Unsupervised clustering

Selecting a distance

- **Unsupervised method need a criteria to assess profile similarity (and dissimilarity) : a distance**

Table 1 Gene expression similarity measures

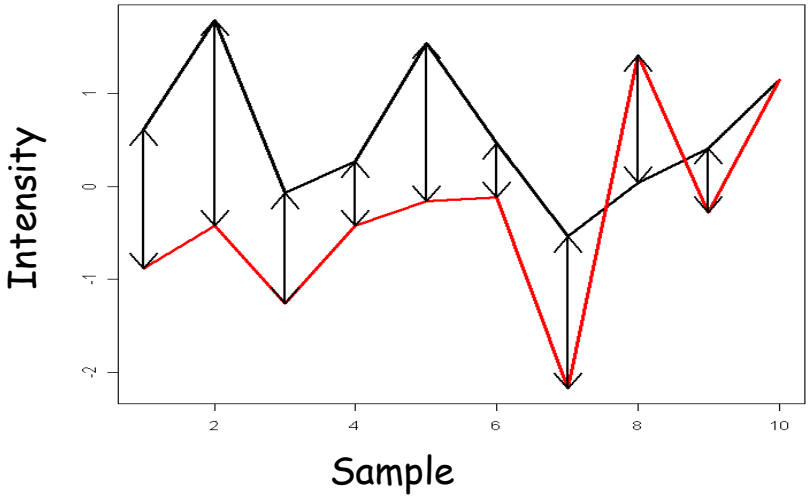
Manhattan distance (city-block distance, L1 norm)	$d_{fg} = \sum_c \lvert e_{fc} - e_{gc} \rvert$
Euclidean distance (L2 norm)	$d_{fg} = \sqrt{\sum_c (e_{fc} - e_{gc})^2}$
Mahalanobis distance	$d_{fg} = (\mathbf{e}_f - \mathbf{e}_g)' \Sigma^{-1} (\mathbf{e}_f - \mathbf{e}_g)$, where Σ is the (full or within-cluster) covariance matrix of the data
Pearson correlation (centered correlation)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c (e_{fc} - \bar{e}_f)(e_{gc} - \bar{e}_g)}{\sqrt{\sum_c (e_{fc} - \bar{e}_f)^2 \sum_c (e_{gc} - \bar{e}_g)^2}}$
Uncentered correlation (angular separation, cosine angle)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c e_{fc} e_{gc}}{\sqrt{\sum_c e_{fc}^2 \sum_c e_{gc}^2}}$
Spellman rank correlation	As Pearson correlation, but replace e_{gc} with the rank of e_{gc} within the expression values of gene g across all conditions $c = 1 \ldots C$
Absolute or squared correlation	$d_{fg} = 1 - \lvert r_{fg} \rvert$ or $d_{fg} = 1 - r_{fg}^2$

d_{fg}, distance between expression patterns for genes f and g. e_{gc}, expression level of gene g under condition c.

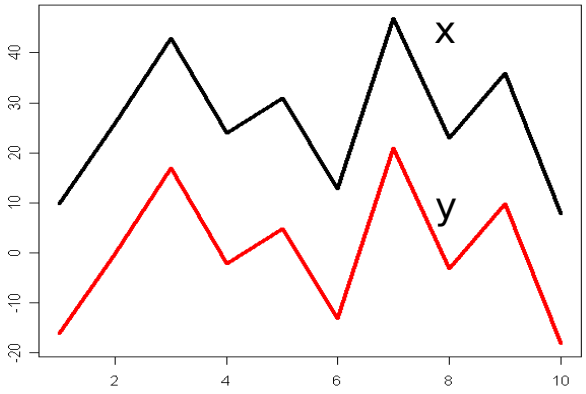
How does gene expression clustering work ? P. D'Haeseleer. Nat Biotech. 23(12). 2005.

Euclidean distance

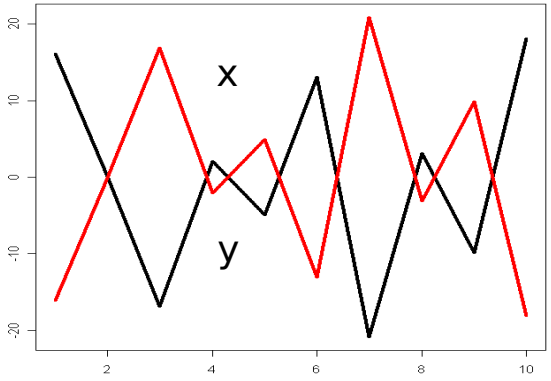
$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



d(x,y)= 82.2



d(x,y)= 80.4



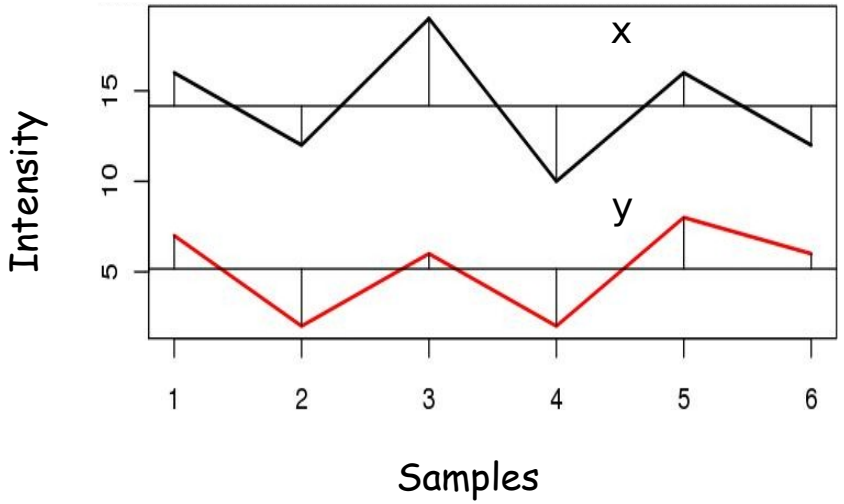
- Limitations
  - Highly sensitive to global expression level
  - Does not measure anti-correlation
  - No upper bound

