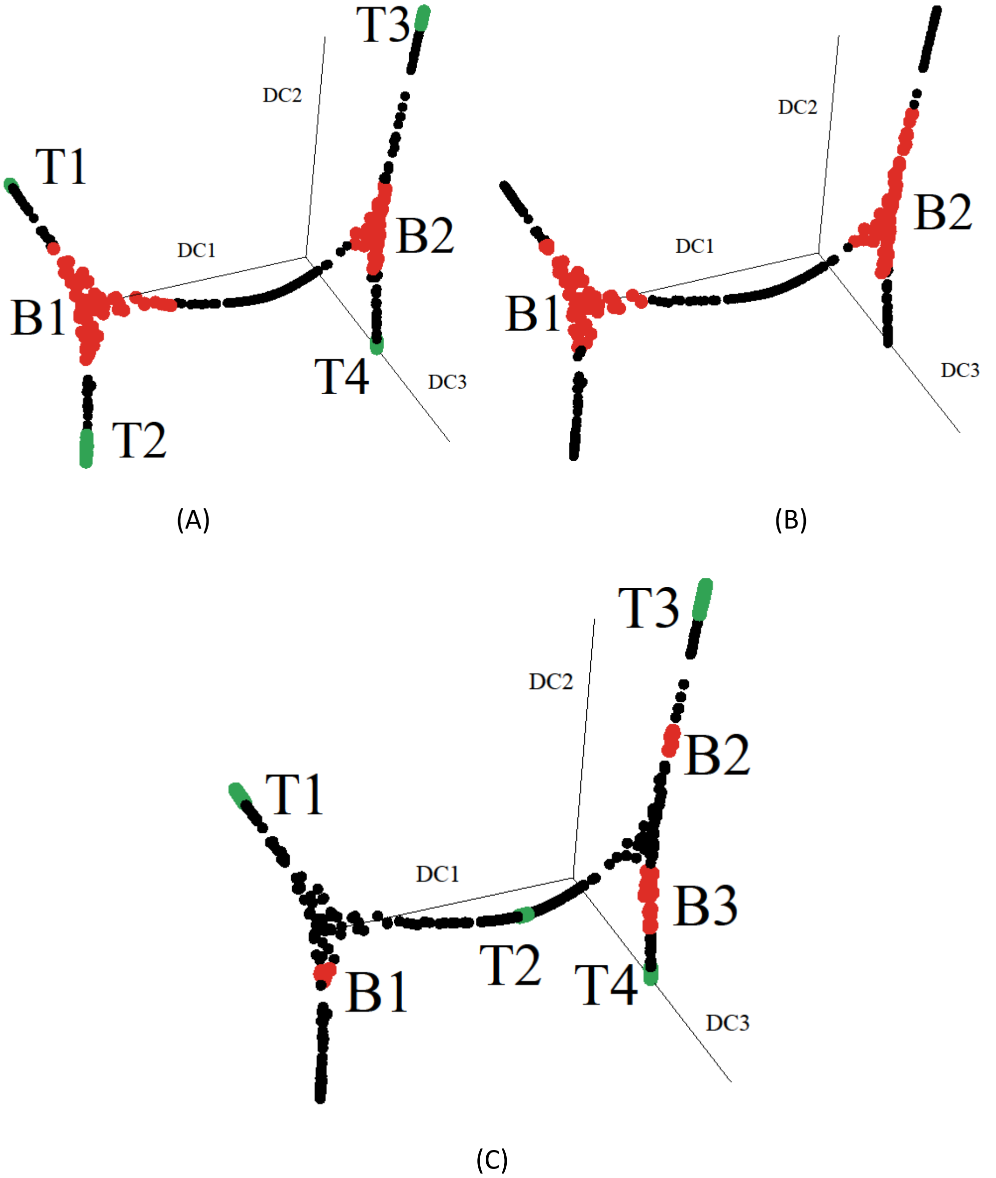
Supplementary Materials for

Model-based branching point detection in single-cell data by K-Branches clustering

