

## Contents

---

- [addpath to all tools](#)
- [initiate one instance of the "cooccurrence\\_clustering\\_analysis" class](#)
- [read data - 10X format](#)
- [filter the data by removing genes and cells \(same as Seurat tutorial on this data\)](#)
- [binarize data](#)
- [cooccurrence clustering](#)
- [re-order pathways for better visualization](#)
- [Visualize the percentages of detection for pathways in individual cells](#)
- [export results to files](#)

### addpath to all tools

```
addpath(genpath('..\tools\'))
```

### initiate one instance of the "cooccurrence\_clustering\_analysis" class

```
pbmc = cooccurrence_clustering_analysis;
```

### read data - 10X format

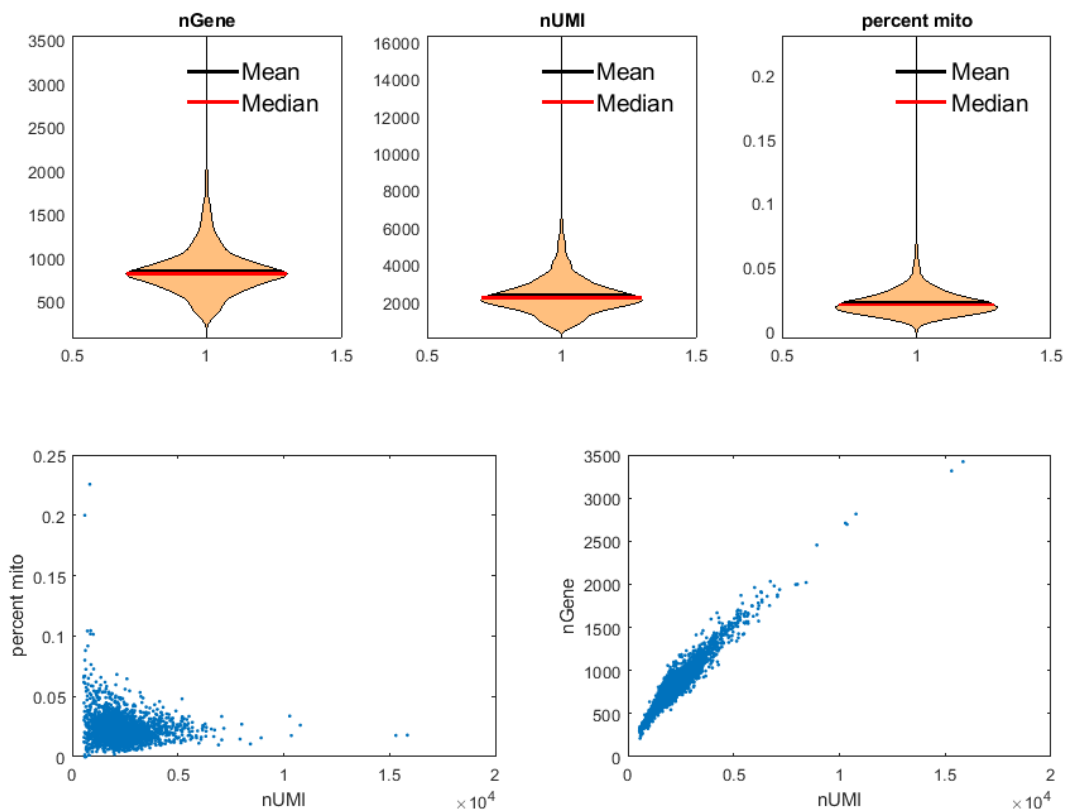
```
pbmc = pbmc.Read10X('Downloads/filtered_gene_bc_matrices/hg19/');
```

```
Loading cell names ... Elapsed time is 0.003551 seconds.  
Loading gene names ... Elapsed time is 0.065075 seconds.  
Loading raw data ... Elapsed time is 1.042454 seconds.  
Converting data of 32738 genes * 2700 cells into matrix ... Elapsed time is 0.175587 seconds.
```

### filter the data by removing genes and cells (same as Seurat tutorial on this data)

```
pbmc.initial_filtering_min_num_cells = 3;  
pbmc.initial_filtering_min_num_genes = 200;  
pbmc.initial_filtering_max_num_genes = 2500;  
pbmc.initial_filtering_max_percent_mito = 0.05;  
pbmc = pbmc.initial_filtering_of_data(1);
```

```
Data after initial filtering 13714 genes * 2638 cells.
```



#### binarize data

```
pbmc.binarization_threshold = 0;
pbmc.binary_data = full(double(pbmc.data>pbmc.binarization_threshold));
```

