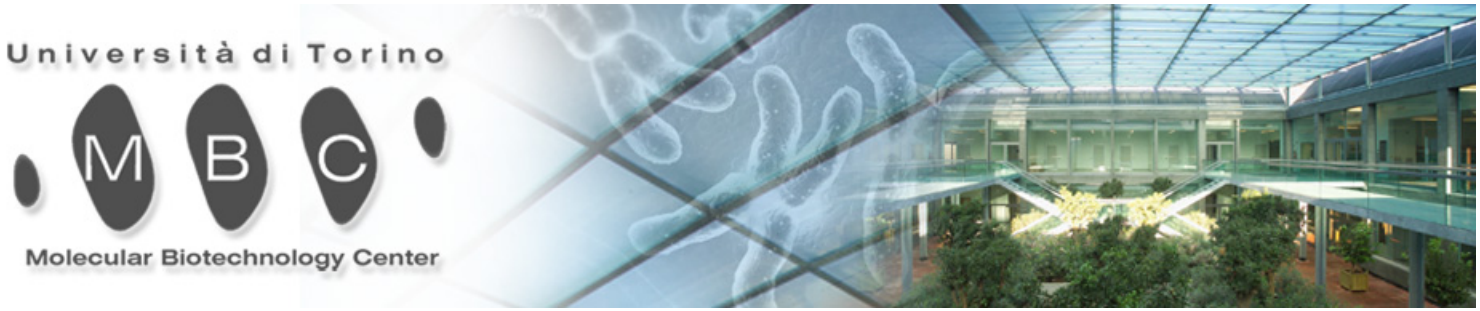


PAMR-SC

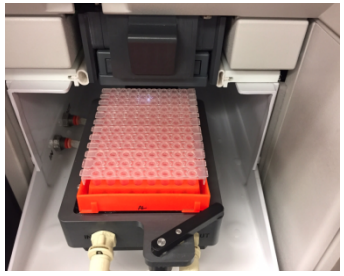
Calogero Raffaele & Alessandrì Luca
University of Torino (Italy)
raffaele.calogero@unito.it

Subpopulation Organization Of Innate T-cells

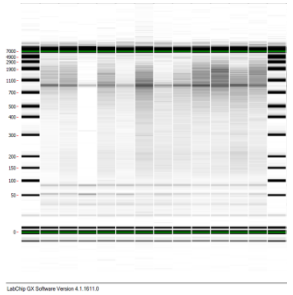
- We have a collaboration with Prof. De Libero, Basel University, for the analysis of single cell data to detect the presence of sub-populations within 4 cell populations, ~500 cells/group:
  - Activated T cells expressing the TCR γδ;
  - Resting TCR γδ cells;
  - Activated MAIT cells,
  - Resting MAIT cells.
- Aim:
  - Identify a set of gene products, that can be use in multi-color FACS analysis to study the detected sub-populations.

Smart-seq2* & Single Cell FACS Sorting, 96 cells/plate

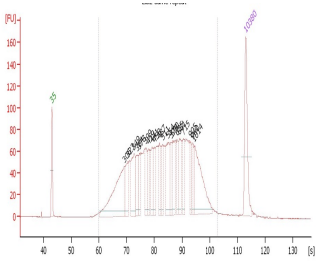
Prepare cells → Sort cells → Lysis, RT, Amplification → Library preparation → Sequence



