Benchmarking digital PCR partition classification methods with empirical and simulated duplex data

Contents

1 Glossary

Table S1: Glossary: overview of frequently used digital PCR terminology, adapted from Table 1 in [2].

term	**description**	**alternative name**
baseline	the fluorescence of the negative partitions	fluorescence noise
channel	part of the light spectrum used to detect signal, typically annotated as emission wave length	color, detection channel, emission channel
cluster	group of partitions that display similar fluorescence intensities	(partition) population
crosstalk	the signal or fluorescence from one channel is mistakenly seen in the wrong place, often indicated by non-orthogonality	
duplex	dPCR reaction in which at least 2 different targets are quantified using 2 different detection channels	
fluorescence intensity	the fluorescence of a partition	fluorescence amplitude, end-point fluorescence, relative fluorescence unit
higher-order multiplexing	dPCR reaction in which more targets are quantified than the number of fluorescent detection channels	intensity multiplexing
lower-order clusters	simple clusters that contain only a single target	
negative control	complex biological specimen that does not contain the target	
negative population	partition group that contains no target	negative cluster
no-template control	sample that contains no targets and is used as a general control for extraneous nucleic acid contamination or non-specific amplification	ntc, blanco, blank sample
partition	the subreaction used for limiting dilution and subsequently measured as positive or negative post reaction	droplet, chamber
positive control	sample that contains target and is used to test if the assay is performing correctly	
positive population	partition group that contains one or more targets. In (non-higher-order) multiplex assays, there can be as many as $2^n - 1$ positive clusters, where n is the number of targets. They can be single positive, double positive, triple, etc.	positive cluster
rain	the partitions that are located within the space between the positive and negative clusters	
resolution	a measure of the separation in fluorescence between positive and negative partitions	peak resolution, separability score
singleplex	assay used to detect one target sequence	
threshold	the line that separates the partition clusters based on fluorescence intensity	

2 Package version

Table S2: Package session information

package	**version**
stats	4.2.2
e1071	1.7-12
dbscan	1.1-11
dpcp	retrieved from `https://github.com/alfodefalco/dPCP` on January 15th, 2023
flowSOM	2.6.0
flowPeaks	1.44.0
flowClust	3.36.0
flowMerge	2.46.0
SamSPECTRAL	1.52.0
calico	retrieved from `https://github.com/billytcl/calico` on January 15th, 2023
ddPCRclust	1.18.0
sn	2.1.0
spatstat	3.0.2
DepthProc	2.1.5

3 Data generation mechanism

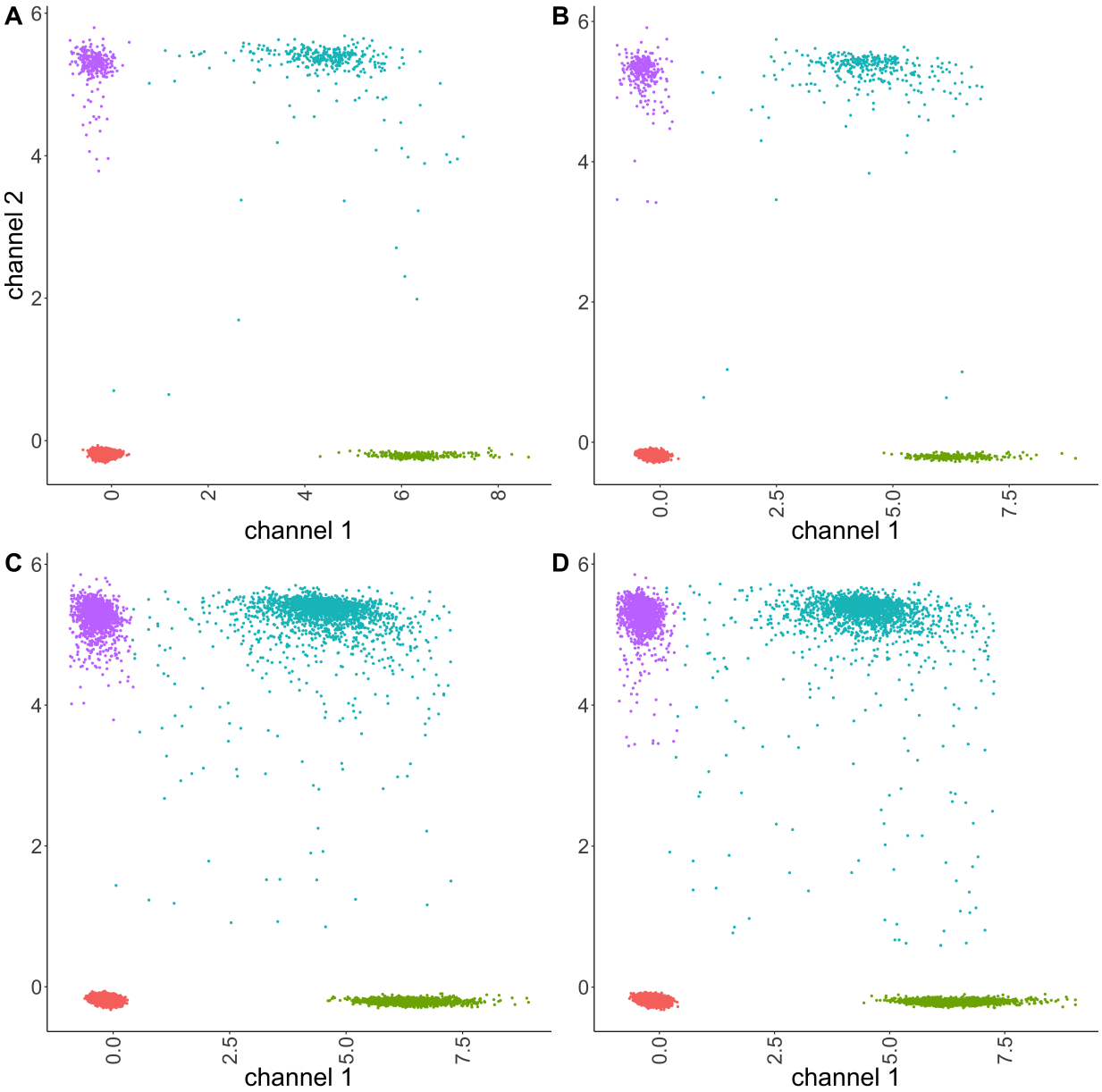

Figure S1: (A) HR dataset without removal of datapoints towards the negative population (B) simulated data using the method described in section Probabilistic model (C) simulated data with no constraints imposed and density 6 times of the original dataset (D) simulated data with constraints imposed and density 6 times of the original dataset. Note: the density is increased to make the contrast resulting from the constraint more visible.

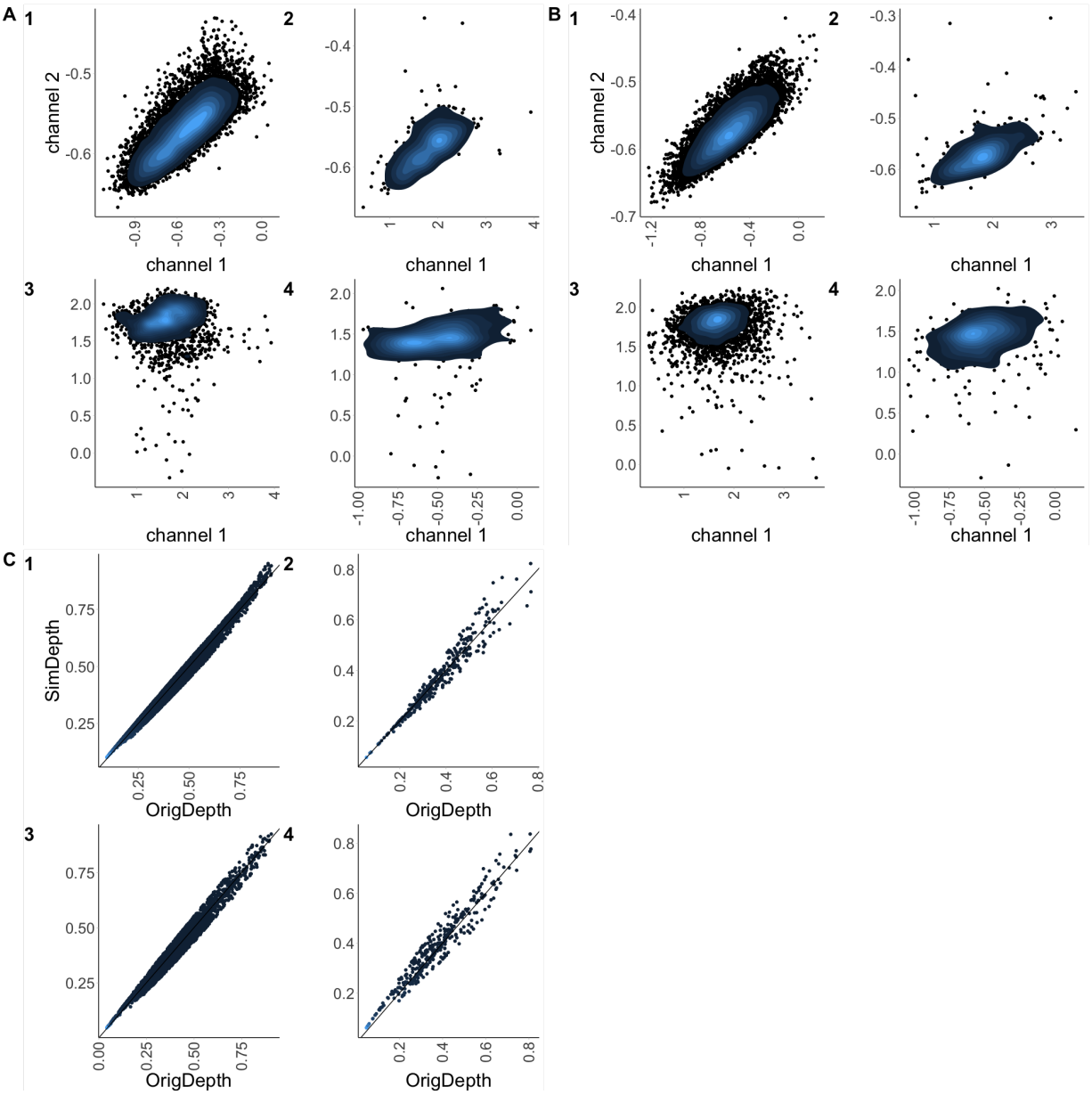

Figure S2: Goodness-of-fit for the MM dataset (A)-(B) Density plots of the original and simulated data respectively and (C) depth-depth plots. 1,2,3,4 represent the negative population, single positive 1, double positive and single positive 2 respectively. The further the data points from the cluster center, the darker those data points are. Depth values indicate how deep a data point is within the data distribution, with more central or typical points having higher depths and outliers having lower depths [1]

.

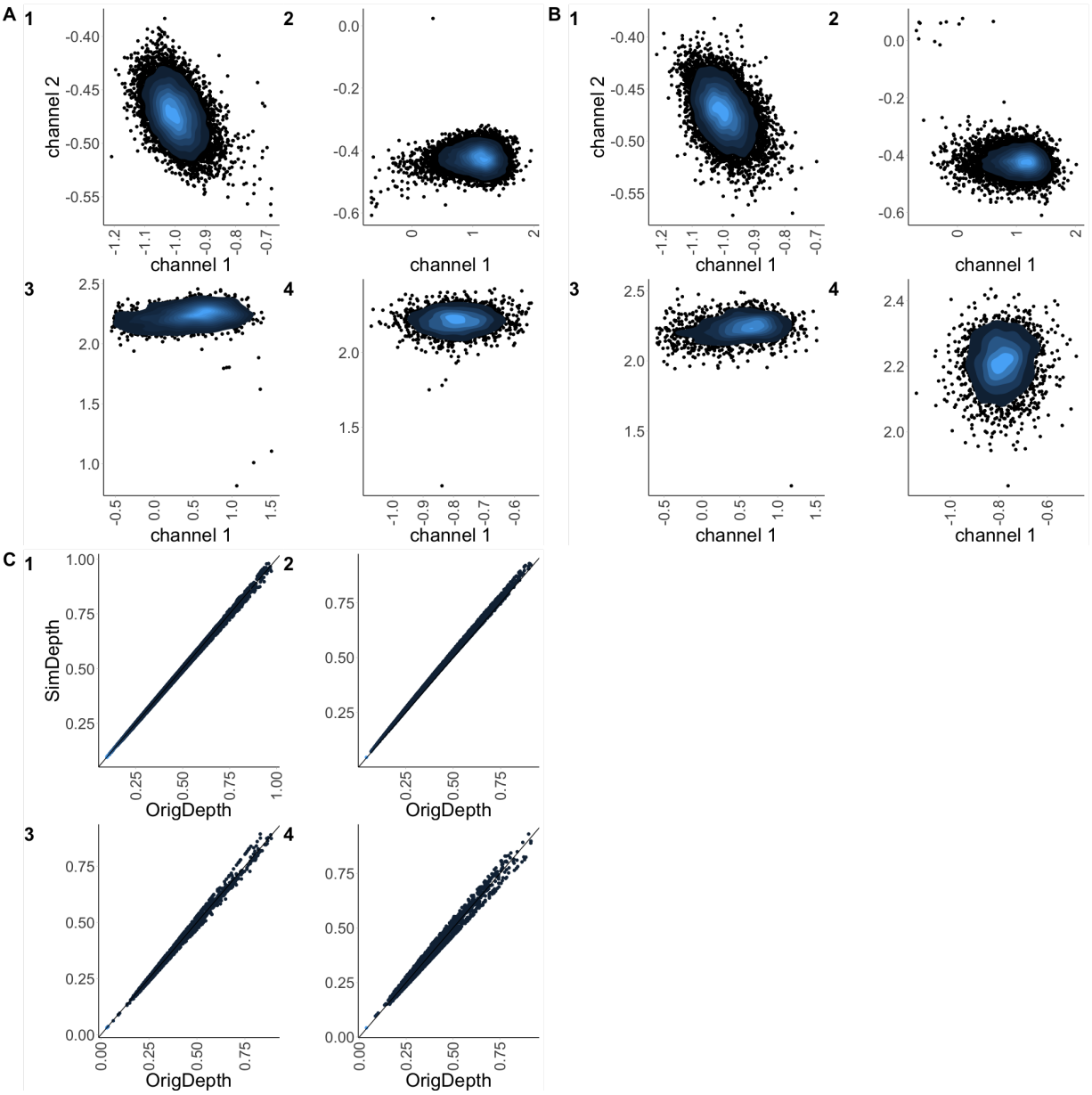

Figure S3: Goodness-of-fit for the LR dataset (A)-(B) Density plots of the original and simulated data respectively and (C) depth-depth plots. 1,2,3,4 represent the negative population, single positive 1, double positive and single positive 2 respectively. The further the data points from the cluster center, the darker those data points are. Depth values indicate how deep a data point is within the data distribution, with more central or typical points having higher depths and outliers having lower depths [1]

.

4 Resolution measurement

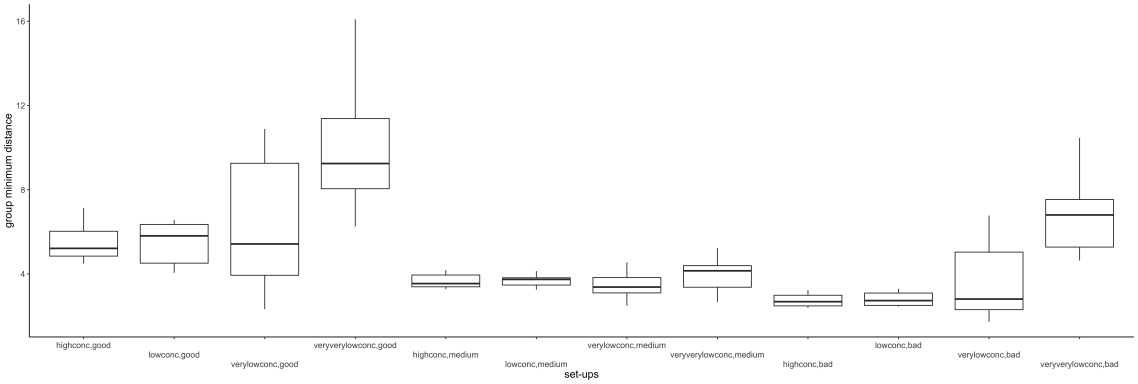

Figure S4: resolution measurement. High concentration to rare concentration refer to different concentration levels. Good, medium and poor correspond to the datasets in Fig.1.

