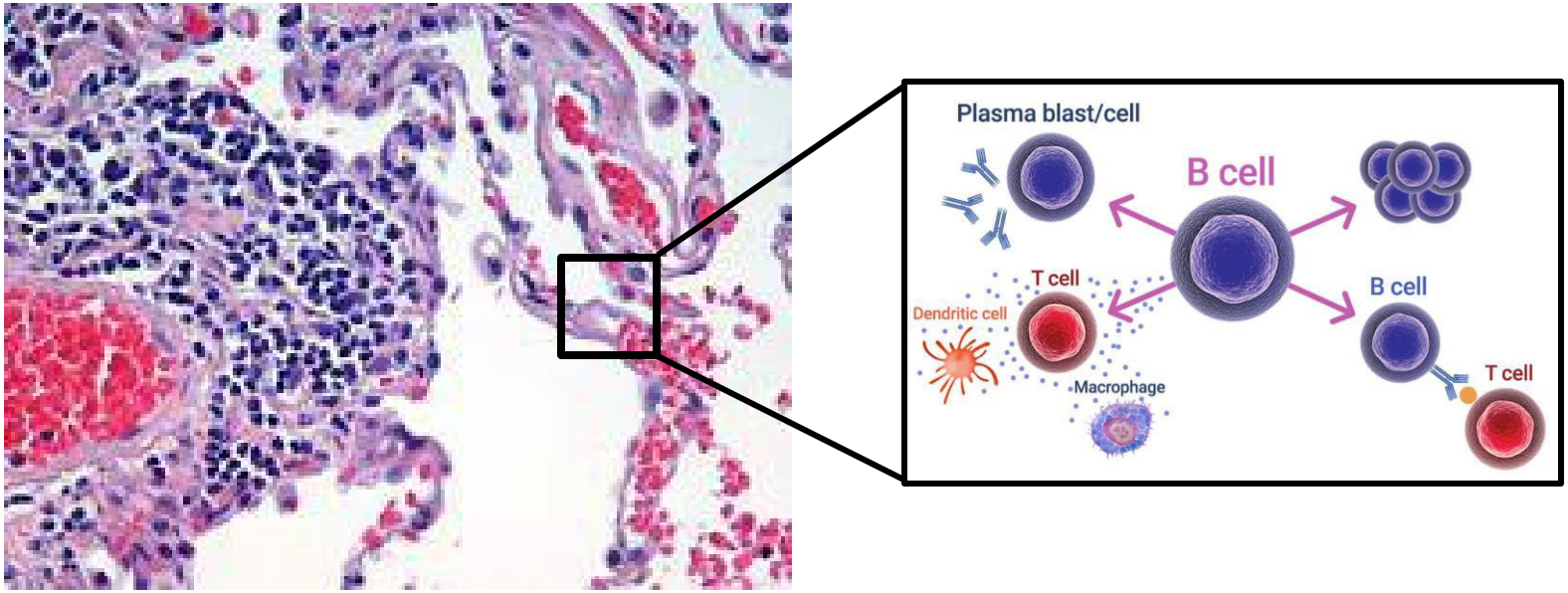
Conclusioni e applicazioni

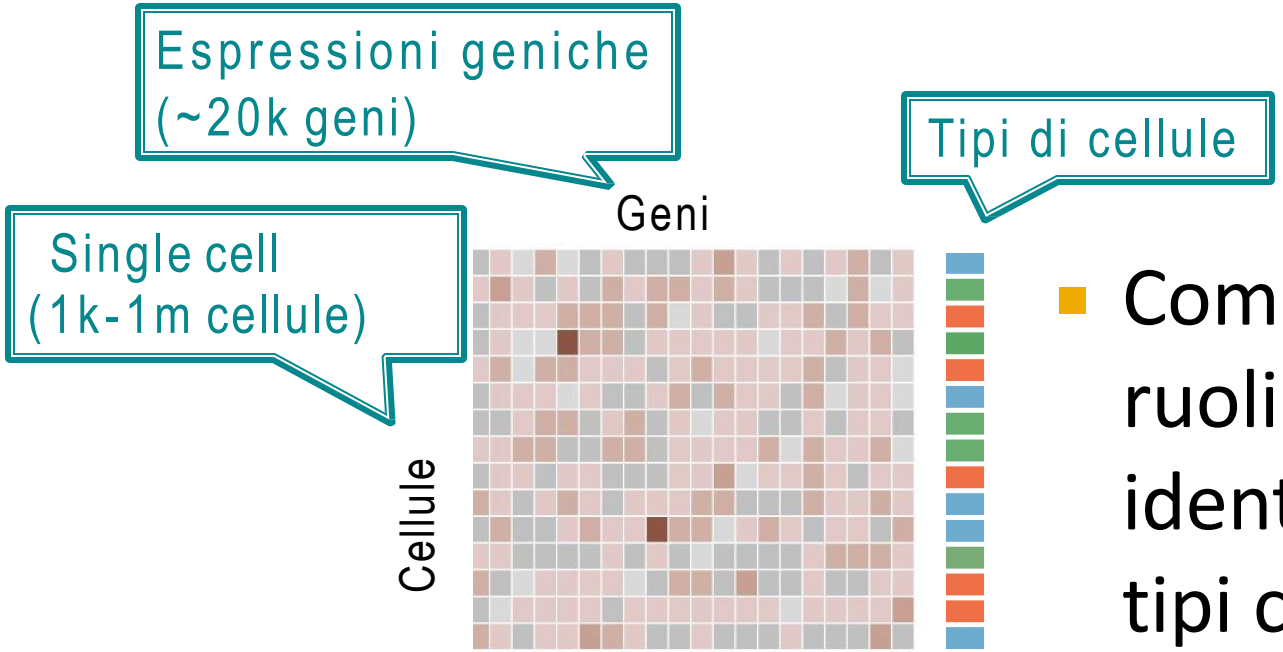
Ogni cellula di un tessuto ha un ruolo specifico



Sfida:

Come determinare i ruoli delle cellule?

