Supplementary Information

Identification of cell types from single-cell transcriptomes using a novel clustering method

Chen Xu[1], Zhengchang Su[1,*]

[1]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

[*]To whom correspondence should be addressed.

1 Supplementary Text

Validation measures

We use three external validation measures, Purity, Adjusted Rand Index (Hubert and Arabie, 1985) and F_1 score (van Rijsbergen, 1974), to evaluate the performance of the clustering methods. Let U be the set of genuine classes (cell types) and V be the set of our computed clusters. Purity first assigns each cluster v_i to the class u_j that is the most frequent in the cluster. Then the total number of correctly assigned objects (cells) is divided by the total number of objects in the dataset (N):

$$Purity = \frac{1}{N}\sum_{i} \left(v_i \cap u_j\right).$$

ARI is one of the most successful measure of the agreement between two partitions with different number of classes/clusters. It is computed by:

$$ARI = \frac{\binom{N}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{N}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]},$$

where 'a' is the number of pairs of objects in the same class in U and the same cluster in V; 'b' is the number of pairs in the same class in U but not the same cluster in V; 'c' is the number of pairs that are not in the same class in U but in the same cluster in V; 'd' is the number of pairs that are neither in the same class in U nor in the same cluster in V. The F_1 score is the harmonic mean of precision and recall:

$$F_1 score = \frac{2 \cdot a}{2 \cdot a + b + c}.$$

Data points that are treated as noise or singletons are excluded from the calculation of Purity. In calculating ARI and F1 score, noise or singletons are treated as individual clusters.

Novelty of SNN-Cliq

The similarity measure in SNN-Cliq is based on the technique of shared nearest neighbor (SNN), which has been applied to several recent clustering applications (Ertöz, *et al*., 2003; Guha, *et al*., 2000; Jarvis and Patrick, 1973). Depending on the problem, different SNN similarity functions were proposed. For example, the similarity between objects x_i and x_j can be simply defined to be the intersection size of their *k*-nearest-neighbor list (Houle, *et al*., 2010). Other functions take the ordering of the nearest neighbors into account.

In a density based clustering approach, Ertöz, *et al*. (2003) took the sum of the similarities of a point's nearest neighbors as a local density measure: $strength(x_i, x_j)=\sum(k+1\text{-}rank(v, x_i))(k+1\text{-}rank(v, x_j))$, where v is a shared neighbor and $rank(v, x_i)$ is the position of v in x_i's list. In our paper, we define a new SNN function that only considers the ordering of the common neighbor that is on average the closest to x_i and x_j (the function is present in Methods 2.1). It emphasizes the closeness between points instead of the local density, thereby not discarding points in very low density regions. In addition, we believe that this SNN function is more tolerant to changes in the parameter k. Finally, our function also extends the concept of SNN to construct a weighted similarity graph.

Furthermore, although the graph clustering step in the SNN-Cliq method is inspired by Zhang et al (Zhang, *et al*., 2009), they differ in many ways due to the differences of target graphs and ultimate goals. Zhang's method aims to cut down a large and dense graph to small parts for the purpose of computational efficiency in further steps; thus, it allows overlaps between resulting subgraphs. However, SNN-Cliq aims to partition a sparse graph into distinct clusters with no overlap in between. We delineate the differences between the two algorithms in the following three points.

First, Zhang's method starts by identifying cliques in a graph, because the graph it deals with is dense and large. By contrast, SNN-Cliq starts by searching for quasi-cliques that allow missing edges between nodes in a subgraph, because the graphs we deal with are usually sparse due to the similarity is calculated by shared nearest neighbor. Second, Zhang's method iteratively combine cliques/subgraphs by checking with two criteria: $|S1 \cap S2| /\min(|S1|, |S2|) >0.9$ and $|S1 \cap S2| /\max(|S1|, |S2|)>0.7$. As a result, S1 and S2 are only merged when the intersection size is large enough in both subgraphs. The high threshold (0.7 and 0.9) used will fail to merge many overlapping subgraphs, but this does not affect their results since their goal is to cut a dense graph instead of a hard clustering. In fact, their resulting subgraphs are still very dense and are similar to the quasi-cliques we find in the first step. In SNN-Cliq, we only require one criterion: $|S1 \cap S2| /\min(|S1|, |S2|) > 0.5$, to merge subgraphs. The purpose of this design is to allow the quasi-cliques to grow into non-spherical clusters. Finally, in the case of a node appearing in multiple clusters, Zhang's method does not assign a node into a particular cluster. By contrast, SNN-Cliq always allocates a node to the nearest cluster to achieve hard clustering.