#### cooccurrence clustering

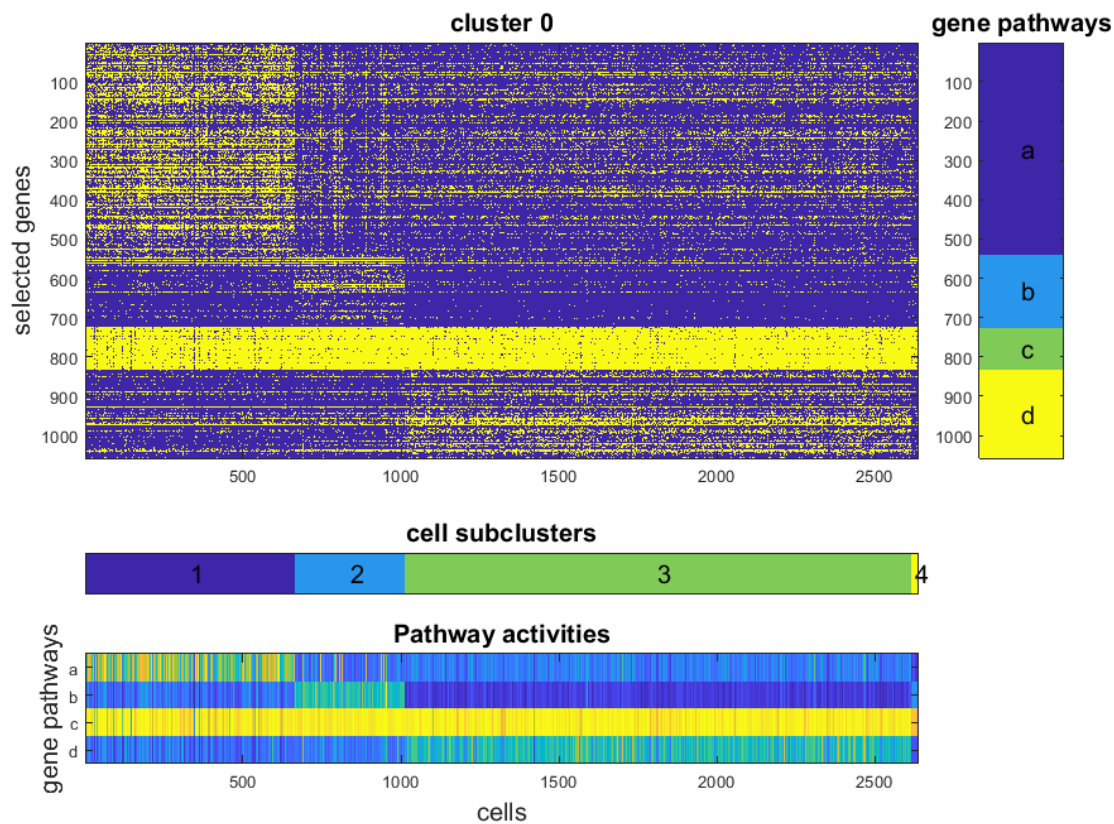
```
pbmc.cooccurrence_min_expressed_cells = 10;           % genes will only be considered if detected in >= minimum number of cells, and undetected in >= minimum number of cells
pbmc.cooccurrence_min_pathway_size = 20;            % only considered gene clusters of size >= this threshold
pbmc.cooccurrence_min_population_size = 10;        % only considered gene clusters of size >= this threshold
pbmc.cooccurrence_snr_merge_threshold = 1.5;      % threshold for merging Louvain communities, based on snr of average detection of each gene cluster
pbmc.cooccurrence_mean_diff_merge_threshold = 0.5; % threshold for merging Louvain communities
pbmc.cooccurrence_mean_ratio_merge_threshold = 2; % threshold for merging Louvain communities

pbmc = pbmc.iterative_cooccurrence_clustering;
```

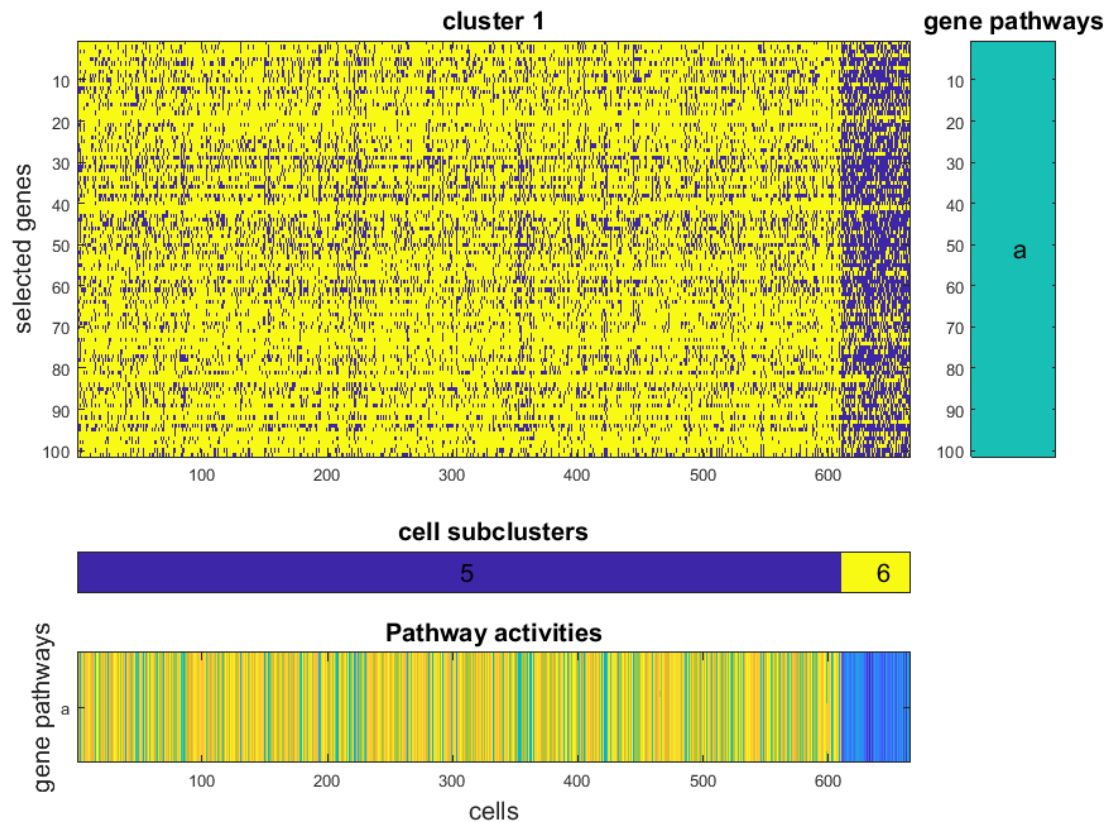
```
Remaining clusters to partition 1
Processing cluster 0 now ...
Processing data subset with 13714 genes and 2638 cells:
Remove genes detected in <10 cells. Remaining 11085 genes. Elapsed time is 0.176278 seconds.
Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 137.727612 seconds.
Compute gene-gene similarity ... Elapsed time is 10.971495 seconds.
Create gene-gene graph for clustering genes ...
Writing graph into file ... 100%Elapsed time is 0.971058 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 1.222256 seconds.
Gene-gene graph contains 7 pathways, 1498 genes in total
Elapsed time is 1.643502 seconds.
```

## step\_02\_cooccurrence\_clustering\_pipeline

```
Create cell-cell graph for clustering cells ...
Writing graph into file ... 100%Elapsed time is 0.307362 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 1.668345 seconds.
Cell-cell graph contains 15 cell types by community detection
Elapsed time is 1.727648 seconds.
Cell-cell graph contains 15 cell types after merging tiny cell clusters
Cell-cell graph contains 4 cell types after merging
Number of useful pathways is 4
```



