Supplementary for SCCNAInfer: a robust and accurate tool to infer the absolute copy number on scDNA-seq data

Liting Zhang¹, Xin Maizie Zhou² and Xian Mallory¹*

¹ Florida State University, Tallahassee, FL 32304, USA
xfan2@fsu.edu
² Vanderbilt University, Nashville, TN 37235, USA

1 Simulation of Spike-in data

This simulation followed a generative model that generated the absolute copy number for the cells for each bin first, and then simulated the read count according to the absolute copy number.

In more detail, we started with cells whose copy number is 2 genome-wide, and added the CNAs in silico to these 100 cells throughout the whole genome.

To mimic the subclonality that the real data has, we added the subclonality to the dataset by dividing the cells evenly to multiple clusters as the default setting, whereas the number of clusters was a number that we varied to test SCCNAInfer’s robustness to different datasets. Here we set the number of clusters to be 1, 2, 4 and 6 whereas 4 was the default number of clusters. For these cluster of cells, if there was more than one cluster, one was designated as the normal cell subclone and the rest were designated as aneuploid subclones.

For the normal cell subclones, we did not add any CNAs to them. For the aneuploid subclones, we first added 10 CNAs that were shared by all cells to mimic the clonal CNAs. Such CNAs were randomly sampled from the whole genome such that each CNA spanned 100-120 bins whereas the bin width was 500kb. An integer was randomly sampled between 1 and 5, to be assigned as the absolute copy number for each CNA.

We then added 30-40 subclonal CNAs to each aneuploid subclone, whereas the number of bins each CNA may span was randomly sampled between 100 to 120 like those clonal CNAs. The number of subclonal CNAs added to each aneuploid subclone was randomly sampled from 30 to 40. These numbers were selected based on our observation of the real datasets. Since in the real dataset, the ploidy of each subclone may differ from each other, we mimicked this process by setting a baseline ploidy for each aneuploid subclone, upon which we added these CNAs. In more detail, for each aneuploid subclone, we randomly sampled a number from 1.5 to 4.5 with 0.1 as the step size. The sampled number was designated as the baseline ploidy for this subclone. We then determined each CNA by randomly sampling a number from 0 to the baseline as the deviation from the baseline. We designated that each CNA has 60% of the chance being an amplification, and 40% of the chance being a deletion. The resulting absolute
copy number was rounded to make sure that it was an integer, and such an integer number served as the ground truth for evaluation.

Since the size of single-cell data is expanding, we varied the number of cells to evaluate SCCNAInfer's robustness to different cell counts. We tested with 50, 100, 500, and 1000 cells, where 100 cells was the default for comparison.

Next, we added noises to the dataset to mimic the sequencing coverage fluctuation. We did this by adding a number to the ground truth absolute copy number for each segment at each cell, whereas each segment in a cell had 50% of the chance to have such added noise. The number added was sampled from a Gaussian distribution whose mean is 0, and standard deviation was a variable that we varied for testing SCCNAInfer's robustness. Here the standard deviations that we used were 0.5, 0.8 and 1 whereas 0.8 was set as the default standard deviation.

Given the noisy absolute copy number profile for each cell, we generated the corresponding read count as follows. Suppose $CN_{ij}$ is the simulated noisy copy number for cell $i$ at bin $j$. For each subclone, we randomly selected a normal cell from (Navin et al., 2011) such that the normal cell’s read count will serve as the reference for all the cells in this subclone. Let $R_j$ represent bin $j$’s read count of this normal cell. The resulting read count for bin $j$ at cell $i$ is $R_j \times CN_{ij}/2$. We then scale the read count of each cell such that the total read count of each cell equals the total read count of the selected normal cell. This is to avoid the inflation of the number of reads resulting from the increase of the ploidy. The final read count of the cells from each subclone served as the input to SCCNAInfer.