2 Supplementary Figures

Supplementary Figure S1

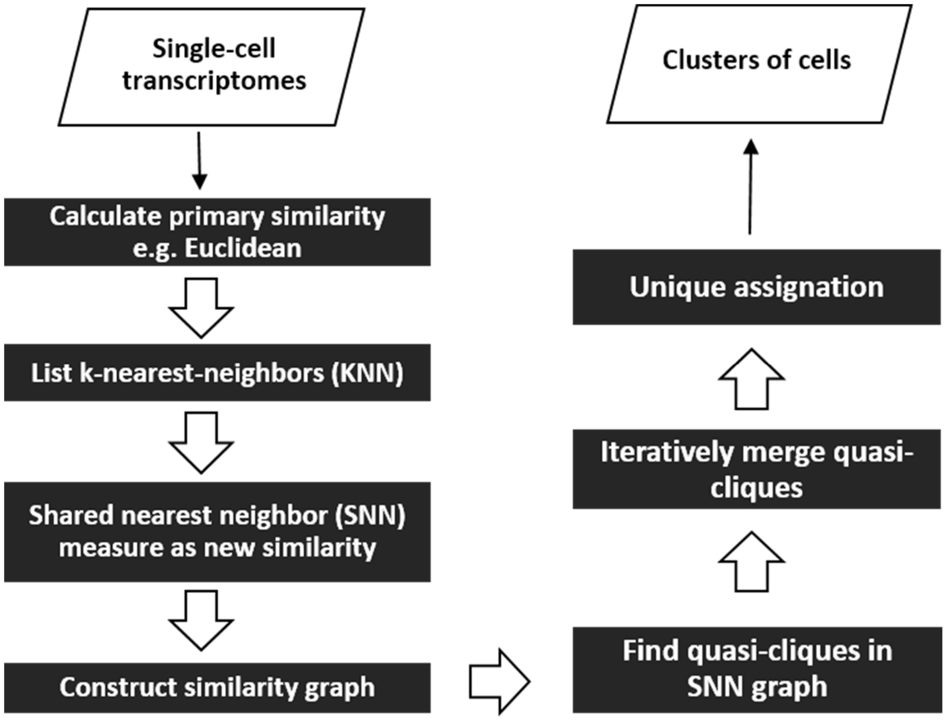

Single-cell transcriptomes
Calculate primary similarity e.g. Euclidean
List k-nearest-neighbors (KNN)
Shared nearest neighbor (SNN) measure as new similarity
Construct similarity graph
Find quasi-cliques in SNN graph
Iteratively merge quasi-cliques
Unique assignation
Clusters of cells

Fig. S1. An overview of the SNN-Cliq algorithm.

