A guide to use the Polytect: an automatic clustering and labeling method for multi-color digital PCR data

Yao Chen

Introduction

Digital PCR (dPCR) is a state-of-the-art quantification method of target nucleic acids. The technology is based on massive partitioning of a reaction mixture into individual PCR reactions. This results in partition-level end-point fluorescence intensities that are subsequently used to classify partitions as positive or negative, i.e., containing or not containing the target nucleic acid(s). Many automatic dPCR partition classification methods have been proposed, but they are limited to the analysis of single or dual color dPCR data. While general-purpose or flow cytometry clustering methods can be directly applied to multi-color dPCR data, these methods do not exploit the approximate prior knowledge on cluster center locations available in dPCR data.

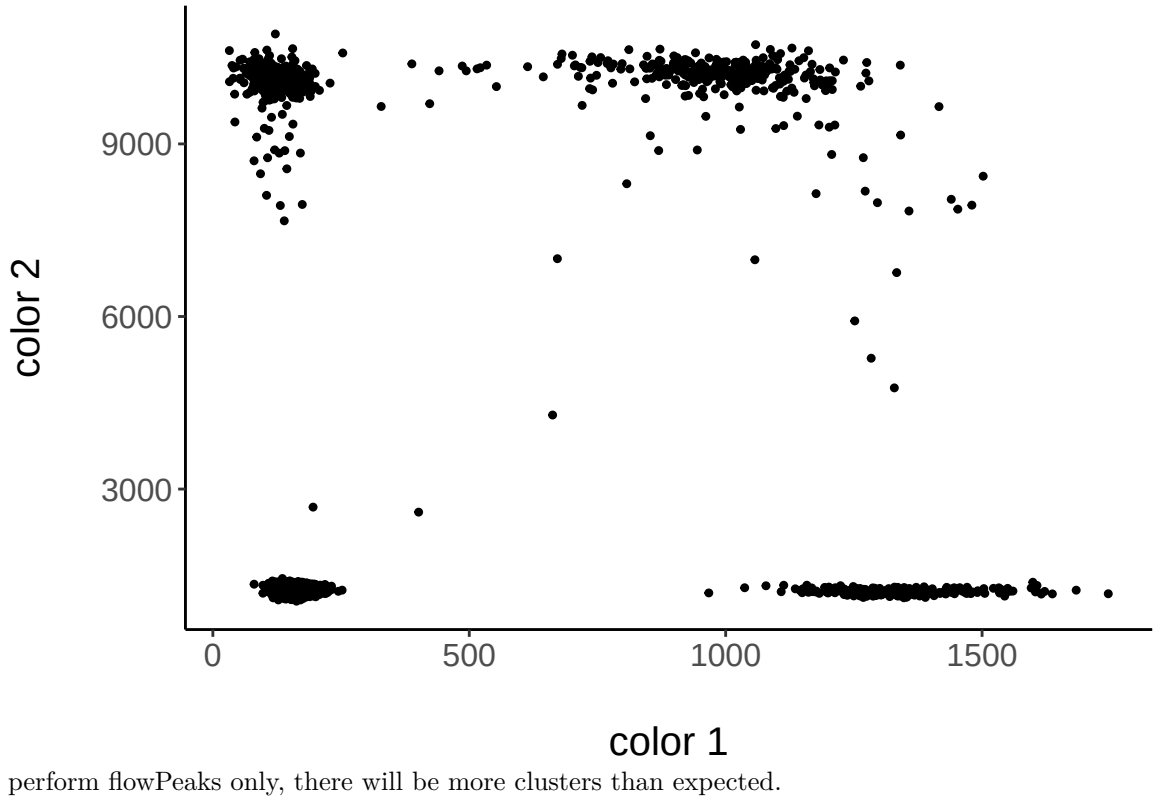
We developed a novel method, `Polytect`, that relies on crude cluster results from flowPeaks, previously shown to offer good partition classification performance, and subsequently refines flowPeaks' results by cluster merging and automatic cluster labeling, exploiting the prior knowledge on cluster center locations.