Nikolaos K. Chlis [1,2], F. Alexander Wolf [1] and Fabian J. Theis [1,2,3]

[1]Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany
[2]School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany
[3]Department of Mathematics, Technical University of Munich, 85748 Garching, Germany

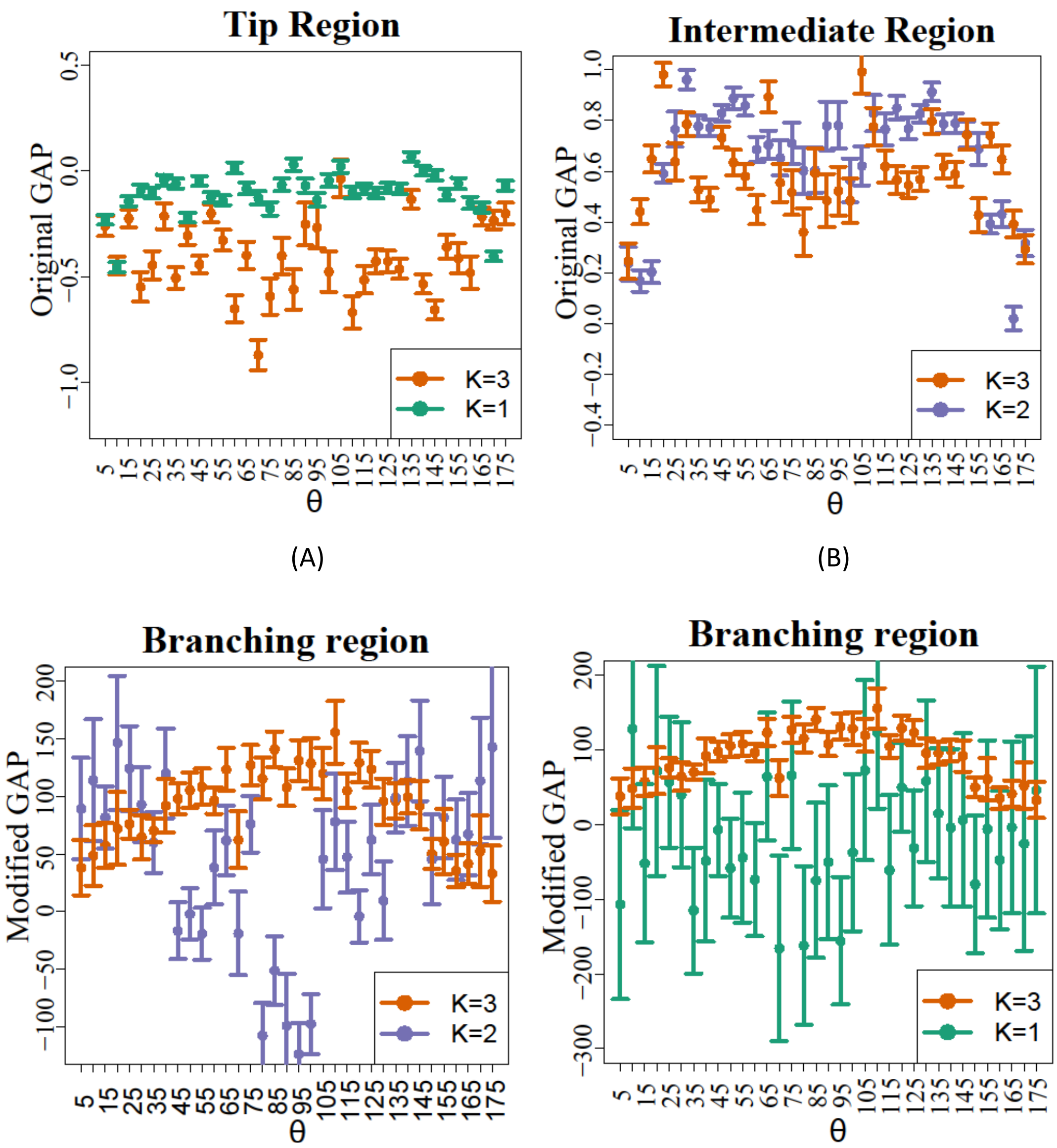


(A)

