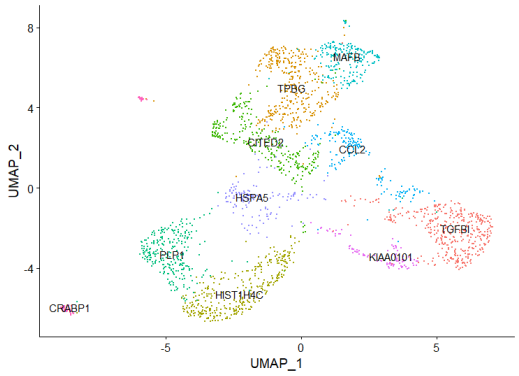
Notes meeting 25-11

Lucas Jansen Klomp

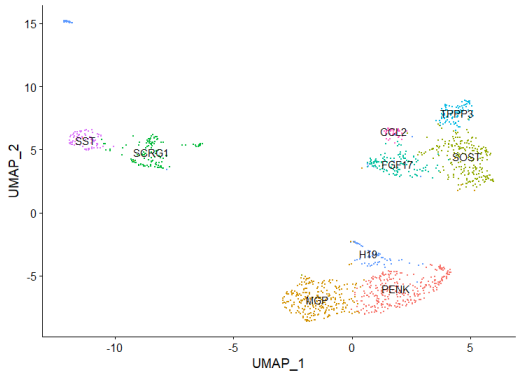
November 2021

1 UMAP analysis of the single-cell data

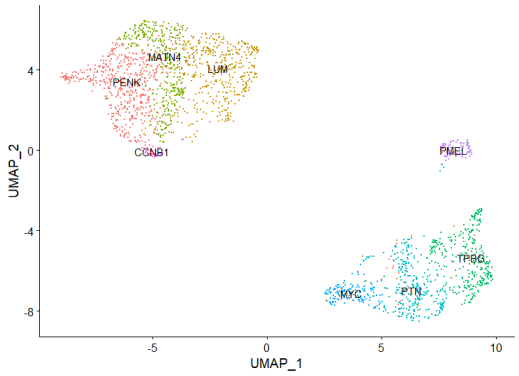
For these notes we use the data set from Wu et al. (2021). Particularly, we use the single cell RNA-seq data as collected at five different time points, namely day 1, 7, 14, 28 and 42. We first perform a dimension-reduction using UMAP (McInnes et al., 2018). This analysis should reveal clusters of similar cells in the data set. The results for the different time points is shown in Figure 1. For each identified cluster, the most highly differentially expressed feature is given. The code can be found on github.



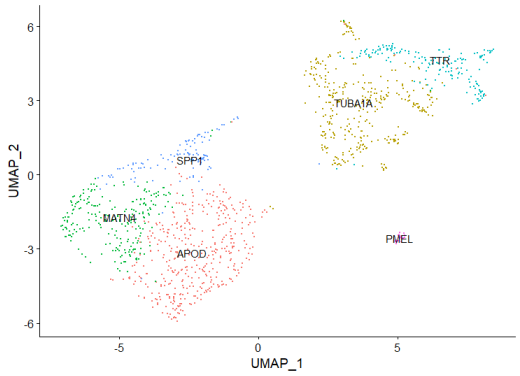
(a) Clusters derived using UMAP for day 1.



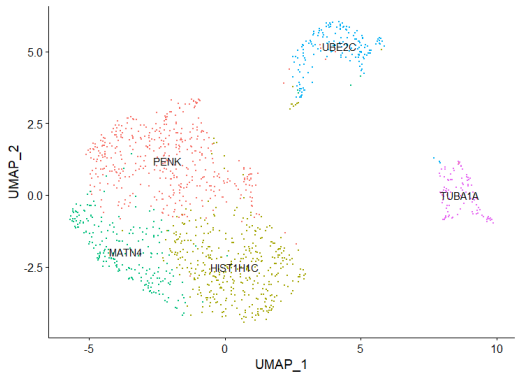
(b) Clusters derived using UMAP for day 7.



(c) Clusters derived using UMAP for day 14.



(d) Clusters derived using UMAP for day 28.