BD FACSAria II™
Special Order System

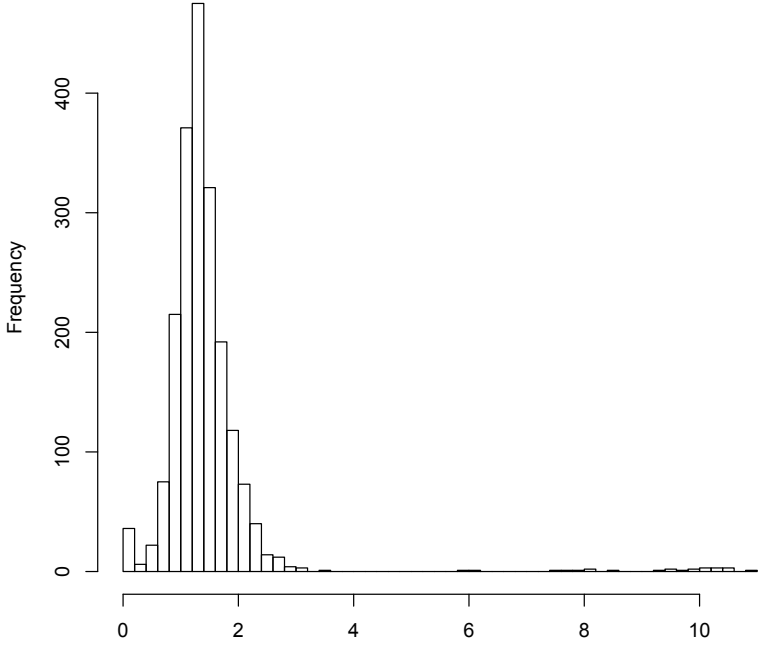


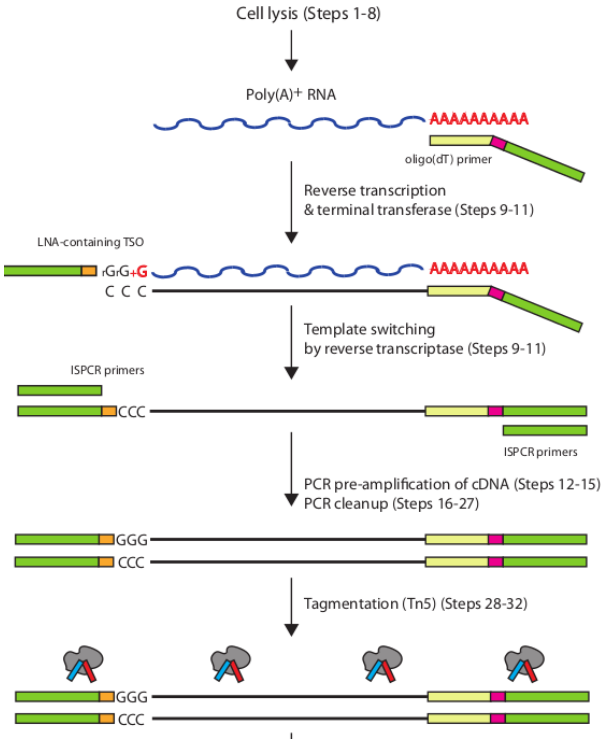
Single cell cDNA



Pooled single cell cDNA library

HiSeq 2000



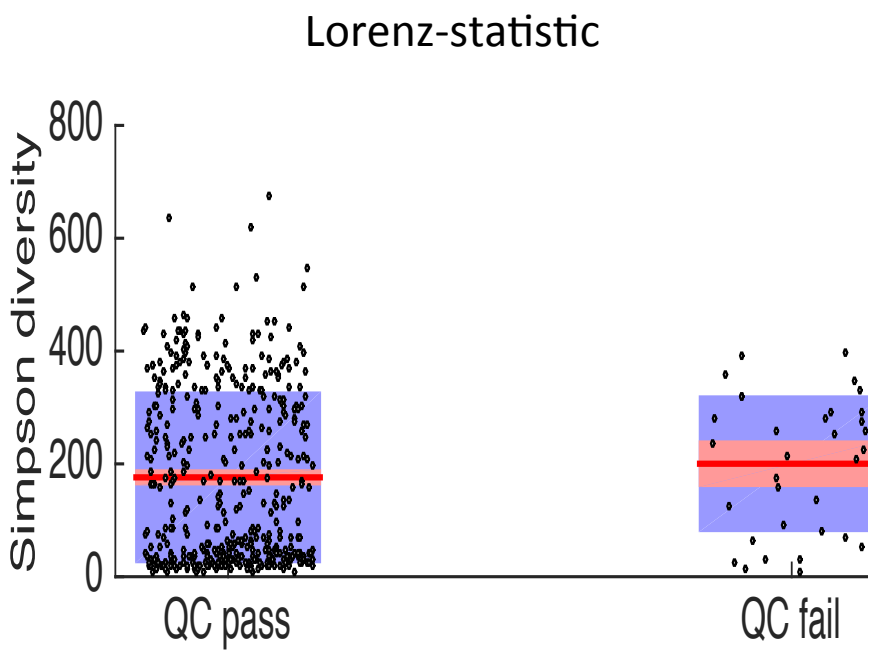


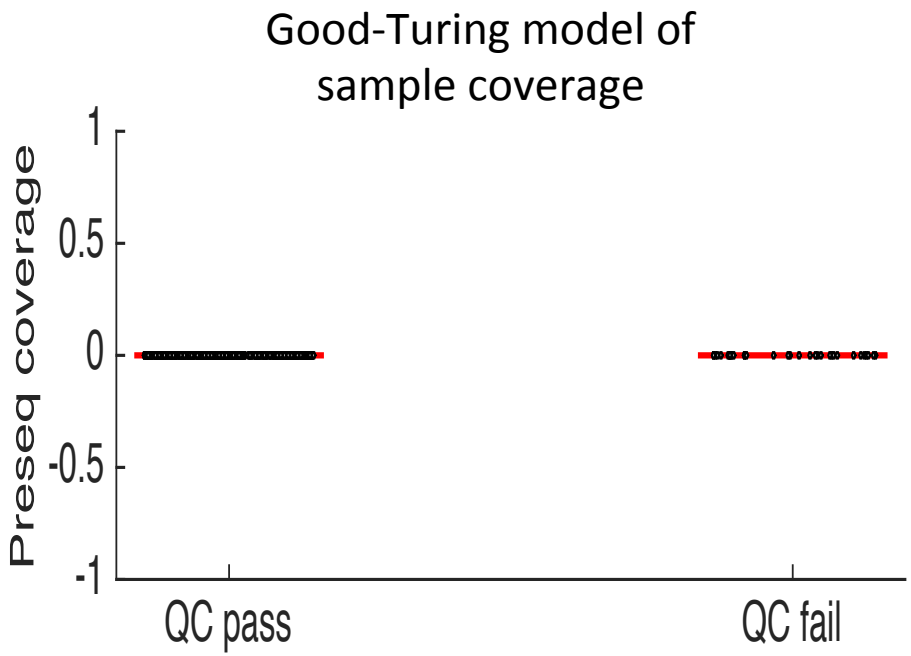
*Adapted from Picelli et al. Nature Methods 10, 1096–1098 (2013)

Workflow

- Data reduction.
- Cluster data to identify sub-populations
- Checking robustness of the sub-population clusters
- Selecting the main gene players in clusters formation.