Since in the real dataset, the number of the cells each subclone differs from each other, given all the default parameters mentioned above, we added two different levels of skewness to the distribution of the cells to each subclone. Specifically, we sampled the number of cells for the four clusters from a Dirichlet distribution whose alpha was [10, 10, 10, 10] for less skewness, and [15, 5, 3, 2] for more skewness.
2 Supplementary Figures

Fig. S1: Performance Assessment of SCCNAInfer via Spike-In Analysis based on the Read Count of a Normal Cell from KTN302 for A. Varied Number of Cluster, and B. Varied Number of Cells. Mean squared errors of the inferred CNAs from Ginkgo, SCOPE and SeCNV are shown in red violins. SCCNAInfer’s results with the application to each of these three methods’ segmentation are shown in blue next to each method, respectively.
Fig. S2: Performance Assessment of SCCNAInfer via Spike-In Analysis based on the read count of a normal cell from KTN302 for **A**, Varied Level of Noise, and **B**, Varied Skewness of Cell Distribution. Mean squared errors of the inferred CNAs from Ginkgo, SCOPE and SeCNV are shown in red violins. SCCNAInfer’s results with the application to each of these three methods’ segmentation are shown in blue next to each method, respectively.
Fig. S3: Heatmap of the CNA call of KTN302 from Aneufinder is shown in the top panel. Heatmap of the CNA call from SCCNAInfer that was applied to Aneufinder’s segmentation is shown in the bottom panel. Each row represents the CNA call of one cell. The left vertical color bar shows the subclone that the cell belongs to. The color legend of the CNA call and the subclone that each cell belongs to are both shown on the bottom.
Fig. S4: Heatmap of the CNA call of KTN302 from Ginkgo is shown in the top panel. Heatmap of the CNA call from SCCNAInfer that was applied to Ginkgo’s segmentation is shown in the bottom panel. Each row represents the CNA call of one cell. The left vertical color bar shows the subclone that the cell belongs to. The color legend of the CNA call and the subclone that each cell belongs to are both shown on the bottom.
Fig. S5: Heatmap of the CNA call of KTN302 from SeCNV is shown in the top panel. Heatmap of the CNA call from SCCNAInfer that was applied to SeCNV’s segmentation is shown in the bottom panel. Each row represents the CNA call of one cell. The left vertical color bar shows the subclone that the cell belongs to. The color legend of the CNA call and the subclone that each cell belongs to are both shown on the bottom.
Fig. S6: Heatmap of the CNA call of KTN129 from Aneufinder is shown in the top panel. Heatmap of the CNA call from SCCNAInfer that was applied to Aneufinder’s segmentation is shown in the bottom panel. Each row represents the CNA call of one cell. The left vertical color bar shows the subclone that the cell belongs to. The color legend of the CNA call and the subclone that each cell belongs to are both shown on the bottom.
Fig. S7: Heatmap of the CNA call of KTN129 from Ginkgo is shown in the top panel. Heatmap of the CNA call from SCCNAInfer that was applied to Ginkgo’s segmentation is shown in the bottom panel. Each row represents the CNA call of one cell. The left vertical color bar shows the subclone that the cell belongs to. The color legend of the CNA call and the subclone that each cell belongs to are both shown on the bottom.
Fig. S8: Heatmap of the CNA call of KTN129 from SCOPE is shown in the top panel. Heatmap of the CNA call from SCCNAInfer that was applied to SCOPE’s segmentation is shown in the bottom panel. Each row represents the CNA call of one cell. The left vertical color bar shows the subclone that the cell belongs to. The color legend of the CNA call and the subclone that each cell belongs to are both shown on the bottom.
Fig. S9: Heatmap of the CNA call of KTN129 from SeCNV is shown in the top panel. Heatmap of the CNA call from SCCNAInfer that was applied to SeCNV’s segmentation is shown in the bottom panel. Each row represents the CNA call of one cell. The left vertical color bar shows the subclone that the cell belongs to. The color legend of the CNA call and the subclone that each cell belongs to are both shown on the bottom.