(B)

(C)

Figure S1: Results of local K-Branches on toy data of differentiation (Haghverdi et al., 2015) for different values of the neighbourhood size S. The exact same three diffusion components were used in all the above examples. (A) For S = 61, as calculated by the proposed heuristic, the four tips (T1, T2, T3, T4, marked in green) and two branching regions (B1, B2, marked in red) are identified successfully. (B) When using a small neighbourhood size of S = 10, the two branching regions are identified but the four tips are not. (C) For a too large value of S = 200, three of the four actual tips are identified. However, the identification of branching regions is not as successful. Moreover, one “artifact” tip is identified (T2).



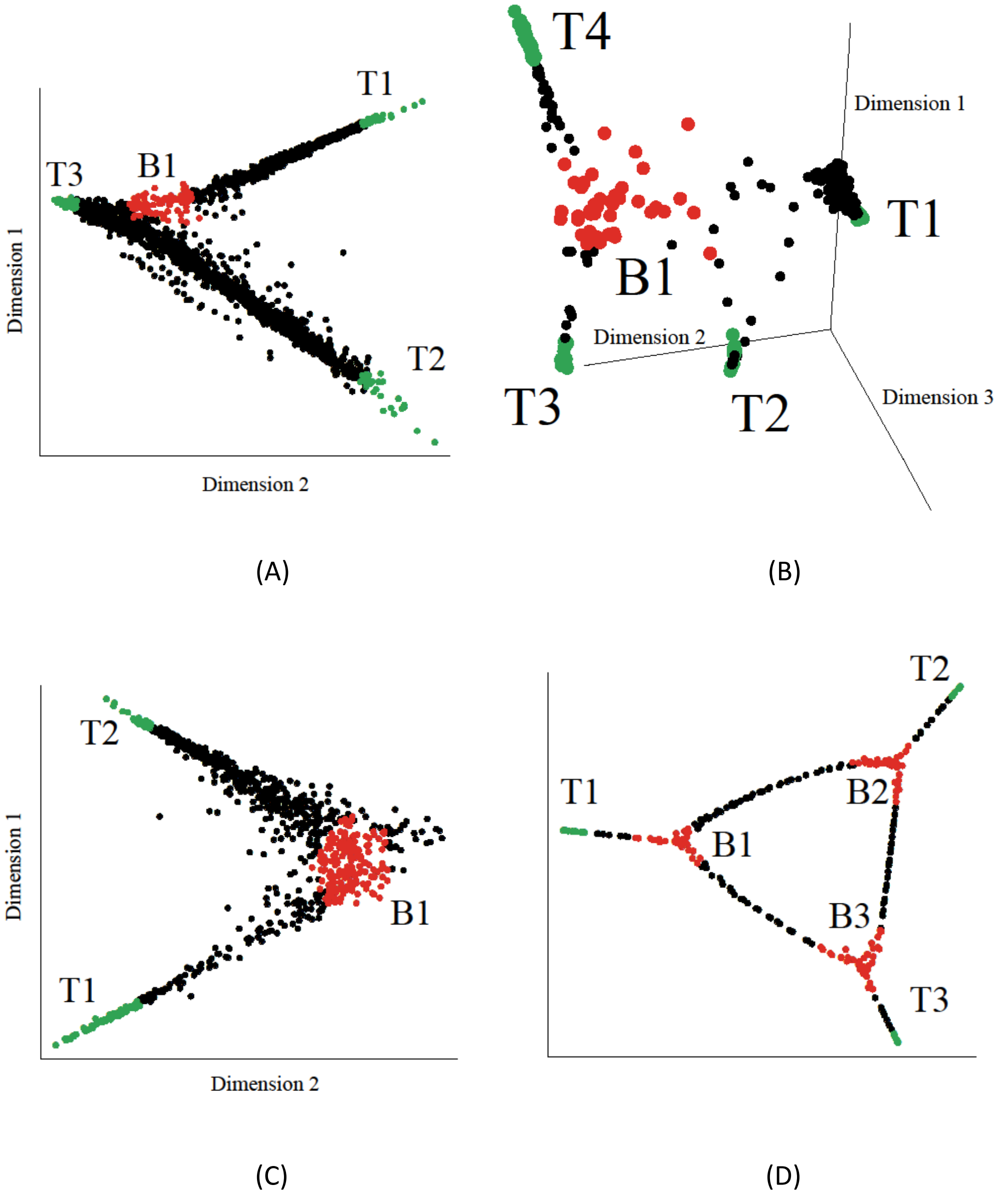
(A)