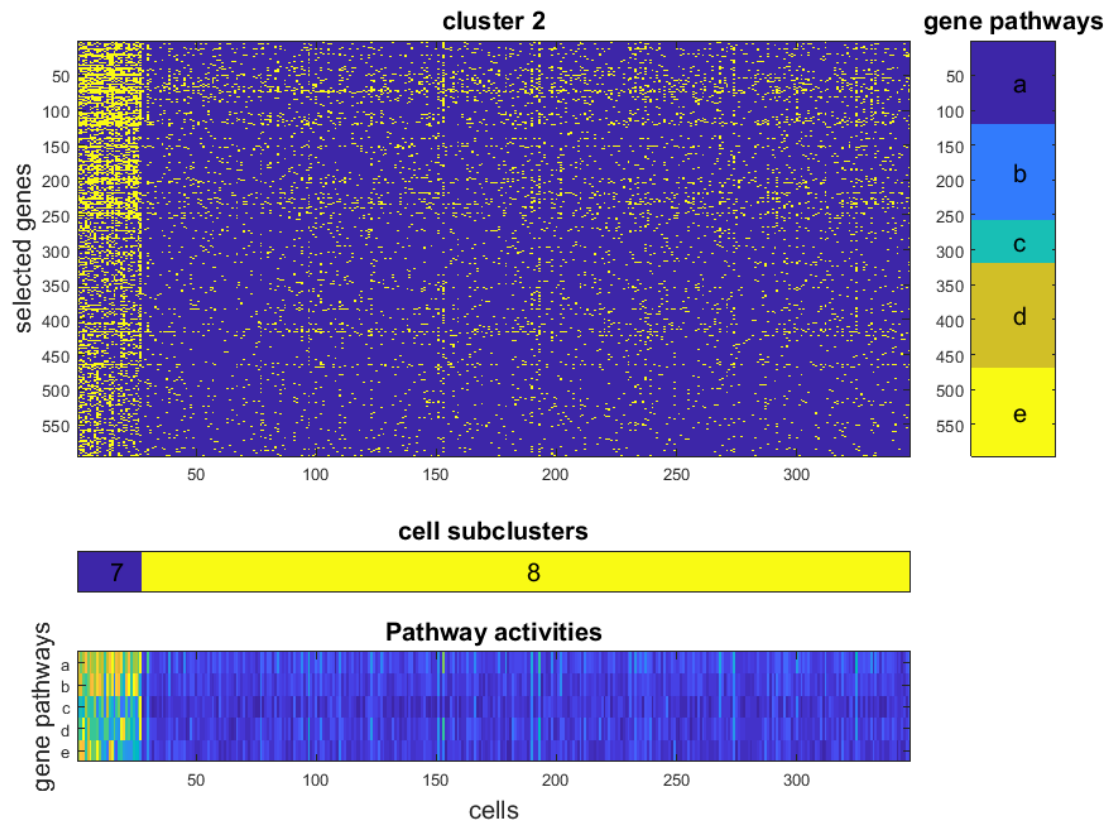
```
Remaining clusters to partition 4
Processing cluster 1 now ...
Processing data subset with 13714 genes and 665 cells:
Remove genes detected in <10 cells. Remaining 7676 genes. Elapsed time is 0.045337 seconds.
Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 45.340451 seconds.
Compute gene-gene similarity ... Elapsed time is 3.605890 seconds.
Create gene-gene graph for clustering genes ...
Writing graph into file ... 100%Elapsed time is 0.390935 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.621711 seconds.
Gene-gene graph contains 5 pathways, 492 genes in total
Elapsed time is 0.866124 seconds.
Create cell-cell graph for clustering cells ...
Writing graph into file ... 100%Elapsed time is 0.074411 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.446810 seconds.
Cell-cell graph contains 8 cell types by community detection
Elapsed time is 0.463286 seconds.
Cell-cell graph contains 8 cell types after merging tiny cell clusters
Cell-cell graph contains 2 cell types after merging
Number of useful pathways is 1
```



```

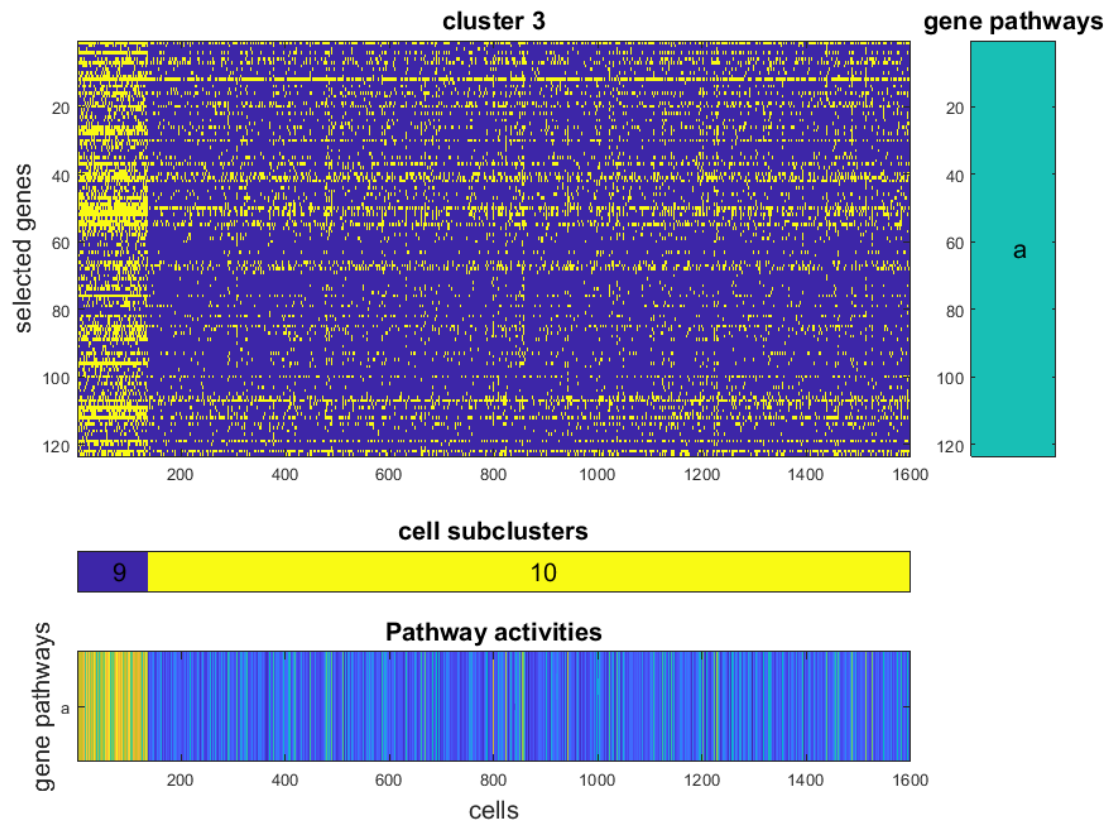
Remaining clusters to partition 5
Processing cluster 2 now ...
Processing data subset with 13714 genes and 347 cells:
Remove genes detected in <10 cells. Remaining 5592 genes. Elapsed time is 0.019993 seconds.
Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 22.492484 seconds.
Compute gene-gene similarity ... Elapsed time is 2.051865 seconds.
Create gene-gene graph for clustering genes ...
Writing graph into file ... 100%Elapsed time is 0.211891 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.665054 seconds.
Gene-gene graph contains 7 pathways, 650 genes in total
Elapsed time is 0.846275 seconds.
Create cell-cell graph for clustering cells ...
Writing graph into file ... 100%Elapsed time is 0.045104 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.350452 seconds.
Cell-cell graph contains 6 cell types by community detection
Elapsed time is 0.359823 seconds.
Cell-cell graph contains 6 cell types after merging tiny cell clusters
Cell-cell graph contains 2 cell types after merging
Number of useful pathways is 5

