# Pathway Enrichment with KNIME

# Outline: https://github.com/InesAssum/EnrichmentNodes

1. Pathway Enrichment using modular R library

2. Pathway Enrichment with KNIME for Users
3. Pathway Enrichment with KNIME for Developers

# Modular R library integrating multiple Pathway Enrichment methods

# Outline: modular framework


**Expression data file (RDS, RData, txt, csv)**
- ID
- statistic
- pvalues
- probability

Optional: GMT-file (KEGG, GO etc.)

**Method:**
- **R function** (possibly command line tool)

**Parameters**
- **general**, e.g. gene set size
- **method specific**, e.g. number of permutations, cutoff

**RDS file**
**list(**
- **Table:** pathways, score, pvals, probability
- **Log file:** method, parameters
- **Method specific output object:** (e.g. incl. leadingEdge in GSEA)

**)**

**Current methods include:** MONA, GSEA (...more coming soon!)

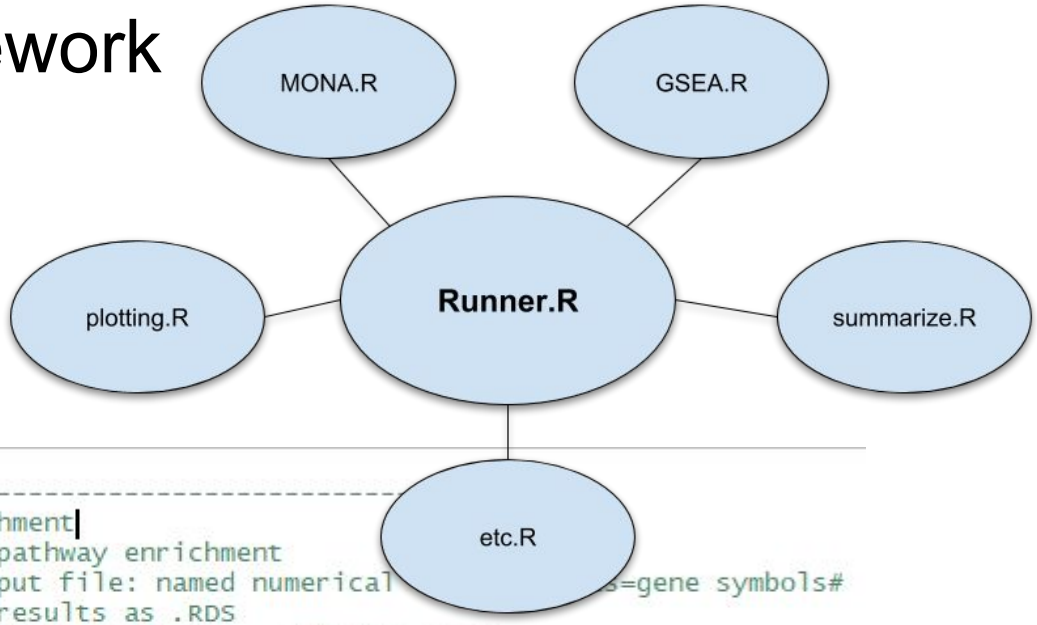
**Suited for:** - big/advanced projects (simulations, benchmarking)

- anyone, who is using R anyway

How to deal with differences in methods?

# Outline: modular framework

- full code available on GitHub
- run locally in R or
- use docker support




```
#!/usr/bin/env Rscript
# ------------------------------------------------------------------------
#' Runner script for multiOMICs pathway enrichment
#' @param method string // select method for pathway enrichment
#' @param data string // file path to .RDS input file: named numerical                s=gene symbols#
#' @param result string // file path to save results as .RDS
#' @param gmt string // pathway definition to use: KEGG, GO... default: KEGG
#' @param minSize integer // min size of GSs to be considered, default: 15
#' @param maxSize integer // max size of GSs to be considered, default: 500
#' @param nperm integer // number of permutations, default: 10000
#' @param mygmt string // custom .gmt file // not required
# ------------------------------------------------------------------------
```