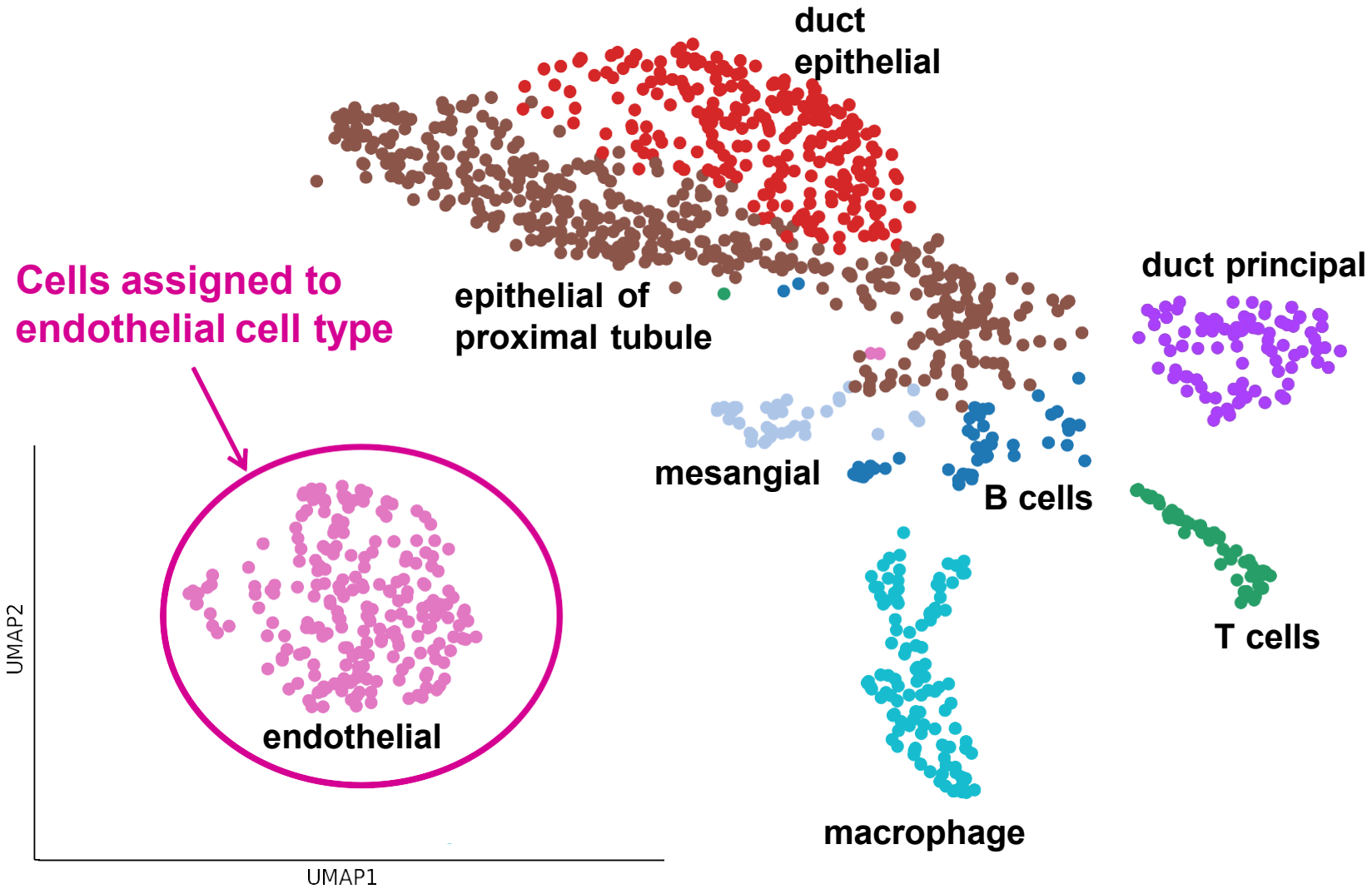
Identifichiamo i tipi cellulari



- Comprendere i ruoli delle cellule: identificazione dei tipi cellulari

- **Attività di identificazione del tipo di cellula: date le espressioni geniche delle cellule, assegnare le cellule ai tipi di cellule**
- Si riduce a una **Attività di clustering:** Raggruppa le cellule in base alle loro somiglianze di espressione genica

Identifichiamo i tipi cellulari



[UMAP. McInnes, Healy, Melville. '18]

Come facciamo il clustering

- **È possibile utilizzare metodi di clustering standard come K-means per risolvere questo problema?**
- **Perché i metodi cluster standard non funzionano bene?**
  - I dati sono molto dimensionali (~20k geni per cellula)
  - I dati sono rumorosi e sparsi (la maggior parte dei valori sarà zero)
  - Il numero di cluster (tipi di cellule) è sconosciuto
  - I tipi di cellula sono organizzati gerarchicamente
    - **La definizione del tipo di cella è provvisoria**
    - **Un tipo di cellula può avere più sottotipi di celle**
    - **Dove mettere una soglia su una definizione di un tipo di cellula?**

Rappresentiamo le cellule come un grafo

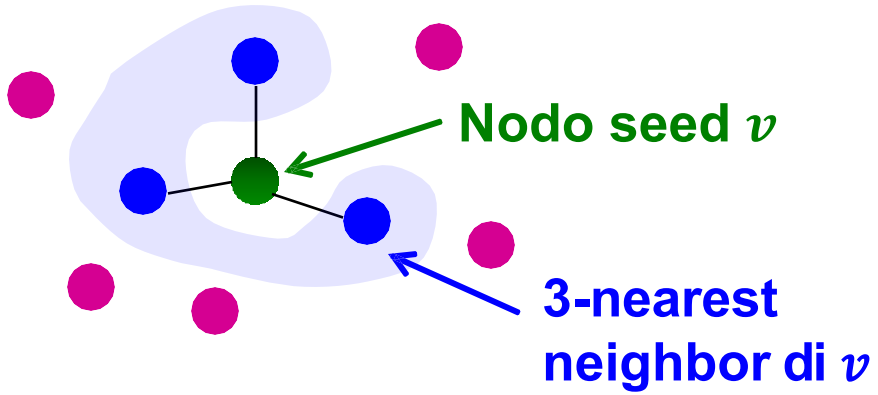
- **Idea:** Costruisci un **grafo** tra i punti dati (cellule) e identifica **comunità gerarchiche**

Perché il grafo è una buona rappresentazione?

