ScisorWiz - Standard Workflow

```
library(ScisorWiz)
```

The *ScisorWiz* package creates a plot utilizing single-cell long-read RNA sequencing data in order to display isoform expression differentiation across multiple cell types for a single gene.

Software required

- samtools
- python version >= 3.7 with the following libraries
  - pandas

Section 1 - Main workflow - Organize data and create the plot

There are two options for running the pipeline to plot the data depending on the files available to the user:

- ScisorWiz_AllInfo
  - Used if the user has output from the scisorseqr package called the AllInfo file.
    - This method is recommended to provide more customizability and specificity to the output plots.
- ScisorWiz_2File
  - Used if the user has gff.gz file and genes.gz file filtered for detected cage and polyA peaks.
    - genes.gz file is a tab-separated file containing readID to geneID mappings and is formatted as follows:

readID	geneID
RGL:GCAGCCAGTAAATGTG:m64013_190223_004143/73663463/ccs.path1	ENSMUSG00000041571.9
OPCs:TCAGGATAGTTCGCAT:m64013_190221_020520/30540150/ccs.path1	ENSMUSG00000038717.8

Pipeline Option 1 - ScisorWiz_AllInfo

Using AllInfo file output from the scisorseqr package, the user can choose the clustering method to utilize for the data on the final plot. For that the user must specify:

- GENCODE annotation file for user data
- AllInfo file derived from scisorseqr
- Cell type file listing user-specified cell types of interest and the display color of each (example of document format below)*
- Gene of interest
- Clustering method**

- Confidence interval (CI) for alternative exon consideration. Default value for exon inclusion rate is .05 (5% < altExon inclusion < 95%)
- Output directory in which the user wants output files stored
- Optional: Mismatches file containing output from the MismatchFinder function (see Section 2). Default value is NULL

*Celltype file is tab separated with each cell type on a new line and cell type names must be written exactly as they appear in the GENCODE file. The file looks like the following:

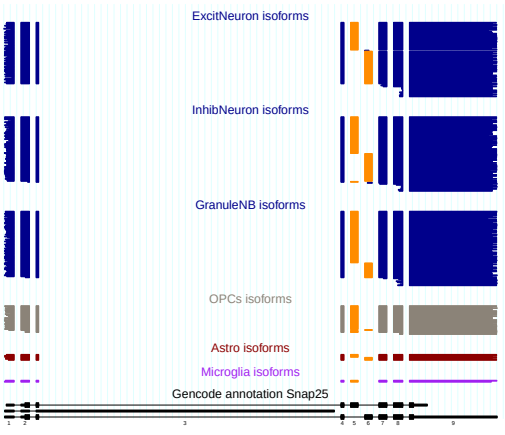
ExcitNeuron	darkblue
InhibNeuron	darkblue
GranuleNB	darkblue
OPCs	antiquewhite4
Astro	darkred
Microglia	purple

**The user can choose from one of the clustering methods by inputting the number that is next to the method as shown below:

1. Intron chain
2. TSS site
3. PolyA site
4. Intron chain, TSS site, and PolyA site

```
gencodeAnno <- system.file("extdata/", "gencode.vM21.annotation.gtf.gz",
                           package = "ScisorWiz")
allInfoFile <- system.file("extdata/", "AllInfo.gz", package = "ScisorWiz")
cTypeFile <- system.file("extdata/", "userInput/celltypeFile",
                         package = "ScisorWiz")

## Run command without plotting mismatches
ScisorWiz_AllInfo(gencodeAnno = gencodeAnnoFile, AllInfoInput = allInfoFile,
                  cellTypeFile = cTypeFile, gene = "Snap25", cluster = 1, ci = .05,
                  outputDir = "extdata/outputDir/")

##_____________________________________OR_____________________________________##

mismatches <- system.file("extdata/", "outputDir/Snap25.mismatches.txt.gz",
                          package = "ScisorWiz")

## Run command with plotting mismatches
ScisorWiz_AllInfo(gencodeAnno = gencodeAnnoFile, AllInfoInput = allInfoFile,
                  cellTypeFile = cTypeFile, gene = "Snap25", cluster = 1, ci = .05
                  outputDir = "extdata/outputDir/", mismatchFile = mismatches)
```

The output after running the data through the ScisorWiz_AllInfo function without mismatches is as follows:



NOTE The example data included in this package does not yield the plot above. This is just an example plot generated from the full dataset.