(B)

(C)

(D)

Figure S2: Repeating the experiment of Figure 2 in the original manuscript, using Gaussian noise with larger variance of σ = 1. (A) The original GAP statistic successfully identifies tip cells in tip regions. (B) The original GAP statistic cannot be used for identification of branching cells, since it would identify a large number of false positive branching cells in intermediate regions. (C) The modified GAP statistic can successfully identify branching region cells. (D) The modified GAP statistic cannot be used for tip cell identification since it would lead to false identification of a large number of branching cells as tip cells.



(A)

(B)

(C)

(D)

Figure S3: Results of local K-Branches using LLE instead of diffusion maps for dimensionality reduction. (A) Results of local K-Branches on the (Paul et al., 2015) data, on dimensions extracted by LLE. Due to the high dimensionality of the dataset, PCA was used as preprocessing and LLE was performed on the first 50 principal components of the data. Local K-Branches successfully identifies three tips (T1, T2, T3, marked in green) and one branching region (B1, marked in red). (B) Results of local K-Branches on the (Guo et al., 2010) data, on dimensions extracted by LLE. Local K-Branches successfully identifies four tips (T1, T2, T3, T4, marked in green). However, the two separate branching regions are merged into one when LLE is used instead of diffusion maps. (C) Results on the (Kouno et al., 2013) data, on dimensions extracted by LLE. Local K branches identifies two tips and one branching region. In contrast, two tips and no branching were identified when diffusion maps were used. (D) Results of local K-Branches on artificial data, on dimensions extracted by LLE. Local K-Branches successfully identifies three tips (T1, T2, T3, marked in green) and three branching regions (B1, B2, B3, marked in red).