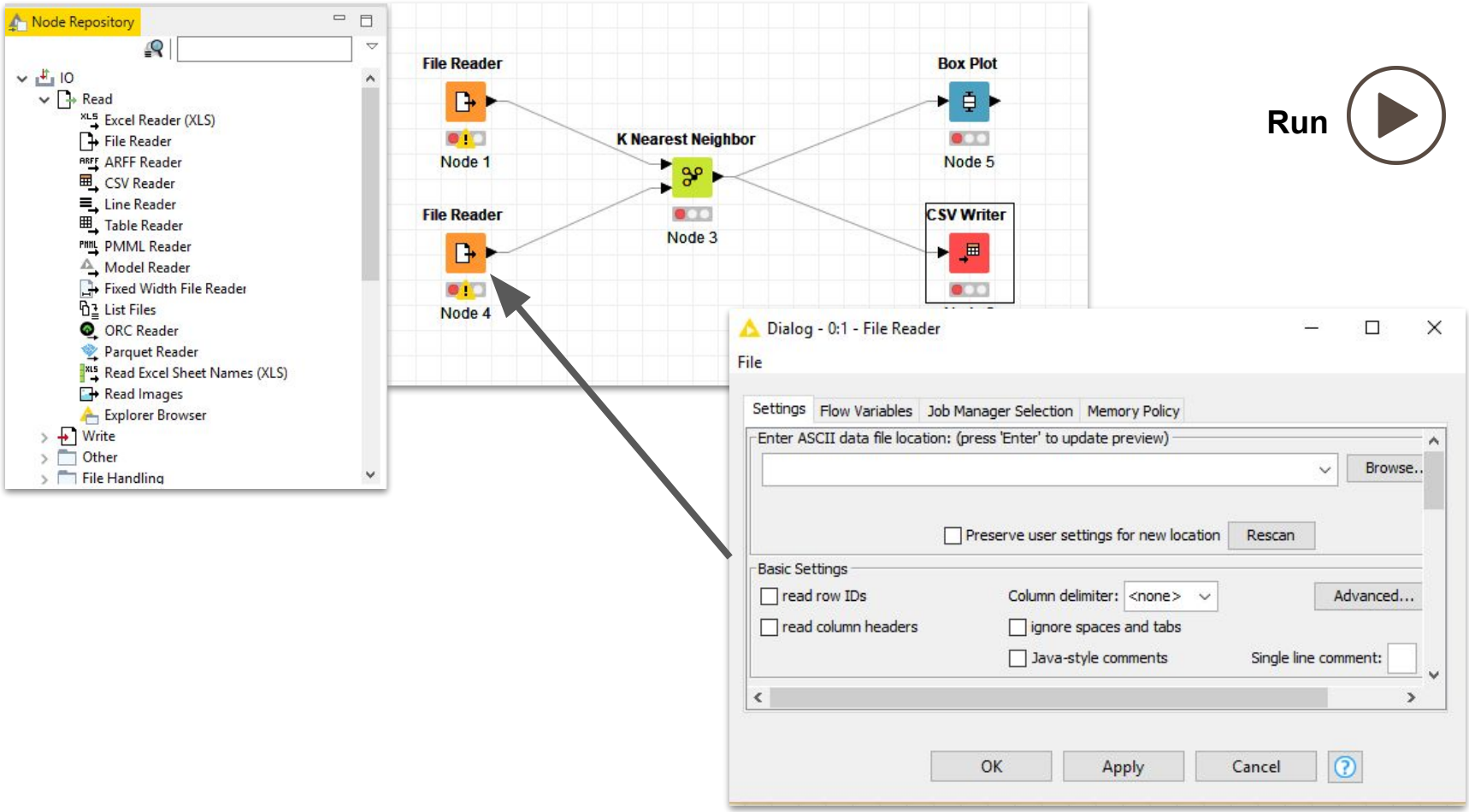
```
C:\Users\krisg\Documents>Rscript run_PEanalysis.R -m fgsea -input data.csv -o result.csv -g KEGG
[1] "Processing....."
[1] "Completed successful! See log file for more information."
```

...but what if you don’t have a programming background ?