- **Rappresentazione naturale:** modelli le **interazioni** cellula-cellula
- Le cellule con espressioni geniche più simili hanno maggiori probabilità di interagire
  - Costruire un grafo basato sulle somiglianze tra le espressioni geniche delle cellule
- Le comunità di rete gerarchiche modellano la gerarchia dei tipi di cellule

Grafo K-NN

- **K nearest neigbor (K-NN) graph:** Grafo diretto con l'insieme di vertici V e uno arco da ogni v ∈ V al suoi K **Oggetti più simili** in V sotto una determinata similarità
  - *Es.*, similarità del coseno, l_2, l_1



Come costruirlo

- **Algoritmo burte force:**
  - $O(n^2)$
  - Pratico solo per piccoli set di dati!

Come calcolare in modo efficiente il grafico K-NN?

- **NN-Descent** [Dong, Charikar, Li. '11]
  - Metodo scalabile per la creazione di un grafo K-NN approssimato
  - Adatto per set di dati su larga scala
  - Il costo empirico è di circa $O\left(n^{1.14}\right)$
  - Adatto per l'implementazione distribuita (ad esempio, Map Reduce)

NN-Descent

NN-Descent è un algoritmo di raffinamento iterativo:

- Inizia con un grafico KNN casuale
  - Ogni nodo sceglie K nodi casuali come vicini più prossimi.
- Perfezionare in modo iterativo l'elenco dei vicini più prossimi di ogni nodo:
  - **Un vicino di casa di un vicino potrebbe anche essere il mio vicino.**
- Continuate così fino alla convergenza.

NN-Descent

- Inizia con insieme K-NN casuale campionando K item per ogni nodo $v \in V$
- Quindi in modo iterativo per ogni nodo $v \in V$:
  - $\boldsymbol{B[v]}$ ... è l'attuale/approssimato K-NN di v
  - $\boldsymbol{R[v]}$ ... è l'attuale reverse del K-NN di v
    - **Reverse K-NN**: $\boldsymbol{R}[v] = \{u \in V | v \in B[u]\}$
  - **Vicinato generico** $\boldsymbol{B^*[v] = B[v] \cup R[v]}$
  - Per ogni $\boldsymbol{u \in B^*[v]}$, verifica la similarità tra $\boldsymbol{v}$ e $\boldsymbol{B^*}[u]$ (Vicini generali di u sono candidati ad essere nuovi vicini di v)
  - Aggiorna l'elenco dei vicini più vicini se la somiglianza è maggiore rispetto all'insieme dei vicini correnti

L'algoritmo

- **NNDescent(V,σ, K):**

  $B[v] = Random\ sample\ of\ K\ items\ V,\ \forall v \in V$

  Loop:

  $R = reverse\ (B)$

  $B^*[v] = B[v] \cup R[v],\ \forall v \in V$

  $c = 0$

  for $v \in V$:

  for $u_1 \in B^*[v], u_2 \in B^*[u_1]$:

  $l = \sigma(v, u_2)$

  $c = c + updateNN\ (B[v], \langle u_2, l \rangle)$

  return B **if** $c = 0$

V … dataset
σ …similarity oracle
K …number of neighbors
$B[v]$ … approximate neighbors of v
$R[v]$ … approximate reverse neighbors of v
$B^*[v]$ … approximate general neighbors of v
c … counter

$B[v]$ è organizzato come un heap
➔ updates cost $O(logK)$

- $reverse(B)$:

  $R[v] = \{u \mid \langle v, \dots \rangle \in B[u]\},\ \forall v \in V$

  return R

- $updateNN(H, \langle u, l, \dots \rangle)$:

  Update KNN heap H

  return 1 if changed, 0 if not

Es. K=2

Vicini:

$$B[s] = \{c, d\}$$

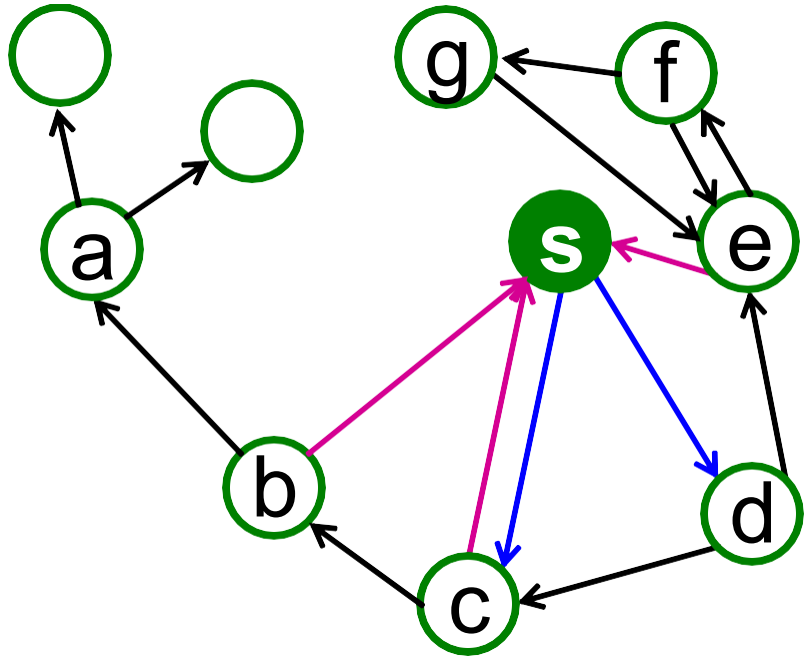
Vicinato inverso

$$R[s] = \{b, c, e\}$$

Vicinato generale

$$B^*[s] = \{b, c, d, e\}$$

$$B^*[b] = \{a, c, s\}$$
$$B^*[c] = \{b, d, s\}$$
$$B^*[d] = \{c, e, s\}$$
$$B^*[e] = \{d, f, g, s\}$$



