Analysis Spatial pattern

Lorette Noiret

September 13, 2017

1 Study goals

1.1 Aims

Understand the spatio-temporal pattern of gene expression and the link this pattern with the phenotypes (pattern of lamination...).
In a first step, we are focusing on each time point separately. This step should also help to identify regions that will be used for single-cell analyses.

- Identifiation of the regions.
  - How many regions can we identify?
  - Should we identify the regions manually (contouring) or automatically?
  - Should we identify the region based on gene expressions (descritptive analysis, clusetering) or based the phenotypic pattern (PLS?)?
  - is it necessary to apply a mask (defined manually) to find those regions (e.g. exclusion cuticule)? Yohanns also suggested to create a mask automatically (by selecting x number of pixels around the 'zero area')
  - should we average the expression of a specific gene in several animals or only consider one animal at a time?
- Analyse the pattern of some key gene expression.
  - which genes are co-expressed together in each region? Clustering is useful to better understand the pattern even if we do not use the results for the single-cell analysis.
  - do the pattern of gene expression (spatial organization of cluster) correlate with the phenotypic pattern?

Other ideas:
Compare the gene expression profiles with the fly of different size.

2 Step1 - Spatial analysis at 12apf

2.1 Data

Genes express in the notum at 12apf.

Source directory: 'Patt_CellProp_Rescaled_2017_sept_06th_test_raw_data'.

Initial dataset:

- Yohanns and Maria selected 10 genes of interest (ara, BH1, bi, esg, eyg, hth, pct, sd, Sr, Ush) measured in 10 individuals. Better understand the limitations associated with data acquisation : expression, max projection, stiching...
- Some phenotypic map are also available (Delamination, proliferation,TissueDeform,TissueStress)
- a file contains the macrocheataetae

Comments and challenges associated with these data:

- The gene barH1 (bh1) was measured on a smaller notum, so comparisons we should perform the clustering with and without it and compare the results
- Nothing tell us that those genes are the one explaining the spatial phenotypic organization (at that time point). Maybe a PLS before clustering could be useful (more than a PCA)
- Should we pool the data from several animals?
- Image includes the growing cuticule, should we exclude it (by applying a mask delimited by Maria or an automatic approach)

2.2 Statistical Analysis

2.2.1 Workflow

1. read image
2. Apply mask (only if UseMask==1)
3. Normalize each image on 0-1 scale (divide all the pixel by the maximum intensity value). Impact normalization?
4. vectorize image
5. TO DO: impact if we do a PCA before hand?
6. remove unwanted (if pixel=0 on all image then it is a background point and I can exclude it to make the code faster)

7. k-mean 2 to 10 clusters (max iteration 1000).
   - Use 3 to 5 replicates to insure that we find a global minimum.
   - Assess quality clustering: silhouette (how well each object lies within cluster), but not doable of full-size image. BIC but based on log-likelihood (gaussian assumption seems wrong). Wilks statistics

2.2.2 Smoothing

The images have taken in high resolution and the nucleus are apparent (black area). We should try to do some interpolation before clustering?

- Smoothing data: using gaussian filter or another one
- resize: decrease the image size with a filter (resizem). Advantage: faster computation, can compute silhouette.
- kriging: spatial interpolation, can take into account anisotropie and discontinuity TO DO, available in Matlab??

(filtering, resize or kriging)?
Challenges: how to choose the optimal parameters (sigma/distance for gaussian filter, scaling and filter for resizing)?
Smoothing with a distance of 15 pixels remove completely the discountinuous aspects, but also blur the small patterns (becoming non-apparent when create the clusters).

2.2.3 Automatic detection of cluster (methods)

Different approach can be tested:

- Hierarchical approach (AHC): tried it but matlab memory bugged. Maybe I did something wrong or should implement it in Python?
- Kmeans I tried euclidian distance, but could try other metric as well. Ask bioinformatician if some metric works better with gene expression
- Kmedoid: interest if we look at the expression qualitatively (expressed / not expressed). Advantage: center of class is a datapoint (use L1 metric). TO DO, available in matlab
- Quid of spatial clustering. Some approach must exist, check litterature

kmeans To identify the regions, we classify the 10 genes expression map using a k-means algotithm (from 2 to 10 clusters).
How to determine the optimal number of cluster?

- Compare pattern with the one defined by Maria
- Look at some automatic features:

  - silhouette
  - Wilks statistic, variance within cluster, distance between centroids
  - Stability : how robust a clustering solution is under pertubation or sub-sampling
  -

Other technical aspects

- Number of repeats: kmeans with n groups performed several times to limit the impact of centroids inititialization. I took 3, maybe should go up to 5.
- Choice metric. Only tested euclidian so far
- Choice color for clusters. Choice of color makes a difference in term of vizualization and apparent quality of clusters

Open questions (to discuss):

- Data are not continuous (cells border). Should we smooth the image first?
- Should we consider another algorithm? I initially tried hierarchical cluster analysis, but the size was to large, and Matlab crashes. Is there a better way to do it (another sotware? another algorithm?)?

2.2.4 Other points

Does a preliminary PCA change the results? Rather than PCA, PLS could help?

2.2.5 Classification approach

2.3 Results

2.3.1 Descriptive analysis

Visually description of the pattern of each gene and phenotype.

name	some functions	comments
ara (araucan)	TF, notum cell fate specification	expressed latterally (except 2 zones (wings or another phenotype later?), midline +/- empty, a few intense spots
bh1 (BarH1)	TF, chaeta morphogenesis; compound eye photoreceptor cell differentiation	Fly is much smaller → problem?, opposition posterior/anterior, only expressed in region clode to neck + few spot close to scutellum
bi (Bifid)	TF, development of several tissues such as brain, eyes and wings	pattern both vertical and horizontal. More expressed in scutellum, 3 empty spots in vertical middle, problem stitching?
esg (Escargot)	TF, maintenance of cell number	scutellum except 2 spots and along horizontal axis (physio?)
eyg (eyegone)	TF (transcriptional repressor), notum development	not in scutelum, pattern not uniform in scutum (vertical pattern),true or artefact?
hth (homothorax)	TF, phenotypes manifest in adult abdominal segment, regulation of cell fate commitment; macromolecule localization, formation of anatomical boundary	signal seems noisy,scutellum, 2 patch in lateral scutum
ptc (patched)	hedgehog receptor activity, cell morphogenesis involved in differentiation; columnar/cuboidal epithelial cell differentiation	expressed on the two posterior/anterior lines, (neck
sd (scalloped)	TF,tissue morphogenesis; regulation cell communication, regul multicellular organismal development, stem cell proliferation	essential posterior line
Sr (stripe)	nucleic acid binding, epithelial cell migration, ectoderm development, determination of muscle attachment site	parallel horizontal strip in scutum
Ush (u-shaped)	TF binding,leading edge cell differentiation, pattern, epithelial cell fate commitment specification process	everywhere except lateral border and 2 spots in scutellum
Delamination		scutellum and horizontal midline
Proliferation		scutellum, horizontal midline + parallel stripe in scutum
Deformation		
Stress		

TF = transcription factor
Comments:

- ask Boris to describe Deformation and stress

- No phenotypic region correspond to the empty lateral spots and 2 spot in scutellum.

2.3.2 Clustering analysis

Using raw data or the mask? Le noir et 20 en plus

Manual versus automatic