Supplementary Figure S2

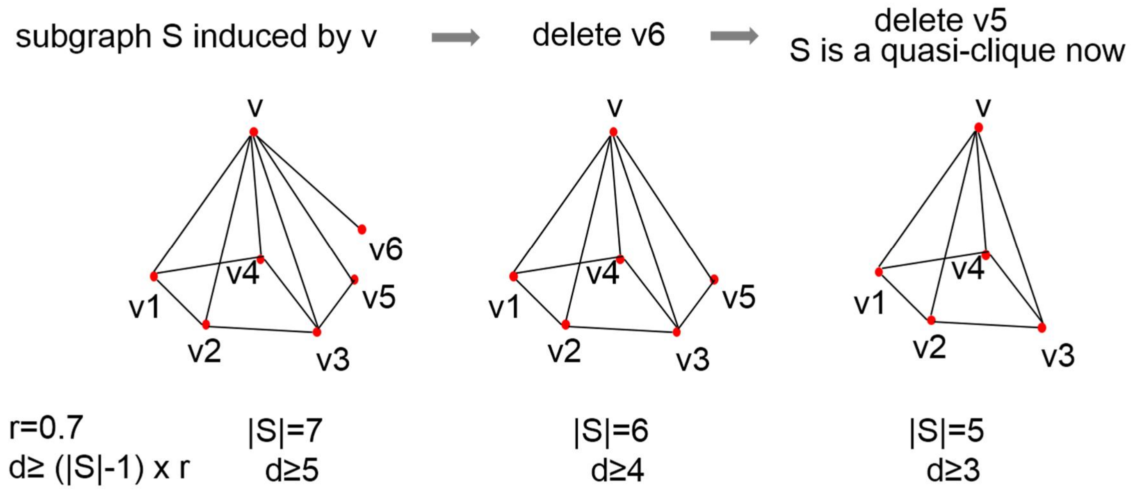


Fig. S2. Finding quasi-clique associated with a node. A schematic to illustrate how to greedily find a quasi-clique associated with a node v in the subgraph induced by v and its neighbors. Initially, S includes seven nodes and v6 has the minimum degree (d_{v6}=1). Since a quasi-clique of r=0.7 requires a node to have at least 5 neighbors (d≥5), v6 is removed from S. In the new S with six nodes, the threshold of degree becomes d≥4. After v5 is deleted, all nodes (v, v1, v2, v3 and v4) connect to enough neighbors (d≥3) and S becomes a quasi-clique.