# KNIME for Users

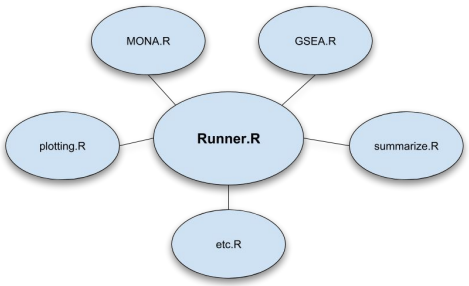
# KNIME Analytics Platform

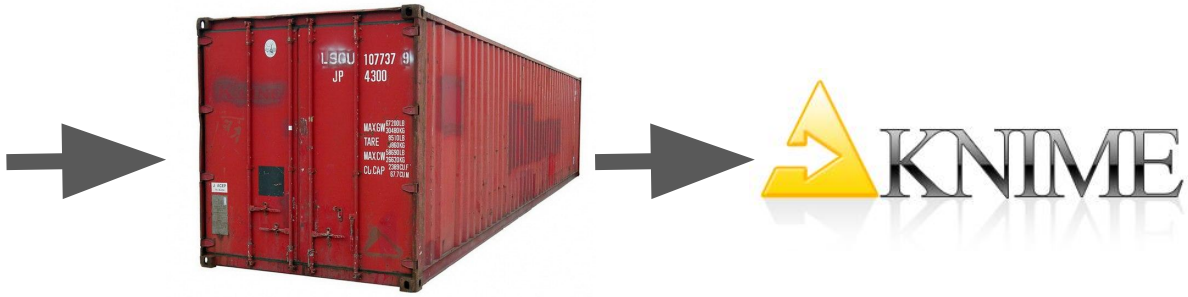
KNIME is a free software for interactive data analysis



# Introducing KNIME EnrichmentNodes

- KNIME Nodes are built on top of modular enrichment R scripts
- EnrichmentNodes use Docker containers that include
  - all software
  - all libraries and other dependencies
  - independent of local OS
- Docker image stored online at dockerhub.io and gets downloaded automatically