Examples

To demonstrate the core functions of this package, we will use the HR digital PCR dataset, which can be accessed using the command data(HR). The HR dataset is a data frame consisting of 18,233 rows and 2 columns. The dataset has 4 clusters. The clustering analysis can be performed using the following commands.

```
library(Polytect)
library(ggplot2)
library(flowPeaks)
data(HR)
head(HR)
#>   channel1  channel2
#> 1 32.25462 10622.230
#> 2 32.55316 10081.106
#> 3 37.41653 10363.010
#> 4 39.65380 10140.198
#> 5 40.31481 10319.560
#> 6 42.58688  9864.532
```

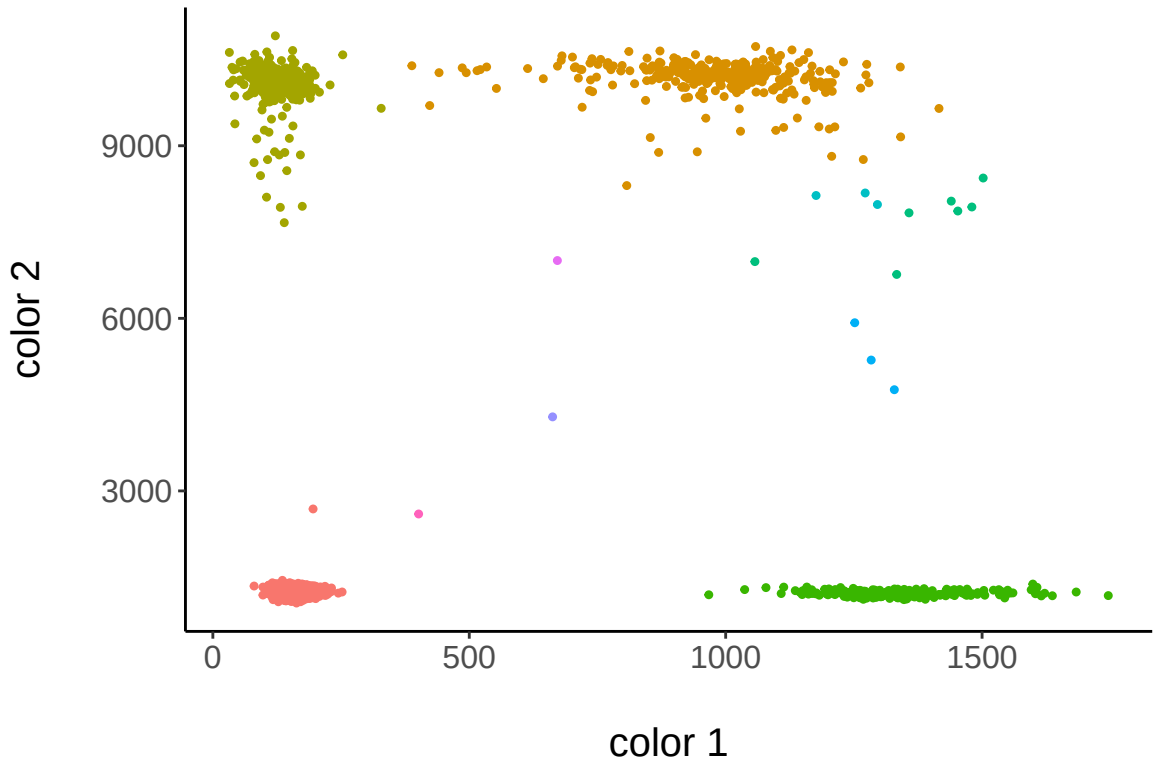
```
ggplot(data=HR, aes(channel1, channel2))+ geom_point(size=0.9,show.legend = FALSE) +
  labs(x = "color 1", y = "color 2") +theme(text = element_text(size = 15), panel.grid.major = element_
                                            panel.grid.minor = element_blank(),
                                            panel.background = element_blank(),
                                            axis.line = element_line(colour = "black"),
                                            plot.margin = margin(t = 20,  # Top margin
                                                                 r = 20,  # Right margin
                                                                 b = 20,  # Bottom margin
                                                                 l = 20),
                                            axis.title.x = element_text(margin = margin(t = 20)),
                                            axis.title.y = element_text(margin = margin(r = 20)))
```



