# Tutorial for R package scDeconv

Yu Liu

12/12/2021

## Introduction

Many DNA methylation (DNAm) data are generated from tissues composed of various cell types and hence cell deconvolution methods are needed to infer their cell compositions. However, a bottleneck for DNAm data is the lack of cell-type-specific DNAm references. On the other hand, scRNA-seq data are being accumulated rapidly with various cell type transcriptomic signatures characterized, and also many paired bulk RNA-DNAm data are publicly available currently. Hence, the R package *scDeconv* was developed to use these resources to solve the reference deficiency problem of DNAm data and deconvolve them from scRNA-seq data in a trans-omics manner. It assumes that paired samples have similar cell compositions and so the cell content information deconvolved from the scRNA-seq and paired RNA data can be transferred to the paired DNAm samples, and then an ensemble model is trained to predict these cell contents with DNAm features, and also adjust the paired RNA deconvolution in a co-training manner. This tutorial will introduce the main functions of *scDeconv*.

## Package installation

Code of *scDeconv* is freely available at https://github.com/yuabrahamliu/scDeconv.

The following commands can be used to install this R package.

```
library(devtools)

install_github('yuabrahamliu/scDeconv')
```

## Data preparation

We will use the data that accompany with *scDeconv* package in this tutorial. They contain a Seurat object generated from scRNA-seq data and preprocessed by the R package *Seurat*. It is a small subset of a human placenta scRNA-seq dataset in ArrayExpress, with experiment code E-MTAB-6701 (droplet-based data), and covers 1388 cells and 21737 genes. We will deconvolve 4 main placental cell types in it, including extravillous trophoblasts (EVTs), fibroblasts (FBs), Hofbauer cells (HBs), and villous cytotrophoblast (VCTs). Among them, EVT and VCT are epithelial trophoblasts with similar origins, while HB cells are fetal macrophages present in placenta. The cell type information is contained in the meta data of the Seurat object. In addition, it also has the read count data and normlized data. We will use this scRNA-seq dataset to generate a RNA deconvolution reference via *scDeconv*.

While the data need to be deconvolved are bulk DNAm data with 311 human placenta samples and 18262 probes. They are collected from 9 different GEO datasets based on the platforms of Illumina 27K and 450K, and have gone through preprocessing with batch difference adjusted, and the shared high-quality probes retained.

To deconvolve these DNAm data with the scRNA-seq data, *scDeconv* needs a paired bulk RNA-bulk DNAm dataset to fulfill the trans-omics deconvolution, and it is also in the accompanied data, with 48 human placenta samples. Its RNA part contains 22188 genes and its DNAm part contains 18626 probes, same as the ones in the DNAm data to be deconvolved, and the batch difference between them have been adjusted with the *ComBat* function in the R package *sva*, using this paired DNAm set as the reference batch.

The paired RNA data are from the platform Affymetrix Human Gene 1.0 ST Array, and the gene expression values are library size normalized values with log2 transformation. While that in the DNAm datasets are beta values. The samples in both the paired dataset and the external DNAm dataset to be deconvolved can be divided into 2 groups. One is the normal sample group, and the other is the disease group with preeclampsia pregnancy complication. This information, as well as the original GEO datasets of the samples, can be found in the meta data frame coupled with this package.

Now, attach *scDeconv* to the R session and take a look at these data.

```
library(scDeconv)

scRNA <- system.file('extdata', 'scRNAseqdat.rds', package = 'scDeconv')
scRNA <- readRDS(scRNA)

pRNA <- system.file('extdata', 'pairedRNAdat.rds', package = 'scDeconv')
pRNA <- readRDS(pRNA)

pDNAm <- system.file('extdata', 'pairedDNAmdat.rds', package = 'scDeconv')
pDNAm <- readRDS(pDNAm)

externalDNAm <- system.file('extdata', 'externalDNAmdat.rds', package = 'scDeconv')
externalDNAm <- readRDS(externalDNAm)

DNAmpd <- system.file('extdata', 'DNAmpd.rds', package = 'scDeconv')
DNAmpd <- readRDS(DNAmpd)
```