5 Examples of simulated data

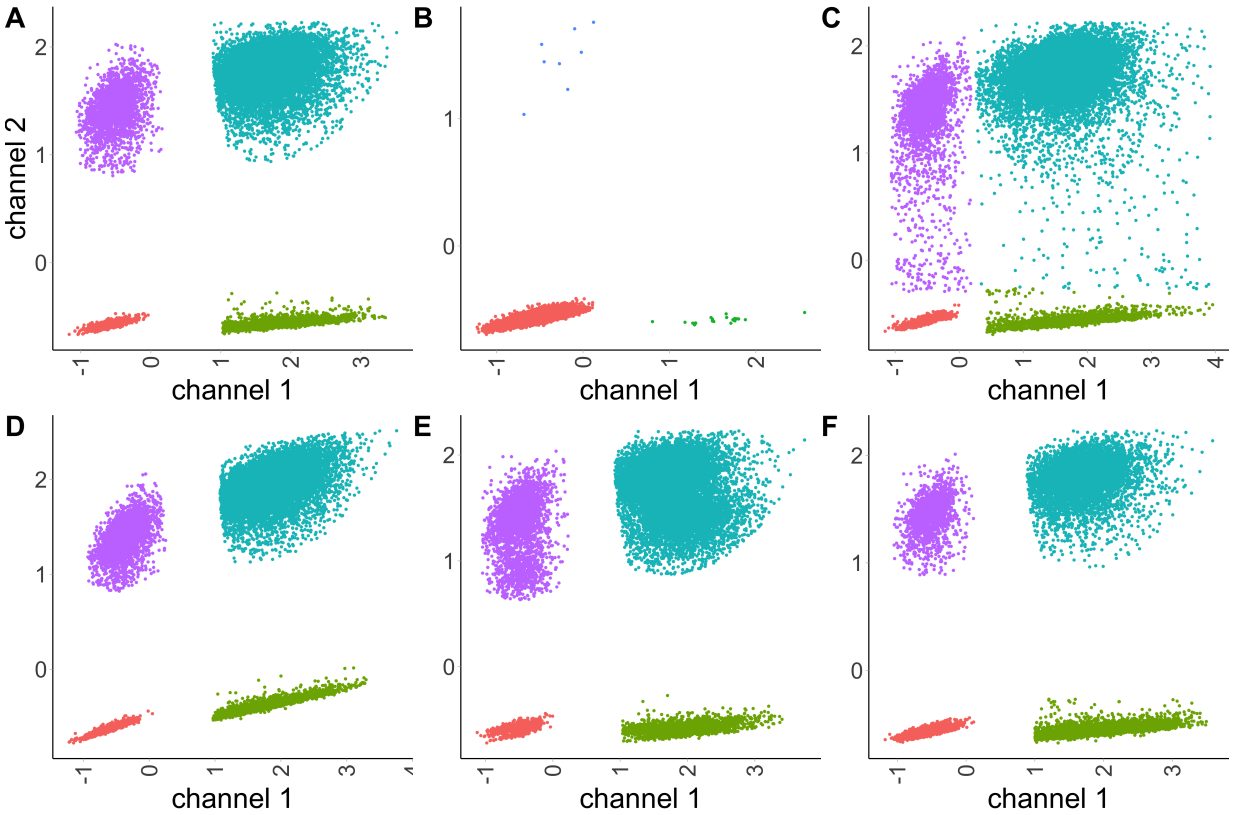

Figure S5: (A) high concentration with no rain (B) rare concentration (C) high concentration with high percentage of rain (D) high concentration with no rain and non-orthogonal assay (E) high concentration with no rain and two modes with small overlapping (F) high concentration with unequal cluster sizes

6 Parameter optimization

The details about the selected optimal parameters are in Table S3. In the following sections, we also showed the best clustering results with the parameters found via default values, manual and automatic search (see Fig.S6 to S16).

method	key parameters	default values	manual/instructed search			automatic search silhouette coefficients		
			HR dataset	MM dataset	LR dataset	HR dataset	MM dataset	LR dataset
DBSCAN	eps; minPts	eps=0.15; minPts=5	eps=1; minPts=20	eps=0.1; minPts=5	eps=1; minPts=20	eps=0.932; minPts=24	eps=0.567; minPts=9	eps=0.945; minPts=100
ddPCRclust	numOfMarkers; sensitivity	numOfMarkers is fixed at 2; sensitivity=1	sensitivity=1.5	sensitivity=1	sensitivity=2	sensitivity=1.357	sensitivity=0.203	sensitivity=1.723
dpcp	eps; minPts in db_combination	eps=0.15; minPts=5	eps=1; minPts=20	eps=0.1; minPts=5	eps=0.1; minPts=100	eps=0.743; minPts=57	eps=0.218; minPts=18	eps=0.062; minPts=72
flowSOM	grid size of the self-organizing map	xdim=10; ydim=10	xdim=5; ydim=10	xdim=5; ydim=5	xdim=5; ydim=10	xdim=5; ydim=24	xdim=11; ydim=20	xdim=11; ydim=20
flowPeaks	tol; h0;h	tol=0.1;h0=1; h=1.5	tol=0.1; h0=10; h=1.5	tol=0.1; h0=1; h=150	tol=0.1; h0=1; h=150	tol=0.491; h0=23.856; h=149.931	tol=0.644; h0=63.152; h=107.306	tol=0.914; h0=6.464; h=2.222
flowClust	initialization of cluster centroids	/	flowClust2Prior function			/	/	/
SamSPECTRAL	normal.sigmas; separation.factor	normal.sigmas =200; separation.factor =0.7	normal.sigmas =200; separation.factor =0.5	normal.sigmas =200; separation.factor =0.5	normal.sigmas =200; separation.factor =0.5	normal.sigmas =124.840; separation.factor =0.260	normal.sigmas =999.930; separation.factor =0.100	normal.sigmas =10.032; separation.factor =0.338
calico	raster_size	raster_size=480	raster_size=200	raster_size=200	raster_size=200	raster_size=200	raster_size=390	raster_size=465

	automatic search adjusted rand index			selected parameters	
HR dataset	MM dataset	LR dataset	HR dataset	MM dataset	LR dataset
eps=0.778; minPts=25	eps=0.184; minPts=23	eps=0.103; minPts=60	eps=1; minPts=20	eps=0.184; minPts=23	eps=0.103; minPts=60
sensitivity=1.596	sensitivity=0.199	sensitivity=0.514	sensitivity=1.5	sensitivity=1	sensitivity=0.514
eps=0.951; minPts=53	eps=0.066; minPts=31	eps=0.129; minPts=100	eps=1; minPts=20	eps=0.1; minPts=5	eps=0.1; minPts=100
xdim=15; ydim=18	xdim=24; ydim=12	xdim=6; ydim=17	xdim=5; ydim=10	xdim=5; ydim=5	xdim=5; ydim=10
tol=0.858; h0=2.905 h=6.270	tol=0; h0=4.669; h=115.785	tol=0.258; h0=0.631; h=19.231	tol=0.1; h0=10; h=1.5	tol=0.1; h0=1; h=150	tol=0.258; h0=0.631; h=19.231
/	/	/	Prior knowledge about the centroids did not really help. We will go with the default setting.		
normal.sigmas =87.938; separation.factor =0.315	normal.sigma =146.133; separation.factor =0.104	normal.sigma =89.688; separation.factor =0.450	normal.sigmas =200; separation.factor =0.7	normal.sigma =146.133; separation.factor =0.104	normal.sigma =89.688; separation.factor =0.450
raster_size=200	raster_size=306	raster_size=354	raster_size=200	raster_size=200	raster_size=200

Table S3: Optimal parameters given by default values, manual search and automatic search with silhouette coefficients or adjusted rand index

DBSCAN

For the HR dataset, the manual search yielded the optimal result (fig.S6B), with the corresponding optimal parameters detailed in Table... For the MM and LR datasets, the best result was achieved through automatic search using the adjusted Rand index (fig.S6H and L). Note that for the datasets MM and LR, the automatic search with silhouette coefficients again selected parameters resulting in only one or two clusters.

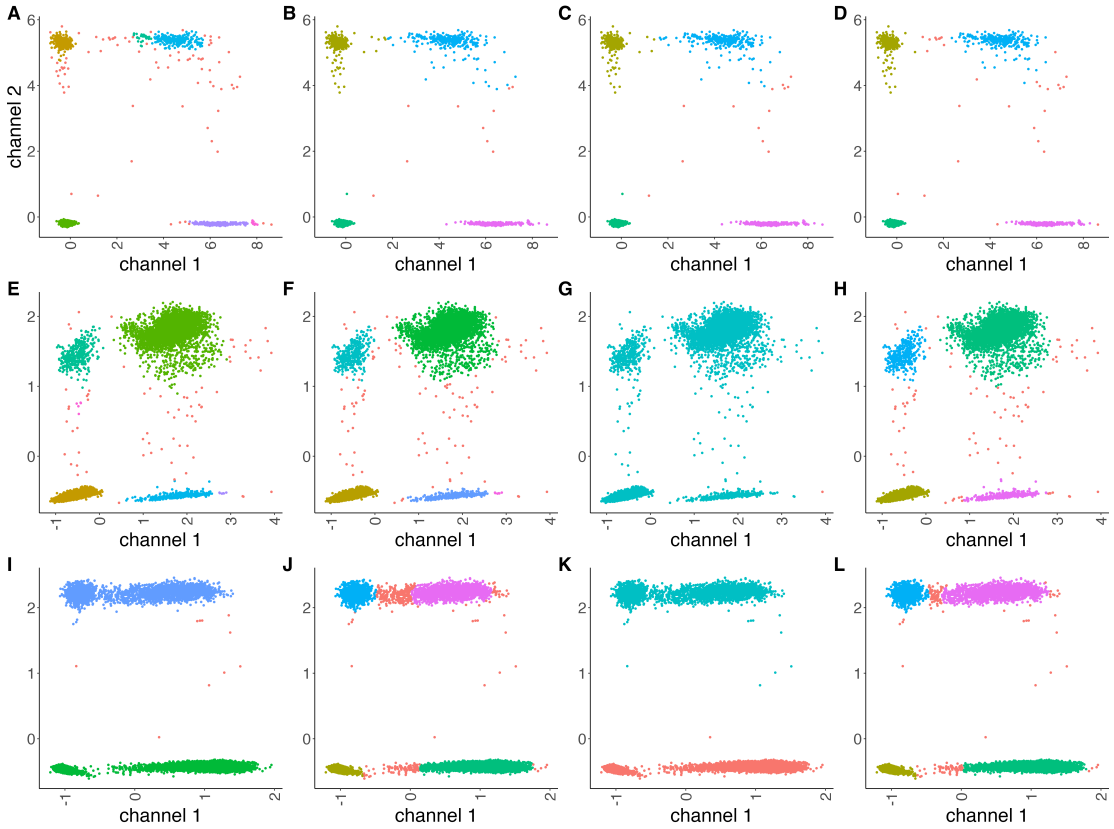

Figure S6: The best clustering results obtained using DBSCAN via different methods. Rows represent distinct datasets ranging from HR to MM and LR. Columns represent various methods, including the default setting, manual search, and automatic search with silhouette coefficients or adjusted Rand index, respectively.

ddPCRclust

For HR and LR datasets, all methods produced clustering results that were quite similar (see Fig.S7A,B,C,I,J and K). However, the automatic search with adjusted Rand index slightly outperformed other methods. In the case of the MM dataset, all methods yielded incorrect clustering outcomes, identifying only two clusters (Fig.S7E,F and G). While we initially attempted these parameters on simulated data, they worked well in some scenarios but generating errors (produce only only one cluster) in others. Consequently, we reverted to the default settings.

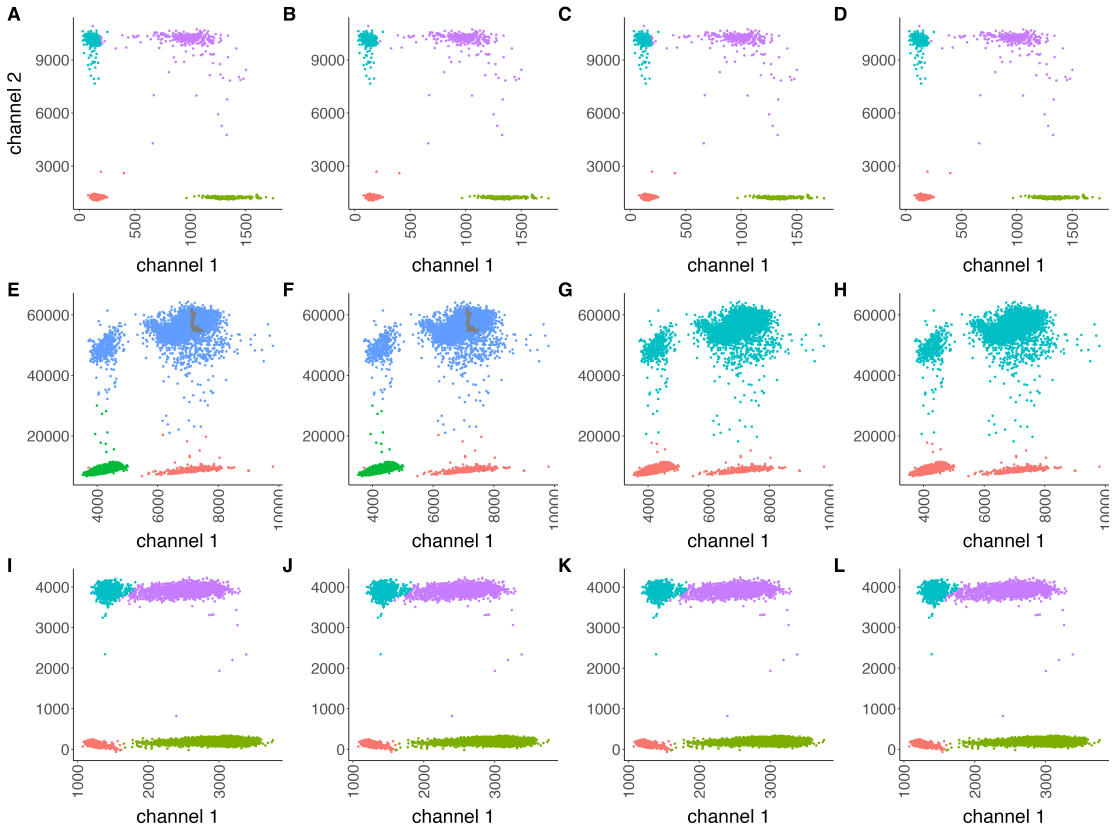

Figure S7: The best clustering results obtained using ddPCRclust via different methods. Rows represent distinct datasets ranging from HR to MM and LR. Columns represent various clustering methods, including the default setting, manual search, and automatic search with silhouette coefficients or adjusted Rand index, respectively.

With the default parameters, ddpcrclust demonstrated functionality across a broader range of scenarios. Nevertheless, it still encountered errors in certain situations. Notably, when applied to the simulated data of the MM dataset, testing was halted due to memory limitations in some scenarios (we ran simulation on high-performance computing). As a result, we made the decision to exclude ddpcrclust from the simulation comparison.

dpcp

dpcp is a two-step approach. It starts with DBSCAN to identify the primary clusters. Then those cluster centroids will be used for the initialization of c-means in the second step. The key parameters are 'eps' and 'minPts' in the first step.

