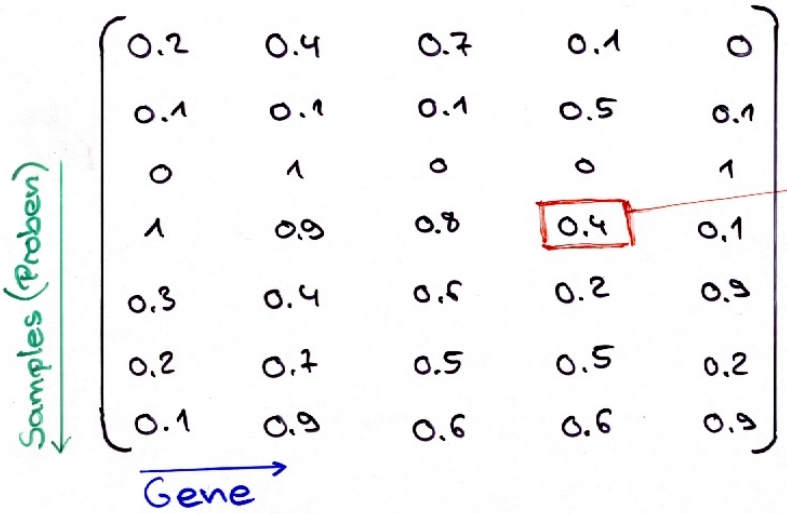


- Zeilen: die untersuchten Samples (Proben). Verschiedene Bedingungen → 5-50 Z., verschiedene Patienten → 50-250 Z.
- Spalten: die überwachten Gene, hängt vom Array ab → 1000-100000 Sp.

Distanzen zwischen zwei Vektoren

$u = (u_1, \ldots, u_n)$ und $v = (v_1, \ldots, v_n)$

- euklidische Distanz

$$d_E(u,v) = \sqrt{(u_1 - v_1)^2 + \ldots + (u_n - v_n)^2}$$

- Manhatten-Distanz

$$d_M(u,v) = |u_1 - v_1| + \ldots + |u_n - v_n|$$

Ähnlichkeiten zwischen Vektoren u, v

- Skalarprodukt

$$S(u,v) = \langle u,v \rangle = u_1 v_1 + \ldots + u_n \cdot v_n$$

- Korrelationskoeffizient (Pearson)

$$s_p(u,v) = \frac{(u_1 - \bar{u})(v_1 - \bar{v}) + \ldots + (u_n - \bar{u})(v_n - \bar{v})}{\sqrt{(u_1 - \bar{u})^2 + \ldots + (v_1 - \bar{v})^2} \cdot \sqrt{(v_1 - \bar{v})^2 + \ldots + (v_n - \bar{v})^2}}$$

$\bar{u}$ ist Mittelwert von $u_1, \ldots, u_n$

Agglomeratives Clustern

- single linkage

$$D(u,k) = \min\{D(i,k), D(j,k)\}$$

- complete linkage

$$D(u,k) = \max\{D(i,k), D(j,k)\}$$

- WPGMA : weighted pair group method using arithmetic averages

$$D(u,k) = \frac{D(i,k) + D(j,k)}{2}$$

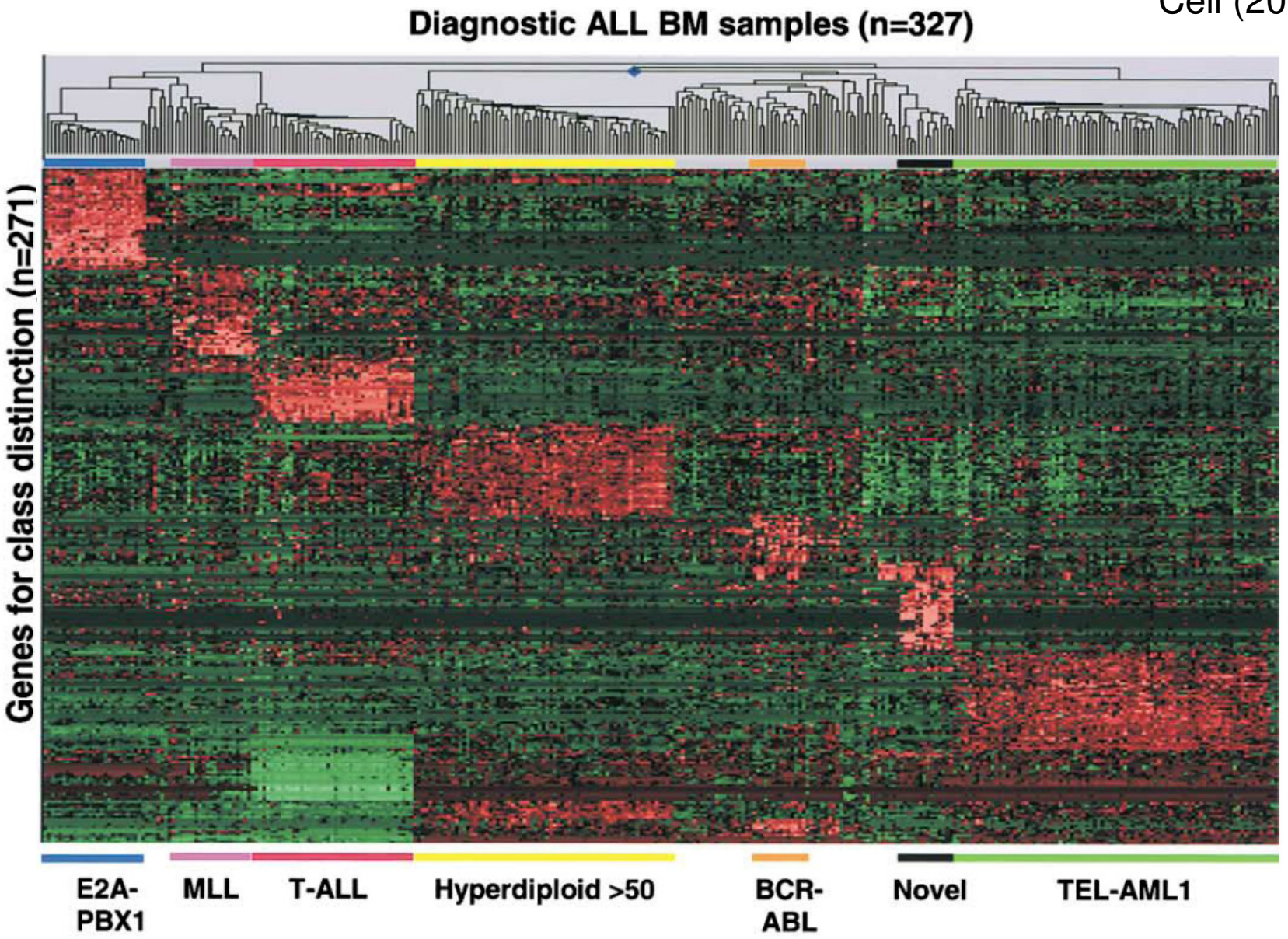
- UPGMA : unweighted pair group method using arithmetic averages