Supplementary Figure S3

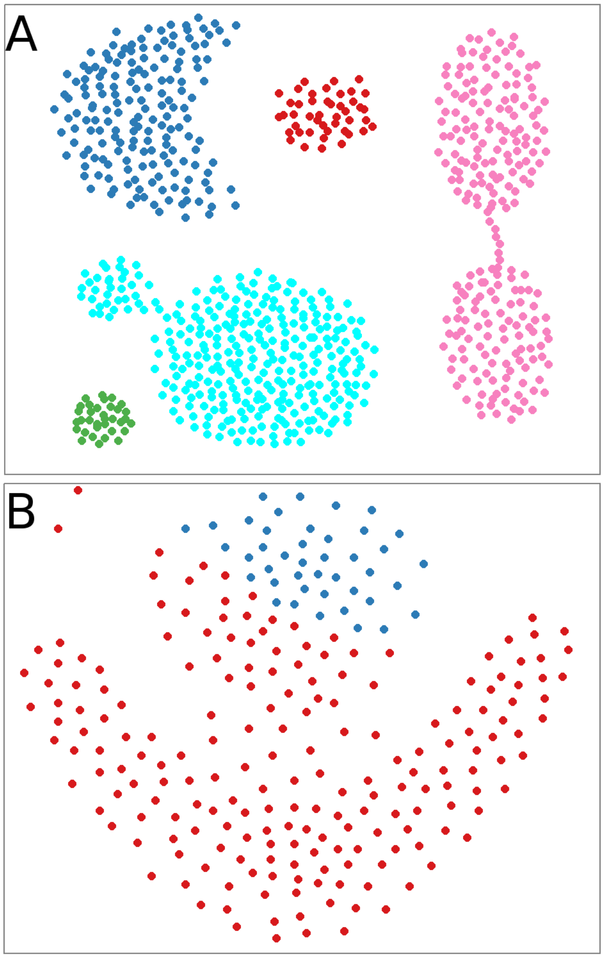
