# Package 'EZscRNA'

July 3, 2019

**Title** Simple Convenience Functions for High-Level scRNA-Seq Analysis

**Version** 0.0.0.9000

**Description** This package serves as a (hopefully) convenient wrapper for EZ(-er) scRNA-seq analysis using best practices. It uses Seurat (v3+), sctransform, and Monocle (v3+) for most analysis steps. It provides a number of high-level functions spanning QC/preprocessing, exploratory data analysis, clustering, cell type inference using user-provided reference data sets, pseudotime trajectories, and visualizations.

**Depends** R (>= 3.5.1),
Seurat (>= 3.0.0),
sctransform (>= 0.2.0)

**Imports** dplyr,
ggplot2,
foreach,
doParallel,
RColorBrewer

**License** GPL-3

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 6.1.1

## R topics documented:


AddClonotype . . . . . 2
AssignCellType . . . . . 2
BatchCCEDA . . . . . 3
InferCellType . . . . . 4
NormScoreCC . . . . . 5
PrepRef . . . . . 5
RunQC . . . . . 6
RunSCT . . . . . 6
VizAnnotatedMarkers . . . . . 8
VizCellType . . . . . 8
VizGeneList . . . . . 8
VizMetaData . . . . . 8
VizVDJDist . . . . . 9

**Index** **10**

## AddClonotype *Add 10X clonotype data to seurat object*

### Description

AddClonotype adds clonotype info from matched 10X VDJ sequencing to the metadata of a given Seurat object.

### Usage

```
AddClonotype(vdj.dir, scrna)
```

### Arguments

| | |
|---|---|
| vdj.dir | String containing path to TCR (VDJ) directory. |
| scrna | Seurat object. |

### Details

This function is admittedly rough and will be rewritten in the future. It does not include specific VDJ genes for each cell, rather just using the final amino acid sequence for inter-sample comparison.

### Value

Seurat object with clonotype data (clonotype_id and cdr3s_aa) added to the metadata for each cell.

## AssignCellType *Infers and assigns cell type for each cell*

### Description

AssignCellType performs correlation-based cell inference using a user-provided reference dataset. It returns either a seurat object with lineage, cell.type, corr, and (potentially) base.lineage columns added to the metadata, or a dataframe containing the top three predicted cell types for each cell along with their correlation values. This dataframe will be saved in the output directory regardless, along with distribution statistics for inferred cell types and lineages if set.

### Usage

```
AssignCellType(scrna, dataset, outdir, lineage = FALSE, assign = TRUE,
  n.cores = 1)
```

### Arguments

| | |
|---|---|
| scrna | Seurat object. |
| dataset | Path to tab-delimited table of gene counts. First column must be gene identifiers of same type as Seurat object. Each subsequent column should contain counts for a cell type with the column header denoting the cell type. Replicates should contain an underscore followed by the replicate number (e.g. BCell_1, BCell_2, etc). |
| outdir | Path to output directory. |
| lineage | Boolean indicating whether or not column names are formatted as lineage followed by more granular specifications (e.g. Tcell.CD4_1, Tcell.CD8_1, etc). If TRUE, will assign the lineage (everything before the) first '.' in the column name to a metadata column called "base.lineage". Columns not containing a '.' will use the entire name, though replicate indicators will be removed (e.g. NK_1 will just be NK). |
| assign | Boolean indicating whether inferred cell types should actually be assigned to seurat object or just returned as a table. TRUE by default. |
| n.cores | Number of cores to use for correlation. Linearly decreases computation time. |

### Details

The reference dataset can be from any source, it should just be normalized so that columns (cell types) are comparable. The majority of this code was written by Allegra Petti - the version here has just been made more generic.

### Value

If assign is TRUE, returns a seurat object with inferred cell type information in the metadata. If FALSE, returns a dataframe of the inferred cell type/lineage information instead.

---

## BatchCCEDA *Exploratory data analysis plots*

---

### Description

BatchCCEDA creates a number of plots to determine the number of PCs to use for PCA/clustering and whether or not cell cycle scores and batch effects should be addressed. Runs and plots an ElbowPlot to determine PCs for later use. Runs and plots PCA for cell cycle genes to show their impact. PCA on variable features can also be plotted by batch to view potential batch effects.