The summary or beginning parts of these data are shown below.

```
#The scRNA-seq data
scRNA
#> An object of class Seurat
#> 21737 features across 1388 samples within 1 assay
#> Active assay: RNA (21737 features, 0 variable features)
head(scRNA@meta.data)
#>                              orig.ident nCount_RNA nFeature_RNA Fetus location
#> FCA7196226_CATGCCTGTCCCTTGT FCA7196226      23559         4736   F27 Placenta
#> FCA7196226_AGCGGTCAGCTGCAAG FCA7196226       8203         3194   F27 Placenta
#> FCA7196226_AATCCAGCATTGGCGC FCA7196226      18224         3602   F27 Placenta
#> FCA7196226_GAACATCTCTTGTACT FCA7196226      18266         4077   F27 Placenta
#> FCA7196226_AGATCTGGTTGCCTCT FCA7196226      17689         4169   F27 Placenta
#> FCA7196226_CTCGAGGAGGGCACTA FCA7196226       6314         2796   F27 Placenta
#>                              annotation
#> FCA7196226_CATGCCTGTCCCTTGT        VCT
#> FCA7196226_AGCGGTCAGCTGCAAG        VCT
#> FCA7196226_AATCCAGCATTGGCGC        VCT
#> FCA7196226_GAACATCTCTTGTACT        VCT
#> FCA7196226_AGATCTGGTTGCCTCT        VCT
#> FCA7196226_CTCGAGGAGGGCACTA        VCT
```

```
#The paired RNA microarray data
pRNA[1:6,1:6]
#>         GSM1940495 GSM1940496 GSM1940499 GSM1940500 GSM1940501 GSM1940502
#> A1BG      7.376765   7.413560   7.285113   7.352853   7.488182   7.300325
#> A1CF      5.522374   5.537723   5.463703   5.671208   5.581101   5.413397
#> A2M      11.622694  11.120252  11.230030  11.327457  11.541988  11.442168
#> A2ML1     5.994014   6.101948   5.982931   6.033332   6.075942   6.036454
#> A3GALT2   6.150027   6.216505   6.046596   6.320198   6.233631   6.119112
#> A4GALT    7.752668   8.059492   7.935501   7.676856   7.830030   7.920985
```

```
#The paired DNAm data
pDNAm[1:6,1:6]
#>            GSM1940495 GSM1940496 GSM1940499 GSM1940500 GSM1940501 GSM1940502
#> cg00000292 0.73358449 0.59264315 0.59496274 0.63138095 0.66699830 0.67046169
#> cg00002426 0.69265004 0.51667233 0.47864586 0.45470146 0.51774379 0.59661450
#> cg00003994 0.16837433 0.12105939 0.22729146 0.22437398 0.32077141 0.24005318
#> cg00007981 0.04251201 0.02582752 0.03510695 0.03843307 0.04369758 0.03930951
#> cg00008493 0.41044308 0.32724144 0.39277370 0.30326993 0.48137703 0.48915648
#> cg00008713 0.05892651 0.04065540 0.06138317 0.08263494 0.09078148 0.08007630
```

```
#The external DNAm data to be deconvolved
externalDNAm[1:6,1:6]
#>             GSM788417  GSM788419  GSM788420  GSM788421  GSM788414  GSM788415
#> cg00000292 0.65961366 0.67591141 0.65709651 0.66077820 0.66847653 0.67436406
#> cg00002426 0.53516824 0.53883284 0.53683120 0.53990206 0.53804637 0.53543344
#> cg00003994 0.17674229 0.16432771 0.16631494 0.16803355 0.16416337 0.16852233
#> cg00007981 0.03553198 0.02934814 0.03189098 0.02765783 0.02768899 0.02764183
#> cg00008493 0.43619912 0.42998023 0.43761396 0.44973257 0.46281094 0.45935308
#> cg00008713 0.07431056 0.06367040 0.06764089 0.06218142 0.05267028 0.05397539
```

```
#The meta data for the paired samples
head(subset(DNAmpd, type == 'paired'))
#>     sampleid  Samplegroup Gestwk  dataset   type
#> 1 GSM1940495 Preeclampsia     37 GSE98224 paired
#> 2 GSM1940496 Preeclampsia     29 GSE98224 paired
#> 3 GSM1940499 Preeclampsia     35 GSE98224 paired
#> 4 GSM1940500 Preeclampsia     31 GSE98224 paired
#> 5 GSM1940501 Preeclampsia     29 GSE98224 paired
#> 6 GSM1940502 Preeclampsia     37 GSE98224 paired
table(subset(DNAmpd, type == 'paired')$Samplegroup)
#>
#>      Control Preeclampsia
#>           18           30
#The meta data for the external samples to be deconvolved
head(subset(DNAmpd, type == 'external'))
#>     sampleid Samplegroup Gestwk  dataset     type
#> 49 GSM788417     Control      8 GSE31781 external
#> 50 GSM788419     Control      8 GSE31781 external
#> 51 GSM788420     Control      8 GSE31781 external
#> 52 GSM788421     Control      9 GSE31781 external
#> 53 GSM788414     Control     12 GSE31781 external
#> 54 GSM788415     Control     12 GSE31781 external
table(subset(DNAmpd, type == 'external')$Samplegroup)
```

```
#>
#>      Control Preeclampsia
#>          240           71
```

## RNA reference generation

We will use the scRNA-seq data and the paired RNA data to construct a RNA deconvolution reference first. This can be achieved via the function `scRef` in the package.

