e-01 0.003086
#> [100,] 58 9.509e-01 0.002945
```

```
#> [1] "lambda min is at location 85"
#> [1] "the leave-out cells in the source subpop is 93"
#> [1] "use 94 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 187 cells and 201 genes..."
#> [1] "evaluation accuracy ElasticNet 0.877005347593583"
#> [1] "evaluation accuracy LDA 0.491978609625668"
#> [1] "done bootstrap 1"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 95.1871657754011"
#> [1] "add 7 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 27.807486631016"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 2.85714285714286"
#> [1] "add 7 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 45"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 24.0601503759398"
#> [1] "add 7 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 34.5864661654135"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 27.5"
#> [1] "add 7 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 45"
#> [1] "Total 187 cells as source subpop"
#> [1] "Total 313 cells in remaining subpops"
#> [1] "subsampling 94 cells for training souce subpop"
#> [1] "subsampling 94 cells in remaining subpops for training"
#> [1] "use 201 genes for training model"
#> [1] "use 201 genes 188 cells for testing model"
#> [1] "rename remaining subpops to 2_3_4"
```

```
#> [1] "there are 94 cells in class 2_3_4 and 94 cells in class 1"
#> [1] "standardizing prediction/target dataset"
#> [1] "performming elasticnet model training..."
#> [1] "performming LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>        Df      %Dev   Lambda
#>   [1,]  0 3.364e-15 0.270400
#>   [2,]  1 1.876e-02 0.258100
#>   [3,]  2 3.608e-02 0.246400
#>   [4,]  2 5.671e-02 0.235200
#>   [5,]  3 7.789e-02 0.224500
#>   [6,]  3 1.019e-01 0.214300
#>   [7,]  3 1.241e-01 0.204500
#>   [8,]  3 1.449e-01 0.195200
#>   [9,]  3 1.642e-01 0.186400
#>  [10,]  3 1.824e-01 0.177900
#>  [11,]  3 1.994e-01 0.169800
#>  [12,]  4 2.158e-01 0.162100
#>  [13,]  5 2.341e-01 0.154700
#>  [14,]  6 2.518e-01 0.147700
#>  [15,]  6 2.689e-01 0.141000
#>  [16,]  8 2.864e-01 0.134600
#>  [17,]  8 3.033e-01 0.128500
#>  [18,]  8 3.194e-01 0.122600
#>  [19,] 10 3.349e-01 0.117000
#>  [20,] 12 3.503e-01 0.111700
#>  [21,] 12 3.654e-01 0.106600
#>  [22,] 12 3.798e-01 0.101800
#>  [23,] 12 3.936e-01 0.097170
#>  [24,] 12 4.068e-01 0.092760
#>  [25,] 12 4.194e-01 0.088540
#>  [26,] 12 4.315e-01 0.084520
#>  [27,] 12 4.430e-01 0.080680
#>  [28,] 14 4.543e-01 0.077010
#>  [29,] 14 4.652e-01 0.073510
#>  [30,] 15 4.770e-01 0.070170
#>  [31,] 17 4.888e-01 0.066980
#>  [32,] 17 5.006e-01 0.063930
#>  [33,] 17 5.119e-01 0.061030
#>  [34,] 17 5.229e-01 0.058250
#>  [35,] 18 5.337e-01 0.055610
#>  [36,] 18 5.441e-01 0.053080
#>  [37,] 19 5.542e-01 0.050670
#>  [38,] 22 5.652e-01 0.048360
#>  [39,] 25 5.771e-01 0.046170
#>  [40,] 26 5.893e-01 0.044070
#>  [41,] 27 6.016e-01 0.042060
#>  [42,] 27 6.135e-01 0.040150
#>  [43,] 27 6.250e-01 0.038330
#>  [44,] 27 6.361e-01 0.036590
```

```
#>  [45,] 27 6.468e-01 0.034920
#>  [46,] 28 6.577e-01 0.033340
#>  [47,] 30 6.684e-01 0.031820
#>  [48,] 31 6.791e-01 0.030370
#>  [49,] 31 6.895e-01 0.028990
#>  [50,] 31 6.994e-01 0.027680
#>  [51,] 30 7.090e-01 0.026420
#>  [52,] 31 7.182e-01 0.025220
#>  [53,] 33 7.272e-01 0.024070
#>  [54,] 33 7.360e-01 0.022980
#>  [55,] 33 7.446e-01 0.021930
#>  [56,] 33 7.529e-01 0.020940
#>  [57,] 34 7.610e-01 0.019980
#>  [58,] 35 7.688e-01 0.019080
#>  [59,] 35 7.763e-01 0.018210
#>  [60,] 35 7.835e-01 0.017380
#>  [61,] 35 7.905e-01 0.016590
#>  [62,] 35 7.973e-01 0.015840
#>  [63,] 35 8.038e-01 0.015120
#>  [64,] 38 8.104e-01 0.014430
#>  [65,] 37 8.168e-01 0.013770
#>  [66,] 37 8.231e-01 0.013150
#>  [67,] 36 8.291e-01 0.012550
#>  [68,] 36 8.349e-01 0.011980
#>  [69,] 36 8.405e-01 0.011440
#>  [70,] 36 8.459e-01 0.010920
#>  [71,] 38 8.515e-01 0.010420
#>  [72,] 39 8.570e-01 0.009946
#>  [73,] 40 8.627e-01 0.009494
#>  [74,] 40 8.682e-01 0.009063
#>  [75,] 41 8.735e-01 0.008651
#>  [76,] 41 8.786e-01 0.008257
#>  [77,] 41 8.835e-01 0.007882
#>  [78,] 40 8.883e-01 0.007524
#>  [79,] 40 8.928e-01 0.007182
#>  [80,] 40 8.972e-01 0.006855
#>  [81,] 40 9.014e-01 0.006544
#>  [82,] 41 9.055e-01 0.006246
#>  [83,] 41 9.094e-01 0.005963
#>  [84,] 41 9.131e-01 0.005692
#>  [85,] 42 9.167e-01 0.005433
#>  [86,] 42 9.202e-01 0.005186
#>  [87,] 42 9.235e-01 0.004950
#>  [88,] 42 9.267e-01 0.004725
#>  [89,] 42 9.298e-01 0.004510
#>  [90,] 42 9.327e-01 0.004305
#>  [91,] 42 9.356e-01 0.004110
#>  [92,] 42 9.383e-01 0.003923
#>  [93,] 42 9.409e-01 0.003745
#>  [94,] 42 9.434e-01 0.003574
#>  [95,] 41 9.458e-01 0.003412
#>  [96,] 42 9.480e-01 0.003257
#>  [97,] 42 9.502e-01 0.003109
```

```
#>  [98,] 42 9.523e-01 0.002968
#>  [99,] 42 9.542e-01 0.002833
#> [100,] 40 9.562e-01 0.002704
#> [1] "lambda min is at location 95"
#> [1] "the leave-out cells in the source subpop is 93"
#> [1] "use 94 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 187 cells and 201 genes..."
#> [1] "evaluation accuracy ElasticNet 0.893048128342246"
#> [1] "evaluation accuracy LDA 0.60427807486631"
#> [1] "done bootstrap 2"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 95.1871657754011"
#> [1] "add 7 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 62.0320855614973"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 1.42857142857143"
#> [1] "add 7 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 45.7142857142857"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 21.8045112781955"
#> [1] "add 7 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 39.0977443609023"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 50"
#> [1] "add 7 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 37.5"

c_selectID <- 2
genes = DEgenes$DE_Subpop2vsRemaining$id[1:200]
```

```
LSOLDA_dat2 <- bootstrap_scGPS(nboots = 2,mixedpop1 = mixedpop2, mixedpop2 = mixedpop2, genes=genes, c_s
     cluster_mixedpop2 = cluster_mixedpop2)
#> [1] "Total 140 cells as source subpop"
#> [1] "Total 360 cells in remaining subpops"
#> [1] "subsampling 70 cells for training souce subpop"
#> [1] "subsampling 70 cells in remaining subpops for training"
#> [1] "use 201 genes for training model"
#> [1] "use 201 genes 140 cells for testing model"
#> [1] "rename remaining subpops to 1_3_4"
#> [1] "there are 70 cells in class 1_3_4 and 70 cells in class 2"
#> [1] "standardizing prediction/target dataset"
#> [1] "performming elasticnet model training..."
#> [1] "performning LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>        Df      %Dev   Lambda
#>   [1,]  0 2.723e-15 0.408700
#>   [2,]  1 4.287e-02 0.390100
#>   [3,]  1 8.221e-02 0.372400
#>   [4,]  2 1.193e-01 0.355500
#>   [5,]  3 1.584e-01 0.339300
#>   [6,]  3 1.960e-01 0.323900
#>   [7,]  3 2.310e-01 0.309200
#>   [8,]  3 2.636e-01 0.295100
#>   [9,]  4 2.942e-01 0.281700
#>  [10,]  4 3.234e-01 0.268900
#>  [11,]  4 3.508e-01 0.256700
#>  [12,]  4 3.765e-01 0.245000
#>  [13,]  4 4.008e-01 0.233900
#>  [14,]  4 4.237e-01 0.223300
#>  [15,]  6 4.469e-01 0.213100
#>  [16,]  6 4.694e-01 0.203400
#>  [17,]  6 4.906e-01 0.194200
#>  [18,]  6 5.107e-01 0.185400
#>  [19,]  6 5.298e-01 0.176900
#>  [20,]  6 5.479e-01 0.168900
#>  [21,]  6 5.651e-01 0.161200
#>  [22,]  6 5.815e-01 0.153900
#>  [23,]  6 5.971e-01 0.146900
#>  [24,]  6 6.119e-01 0.140200
#>  [25,]  7 6.262e-01 0.133800
#>  [26,]  7 6.399e-01 0.127800
#>  [27,]  7 6.530e-01 0.121900
#>  [28,]  7 6.655e-01 0.116400
#>  [29,]  9 6.777e-01 0.111100
#>  [30,]  9 6.895e-01 0.106100
#>  [31,] 10 7.009e-01 0.101200
#>  [32,] 10 7.118e-01 0.096640
#>  [33,] 10 7.223e-01 0.092250
#>  [34,] 10 7.323e-01 0.088060
#>  [35,] 10 7.419e-01 0.084050
```

```
#>  [36,] 10 7.512e-01 0.080230
#>  [37,] 10 7.601e-01 0.076590
#>  [38,] 10 7.686e-01 0.073110
#>  [39,] 10 7.769e-01 0.069780
#>  [40,] 10 7.848e-01 0.066610
#>  [41,] 10 7.924e-01 0.063580
#>  [42,] 11 7.998e-01 0.060690
#>  [43,] 10 8.070e-01 0.057940
#>  [44,] 10 8.138e-01 0.055300
#>  [45,] 10 8.203e-01 0.052790
#>  [46,] 10 8.265e-01 0.050390
#>  [47,] 10 8.326e-01 0.048100
#>  [48,] 10 8.383e-01 0.045910
#>  [49,] 11 8.439e-01 0.043830
#>  [50,] 11 8.496e-01 0.041830
#>  [51,] 11 8.551e-01 0.039930
#>  [52,] 11 8.604e-01 0.038120
#>  [53,] 11 8.655e-01 0.036390
#>  [54,] 10 8.704e-01 0.034730
#>  [55,] 11 8.752e-01 0.033150
#>  [56,] 12 8.799e-01 0.031650
#>  [57,] 13 8.844e-01 0.030210
#>  [58,] 14 8.888e-01 0.028830
#>  [59,] 14 8.931e-01 0.027520
#>  [60,] 14 8.972e-01 0.026270
#>  [61,] 16 9.011e-01 0.025080
#>  [62,] 17 9.050e-01 0.023940
#>  [63,] 17 9.087e-01 0.022850
#>  [64,] 17 9.123e-01 0.021810
#>  [65,] 17 9.158e-01 0.020820
#>  [66,] 17 9.191e-01 0.019870
#>  [67,] 17 9.223e-01 0.018970
#>  [68,] 18 9.255e-01 0.018110
#>  [69,] 18 9.285e-01 0.017290
#>  [70,] 20 9.314e-01 0.016500
#>  [71,] 20 9.342e-01 0.015750
#>  [72,] 20 9.370e-01 0.015030
#>  [73,] 21 9.396e-01 0.014350
#>  [74,] 20 9.421e-01 0.013700
#>  [75,] 20 9.446e-01 0.013080
#>  [76,] 21 9.469e-01 0.012480
#>  [77,] 21 9.492e-01 0.011910
#>  [78,] 20 9.514e-01 0.011370
#>  [79,] 20 9.534e-01 0.010860
#>  [80,] 20 9.554e-01 0.010360
#>  [81,] 20 9.573e-01 0.009892
#>  [82,] 21 9.592e-01 0.009442
#>  [83,] 21 9.610e-01 0.009013
#>  [84,] 23 9.628e-01 0.008603
#>  [85,] 23 9.644e-01 0.008212
#>  [86,] 23 9.660e-01 0.007839
#>  [87,] 24 9.675e-01 0.007483
#>  [88,] 24 9.690e-01 0.007143
```