Pearson correlation coefficient (*r*)

- Obtained by dividing the covariance of the genes by the product of their standard deviations
  - Covariance is a measure of how much two variables change together

$$cov(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$r = \frac{cov(x, y)}{\sqrt{var(x)\, var(y)}}$$



- Limits : -1 and 1

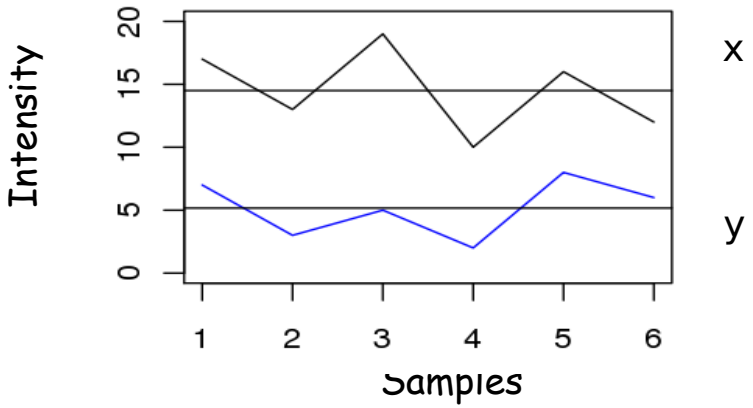
- Distance is obtained using:

$$d = (1 - r) / 2$$

Spearman correlation coefficient (*rho*)

- Used to measure the degree of correspondence between two rankings

$$rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$



Intensities

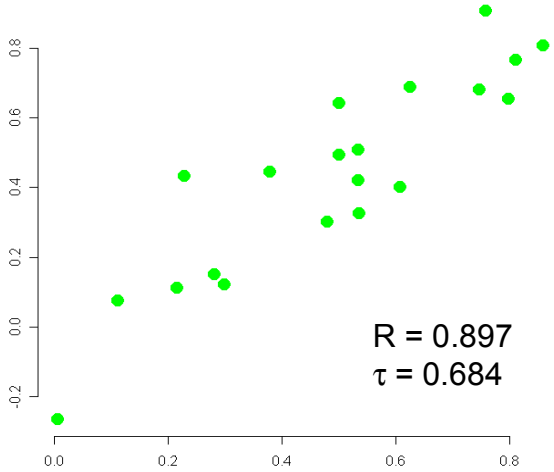
	1	2	3	4	5	6
x	17	13	19	10	16	12
y	7	3	5	2	8	6

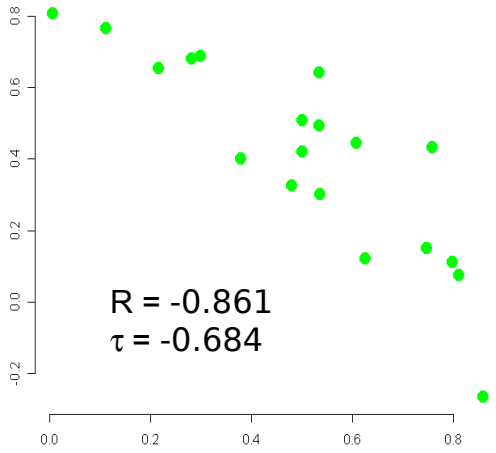
Ranks

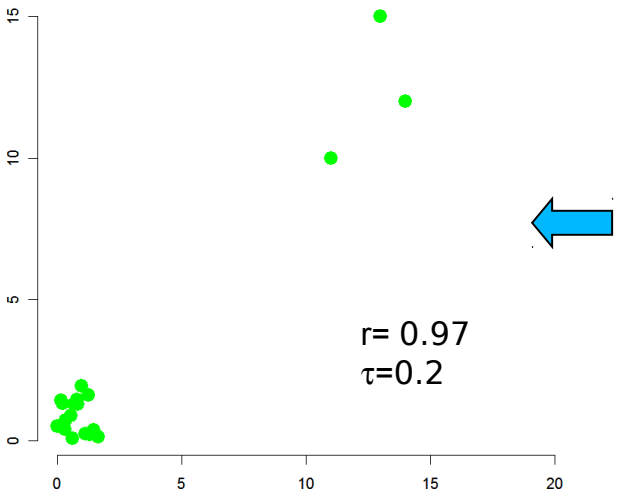
	1	2	3	4	5	6
x	5	3	6	1	4	2
y	5	2	3	1	6	4
d	0	1	3	0	-2	-2