```



```

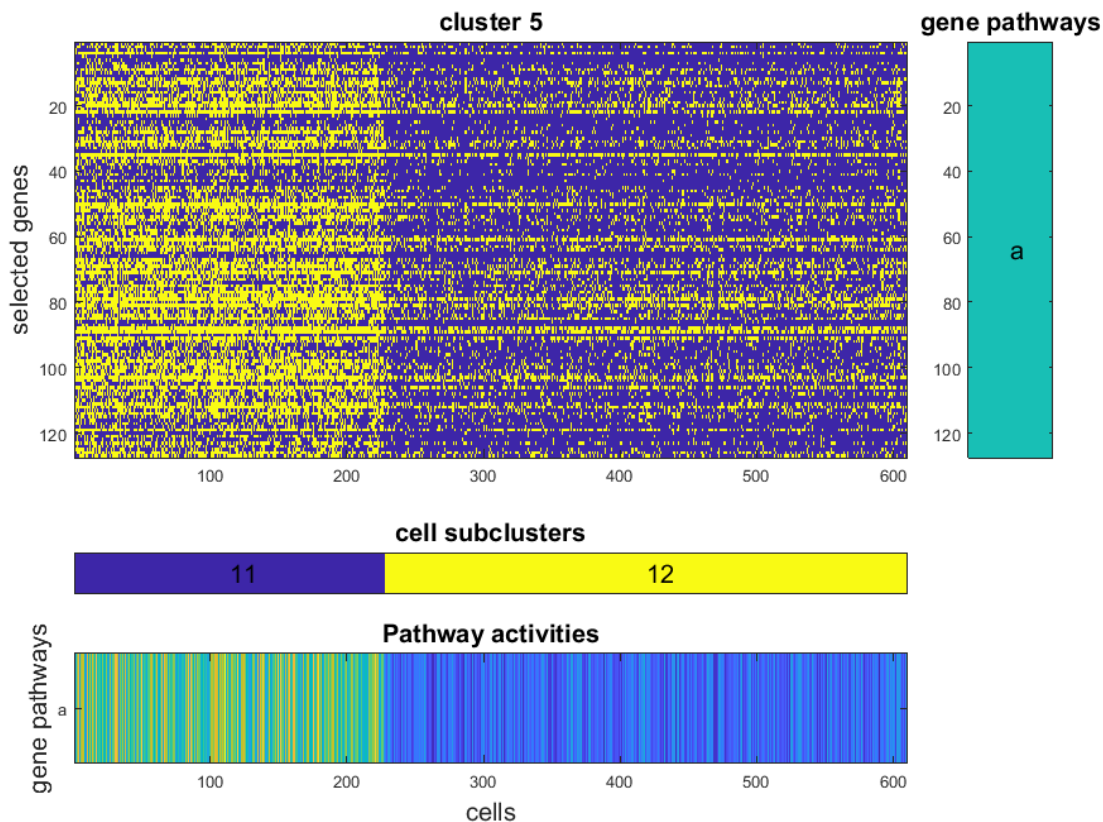
Remaining clusters to partition 6
Processing cluster 3 now ...
Processing data subset with 13714 genes and 1600 cells:
Remove genes detected in <10 cells. Remaining 9646 genes. Elapsed time is 0.098636 seconds.
Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 85.974532 seconds.
Compute gene-gene similarity ... Elapsed time is 6.617934 seconds.
Create gene-gene graph for clustering genes ...
Writing graph into file ... 100%Elapsed time is 0.575993 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.830773 seconds.
Gene-gene graph contains 6 pathways, 374 genes in total
Elapsed time is 1.160087 seconds.
Create cell-cell graph for clustering cells ...
Writing graph into file ... 100%Elapsed time is 0.183530 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.889321 seconds.
Cell-cell graph contains 10 cell types by community detection
Elapsed time is 0.924508 seconds.
Cell-cell graph contains 9 cell types after merging tiny cell clusters
creating a total of 8 edges ...      8
Cell-cell graph contains 2 cell types after merging
Number of useful pathways is 1
    
```



Remaining clusters to partition 7  
 Processing cluster 4 now ...  
 Processing data subset with 13714 genes and 26 cells:  
 Remove genes detected in <10 cells. Remaining 74 genes. Elapsed time is 0.000780 seconds.  
 Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 0.010259 seconds.  
 Compute gene-gene similarity ... Elapsed time is 0.000599 seconds.  
 Create gene-gene graph for clustering genes ...  
 Writing graph into file ... 111% Elapsed time is 0.001581 seconds.  
 Running ModularityOptimizer for clustering ... Elapsed time is 0.147289 seconds.  
 Gene-gene graph contains 0 pathways, 0 genes in total  
 Elapsed time is 0.151635 seconds.