We used 'dbscan_combination' function to identify the optimal input parameters. We selected those parameters with which the primary clusters are distinct (see Fig.S8,S9 and S10).

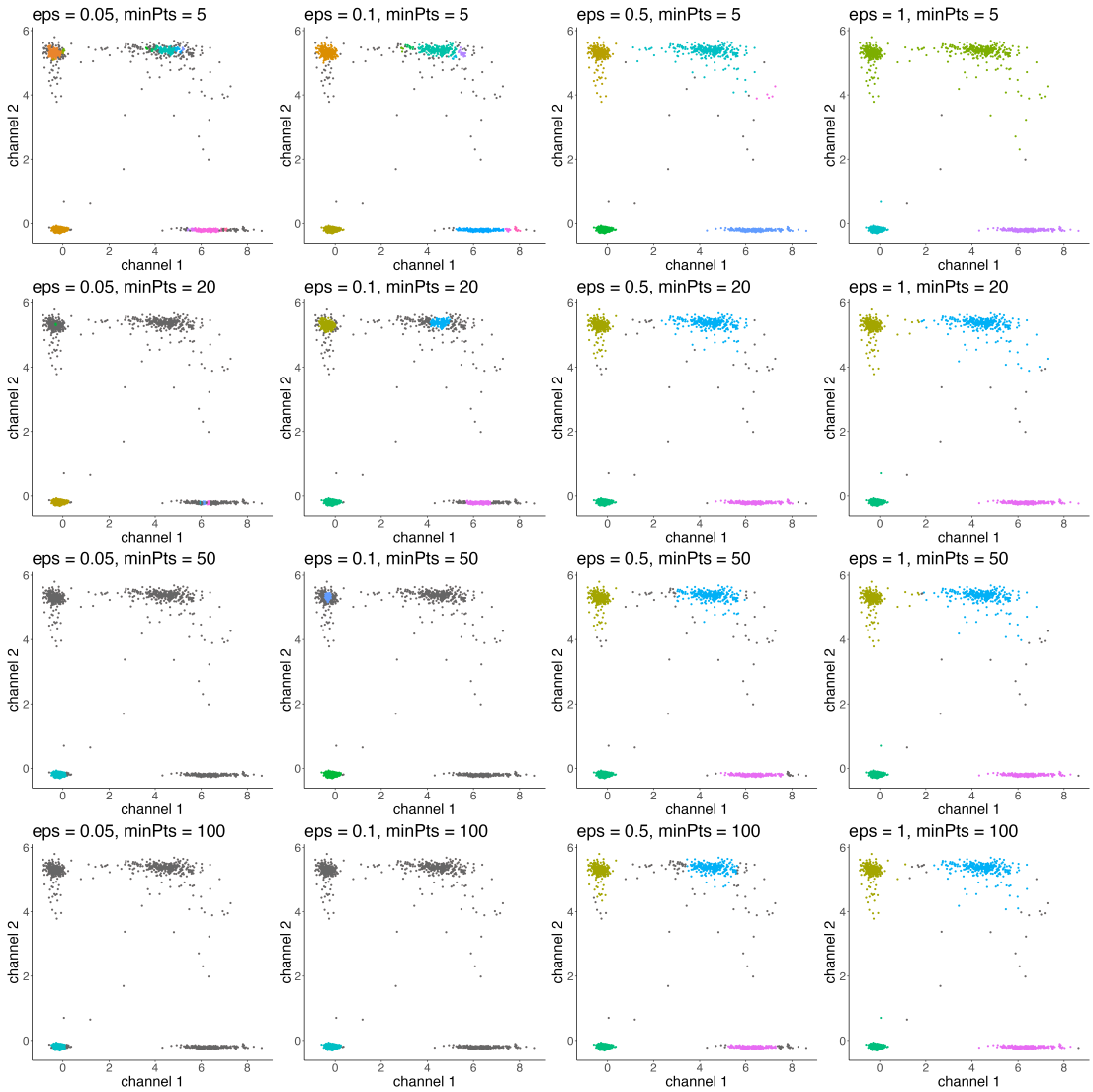

Figure S8: DBSCAN clustering results with varying 'eps' and 'minPts' for HR dataset

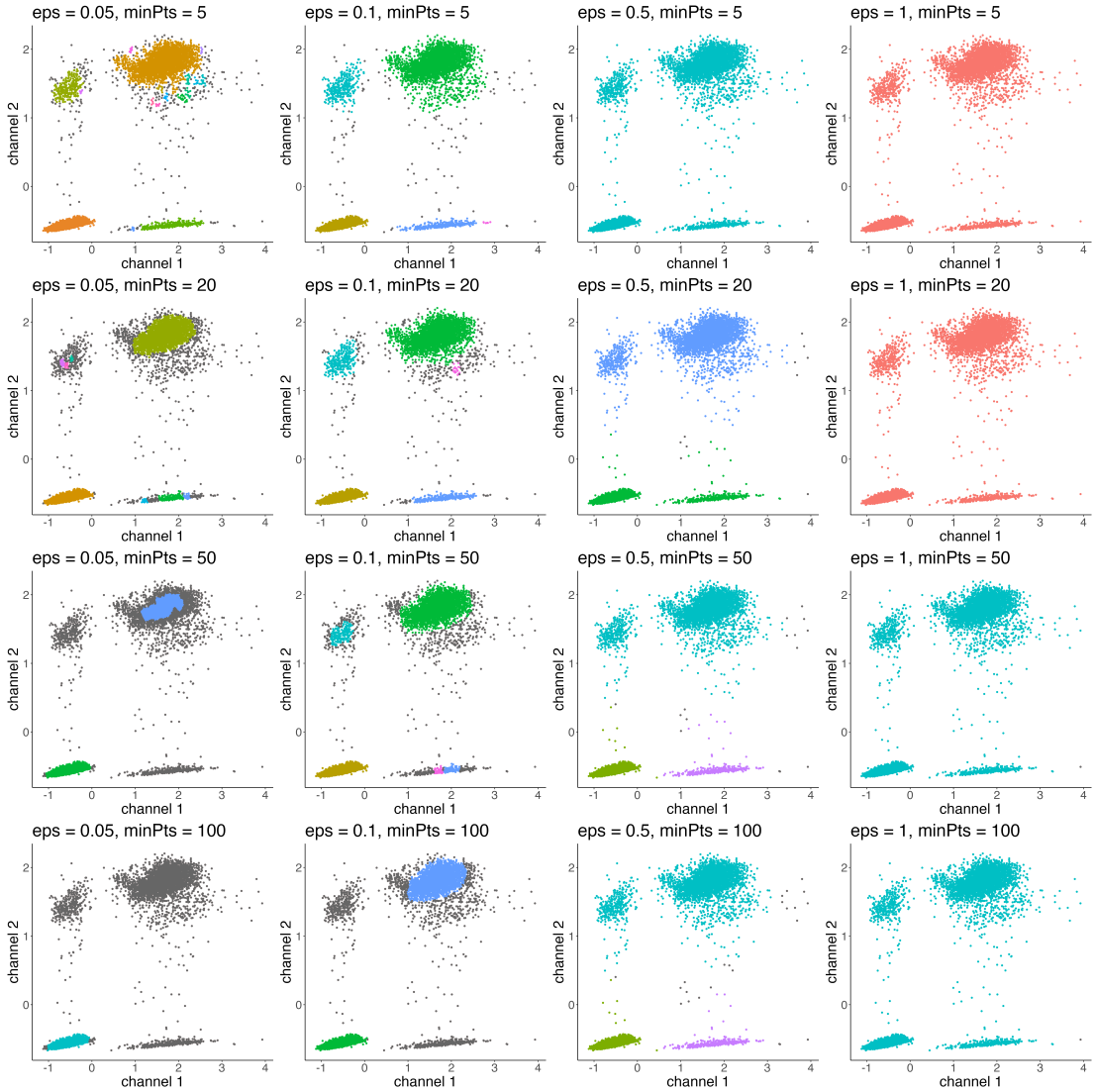

Figure S9: DBSCAN clustering results with varying 'eps' and 'minPts' for MM dataset

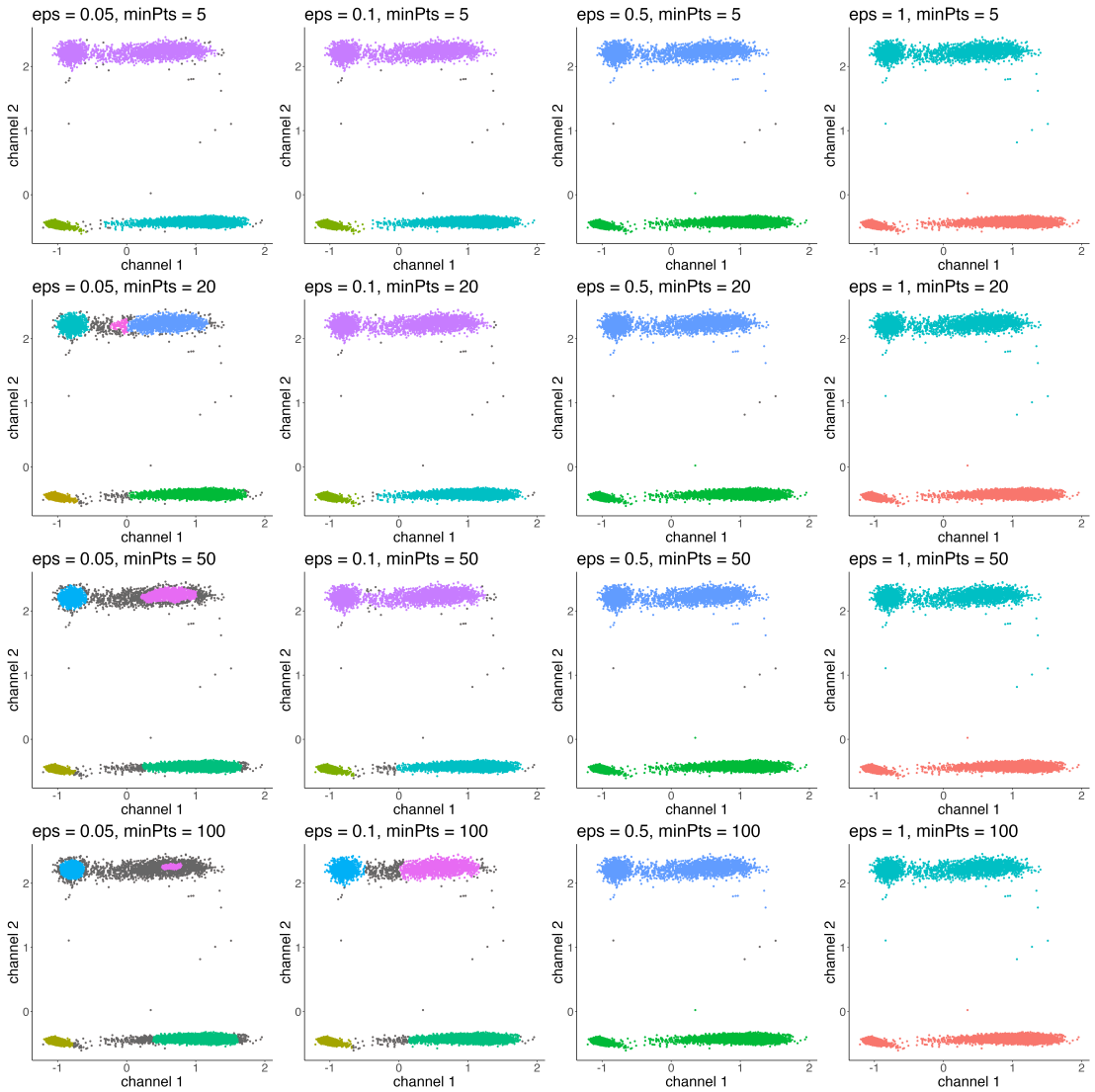

Figure S10: DBSCAN clustering results with varying 'eps' and 'minPts' for LR dataset

We also performed the automatic search and compared the results with default setting and dbscan_combination (see Fig.S11). The results show that for HR and MM datasets, all methods were the same. For the LR dataset, the default setting did not work because the primary clusters cannot be correctly identified. All other 3 methods gave the same results. In this case, we go with the parameters found with dbscan_combination.

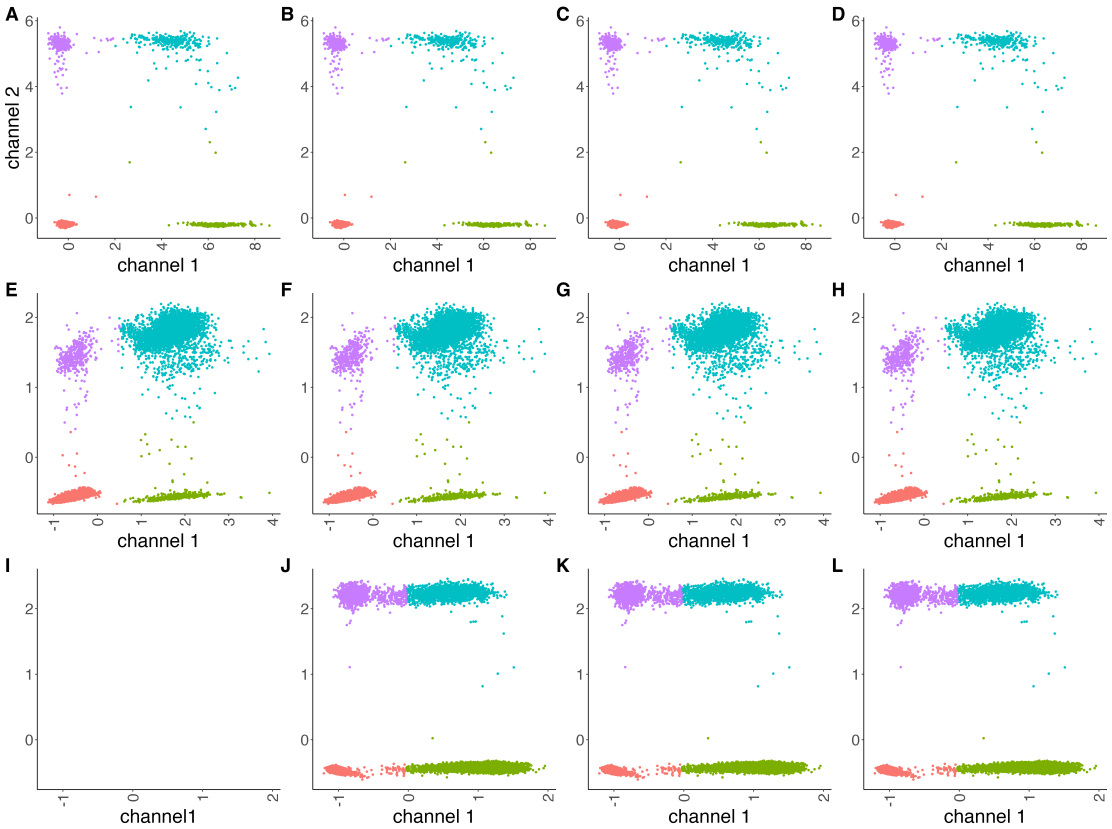

Figure S11: The best clustering results obtained using dpcp via different methods. Rows represent distinct datasets ranging from HR to MM and LR. Columns represent various clustering methods, including the default setting, dbscan_combination search, and automatic search with silhouette coefficients or adjusted Rand index, respectively. Note that with the default setting, the primary clusters are not distinct for LR dataset. dpcp thus gave an error.

flowSOM

For all three datasets, manual search gave the best clustering results (see Fig.S12).