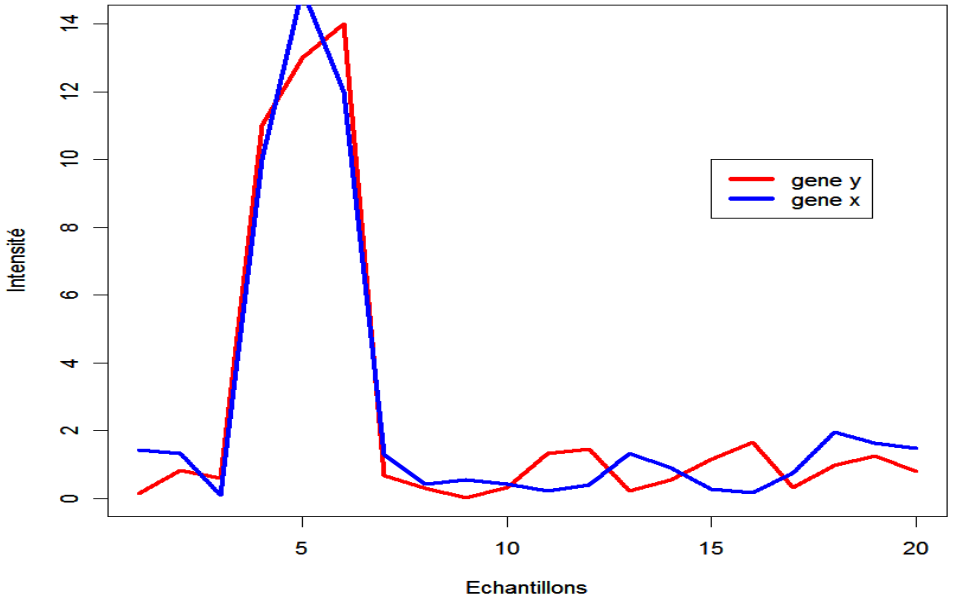
rho = 1 - 108/210 =0.485

Pearson and spearman









Hierarchical clustering

Distance matrix

Expression matrix

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
g1	-0.733	-1.098	-0.741	1.778	0.119	-0.510	0.446	0.225	0.244	0.006
g2	-0.660	-0.772	0.586	1.201	-1.785	-0.108	-0.191	0.160	1.585	1.840
g3	0.753	-1.088	-1.672	-0.659	-0.709	-0.496	0.158	-0.024	0.355	1.418
g4	0.498	1.006	0.331	-0.960	0.925	-1.174	-0.790	-0.334	0.338	-0.105
g5	-1.020	-0.450	1.332	0.153	-0.611	0.515	-0.323	-1.014	0.831	0.656
g6	0.616	-0.366	0.038	2.375	-0.526	1.089	-1.356	0.591	-2.298	0.204
g7	0.670	0.691	-0.534	1.611	1.376	-0.334	-0.420	1.892	0.234	0.788
g8	-0.978	-0.279	-0.654	-0.337	-0.722	2.244	0.376	-0.145	-1.190	0.634
g9	2.669	0.790	-0.445	0.712	-1.073	1.010	-1.598	-1.398	0.944	-0.483
g10	0.318	0.414	1.462	0.251	1.033	-1.032	0.841	0.662	-0.209	-0.485
g11	-0.737	0.139	0.439	1.524	-0.337	0.835	-0.006	-1.082	-0.796	0.943
g12	-0.771	-0.577	0.221	-1.217	0.302	0.804	-0.424	-0.465	-0.826	0.258
g13	1.355	2.217	0.566	-1.470	-0.051	1.060	1.871	0.552	-0.559	-0.999
g14	0.640	2.381	-0.928	0.577	2.518	-0.164	-1.731	0.302	1.725	-2.903
g15	0.896	0.132	-0.368	1.224	-0.306	-1.046	-0.869	-0.969	1.453	-1.264
g16	0.093	-0.749	0.923	-0.157	1.180	-0.520	-0.098	-0.429	-0.011	-0.450
g17	2.415	-2.154	-0.900	-0.969	-0.970	-2.180	0.619	0.057	0.170	-0.979
g18	-0.580	-2.411	0.959	1.523	0.047	-1.019	-0.727	0.741	-0.294	1.205
g19	-0.085	1.080	-0.350	-0.332	-0.493	-0.547	-0.478	-0.304	0.766	-0.271
g20	-1.002	-0.954	-0.674	-0.737	-0.427	0.919	1.140	-0.468	0.041	-2.620

Distance matrix (samples)

	E1	E2	E3	E4	E5	E6	E7	E8	E9
E2	6.370								
E3	6.867	6.548							
E4	7.277	8.038	6.120						
E5	6.531	5.244	5.461	6.797					
E6	7.635	6.183	5.893	6.964	7.054				
E7	6.869	7.203	5.177	7.990	6.463	5.958			
E8	5.839	6.323	5.027	5.631	4.373	6.270	4.755		
E9	5.804	6.248	6.080	6.676	5.942	7.574	6.659	6.038	
E10	8.250	9.004	5.719	6.475	8.211	6.866	6.784	5.935	7.655

Distance matrix (genes)

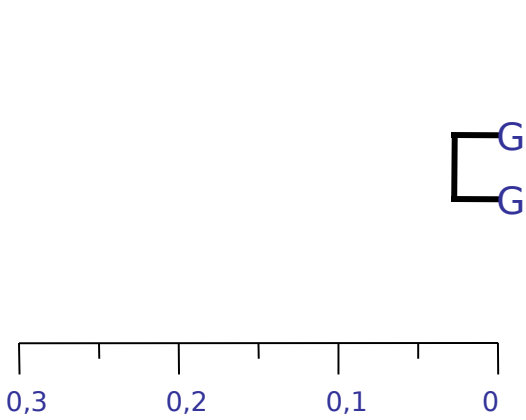
	g1	g2	g3	g4	g5	g6	g7	g8	g9	g10	g11	g12	g13	g14	g15	g16	g17	g18	g19
g2	3.401																		
g3	3.443	3.716																	
g4	4.188	4.864	3.873																
g5	3.452	2.650	4.035	3.658															
g6	4.022	5.078	5.138	5.343	4.824														
g7	3.315	4.622	4.291	3.781	4.823	4.122													
g8	4.012	4.525	3.919	4.885	3.542	4.088	5.035												
g9	5.257	5.176	4.681	4.383	4.707	4.884	5.192	5.339											
g10	3.532	4.772	4.615	2.694	3.769	4.882	3.400	4.886	5.445										
g11	2.979	3.464	4.190	4.201	2.473	3.240	4.200	2.881	4.662	3.919									
g12	3.791	4.470	3.530	3.183	2.706	4.452	4.654	2.410	5.073	3.684	3.084								
g13	5.822	6.447	5.454	4.234	5.341	6.115	5.324	4.699	5.381	3.886	5.275	4.544							
g14	6.061	7.599	6.991	4.523	6.595	6.972	5.051	7.249	5.527	5.393	6.699	6.318	6.078						
g15	3.363	4.313	4.191	3.272	3.867	5.032	4.238	5.478	3.201	3.931	4.199	4.620	5.579	4.442					
g16	3.072	4.431	3.868	2.337	2.863	4.625	3.849	4.244	4.811	2.063	3.416	2.488	4.658	5.311	3.342				
g17	4.858	5.819	3.673	4.765	5.786	6.407	5.866	6.299	5.462	4.919	6.246	5.325	6.026	7.351	4.398	4.449			
g18	2.887	3.418	4.230	4.844	3.695	4.171	4.187	5.013	6.206	4.118	3.814	4.191	7.094	7.609	4.766	3.486	5.324		
g19	3.410	3.778	3.288	1.979	2.969	4.891	3.781	3.997	3.749	3.235	3.472	3.103	4.101	4.730	2.573	2.937	4.744	4.749	
g20	4.082	5.730	4.873	4.849	4.366	5.830	5.959	3.905	5.731	4.591	4.766	3.640	4.734	6.100	4.552	3.876	5.136	5.809	4.011