(e) Clusters derived using UMAP for day 42.

Figure 1: Clusters in scRNA data derived from iPSCs captured at different time points.

2 Naive network construction from scRNA-seq data

Next, we describe a simple method to derive an undirected co-expression network from the single cell RNA sequencing data. For this, we use the log-normalized data. We use a method similar to the network construction with hard thresholding as implemented in WGCNA (Langfelder and Horvath, 2008). We have samples $x_j \in \mathbb{R}^N$ for each feature $j \in \{1, ..., K\}$. We use Pearson's correlation coefficient to derive the similarity matrix S where

$$S_{i,j} = \rho_{x_i, x_j} = \frac{\text{cov}(x_i, x_j)}{\sigma_{x_i} \sigma_{x_j}}.$$

An undirected network can be constructed from the similarity matrix by constructing the adjacency matrix A with elements

$$A_{i,j} = \begin{cases} 1 \text{ if } |S_{i,j}| \geq \tau, \\ 0 \text{ otherwise.} \end{cases}.$$

Here, $A_{i,j} = 1$ if a connection exists in the graph between nodes i and j. The adjacency matrix is sufficient to define the undirected co-expression network. We set $\tau = 0.35$. This value was chosen through **trial and error** as choosing larger values of τ tends to yield very sparse networks, while lower values of τ tend to yield very dense networks.

As the data set is quite large we have opted to remove features where little expression is found across all samples. This is mainly done to avoid problems with the amount of working memory required to compute the similarity matrix. We remove features from the data set where the sum of the normalized expressions across all samples is lower than a certain threshold.

Alternatively, it might be a good option to instead remove features for which the expression has a low variance, as these are perhaps less interesting when constructing a co-expression network. However, we have not tested this.

3 Extracting important nodes from undirected co-expression networks

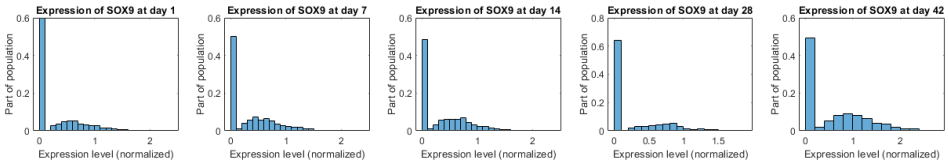
In order to extract important nodes from the obtained co-expression networks, we use the PageRank algorithm as implemented in MATLAB (Page et al., 1999). In Table 1 the 15 features with the highest PageRank are given for the available time points using networks which are constructed as described in Section 2.

Table 1: Ranking of the features with the highest PageRank for each of the considered time points.

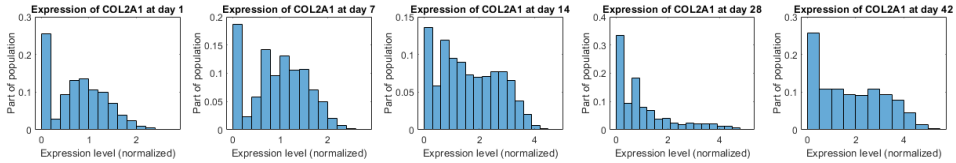
	D1	D7	D14	D28	D42
1	TPBG	CLU	PENK	LGALS1	EIF1
2	CNTNAP2	RPL10	COL1A2	TUBA1A	SOX2
3	S100A11	CNTNAP2	S100A11	BGN	PENK
4	COL3A1	PIFO	MGP	EMP3	SPARC
5	PLP1	TPBG	COL9A3	TIMP1	GAP43
6	CITED2	MGP	BGN	CENPV	STMN4
7	TMSB10	GSTP1	HTRA1	MAP1B	JUNB
8	TGFBI	PENK	COL3A1	TUBB2B	GPM6A
9	PRRX1	ARL4C	SPATS2L	LAPTM4A	HAPLN1
10	PCOLCE	TPPP3	FZD3	STMN1	CD59
11	COL1A1	C1orf194	MGST1	S100A11	PCOLCE
12	MAFB	COL1A1	SPARC	COL1A2	COL1A2
13	H2AFZ	C11orf88	COL1A1	COL6A2	COL1A1
14	TUBA1B	C1orf192	PMP22	TMSB15A	NR2F1
15	ARL4C	HMGB1	SOX2	PLP2	BEX1