```
#>  [89,] 24 9.704e-01 0.006818
#>  [90,] 24 9.717e-01 0.006508
#>  [91,] 24 9.730e-01 0.006212
#>  [92,] 24 9.742e-01 0.005930
#>  [93,] 24 9.753e-01 0.005660
#>  [94,] 24 9.764e-01 0.005403
#>  [95,] 24 9.775e-01 0.005158
#>  [96,] 25 9.785e-01 0.004923
#>  [97,] 25 9.795e-01 0.004699
#>  [98,] 25 9.804e-01 0.004486
#>  [99,] 25 9.813e-01 0.004282
#> [100,] 25 9.821e-01 0.004087
#> [1] "lambda min is at location 42"
#> [1] "the leave-out cells in the source subpop is 70"
#> [1] "use 70 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 140 cells and 201 genes..."
#> [1] "evaluation accuracy ElasticNet 0.957142857142857"
#> [1] "evaluation accuracy LDA 0.8428571428571143"
#> [1] "done bootstrap 1"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 0.53475935828877"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 12.8342245989305"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 83.5714285714286"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 92.1428571428571"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is "
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 24.812030075188"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
```

```
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 22.5"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 45"
#> [1] "Total 140 cells as source subpop"
#> [1] "Total 360 cells in remaining subpops"
#> [1] "subsampling 70 cells for training souce subpop"
#> [1] "subsampling 70 cells in remaining subpops for training"
#> [1] "use 201 genes for training model"
#> [1] "use 201 genes 140 cells for testing model"
#> [1] "rename remaining subpops to 1_3_4"
#> [1] "there are 70 cells in class 1_3_4 and 70 cells in class 2"
#> [1] "standardizing prediction/target dataset"
#> [1] "performning elasticnet model training..."
#> [1] "performming LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>        Df      %Dev   Lambda
#>   [1,]  0 2.723e-15 0.363200
#>   [2,]  2 3.710e-02 0.346700
#>   [3,]  2 7.408e-02 0.331000
#>   [4,]  2 1.081e-01 0.315900
#>   [5,]  2 1.395e-01 0.301600
#>   [6,]  3 1.692e-01 0.287800
#>   [7,]  3 1.984e-01 0.274800
#>   [8,]  4 2.279e-01 0.262300
#>   [9,]  5 2.556e-01 0.250400
#>  [10,]  5 2.815e-01 0.239000
#>  [11,]  5 3.058e-01 0.228100
#>  [12,]  5 3.285e-01 0.217700
#>  [13,]  5 3.497e-01 0.207800
#>  [14,]  5 3.697e-01 0.198400
#>  [15,]  5 3.885e-01 0.189400
#>  [16,]  5 4.062e-01 0.180800
#>  [17,]  5 4.228e-01 0.172600
#>  [18,]  5 4.385e-01 0.164700
#>  [19,]  6 4.534e-01 0.157200
#>  [20,]  8 4.678e-01 0.150100
#>  [21,]  9 4.820e-01 0.143300
#>  [22,] 10 4.959e-01 0.136800
#>  [23,] 10 5.093e-01 0.130500
#>  [24,] 12 5.225e-01 0.124600
#>  [25,] 13 5.354e-01 0.118900
#>  [26,] 13 5.479e-01 0.113500
#>  [27,] 13 5.598e-01 0.108400
#>  [28,] 13 5.712e-01 0.103400
#>  [29,] 14 5.834e-01 0.098750
#>  [30,] 13 5.959e-01 0.094260
#>  [31,] 13 6.078e-01 0.089970
#>  [32,] 13 6.191e-01 0.085880
```

```
#>  [33,] 13 6.301e-01 0.081980
#>  [34,] 14 6.407e-01 0.078250
#>  [35,] 14 6.508e-01 0.074700
#>  [36,] 15 6.611e-01 0.071300
#>  [37,] 15 6.710e-01 0.068060
#>  [38,] 17 6.806e-01 0.064970
#>  [39,] 18 6.901e-01 0.062010
#>  [40,] 19 6.992e-01 0.059200
#>  [41,] 19 7.079e-01 0.056510
#>  [42,] 20 7.164e-01 0.053940
#>  [43,] 20 7.248e-01 0.051490
#>  [44,] 21 7.331e-01 0.049150
#>  [45,] 21 7.412e-01 0.046910
#>  [46,] 22 7.495e-01 0.044780
#>  [47,] 22 7.575e-01 0.042740
#>  [48,] 23 7.655e-01 0.040800
#>  [49,] 22 7.731e-01 0.038950
#>  [50,] 22 7.802e-01 0.037180
#>  [51,] 26 7.877e-01 0.035490
#>  [52,] 26 7.957e-01 0.033870
#>  [53,] 25 8.032e-01 0.032330
#>  [54,] 25 8.103e-01 0.030860
#>  [55,] 25 8.172e-01 0.029460
#>  [56,] 25 8.239e-01 0.028120
#>  [57,] 24 8.304e-01 0.026840
#>  [58,] 24 8.366e-01 0.025620
#>  [59,] 25 8.426e-01 0.024460
#>  [60,] 25 8.484e-01 0.023350
#>  [61,] 24 8.539e-01 0.022290
#>  [62,] 25 8.592e-01 0.021270
#>  [63,] 25 8.644e-01 0.020310
#>  [64,] 25 8.694e-01 0.019380
#>  [65,] 25 8.743e-01 0.018500
#>  [66,] 25 8.790e-01 0.017660
#>  [67,] 25 8.835e-01 0.016860
#>  [68,] 25 8.879e-01 0.016090
#>  [69,] 25 8.921e-01 0.015360
#>  [70,] 25 8.962e-01 0.014660
#>  [71,] 25 9.002e-01 0.014000
#>  [72,] 25 9.040e-01 0.013360
#>  [73,] 26 9.077e-01 0.012750
#>  [74,] 27 9.115e-01 0.012170
#>  [75,] 28 9.151e-01 0.011620
#>  [76,] 29 9.186e-01 0.011090
#>  [77,] 29 9.220e-01 0.010590
#>  [78,] 29 9.252e-01 0.010110
#>  [79,] 29 9.283e-01 0.009647
#>  [80,] 29 9.313e-01 0.009209
#>  [81,] 28 9.342e-01 0.008790
#>  [82,] 28 9.369e-01 0.008391
#>  [83,] 28 9.395e-01 0.008010
#>  [84,] 30 9.422e-01 0.007645
#>  [85,] 30 9.447e-01 0.007298
```

```
#>  [86,] 29 9.472e-01 0.006966
#>  [87,] 29 9.495e-01 0.006650
#>  [88,] 29 9.518e-01 0.006347
#>  [89,] 29 9.539e-01 0.006059
#>  [90,] 29 9.559e-01 0.005784
#>  [91,] 30 9.579e-01 0.005521
#>  [92,] 30 9.597e-01 0.005270
#>  [93,] 30 9.615e-01 0.005030
#>  [94,] 30 9.632e-01 0.004802
#>  [95,] 30 9.648e-01 0.004583
#>  [96,] 30 9.664e-01 0.004375
#>  [97,] 29 9.679e-01 0.004176
#>  [98,] 30 9.693e-01 0.003986
#>  [99,] 30 9.707e-01 0.003805
#> [100,] 30 9.720e-01 0.003632
#> [1] "lambda min is at location 92"
#> [1] "the leave-out cells in the source subpop is 70"
#> [1] "use 70 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 140 cells and 201 genes..."
#> [1] "evaluation accuracy ElasticNet 0.957142857142857"
#> [1] "evaluation accuracy LDA 0.857142857142857"
#> [1] "done bootstrap 2"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 11.2299465240642"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 19.7860962566845"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 99.2857142857143"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 98.5714285714286"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 8.27067669172932"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 15.7894736842105"
#> [1] "predicting from source to target subpop 4..."
```

```
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 32.5"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 65"

c_selectID <- 3
genes = DEgenes$DE_Subpop3vsRemaining$id[1:200]
LSOLDA_dat3 <- bootstrap_scGPS(nboots = 2,mixedpop1 = mixedpop2, mixedpop2 = mixedpop2, genes=genes, c_
     cluster_mixedpop2 = cluster_mixedpop2)
#> [1] "Total 133 cells as source subpop"
#> [1] "Total 367 cells in remaining subpops"
#> [1] "subsampling 66 cells for training souce subpop"
#> [1] "subsampling 66 cells in remaining subpops for training"
#> [1] "use 200 genes for training model"
#> [1] "use 200 genes 132 cells for testing model"
#> [1] "rename remaining subpops to 1_2_4"
#> [1] "there are 66 cells in class 1_2_4 and 66 cells in class 3"
#> [1] "standardizing prediction/target dataset"
#> [1] "performming elasticnet model training..."
#> [1] "performning LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>        Df       %Dev   Lambda
#>   [1,]  0 -2.403e-15 0.264700
#>   [2,]  2  2.194e-02 0.252700
#>   [3,]  3  5.066e-02 0.241200
#>   [4,]  5  7.816e-02 0.230300
#>   [5,]  7  1.074e-01 0.219800
#>   [6,]  7  1.369e-01 0.209800
#>   [7,]  8  1.643e-01 0.200300
#>   [8,]  8  1.903e-01 0.191200
#>   [9,]  8  2.146e-01 0.182500
#>  [10,]  8  2.374e-01 0.174200
#>  [11,]  9  2.588e-01 0.166300
#>  [12,] 10  2.805e-01 0.158700
#>  [13,] 11  3.014e-01 0.151500
#>  [14,] 12  3.216e-01 0.144600
#>  [15,] 11  3.415e-01 0.138000
#>  [16,] 11  3.603e-01 0.131800
#>  [17,] 13  3.785e-01 0.125800
#>  [18,] 14  3.968e-01 0.120100
#>  [19,] 14  4.143e-01 0.114600
#>  [20,] 15  4.314e-01 0.109400
#>  [21,] 15  4.481e-01 0.104400
#>  [22,] 15  4.641e-01 0.099670
#>  [23,] 16  4.798e-01 0.095140
#>  [24,] 17  4.962e-01 0.090820
```

```
#>  [25,] 20  5.124e-01 0.086690
#>  [26,] 20  5.284e-01 0.082750
#>  [27,] 20  5.436e-01 0.078990
#>  [28,] 20  5.581e-01 0.075400
#>  [29,] 21  5.722e-01 0.071970
#>  [30,] 20  5.856e-01 0.068700
#>  [31,] 22  5.986e-01 0.065580
#>  [32,] 22  6.112e-01 0.062600
#>  [33,] 24  6.237e-01 0.059750
#>  [34,] 24  6.357e-01 0.057040
#>  [35,] 24  6.472e-01 0.054440
#>  [36,] 24  6.583e-01 0.051970
#>  [37,] 24  6.689e-01 0.049610
#>  [38,] 25  6.793e-01 0.047350
#>  [39,] 27  6.897e-01 0.045200
#>  [40,] 28  6.997e-01 0.043150
#>  [41,] 29  7.095e-01 0.041180
#>  [42,] 31  7.193e-01 0.039310
#>  [43,] 33  7.292e-01 0.037530
#>  [44,] 34  7.388e-01 0.035820
#>  [45,] 35  7.482e-01 0.034190
#>  [46,] 35  7.572e-01 0.032640
#>  [47,] 36  7.660e-01 0.031150
#>  [48,] 37  7.748e-01 0.029740
#>  [49,] 37  7.837e-01 0.028390
#>  [50,] 39  7.926e-01 0.027100
#>  [51,] 39  8.012e-01 0.025870
#>  [52,] 41  8.094e-01 0.024690
#>  [53,] 42  8.178e-01 0.023570
#>  [54,] 40  8.258e-01 0.022500
#>  [55,] 41  8.334e-01 0.021470
#>  [56,] 42  8.407e-01 0.020500
#>  [57,] 42  8.477e-01 0.019570
#>  [58,] 43  8.544e-01 0.018680
#>  [59,] 44  8.609e-01 0.017830
#>  [60,] 45  8.671e-01 0.017020
#>  [61,] 45  8.731e-01 0.016240
#>  [62,] 45  8.788e-01 0.015510
#>  [63,] 46  8.842e-01 0.014800
#>  [64,] 48  8.894e-01 0.014130
#>  [65,] 48  8.945e-01 0.013490
#>  [66,] 47  8.993e-01 0.012870
#>  [67,] 47  9.039e-01 0.012290
#>  [68,] 47  9.082e-01 0.011730
#>  [69,] 47  9.124e-01 0.011200
#>  [70,] 47  9.164e-01 0.010690
#>  [71,] 48  9.202e-01 0.010200
#>  [72,] 48  9.238e-01 0.009738
#>  [73,] 48  9.272e-01 0.009295
#>  [74,] 48  9.305e-01 0.008873
#>  [75,] 49  9.337e-01 0.008470
#>  [76,] 49  9.367e-01 0.008085
#>  [77,] 49  9.395e-01 0.007717
```