Remaining clusters to partition 6  
 Processing cluster 5 now ...  
 Processing data subset with 13714 genes and 610 cells:  
 Remove genes detected in <10 cells. Remaining 7544 genes. Elapsed time is 0.036292 seconds.  
 Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 41.610551 seconds.  
 Compute gene-gene similarity ... Elapsed time is 3.352553 seconds.  
 Create gene-gene graph for clustering genes ...  
 Writing graph into file ... 100% Elapsed time is 0.357293 seconds.  
 Running ModularityOptimizer for clustering ... Elapsed time is 0.547415 seconds.  
 Gene-gene graph contains 3 pathways, 171 genes in total  
 Elapsed time is 0.780370 seconds.  
 Create cell-cell graph for clustering cells ...  
 Writing graph into file ... 100% Elapsed time is 0.066126 seconds.  
 Running ModularityOptimizer for clustering ... Elapsed time is 0.402017 seconds.  
 Cell-cell graph contains 10 cell types by community detection  
 Elapsed time is 0.416640 seconds.  
 Cell-cell graph contains 8 cell types after merging tiny cell clusters  
 creating a total of 7 edges ... 7

Cell-cell graph contains 2 cell types after merging  
 Number of useful pathways is 1



Remaining clusters to partition 7  
 Processing cluster 6 now ...  
 Processing data subset with 13714 genes and 55 cells:  
 Remove genes detected in <10 cells. Remaining 428 genes. Elapsed time is 0.001657 seconds.  
 Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 0.116097 seconds.  
 Compute gene-gene similarity ... Elapsed time is 0.009520 seconds.  
 Create gene-gene graph for clustering genes ...  
 Writing graph into file ... 100% Elapsed time is 0.012185 seconds.  
 Running ModularityOptimizer for clustering ... Elapsed time is 0.240120 seconds.  
 Gene-gene graph contains 2 pathways, 126 genes in total  
 Elapsed time is 0.252659 seconds.  
 Create cell-cell graph for clustering cells ...  
 Writing graph into file ... 100% Elapsed time is 0.006548 seconds.  
 Running ModularityOptimizer for clustering ... Elapsed time is 0.193435 seconds.  
 Cell-cell graph contains 3 cell types by community detection  
 Elapsed time is 0.197232 seconds.  
 Cell-cell graph contains 2 cell types after merging tiny cell clusters  
 creating a total of 1 edges ... 1  
 Cell-cell graph contains 1 cell types after merging

Remaining clusters to partition 6  
 Processing cluster 7 now ...  
 Processing data subset with 13714 genes and 27 cells:  
 Remove genes detected in <10 cells. Remaining 857 genes. Elapsed time is 0.000830 seconds.  
 Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 0.440201 seconds.  
 Compute gene-gene similarity ... Elapsed time is 0.037616 seconds.

## step\_02\_cooccurrence\_clustering\_pipeline

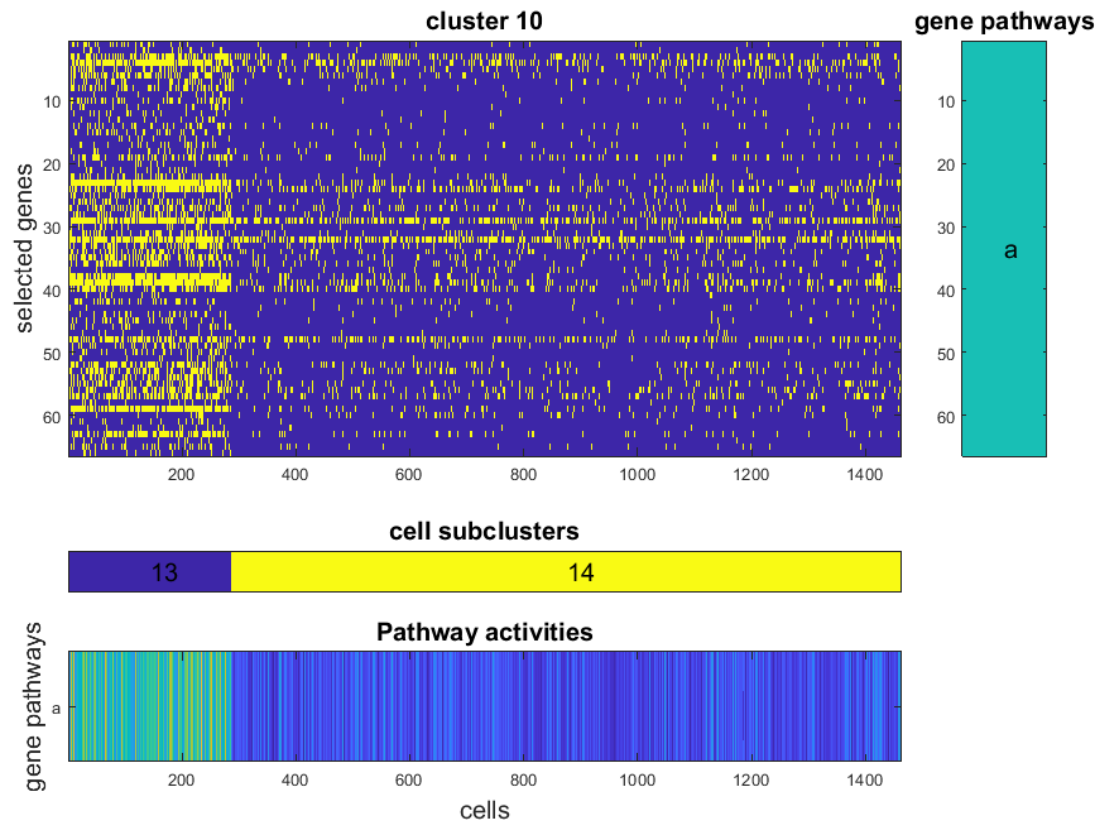
```
Create gene-gene graph for clustering genes ...
Writing graph into file ... 101%Elapsed time is 0.006394 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.224279 seconds.
Gene-gene graph contains 0 pathways, 0 genes in total
Elapsed time is 0.245747 seconds.

Remaining clusters to partition 5
Processing cluster 8 now ...
Processing data subset with 13714 genes and 320 cells:
Remove genes detected in <10 cells. Remaining 4868 genes. Elapsed time is 0.016674 seconds.
Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 16.433138 seconds.
Compute gene-gene similarity ... Elapsed time is 1.327371 seconds.
Create gene-gene graph for clustering genes ...
Writing graph into file ... 100%Elapsed time is 0.149201 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.439159 seconds.
Gene-gene graph contains 1 pathways, 23 genes in total
Elapsed time is 0.574739 seconds.
Create cell-cell graph for clustering cells ...
Writing graph into file ... 100%Elapsed time is 0.047160 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.273379 seconds.
Cell-cell graph contains 8 cell types by community detection
Elapsed time is 0.283181 seconds.
Cell-cell graph contains 2 cell types after merging tiny cell clusters
creating a total of 1 edges ... 1
Cell-cell graph contains 1 cell types after merging

Remaining clusters to partition 4
Processing cluster 9 now ...
Processing data subset with 13714 genes and 137 cells:
Remove genes detected in <10 cells. Remaining 3058 genes. Elapsed time is 0.006839 seconds.
Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 5.930400 seconds.
Compute gene-gene similarity ... Elapsed time is 0.500944 seconds.
Create gene-gene graph for clustering genes ...
Writing graph into file ... 100%Elapsed time is 0.061691 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.326695 seconds.
Gene-gene graph contains 0 pathways, 0 genes in total
Elapsed time is 0.405420 seconds.

Remaining clusters to partition 3
Processing cluster 10 now ...
Processing data subset with 13714 genes and 1463 cells:
Remove genes detected in <10 cells. Remaining 9382 genes. Elapsed time is 0.091390 seconds.
Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 95.092016 seconds.
Compute gene-gene similarity ... Elapsed time is 6.519133 seconds.
Create gene-gene graph for clustering genes ...
Writing graph into file ... 100%Elapsed time is 0.595015 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.646312 seconds.
Gene-gene graph contains 3 pathways, 129 genes in total
Elapsed time is 0.977788 seconds.
Create cell-cell graph for clustering cells ...
Writing graph into file ... 100%Elapsed time is 0.175151 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.718753 seconds.
Cell-cell graph contains 19 cell types by community detection
Elapsed time is 0.753739 seconds.
Cell-cell graph contains 15 cell types after merging tiny cell clusters
Cell-cell graph contains 2 cell types after merging
Number of useful pathways is 1
```