We provide the scRNA-seq data to `scRef` via its parameter `Seuratobj`, and set another parameter `targetcelltypes` as `c('EVT', 'FB', 'HB', 'VCT')`, meaning these 4 cell types in the scRNA-seq data will be covered to generate the reference. The parameter `celltypecolname` is set as "annotation", indicating the cell type information for each single cell is stored in the column "annotation" of the scRNA-seq data.

Because the first step of reference making is to synthesize several pseudo-bulk RNA-seq samples for each cell type from the scRNA-seq data, the parameter `pseudobulknum` is used to set how many such samples will be made for each cell type, and we set it as 100 here, meaning each cell type will get 100 pseudo-bulk RNA samples via sampling from the scRNA-seq data, and the 4 cell types will totally get 100*4 = 400 such samples.

Then, the synthesized samples will go through several steps to get the final reference, while if the bulk RNA data need to be deconvolved (the paired RNA data `pRNA` here) is provided to the parameter `targetdat`, a batch adjustment step will be included to remove the batch difference or platform difference between the scRNA-seq and the bulk RNA data. While if no data is provided to it, this step will be skipped, but other processing will still be performed to get the reference. Because the values in `pRNA` are log2 transformed values, we set the parameter `targetlogged` as TRUE.

```
refres <- scRef(Seuratobj = scRNA,
                targetcelltypes = c('EVT', 'FB', 'HB', 'VCT'),
                celltypecolname = 'annotation',
                pseudobulknum = 100,

                targetdat = pRNA,
                targetlogged = TRUE)
```

The result `refres` is a list containing 2 slots. The one named "ref" is the RNA reference generated.

```
head(refres$ref)
#>               EVT        FB        HB       VCT
#> ZNF490 15.350183 14.510841 14.510841 14.510841
#> CP     21.066156 20.387364 20.387364 20.387364
#> ABCC2  14.769732 14.266494 14.266494 14.266494
#> SULF1  12.160773 12.160773 16.972445 12.160773
#> PADI1  26.177981 10.876944  6.304278  4.723937
#> TLR8    8.157669  7.737533 25.150155  8.579287
```

The other slot named "targetnolog" is the adjusted bulk RNA data to be deconvolved, and the values in it are non-log transformed values

```
refres$targetnolog[1:6,1:6]
#>        GSM1940495 GSM1940496 GSM1940499 GSM1940500 GSM1940501 GSM1940502
#> ZNF490  18.416815  13.522850  14.930505  10.946805  13.748425   8.916959
#> CP       0.000000   0.000000 149.774824   0.000000   0.000000   0.000000
```

```
#> ABCC2   14.707393   7.098110   7.169431  16.389123  12.370262  10.784374
#> SULF1    9.567271  11.716441  34.076833   2.233104   9.400352   3.464205
#> PADI1    3.824240   7.766002  10.525161  15.920500  18.052632   3.175283
#> TLR8     7.326088  14.035631  16.531961   4.667870   4.065535  15.483962
```

## Bulk DNAm data deconvolution with RNA reference

After getting the RNA reference, we use it to deconvolve the external DNAm data `externalDNAm` via the function `epDeconv`. It also needs the paired bulk RNA-bulk DNAm dataset and for the RNA part, it is the adjusted bulk RNA data returned by `scRef`, and we provide it to the parameter `rnamat`, while for the DNAm part, we provide the data `pDNAm` to `methylmat`. Because the values in the adjusted RNA data are non-log transformed values, we set the parameter `rnamatlogged` as FALSE.