Data reduction

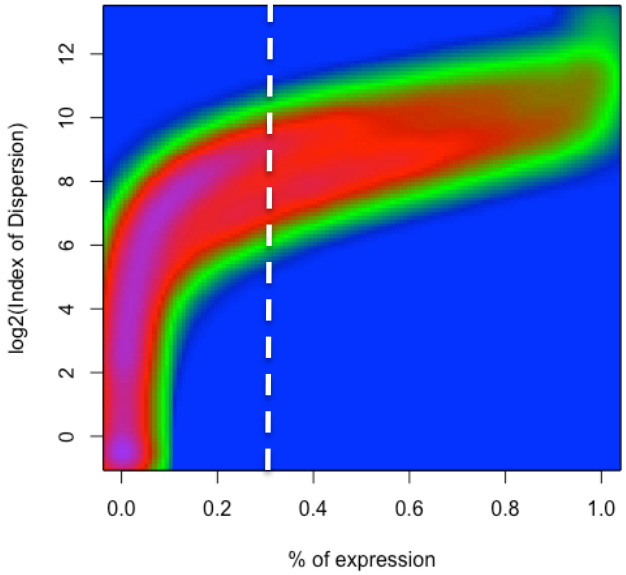


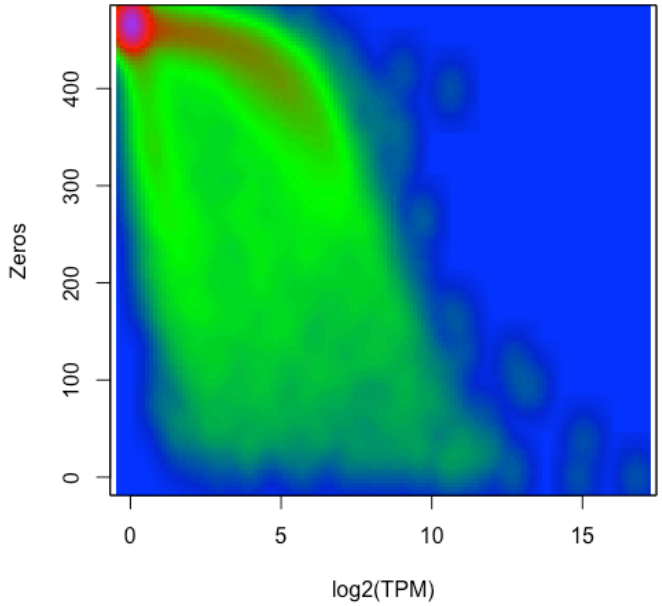


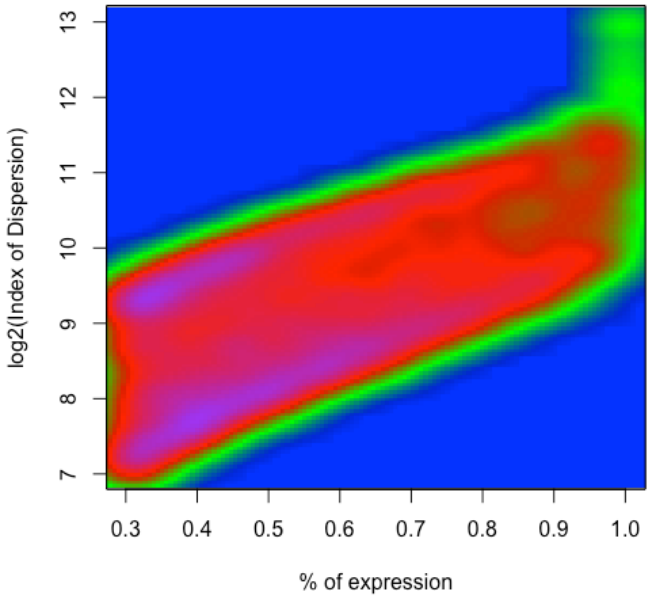
*Diaz et al. Bioinformatics (2016) 32 (14): 2219-2220

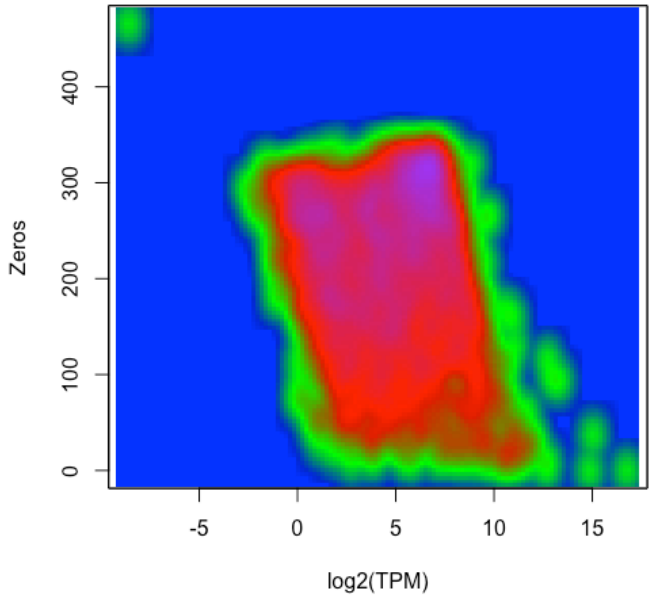
#Daley et al. Bioinformatics (2014) 30 (22): 3159-65