Remaining clusters to partition 4  
 Processing cluster 11 now ...  
 Processing data subset with 13714 genes and 227 cells:  
 Remove genes detected in <10 cells. Remaining 5324 genes. Elapsed time is 0.014972 seconds.  
 Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 18.881118 seconds.  
 Compute gene-gene similarity ... Elapsed time is 1.578075 seconds.  
 Create gene-gene graph for clustering genes ...  
 Writing graph into file ... 100% Elapsed time is 0.170057 seconds.  
 Running ModularityOptimizer for clustering ... Elapsed time is 0.428300 seconds.  
 Gene-gene graph contains 0 pathways, 0 genes in total  
 Elapsed time is 0.577106 seconds.

Remaining clusters to partition 3  
 Processing cluster 12 now ...  
 Processing data subset with 13714 genes and 383 cells:  
 Remove genes detected in <10 cells. Remaining 5401 genes. Elapsed time is 0.023067 seconds.  
 Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 20.651082 seconds.  
 Compute gene-gene similarity ... Elapsed time is 1.645740 seconds.  
 Create gene-gene graph for clustering genes ...  
 Writing graph into file ... 100% Elapsed time is 0.183222 seconds.  
 Running ModularityOptimizer for clustering ... Elapsed time is 0.392395 seconds.  
 Gene-gene graph contains 0 pathways, 0 genes in total  
 Elapsed time is 0.545966 seconds.

Remaining clusters to partition 2  
 Processing cluster 13 now ...  
 Processing data subset with 13714 genes and 287 cells:  
 Remove genes detected in <10 cells. Remaining 5149 genes. Elapsed time is 0.017176 seconds.  
 Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 18.021172 seconds.  
 Compute gene-gene similarity ... Elapsed time is 1.475066 seconds.

## step\_02\_cooccurrence\_clustering\_pipeline

```
Create gene-gene graph for clustering genes ...
Writing graph into file ... 100%Elapsed time is 0.168623 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.438410 seconds.
Gene-gene graph contains 0 pathways, 0 genes in total
Elapsed time is 0.580326 seconds.

Remaining clusters to partition 1
Processing cluster 14 now ...
Processing data subset with 13714 genes and 1176 cells:
Remove genes detected in <10 cells. Remaining 8838 genes. Elapsed time is 0.075252 seconds.
Iterate 10 random permutations for gene-gene similarity threshold ... 10 Elapsed time is 67.357760 seconds.
Compute gene-gene similarity ... Elapsed time is 5.311134 seconds.
Create gene-gene graph for clustering genes ...
Writing graph into file ... 100%Elapsed time is 0.476414 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.535841 seconds.
Gene-gene graph contains 1 pathways, 28 genes in total
Elapsed time is 0.818421 seconds.
Create cell-cell graph for clustering cells ...
Writing graph into file ... 100%Elapsed time is 0.179597 seconds.
Running ModularityOptimizer for clustering ...Elapsed time is 0.531841 seconds.
Cell-cell graph contains 15 cell types by community detection
Elapsed time is 0.558412 seconds.
Cell-cell graph contains 2 cell types after merging tiny cell clusters
creating a total of 1 edges ... 1
Cell-cell graph contains 1 cell types after merging
```