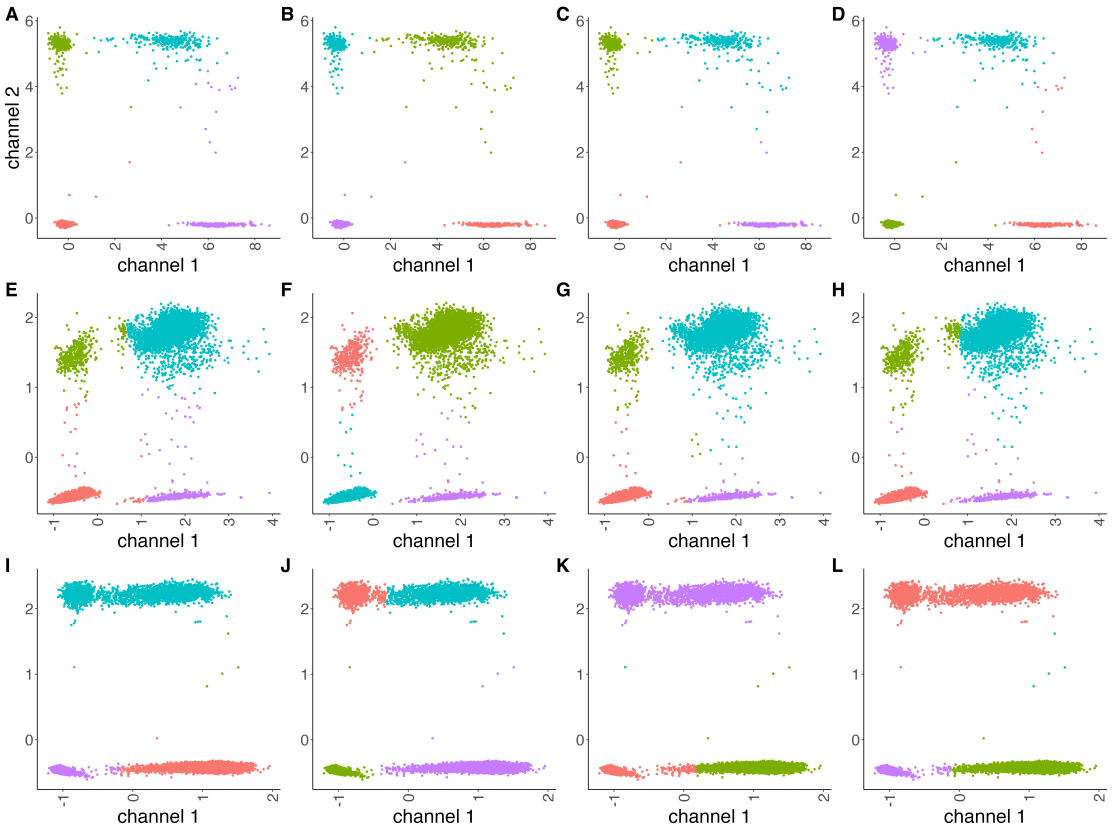

Figure S12: The best clustering results obtained using flowSOM via different methods. Rows represent distinct datasets ranging from HR to MM and LR. Columns represent various clustering methods, including the default setting, manual search, and automatic search with silhouette coefficients or adjusted Rand index, respectively.

flowpeaks

For HR and MM datasets, the manual search yielded the optimal result (fig.S13B and F). For the LR dataset, the best result was achieved through automatic search using the adjusted Rand index (fig.S13L). Note that for the dataset MM, the automatic search with silhouette coefficients selected parameters resulting in only two clusters. This highlights a potential challenge when relying solely on automatic parameter search without prior knowledge of the true grouping.

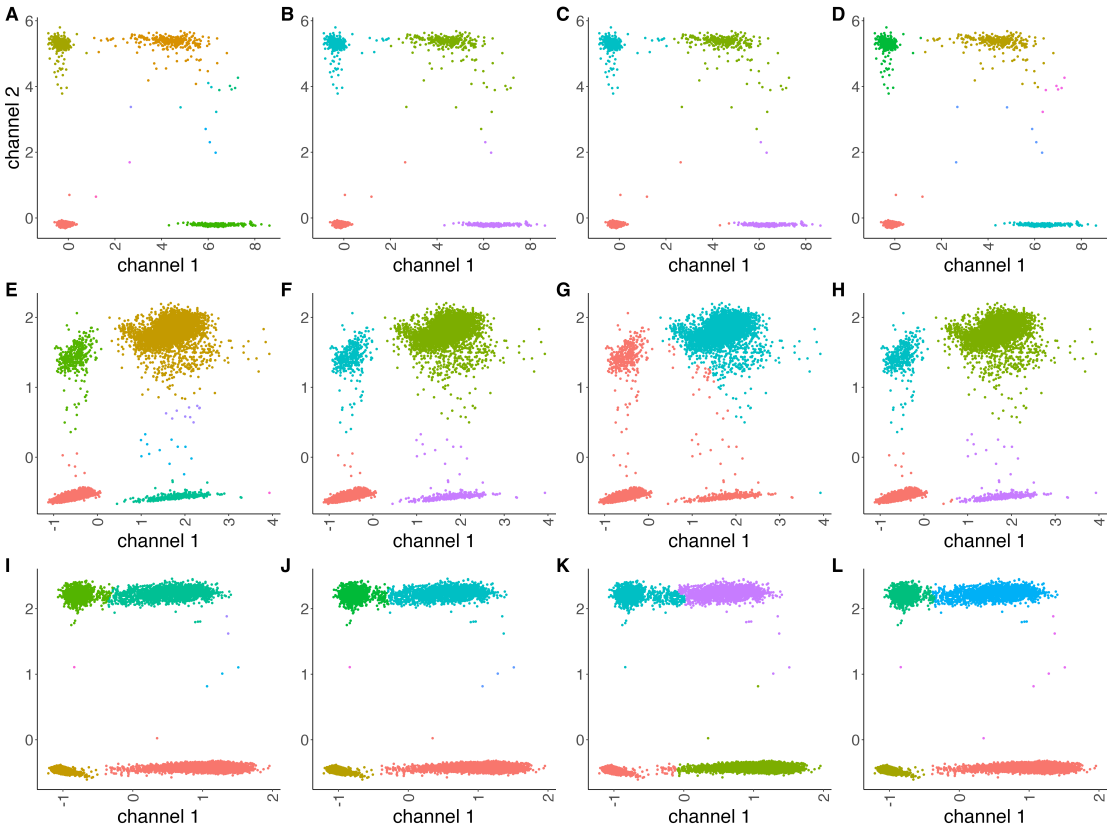

Figure S13: The best clustering results obtained using flowPeaks via different methods. Rows represent distinct datasets ranging from HR to MM and LR. Columns represent various clustering methods, including the default setting, manual search, and automatic search with silhouette coefficients or adjusted Rand index, respectively.

flowclust and flowmerge

The key parameter is the initialization of the cluster centroids. For this method, we used the 'flowClust2Prior' function in the 'flowStats' R package, which generates a prior specification based on a flowClust fit object and can be passed to a second round of flowClust() with usePrior="yes".

To our surprise, the result shows that feeding the prior of the centroids does not help with clustering (Fig.S14). Even the cluster number is sometimes incorrect (3 instead of 4 clusters found). Based on those results, we chose to stay with the default setting and did not feed any prior for flowClust and flowmerge which is based on the results of flowClust model.

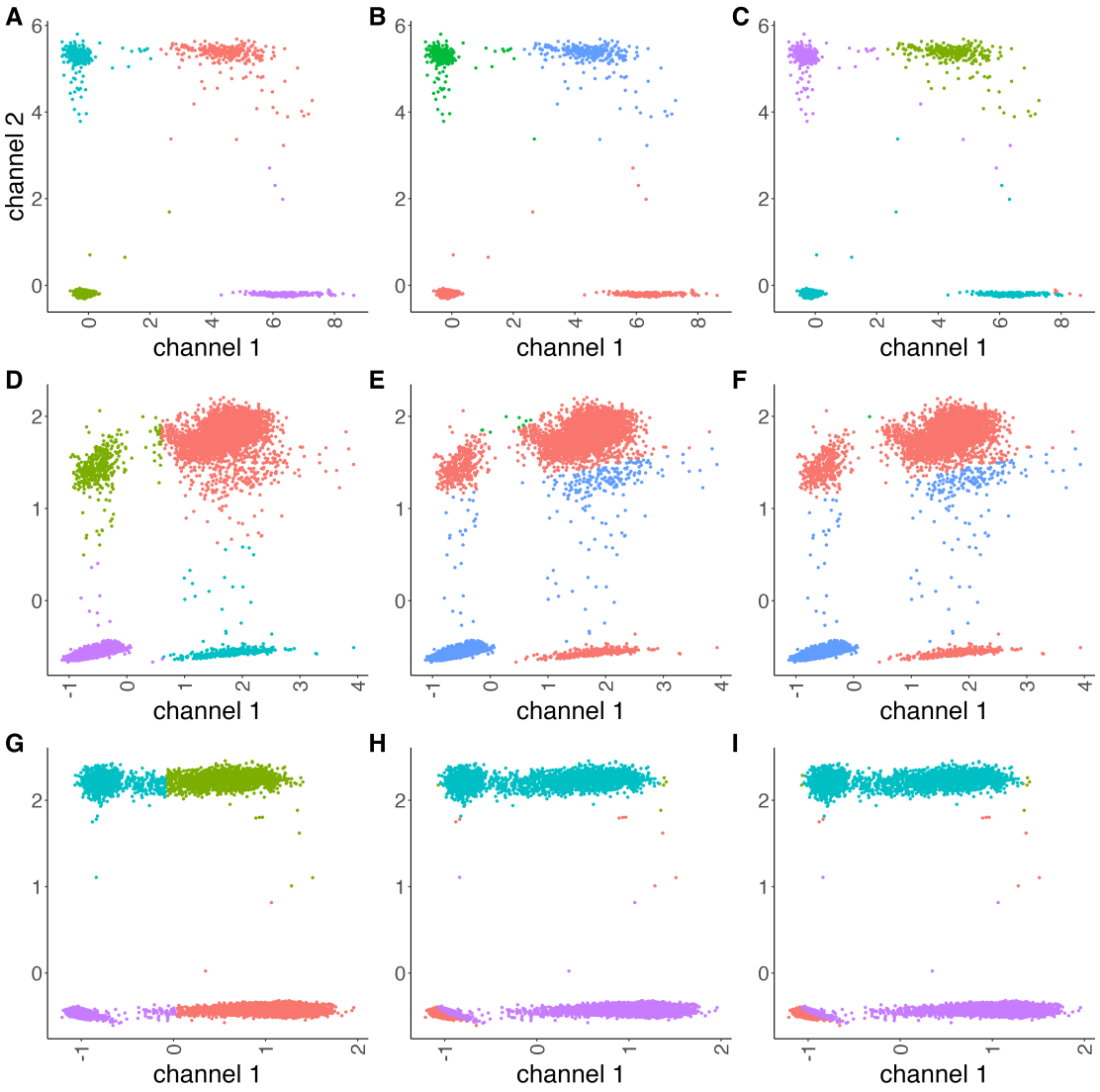

Figure S14: The best clustering results obtained using flowClust via different methods. Rows represent distinct datasets ranging from HR to MM and LR. Columns represent the default setting without prior, default prior by using 'flowClust2Prior' function, and modified prior by feeding directly the correct centroids.

SamSPECTRAL

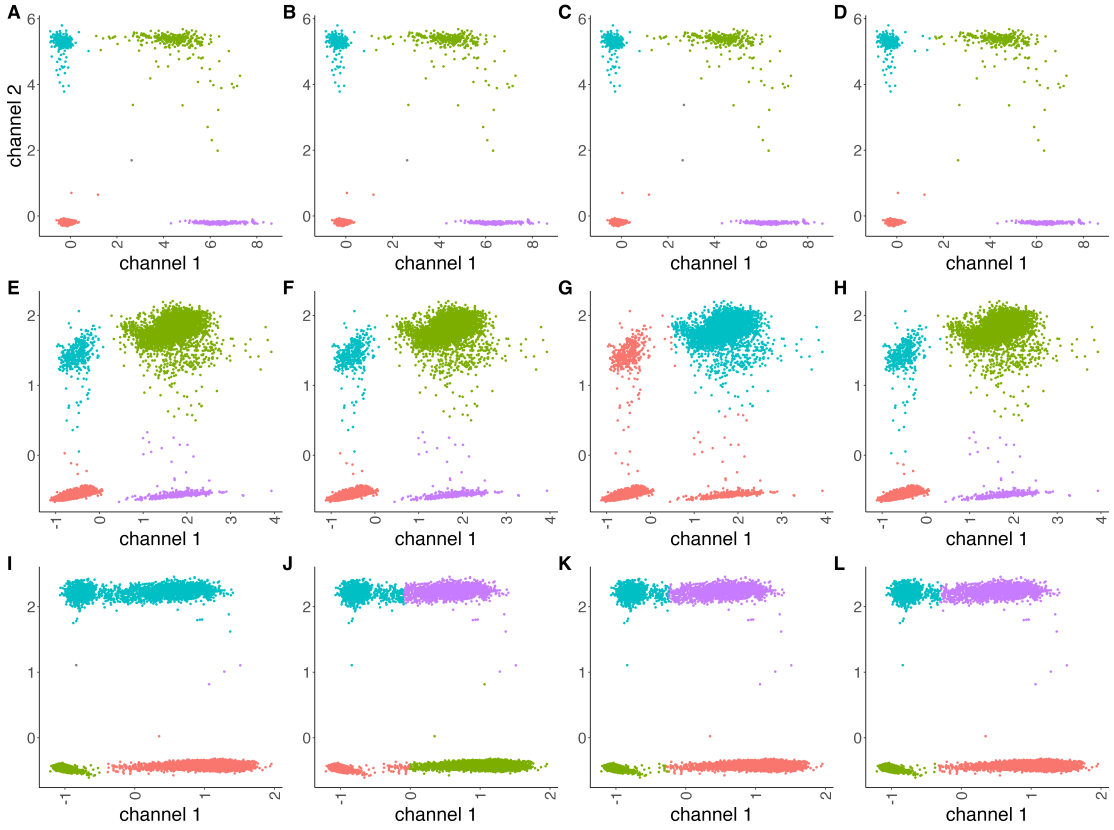

Figure S15: The best clustering results obtained using SamSPECTRAL via different methods. Rows represent distinct datasets ranging from HR to MM and LR. Columns represent various clustering methods, including the default setting, manual search, and automatic search with silhouette coefficients or adjusted Rand index, respectively.

For HR dataset, the default setting and manual search gave the best results. For other datasets, automatic search with adjusted rand index is the best (see Fig.S15).

calico

The key parameter is the raster size in the gridding step to reduce the differences in cluster sizes.

For all datasets, default setting, manual search or automatic search yielded the same results (Fig.S16). This method tends to misclassify the data points on the edge. And the adjustment of parameters does not seem to improve it. In this case, we will go with the default setting.