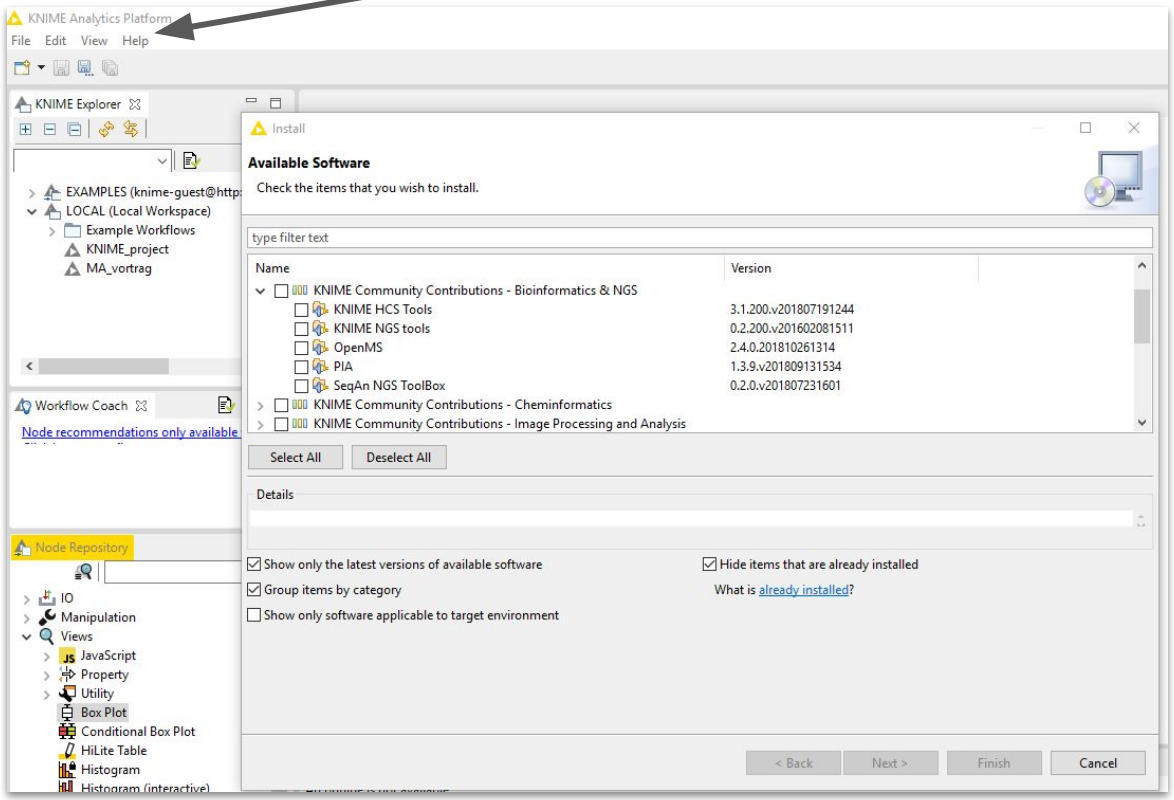



# KNIME Pathway Enrichment with “EnrichmentNodes”

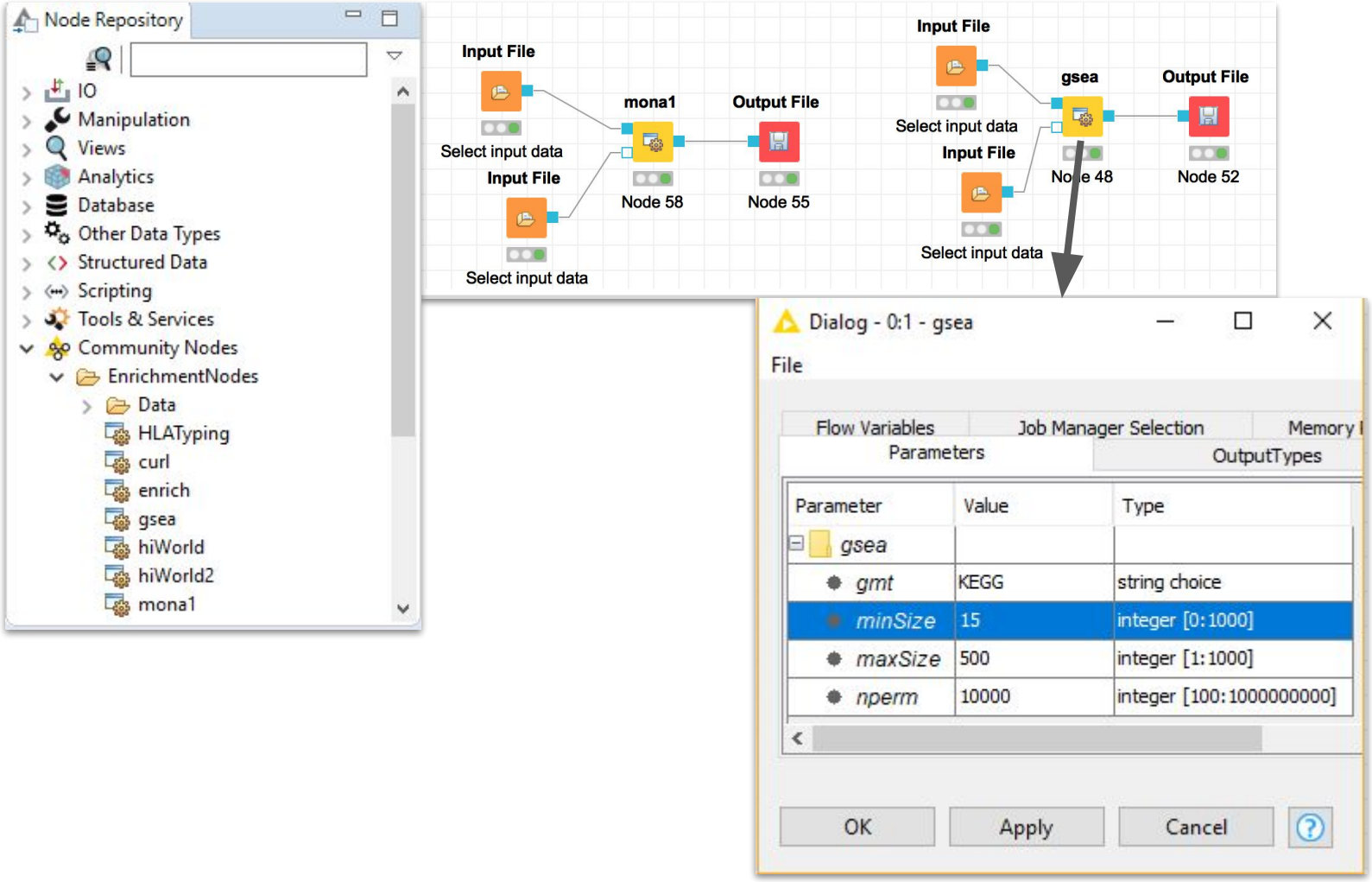
Requirements: KNIME, Docker, CPU supporting VT-x or AMD-v