```
#>  [78,] 48  9.423e-01 0.007367
#>  [79,] 48  9.449e-01 0.007032
#>  [80,] 48  9.474e-01 0.006712
#>  [81,] 48  9.497e-01 0.006407
#>  [82,] 48  9.520e-01 0.006116
#>  [83,] 48  9.542e-01 0.005838
#>  [84,] 48  9.562e-01 0.005572
#>  [85,] 48  9.582e-01 0.005319
#>  [86,] 48  9.601e-01 0.005077
#>  [87,] 48  9.619e-01 0.004847
#>  [88,] 48  9.636e-01 0.004626
#>  [89,] 48  9.652e-01 0.004416
#>  [90,] 49  9.668e-01 0.004215
#>  [91,] 49  9.683e-01 0.004024
#>  [92,] 50  9.697e-01 0.003841
#>  [93,] 50  9.711e-01 0.003666
#>  [94,] 50  9.724e-01 0.003500
#>  [95,] 50  9.736e-01 0.003341
#>  [96,] 50  9.748e-01 0.003189
#>  [97,] 50  9.760e-01 0.003044
#>  [98,] 50  9.770e-01 0.002906
#>  [99,] 50  9.781e-01 0.002773
#> [100,] 51  9.791e-01 0.002647
#> [1] "lambda min is at location 38"
#> [1] "the leave-out cells in the source subpop is 67"
#> [1] "use 66 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 133 cells and 200 genes..."
#> [1] "evaluation accuracy ElasticNet 0.962406015037594"
#> [1] "evaluation accuracy LDA 0.819548872180451"
#> [1] "done bootstrap 1"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 19.7860962566845"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 41.7112299465241"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 1.42857142857143"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 5"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 91.7293233082707"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
```

```
#> [1] "class probability prediction LDA for target subpop 3 is 74.4360902255639"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 27.5"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 32.5"
#> [1] "Total 133 cells as source subpop"
#> [1] "Total 367 cells in remaining subpops"
#> [1] "subsampling 66 cells for training souce subpop"
#> [1] "subsampling 66 cells in remaining subpops for training"
#> [1] "use 200 genes for training model"
#> [1] "use 200 genes 132 cells for testing model"
#> [1] "rename remaining subpops to 1_2_4"
#> [1] "there are 66 cells in class 1_2_4 and 66 cells in class 3"
#> [1] "standardizing prediction/target dataset"
#> [1] "performning elasticnet model training..."
#> [1] "performming LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>        Df       %Dev   Lambda
#>   [1,]  0 -2.403e-15 0.311000
#>   [2,]  1  2.482e-02 0.296900
#>   [3,]  1  4.755e-02 0.283400
#>   [4,]  2  7.656e-02 0.270500
#>   [5,]  2  1.035e-01 0.258200
#>   [6,]  3  1.288e-01 0.246500
#>   [7,]  3  1.584e-01 0.235300
#>   [8,]  4  1.872e-01 0.224600
#>   [9,]  4  2.165e-01 0.214400
#>  [10,]  6  2.453e-01 0.204600
#>  [11,]  6  2.735e-01 0.195300
#>  [12,]  7  3.001e-01 0.186400
#>  [13,]  8  3.252e-01 0.178000
#>  [14,]  8  3.488e-01 0.169900
#>  [15,]  8  3.710e-01 0.162200
#>  [16,]  8  3.920e-01 0.154800
#>  [17,]  8  4.118e-01 0.147800
#>  [18,]  9  4.313e-01 0.141000
#>  [19,]  9  4.506e-01 0.134600
#>  [20,] 10  4.690e-01 0.128500
#>  [21,] 11  4.875e-01 0.122700
#>  [22,] 11  5.051e-01 0.117100
#>  [23,] 12  5.220e-01 0.111800
#>  [24,] 12  5.382e-01 0.106700
#>  [25,] 12  5.535e-01 0.101800
#>  [26,] 12  5.682e-01 0.097210
#>  [27,] 12  5.821e-01 0.092790
#>  [28,] 12  5.955e-01 0.088570
#>  [29,] 12  6.082e-01 0.084550
```

```
#>  [30,] 12  6.204e-01 0.080710
#>  [31,] 12  6.320e-01 0.077040
#>  [32,] 12  6.432e-01 0.073540
#>  [33,] 12  6.539e-01 0.070190
#>  [34,] 12  6.641e-01 0.067000
#>  [35,] 15  6.741e-01 0.063960
#>  [36,] 15  6.842e-01 0.061050
#>  [37,] 15  6.938e-01 0.058280
#>  [38,] 16  7.031e-01 0.055630
#>  [39,] 17  7.121e-01 0.053100
#>  [40,] 17  7.215e-01 0.050690
#>  [41,] 18  7.307e-01 0.048380
#>  [42,] 18  7.395e-01 0.046180
#>  [43,] 18  7.480e-01 0.044080
#>  [44,] 18  7.562e-01 0.042080
#>  [45,] 18  7.640e-01 0.040170
#>  [46,] 18  7.716e-01 0.038340
#>  [47,] 19  7.794e-01 0.036600
#>  [48,] 19  7.870e-01 0.034940
#>  [49,] 20  7.944e-01 0.033350
#>  [50,] 20  8.014e-01 0.031830
#>  [51,] 20  8.083e-01 0.030390
#>  [52,] 21  8.149e-01 0.029000
#>  [53,] 23  8.219e-01 0.027690
#>  [54,] 23  8.287e-01 0.026430
#>  [55,] 25  8.354e-01 0.025230
#>  [56,] 26  8.419e-01 0.024080
#>  [57,] 26  8.482e-01 0.022990
#>  [58,] 26  8.543e-01 0.021940
#>  [59,] 27  8.601e-01 0.020940
#>  [60,] 27  8.659e-01 0.019990
#>  [61,] 27  8.714e-01 0.019080
#>  [62,] 28  8.768e-01 0.018220
#>  [63,] 28  8.819e-01 0.017390
#>  [64,] 28  8.868e-01 0.016600
#>  [65,] 28  8.915e-01 0.015840
#>  [66,] 28  8.960e-01 0.015120
#>  [67,] 28  9.003e-01 0.014440
#>  [68,] 29  9.045e-01 0.013780
#>  [69,] 32  9.086e-01 0.013150
#>  [70,] 35  9.125e-01 0.012560
#>  [71,] 35  9.162e-01 0.011980
#>  [72,] 36  9.199e-01 0.011440
#>  [73,] 35  9.233e-01 0.010920
#>  [74,] 35  9.266e-01 0.010420
#>  [75,] 35  9.298e-01 0.009950
#>  [76,] 34  9.329e-01 0.009498
#>  [77,] 35  9.358e-01 0.009066
#>  [78,] 35  9.386e-01 0.008654
#>  [79,] 35  9.413e-01 0.008260
#>  [80,] 35  9.439e-01 0.007885
#>  [81,] 37  9.463e-01 0.007527
#>  [82,] 37  9.487e-01 0.007185
```

```
#>  [83,] 37  9.510e-01 0.006858
#>  [84,] 38  9.532e-01 0.006546
#>  [85,] 39  9.553e-01 0.006249
#>  [86,] 39  9.573e-01 0.005965
#>  [87,] 39  9.593e-01 0.005694
#>  [88,] 39  9.611e-01 0.005435
#>  [89,] 41  9.629e-01 0.005188
#>  [90,] 41  9.645e-01 0.004952
#>  [91,] 41  9.661e-01 0.004727
#>  [92,] 41  9.677e-01 0.004512
#>  [93,] 41  9.691e-01 0.004307
#>  [94,] 41  9.705e-01 0.004111
#>  [95,] 41  9.719e-01 0.003924
#>  [96,] 41  9.731e-01 0.003746
#>  [97,] 40  9.744e-01 0.003576
#>  [98,] 40  9.755e-01 0.003413
#>  [99,] 40  9.766e-01 0.003258
#> [100,] 40  9.777e-01 0.003110
#> [1] "lambda min is at location 63"
#> [1] "the leave-out cells in the source subpop is 67"
#> [1] "use 66 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 133 cells and 200 genes..."
#> [1] "evaluation accuracy ElasticNet 0.849624060150376"
#> [1] "evaluation accuracy LDA 0.75187969924812"
#> [1] "done bootstrap 2"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 44.9197860962567"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 36.3636363636364"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 1.42857142857143"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 12.8571428571429"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 95.4887218045113"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 80.4511278195489"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 45"
```

```
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 40"

c_selectID <- 4
genes = DEgenes$DE_Subpop4vsRemaining$id[1:200]
LSOLDA_dat4 <- bootstrap_scGPS(nboots = 2,mixedpop1 = mixedpop2, mixedpop2 = mixedpop2, genes=genes, c_s
     cluster_mixedpop2 = cluster_mixedpop2)
#> [1] "Total 40 cells as source subpop"
#> [1] "Total 460 cells in remaining subpops"
#> [1] "subsampling 20 cells for training souce subpop"
#> [1] "subsampling 20 cells in remaining subpops for training"
#> [1] "use 201 genes for training model"
#> [1] "use 201 genes 40 cells for testing model"
#> [1] "rename remaining subpops to 1_2_3"
#> [1] "there are 20 cells in class 1_2_3 and 20 cells in class 4"
#> [1] "removing 1 genes with no variance"
#> [1] "standardizing prediction/target dataset"
#> [1] "performming elasticnet model training..."
#> [1] "performning LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>        Df    %Dev   Lambda
#>   [1,]  0 0.00000 0.346100
#>   [2,]  1 0.03072 0.330300
#>   [3,]  1 0.05881 0.315300
#>   [4,]  1 0.08457 0.301000
#>   [5,]  2 0.10880 0.287300
#>   [6,]  2 0.13270 0.274200
#>   [7,]  2 0.15490 0.261800
#>   [8,]  2 0.17550 0.249900
#>   [9,]  2 0.19460 0.238500
#>  [10,]  3 0.21240 0.227700
#>  [11,]  4 0.23120 0.217300
#>  [12,]  4 0.25630 0.207500
#>  [13,]  4 0.28010 0.198000
#>  [14,]  4 0.30270 0.189000
#>  [15,]  4 0.32410 0.180400
#>  [16,]  5 0.34550 0.172200
#>  [17,]  5 0.36750 0.164400
#>  [18,]  5 0.38860 0.156900
#>  [19,]  7 0.40990 0.149800
#>  [20,]  7 0.43340 0.143000
#>  [21,]  8 0.45610 0.136500
#>  [22,]  9 0.48000 0.130300
#>  [23,]  9 0.50280 0.124400
#>  [24,]  9 0.52440 0.118700
#>  [25,]  8 0.54380 0.113300
#>  [26,]  7 0.56190 0.108200
#>  [27,]  8 0.57940 0.103300
#>  [28,]  8 0.59660 0.098560
```