If we perform flowPeaks only, there will be more clusters than expected.

```
data_scaled<-apply(HR,2,function(x) (x-min(x))/(max(x)-min(x)))
data_input<-as.matrix(data_scaled)
fp<-flowPeaks(data_input)
#>          step 0, set the intial seeds, tot.wss=0.0767292
#>          step 1, do the rough EM, tot.wss=0.0519279 at 0.059886 sec
#>          step 2, do the fine transfer of Hartigan-Wong Algorithm
#>                    tot.wss=0.0483346 at 0.11451 sec
```

```
ggplot(data=HR, aes(channel1, channel2,colour = factor(fp$peaks.cluster)))+ geom_point(size=0.9,show.leg
                                                    panel.grid.minor = element_blank(),
                                                    panel.background = element_blank(),
                                                    axis.line = element_line(colour = "black"),
                                                    plot.margin = margin(t = 20,  # Top margin
                                                                         r = 20,  # Right margin
                                                                         b = 20,  # Bottom margin
                                                                         l = 20),
                                                    axis.title.x = element_text(margin = margin(t = 20)),
                                                    axis.title.y = element_text(margin = margin(r = 20)))
```



The main function that performs flowPeaks first, then merges the excess clusters.

```
result<-polytect_clust(data=HR,cluster_num=4,type="2color")
#>         step 0, set the intial seeds, tot.wss=0.0767292
#>         step 1, do the rough EM, tot.wss=0.0519279 at 0.058716 sec
#>         step 2, do the fine transfer of Hartigan-Wong Algorithm
#>                  tot.wss=0.0483346 at 0.113486 sec
```

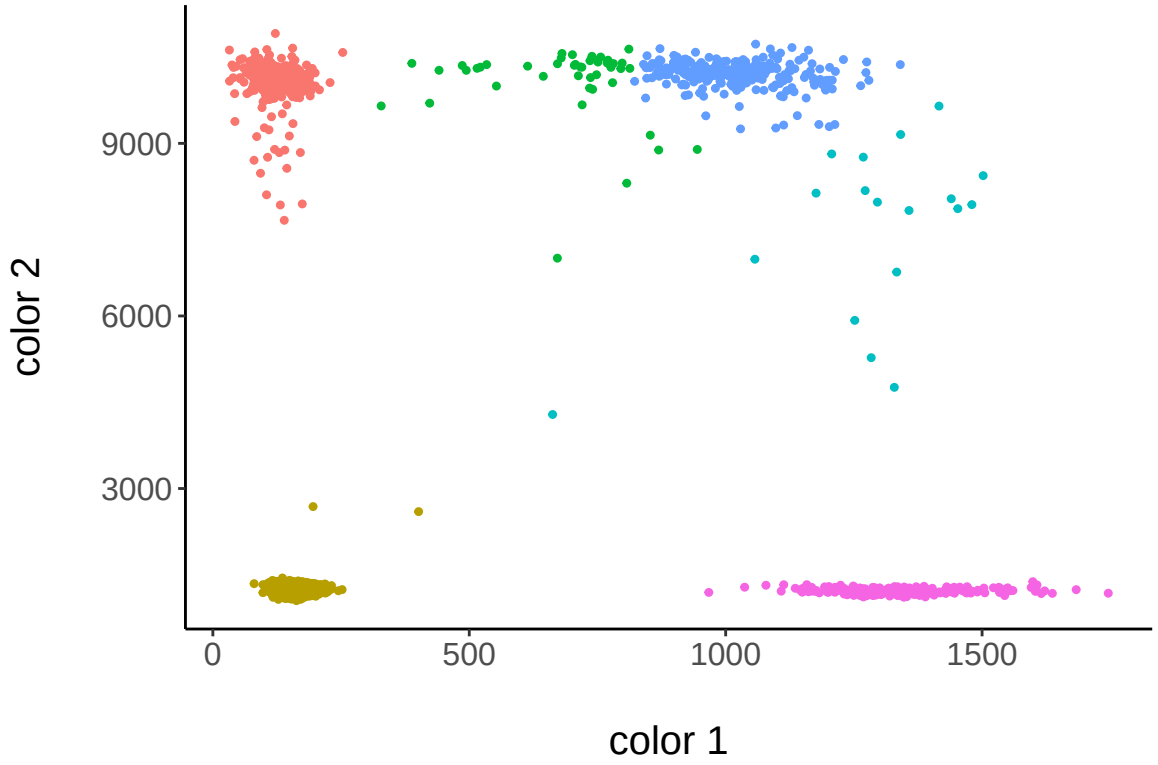
```
print(head(result))
#>   channel1  channel2 cluster
#> 1 32.25462 10622.230       3
#> 2 32.55316 10081.106       3
#> 3 37.41653 10363.010       3
#> 4 39.65380 10140.198       3
#> 5 40.31481 10319.560       3
#> 6 42.58688  9864.532       3
```

