Highly variable genes documentation

Franziska Heinkele

9 Juni 2019

Load data:

```
library(readr)
  Untreated = readRDS(paste0(wd,"/data/NCI_TPW_gep_untreated.rds"))
  Treated = readRDS(paste0(wd,"/data/NCI_TPW_gep_treated.rds"))
  Basal = readRDS(paste0(wd,"/data/CCLE_basalexpression.rds"))
  Copynumber = readRDS(paste0(wd,"/data/CCLE_copynumber.rds"))
  Mutations = readRDS(paste0(wd,"/data/CCLE_mutations.rds"))
  Sensitivity = readRDS(paste0(wd,"/data/NegLogGI50.rds"))
  Drug_annotation = read_tsv(paste0(wd,"/data/drug_annotation.tsv"))
  Cellline_annotation = read_tsv(paste0(wd,"/data/cellline_annotation.tsv"))
  Metadata = read_tsv(paste0(wd,"/data/NCI_TPW_metadata.tsv"))
```

BROAD ANALYSIS

HIGHLY VARIABLE GENES

We aim to find genes which vary the most in Untreated and Treated celllines. Therefore we work with Seurat packages, which need to be installed: install.packages('Seurat') For the seurat library to be available, digest packages need to be installed: install.packages("digest")

```
library(Seurat)
```

In the following code, we use the abbreviation **"BroadAnU"** representative for "Broad analysis untreated" to achieve a better readability. First we create a new object that contains the Untreated data. Then 2000 most variable genes are selected from the Untreated data.

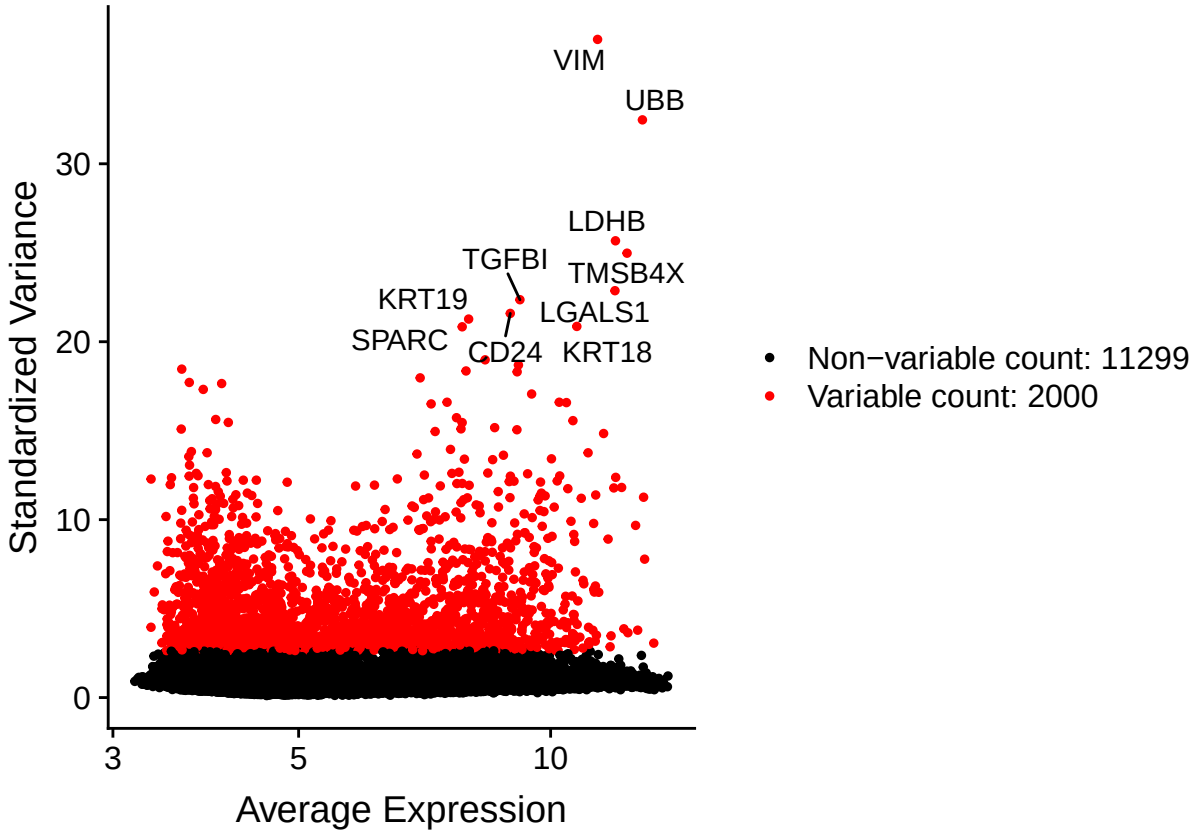
```
BroadAnU <- CreateSeuratObject(counts = Untreated, project = "BroadAnU")
BroadAnU <- FindVariableFeatures(BroadAnU, selection.method = "vst", nfeatures= 2000)
```

Next we want to identify the 10 most highly variable genes.

```
top10Untreated <- head(VariableFeatures(BroadAnU), 10)
```