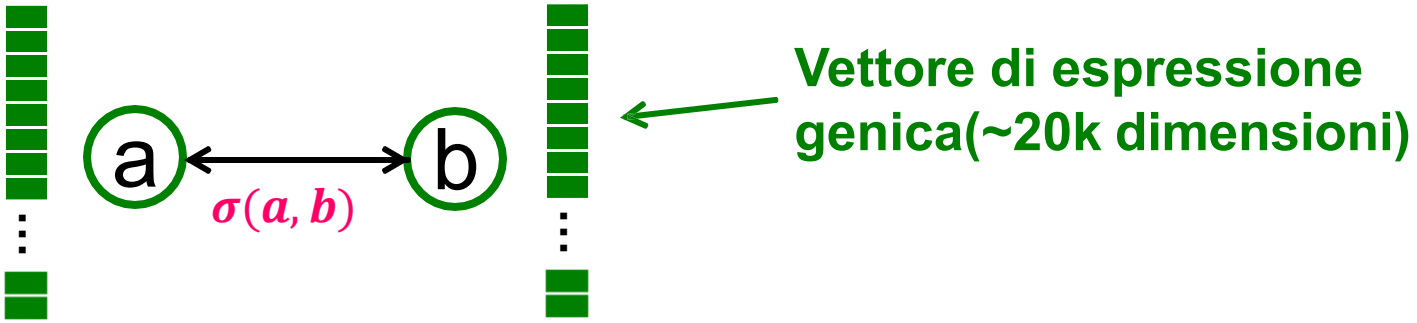
Vicinato generale $B^*[s]$
Contiene i nuovi candiati per $B[s]$

Verificheremo {a, b, e, f, g} come prossimi candidati per B[s]:
Calcolare σ(s, a), σ(s, b), σ(s, e), σ(s, f), σ(s, g) e aggiorniamo i NNs di s

Le frecce indicano i vicini di un particolare nodo. Ad esempio, la freccia da b a s indica che b ha selezionato s come vicino (ma non è necessario che sia vero il contrario).

Identificazione dei tipi cellulari

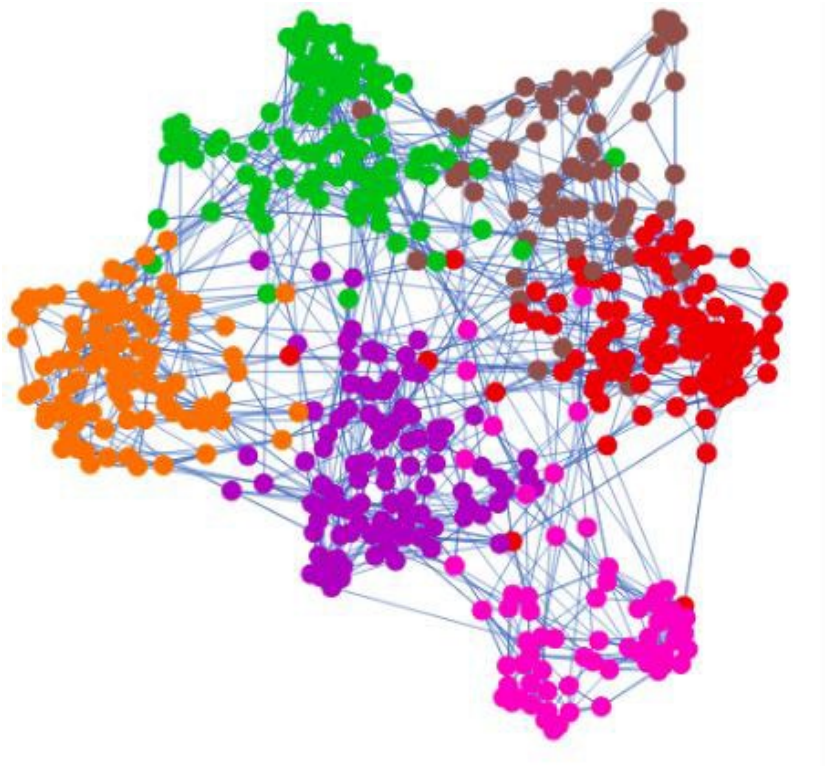
- **Quale misura di similarità σ usare?**
  - Le cellule vengono confrontate in base ai loro profili di espressione genica
  - **Sfida:** Il numero di geni è molto alto



  - **Approccio**: Applichiamo **SVD** (circa 50 dimensioni) a quindi calcoliamo la distanza l_2 nello spazio a più bassa dimensione

E dopo? Come identifichiamo i cluster?