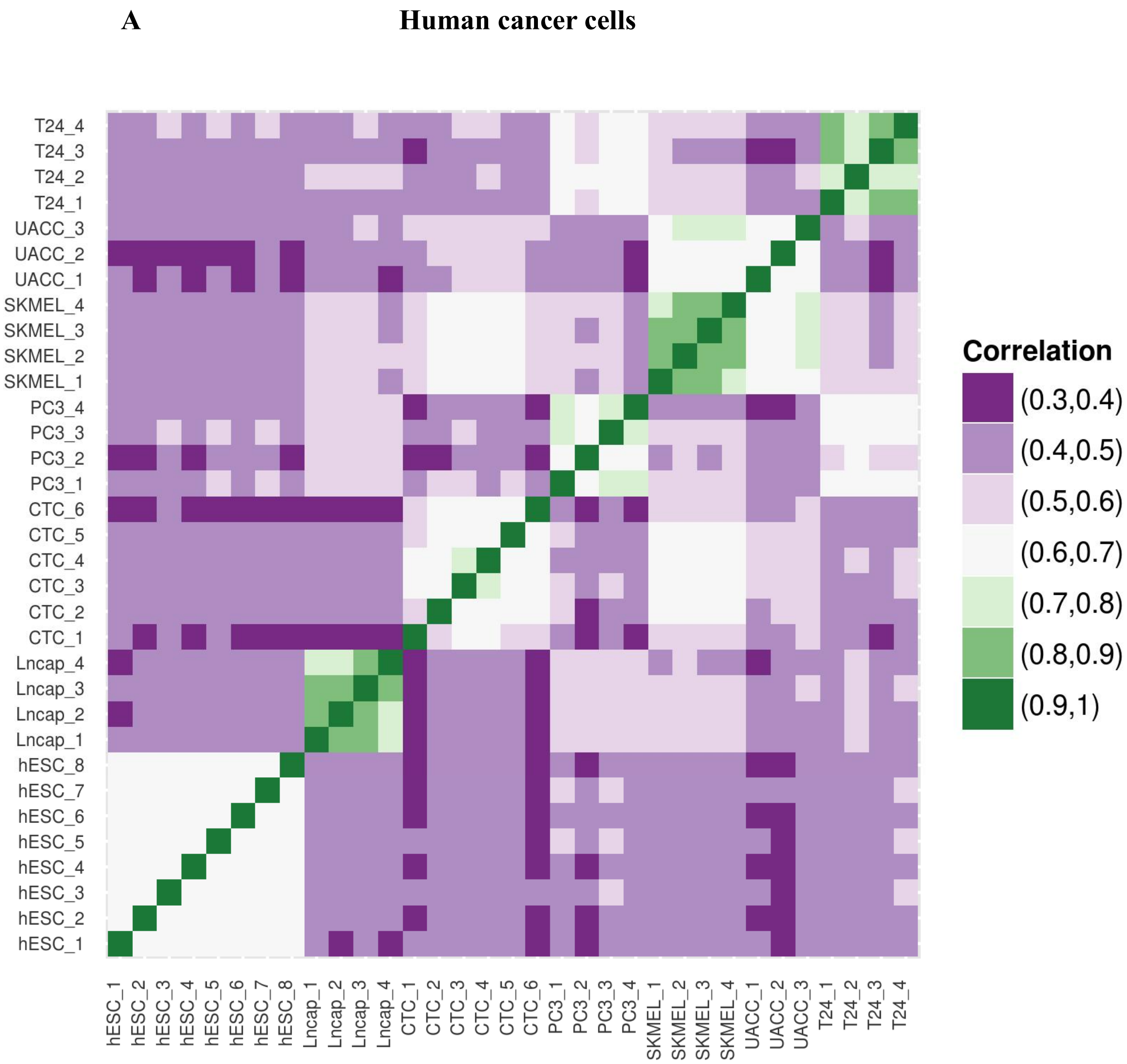
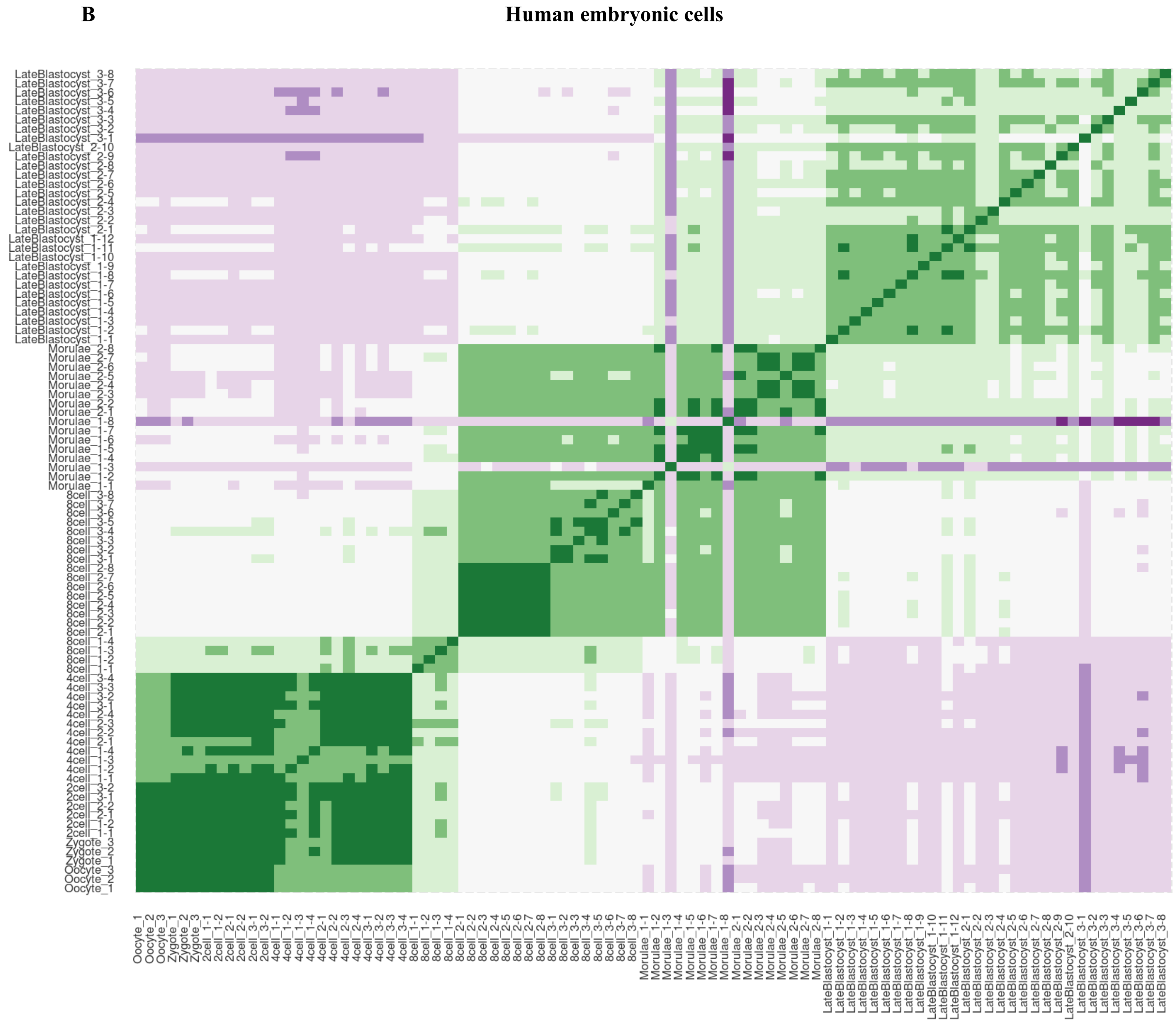