- Install extensions in KNIME: GKN, EnrichmentNodes (Help->Install software…)

- Docker loads automatically
- alternatively build your own version locally (temporary)



# Using KNIME


Node Repository
IO
Manipulation
Views
Analytics
Database
Other Data Types
Structured Data
Scripting
Tools & Services
Community Nodes
EnrichmentNodes
Data
HLATyping
curl
enrich
gsea
hiWorld
hiWorld2
mona1
Input File
Select input data
Input File
Select input data
mona1
Node 58
Output File
Node 55
Input File
Select input data
Input File
Select input data
gsea
Node 48
Output File
Node 52
Dialog - 0:1 - gsea
File
Flow Variables
Job Manager Selection
Memory
Parameters
OutputTypes
Parameter
Value
Type
gsea
gmt
KEGG
string choice
minSize
15
integer [0:1000]
maxSize
500
integer [1:1000]
nperm
10000
integer [100:1000000000]
OK
Apply
Cancel

# KNIME for Developers

# Integrating own Methods into KNIME Nodes

**First Step:** integrate your R script/tool into our modular R framework

```
# ------------------------------------------------------------------------
#' Script to run GSEA analysis
#'
#' @author Ines Assum
#' @param input string // .RDS input file: named numerical vector (gene symbols)
#' @param output string // file path to save results as .RDS
#' @param gmt string // Gene set definition (KEGG / GO / custom [mygmt])
#' @param minSize integer // min size of GSs to be considered, default: 15
#' @param maxSize integer // max size of GSs to be considered, default: 500
#' @param nperm integer // number of permutations, default: 10000
#' @param mygmt string // custom .gmt file // not necessary
# ------------------------------------------------------------------------

run_fgsea <- function(input, output, gmt, minSize, maxSize, nperm, mygmt){
  library(fgsea)
  pathways <- gmtPathways(gmt)
  GSEA <- fgsea(pathways,
                rank,
                nperm,
                minSize=minSize,
                maxSize=maxSize)
}
```

# Integrating own Methods into KNIME

**Requirements:** Your Tool/R-Script, KNIME SDK, Ant and GenericKnimeNodes (GKN)

1. **CTD-file**

```
<tool name="YourTool.R" >
    <description>Draw plot for a nucleotide sequence.</description>
        <cli>
        <clielement optionIdentifier="-sequence">
            <mapping referenceName="YourTool.sequence" />
        </clielement>
        <clielement optionIdentifier="-outfile">
            <mapping referenceName="YourTool.outputfile" />
        </clielement>
    </cli>
    <PARAMETERS >
        <NODE name="YourTool" description="Draw plot for a nucleotide sequence.">
            <ITEM name="sequence" value="" description="sequence filename" supported_formats="*.fasta"/>
            <ITEM name="outputfile" value="" description="Output file." supported_formats="*.ps"/>
        </NODE>
    </PARAMETERS>
</tool>
```