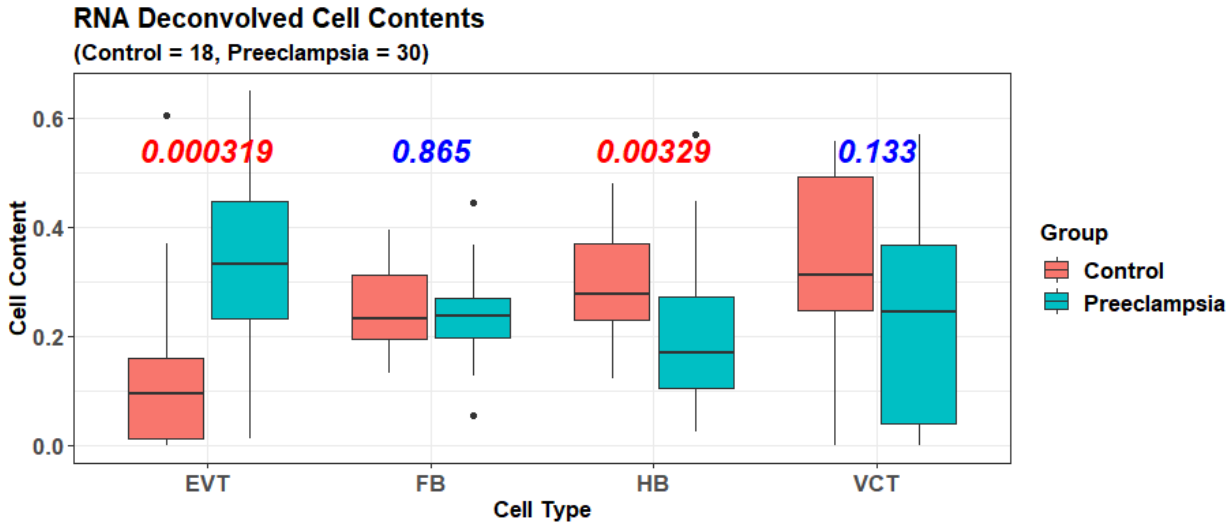
To deconvolve `externalDNAm` with the RNA reference `refres$ref`, `epDeconv` will construct an ensemble model on the paired RNA and DNAm data in a co-training manner, and then use this model to predict the cell contents for `externalDNAm`. For the number of base learners of the ensemble, it is defined by the parameter `learnernum` and we set it as 10 here. While because we want the 4 cell contents deconvolved can have a sum of 1 for each DNAm sample, we set the parameter `resscale` as TRUE.

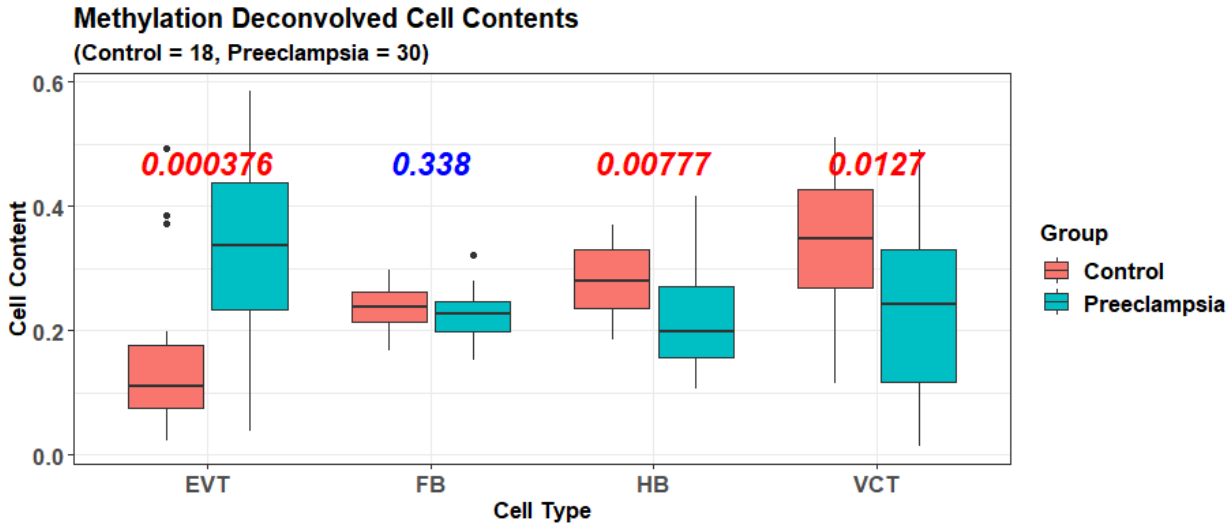
If we want to have box plots to show the deconvolution results for the paired RNA, paired DNAm, and external DNAm data, we can set the parameter `plot` as TRUE, and also provide the meta data frame of the paired samples to `pddat`, and that of the external samples to `targetmethylpddat`, so that the sample group information can be transferred to the function and the cell content difference can also be shown in the plot.