Data reduction



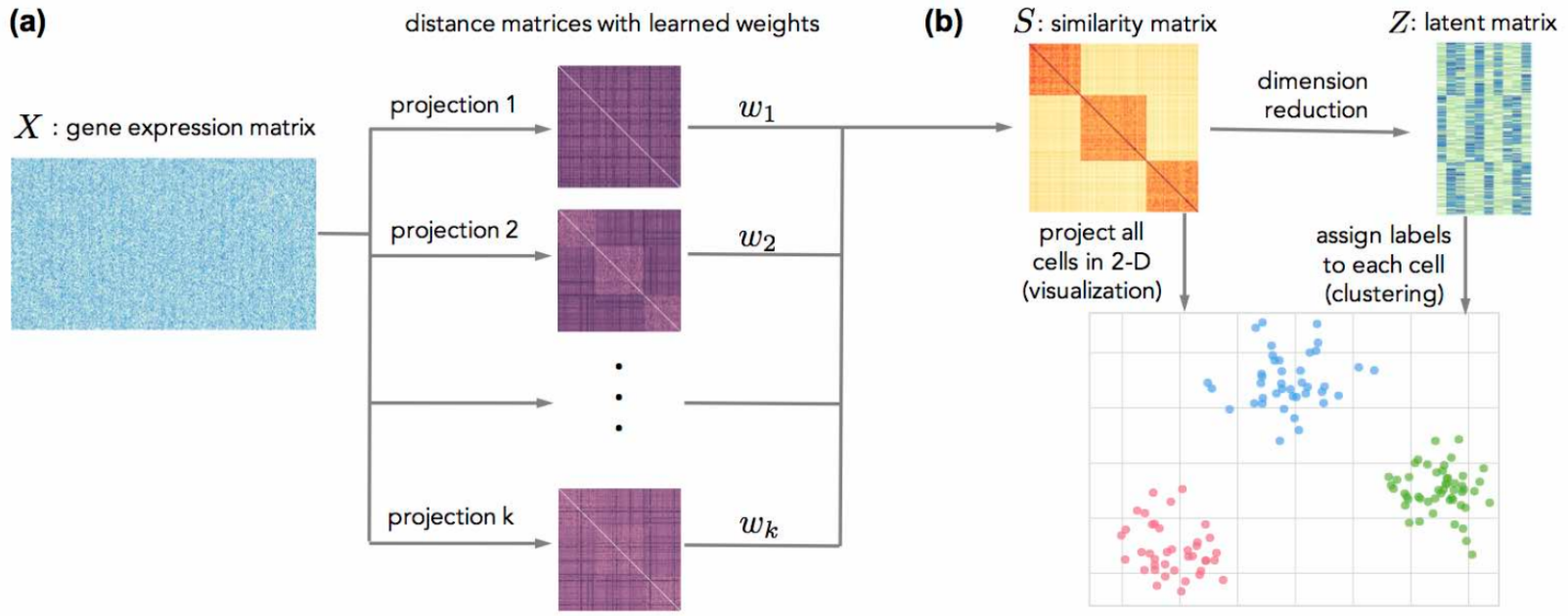






Detecting cell heterogeneity

SIMLR



http://bioconductor.org/packages/release/bioc/html/SIMLR.html

Bo Wanget al Biorxiv.org doi: http://dx.doi.org/10.1101/052225

Detecting cell heterogeneity

Table 2. NMI values for the four single-cell data sets. Higher values indicate better performance.

Data set	PCA	Laplacian	MDS	t-SNE	Sammon	PPCA	FA	ZIFA	SIMLR
Buettner	0.56	0.27	0.56	0.32	0.05	0.59	0.67	0.65	0.89
Kolodziejczyk	0.77	0.62	0.77	0.77	0.40	0.76	0.72	0.72	0.95