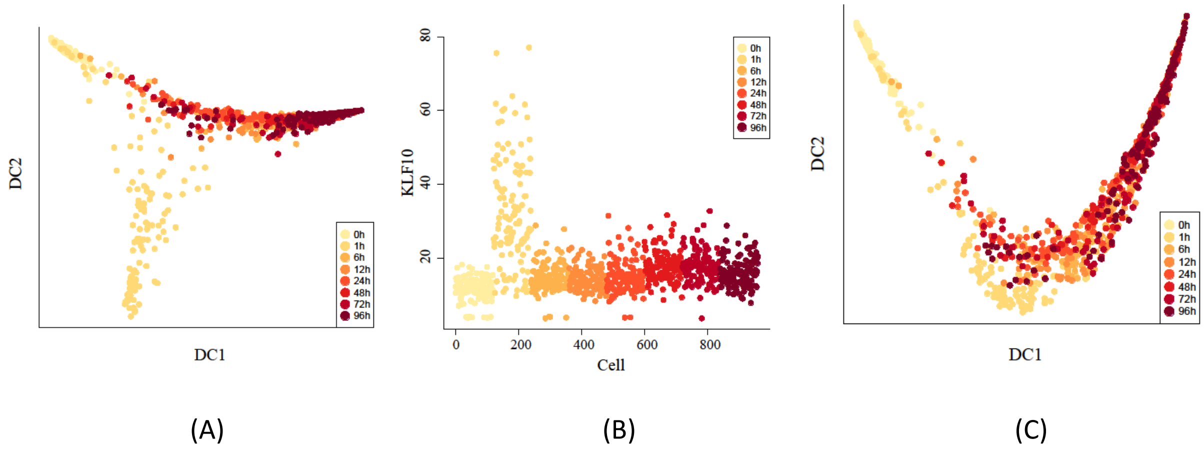


Figure S4: Analysis of multiple time point single-cell qPCR dataset of human myeloid monocytic leukemia cells (Kouno et al., Genome Biology, 14:r118, 2013). The original dataset contains 45 genes and 960 cells. Data were collected at multiple time points, where THP-1 human myeloid monocytic leukemia cells were undergoing differentiation into macrophages. As reported in the original publication the second stage of "Early response" cells (1h) seems significantly different than all other stages, which is reflected as branching in the diffusion map representation (A). However, no branching should be present in the dataset. As such, we have identified that it is an artifact of high level of KLF10, which is only expressed in the Early response cells (B). We removed KLF10 and recalculated the diffusion map, which now accurately depicts the differentiation of "Native" cells (0h) to macrophages (96h) without branching. If KLF10 was not removed, the performance of local K-Branches, DPT and TSCAN was as follows: Local K-Branches achieved 0.44 precision and 0.17 recall, while DPT achieved 0.67 precision and 0.008 recall. Finally, TSCAN achieved 0.57 precision and 0.31 recall.

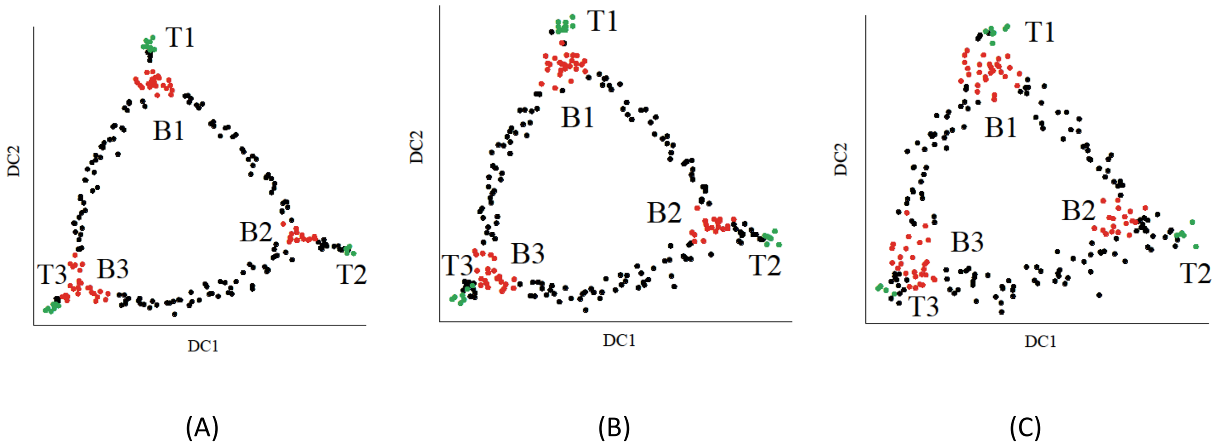


Figure S5: Results of local K-Branches on the toy data containing a loop for different levels of added noise. Three tips and three branching regions are identified in all settings. In all cases the added noise is zero-mean Gaussian with varying standard deviation σ. (A) σ = 0.002 (B) σ = 0.004 (C) σ = 0.008.