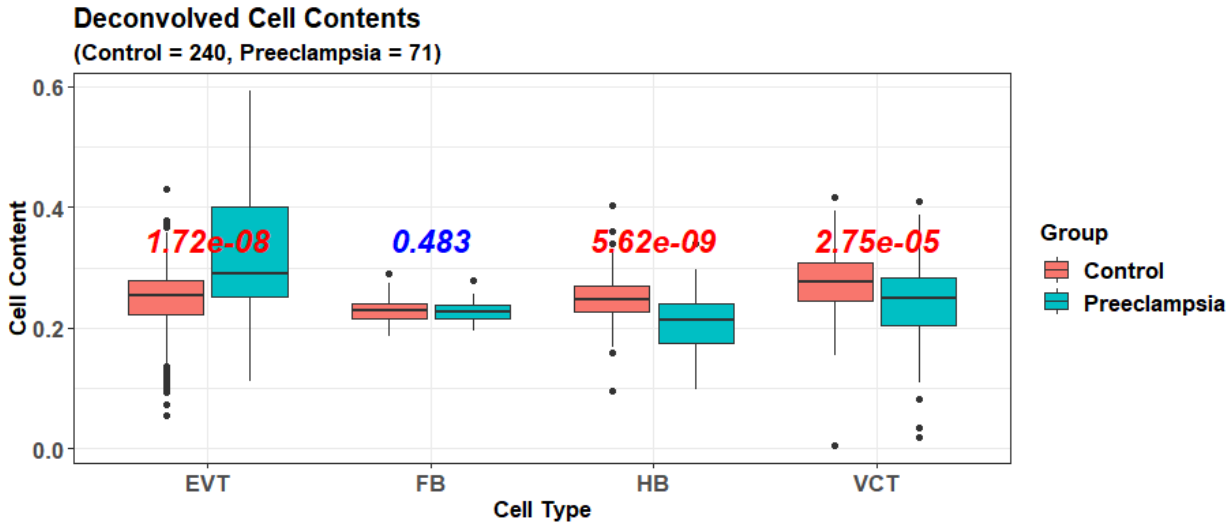
```
dnamres <- epDeconv(rnaref = refres$ref,
                    rnamat = refres$targetnolog,
                    rnamatlogged = FALSE,

                    methylmat = pDNAm,
                    learnernum = 10,
                    resscale = TRUE,

                    targetmethyldat = externalDNAm,
                    plot = TRUE,
                    pddat = subset(DNAmpd, type == 'paired'),
                    targetmethylpddat = subset(DNAmpd, type == 'external'))
```

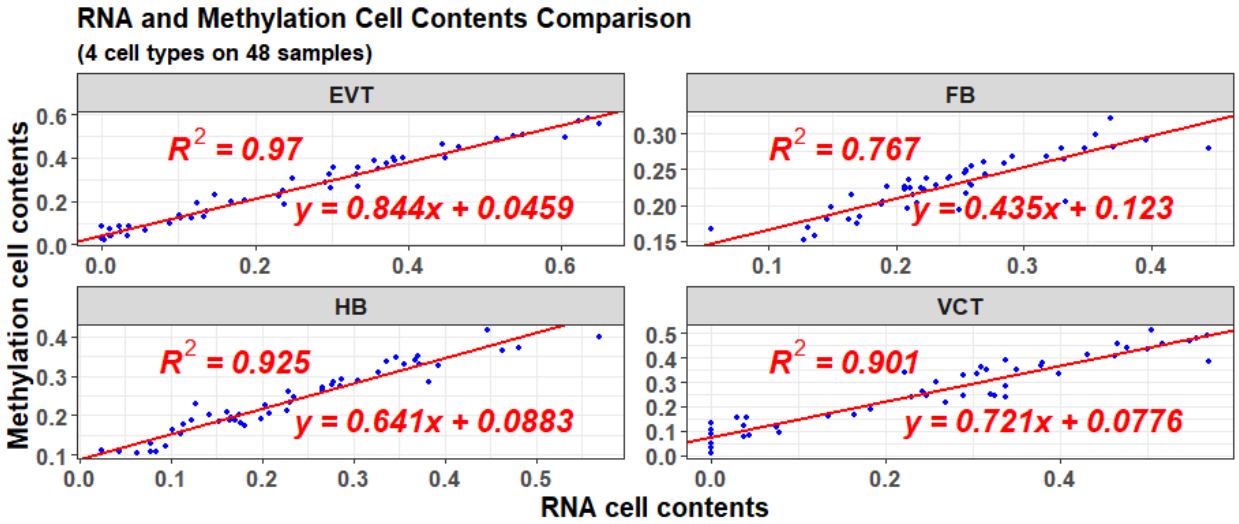





From the box plots, we can see that in all the 3 datasets, preeclampsia samples have a much higher EVT cell content than normal samples, while their HB and VCT cells are largely reduced.