Or you can use any initial clustering results as an input to the polytect_merge function

```
## it is advised to standardize the data
dist_matrix <- dist(data_input)
hc <- hclust(dist_matrix, method = "ward.D2")
# the number of clusters is specified at 6, which is larger than 4
hc_clusters <- cutree(hc, k = 6)

ggplot(data=HR, aes(channel1, channel2,colour = factor(hc_clusters)))+ geom_point(size=0.9,show.legend =
                                                    panel.grid.minor = element_blank(),
                                                    panel.background = element_blank(),
                                                    axis.line = element_line(colour = "black"),
                                                    plot.margin = margin(t = 20,  # Top margin
                                                                         r = 20,  # Right margin
```

```
                                                    b = 20,  # Bottom margin
                                                    l = 20),
                                axis.title.x = element_text(margin = margin(t = 20)),
                                axis.title.y = element_text(margin = margin(r = 20)))
```

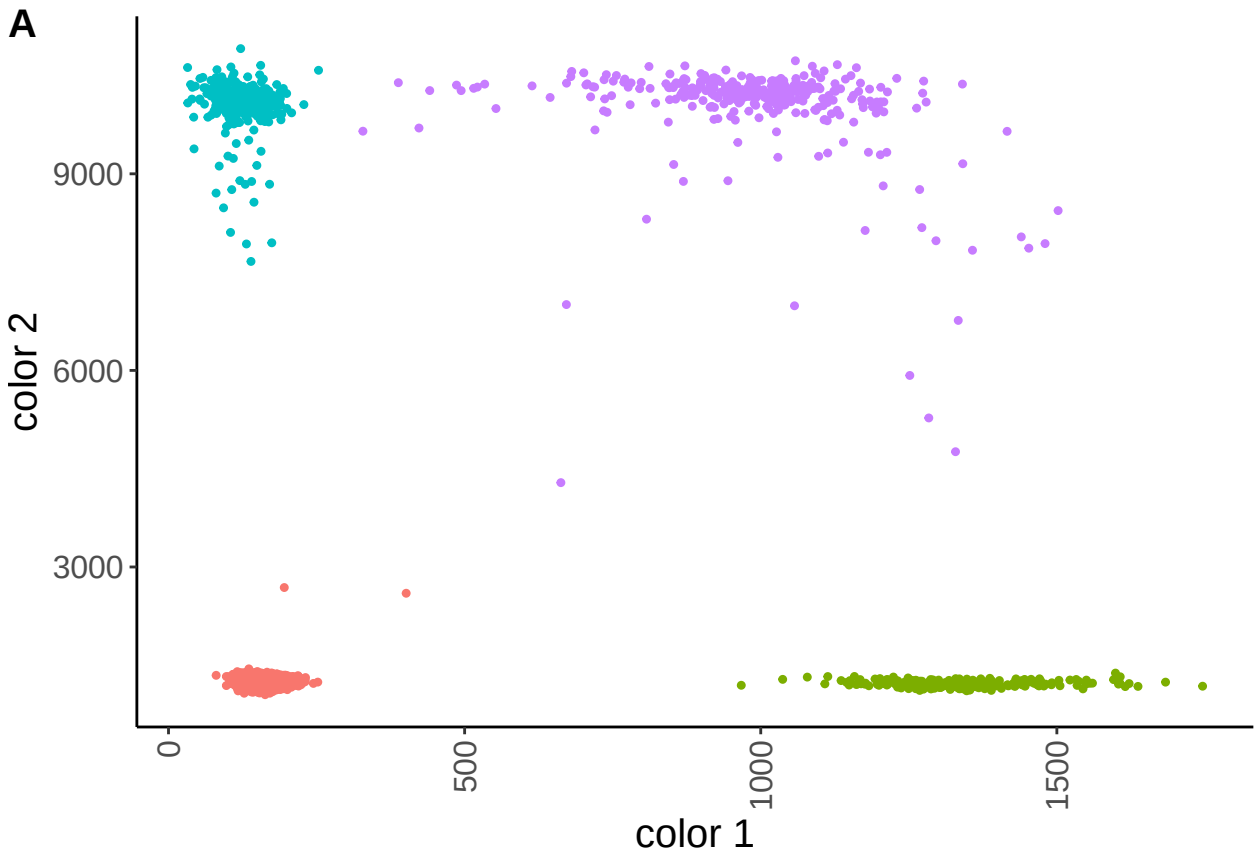



```
hc_parse<-list()
hc_parse$cluster<-hc_clusters

result<-polytect_merge(data=HR,cluster_num=4,base_clust=hc_parse,type="2color")
print(head(result))
#>    channel1  channel2 cluster
#> 1 32.25462 10622.230       3
#> 2 32.55316 10081.106       3
#> 3 37.41653 10363.010       3
#> 4 39.65380 10140.198       3
#> 5 40.31481 10319.560       3
#> 6 42.58688  9864.532       3
```