$$D(u,k) = \frac{n_i \cdot D(i,k) + n_j \cdot D(j,k)}{n_i + n_j}$$

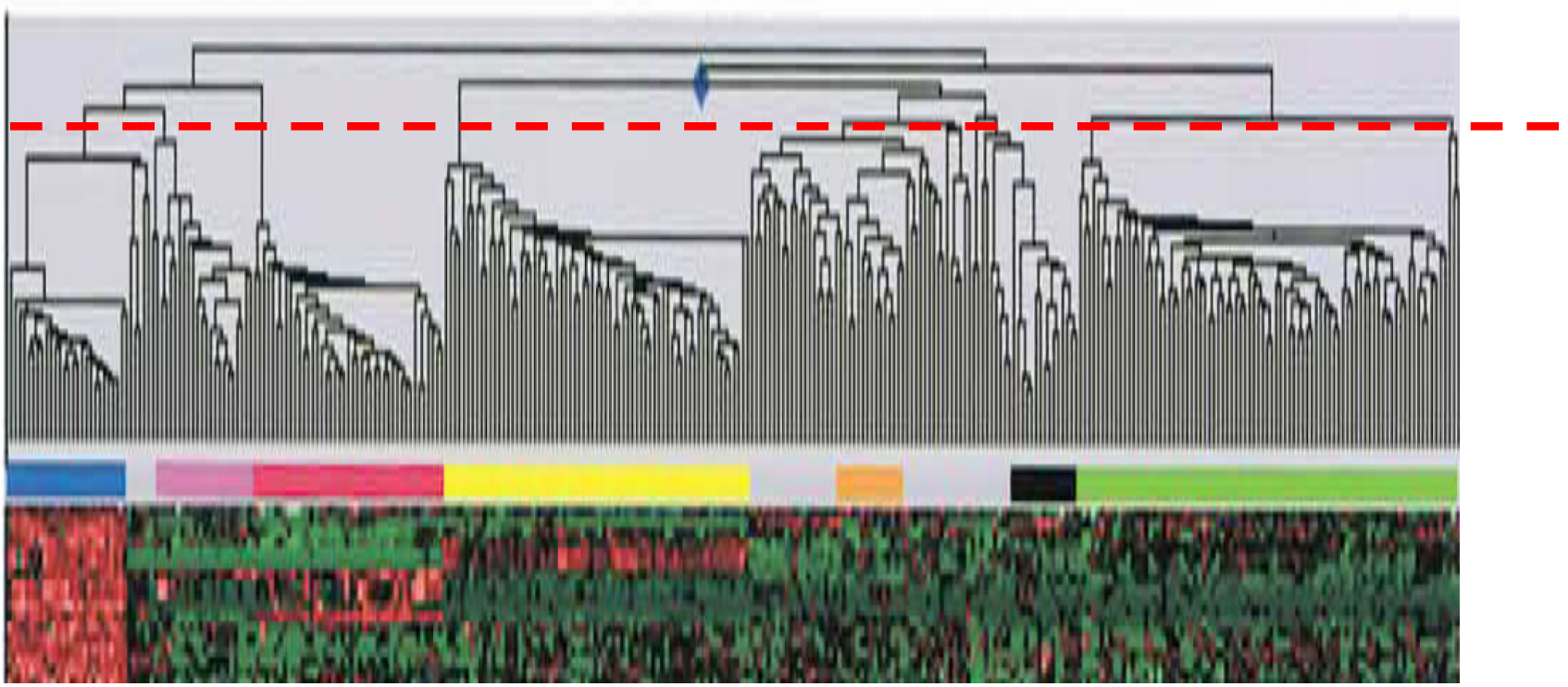
mit $n_i = |C_i|$, $n_j = |C_j|$ Größe der zugehörigen Cluster

Hierarchisches Clustern

Yeoh *et al*, Cancer Cell (2002)



Hierarchisches Clustern



- uns interessieren nur die letzten (3-20) Schritte des Clusterings