### Usage

```
BatchCCEDA(scrna, outdir, npcs = 50, batch = FALSE)
```

### Arguments

| | |
|---|---|
| scrna | Seurat object. |
| outdir | Path to output directory for plots. |
| npcs | Number of PCs to use for PCA and ElbowPlot. 50 by default. |
| batch | Boolean indicating whether batch should be investigated. Requires batch metadata for each cell. FALSE by default. |

## Value

A Seurat object with a PCA for cell cycle genes stored with `reduction.name = "cc"`.

---

`InferCellType` *Infer cell type using reference dataset*

---

## Description

`InferType` utilizes a reference dataset to perform correlations of each cell in a seurat object with each sample in the reference dataset. It retuns a table containing the most likely cell type for each cell based on correlation values from the reference dataset.

## Usage

```
InferCellType(scrna, dataset, outdir, lineage = FALSE, n.cores = 1)
```

## Arguments

| | |
|---|---|
| `scrna` | Seurat object. |
| `dataset` | Path to tab-delimited table of gene counts. First column must be gene identifiers that match those of the Seurat object. Each subsequent column should contain counts for a cell type with the column header denoting the cell type. Replicates should contain an underscore followed by the replicate number (e.g. Bcell_1, Bcell_2, etc). |
| `outdir` | Path to output directory. |
| `lineage` | Boolean indicating whether or not column names are formatted as lineage followed by more granular specifications (e.g. Tcell.CD4_1, Tcell.CD8_1, etc). If TRUE, will assign the lineage (everything before the) first '.' in the column name to a metadata column called "base.lineage". Columns not containing a '.' will use the entire name, though replicate indicators will be removed (e.g. NK_1 will just be NK). |
| `n.cores` | Number of cores to use for correlation. Linearly decreases computation time. |

## Details

The reference dataset can be from any source, it should just be normalized so that columns (cell types) are comparable.

## Value

A table containing the top three predicted cell types for each cell.

## NormScoreCC — *Normalize counts and score cell cycle for each cell*

### Description

NormScoreCC returns a Seurat object with normalized counts and adds cell cycle scores for each gene based on Seurat's cell cycle gene lists.

### Usage

```
NormScoreCC(scrna)
```

### Arguments

| | |
|---|---|
| scrna | Seurat object to score cell cycle genes for each cell. |

### Details

The Seurat authors state (https://github.com/satijalab/seurat/issues/1679) that counts should always be normalized before cell cycle or module scoring. This is particularly important if one is using the SCTranform function for data normalization, scaling, and therefore, regression.

### Value

Seurat object with cell cycle scores (S.Score, G2M.Score) and Phase added to metadata for each cell.

## PrepRef — *Load and prepare reference dataset for cell type inference*

### Description

PrepRef processes a reference dataset of normalized gene counts for correlation analysis with a seurat object. Removes unnecessary genes from the reference dataset and matches the row ordering of the seurat object.

### Usage

```
PrepRef(scrna, dataset)
```

### Arguments

| | |
|---|---|
| scrna | Seurat object. |
| dataset | Path to tab-delimited table of gene counts. First column must be gene identifiers that match those of the Seurat object. Each subsequent column should contain counts for a cell type with the column header denoting the cell type. Replicates contain an underscore followed by the replicate number (e.g. BCell_1, BCell_2, etc). |

## Value

A list of two dataframes: "ref" - sorted genes from the reference dataset also found in the seurat object, and "sc.sub" - sorted genes from the seurat object also found in the reference dataset.

---

RunQC *Creates basic QC plots*

---

## Description

RunQC saves 3 QC plots showing gene counts, read counts, and percent mitochondrial reads per cell to help determine filters. It returns a Seurat object with percent mitochondrial reads added to the metadata.

## Usage

```
RunQC(scrna, outdir)
```

## Arguments

| | |
|---|---|
| scrna | Seurat object. |
| outdir | Path to output directory. |

## Value

Seurat object with percent mitochondrial reads added to the metadata for each cell.

---

RunSCT *Normalize, scale, and regress out wanted variation*

---

## Description

SCT runs SCTransform on a Seurat object, followed by PCA, UMAP, and clustering. Produces PCA and UMAP dimplots based on user-provided list of metadata columns to use for grouping. Also finds marker genes and saves the output as a table along with a heatmap of the top 10 upregulated genes in each cluster.