The clustering results can be visualized by 2-d plots.

```
polytect_plot(result)
```

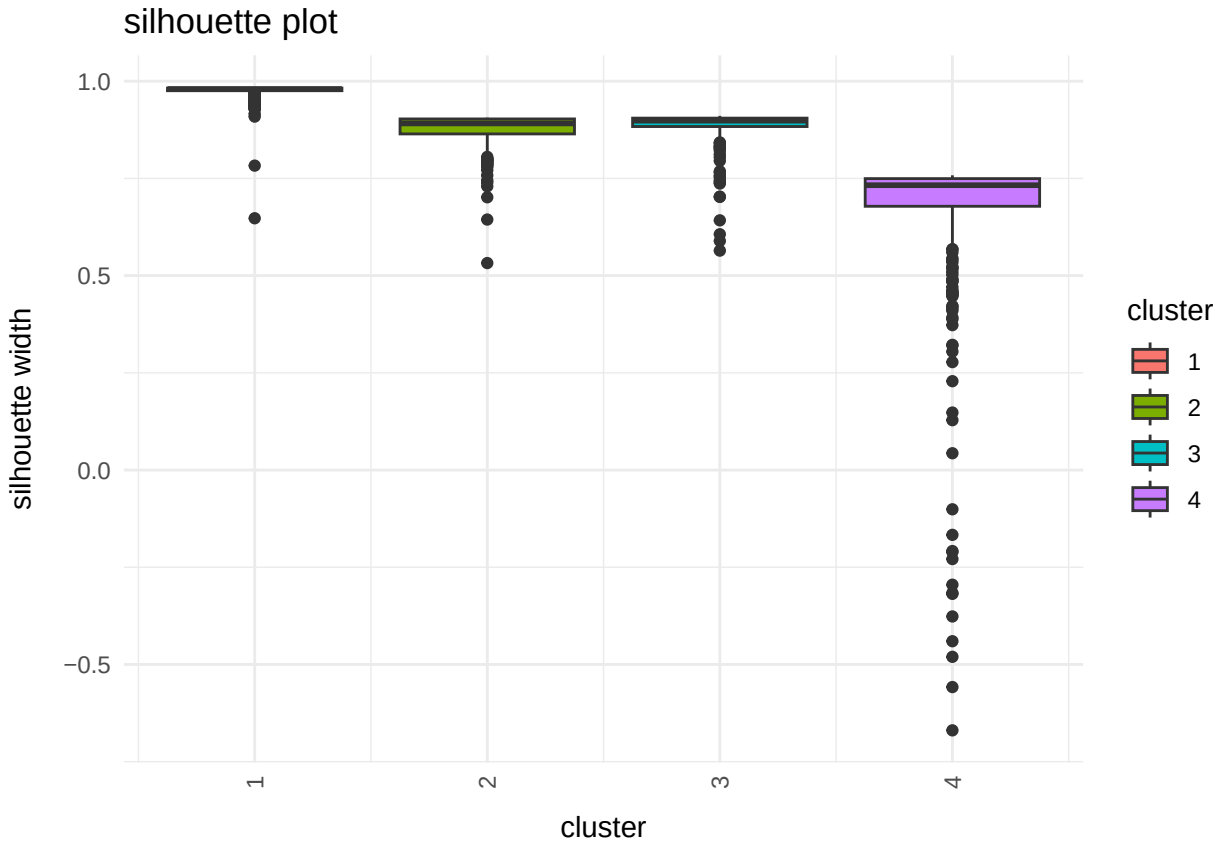


You can also summarise the results, which will give you cluster centers, group sizes and silhouette coefficients for each group

```
result_summary<-polytect_summary(result)
print(result_summary)
#>   cluster mean_channel1 mean_channel2 cluster_size silhouette_coef
#> 1       1      155.2730      1249.602        17342            0.98
#> 2       2     1330.4427      1216.940          259            0.87