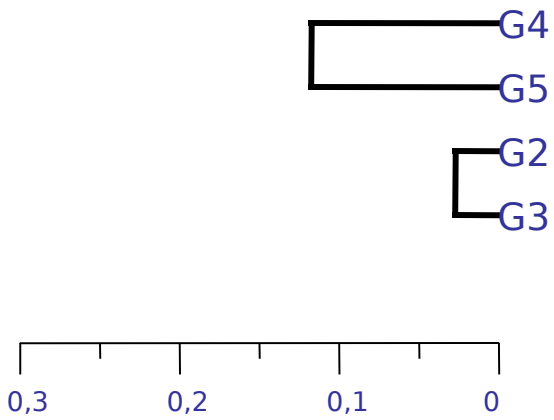
Hierarchical clustering

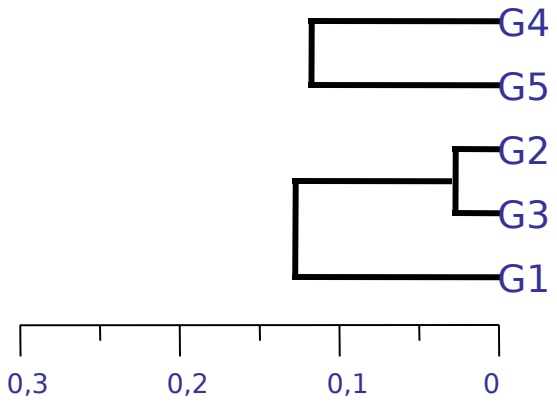
	G1	G2	G3	G4	G5	G6
G1	0					
G2	0,1	0				
G3	0,16	0,03	0			
G4	0,9	0,8	0,5	0		
G5	0,25	0,4	0,77	0,12	0	
G6	1,3	1,5	0,95	1,0	1,1	0

	G1	G2-G3	G4	G5	G6
G1	0				
G2-G3	0,13	0			
G4	0,9	0,65	0		
G5	0,25	0,58	0,12	0	
G6	1,3	1,22	1,0	1,1	0

	G1	G2-G3	G4-G5	G6
G1	0			
G2-G3	0,13	0		
G4-G5	0,57	0,61	0	
G6	1,3	1,2	1,0	0

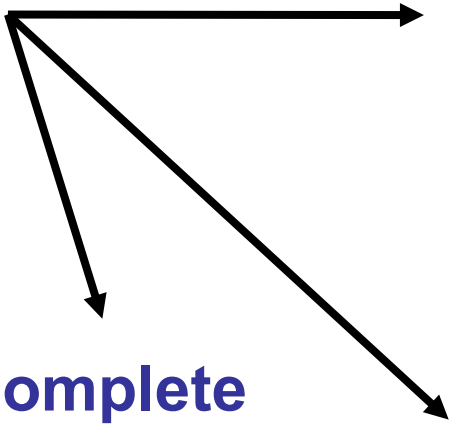






Agglomeration methods

	G1	G2	G3	G4	G5	G6
G1	0					
G2	0,1	0				
G3	0,16	0,03	0			
G4	0,9	0,8	0,5	0		
G5	0,25	0,4	0,77	0,12	0	
G6	1,3	1,5	0,95	1,0	1,1	0



Single

	G1	G2-G3	G4	G5	G6
G1	0				
G2-G3	0,1	0			
G4	0,9	0,5	0		
G5	0,25	0,4	0,12	0	
G6	1,3	0,95	1,0	1,1	0

Complete

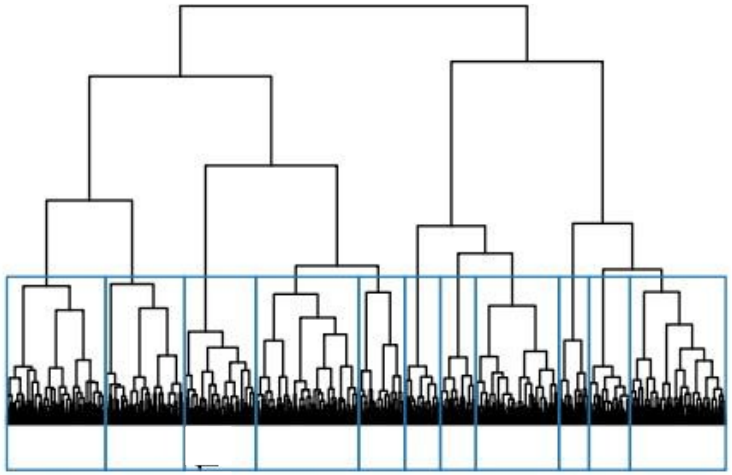
	G1	G2-G3	G4	G5	G6
G1	0				
G2-G3	0,16	0			
G4	0,9	0,8	0		
G5	0,25	0,77	0,12	0	
G6	1,3	1,5	1,0	1,1	0