```
#>  [29,]  8 0.61290 0.094080
#>  [30,]  8 0.62850 0.089800
#>  [31,]  8 0.64330 0.085720
#>  [32,]  8 0.65750 0.081830
#>  [33,]  8 0.67100 0.078110
#>  [34,]  8 0.68390 0.074560
#>  [35,]  8 0.69620 0.071170
#>  [36,]  8 0.70800 0.067930
#>  [37,]  8 0.71930 0.064850
#>  [38,]  9 0.73010 0.061900
#>  [39,] 10 0.74070 0.059090
#>  [40,] 10 0.75080 0.056400
#>  [41,] 11 0.76090 0.053840
#>  [42,] 11 0.77080 0.051390
#>  [43,] 11 0.78020 0.049050
#>  [44,] 12 0.78930 0.046820
#>  [45,] 12 0.79810 0.044700
#>  [46,] 12 0.80660 0.042660
#>  [47,] 12 0.81470 0.040730
#>  [48,] 12 0.82240 0.038870
#>  [49,] 12 0.82980 0.037110
#>  [50,] 13 0.83700 0.035420
#>  [51,] 13 0.84390 0.033810
#>  [52,] 13 0.85050 0.032270
#>  [53,] 13 0.85690 0.030810
#>  [54,] 13 0.86290 0.029410
#>  [55,] 13 0.86880 0.028070
#>  [56,] 13 0.87430 0.026790
#>  [57,] 14 0.87980 0.025580
#>  [58,] 14 0.88510 0.024410
#>  [59,] 16 0.89040 0.023300
#>  [60,] 16 0.89550 0.022250
#>  [61,] 16 0.90030 0.021230
#>  [62,] 16 0.90490 0.020270
#>  [63,] 16 0.90930 0.019350
#>  [64,] 16 0.91350 0.018470
#>  [65,] 16 0.91740 0.017630
#>  [66,] 16 0.92120 0.016830
#>  [67,] 16 0.92480 0.016060
#>  [68,] 16 0.92830 0.015330
#>  [69,] 16 0.93160 0.014640
#>  [70,] 16 0.93470 0.013970
#>  [71,] 16 0.93770 0.013340
#>  [72,] 16 0.94050 0.012730
#>  [73,] 16 0.94320 0.012150
#>  [74,] 16 0.94580 0.011600
#>  [75,] 16 0.94820 0.011070
#>  [76,] 17 0.95060 0.010570
#>  [77,] 17 0.95290 0.010090
#>  [78,] 17 0.95510 0.009629
#>  [79,] 17 0.95710 0.009192
#>  [80,] 17 0.95910 0.008774
#>  [81,] 17 0.96100 0.008375
```

```
#>  [82,] 17 0.96280 0.007995
#>  [83,] 17 0.96450 0.007631
#>  [84,] 17 0.96610 0.007284
#>  [85,] 17 0.96770 0.006953
#>  [86,] 17 0.96910 0.006637
#>  [87,] 17 0.97060 0.006336
#>  [88,] 18 0.97190 0.006048
#>  [89,] 18 0.97320 0.005773
#>  [90,] 18 0.97440 0.005510
#>  [91,] 18 0.97560 0.005260
#>  [92,] 18 0.97670 0.005021
#>  [93,] 18 0.97780 0.004793
#>  [94,] 18 0.97880 0.004575
#>  [95,] 18 0.97970 0.004367
#>  [96,] 18 0.98070 0.004168
#>  [97,] 18 0.98150 0.003979
#>  [98,] 18 0.98240 0.003798
#>  [99,] 18 0.98320 0.003625
#> [100,] 18 0.98390 0.003461
#> [1] "lambda min is at location 58"
#> [1] "the leave-out cells in the source subpop is 20"
#> [1] "use 20 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 40 cells and 200 genes..."
#> [1] "evaluation accuracy ElasticNet 0.6"
#> [1] "evaluation accuracy LDA 0.65"
#> [1] "done bootstrap 1"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 62.0320855614973"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 67.9144385026738"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 75.7142857142857"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 84.2857142857143"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 47.3684210526316"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 60.1503759398496"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 77.5"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 25"
```

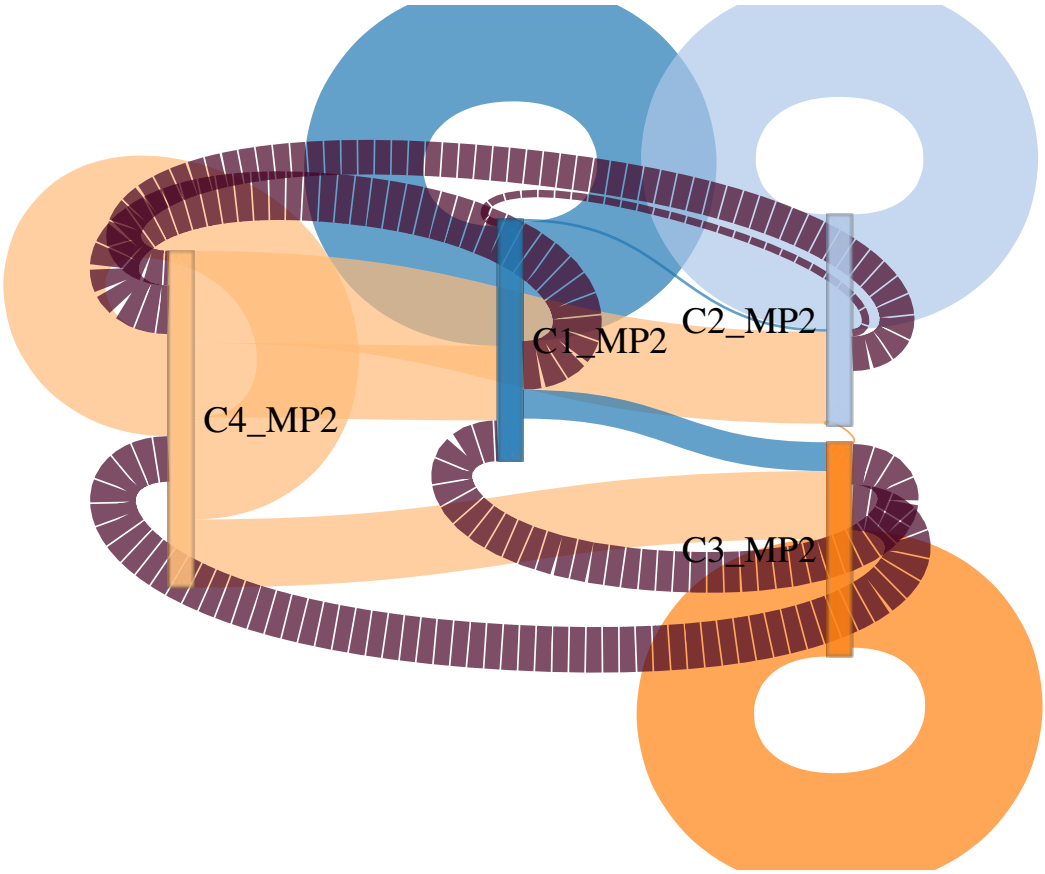
```
#> [1] "Total 40 cells as source subpop"
#> [1] "Total 460 cells in remaining subpops"
#> [1] "subsampling 20 cells for training souce subpop"
#> [1] "subsampling 20 cells in remaining subpops for training"
#> [1] "use 201 genes for training model"
#> [1] "use 201 genes 40 cells for testing model"
#> [1] "rename remaining subpops to 1_2_3"
#> [1] "there are 20 cells in class 1_2_3 and 20 cells in class 4"
#> [1] "standardizing prediction/target dataset"
#> [1] "performning elasticnet model training..."
#> [1] "performning LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>        Df    %Dev   Lambda
#>   [1,]  0 0.00000 0.333800
#>   [2,]  1 0.02858 0.318600
#>   [3,]  1 0.05474 0.304100
#>   [4,]  1 0.07880 0.290300
#>   [5,]  2 0.10480 0.277100
#>   [6,]  2 0.13020 0.264500
#>   [7,]  2 0.15370 0.252500
#>   [8,]  2 0.17570 0.241000
#>   [9,]  2 0.19630 0.230000
#>  [10,]  2 0.21560 0.219600
#>  [11,]  2 0.23370 0.209600
#>  [12,]  2 0.25080 0.200100
#>  [13,]  2 0.26700 0.191000
#>  [14,]  2 0.28230 0.182300
#>  [15,]  3 0.29790 0.174000
#>  [16,]  3 0.31270 0.166100
#>  [17,]  4 0.32710 0.158600
#>  [18,]  4 0.34210 0.151400
#>  [19,]  4 0.35630 0.144500
#>  [20,]  4 0.36990 0.137900
#>  [21,]  6 0.38620 0.131600
#>  [22,]  7 0.40750 0.125700
#>  [23,]  7 0.42780 0.119900
#>  [24,]  7 0.44700 0.114500
#>  [25,]  7 0.46520 0.109300
#>  [26,]  7 0.48260 0.104300
#>  [27,]  7 0.49910 0.099580
#>  [28,]  7 0.51490 0.095060
#>  [29,]  7 0.53010 0.090740
#>  [30,]  7 0.54480 0.086610
#>  [31,]  7 0.55880 0.082680
#>  [32,]  8 0.57240 0.078920
#>  [33,]  9 0.58620 0.075330
#>  [34,]  9 0.60110 0.071910
#>  [35,]  9 0.61540 0.068640
#>  [36,] 10 0.62930 0.065520
#>  [37,] 10 0.64290 0.062540
```

```
#>  [38,] 11 0.65660 0.059700
#>  [39,] 10 0.66980 0.056990
#>  [40,] 11 0.68260 0.054390
#>  [41,] 13 0.69670 0.051920
#>  [42,] 13 0.71030 0.049560
#>  [43,] 12 0.72300 0.047310
#>  [44,] 13 0.73500 0.045160
#>  [45,] 12 0.74640 0.043110
#>  [46,] 14 0.75760 0.041150
#>  [47,] 14 0.76890 0.039280
#>  [48,] 14 0.77980 0.037490
#>  [49,] 14 0.79010 0.035790
#>  [50,] 14 0.79990 0.034160
#>  [51,] 15 0.80930 0.032610
#>  [52,] 15 0.81820 0.031130
#>  [53,] 15 0.82660 0.029710
#>  [54,] 14 0.83470 0.028360
#>  [55,] 14 0.84230 0.027070
#>  [56,] 14 0.84960 0.025840
#>  [57,] 14 0.85650 0.024670
#>  [58,] 14 0.86310 0.023550
#>  [59,] 14 0.86940 0.022480
#>  [60,] 14 0.87530 0.021450
#>  [61,] 14 0.88100 0.020480
#>  [62,] 14 0.88640 0.019550
#>  [63,] 14 0.89160 0.018660
#>  [64,] 14 0.89650 0.017810
#>  [65,] 14 0.90120 0.017000
#>  [66,] 15 0.90570 0.016230
#>  [67,] 15 0.91000 0.015490
#>  [68,] 16 0.91420 0.014790
#>  [69,] 16 0.91810 0.014120
#>  [70,] 16 0.92180 0.013470
#>  [71,] 16 0.92540 0.012860
#>  [72,] 16 0.92880 0.012280
#>  [73,] 16 0.93210 0.011720
#>  [74,] 16 0.93520 0.011190
#>  [75,] 16 0.93810 0.010680
#>  [76,] 16 0.94090 0.010190
#>  [77,] 16 0.94360 0.009729
#>  [78,] 17 0.94620 0.009287
#>  [79,] 17 0.94860 0.008865
#>  [80,] 17 0.95100 0.008462
#>  [81,] 18 0.95320 0.008077
#>  [82,] 18 0.95540 0.007710
#>  [83,] 18 0.95740 0.007360
#>  [84,] 18 0.95930 0.007025
#>  [85,] 18 0.96120 0.006706
#>  [86,] 19 0.96300 0.006401
#>  [87,] 19 0.96470 0.006110
#>  [88,] 19 0.96630 0.005833
#>  [89,] 19 0.96780 0.005567
#>  [90,] 19 0.96930 0.005314
```

```
#>  [91,] 20 0.97070 0.005073
#>  [92,] 20 0.97200 0.004842
#>  [93,] 20 0.97330 0.004622
#>  [94,] 20 0.97450 0.004412
#>  [95,] 20 0.97570 0.004212
#>  [96,] 20 0.97680 0.004020
#>  [97,] 20 0.97780 0.003837
#>  [98,] 19 0.97880 0.003663
#>  [99,] 18 0.97980 0.003497
#> [100,] 18 0.98070 0.003338
#> [1] "lambda min is at location 89"
#> [1] "the leave-out cells in the source subpop is 20"
#> [1] "use 20 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 40 cells and 201 genes..."
#> [1] "evaluation accuracy ElasticNet 0.775"
#> [1] "evaluation accuracy LDA 0.8"
#> [1] "done bootstrap 2"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 56.6844919786096"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 47.0588235294118"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 70.7142857142857"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 80.7142857142857"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 60.9022556390977"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 60.1503759398496"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 85"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 90"
```

```r
#prepare table input for sankey plot

LASSO_C1S2  <- reformat_LASSO(c_selectID=1, mp_selectID = 2, LSOLDA_dat=LSOLDA_dat1,
                          nPredSubpop = length(unique(colData(mixedpop2)[,1])),
                          Nodes_group ="#7570b3")