#> 3       3      122.9398     10046.944          291            0.88
#> 4       4      988.5556     10044.658          341            0.65
```

You can also plot the individual silhouette coefficient in each cluster

```
sil_plot(result)
```



There is a function to calculate the concentration of the targets.

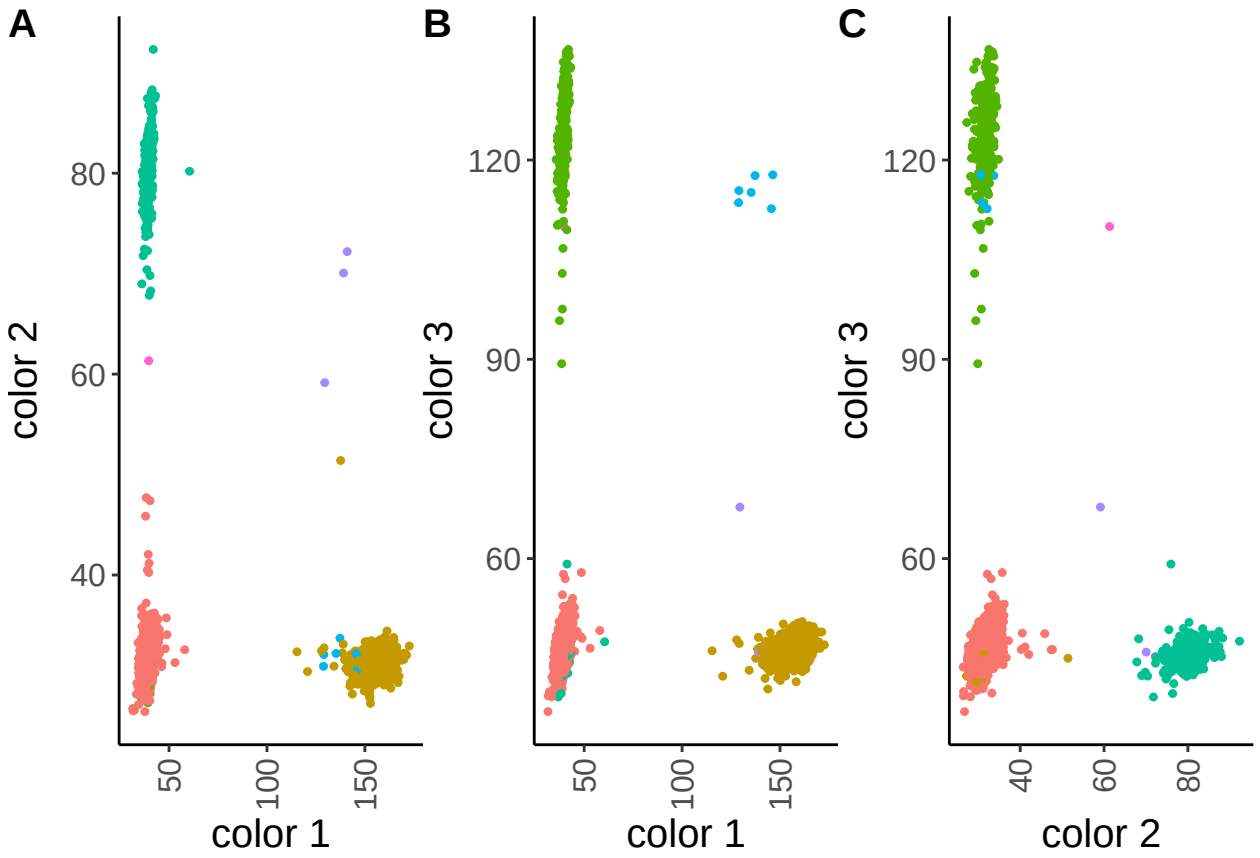
```
target_conc<-conc_cal(result,sampvol=0.91,volmix=20,voltemp=20,type="2color")
print(target_conc)
#>   target concentration
#> 1      1      36.77032
#> 2      2      36.77032
```