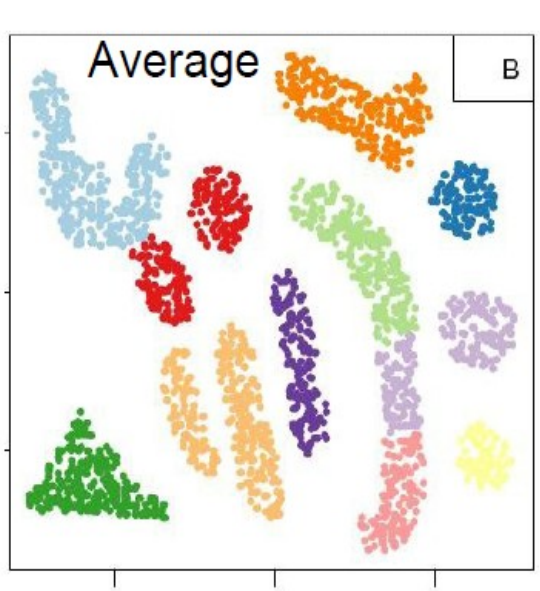
Average

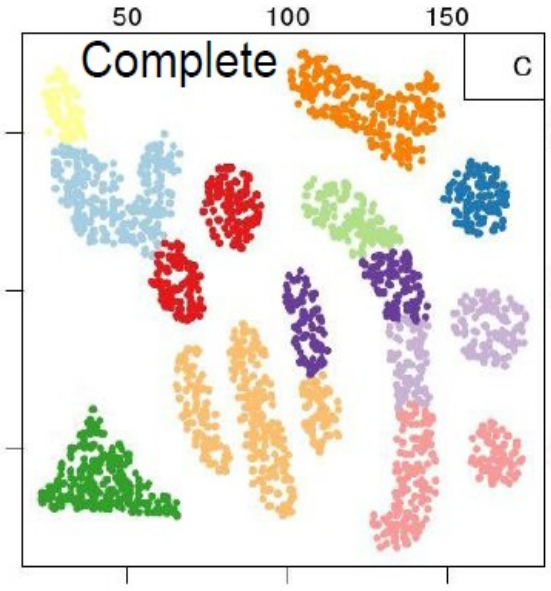
	G1	G2-G3	G4	G5	G6
G1	0				
G2-G3	0,13	0			
G4	0,9	0,65	0		
G5	0,25	0,58	0,12	0	
G6	1,3	1,22	1,0	1,1	0

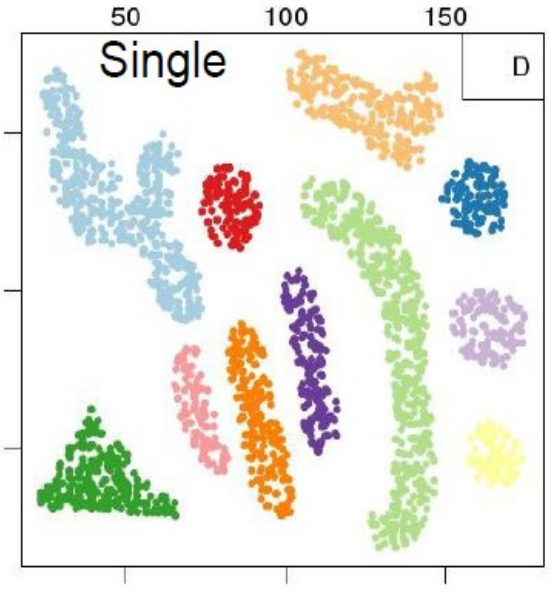
Which agglomeration method ?

- An artificial 2D dataset
  - Here single criteria performs best
- Microarrays
  - Most generally average