- Una volta creato il grafo K-NN delle celle, come facciamo **a identificare le comunità di rete?**



Torniamo all’identificazione dei tipi cellulari

Input:

Dati di espressione genica di singole cellule

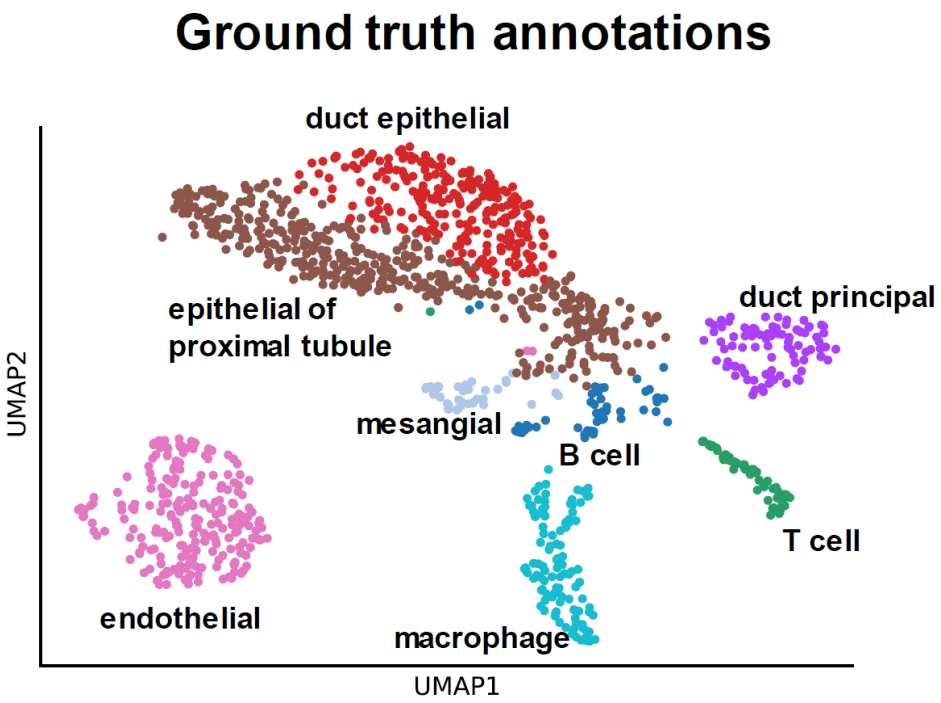
Passi:

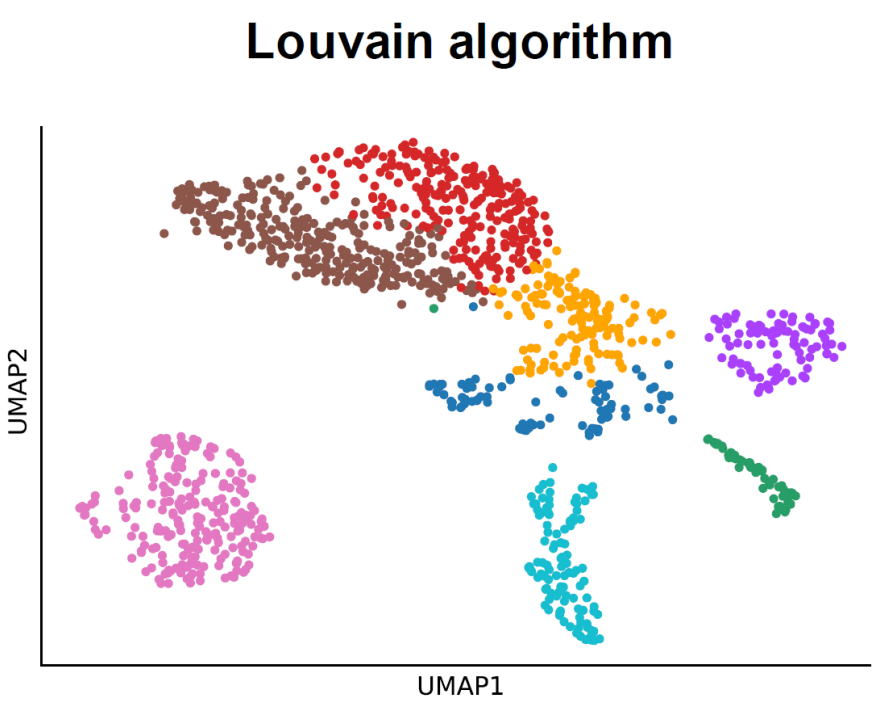
1) Applicare SVD ai dati di espressione genica cellulare(~50 dim)

2) Creare K-NN (K=15) grafico tra le espressioni geniche delle cellule a bassa intensità

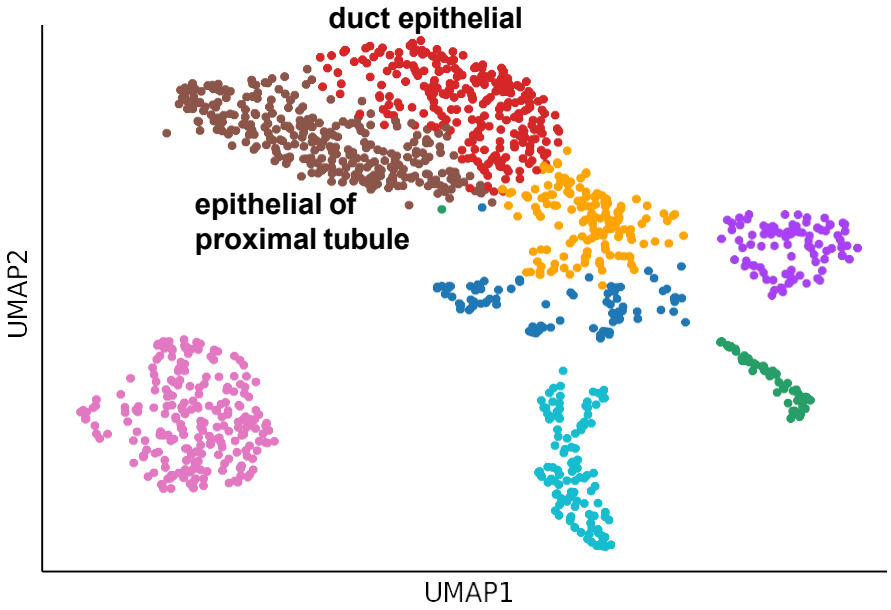
3) Applicare l'algoritmo di Louvain per identificare i cluster