LASSO_C2S2  <- reformat_LASSO(c_selectID=2, mp_selectID =2, LSOLDA_dat=LSOLDA_dat2,
                          nPredSubpop = length(unique(colData(mixedpop2)[,1])),
                          Nodes_group ="#1b9e77")

LASSO_C3S2  <- reformat_LASSO(c_selectID=3, mp_selectID =2, LSOLDA_dat=LSOLDA_dat3,
                          nPredSubpop = length(unique(colData(mixedpop2)[,1])),
                          Nodes_group ="#e7298a")

LASSO_C4S2  <- reformat_LASSO(c_selectID=4, mp_selectID =2, LSOLDA_dat=LSOLDA_dat4,
                          nPredSubpop = length(unique(colData(mixedpop2)[,1])),
                          Nodes_group ="#00FFFF")

combined <- rbind(LASSO_C1S2,LASSO_C2S2,LASSO_C3S2, LASSO_C4S2 )
combined <- combined[is.na(combined$Value) != TRUE,]

nboots = 2
#links: source, target, value
#source: node, nodegroup
combined_D3obj <-list(Nodes=combined[,(nboots+3):(nboots+4)], Links=combined[,c((nboots+2):(nboots+1),n

library(networkD3)

Node_source <- as.vector(sort(unique(combined_D3obj$Links$Source)))
Node_target <- as.vector(sort(unique(combined_D3obj$Links$Target)))
Node_all <-unique(c(Node_source, Node_target))

#assign IDs for Source (start from 0)
Source <-combined_D3obj$Links$Source
Target <- combined_D3obj$Links$Target

for(i in 1:length(Node_all)){
  Source[Source==Node_all[i]] <-i-1
  Target[Target==Node_all[i]] <-i-1
}

combined_D3obj$Links$Source <- as.numeric(Source)
combined_D3obj$Links$Target <- as.numeric(Target)
combined_D3obj$Links$LinkColor <- combined$NodeGroup

#prepare node info
node_df <-data.frame(Node=Node_all)
node_df$id <-as.numeric(c(0, 1:(length(Node_all)-1)))

suppressMessages(library(dplyr))
Color <- combined %>% count(Node, color=NodeGroup) %>% select(2)
node_df$color <- Color$color
```

```
suppressMessages(library(networkD3))
p1<-sankeyNetwork(Links =combined_D3obj$Links, Nodes = node_df,  Value = "Value", NodeGroup ="color", Li
                  fontSize = 22 )
p1
```




```
#saveNetwork(p1, file = paste0(path,'Subpopulation_Net.html'))
##R Setting Information
#sessionInfo()
#rmarkdown::render("/Users/quan.nguyen/Documents/Powell_group_MacQuan/AllCodes/scGPS/vignettes/vignette
#rmarkdown::render("/Users/quan.nguyen/Documents/Powell_group_MacQuan/AllCodes/scGPS/vignettes/vignette
```

4.3 Plot the relationship between clusters in two samples

Here we look at one example use case to find relationship between clusters within one sample or between two sample

```
#run prediction for 3 clusters
cluster_mixedpop1 <- colData(mixedpop1)[,1]
cluster_mixedpop2 <- colData(mixedpop2)[,1]
row_cluster <-length(unique(colData(mixedpop2)[,1]))

c_selectID <- 1
genes = DEgenes$DE_Subpop1vsRemaining$id[1:200] #top 200 gene markers distinguishing cluster 1
LSOLDA_dat1 <- bootstrap_scGPS(nboots = 2, mixedpop1 = mixedpop1, mixedpop2 = mixedpop2, genes=genes, c
#> [1] "Total 224 cells as source subpop"
```

```
#> [1] "Total 366 cells in remaining subpops"
#> [1] "subsampling 112 cells for training souce subpop"
#> [1] "subsampling 112 cells in remaining subpops for training"
#> [1] "use 191 genes for training model"
#> [1] "use 191 genes 224 cells for testing model"
#> [1] "rename remaining subpops to 2_3"
#> [1] "there are 112 cells in class 2_3 and 112 cells in class 1"
#> [1] "standardizing prediction/target dataset"
#> [1] "performning elasticnet model training..."
#> [1] "performning LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>       Df       %Dev    Lambda
#>  [1,]  0 -2.563e-15 2.894e-01
#>  [2,]  3  4.988e-02 2.637e-01
#>  [3,]  4  1.068e-01 2.402e-01
#>  [4,]  4  1.615e-01 2.189e-01
#>  [5,]  5  2.097e-01 1.994e-01
#>  [6,]  5  2.534e-01 1.817e-01
#>  [7,]  5  2.925e-01 1.656e-01
#>  [8,]  6  3.284e-01 1.509e-01
#>  [9,]  6  3.611e-01 1.375e-01
#> [10,]  7  3.910e-01 1.253e-01
#> [11,]  8  4.198e-01 1.141e-01
#> [12,]  9  4.461e-01 1.040e-01
#> [13,]  9  4.709e-01 9.475e-02
#> [14,]  9  4.935e-01 8.633e-02
#> [15,] 10  5.144e-01 7.866e-02
#> [16,] 10  5.346e-01 7.168e-02
#> [17,] 12  5.534e-01 6.531e-02
#> [18,] 15  5.751e-01 5.951e-02
#> [19,] 18  5.974e-01 5.422e-02
#> [20,] 20  6.199e-01 4.940e-02
#> [21,] 21  6.419e-01 4.501e-02
#> [22,] 23  6.625e-01 4.102e-02
#> [23,] 25  6.814e-01 3.737e-02
#> [24,] 24  6.994e-01 3.405e-02
#> [25,] 24  7.161e-01 3.103e-02
#> [26,] 27  7.321e-01 2.827e-02
#> [27,] 28  7.468e-01 2.576e-02
#> [28,] 28  7.607e-01 2.347e-02
#> [29,] 28  7.736e-01 2.139e-02
#> [30,] 28  7.856e-01 1.949e-02
#> [31,] 30  7.970e-01 1.775e-02
#> [32,] 31  8.077e-01 1.618e-02
#> [33,] 32  8.180e-01 1.474e-02
#> [34,] 34  8.285e-01 1.343e-02
#> [35,] 34  8.384e-01 1.224e-02
#> [36,] 36  8.479e-01 1.115e-02
#> [37,] 40  8.575e-01 1.016e-02
#> [38,] 41  8.672e-01 9.257e-03
```

```
#> [39,] 41  8.763e-01 8.435e-03
#> [40,] 43  8.851e-01 7.686e-03
#> [41,] 45  8.935e-01 7.003e-03
#> [42,] 46  9.016e-01 6.381e-03
#> [43,] 48  9.094e-01 5.814e-03
#> [44,] 49  9.167e-01 5.297e-03
#> [45,] 52  9.237e-01 4.827e-03
#> [46,] 53  9.302e-01 4.398e-03
#> [47,] 54  9.363e-01 4.007e-03
#> [48,] 56  9.418e-01 3.651e-03
#> [49,] 56  9.469e-01 3.327e-03
#> [50,] 57  9.515e-01 3.031e-03
#> [51,] 56  9.558e-01 2.762e-03
#> [52,] 58  9.597e-01 2.517e-03
#> [53,] 60  9.633e-01 2.293e-03
#> [54,] 60  9.666e-01 2.089e-03
#> [55,] 60  9.696e-01 1.904e-03
#> [56,] 60  9.723e-01 1.735e-03
#> [57,] 61  9.748e-01 1.581e-03
#> [58,] 61  9.771e-01 1.440e-03
#> [59,] 61  9.791e-01 1.312e-03
#> [60,] 61  9.810e-01 1.196e-03
#> [61,] 60  9.827e-01 1.089e-03
#> [62,] 62  9.842e-01 9.926e-04
#> [63,] 62  9.856e-01 9.045e-04
#> [64,] 62  9.869e-01 8.241e-04
#> [65,] 62  9.881e-01 7.509e-04
#> [66,] 62  9.892e-01 6.842e-04
#> [67,] 63  9.901e-01 6.234e-04
#> [68,] 64  9.910e-01 5.680e-04
#> [69,] 65  9.918e-01 5.176e-04
#> [70,] 66  9.925e-01 4.716e-04
#> [71,] 67  9.932e-01 4.297e-04
#> [72,] 67  9.938e-01 3.915e-04
#> [73,] 67  9.943e-01 3.567e-04
#> [74,] 67  9.949e-01 3.250e-04
#> [75,] 67  9.953e-01 2.962e-04
#> [76,] 67  9.957e-01 2.699e-04
#> [77,] 67  9.961e-01 2.459e-04
#> [78,] 67  9.965e-01 2.240e-04
#> [79,] 67  9.968e-01 2.041e-04
#> [80,] 67  9.971e-01 1.860e-04
#> [81,] 67  9.973e-01 1.695e-04
#> [82,] 67  9.976e-01 1.544e-04
#> [83,] 66  9.978e-01 1.407e-04
#> [84,] 66  9.980e-01 1.282e-04
#> [85,] 66  9.981e-01 1.168e-04
#> [86,] 66  9.983e-01 1.064e-04
#> [87,] 66  9.985e-01 9.698e-05
#> [88,] 66  9.986e-01 8.837e-05
#> [89,] 66  9.987e-01 8.052e-05
#> [90,] 66  9.988e-01 7.336e-05
#> [91,] 67  9.989e-01 6.685e-05
```

```
#> [92,] 67  9.990e-01 6.091e-05
#> [1] "lambda min is at location 30"
#> [1] "the leave-out cells in the source subpop is 112"
#> [1] "use 112 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 224 cells and 191 genes..."
#> [1] "evaluation accuracy ElasticNet 0.8660714285714291"
#> [1] "evaluation accuracy LDA 0.5446428571428571"
#> [1] "done bootstrap 1"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 44.91978609625671"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 56.14973262032091"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 89.28571428571431"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 41.42857142857141"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 27.067669172932334"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 60.902255639097746"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 50"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 47.5"
#> [1] "Total 224 cells as source subpop"
#> [1] "Total 366 cells in remaining subpops"
#> [1] "subsampling 112 cells for training souce subpop"
#> [1] "subsampling 112 cells in remaining subpops for training"
#> [1] "use 191 genes for training model"
#> [1] "use 191 genes 224 cells for testing model"
```

```
#> [1] "rename remaining subpops to 2_3"
#> [1] "there are 112 cells in class 2_3 and 112 cells in class 1"
#> [1] "removing 3 genes with no variance"
#> [1] "standardizing prediction/target dataset"
#> [1] "performming elasticnet model training..."
#> [1] "performning LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>       Df       %Dev    Lambda
#>  [1,]  0 -2.563e-15 2.900e-01
#>  [2,]  2  4.555e-02 2.642e-01
#>  [3,]  5  9.811e-02 2.408e-01
#>  [4,]  5  1.514e-01 2.194e-01
#>  [5,]  5  1.975e-01 1.999e-01
#>  [6,]  5  2.379e-01 1.821e-01
#>  [7,]  5  2.736e-01 1.659e-01
#>  [8,]  5  3.053e-01 1.512e-01
#>  [9,]  5  3.337e-01 1.378e-01
#> [10,]  5  3.591e-01 1.255e-01
#> [11,]  7  3.829e-01 1.144e-01
#> [12,]  8  4.084e-01 1.042e-01
#> [13,] 11  4.353e-01 9.496e-02
#> [14,] 12  4.614e-01 8.652e-02
#> [15,] 15  4.864e-01 7.884e-02
#> [16,] 16  5.111e-01 7.183e-02
#> [17,] 16  5.338e-01 6.545e-02
#> [18,] 17  5.546e-01 5.964e-02
#> [19,] 17  5.739e-01 5.434e-02
#> [20,] 20  5.919e-01 4.951e-02
#> [21,] 21  6.092e-01 4.511e-02
#> [22,] 21  6.250e-01 4.111e-02
#> [23,] 21  6.395e-01 3.745e-02
#> [24,] 24  6.532e-01 3.413e-02
#> [25,] 28  6.679e-01 3.109e-02
#> [26,] 31  6.835e-01 2.833e-02
#> [27,] 32  7.002e-01 2.582e-02
#> [28,] 34  7.160e-01 2.352e-02
#> [29,] 38  7.317e-01 2.143e-02
#> [30,] 38  7.476e-01 1.953e-02
#> [31,] 41  7.623e-01 1.779e-02
#> [32,] 44  7.767e-01 1.621e-02
#> [33,] 45  7.904e-01 1.477e-02
#> [34,] 45  8.037e-01 1.346e-02
#> [35,] 46  8.163e-01 1.226e-02
#> [36,] 47  8.282e-01 1.117e-02
#> [37,] 49  8.399e-01 1.018e-02
#> [38,] 50  8.510e-01 9.278e-03
#> [39,] 54  8.618e-01 8.453e-03
#> [40,] 56  8.730e-01 7.702e-03
#> [41,] 57  8.835e-01 7.018e-03
#> [42,] 58  8.932e-01 6.395e-03
```