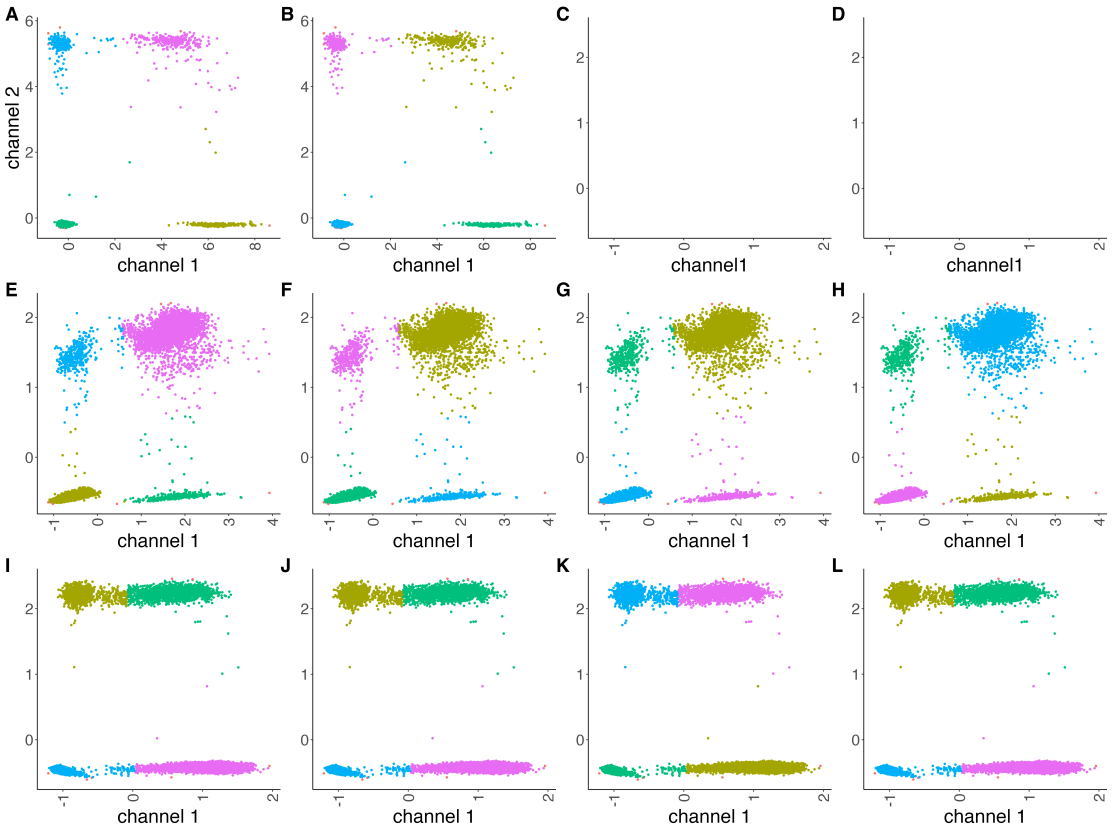

Figure S16: The best clustering results obtained using calico via different methods. Rows represent distinct datasets ranging from HR to MM and LR. Columns represent various clustering methods, including the default setting, manual search, and automatic search with silhouette coefficients or adjusted Rand index, respectively. Note that for the dataset HR, the automatic search gave an error. A possible reason is that for this high-resolution dataset with well-separated clusters, raster sizes will have very little impact on the clustering performance. The output of silhouette coefficients or adjusted rand index did not change. Thus the algorithm failed to find the optimal parameter.

7 Simulation results

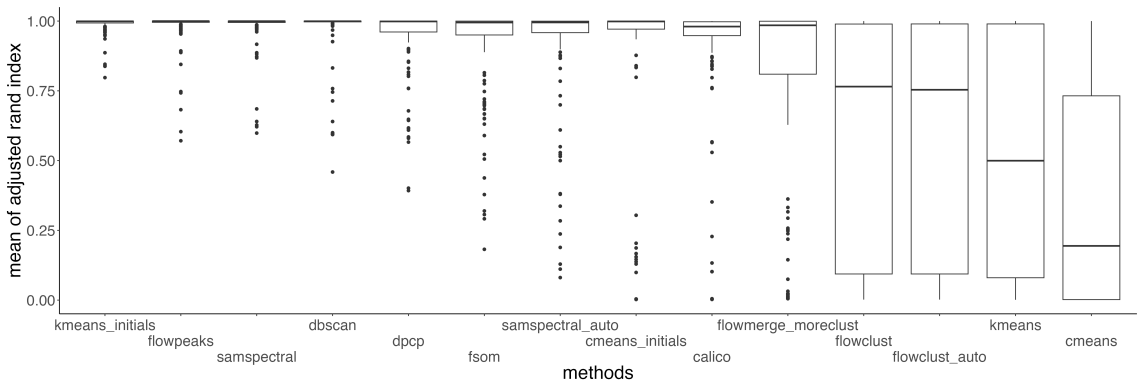

Figure S17: Averaged adjusted rand index (ARI) across the 150 factor combinations using the optimal parameters

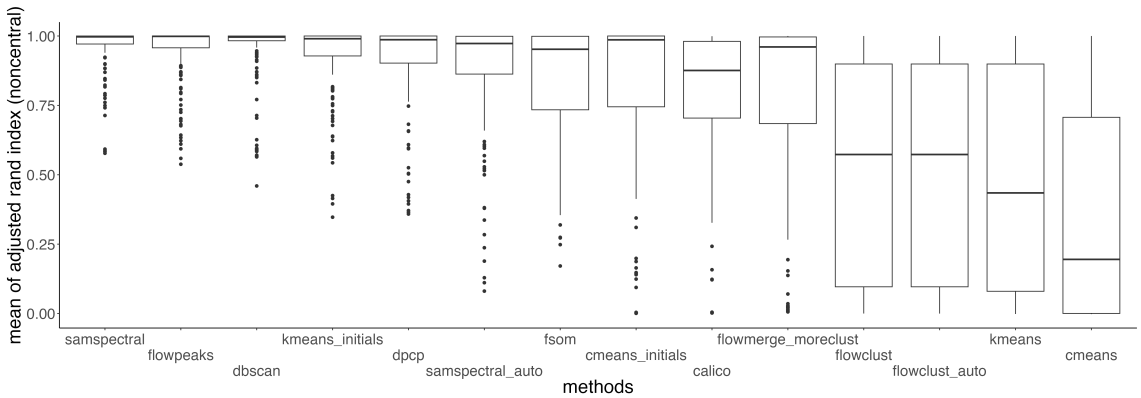

Figure S18: Adjusted rand index of data points at edge across the 150 factor combinations using the optimal parameters

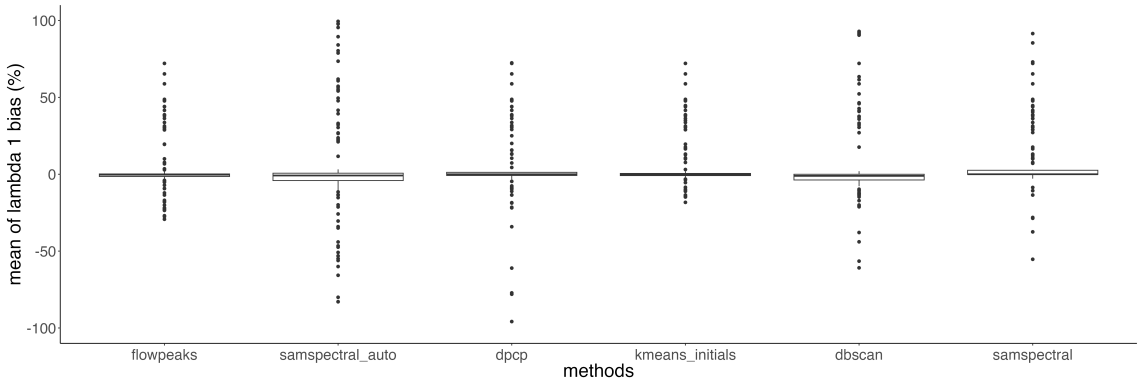

Figure S19: Relative bias of λ_1 using the optimal parameters. Note: cmeans, kmeans, flowclust, flowclust in automatic mode, flowmerge, flowSOM and calico were omitted from the bias analysis since their overall performance is poor.

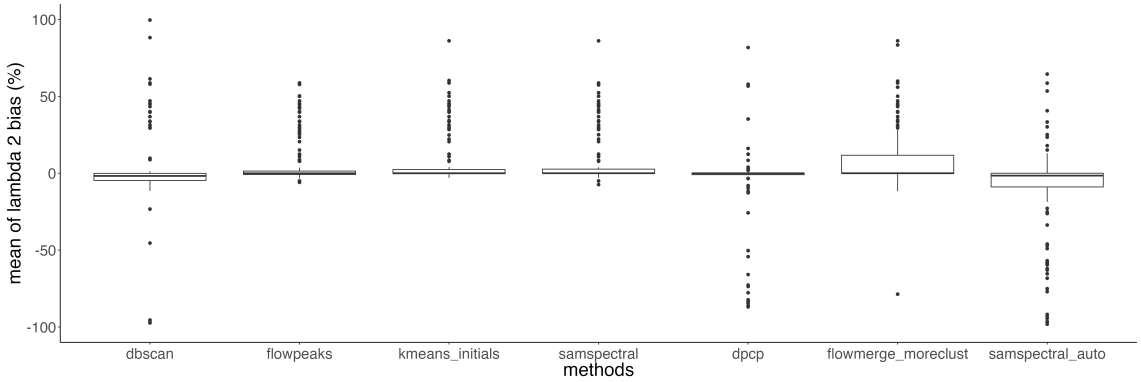

Figure S20: Relative bias of λ_2 using the optimal parameters. Note: cmeans, kmeans, flowclust, flowclust in automatic mode, flowSOM and calico were omitted from the bias analysis since their overall performance is poor.

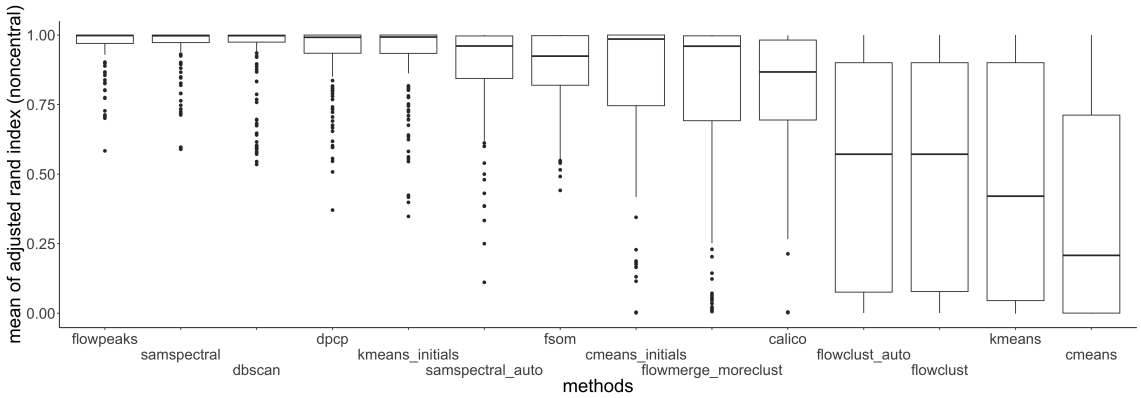

Figure S21: Adjusted rand index of data points at edge across the 150 factor combinations

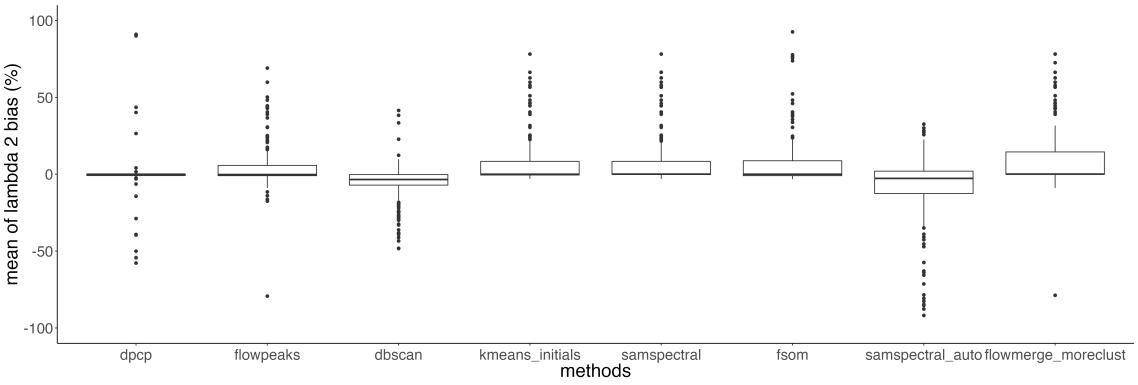

Figure S22: Relative bias of λ_2. Note: cmeans, kmeans, flowclust, flowclust in automatic mode and calico were omitted from the bias analysis since their overall performance is poor.

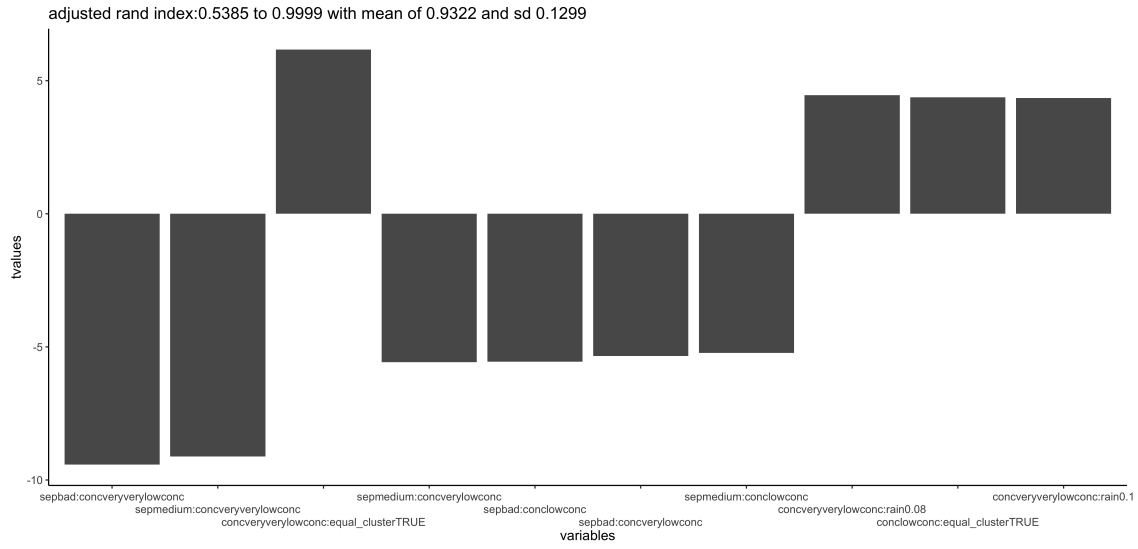

Figure S23: The top 10 variables with the largest absolute t-values of calico when fitting the adjusted rand index as the response variable. A positive t-value means this variable has a positive impact. For adjusted rand index, the higher the better.

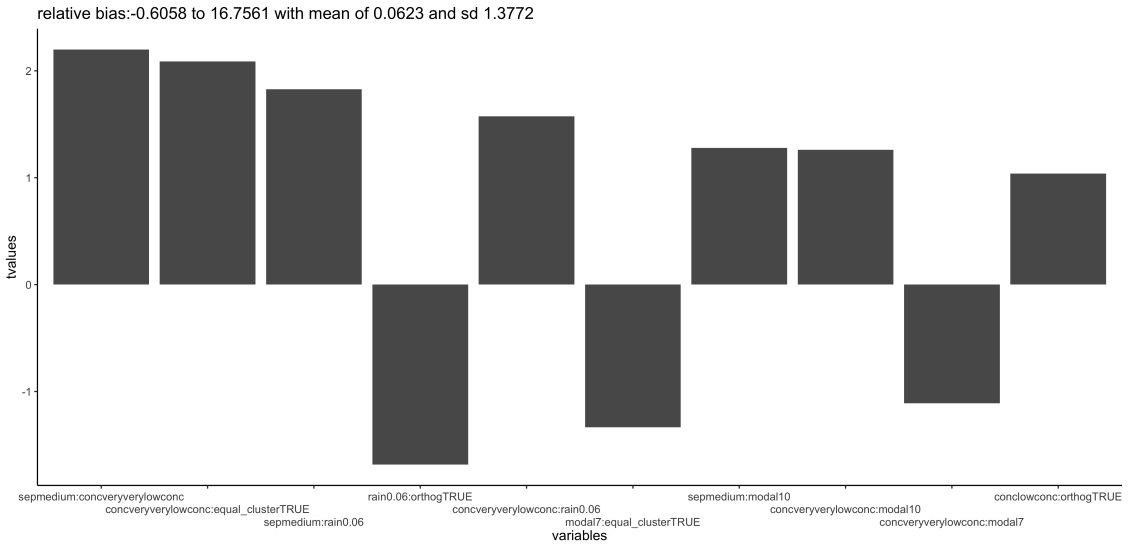

Figure S24: The top 10 variables with the largest absolute t-values of calico when fitting the relative bias of λ_1 as the response variable. A positive t-value means this variable has a positive impact. For relative bias, the lower the better.