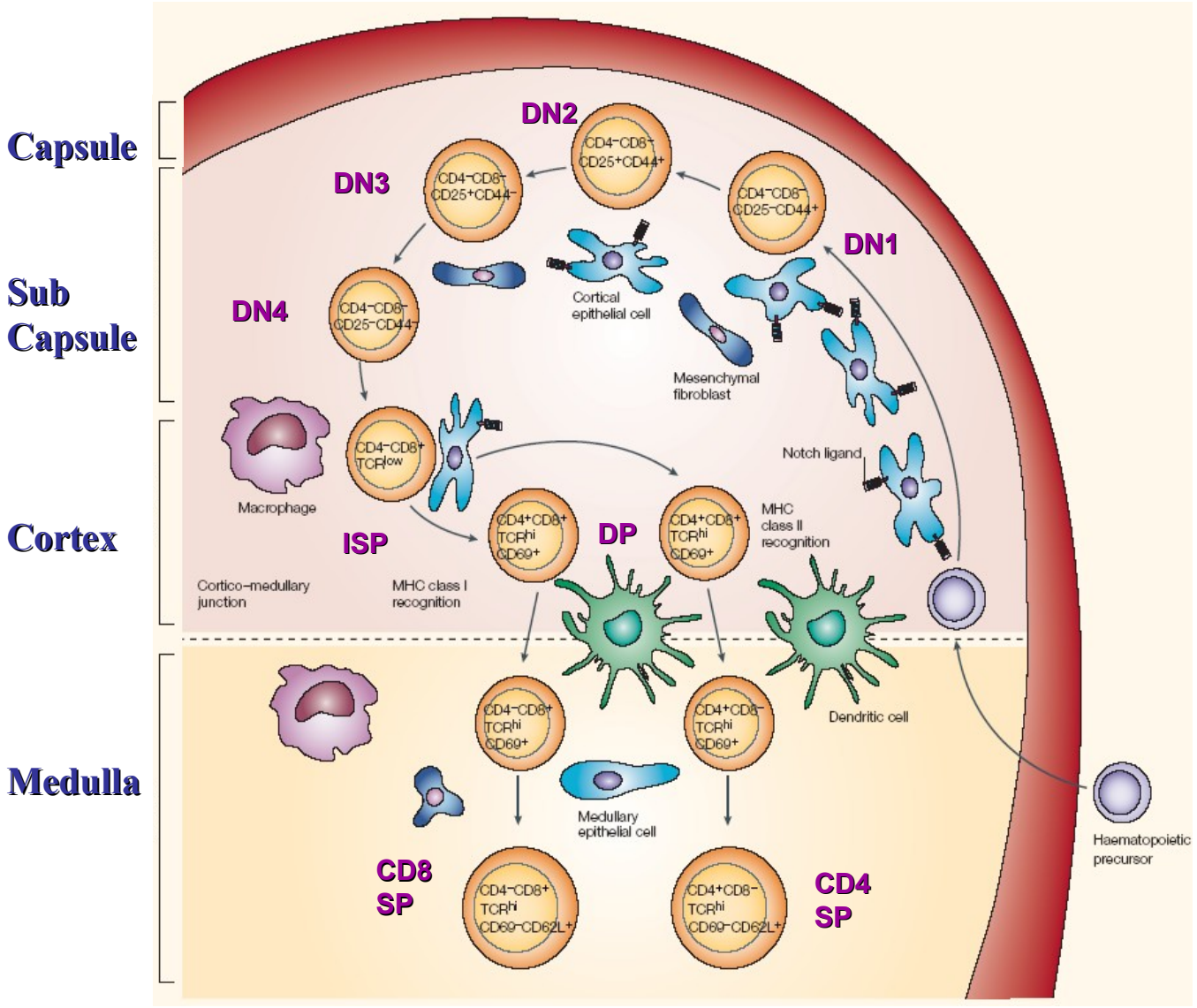








The Thymus project

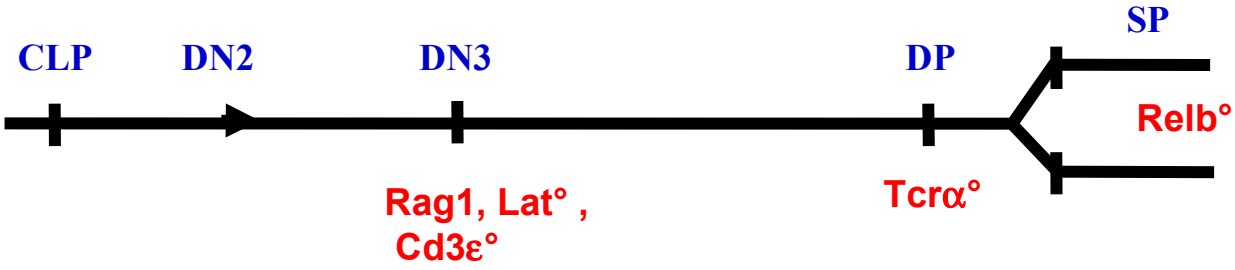


(From **Zúñiga-Pflücker JC**, *Nature Reviews Immunology* 4, 67-72; 2004)

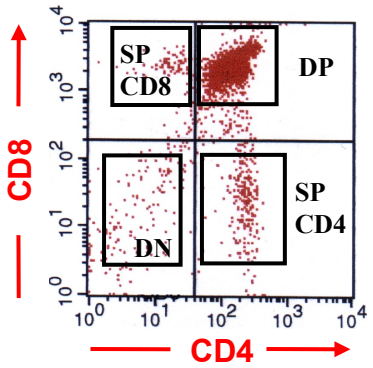
Genetically engineered mice

Thymi from WT and KO mice

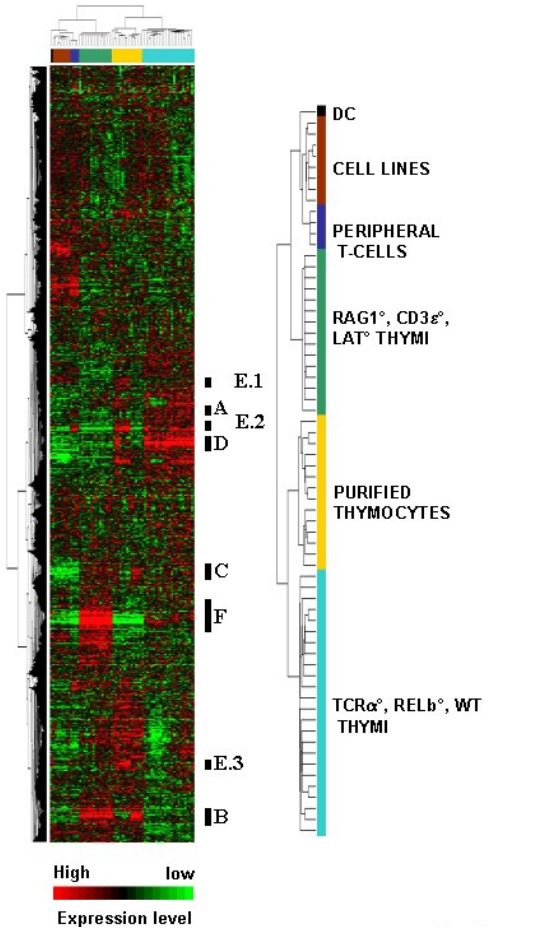
- ✓ Thymi from KO mice are enriched in particular cell types