```
#> [43,] 59  9.022e-01 5.827e-03
#> [44,] 62  9.107e-01 5.309e-03
#> [45,] 63  9.185e-01 4.837e-03
#> [46,] 64  9.257e-01 4.408e-03
#> [47,] 65  9.323e-01 4.016e-03
#> [48,] 64  9.382e-01 3.659e-03
#> [49,] 65  9.436e-01 3.334e-03
#> [50,] 65  9.486e-01 3.038e-03
#> [51,] 67  9.531e-01 2.768e-03
#> [52,] 68  9.573e-01 2.522e-03
#> [53,] 68  9.610e-01 2.298e-03
#> [54,] 69  9.645e-01 2.094e-03
#> [55,] 69  9.676e-01 1.908e-03
#> [56,] 70  9.705e-01 1.738e-03
#> [57,] 71  9.731e-01 1.584e-03
#> [58,] 71  9.754e-01 1.443e-03
#> [59,] 72  9.776e-01 1.315e-03
#> [60,] 72  9.796e-01 1.198e-03
#> [61,] 72  9.814e-01 1.092e-03
#> [62,] 72  9.830e-01 9.948e-04
#> [63,] 73  9.845e-01 9.064e-04
#> [64,] 74  9.859e-01 8.259e-04
#> [65,] 73  9.872e-01 7.525e-04
#> [66,] 73  9.883e-01 6.857e-04
#> [67,] 71  9.893e-01 6.248e-04
#> [68,] 71  9.903e-01 5.693e-04
#> [69,] 72  9.911e-01 5.187e-04
#> [70,] 72  9.919e-01 4.726e-04
#> [71,] 72  9.926e-01 4.306e-04
#> [72,] 72  9.933e-01 3.924e-04
#> [73,] 71  9.939e-01 3.575e-04
#> [74,] 71  9.944e-01 3.258e-04
#> [75,] 71  9.949e-01 2.968e-04
#> [76,] 70  9.953e-01 2.704e-04
#> [77,] 70  9.957e-01 2.464e-04
#> [78,] 71  9.961e-01 2.245e-04
#> [79,] 72  9.965e-01 2.046e-04
#> [80,] 71  9.968e-01 1.864e-04
#> [81,] 71  9.971e-01 1.698e-04
#> [82,] 72  9.973e-01 1.548e-04
#> [83,] 73  9.976e-01 1.410e-04
#> [84,] 73  9.978e-01 1.285e-04
#> [85,] 73  9.980e-01 1.171e-04
#> [86,] 73  9.981e-01 1.067e-04
#> [87,] 73  9.983e-01 9.719e-05
#> [88,] 73  9.985e-01 8.856e-05
#> [89,] 74  9.986e-01 8.069e-05
#> [90,] 75  9.987e-01 7.352e-05
#> [91,] 75  9.988e-01 6.699e-05
#> [92,] 75  9.989e-01 6.104e-05
#> [93,] 76  9.990e-01 5.562e-05
#> [1] "lambda min is at location 21"
#> [1] "the leave-out cells in the source subpop is 112"
```

```
#> [1] "use 112 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 224 cells and 188 genes..."
#> [1] "evaluation accuracy ElasticNet 0.879464285714286"
#> [1] "evaluation accuracy LDA 0.65625"
#> [1] "done bootstrap 2"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 52.4064171122995"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 25.668449197861"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 47.8571428571429"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 93.5714285714286"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 26.3157894736842"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 21.0526315789474"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 27.5"
#> [1] "add 6 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 67.5"


c_selectID <- 2
genes = DEgenes$DE_Subpop2vsRemaining$id[1:200]
LSOLDA_dat2 <- bootstrap_scGPS(nboots = 2,mixedpop1 = mixedpop1, mixedpop2 = mixedpop2, genes=genes, c_
     cluster_mixedpop2 = cluster_mixedpop2)
#> [1] "Total 203 cells as source subpop"
#> [1] "Total 387 cells in remaining subpops"
#> [1] "subsampling 102 cells for training souce subpop"
#> [1] "subsampling 102 cells in remaining subpops for training"
#> [1] "use 192 genes for training model"
#> [1] "use 192 genes 204 cells for testing model"
#> [1] "rename remaining subpops to 1_3"
#> [1] "there are 102 cells in class 1_3 and 102 cells in class 2"
#> [1] "removing 5 genes with no variance"
#> [1] "standardizing prediction/target dataset"
#> [1] "performming elasticnet model training..."
```

```
#> [1] "performming LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>       Df      %Dev    Lambda
#>  [1,]  0 1.281e-15 3.550e-01
#>  [2,]  1 6.199e-02 3.235e-01
#>  [3,]  1 1.146e-01 2.948e-01
#>  [4,]  1 1.599e-01 2.686e-01
#>  [5,]  1 1.995e-01 2.447e-01
#>  [6,]  1 2.343e-01 2.230e-01
#>  [7,]  2 2.674e-01 2.032e-01
#>  [8,]  2 3.010e-01 1.851e-01
#>  [9,]  4 3.345e-01 1.687e-01
#> [10,]  4 3.659e-01 1.537e-01
#> [11,]  4 3.940e-01 1.400e-01
#> [12,]  5 4.223e-01 1.276e-01
#> [13,]  6 4.491e-01 1.163e-01
#> [14,]  6 4.736e-01 1.059e-01
#> [15,]  8 4.963e-01 9.652e-02
#> [16,]  8 5.190e-01 8.795e-02
#> [17,]  8 5.393e-01 8.013e-02
#> [18,] 11 5.584e-01 7.302e-02
#> [19,] 12 5.778e-01 6.653e-02
#> [20,] 15 5.963e-01 6.062e-02
#> [21,] 15 6.131e-01 5.523e-02
#> [22,] 20 6.309e-01 5.033e-02
#> [23,] 20 6.499e-01 4.586e-02
#> [24,] 20 6.674e-01 4.178e-02
#> [25,] 21 6.841e-01 3.807e-02
#> [26,] 21 6.998e-01 3.469e-02
#> [27,] 22 7.151e-01 3.161e-02
#> [28,] 22 7.294e-01 2.880e-02
#> [29,] 23 7.431e-01 2.624e-02
#> [30,] 24 7.565e-01 2.391e-02
#> [31,] 27 7.694e-01 2.179e-02
#> [32,] 29 7.832e-01 1.985e-02
#> [33,] 31 7.968e-01 1.809e-02
#> [34,] 33 8.101e-01 1.648e-02
#> [35,] 34 8.228e-01 1.502e-02
#> [36,] 36 8.348e-01 1.368e-02
#> [37,] 37 8.466e-01 1.247e-02
#> [38,] 38 8.576e-01 1.136e-02
#> [39,] 41 8.684e-01 1.035e-02
#> [40,] 42 8.788e-01 9.430e-03
#> [41,] 43 8.883e-01 8.593e-03
#> [42,] 45 8.975e-01 7.829e-03
#> [43,] 46 9.060e-01 7.134e-03
#> [44,] 46 9.138e-01 6.500e-03
#> [45,] 48 9.210e-01 5.923e-03
#> [46,] 50 9.278e-01 5.396e-03
#> [47,] 52 9.341e-01 4.917e-03
```

```
#> [48,] 51 9.399e-01 4.480e-03
#> [49,] 53 9.452e-01 4.082e-03
#> [50,] 53 9.499e-01 3.720e-03
#> [51,] 53 9.543e-01 3.389e-03
#> [52,] 54 9.584e-01 3.088e-03
#> [53,] 54 9.620e-01 2.814e-03
#> [54,] 54 9.653e-01 2.564e-03
#> [55,] 55 9.684e-01 2.336e-03
#> [56,] 55 9.712e-01 2.128e-03
#> [57,] 53 9.737e-01 1.939e-03
#> [58,] 53 9.761e-01 1.767e-03
#> [59,] 53 9.782e-01 1.610e-03
#> [60,] 53 9.801e-01 1.467e-03
#> [61,] 53 9.818e-01 1.337e-03
#> [62,] 53 9.834e-01 1.218e-03
#> [63,] 53 9.849e-01 1.110e-03
#> [64,] 53 9.862e-01 1.011e-03
#> [65,] 53 9.874e-01 9.214e-04
#> [66,] 55 9.885e-01 8.395e-04
#> [67,] 55 9.895e-01 7.649e-04
#> [68,] 55 9.904e-01 6.970e-04
#> [69,] 56 9.913e-01 6.351e-04
#> [70,] 57 9.920e-01 5.786e-04
#> [71,] 57 9.927e-01 5.272e-04
#> [72,] 57 9.934e-01 4.804e-04
#> [73,] 57 9.940e-01 4.377e-04
#> [74,] 58 9.945e-01 3.988e-04
#> [75,] 59 9.950e-01 3.634e-04
#> [76,] 61 9.954e-01 3.311e-04
#> [77,] 61 9.958e-01 3.017e-04
#> [78,] 61 9.962e-01 2.749e-04
#> [79,] 61 9.965e-01 2.505e-04
#> [80,] 61 9.968e-01 2.282e-04
#> [81,] 61 9.971e-01 2.080e-04
#> [82,] 62 9.974e-01 1.895e-04
#> [83,] 62 9.976e-01 1.726e-04
#> [84,] 62 9.978e-01 1.573e-04
#> [85,] 62 9.980e-01 1.433e-04
#> [86,] 62 9.982e-01 1.306e-04
#> [87,] 62 9.983e-01 1.190e-04
#> [88,] 62 9.985e-01 1.084e-04
#> [89,] 62 9.986e-01 9.879e-05
#> [90,] 62 9.987e-01 9.002e-05
#> [91,] 62 9.989e-01 8.202e-05
#> [92,] 62 9.990e-01 7.473e-05
#> [93,] 63 9.990e-01 6.809e-05
#> [1] "lambda min is at location 28"
#> [1] "the leave-out cells in the source subpop is 101"
#> [1] "use 102 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 203 cells and 187 genes..."
#> [1] "evaluation accuracy ElasticNet 0.866995073891626"
```

```
#> [1] "evaluation accuracy LDA 0.605911330049261"
#> [1] "done bootstrap 1"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 67.9144385026738"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 47.0588235294118"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 7.14285714285714"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 16.4285714285714"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 68.4210526315789"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 63.1578947368421"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 37.5"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 40"
#> [1] "Total 203 cells as source subpop"
#> [1] "Total 387 cells in remaining subpops"
#> [1] "subsampling 102 cells for training souce subpop"
#> [1] "subsampling 102 cells in remaining subpops for training"
#> [1] "use 192 genes for training model"
#> [1] "use 192 genes 204 cells for testing model"
#> [1] "rename remaining subpops to 1_3"
#> [1] "there are 102 cells in class 1_3 and 102 cells in class 2"
#> [1] "removing 3 genes with no variance"
#> [1] "standardizing prediction/target dataset"
#> [1] "performming elasticnet model training..."
#> [1] "performming LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
```