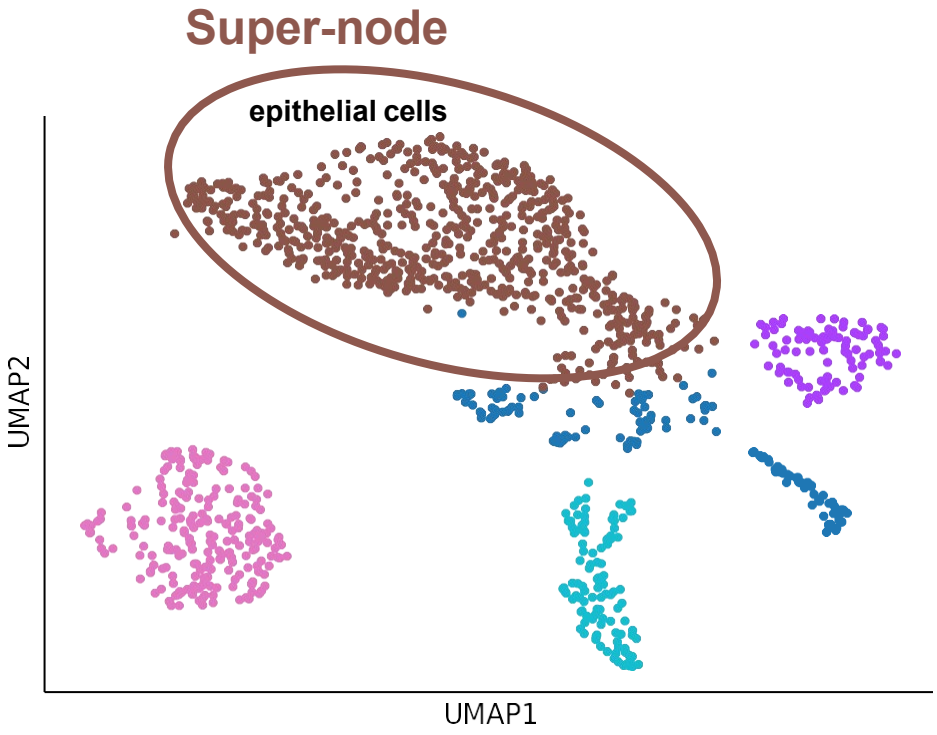
Identificazione dei tipi cellulari

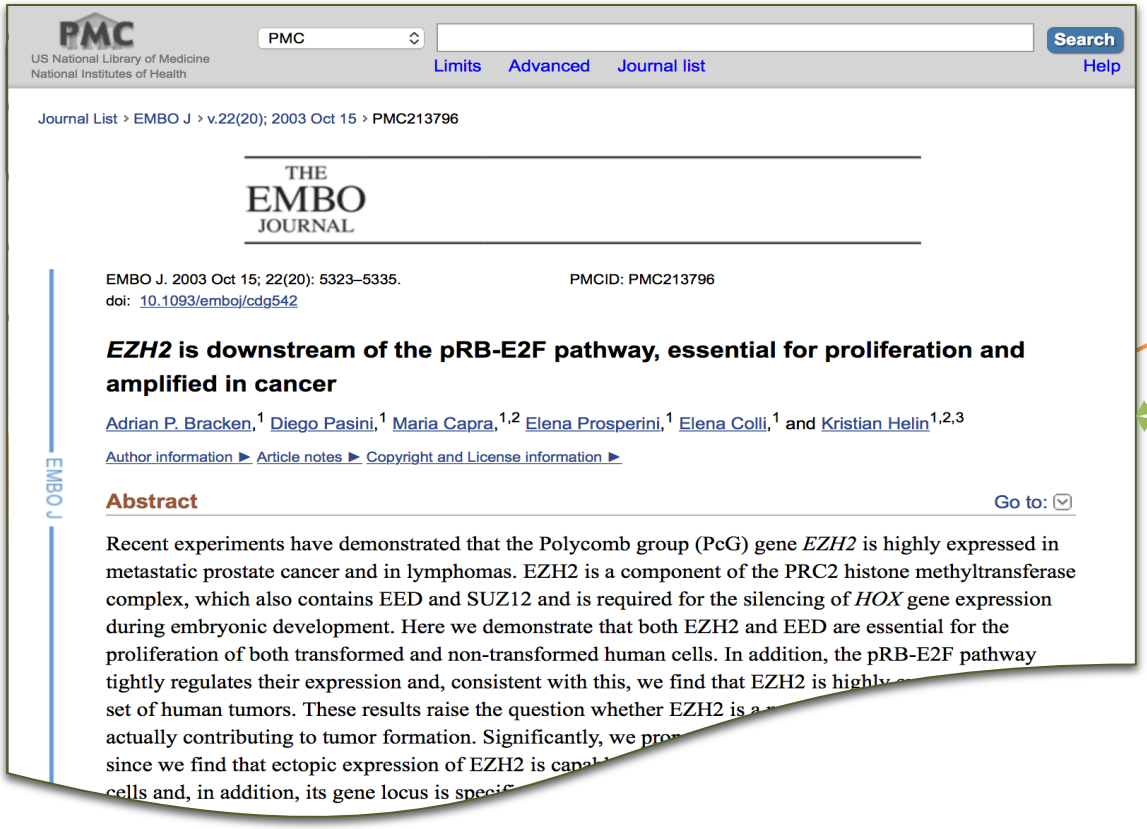




Cluster


duct epithelial
epithelial of
proximal tubule
UMAP2
UMAP1


Super-node
epithelial cells
UMAP2
UMAP1



PMC
US National Library of Medicine
National Institutes of Health

PMC Search
Limits Advanced Journal list Help

Journal List › EMBO J › v.22(20); 2003 Oct 15 › PMC213796

THE EMBO JOURNAL

EMBO J. 2003 Oct 15; 22(20): 5323–5335. PMCID: PMC213796
doi: 10.1093/emboj/cdg542