Pollen	0.83	0.80	0.83	0.93	0.75	0.83	0.82	0.79	0.94
Usoskin	0.39	0.48	0.39	0.69	0.41	0.40	0.36	0.41	0.69

Table 3. NNE values for the four single-cell data sets. Lower values indicate better performance.

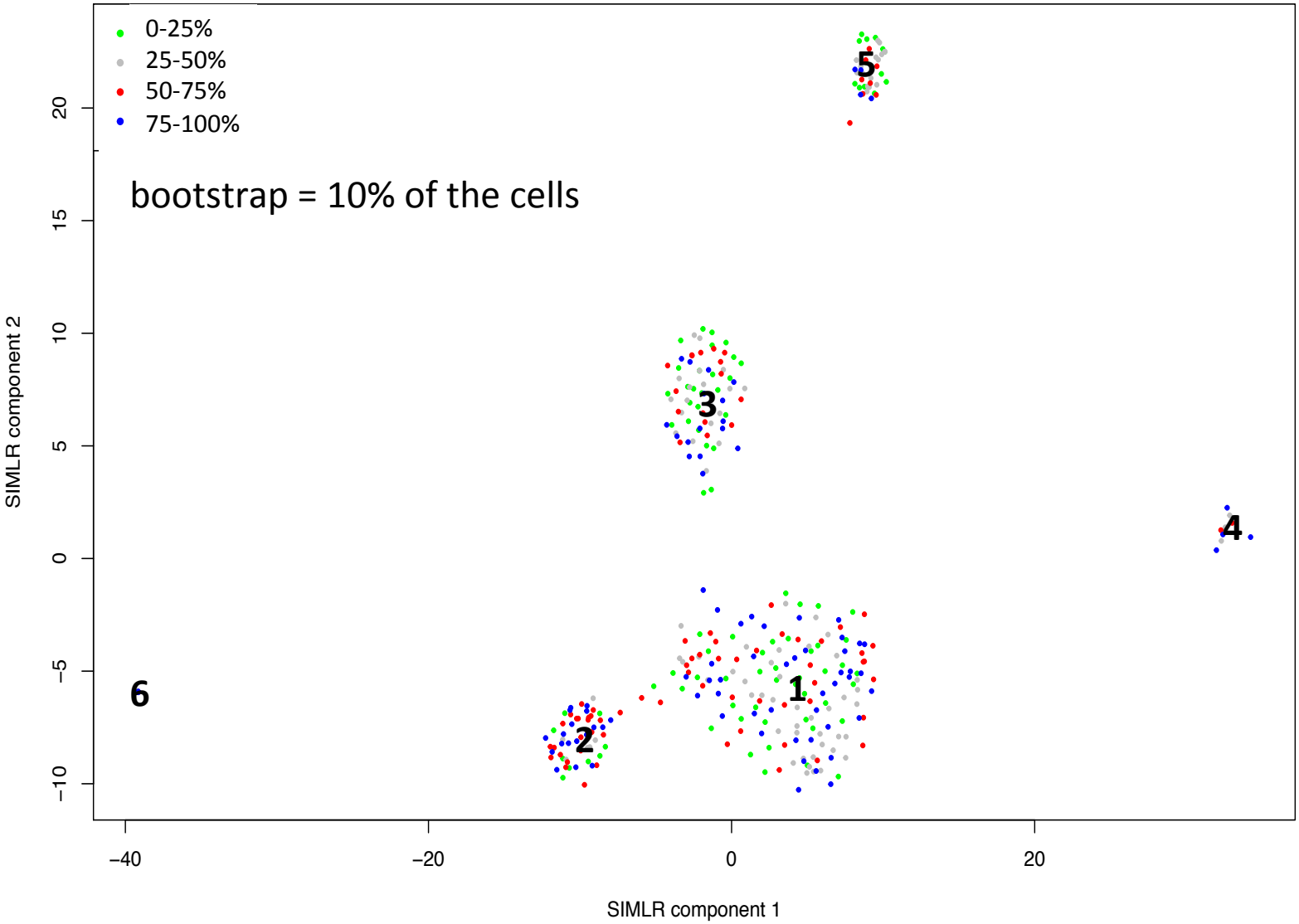
Data set	PCA	Laplacian	MDS	t-SNE	Sammon	PPCA	FA	ZIFA	SIMLR
Buettner	0.21	0.38	0.22	0.18	0.21	0.20	0.11	0.16	0.05
Kolodziejczyk	0.0016	0.0278	0.0016	0.0014	0.018	0.0016	0.009	0.01	0.0018
Pollen	0.052	0.156	0.055	0.023	0.11	0.056	0.075	0.075	0.020
Usoskin	0.30	0.11	0.29	0.072	0.20	0.29	0.31	0.28	0.063