4 Distinguising chondrocytes

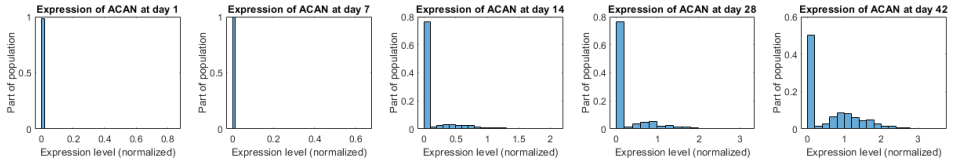
Next, we attempt to distinguish that have differentiated into chondrocytes in the data set. To do this, we consider three markers of chondrocytes: SOX9, Collagen Type II Alpha 1 Chain (COL2A1) and Aggrecan (ACAN). First, we consider the expression of these three markers at each time point. A histogram of the expression for each of the three markers is given in Figure 2.



(a)



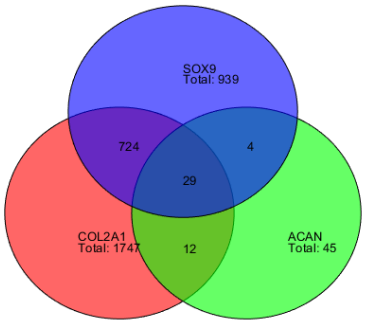
(b)



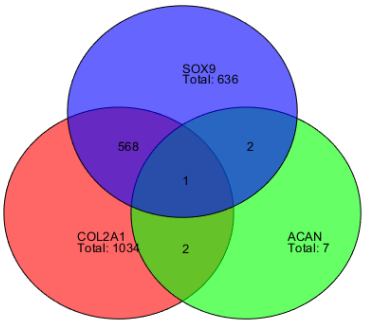
(c)

Figure 2: Distribution of expression levels for three chondrocyte markers.

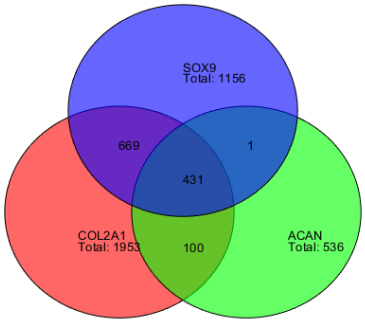
Next, we investigate whether these three markers are often simultaneously expressed. To do this, we provide Venn diagrams for each of the five time points in Figure 3.



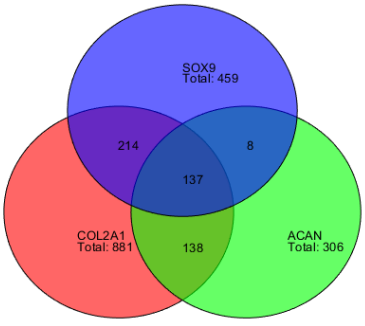
(a) Venn diagram for day 1.



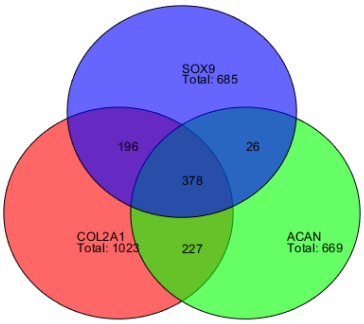
(b) Venn diagram for day 7.



(c) Venn diagram for day 14.



(d) Venn diagram for day 28.



(e) Venn diagram for day 42.

Figure 3: Venn diagrams indicating the combined expression of three chondrocyte markers.

References

Langfelder, P. and Horvath, S. (2008). WGCNA: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Wu, C.-L., Dicks, A., Steward, N., Tang, R., Katz, D. B., Choi, Y.-R., and Guilak, F. (2021). Single cell transcriptomic analysis of human pluripotent stem cell chondrogenesis. *Nature communications*, 12(1):1–18.