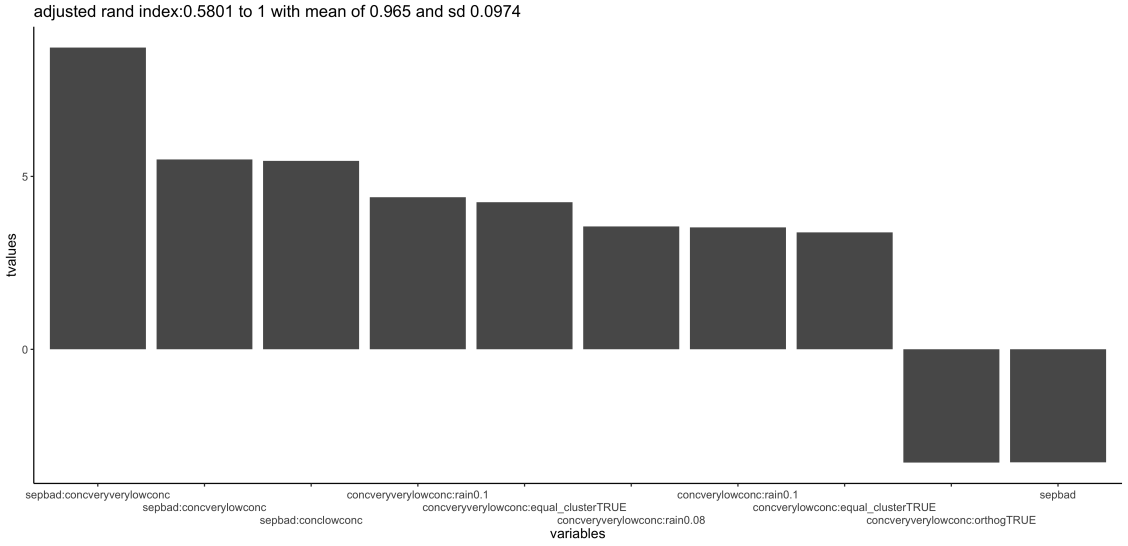

Figure S25: The top 10 variables with the largest absolute t-values of dbscan when fitting the adjusted rand index as the response variable. A positive t-value means this variable has a positive impact. For adjusted rand index, the higher the better.

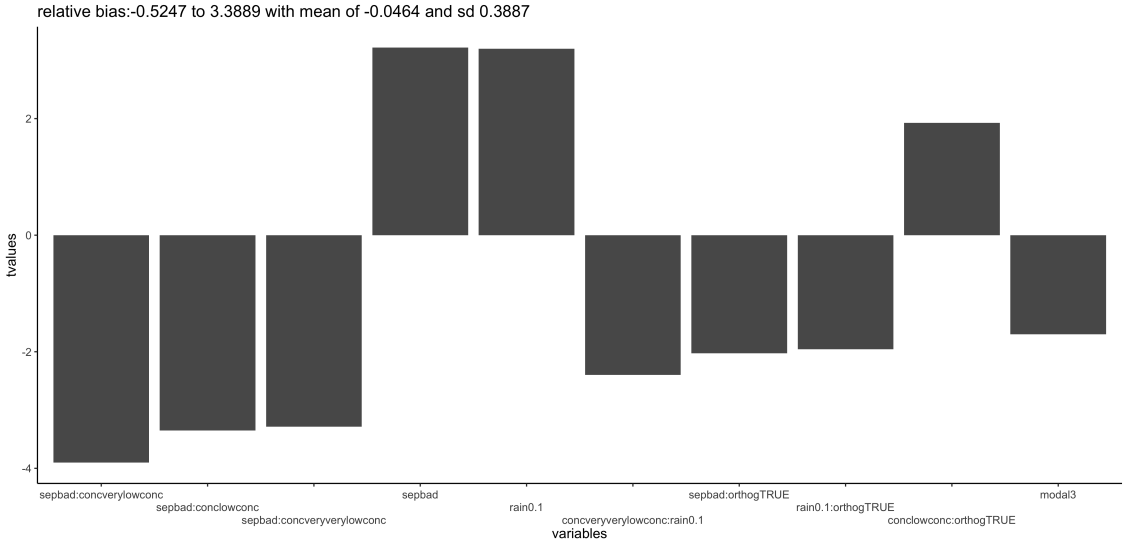

Figure S26: The top 10 variables with the largest absolute t-values of dbscan when fitting the relative bias of λ_1 as the response variable. A positive t-value means this variable has a positive impact. For relative bias, the lower the better.

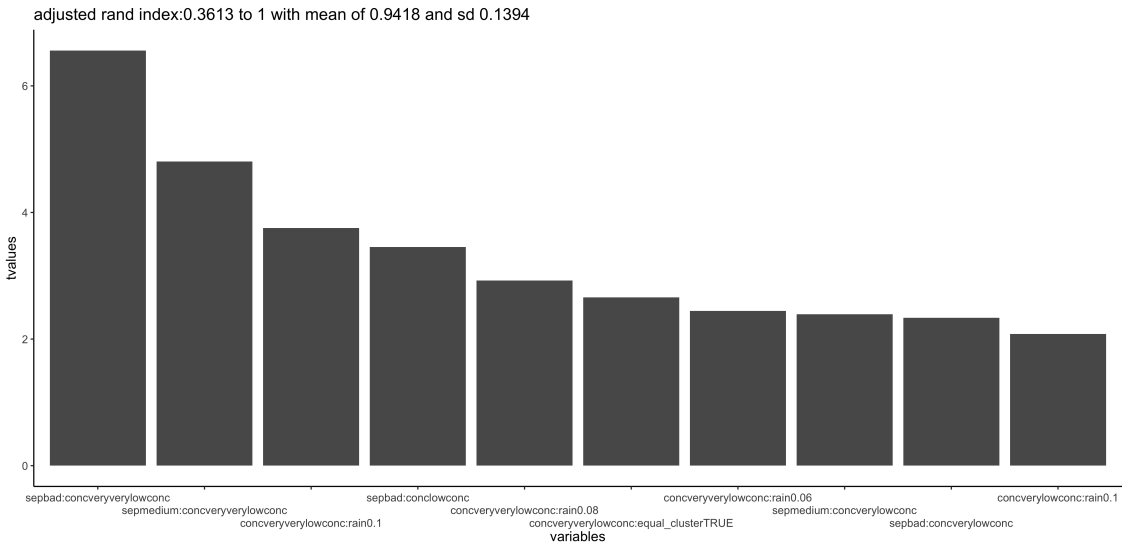

Figure S27: The top 10 variables with the largest absolute t-values of dpcp when fitting the adjusted rand index as the response variable. A positive t-value means this variable has a positive impact. For adjusted rand index, the higher the better.

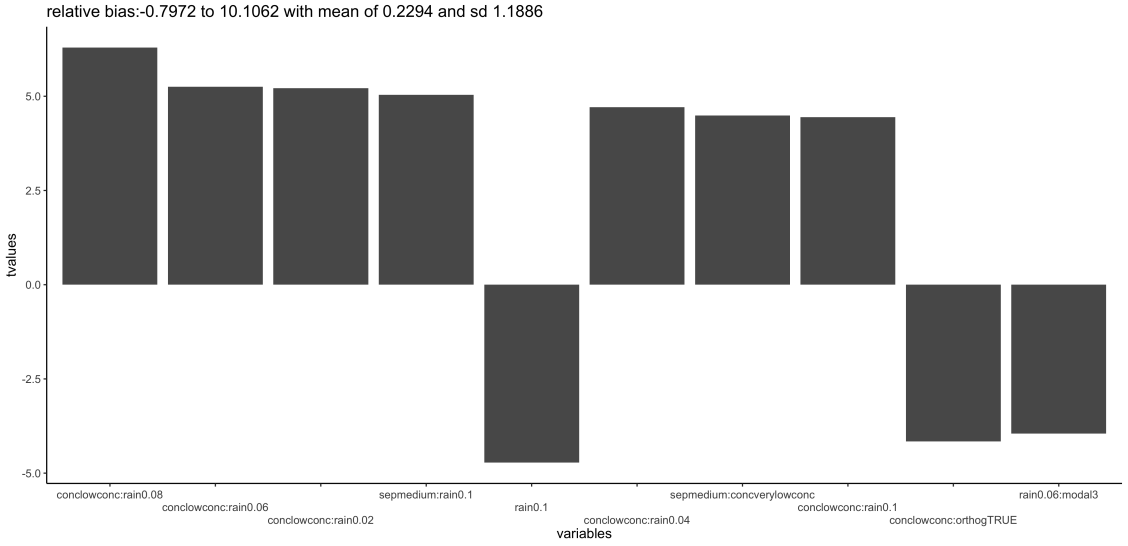

Figure S28: The top 10 variables with the largest absolute t-values of dpcp when fitting the relative bias of λ_1 as the response variable. A positive t-value means this variable has a positive impact. For relative bias, the lower the better.

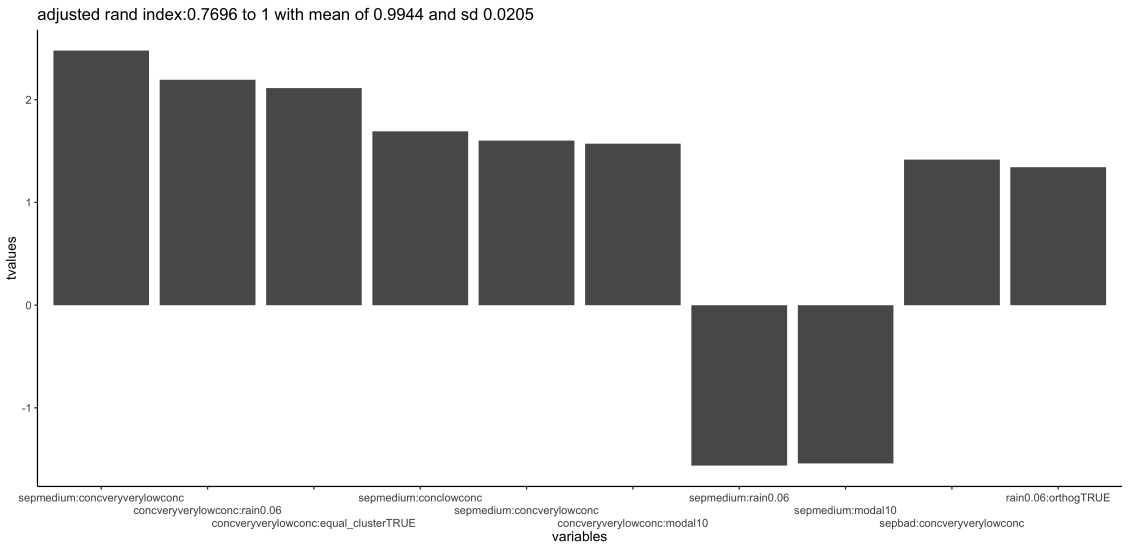

Figure S29: The top 10 variables with the largest absolute t-values of flowpeaks when fitting the adjusted rand index as the response variable. A positive t-value means this variable has a positive impact. For adjusted rand index, the higher the better.

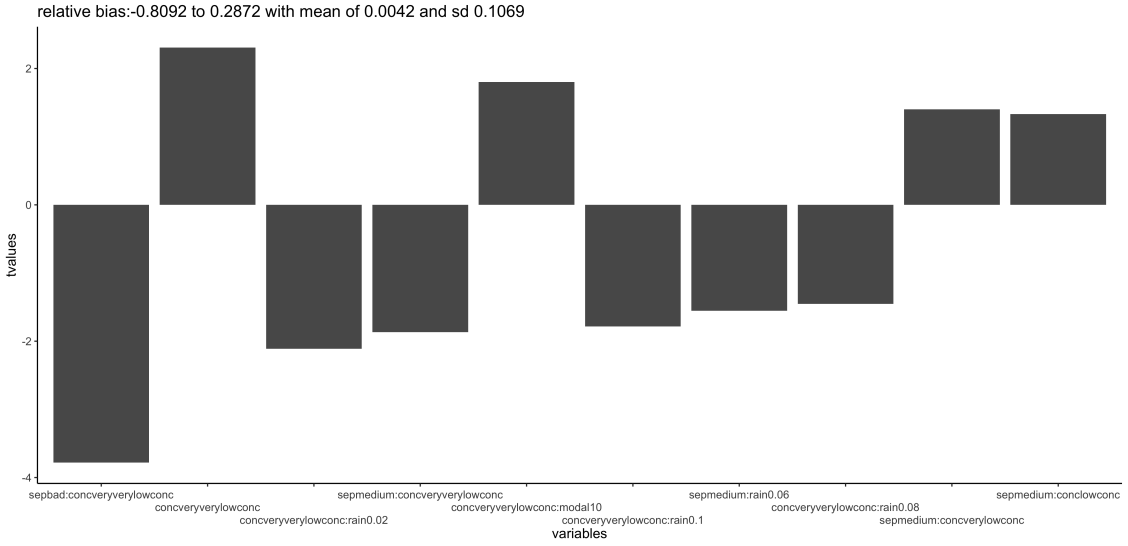

Figure S30: The top 10 variables with the largest absolute t-values of flowpeaks when fitting the relative bias of λ_1 as the response variable. A positive t-value means this variable has a positive impact. For relative bias, the lower the better.

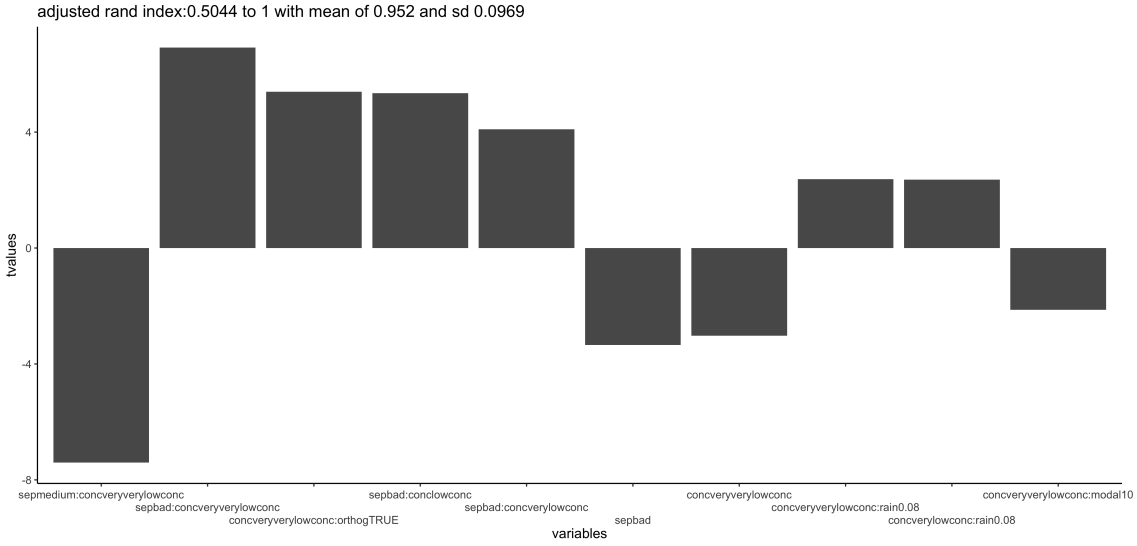

Figure S31: The top 10 variables with the largest absolute t-values of flowSOM when fitting the adjusted rand index as the response variable. A positive t-value means this variable has a positive impact. For adjusted rand index, the higher the better.