First we plot variable features without labels, in a second step we add labels to the 10 most variable genes.

```
plotUntreated <- VariableFeaturePlot(BroadAnU)
plotUntreatedlabeled <- LabelPoints(plot = plotUntreated, points = top10Untreated,
                                    repel = TRUE, xnudge = 0, ynudge = 0)
plotUntreatedlabeled
```



For a comparison, we do the same procedure with the Treated data:

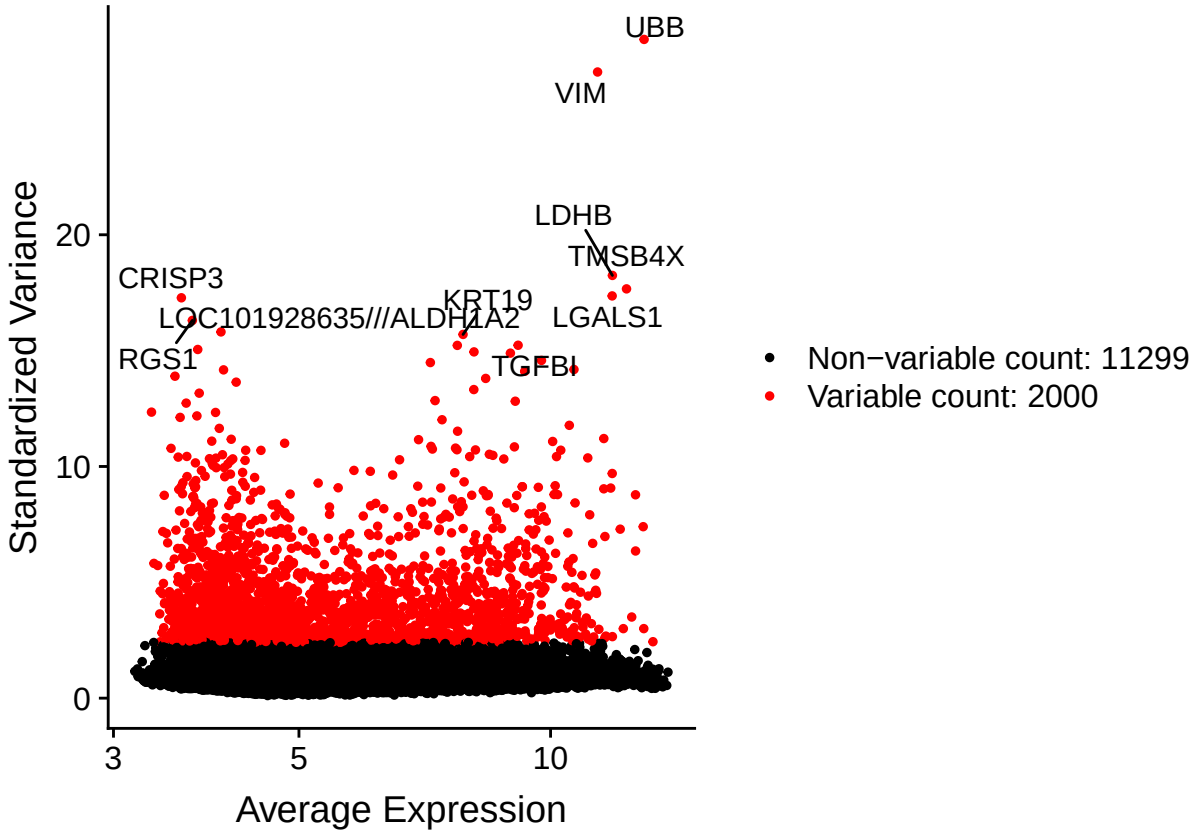
```
BroadAnT <- CreateSeuratObject(counts = Treated, project = "BroadAnT")
BroadAnT <- FindVariableFeatures(BroadAnT, selection.method = "vst", nfeatures= 2000)
```

Identify the 10 most highly variable genes:

```
top10Treated <- head(VariableFeatures(BroadAnT), 10)
```

Plot variable features with and without labels

```
plotTreated <- VariableFeaturePlot(BroadAnT)
plotTreatedlabeled <- LabelPoints(plot = plotTreated, points = top10Treated,
                                  repel = TRUE, xnudge = 0, ynudge = 0)
plotTreatedlabeled
```



Next we test, whether there are matches between Treated and Untreated most variable gene expression:

```
top10Untreated %in% top10Treated
```

```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE
```

7 matches were found! Which genes are in the Untreated but not in the Treated top 10?

```
setdiff(top10Untreated, top10Treated)
```

```
## [1] "CD24"  "KRT18" "SPARC"
```

Which genes are in the Treated but not in the Untreated top 10?

```
setdiff(top10Treated, top10Untreated)
```

```
## [1] "CRISP3"                  "RGS1"
## [3] "LOC101928635///ALDH1A2"
```

We can conclude, that the treatment did not change the high variation for the matching genes. The matching genes are probably not clearly affected by the drugs since their expression varies in treated celllines as well as in unterated celllines.