This package can also handle 3-up to 6-color dPCR data. We first perform flowPeaks only.

```
data(BPV)
data_scaled<-apply(BPV,2,function(x) (x-min(x))/(max(x)-min(x)))
data_input<-as.matrix(data_scaled)
fp<-flowPeaks(data_input)
#>         step 0, set the intial seeds, tot.wss=1.00728
#>         step 1, do the rough EM, tot.wss=0.721981 at 0.19157 sec
#>         step 2, do the fine transfer of Hartigan-Wong Algorithm
#>                  tot.wss=0.708295 at 0.316023 sec
```

```
table(fp$peaks.cluster)
#>
#>     1     2     3     4     5     6     7
#> 24401   482   323   245     6     3     1
```

```
df_data<-as.data.frame(cbind(BPV,cluster=fp$peaks.cluster))
polytect_plot(df_data)
```

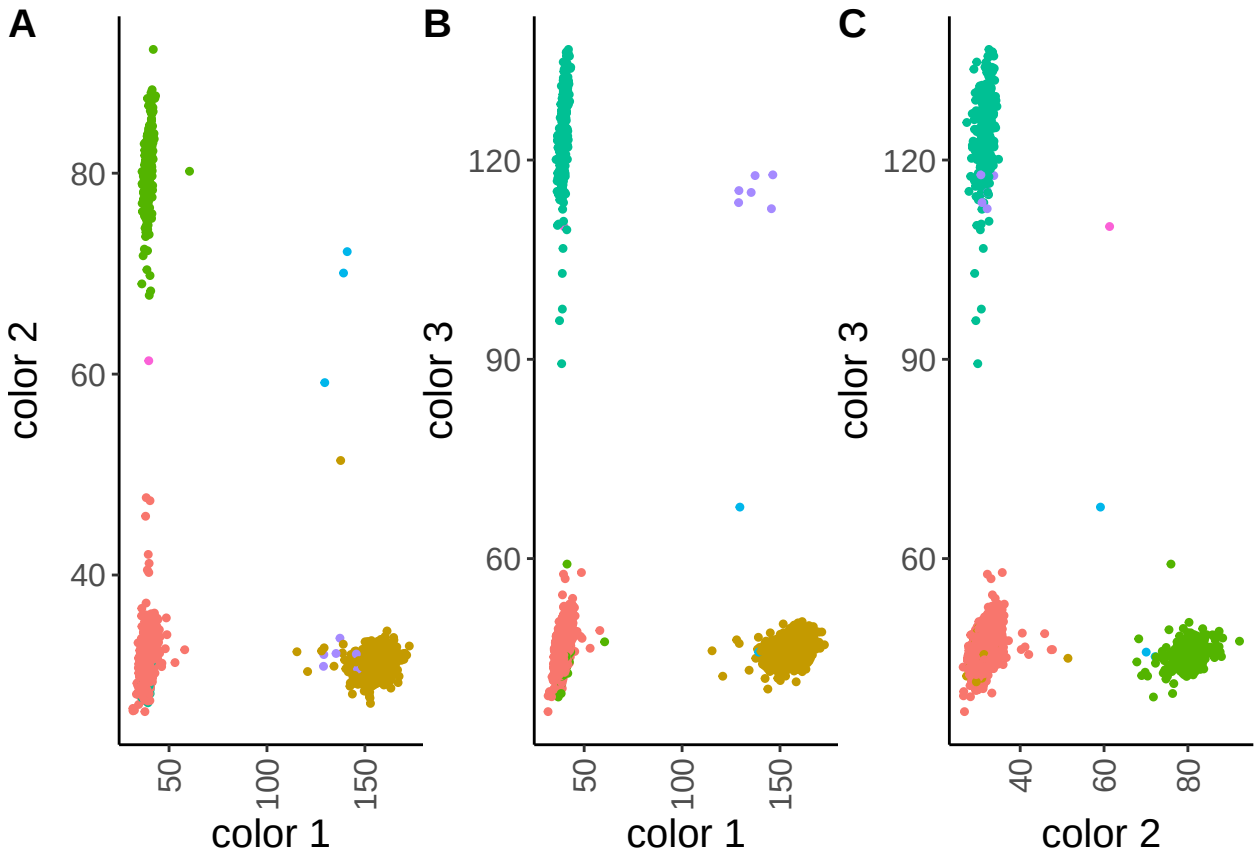


Then the main function.

```
result<-polytect_clust(data=BPV,cluster_num=8,type="3color")
#>          step 0, set the intial seeds, tot.wss=1.00728
#>          step 1, do the rough EM, tot.wss=0.721981 at 0.190087 sec
#>          step 2, do the fine transfer of Hartigan-Wong Algorithm
#>                    tot.wss=0.708295 at 0.312727 sec
```

```
table(result$cluster)
#>
#>     1     2     3     4     5     6     7
#> 24401   482   245   323     3     6     1
```

```
polytect_plot(result)
```



Polytect does not change the clustering result of flowPeaks. What it did is relabeling.

Session Information

The following information was obtained from the R session that generated this vignette:

```
sessionInfo()
#> R version 4.2.2 (2022-10-31)
#> Platform: x86_64-apple-darwin17.0 (64-bit)
#> Running under: macOS Big Sur ... 10.16
#>
#> Matrix products: default
#> BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
#> LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