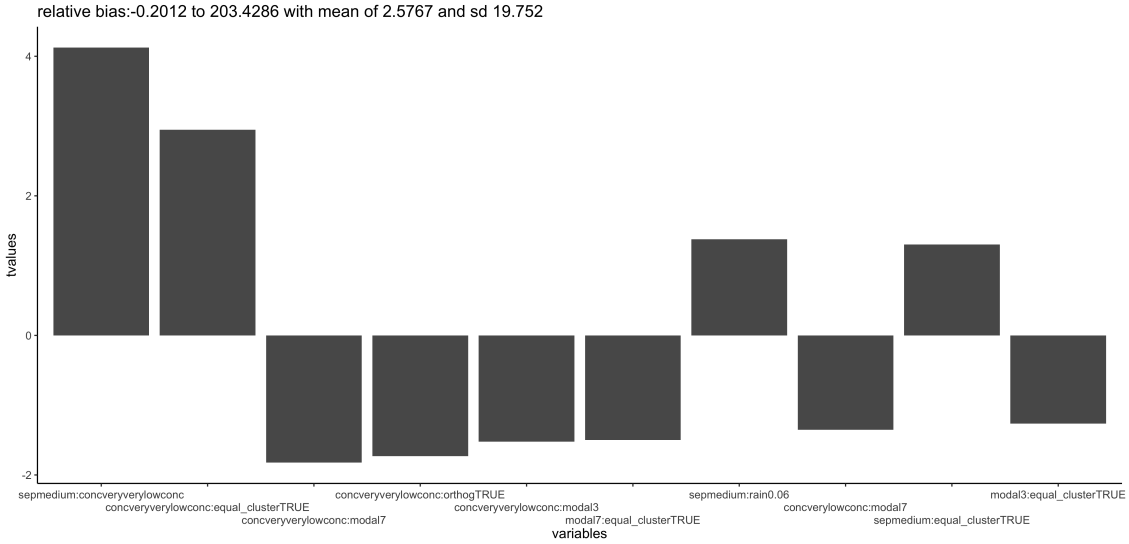

Figure S32: The top 10 variables with the largest absolute t-values of flowSOM when fitting the relative bias of λ_1 as the response variable. A positive t-value means this variable has a positive impact. For relative bias, the lower the better.

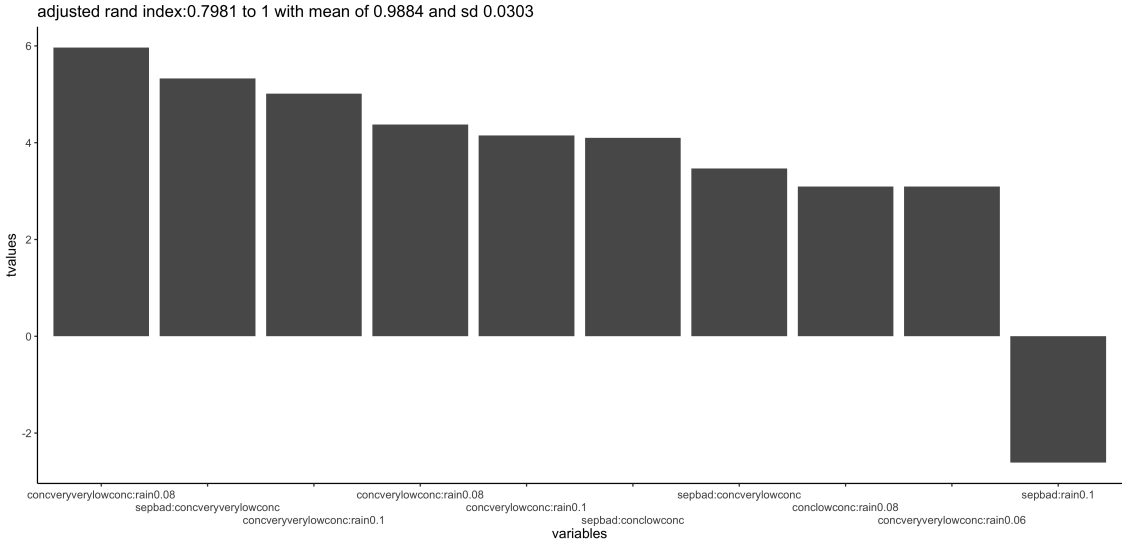

Figure S33: The top 10 variables with the largest absolute t-values of kmeans with initials when fitting the adjusted rand index as the response variable. A positive t-value means this variable has a positive impact. For adjusted rand index, the higher the better.

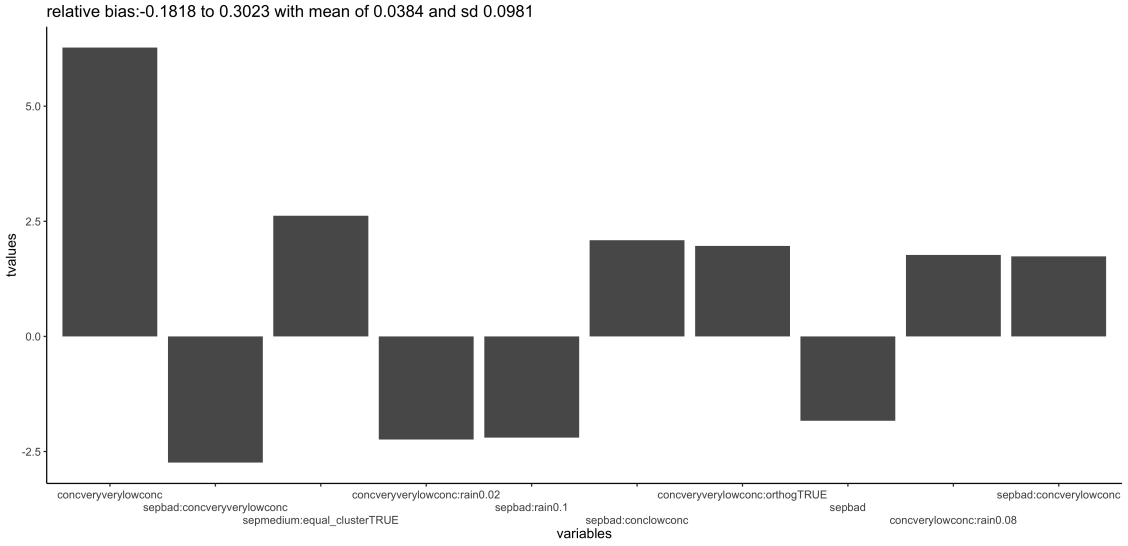

Figure S34: The top 10 variables with the largest absolute t-values of kmeans with initials when fitting the relative bias of λ_1 as the response variable. A positive t-value means this variable has a positive impact. For relative bias, the lower the better.

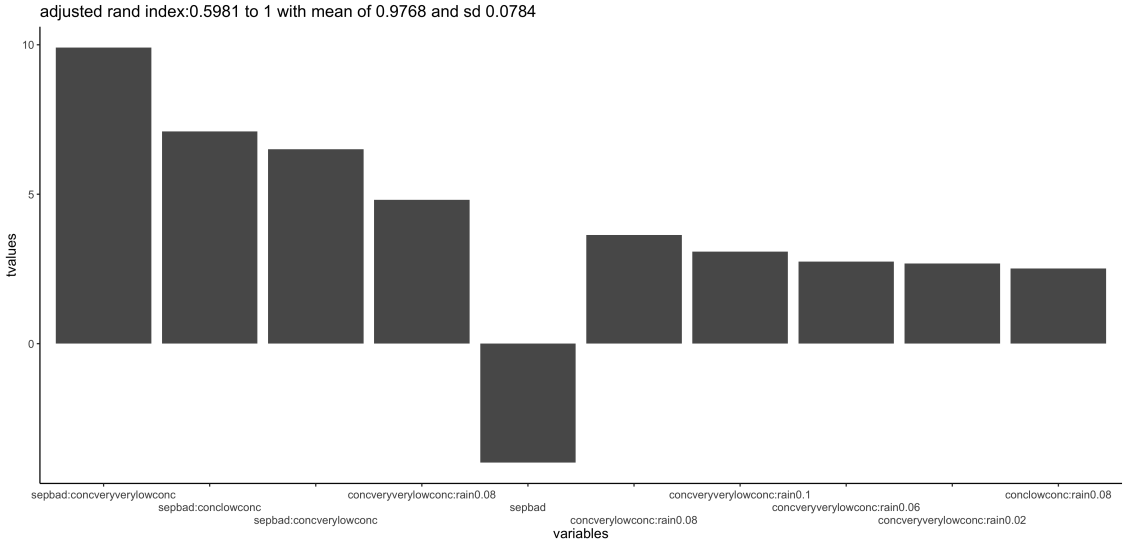

Figure S35: The top 10 variables with the largest absolute t-values of samSPECTRAL when fitting the adjusted rand index as the response variable. A positive t-value means this variable has a positive impact. For adjusted rand index, the higher the better.

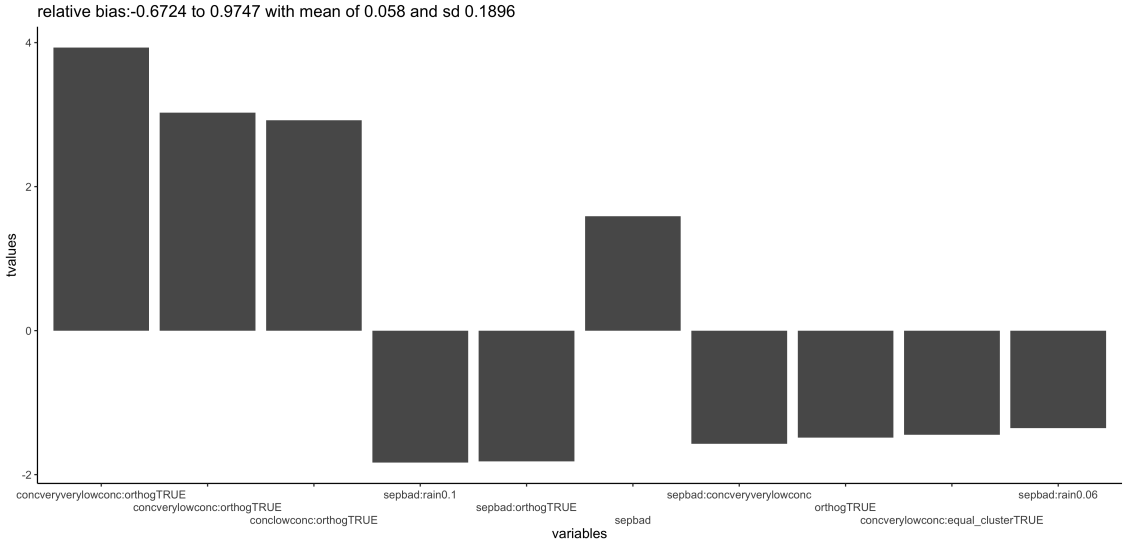

Figure S36: The top 10 variables with the largest absolute t-values of samSPECTRAL when fitting the relative bias of λ_1 as the response variable. A positive t-value means this variable has a positive impact. For relative bias, the lower the better.

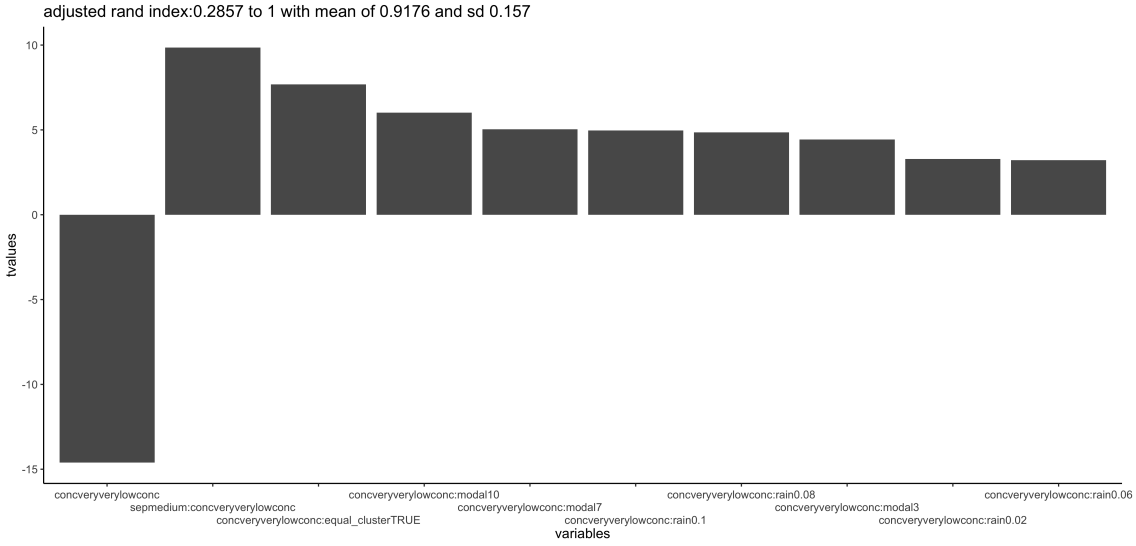

Figure S37: The top 10 variables with the largest absolute t-values of samSPECTRAL in automatic mode when fitting the adjusted rand index as the response variable. A positive t-value means this variable has a positive impact. For adjusted rand index, the higher the better.

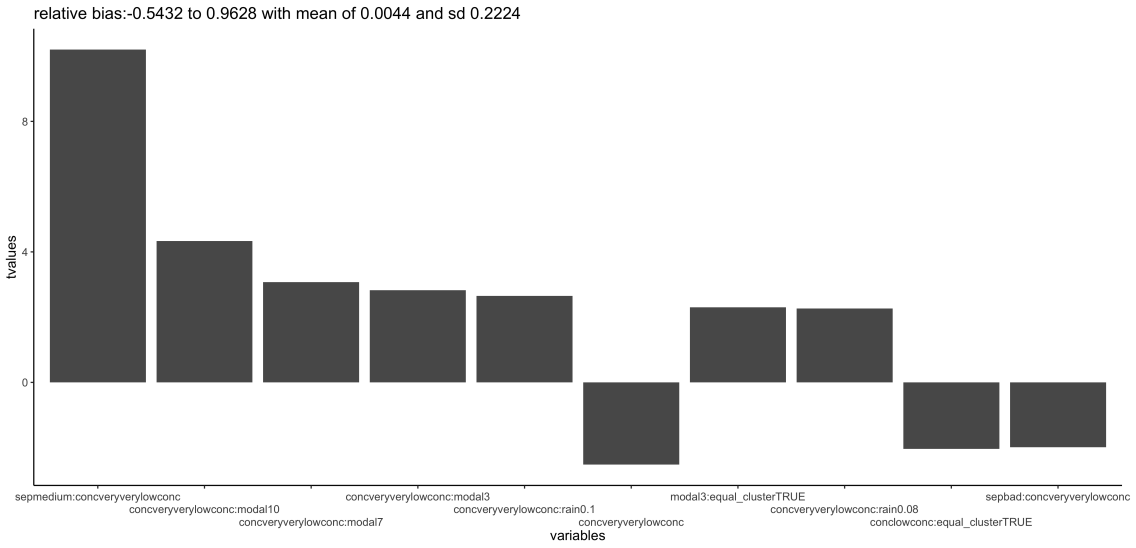

Figure S38: The top 10 variables with the largest absolute t-values of samSPECTRAL in automatic mode when fitting the relative bias of λ_1 as the response variable. A positive t-value means this variable has a positive impact. For relative bias, the lower the better.