Pipeline Option 2 - ScisorWiz_2File

Without the AllInfo file from the scisorseqr pipeline, the user can still run ScisorWiz on their gff.gz and genes.gz files which are filtered for detected cage and PolyA peaks. For that the user must specify:

- GENCODE annotation file for user data
- gff.gz file containing read-specific information
- genes.gz file
- Cell type file listing user-specified cell types of interest and the display color of each
- Gene of interest
- Confidence interval (CI) for alternative exon consideration. Default value for exon inclusion rate is .05 (5% < altExon inclusion < 95%)
- Output directory in which the user wants output files stored
- Optional: Mismatches file containing output from the MismatchFinder function (see Section 2). Default value is NULL

```
## Run command without plotting mismatches
ScisorWiz_2File(gencodeAnno = "gencodeAnnoFile.gz", gffInput = "CagePolyA.gff.gz",
                genesInput = "reads2genes.gz",cellTypeFile = "cellTypeFile_Snap25.tab",
                gene = "Snap25", ci = .05, outputDir = "extdata/outputDir/")

##_____________________________________OR_____________________________________##

## Run command with plotting mismatches
ScisorWiz_2File(gencodeAnno = "gencodeAnnoFile.gz", gffInput = "CagePolyA.gff.gz",
                genesInput = "reads2genes.gz", cellTypeFile = "cellTypeFile_Snap25.tab",
                gene = "Snap25", ci = .05, outputDir = "extdata/outputDir/",
                mismatchFile = "Snap25.mismatches.txt.gz")
```

Section 2 - Optional Pre-processing

Prior to using ScisorWiz main pipeline, the user has the option to run the MismatchFinder function. MismatchFinder will specify any SNVs, insertions, or deletions in the data as compared to the reference genome. For that the user must specify:

- Sorted .bam file
- Reference .fasta file
- GENCODE annotation for the data
- Gene of interest
- Output directory in which the user wants the mismatch file stored

```
## Run command
MismatchFinder(BAM = "sorted.bestperRead.mapping.bam", fasta = "mm10.fa",
               gencodeAnno = "gencode.vM21.annotation.gtf.gz", gene = "Snap25",
               outputDir = "extdata/outputDir/")
```

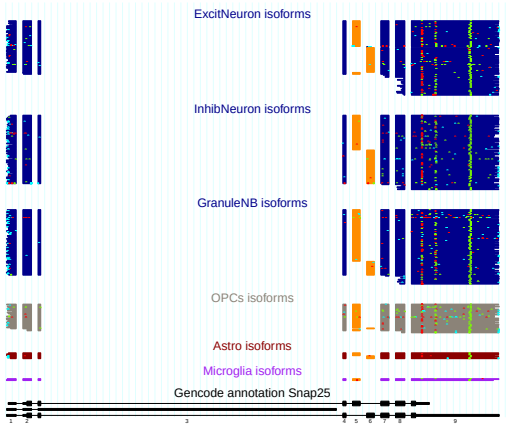
User will see the output directory with a subdirectory specifically named for the gene of interest:

- outputDir
  - Snap25
    - Snap25.info.tab (Provides chromosome, start, and end for MismatchFinder script)
    - Snap25.mismatches.txt.gz (One line per readID with following structure:)

chrom	readID	SNV
chr2	m64013_190219_195127/88146585/ccs	136713563_G|A;136764203_C|A

insertion	deletion
136781282_136781283_A;136781282_136781283_C	136781219_C;136781421_C

The output after running the data through the MismatchFinder function and then the ScisorWiz_AllInfo function is as follows:



NOTE The example data included in this package does not yield the plot above. This is just an example plot generated from the full dataset.

Done!