Fig. S3. Performance of HCS on synthetic 2-D datasets (A) (Gionis *et al.*, 2007) and (B) (Fu and Medico, 2007). Points are colored according to the cluster to which they are assigned.

Supplementary Figure S4




B
Human embryonic cells
LateBlastocyst_3-8
LateBlastocyst_3-7
LateBlastocyst_3-6
LateBlastocyst_3-5
LateBlastocyst_3-4
LateBlastocyst_3-3
LateBlastocyst_3-2
LateBlastocyst_3-1
LateBlastocyst_2-10
LateBlastocyst_2-9
LateBlastocyst_2-8
LateBlastocyst_2-7
LateBlastocyst_2-6
LateBlastocyst_2-5
LateBlastocyst_2-4
LateBlastocyst_2-3
LateBlastocyst_2-2
LateBlastocyst_2-1
LateBlastocyst_1-12
LateBlastocyst_1-11
LateBlastocyst_1-10
LateBlastocyst_1-9
LateBlastocyst_1-8
LateBlastocyst_1-7
LateBlastocyst_1-6
LateBlastocyst_1-5
LateBlastocyst_1-4
LateBlastocyst_1-3
LateBlastocyst_1-2
LateBlastocyst_1-1
Morulae_2-8
Morulae_2-7
Morulae_2-6
Morulae_2-5
Morulae_2-4
Morulae_2-3
Morulae_2-2
Morulae_2-1
Morulae_1-8
Morulae_1-7
Morulae_1-6
Morulae_1-5
Morulae_1-4
Morulae_1-3
Morulae_1-2
Morulae_1-1
8cell_3-8
8cell_3-7
8cell_3-6
8cell_3-5
8cell_3-4
8cell_3-3
8cell_3-2
8cell_3-1
8cell_2-8
8cell_2-7
8cell_2-6
8cell_2-5
8cell_2-4
8cell_2-3
8cell_2-2
8cell_2-1
8cell_1-4
8cell_1-3
8cell_1-2
8cell_1-1
4cell_3-4
4cell_3-3
4cell_3-2
4cell_3-1
4cell_2-4
4cell_2-3
4cell_2-2
4cell_2-1
4cell_1-4
4cell_1-3
4cell_1-2
4cell_1-1
2cell_3-2
2cell_3-1
2cell_2-2
2cell_2-1
2cell_1-2
2cell_1-1
Zygote_3
Zygote_2
Zygote_1
Oocyte_3
Oocyte_2
Oocyte_1
Oocyte_1
Oocyte_2
Oocyte_3
Zygote_1
Zygote_2
Zygote_3
2cell_1-1
2cell_1-2
2cell_2-1
2cell_2-2
2cell_3-1
2cell_3-2
4cell_1-1
4cell_1-2
4cell_1-3
4cell_1-4
4cell_2-1
4cell_2-2
4cell_2-3
4cell_2-4
4cell_3-1
4cell_3-2
4cell_3-3
4cell_3-4
8cell_1-1
8cell_1-2
8cell_1-3
8cell_1-4
8cell_2-1
8cell_2-2
8cell_2-3
8cell_2-4
8cell_2-5
8cell_2-6
8cell_2-7
8cell_2-8
8cell_3-1
8cell_3-2
8cell_3-3
8cell_3-4
8cell_3-5
8cell_3-6
8cell_3-7
8cell_3-8
Morulae_1-1
Morulae_1-2
Morulae_1-3
Morulae_1-4
Morulae_1-5
Morulae_1-6
Morulae_1-7
Morulae_1-8
Morulae_2-1
Morulae_2-2
Morulae_2-3
Morulae_2-4
Morulae_2-5
Morulae_2-6
Morulae_2-7
Morulae_2-8
LateBlastocyst_1-1
LateBlastocyst_1-2
LateBlastocyst_1-3
LateBlastocyst_1-4
LateBlastocyst_1-5
LateBlastocyst_1-6
LateBlastocyst_1-7
LateBlastocyst_1-8
LateBlastocyst_1-9
LateBlastocyst_1-10
LateBlastocyst_1-11
LateBlastocyst_1-12
LateBlastocyst_2-1
LateBlastocyst_2-2
LateBlastocyst_2-3
LateBlastocyst_2-4
LateBlastocyst_2-5
LateBlastocyst_2-6
LateBlastocyst_2-7
LateBlastocyst_2-8
LateBlastocyst_2-9
LateBlastocyst_2-10
LateBlastocyst_3-1
LateBlastocyst_3-2
LateBlastocyst_3-3
LateBlastocyst_3-4
LateBlastocyst_3-5
LateBlastocyst_3-6
LateBlastocyst_3-7
LateBlastocyst_3-8