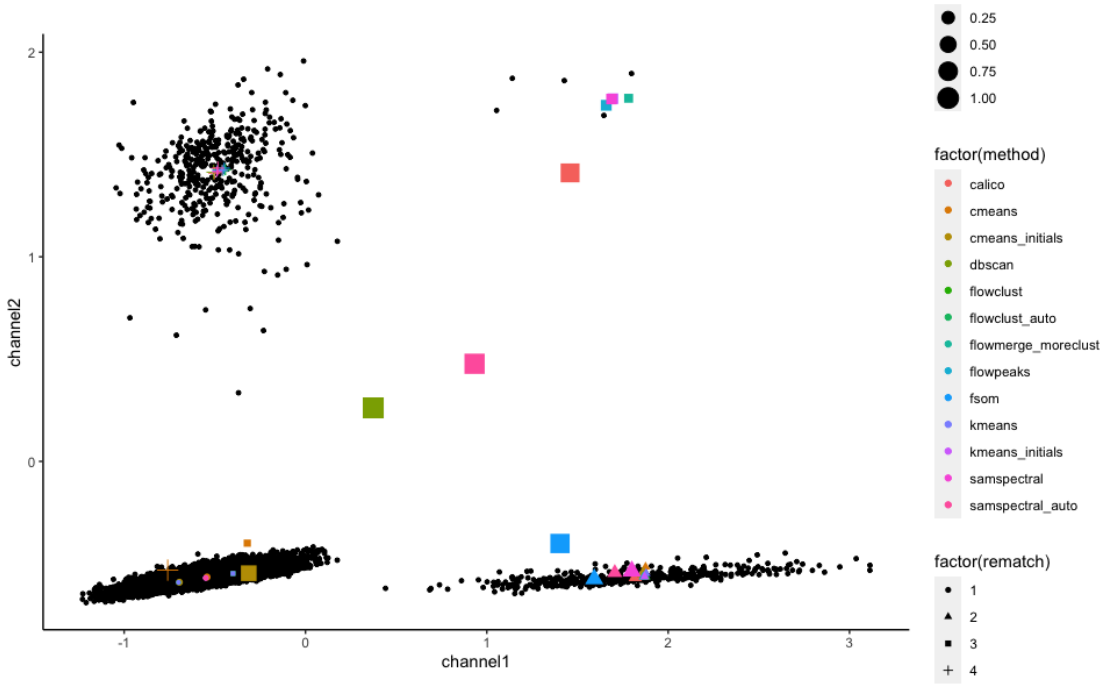

Figure S39: Cluster centers identified by the algorithms under the scenario of good resolution, very low concentration, orthogonality, unimodal, and equal size of target 1 and 2. Each clustering method is represented by a different color, with symbols indicating the clusters. The size of the symbols reflects the variation of the estimates.

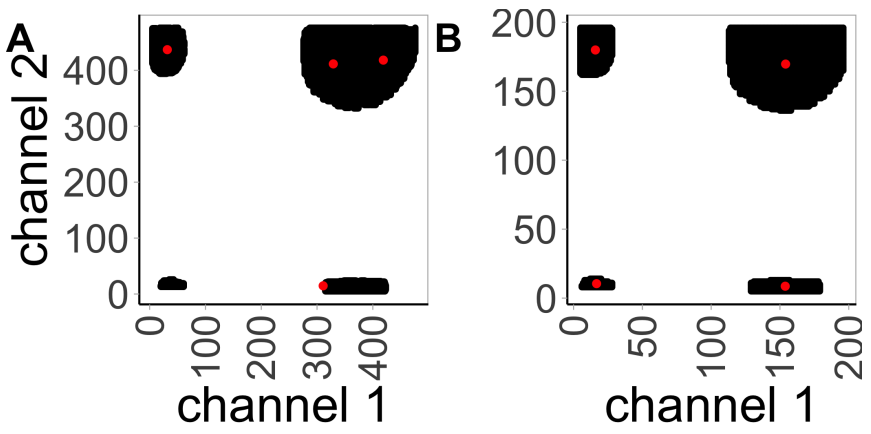

Figure S40: (A) centers found (indicated by red round circles) in calico using 480*480 grids (B) centers found in calico using 200*200 grids. Note that using 480*480 grids, at the high concentration levels, the single positive and double positive clusters are much larger than the negative population (at least 10 times), while using 200*200 grids, the size of single positives and negative population is more comparable (3 to 5 times bigger).

8 Empirical data analysis

Table S4: Performance metrics from the resampling study of the MM dataset

method	$\frac{\hat{\lambda}_1-\lambda_1}{\lambda_1}$ (%)	$\frac{\hat{\lambda}_2-\lambda_2}{\lambda_2}$ (%)	ARI	ARI non-central	number of clusters
dpcp	-0.01‖-0.91	-0.45‖-7.03	0.997‖0.996	0.967‖0.968	/
flowSOM	-0.10‖-2.01	-0.68‖-1.11	0.994‖0.989	0.941‖0.910	/
kmeans with initials	-0.57	-0.62	0.995	0.951	/
kmeans	-0.61	-0.62	0.995	0.950	/
flowclust	-0.61	-0.62	0.995	0.951	/
flowclust auto	-0.61	-0.62	0.995	0.951	4
flowmerge	-0.62	-0.77	0.996	0.973	4.06
calico	-0.78‖-1.00	-0.75‖-0.59	0.994‖0.992	0.941‖0.942	/
cmeans with initials	-0.95	-0.67	0.994	0.943	/
SamSPECTRAL	1.32‖-0.74	-0.67‖-0.39	0.993‖0.997	0.975‖0.981	/
flowpeaks	2.07‖-0.35	-0.09‖-0.90	0.974‖0.997	0.942‖0.974	4.43‖6.37
SamSPECTRAL auto	2.07‖-31.54	-1.79‖-30.03	0.978‖0.965	0.947‖0.932	3.83‖7.4
DBSCAN	-2.61‖-2.00	-3.08‖-2.46	0.995‖0.996	0.959‖0.966	5.01‖6.47
ddPCRclust	17.04	1.94	0.930	0.873	/
cmeans	42.70	121.03	0.544	0.528	/

Average relative bias of λ_1 and λ_2, the ARI calculated for all resampled 10000 data points and for the data points on the edge only, and the average number of clusters identified. '‖' represents 'optimal parameter‖default parameter'. For those methods which have only value, either no optimal parameters are available or the optimal ones coincide with the default ones. The methods are ranked from low to high relative bias (sum of the absolute relative biases $|\lambda_1| + |\lambda_2|$) based on the results with optimal parameters. '/': the number of clusters is pre-defined.

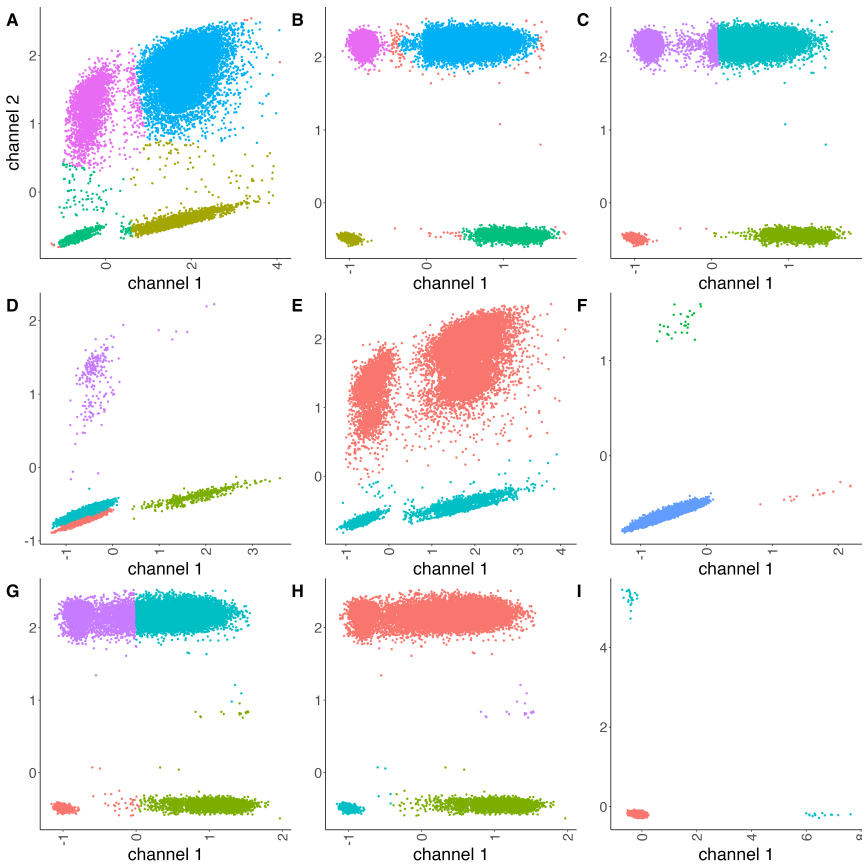

Figure S41: Clustering results with optimal parameters applied to scenarios where methods fail with default parameters: calico (A), DBSCAN (B), dpcp (C), flowmerge (D), flowpeaks (E), flowSOM (F), kmeans with initials (G), samspectral (H), sampectral in automatic mode (I).

Table S5: Performance metrics from the resampling study of the LR dataset

method	$\frac{\hat{\lambda}_1-\lambda_1}{\lambda_1}$ (%)	$\frac{\hat{\lambda}_2-\lambda_2}{\lambda_2}$ (%)	ARI	ARI non-central	number of clusters
flowmerge	-0.21	0	0.999	0.986	4
flowpeaks	-1.05‖-1.09	-0.15‖0.19	0.996‖0.995	0.964‖0.963	5.09‖6.24
kmeans with initials	-2.88	-0.024	0.985	0.872	/
kmeans	-2.88	-0.024	0.985	0.872	/
flowclust	-2.88	-0.024	0.985	0.872	/
flowclust auto	-2.88	-0.024	0.985	0.872	4
cmeans with initials	-2.97	-0.024	0.984	0.869	/
calico	-2.97‖-2.96	-0.13‖0.19	0.984‖0.984	0.869‖0.866	/
dpcp	-3.35‖-7.63	0‖0.32	0.985‖0.976	0.889‖0.890	/
SamSPECTRAL auto	-2.45‖-5.72	-2.03‖-7.56	0.985‖0.985	0.914‖0.928	3.95‖4.57
SamSPECTRAL	-7.80‖-6.80	0‖0.27	0.987‖0.989	0.958‖0.964	/
flowSOM	-9.54‖-12.40	-0.04‖0.26	0.967‖0.965	0.870‖0.866	/
DBSCAN	-5.54‖-12.95	-9.28‖-0.01	0.986‖0.968	0.899‖0.961	5‖3.95
cmeans	13.73	57.88	0.874	0.887	/
ddPCRclust	93.56‖102.60	0.02‖0.382	0.753‖0.725	0.736‖0.710	/

Average relative bias of λ_1 and λ_2, the ARI calculated for all resampled 10000 data points and for the data points on the edge only, and the average number of clusters identified. '‖' represents 'optimal parameter‖default parameter'. For those methods which have only value, either no optimal parameters are available or the optimal ones coincide with the default ones. The methods are ranked from low to high relative bias (sum of the absolute relative biases $|\lambda_1| + |\lambda_2|$) based on the results with optimal parameters. '/': the number of clusters is pre-defined.

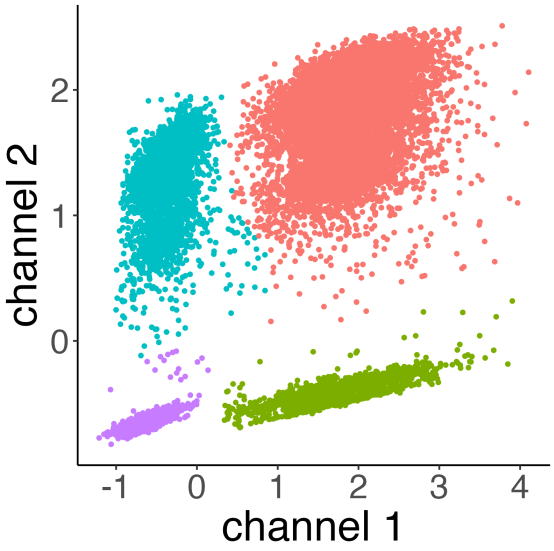

Figure S42: flowpeaks applied with optimal parameters of LR dataset (not MM dataset) to the scenario where this method failed

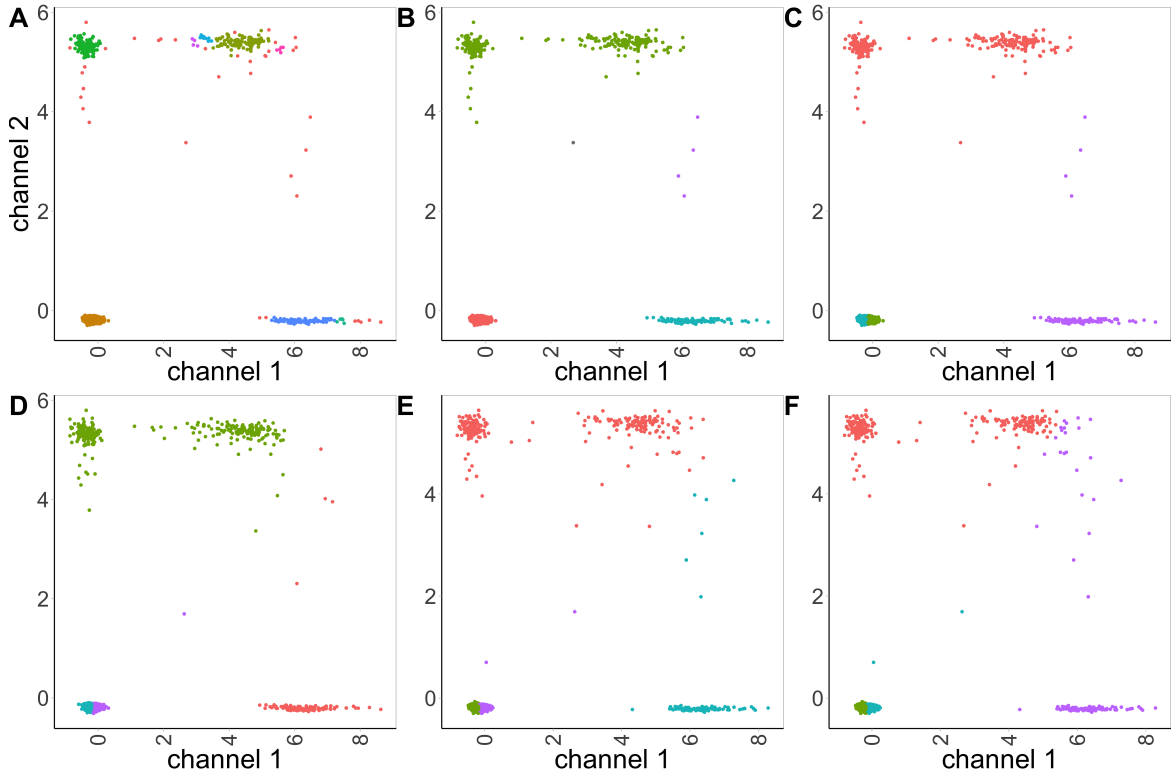

Figure S43: (A) DBSCAN (B) SamSPECTRAL (C) flowclust (D) flowclust in automatic mode (E) kmeans (F) cmeans clustering results for 10 000 samples of HR dataset

method	**runtime (ms)**
kmeans with initials	6
cmeans with initials	99
cmeans	115
calico	153
dpcp	261
DBSCAN	284
kmeans	425
flowpeaks	586
flowSOM	635
ddPCRclust	1188
flowclust	1591
flowclust auto	2574
flowmerge	7544
SamSPECTRAL	10089
SamSPECTRAL auto	13805

Table S6: Runtimes of clustering methods. The runtime is averaged over 100 resampled data points of the HR dataset. The runtime is estimated in *R* (version 4.2.2) on MacBook Air (M2, 2022, 8 GB memory, OSX version 12.5)

References

[1] Rong Yan Liu, Jon Michael Parelius, and Kesar Singh. Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics*, 27:783–858, 1999.

[2] Matthijs Vynck, Yao Chen, David Gleerup, Jo Vandesompele, Wim Trypsteen, Antoon Lievens, Olivier Thas, and Ward De Spiegelaere. Digital pcr partition classification. *Clinical Chemistry*, page hvad063, 2023.