## Usage

```
RunSCT(scrna, outdir, npcs = 50, res = 0.8, min.dist = 0.3,
  n.neighbors = 30, regress = NULL, groups = NULL,
  groups.pca = NULL, groups.label = NULL, groups.legend = NULL,
  ccpca = FALSE, test = "wilcox", logfc.thresh = 0.25,
  min.pct = 0.1)
```

## Arguments

| | |
|---|---|
| scrna | Seurat object. |
| outdir | Path to output directory. |
| npcs | Number of principle components to use for UMAP and clustering. Default is 50, as SCTransform tends to do better with more. |
| res | Numeric value denoting resolution to use for clustering. Default is 0.8 (Seurat default). Increasing this value will speed up clustering, but may decrease the numbers of distinct clusters. Values of 0.5-3 are sensible. |
| min.dist | Number that controls how tighly the embedding is allowed to compress points together in RunUMAP. Increasing may be beneficial for large datasets. Default is 0.3 (Seurat default). |
| n.neighbors | Integer that determines the number of neighboring points used in local approximations of manifold structure in RunUMAP. Altering it may be beneficial for large datasets (though it isn't stated how it should be changed). Values of 5-50 are considered sensical. Default is 30 (Seurat default). |
| regress | Vector of metadata variables to regress during data scaling. Must match column headers in metadata. |
| groups | Vector of metadata variables to use for grouping in UMAP and PCA dimplots. Must match column headers in metadata. |
| groups.pca | Vector of boolean values to determine if DimPlots PCA reductions should also be created for each group. NULL by default ( only UMAP reductions will be shown for each group). If provided, must be same length as groups parameter. |
| groups.label | Vector of boolean values to determine if DimPlots should show labels for each group. NULL by default (labels will not be shown). If provided, must be same length as groups parameter. |
| groups.legend | Vector of boolean values to determine if DimPlots should show legends for each group. NULL by default (legends will be shown). If provided, must be same length as groups parameter. |
| ccpca | Boolean to indicate whether PCA using only cell cycle genes should be performed and plotted by sample identity and Phase. |
| test | Denotes which DE test to use for marker finding. Options are: "wilcox" (default), "bimod", "roc", "t", "negbinom", "poisson", "LR", "MAST", "DESeq2". |
| logfc.thresh | Value that limits DE testing to genes that show, on average, at least X-fold difference (log-scale) between two groups of cells. Increasing speeds up function at cost of potentially missing weaker signals. Default is 0.25 (Seurat default). |
| min.pct | Value that limits DE testing to genes detected in a minimum fraction of cells in either population. Default is 0.1 (Seurat default). |

## Value

A Seurat object with normalized, scaled counts and assigned clusters. If ccpca = TRUE, an additional PCA named "cc" will also be present.

## VizAnnotatedMarkers *Visualize an annotated marker list*

### Description

Visualize an annotated marker list

### Usage

```
VizAnnotatedMarkers(scrna, marker_file, outdir)
```

## VizCellType *Visualize inferred cell types*

### Description

Visualize inferred cell types

### Usage

```
VizCellType(scrna, outdir)
```

## VizGeneList *Visualize a list of genes*

### Description

Visualize a list of genes

### Usage

```
VizGeneList(scrna, genelist, outdir)
```

## VizMetaData *Visualize by metadata variables*

### Description

Visualize by metadata variables

### Usage

```
VizMetaData(scrna, vars, outdir)
```

VizVDJDist *Visualize clonotype distributions*

## Usage

```
VizVDJDist(scrna, outdir, g.by = NULL, o.by = NULL, n.clono.c = 10,
  n.clono.g = NULL)
```

## Arguments

| | |
|---|---|
| scrna | Seurat object with clonotype data added to metadata with AddClonotype. |
| outdir | Path to output directory. |
| g.by | Metadata column to group samples by. If not provided, only histograms of clonotypes will be saved. |
| o.by | Vector containing names of members of each group to sort by within the group. Ignored if g_by is NULL. Should contain one instance of each potential value in g_by column if provided. |
| n.clono.c | Number of top clonotypes to plot for comparison barchart. Default is 10. Ignored if (group_by) is NULL.} |
| | \item{n.clono.g}{Number of clonotypes to show in group-specific histograms. All are shown by default.} } { VizVDJDist visualizes clonotype distributions for each sample in a seurat object as histograms as well as barcharts comparing clonotype proportions between them. } { Rows with } |

# Index

AddClonotype, 2
AssignCellType, 2

BatchCCEDA, 3

InferCellType, 4

NormScoreCC, 5

PrepRef, 5

RunQC, 6
RunSCT, 6

VizAnnotatedMarkers, 8
VizCellType, 8
VizGeneList, 8
VizMetaData, 8
VizVDJDist, 9