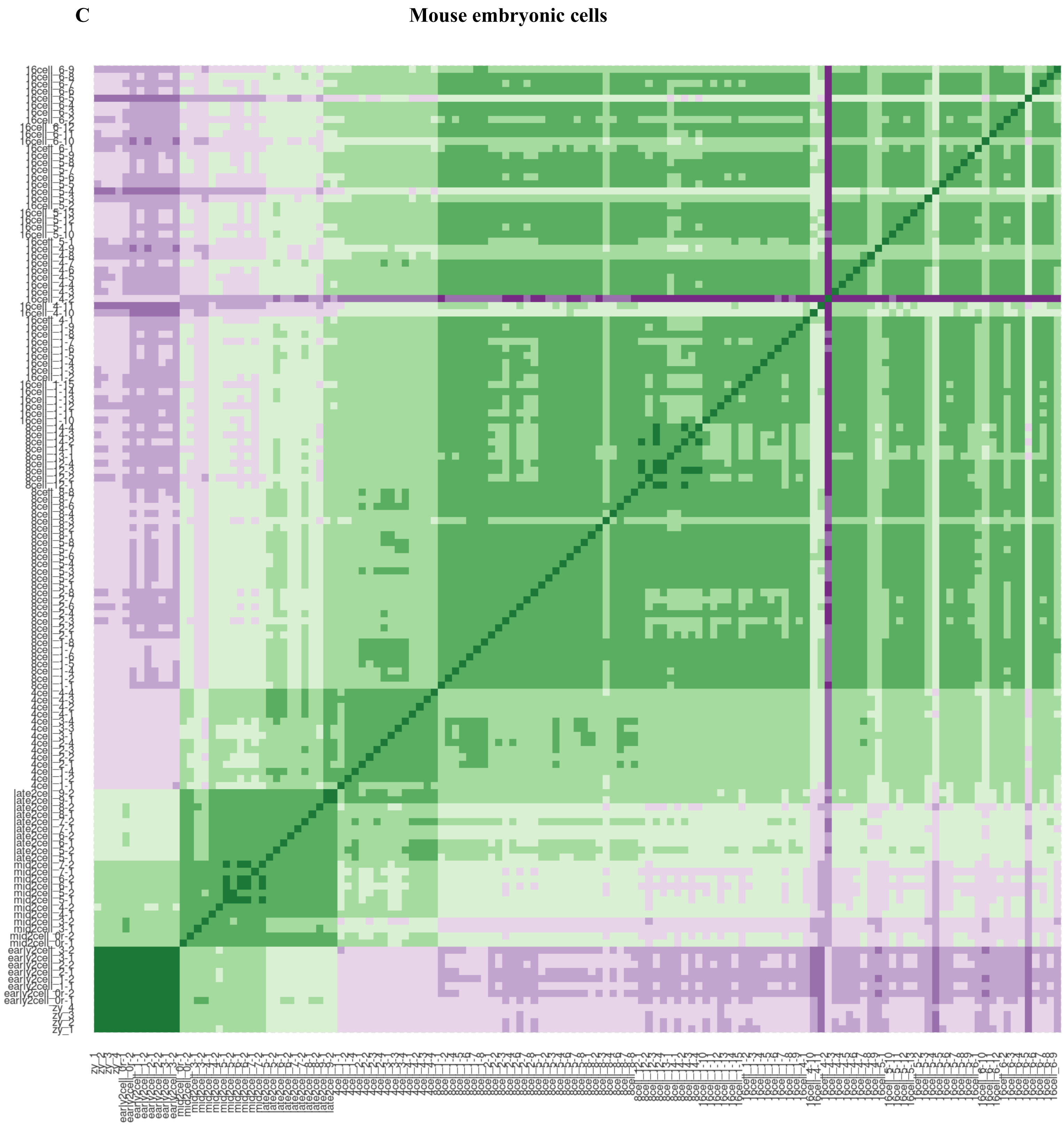


Fig. S4. Variability of single cell transcriptomes. (A-C) Heatmaps show the Pearson's correlation coefficient between gene expression levels in cells, calculated by log transformed RPKMs. Green represents high correlation and purple represents low correlation.

REFERENCES

Ertöz, L., Steinbach, M. and Kumar, V. (2003) Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data, *in Proceddings of Second SIAM Interational Conference on Data Mining*.

Fu,L. and Medico,E. (2007) FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics*, **8**, 3.

Gionis,A. *et al.* (2007) Clustering aggregation. *ACM Trans. Knowl. Discov. Data*, **1**, 4–es.

Guha, S., Rastogi, R. and Shim, K. (2000) Rock: A robust clustering algorithm for categorical attributes, *Information Systems*, **25**, 345-366.

Houle, M.E*., et al.* (2010) Can shared-neighbor distances defeat the curse of dimensionality? In, Gertz,M. and Ludäscher,B. (eds), *Scientific and Statistical Database Management: 22nd International Conference, SSDBM 2010, Heidelberg, Germany, June 30–July 2, 2010. Proceedings*. Springer-Verlag, pp. 482–500.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

Jarvis, R.A. and Patrick, E.A. (1973) Clustering Using a Similarity Measure Based on Shared Near Neighbors, *IEEE Transactions on Computers*, **C-22**, 1025-1034.

Van Rijsbergen,C.J. (1974) FOUNDATION OF EVALUATION. *J. Doc.*, **30**, 365–373.

Zhang, S*., et al.* (2009) Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes., *Nucleic acids research*, **37**, e72.