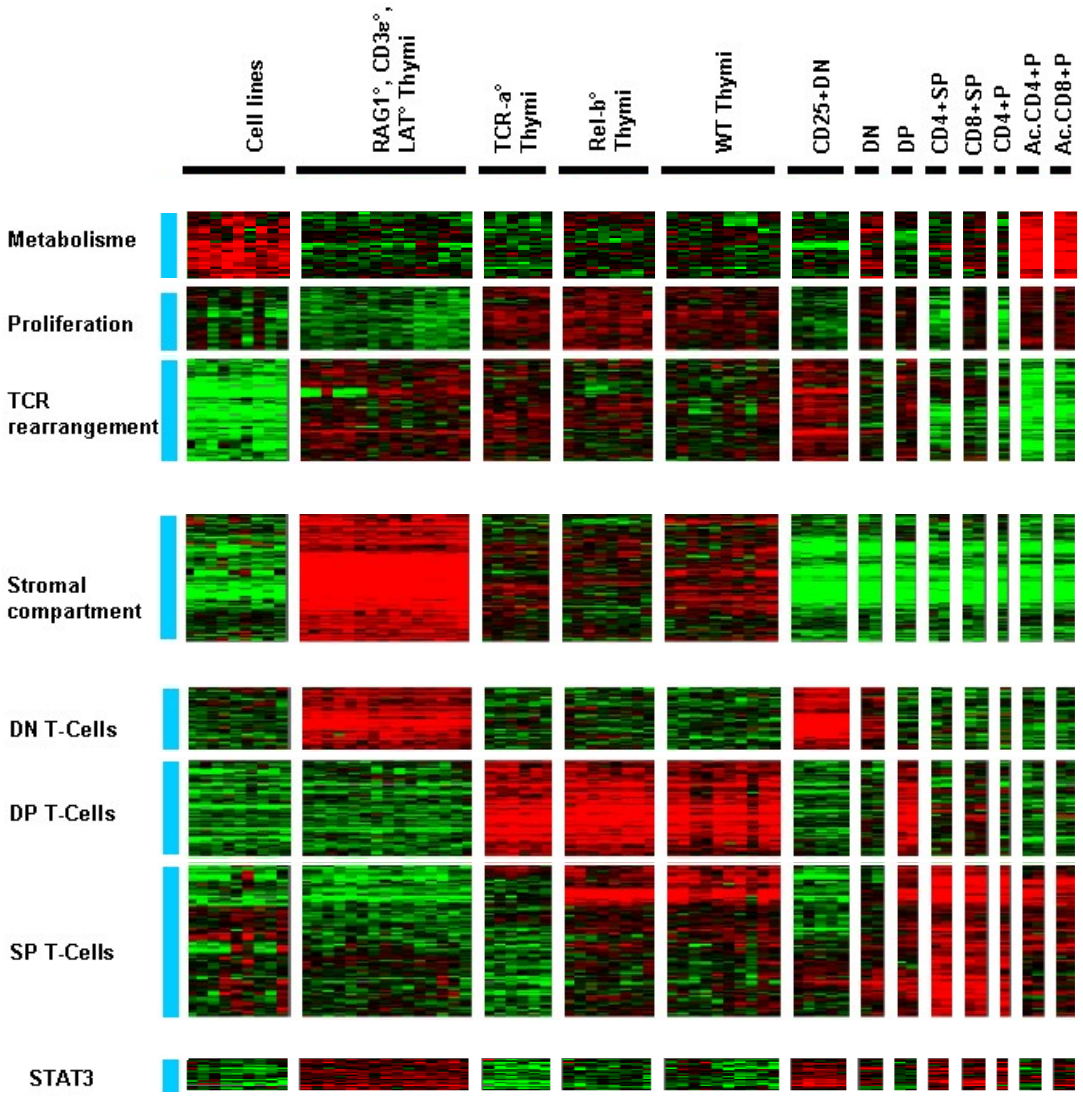


Purified thymocytes



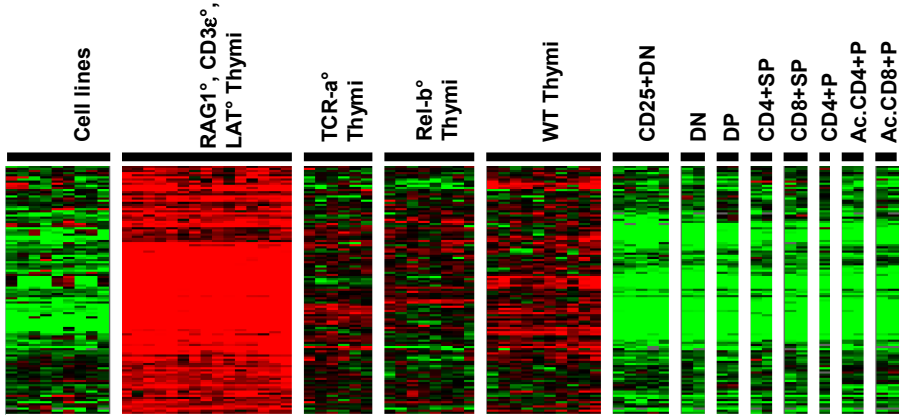
Virtual microdissection





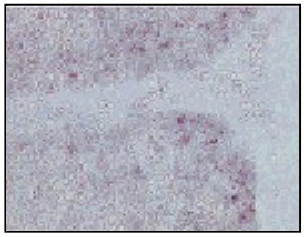
Puthier et al. A General Survey of Thymocyte Differentiation by Transcriptional Analysis of Knockout Mouse Models. **J. Immunol, 2004, 173(10):6109**