In addition, 4 scatter plots are also generated for the cell types as `plot` is set as TRUE, and they compare the deconvolution results between the paired RNA data and the paired DNAm data. Because `epDeconv` constructs the model based on the assumption that paired samples have similar cell compositions, the results predicted by it shows a high correlation between the RNA and DNAm data.




The concrete values of the deconvolution results can be seen from the result `dnamres`. Its slots "rnacellconts" and "methylcellconts" contain the results for the paired data, while the slot "methyltargetcellconts" contains the results for the external DNAm data.

```
#Result for the paired RNA microarray data
head(dnamres$rnacellconts)
#>                  EVT        FB          HB        VCT
#> GSM1940495 0.33367213 0.1931336 0.22831875 0.2465819
#> GSM1940496 0.33135710 0.2586181 0.16554716 0.2454366
#> GSM1940499 0.39360023 0.2416980 0.37026306 0.0000000
#> GSM1940500 0.38089687 0.2550985 0.02477827 0.3369090
#> GSM1940501 0.24784508 0.3179029 0.11330019 0.3205049
#> GSM1940502 0.05616103 0.2068742 0.35558339 0.3797657
```

```
#Result for the paired DNAm data
head(dnamres$methylcellconts)
#>                  EVT        FB        HB        VCT
#> GSM1940495 0.26947568 0.2267247 0.2616417 0.24215789
#> GSM1940496 0.32627289 0.2291105 0.1934914 0.25112522
#> GSM1940499 0.39789689 0.2396586 0.3487757 0.01366883
#> GSM1940500 0.40208801 0.2484923 0.1127943 0.23662538
#> GSM1940501 0.30573949 0.2670643 0.1772920 0.24990424
#> GSM1940502 0.06658484 0.2262086 0.3310046 0.37620193
```

```
#Result for the external DNAm data
head(dnamres$methyltargetcellcounts)
#>                 EVT        FB        HB        VCT
#> GSM788417 0.2817118 0.2298730 0.2091309 0.2792843
#> GSM788419 0.3122537 0.2109420 0.2037660 0.2730383
#> GSM788420 0.2960548 0.2214043 0.2059084 0.2766325
#> GSM788421 0.2852733 0.2101249 0.1959775 0.3086243
#> GSM788414 0.2978430 0.2068120 0.1911144 0.3042306
#> GSM788415 0.3099221 0.2017879 0.1769487 0.3113413
```

While the trained ensemble model is also in `dnamres`, and its slot "modellist" is the base learners of the model, while "normweights" is the base learner weights for the ensemble.

If `externalDNAm` is not provided to `epDeconv`, it will not influence the ensemble model training on the paired data, and the same model can still be returned. Then, `externalDNAm` can be transferred together with the `epDeconv` result to the function `methylpredict` to predict the external sample cell contents.

```
externalcellconts <- methylpredict(model = dnamres,
                                   targetmethyldat = externalDNAm,
                                   resscale = TRUE,
                                   adjustminus = TRUE)

head(externalcellconts)
#>                  EVT        FB        HB       VCT
#> GSM788417 0.2817118 0.2298730 0.2091309 0.2792843
#> GSM788419 0.3122537 0.2109420 0.2037660 0.2730383
#> GSM788420 0.2960548 0.2214043 0.2059084 0.2766325
#> GSM788421 0.2852733 0.2101249 0.1959775 0.3086243
#> GSM788414 0.2978430 0.2068120 0.1911144 0.3042306
#> GSM788415 0.3099221 0.2017879 0.1769487 0.3113413
```

In addition, *scDeconv* also contains other useful functions such as *refDeconv* to deconvolve bulk data using reference from the same omics (single-omics deconvolution), and *celldiff* to select cell-type-specific inter-group differential features from bulk data, and *enrichwrapper* to annotate differential DNAm feature function using a correlation-based method, etc. They will not be covered by this tutorial to make it clearer, but the users can explore them via the help documents.