http://bioconductor.org/packages/release/bioc/html/SIMLR.html
Bo Wanget al Biorxiv.org doi: http://dx.doi.org/10.1101/052225

Critical points in data analysis

- Detecting cell heterogeneity:
  - Defining the optimal number of clusters.
  - Stability of the clusters.
- Detecting the main players, i.e. genes, in cluster organization:
  - Stability of the clusters.

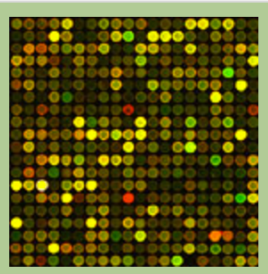
Clusters stability: cells effect



Cells bootstrap

- Allows the evaluation of cluster stability and selection of the optimal number of clusters.
- Scoring cluster quality:
  - normalized mutual information (NMI)
- Bootstrap procedure was implemented using BiocParallel.

Detecting the main gene players



PAM: Prediction Analysis for Microarrays

Class Prediction and Survival Analysis for Genomic Expression Data Mining

Features:

- Performs sample classifica
  via "nearest shrunken cent
  "Diagnosis of multiple canc
  PNAS 2002 99:6567-6572

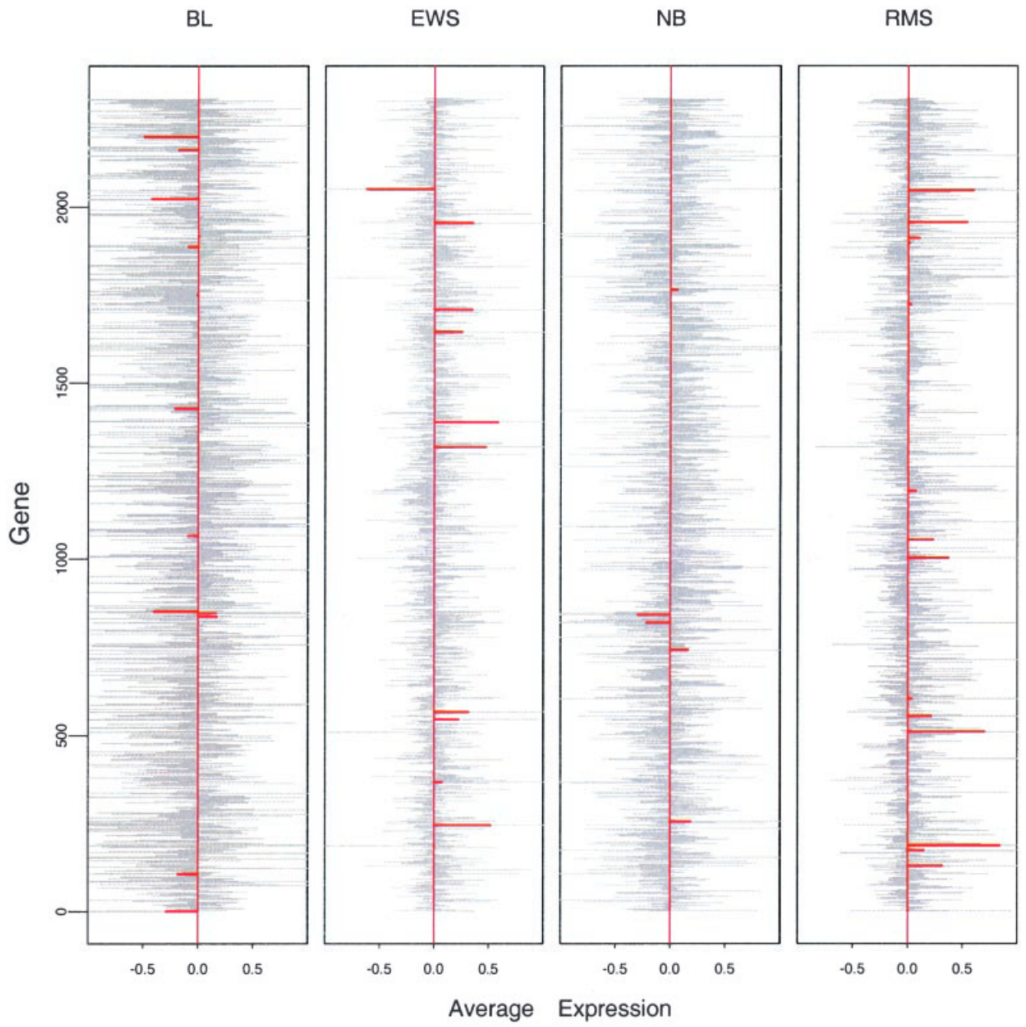


Fig. 1. Centroids (grey) and shrunken centroids (red) for the SRBCT dataset. The overall centroid has been subtracted from the centroid from each class. The horizontal units are log ratios of expression. From left to right, the numbers of training samples for each class are 8, 23, 12, and 20. The order of the genes is arbitrary.

Feature selection

We use the Index of dispersion D for each gene i in cluster j (1) and in all data set (2) :