***EZH2* is downstream of the pRB-E2F pathway, essential for proliferation and amplified in cancer**

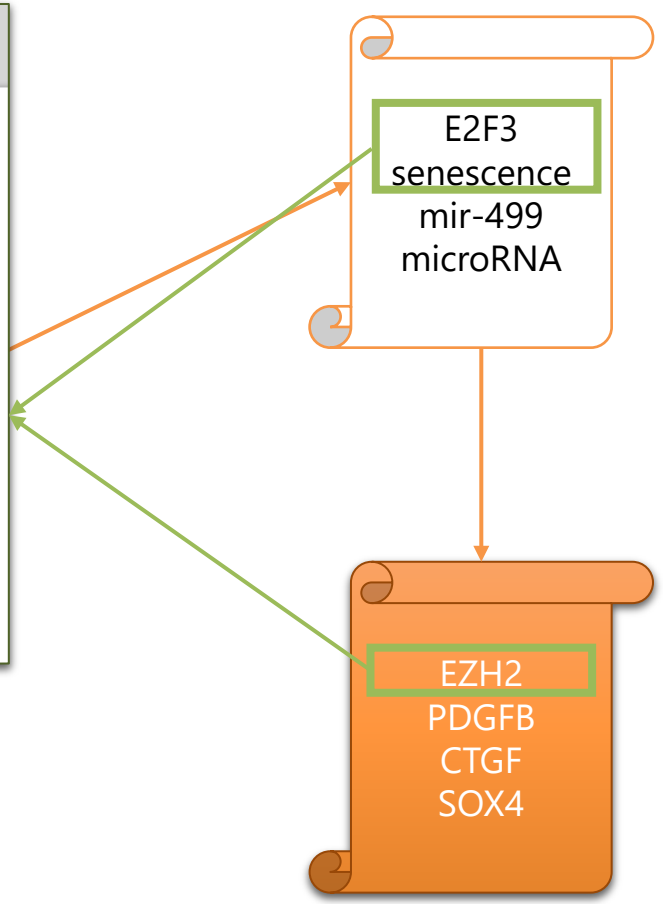
Adrian P. Bracken,[1] Diego Pasini,[1] Maria Capra,[1,2] Elena Prosperini,[1] Elena Colli,[1] and Kristian Helin[1,2,3]

Author information ► Article notes ► Copyright and License information ►

EMBO J

Abstract Go to:

Recent experiments have demonstrated that the Polycomb group (PcG) gene *EZH2* is highly expressed in metastatic prostate cancer and in lymphomas. EZH2 is a component of the PRC2 histone methyltransferase complex, which also contains EED and SUZ12 and is required for the silencing of *HOX* gene expression during embryonic development. Here we demonstrate that both EZH2 and EED are essential for the proliferation of both transformed and non-transformed human cells. In addition, the pRB-E2F pathway tightly regulates their expression and, consistent with this, we find that EZH2 is highly
set of human tumors. These results raise the question whether EZH2 is a
actually contributing to tumor formation. Significantly, we pro
since we find that ectopic expression of EZH2 is capa
cells and, in addition, its gene locus is speci



Problema

- Supponiamo di avere un insieme di documenti ognuno dei quali ricopre un insieme di concetti.
- Selezionare k documenti modo da coprire più concetti possibile:
  - Set Cover
  - Soluzione Greedy (performance garantita)

Modello astratto

- Supponiamo di avere un set di documenti D
  - Ogni documento d copre un insieme di X_d parole/topic o entità
- Per un insieme di documenti $A \subseteq D$ definiamo

$$F(A) = \left| \bigcup_{i \in A} X_i \right|$$

Vogliamo

$$\max_{|A| \leq k} F(A)$$

Set Cover

- Dato un universo di elementi $W = \{w_1, \dots, w_n\}$ siano $X_1, \dots, X_m \subseteq W$
- Trovare k insiemi X_i che ricoprono al massimo W
  - Trovare k insieme X_i la cui dimensione dell'unione è la piu' grande possibile
  - Problema NP-Completo

Euristica Greedy (approssimazione garantita quando?)

$A_o = \{\}$

For $i = 1 \dots k$

trova l'insieme d $\max F(A_{i-1} \cup \{d\})$

$A_i = A_i \cup \{d\}$

Teorema

La procedura greedy produce una soluzione A dove

$F(A) > \left(1 - \frac{1}{e}\right) * F(A_{OPT})$ quando F() ha le seguenti proprietà

- F è **monotona** se $A \subseteq B \; allora \; F(A) \leq F(B)$ e F({})=0
- F è **submodulare:** aggiungere un elemento all'insieme da meno miglioramento che aggiungerlo ad uno dei suoi sottoinsiemi

Submodularità

- $\forall\, A, B \subseteq W$:

$$F(A) + F(B) \geq F(A \cup B) + F(A \cap B)$$

Oppure

- $\forall\, A \subseteq B, d\ in\ W - B$:

$$F(A \cup \{d\}) - F(A) \geq F(B \cup \{d\}) - F(B)$$

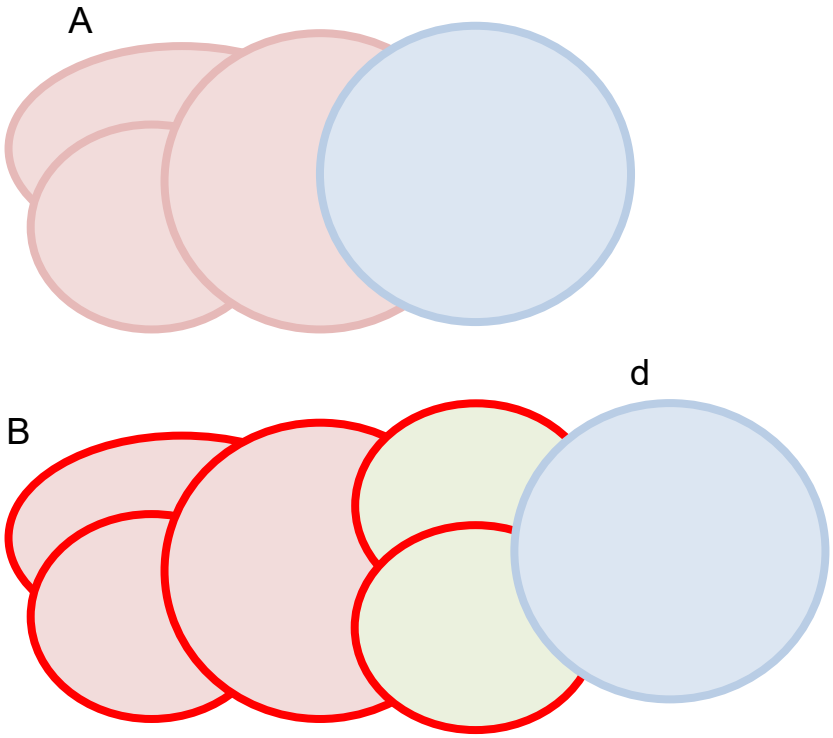
Ha un andamento decrescente

Esempio: Set Cover

$d_1, d_2, \ldots, d_m$ Insiemi

$$F(A) = \left| \bigcup_{i \in A} d_i \right|$$

F e' sub-modulare



Proprietà

- Siano $F_1, F_2, \dots F_m$ m funzioni sub-modulairi e siano $\lambda_1 \lambda_2 \dots \lambda_m > 0$ allora:

$$F(\mathrm{A}) = \sum_{i=1\dots m} \lambda_i F_i(A)$$

È una funzione sub-modulare

- La media di un insieme di funzioni submodulari è submodulare:

$$F(\mathrm{A}) = \sum_{i} P(i) * F_i(A)$$

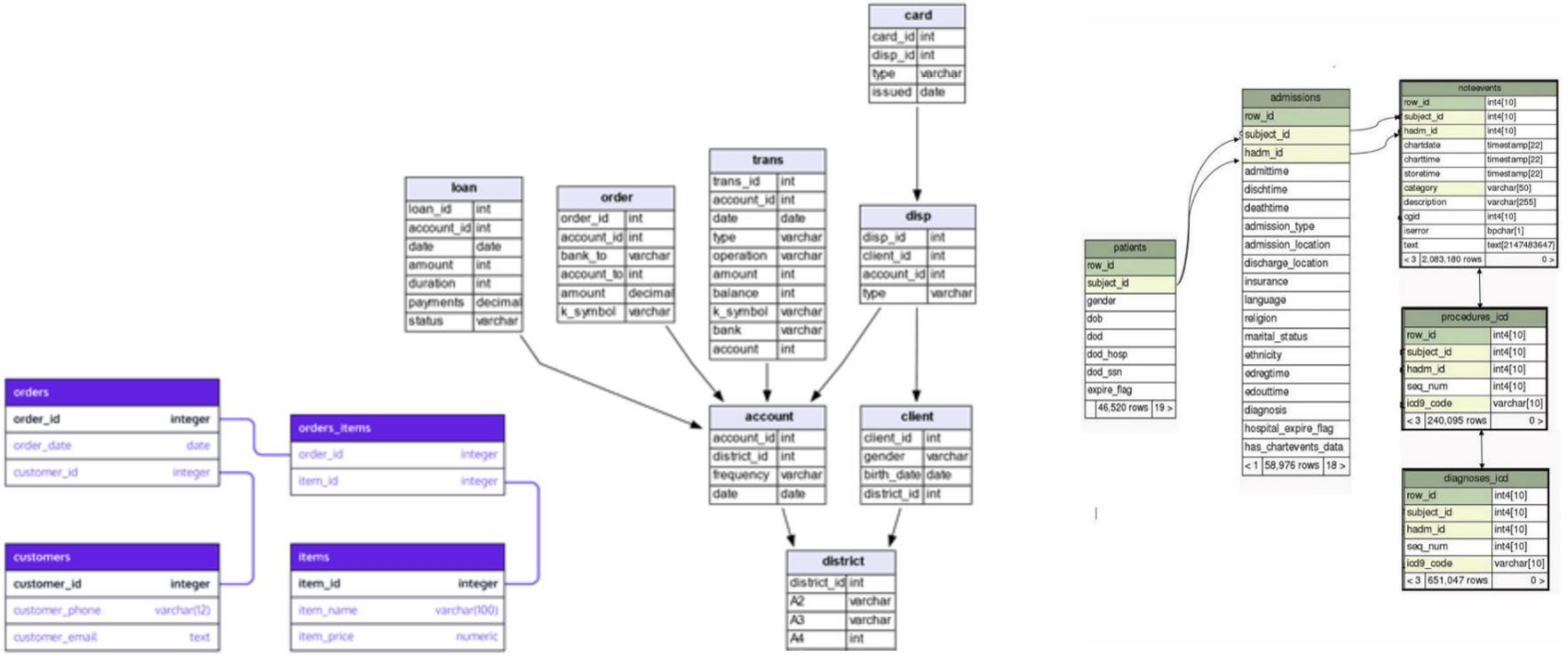
- Importanza del concetto
  - Ogni concetto c ha un importanza w_c
- Funzione di ricoprimento del doc
  - $\text{cover}_d(c) =$ probabilità che il documento d copra c concetto
  - Funzione di ricoprimento:

$$cover_A(c) = 1 - \prod_{d \in A}(1 - cover_d(c))$$

Funzione obiettivo submodulare

$$\max_{A:|A| \leq k} F(A) = \sum_c w_c cover_A(c)$$

Relational Deep Learning



Commerce Finance Health care

I dati relazionali sono grafi