```
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>       Df      %Dev    Lambda
#>  [1,]  0 1.281e-15 3.206e-01
#>  [2,]  1 5.048e-02 2.921e-01
#>  [3,]  1 9.308e-02 2.661e-01
#>  [4,]  2 1.388e-01 2.425e-01
#>  [5,]  2 1.787e-01 2.209e-01
#>  [6,]  2 2.133e-01 2.013e-01
#>  [7,]  4 2.460e-01 1.834e-01
#>  [8,]  4 2.773e-01 1.671e-01
#>  [9,]  4 3.051e-01 1.523e-01
#> [10,]  4 3.298e-01 1.388e-01
#> [11,]  4 3.519e-01 1.264e-01
#> [12,]  6 3.758e-01 1.152e-01
#> [13,]  7 3.980e-01 1.050e-01
#> [14,]  7 4.186e-01 9.564e-02
#> [15,] 10 4.393e-01 8.715e-02
#> [16,] 10 4.627e-01 7.940e-02
#> [17,] 13 4.851e-01 7.235e-02
#> [18,] 14 5.081e-01 6.592e-02
#> [19,] 14 5.292e-01 6.007e-02
#> [20,] 15 5.489e-01 5.473e-02
#> [21,] 16 5.672e-01 4.987e-02
#> [22,] 19 5.858e-01 4.544e-02
#> [23,] 20 6.046e-01 4.140e-02
#> [24,] 20 6.216e-01 3.772e-02
#> [25,] 23 6.378e-01 3.437e-02
#> [26,] 24 6.538e-01 3.132e-02
#> [27,] 27 6.698e-01 2.854e-02
#> [28,] 29 6.855e-01 2.600e-02
#> [29,] 32 7.020e-01 2.369e-02
#> [30,] 34 7.189e-01 2.159e-02
#> [31,] 34 7.351e-01 1.967e-02
#> [32,] 37 7.512e-01 1.792e-02
#> [33,] 40 7.684e-01 1.633e-02
#> [34,] 40 7.844e-01 1.488e-02
#> [35,] 41 7.995e-01 1.356e-02
#> [36,] 44 8.136e-01 1.235e-02
#> [37,] 46 8.269e-01 1.126e-02
#> [38,] 48 8.402e-01 1.026e-02
#> [39,] 52 8.530e-01 9.344e-03
#> [40,] 52 8.650e-01 8.514e-03
#> [41,] 53 8.763e-01 7.758e-03
#> [42,] 56 8.868e-01 7.069e-03
#> [43,] 58 8.966e-01 6.441e-03
#> [44,] 57 9.056e-01 5.868e-03
#> [45,] 58 9.138e-01 5.347e-03
#> [46,] 58 9.213e-01 4.872e-03
#> [47,] 60 9.282e-01 4.439e-03
#> [48,] 60 9.346e-01 4.045e-03
#> [49,] 62 9.403e-01 3.686e-03
#> [50,] 64 9.456e-01 3.358e-03
```

```
#> [51,] 66 9.505e-01 3.060e-03
#> [52,] 66 9.549e-01 2.788e-03
#> [53,] 67 9.589e-01 2.540e-03
#> [54,] 67 9.626e-01 2.315e-03
#> [55,] 66 9.659e-01 2.109e-03
#> [56,] 66 9.689e-01 1.922e-03
#> [57,] 66 9.717e-01 1.751e-03
#> [58,] 66 9.742e-01 1.595e-03
#> [59,] 67 9.765e-01 1.454e-03
#> [60,] 67 9.786e-01 1.325e-03
#> [61,] 68 9.805e-01 1.207e-03
#> [62,] 68 9.822e-01 1.100e-03
#> [63,] 68 9.838e-01 1.002e-03
#> [64,] 67 9.852e-01 9.129e-04
#> [65,] 67 9.865e-01 8.318e-04
#> [66,] 67 9.877e-01 7.579e-04
#> [67,] 67 9.888e-01 6.906e-04
#> [68,] 67 9.898e-01 6.293e-04
#> [69,] 67 9.907e-01 5.734e-04
#> [70,] 67 9.915e-01 5.224e-04
#> [71,] 67 9.922e-01 4.760e-04
#> [72,] 67 9.929e-01 4.337e-04
#> [73,] 68 9.936e-01 3.952e-04
#> [74,] 68 9.941e-01 3.601e-04
#> [75,] 68 9.946e-01 3.281e-04
#> [76,] 68 9.951e-01 2.989e-04
#> [77,] 68 9.955e-01 2.724e-04
#> [78,] 68 9.959e-01 2.482e-04
#> [79,] 69 9.963e-01 2.261e-04
#> [80,] 69 9.966e-01 2.061e-04
#> [81,] 68 9.969e-01 1.877e-04
#> [82,] 69 9.972e-01 1.711e-04
#> [83,] 69 9.974e-01 1.559e-04
#> [84,] 69 9.977e-01 1.420e-04
#> [85,] 69 9.979e-01 1.294e-04
#> [86,] 69 9.981e-01 1.179e-04
#> [87,] 69 9.982e-01 1.074e-04
#> [88,] 70 9.984e-01 9.789e-05
#> [89,] 70 9.985e-01 8.920e-05
#> [90,] 70 9.987e-01 8.127e-05
#> [91,] 70 9.988e-01 7.405e-05
#> [92,] 70 9.989e-01 6.747e-05
#> [93,] 70 9.990e-01 6.148e-05
#> [94,] 70 9.991e-01 5.602e-05
#> [1] "lambda min is at location 30"
#> [1] "the leave-out cells in the source subpop is 101"
#> [1] "use 102 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 203 cells and 189 genes..."
#> [1] "evaluation accuracy ElasticNet 0.866995073891626"
#> [1] "evaluation accuracy LDA 0.566502463054187"
#> [1] "done bootstrap 2"
```

```
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 62.0320855614973"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 36.8983957219251"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 30"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 79.2857142857143"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 71.4285714285714"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 39.0977443609023"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "add 1 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "Replacing missing genes by NA..."
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 37.5"
#> [1] "add 5 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 60"

c_selectID <- 3
genes = DEgenes$DE_Subpop3vsRemaining$id[1:200]
LSOLDA_dat3 <- bootstrap_scGPS(nboots = 2,mixedpop1 = mixedpop1, mixedpop2 = mixedpop2, genes=genes, c_
     cluster_mixedpop2 = cluster_mixedpop2)
#> [1] "Total 163 cells as source subpop"
#> [1] "Total 427 cells in remaining subpops"
#> [1] "subsampling 82 cells for training souce subpop"
#> [1] "subsampling 82 cells in remaining subpops for training"
#> [1] "use 196 genes for training model"
#> [1] "use 196 genes 164 cells for testing model"
#> [1] "rename remaining subpops to 2_1"
#> [1] "there are 82 cells in class 2_1 and 82 cells in class 3"
#> [1] "removing 3 genes with no variance"
#> [1] "standardizing prediction/target dataset"
#> [1] "performming elasticnet model training..."
```

```
#> [1] "performming LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
#>
#>        Df      %Dev   Lambda
#>   [1,]  0 2.403e-15 0.182500
#>   [2,]  2 1.147e-02 0.174200
#>   [3,]  2 2.352e-02 0.166300
#>   [4,]  2 3.459e-02 0.158800
#>   [5,]  2 4.480e-02 0.151500
#>   [6,]  2 5.423e-02 0.144700
#>   [7,]  2 6.298e-02 0.138100
#>   [8,]  3 7.247e-02 0.131800
#>   [9,]  3 8.337e-02 0.125800
#>  [10,]  3 9.347e-02 0.120100
#>  [11,]  3 1.028e-01 0.114600
#>  [12,]  6 1.133e-01 0.109400
#>  [13,]  6 1.246e-01 0.104500
#>  [14,]  6 1.351e-01 0.099700
#>  [15,]  7 1.452e-01 0.095170
#>  [16,]  7 1.570e-01 0.090850
#>  [17,]  9 1.716e-01 0.086720
#>  [18,] 10 1.874e-01 0.082780
#>  [19,] 10 2.025e-01 0.079010
#>  [20,] 10 2.167e-01 0.075420
#>  [21,] 14 2.330e-01 0.071990
#>  [22,] 17 2.520e-01 0.068720
#>  [23,] 18 2.707e-01 0.065600
#>  [24,] 19 2.883e-01 0.062620
#>  [25,] 20 3.054e-01 0.059770
#>  [26,] 21 3.220e-01 0.057050
#>  [27,] 24 3.381e-01 0.054460
#>  [28,] 26 3.541e-01 0.051990
#>  [29,] 26 3.700e-01 0.049620
#>  [30,] 26 3.850e-01 0.047370
#>  [31,] 27 3.995e-01 0.045210
#>  [32,] 30 4.143e-01 0.043160
#>  [33,] 31 4.286e-01 0.041200
#>  [34,] 32 4.423e-01 0.039320
#>  [35,] 32 4.556e-01 0.037540
#>  [36,] 32 4.683e-01 0.035830
#>  [37,] 35 4.813e-01 0.034200
#>  [38,] 36 4.947e-01 0.032650
#>  [39,] 36 5.079e-01 0.031160
#>  [40,] 38 5.207e-01 0.029750
#>  [41,] 39 5.330e-01 0.028400
#>  [42,] 38 5.454e-01 0.027110
#>  [43,] 39 5.573e-01 0.025870
#>  [44,] 40 5.688e-01 0.024700
#>  [45,] 41 5.802e-01 0.023570
#>  [46,] 42 5.913e-01 0.022500
#>  [47,] 45 6.025e-01 0.021480
```

```
#>  [48,] 47 6.140e-01 0.020500
#>  [49,] 49 6.253e-01 0.019570
#>  [50,] 50 6.368e-01 0.018680
#>  [51,] 51 6.481e-01 0.017830
#>  [52,] 52 6.596e-01 0.017020
#>  [53,] 53 6.709e-01 0.016250
#>  [54,] 58 6.826e-01 0.015510
#>  [55,] 59 6.947e-01 0.014810
#>  [56,] 61 7.067e-01 0.014130
#>  [57,] 62 7.185e-01 0.013490
#>  [58,] 62 7.298e-01 0.012880
#>  [59,] 62 7.406e-01 0.012290
#>  [60,] 63 7.510e-01 0.011730
#>  [61,] 63 7.610e-01 0.011200
#>  [62,] 64 7.707e-01 0.010690
#>  [63,] 64 7.801e-01 0.010200
#>  [64,] 65 7.891e-01 0.009741
#>  [65,] 66 7.978e-01 0.009298
#>  [66,] 65 8.063e-01 0.008876
#>  [67,] 67 8.146e-01 0.008472
#>  [68,] 67 8.226e-01 0.008087
#>  [69,] 67 8.303e-01 0.007720
#>  [70,] 67 8.377e-01 0.007369
#>  [71,] 67 8.447e-01 0.007034
#>  [72,] 70 8.515e-01 0.006714
#>  [73,] 70 8.581e-01 0.006409
#>  [74,] 71 8.644e-01 0.006118
#>  [75,] 71 8.704e-01 0.005840
#>  [76,] 71 8.762e-01 0.005574
#>  [77,] 71 8.817e-01 0.005321
#>  [78,] 71 8.870e-01 0.005079
#>  [79,] 71 8.920e-01 0.004848
#>  [80,] 71 8.969e-01 0.004628
#>  [81,] 71 9.015e-01 0.004417
#>  [82,] 70 9.059e-01 0.004217
#>  [83,] 71 9.101e-01 0.004025
#>  [84,] 72 9.141e-01 0.003842
#>  [85,] 72 9.179e-01 0.003667
#>  [86,] 72 9.216e-01 0.003501
#>  [87,] 74 9.251e-01 0.003342
#>  [88,] 74 9.286e-01 0.003190
#>  [89,] 74 9.318e-01 0.003045
#>  [90,] 74 9.349e-01 0.002906
#>  [91,] 74 9.379e-01 0.002774
#>  [92,] 74 9.407e-01 0.002648
#>  [93,] 74 9.434e-01 0.002528
#>  [94,] 74 9.460e-01 0.002413
#>  [95,] 75 9.484e-01 0.002303
#>  [96,] 75 9.508e-01 0.002199
#>  [97,] 76 9.530e-01 0.002099
#>  [98,] 76 9.551e-01 0.002003
#>  [99,] 76 9.572e-01 0.001912
#> [100,] 76 9.591e-01 0.001825
```

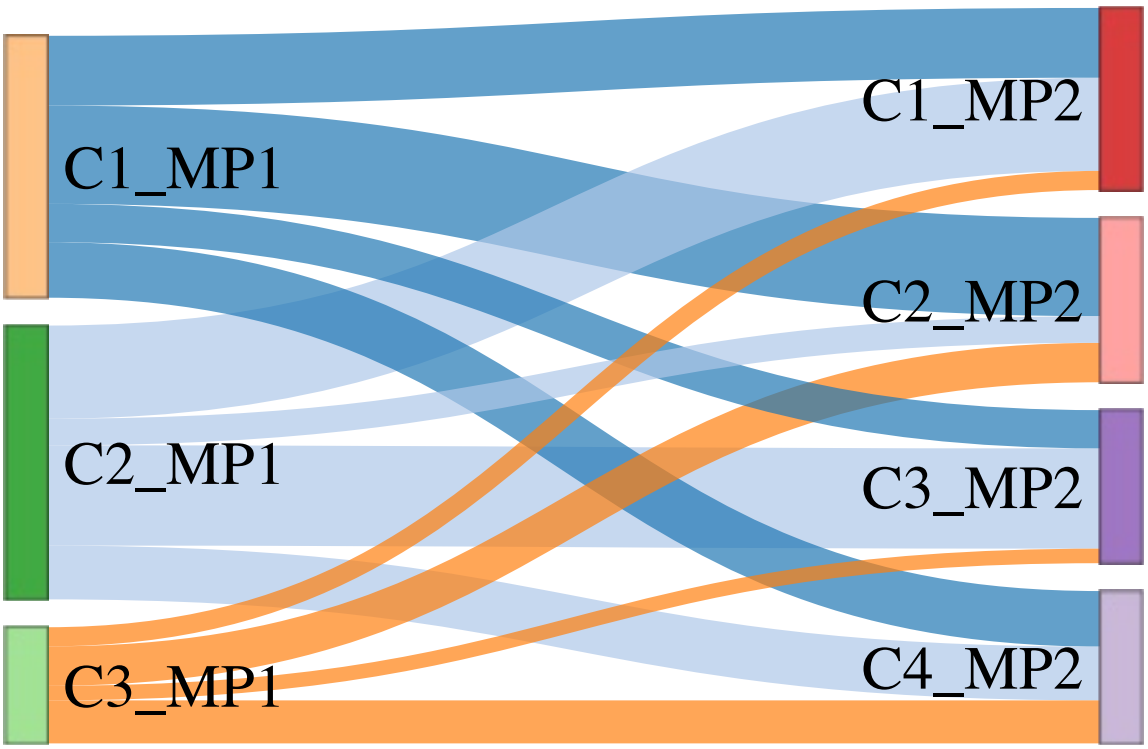
```
#> [1] "lambda min is at location 34"
#> [1] "the leave-out cells in the source subpop is 81"
#> [1] "use 82 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 163 cells and 193 genes..."
#> [1] "evaluation accuracy ElasticNet 0.711656441717791"
#> [1] "evaluation accuracy LDA 0.515337423312883"
#> [1] "done bootstrap 1"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 14.4385026737968"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 35.8288770053476"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 17.1428571428571"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 62.8571428571429"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 11.2781954887218"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 34.5864661654135"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 25"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 47.5"
#> [1] "Total 163 cells as source subpop"
#> [1] "Total 427 cells in remaining subpops"
#> [1] "subsampling 82 cells for training souce subpop"
#> [1] "subsampling 82 cells in remaining subpops for training"
#> [1] "use 196 genes for training model"
#> [1] "use 196 genes 164 cells for testing model"
#> [1] "rename remaining subpops to 2_1"
#> [1] "there are 82 cells in class 2_1 and 82 cells in class 3"
#> [1] "removing 3 genes with no variance"
#> [1] "standardizing prediction/target dataset"
#> [1] "performning elasticnet model training..."
#> [1] "performning LDA model training..."
#> [1] "extracting deviance and best gene features..."
#>
#> Call:  glmnet(x = as.matrix(dataset[, -which(colnames(dataset) == "Cluster_class")]),      y = y_cat
```