$$D_{ij} = log \frac{\sigma_{ij}^2}{\mu_{ij}} \quad (1)$$

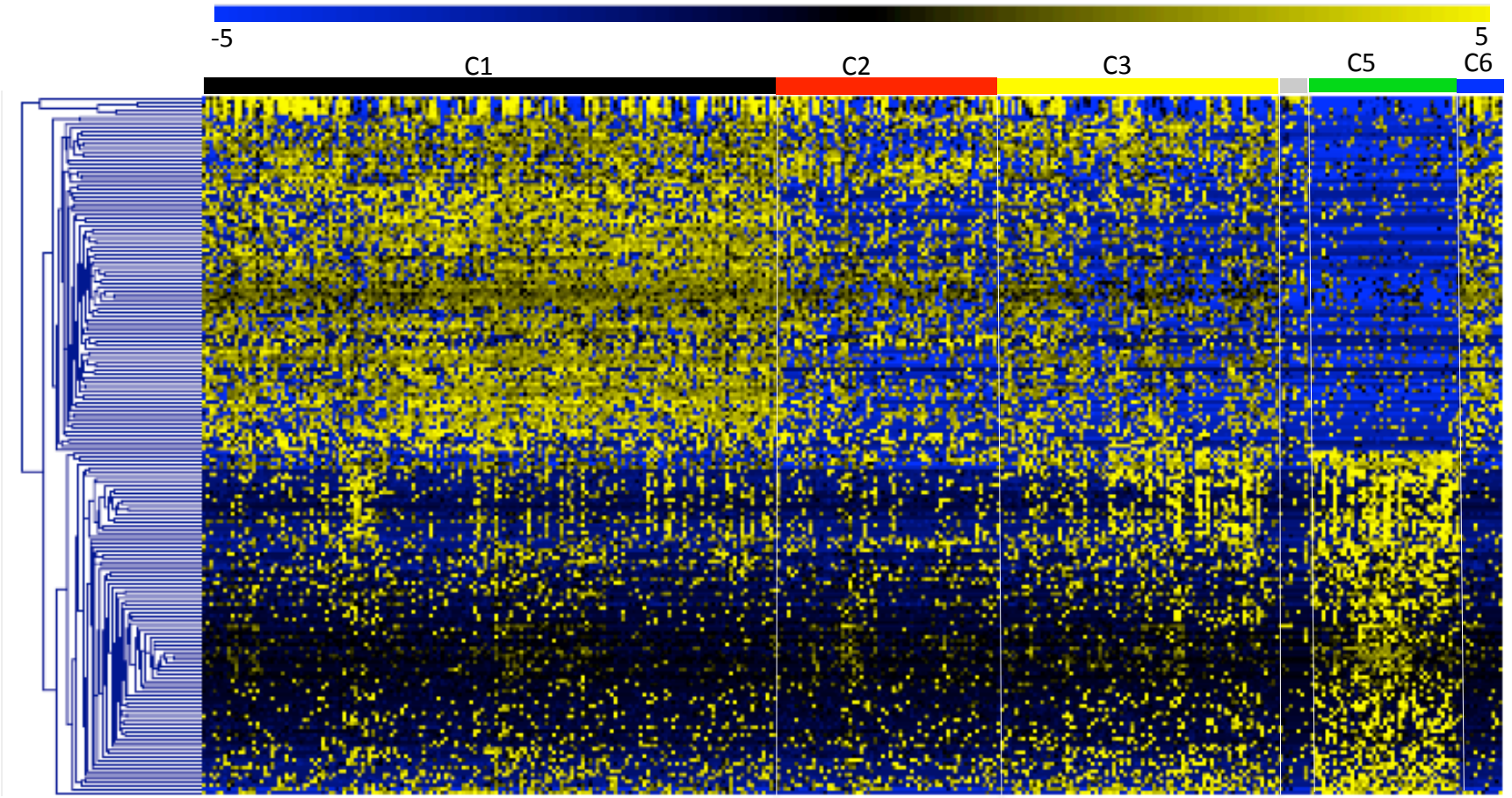
$$D_i = log \frac{\sigma_i^2}{\mu_i} \quad (2)$$

We increase of an amount S_i , % of D_{ij}, the D_{ij} and if $\Delta D_{ij} \leq 0$ the gene is purged:

$$\Delta D_{ij} = D_i - (D_{ij} + S_i) \qquad (3)$$

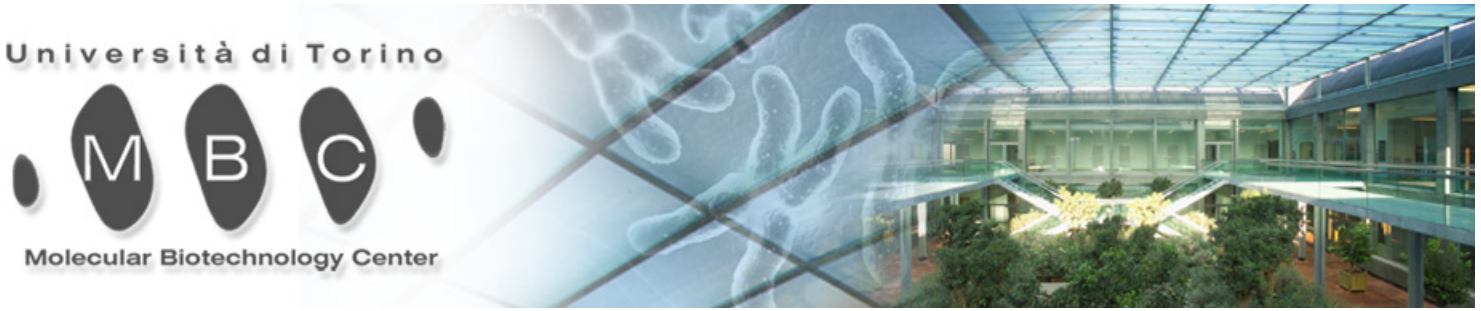
We run SIMLR on the purged genes set and we estimate the cluster stability with respect to full dataset.

HCL of 194 genes Mait activated



On-going

- Finalizing package …..



SIgN

Francesca Zolezzi (Galderma)
Josephine Lum
Bhairav Paleja

Universität Basel

Gennaro De Libero
Lucia Mori

B&Gu
Turin

Raffaele Calogero
Luca Alessandrì