## re-order pathways for better visualization

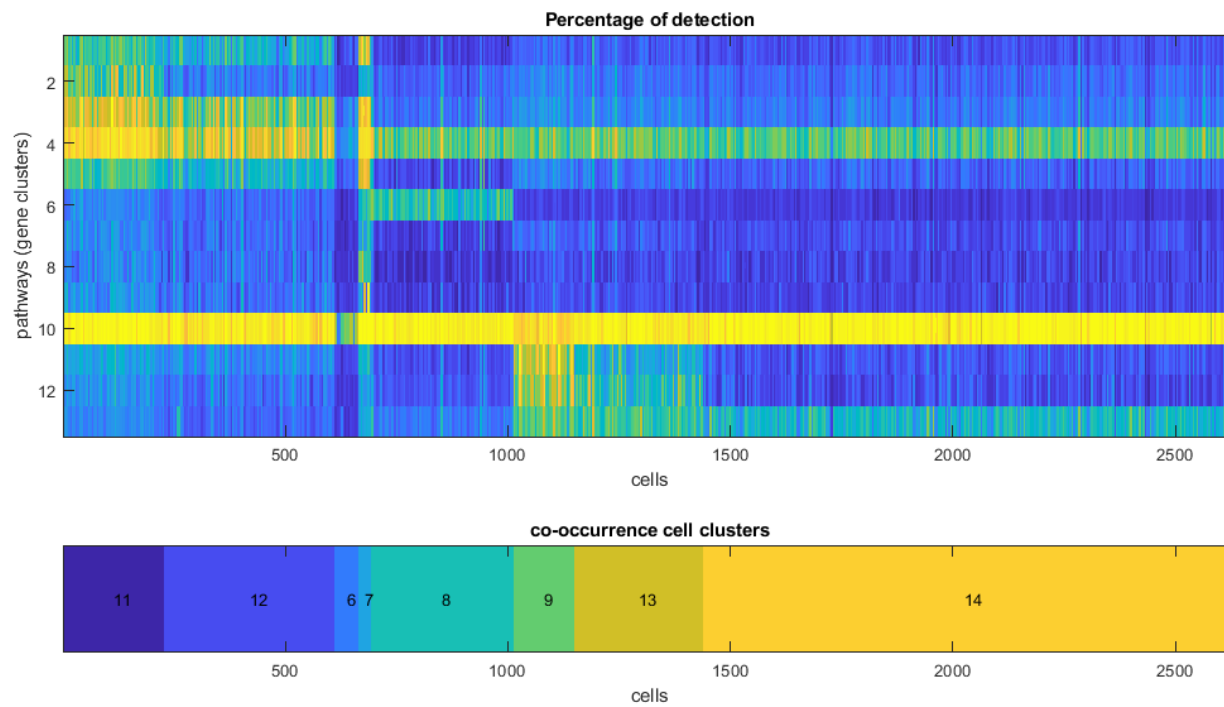
```
pathway_in_cell_detection_rate = pbmc.pathway_in_cell_detection_rate;
scores = zeros(size(pathway_in_cell_detection_rate,1),max(pbmc.cell_labels));
for k=1:max(pbmc.cell_labels)
    for l=k+1:max(pbmc.cell_labels)
        tmp = get_correlations(pathway_in_cell_detection_rate(:,ismember(pbmc.cell_labels,[k,l])),pbmc.cell_labels(ismember(pbmc.cell_labels,[k,l])), 'snr');
        scores(:,k) = scores(:,k) - tmp;
        scores(:,l) = scores(:,l) + tmp;
    end
end
[max_score,signature_for_which_celltype] = max(scores,[],2);
pathway_order = [];
for k=1:max(pbmc.cell_labels)
    tmp_pathway = find(signature_for_which_celltype==k);
    [~,I] = sort(max_score(tmp_pathway),'descend');
    pathway_order = [pathway_order;tmp_pathway(I)];
end
pbmc.pathway_in_cell_detection_rate = pbmc.pathway_in_cell_detection_rate(pathway_order,:);
pbmc.gene_pathway_indicator = pbmc.gene_pathway_indicator(:,pathway_order);

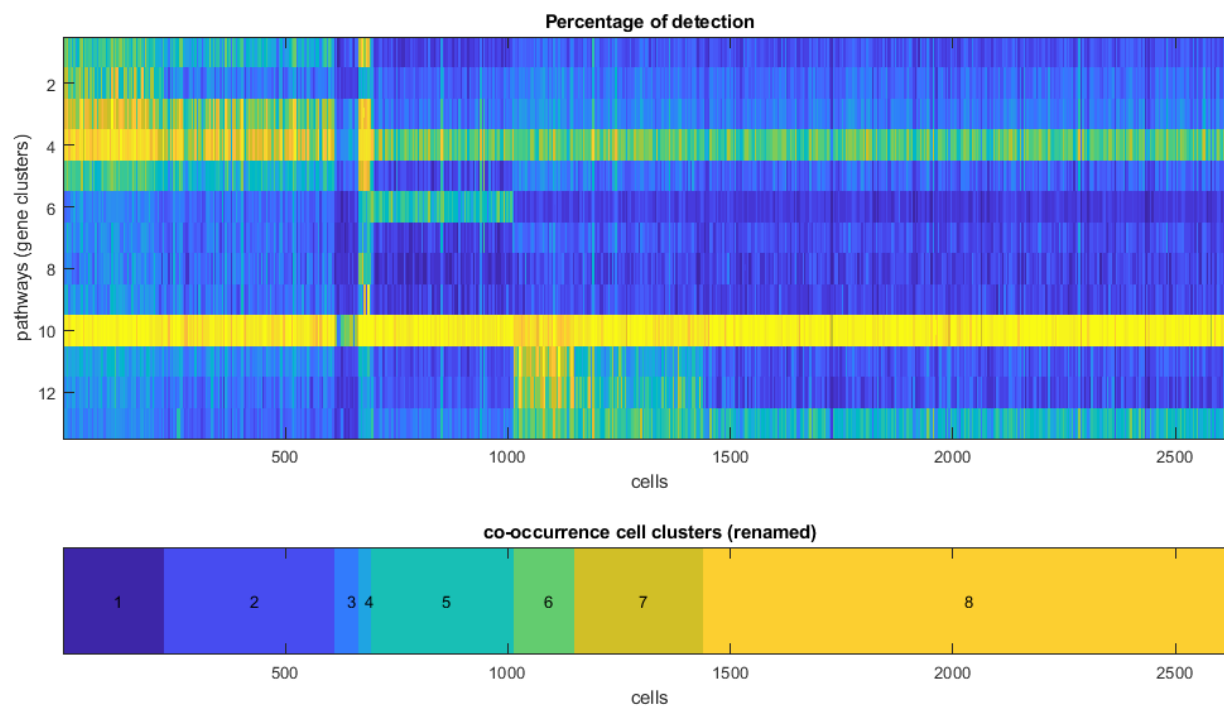
save([mfilename, '_result.mat']);
```