#>
#> locale:
#> [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#> [1] flowPeaks_1.44.0 ggplot2_3.5.1    Polytect_0.99.0
#>
#> loaded via a namespace (and not attached):
#>  [1] tinytex_0.51        tidyselect_1.2.1    xfun_0.44
#>  [4] DiceKriging_1.6.0   splines_4.2.2       lattice_0.22-5
#>  [7] colorspace_2.1-0    vctrs_0.6.5         generics_0.1.3
```

```
#> [10] stats4_4.2.2         htmltools_0.5.8.1   yaml_2.3.8
#> [13] utf8_1.2.4           survival_3.5-7      rlang_1.1.4
#> [16] pillar_1.9.0         withr_3.0.0         sn_2.1.1
#> [19] glue_1.7.0           tidyverse_2.0.0     lhs_1.1.6
#> [22] lifecycle_1.0.4      munsell_0.5.1       gtable_0.3.5
#> [25] mvtnorm_1.2-4        evaluate_0.24.0     labeling_0.4.3
#> [28] knitr_1.47           parallelMap_1.5.1   fastmap_1.2.0
#> [31] parallel_4.2.2       fansi_1.0.6         highr_0.11
#> [34] Rcpp_1.0.12          backports_1.5.0     scales_1.3.0
#> [37] checkmate_2.3.1      farver_2.1.2        fastmatch_1.1-4
#> [40] mnormt_2.1.1         digest_0.6.35       stringi_1.8.3
#> [43] BBmisc_1.13          dplyr_1.1.4         numDeriv_2016.8-1.1
#> [46] cowplot_1.1.3        grid_4.2.2          cli_3.6.2
#> [49] tools_4.2.2          magrittr_2.0.3      tibble_3.2.1
#> [52] mlr_2.19.2           mlrMBO_1.1.5.1      pkgconfig_2.0.3
#> [55] Matrix_1.6-4         data.table_1.15.4   smoof_1.6.0.3
#> [58] ParamHelpers_1.14.1 rmarkdown_2.27      rstudioapi_0.16.0
#> [61] R6_2.5.1             compiler_4.2.2
```

References

Ge Y, Sealfon S (2012). "flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding." Bioinformatics. R package version 4.4.0.

Lun A (2024). bluster: Clustering Algorithms for Bioconductor. R package version 1.14.0.