```
#>
#>        Df      %Dev   Lambda
#>   [1,]  0 2.403e-15 0.178700
#>   [2,]  2 8.267e-03 0.170600
#>   [3,]  3 1.799e-02 0.162900
#>   [4,]  4 2.857e-02 0.155500
#>   [5,]  5 3.929e-02 0.148400
#>   [6,]  5 5.253e-02 0.141700
#>   [7,]  6 6.557e-02 0.135200
#>   [8,]  9 7.867e-02 0.129100
#>   [9,] 10 9.412e-02 0.123200
#>  [10,] 11 1.091e-01 0.117600
#>  [11,] 11 1.234e-01 0.112300
#>  [12,] 11 1.368e-01 0.107200
#>  [13,] 11 1.492e-01 0.102300
#>  [14,] 12 1.621e-01 0.097640
#>  [15,] 13 1.746e-01 0.093200
#>  [16,] 14 1.870e-01 0.088960
#>  [17,] 15 1.995e-01 0.084920
#>  [18,] 18 2.136e-01 0.081060
#>  [19,] 19 2.282e-01 0.077380
#>  [20,] 20 2.423e-01 0.073860
#>  [21,] 22 2.565e-01 0.070500
#>  [22,] 24 2.720e-01 0.067300
#>  [23,] 26 2.873e-01 0.064240
#>  [24,] 28 3.045e-01 0.061320
#>  [25,] 29 3.213e-01 0.058530
#>  [26,] 28 3.371e-01 0.055870
#>  [27,] 28 3.520e-01 0.053330
#>  [28,] 27 3.660e-01 0.050910
#>  [29,] 29 3.803e-01 0.048590
#>  [30,] 29 3.940e-01 0.046390
#>  [31,] 30 4.071e-01 0.044280
#>  [32,] 32 4.200e-01 0.042260
#>  [33,] 34 4.330e-01 0.040340
#>  [34,] 35 4.465e-01 0.038510
#>  [35,] 38 4.594e-01 0.036760
#>  [36,] 39 4.726e-01 0.035090
#>  [37,] 40 4.860e-01 0.033490
#>  [38,] 41 4.994e-01 0.031970
#>  [39,] 42 5.124e-01 0.030520
#>  [40,] 42 5.249e-01 0.029130
#>  [41,] 43 5.374e-01 0.027810
#>  [42,] 45 5.497e-01 0.026540
#>  [43,] 48 5.624e-01 0.025340
#>  [44,] 49 5.756e-01 0.024190
#>  [45,] 52 5.885e-01 0.023090
#>  [46,] 52 6.016e-01 0.022040
#>  [47,] 52 6.142e-01 0.021040
#>  [48,] 53 6.262e-01 0.020080
#>  [49,] 55 6.379e-01 0.019170
#>  [50,] 56 6.493e-01 0.018300
#>  [51,] 56 6.604e-01 0.017460
```

```
#>  [52,] 58 6.714e-01 0.016670
#>  [53,] 59 6.824e-01 0.015910
#>  [54,] 62 6.932e-01 0.015190
#>  [55,] 62 7.037e-01 0.014500
#>  [56,] 63 7.139e-01 0.013840
#>  [57,] 63 7.238e-01 0.013210
#>  [58,] 65 7.334e-01 0.012610
#>  [59,] 68 7.433e-01 0.012040
#>  [60,] 68 7.530e-01 0.011490
#>  [61,] 68 7.625e-01 0.010970
#>  [62,] 68 7.716e-01 0.010470
#>  [63,] 68 7.804e-01 0.009993
#>  [64,] 69 7.890e-01 0.009539
#>  [65,] 69 7.973e-01 0.009106
#>  [66,] 68 8.054e-01 0.008692
#>  [67,] 70 8.133e-01 0.008297
#>  [68,] 71 8.210e-01 0.007920
#>  [69,] 72 8.285e-01 0.007560
#>  [70,] 73 8.358e-01 0.007216
#>  [71,] 73 8.429e-01 0.006888
#>  [72,] 73 8.496e-01 0.006575
#>  [73,] 74 8.561e-01 0.006276
#>  [74,] 75 8.624e-01 0.005991
#>  [75,] 75 8.685e-01 0.005719
#>  [76,] 78 8.745e-01 0.005459
#>  [77,] 80 8.802e-01 0.005211
#>  [78,] 80 8.856e-01 0.004974
#>  [79,] 81 8.909e-01 0.004748
#>  [80,] 82 8.959e-01 0.004532
#>  [81,] 82 9.006e-01 0.004326
#>  [82,] 83 9.052e-01 0.004129
#>  [83,] 82 9.096e-01 0.003942
#>  [84,] 83 9.137e-01 0.003762
#>  [85,] 83 9.177e-01 0.003591
#>  [86,] 85 9.215e-01 0.003428
#>  [87,] 86 9.251e-01 0.003272
#>  [88,] 86 9.286e-01 0.003124
#>  [89,] 87 9.319e-01 0.002982
#>  [90,] 87 9.351e-01 0.002846
#>  [91,] 87 9.381e-01 0.002717
#>  [92,] 87 9.410e-01 0.002593
#>  [93,] 87 9.437e-01 0.002475
#>  [94,] 88 9.463e-01 0.002363
#>  [95,] 89 9.488e-01 0.002256
#>  [96,] 89 9.512e-01 0.002153
#>  [97,] 89 9.534e-01 0.002055
#>  [98,] 89 9.556e-01 0.001962
#>  [99,] 89 9.576e-01 0.001873
#> [100,] 89 9.596e-01 0.001787
#> [1] "lambda min is at location 44"
#> [1] "the leave-out cells in the source subpop is 81"
#> [1] "use 82 target subpops cells for leave-out test set"
#> [1] "standardizing the leave-out target and source subpops..."
```

```
#> [1] "start ElasticNet prediction for estimating accuracy..."
#> [1] "start LDA prediction for estimating accuracy for 163 cells and 193 genes..."
#> [1] "evaluation accuracy ElasticNet 0.693251533742331"
#> [1] "evaluation accuracy LDA 0.570552147239264"
#> [1] "done bootstrap 2"
#> [1] "standardizing target subpops before prediction..."
#> [1] "predicting from source to target subpop 1..."
#> [1] "number of cells in the target subpop 1 is 187"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 1 is 12.2994652406417"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 1 is 36.3636363636364"
#> [1] "predicting from source to target subpop 2..."
#> [1] "number of cells in the target subpop 2 is 140"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 2 is 37.8571428571429"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 2 is 80"
#> [1] "predicting from source to target subpop 3..."
#> [1] "number of cells in the target subpop 3 is 133"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 3 is 9.02255639097744"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 3 is 32.3308270676692"
#> [1] "predicting from source to target subpop 4..."
#> [1] "number of cells in the target subpop 4 is 40"
#> [1] "running elasticNet classification..."
#> [1] "class probability prediction ElasticNet for target subpop 4 is 35"
#> [1] "add 2 random indexes for genes in model but not in target subpop, later to be replaced by NA"
#> [1] "running LDA classification..."
#> [1] "class probability prediction LDA for target subpop 4 is 47.5"

#prepare table input for sankey plot

LASSO_C1S1  <- reformat_LASSO(c_selectID=1, mp_selectID = 1, LSOLDA_dat=LSOLDA_dat1,
                          nPredSubpop = row_cluster, Nodes_group = "#7570b3")

LASSO_C2S1  <- reformat_LASSO(c_selectID=2, mp_selectID = 1, LSOLDA_dat=LSOLDA_dat2,
                          nPredSubpop = row_cluster, Nodes_group = "#1b9e77")

LASSO_C3S1  <- reformat_LASSO(c_selectID=3, mp_selectID = 1, LSOLDA_dat=LSOLDA_dat3,
                          nPredSubpop = row_cluster, Nodes_group = "#e7298a")


combined <- rbind(LASSO_C1S1,LASSO_C2S1,LASSO_C3S1)

nboots = 2
#links: source, target, value
#source: node, nodegroup
combined_D3obj <-list(Nodes=combined[,(nboots+3):(nboots+4)], Links=combined[,c((nboots+2):(nboots+1),n
```

```r
combined <- combined[is.na(combined$Value) != TRUE,]


library(networkD3)

Node_source <- as.vector(sort(unique(combined_D3obj$Links$Source)))
Node_target <- as.vector(sort(unique(combined_D3obj$Links$Target)))
Node_all <-unique(c(Node_source, Node_target))

#assign IDs for Source (start from 0)
Source <-combined_D3obj$Links$Source
Target <- combined_D3obj$Links$Target

for(i in 1:length(Node_all)){
  Source[Source==Node_all[i]] <-i-1
  Target[Target==Node_all[i]] <-i-1
}

combined_D3obj$Links$Source <- as.numeric(Source)
combined_D3obj$Links$Target <- as.numeric(Target)
combined_D3obj$Links$LinkColor <- combined$NodeGroup

#prepare node info
node_df <-data.frame(Node=Node_all)
node_df$id <-as.numeric(c(0, 1:(length(Node_all)-1)))

suppressMessages(library(dplyr))
n <- length(unique(node_df$Node))
getPalette = colorRampPalette(RColorBrewer::brewer.pal(9, "Set1"))
Color = getPalette(n)
node_df$color <- Color
suppressMessages(library(networkD3))
p1<-sankeyNetwork(Links =combined_D3obj$Links, Nodes = node_df,  Value = "Value", NodeGroup ="color", Li
                  fontSize = 22 )
p1
```




```
#saveNetwork(p1, file = paste0(path,'Subpopulation_Net.html'))
##R Setting Information
#sessionInfo()
#rmarkdown::render("/Users/quan.nguyen/Documents/Powell_group_MacQuan/AllCodes/scGPS/vignettes/vignette
#rmarkdown::render("/Users/quan.nguyen/Documents/Powell_group_MacQuan/AllCodes/scGPS/vignettes/vignette
```

4.4 Annotation: scGPS prediction can be used to compare scGPS clusters with a reference dataset to see which cluster is most similar to the reference