## Visualize the percentages of detection for pathways in individual cells

```
[~,I] = sort(pbmc.cell_labels);
h = figure(30); set(h, 'position', [100 100 1200 600]);
subplot(6,1,1:4); imagesc(pbmc.pathway_in_cell_detection_rate(:, I)./max(pbmc.pathway_in_cell_detection_rate(:, I),[],2)); ylabel('pathways (gene clusters)'); xlabel('cells'); title('Percentage of detection')
h = subplot(4,1,4); imagesc(pbmc.cell_labels(I)); h.YTick=[];
for k=1:max(pbmc.cell_labels)
    text(mean(find(pbmc.cell_labels(I)==k)), 1, num2str(unique(pbmc.cell_labels_history(pbmc.cell_labels==k,end))));
end
title('co-occurrence cell clusters'); xlabel('cells')

h = figure(40); set(h, 'position', [100 100 1200 600]);
subplot(6,1,1:4); imagesc(pbmc.pathway_in_cell_detection_rate(:, I)./max(pbmc.pathway_in_cell_detection_rate(:, I),[],2)); ylabel('pathways (gene clusters)'); xlabel('cells'); title('Percentage of detection')
h = subplot(4,1,4); imagesc(pbmc.cell_labels(I)); h.YTick=[];
for k=1:max(pbmc.cell_labels)
    text(mean(find(pbmc.cell_labels(I)==k)), 1, num2str(k));
end
title('co-occurrence cell clusters (renamed)'); xlabel('cells')
```

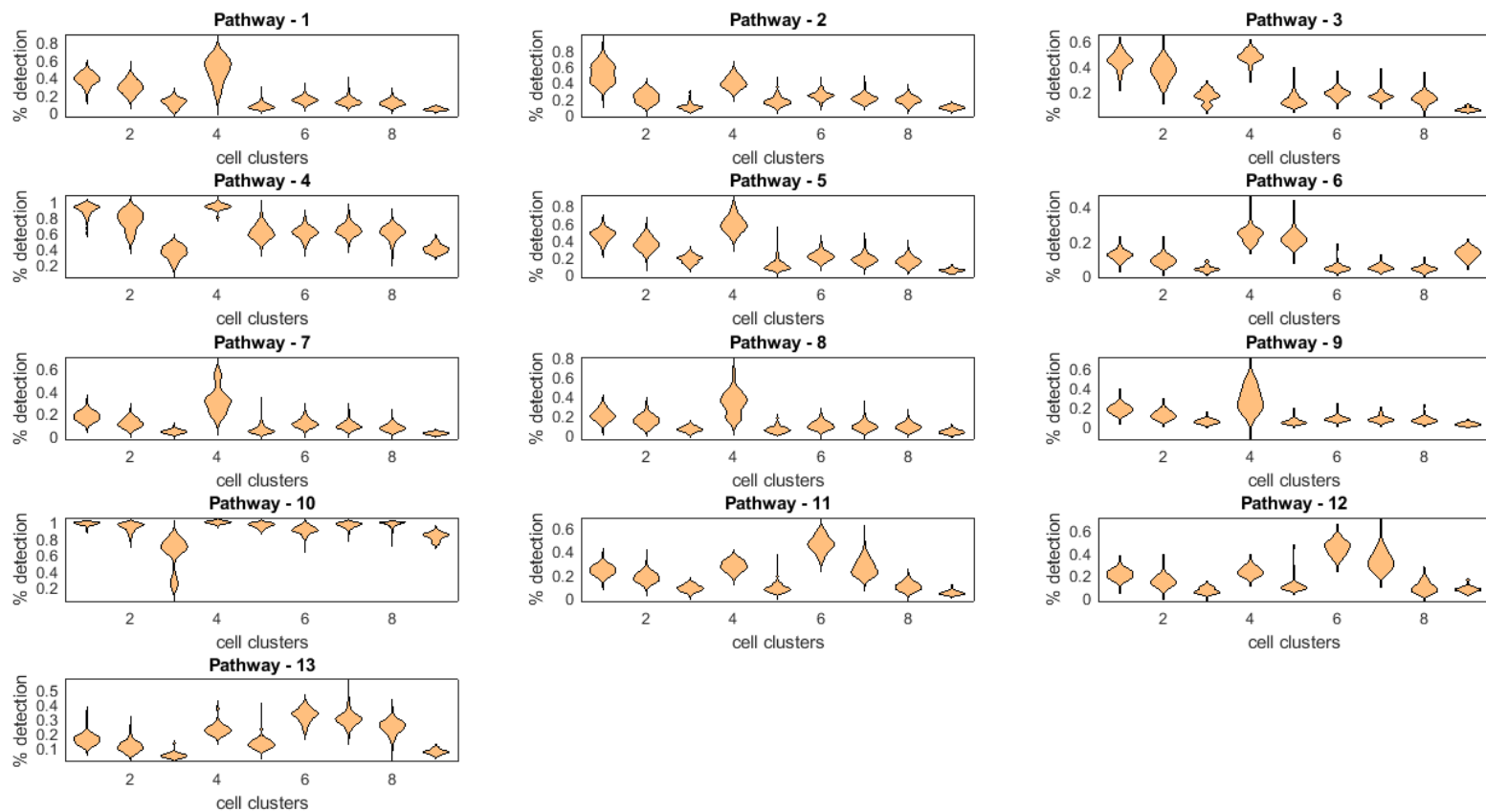




```

h = figure(5); set(h, 'units','normalized','outerposition',[0 0 1 1])
for i=1:size(pbmc.pathway_in_cell_detection_rate,1)
    tmp = cell(0);
    for k=1:max(pbmc.cell_labels)
        tmp{k} = pbmc.pathway_in_cell_detection_rate(i,pbmc.cell_labels==k);
    end
    subplot(5, ceil(size(pbmc.pathway_in_cell_detection_rate,1)/5),i); violin(tmp,'mc',[],'medc',[]);
    title(num2str(i,'Pathway - %d'))
    xlabel('cell clusters');
    ylabel('% detection');
    drawnow
end

```



#### export results to files

```
% pathway_names = strcat('pathway-', regexp(num2str(1:size(pbmc.gene_pathway_indicator,2),'%05d'),'','split')); pathway_names(end)=[];
% write_to_txt_v2([filename,'_result_gene_pathway_indicator.csv'], [{' '}, pathway_names], pbmc.gene_names, pbmc.gene_pathway_indicator, '');
% write_to_txt_v2([filename,'_result_pathway_in_cell_detection_rate.csv'], [{' '}, pbmc.cell_names], pathway_names, pbmc.pathway_in_cell_detection_rate, '');
% write_to_txt_v2([filename,'_result_cell_clusters.csv'], [{' '}, pbmc.cell_names], {'cluster idx'}, pbmc.cell_labels, '');
```