Virtual microdissection



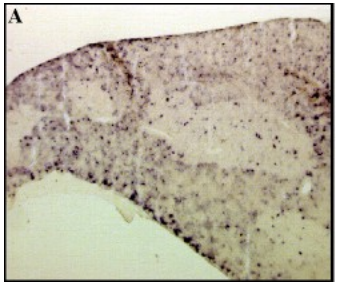
SDF1

(Conrad CB, Eur.J. Immunol, 2000, 30:3371)



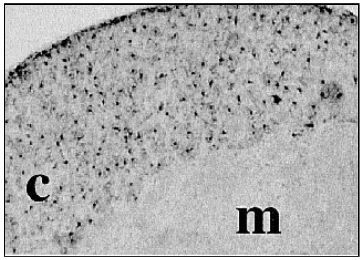
TECK

(Wurbel MA, Eur.J. Immunol, 2000, 30:262)



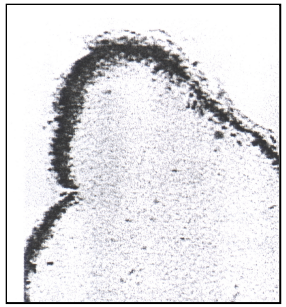
TSSP

(Carrier A et al, Immunogenetics, 1999, 50 :255-270)



SPATIAL

(Flommerfelt FA, Genes & Immunity, 2000, 1:391)



CD83

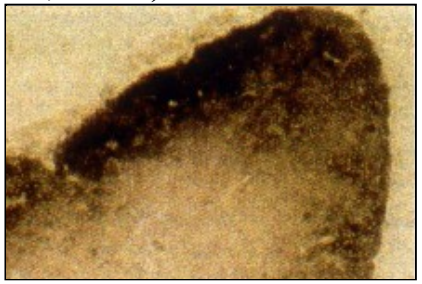
(Fujimoto Y, Cell, 2002, 108: 755

Ly75/CD205

Kraal G, J. Exp. Med, 1986, 163:981)

TSCOT

(Kim GM, J. Immunol, 2000, 164:3185)

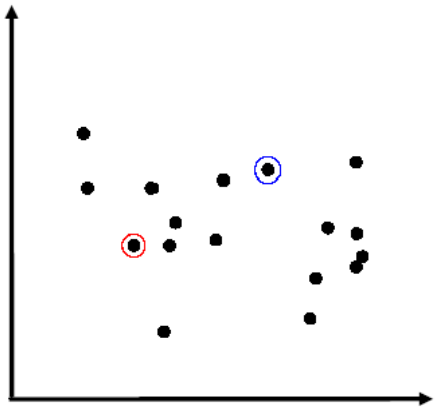


K- means algorithm

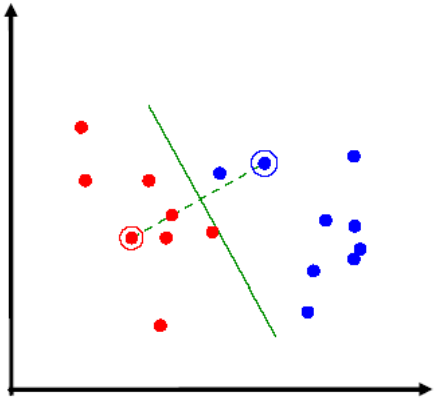
- K-means
  - Méthode de partitionnement
    - Définir k-classes de patients ou de gènes pour les analyser individuellement
  - Définir a priori le nombre de classes (k)
    - Cet argument peut être assez difficile à déterminer lorsqu'on s'intéresse aux gènes (voire aux échantillons)
  - L'initialisation se fait au hasard (plusieurs solutions possibles)

K-means algorithm

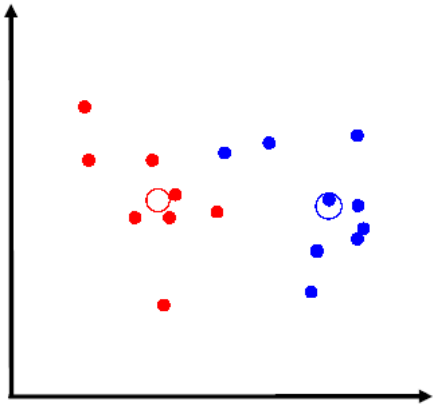
- The number of classes is given as input to the algorithm (here k=2)



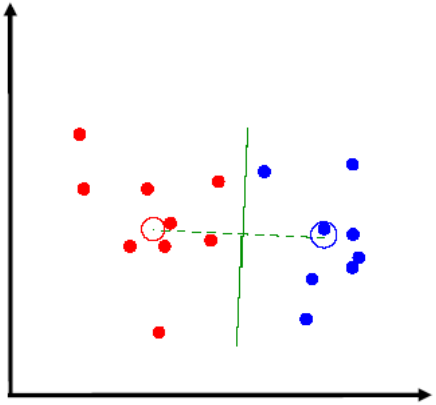
Arbitrary selection of k centers
k=2



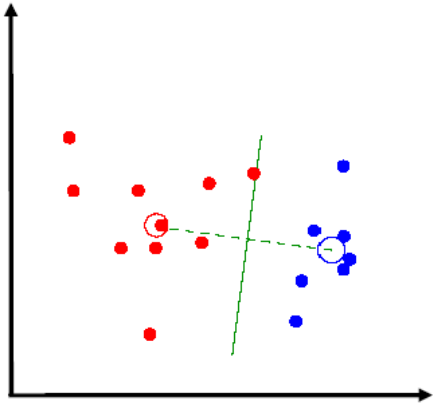
each point is allocated to the closest center



Centers are calculated



each point is allocated to the



Centers are re-calculated

• • •