# Integrating own Methods into KNIME

## 2. Specifying plugin.properties file

```
# the package of the plugin
pluginPackage=de.enrichment

# the name of the plugin
pluginName=EnrichmentNodes

# the version of the plugin
pluginVersion=1.0.0.0

# the path (starting from KNIMEs Community Nodes node)
nodeRepositoyRoot=community

executor=com.genericworkflownodes.knime.execution.impl.LocalDockerToolExecutor
commandGenerator=com.genericworkflownodes.knime.execution.impl.DockerCommandGenerator

#docker specific configurations
dockerMachine=default

#tool specific configurations
tool.monal.dockerImage=enrich
tool.gsea.dockerImage=enrich
tool.enrich.dockerImage=enrich
```

# Integrating own Methods into KNIME Nodes

**3. Provide predefined directory**

```
Plugin_dir
    |
    |--- plugin.properties   (description file)
    |
    |--- descriptors
    |      |-- yourTool.ctd
    |      |-- mime.types
    |
    |--- DESCRIPTION         (short description of project)
    |
    |--- LICENSE             (Licensing information)
    |
    |--- COPYRIGHT           (Copyright information)
```

**4. Run ant**

**5. Import and run on KNIME SDK**

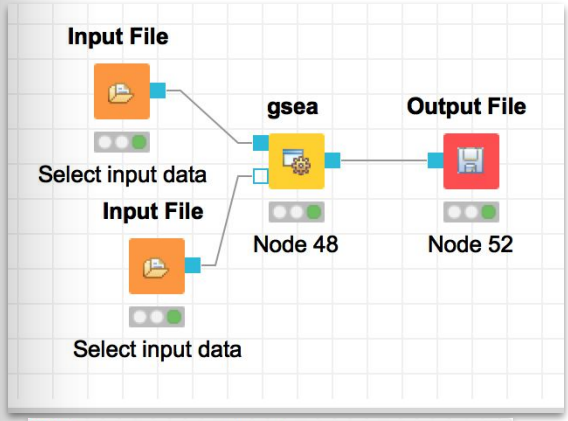
**Voila!**

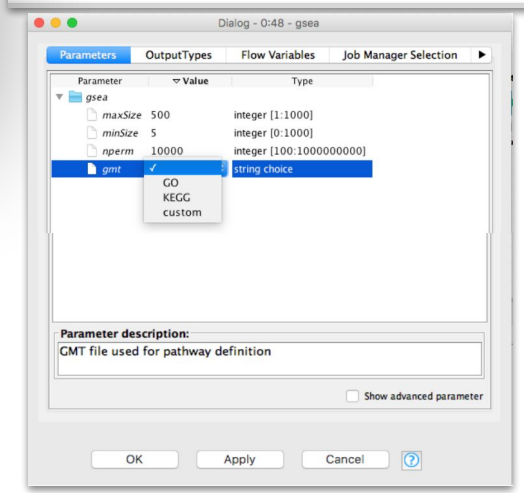
# Example

```
# ------------------------------------------------------------------------------
#' Script to run GSEA analysis
#'
#' @author Ines Assum
#' @param input string // .RDS input file: named numerical vector (gene symbols)
#' @param output string // file path to save results as .RDS
#' @param gmt string // Gene set definition (KEGG / GO / custom [mygmt])
#' @param minSize integer // min size of GSs to be considered, default: 15
#' @param maxSize integer // max size of GSs to be considered, default: 500
#' @param nperm integer // number of permutations, default: 10000
#' @param mygmt string // custom .gmt file // not necessary
# ------------------------------------------------------------------------------

run_fgsea <- function(input, output, gmt, minSize, maxSize, nperm, mygmt){
  library(fgsea)
  pathways <- gmtPathways(gmt)
  GSEA <- fgsea(pathways,
                rank,
                nperm,
                minSize=minSize,
                maxSize=maxSize)
}
```

```
            <ITEM description="GMT file used for pathway definition" name="gmt" restrictions="GO,KEGG,c
            <ITEM description="min Size" name="minSize" restrictions="0:1000" type="int" value="15"/>
            <ITEM description="max Size" name="maxSize" restrictions="1:1000" type="int" value="500"/>
            <ITEM description="Permutations" name="nperm" restrictions="100:1000000000" type="int" valu
        </NODE>
    </PARAMETERS>
    <cli>
        <clielement optionIdentifier="-gmt">
            <mapping referenceName="gsea.gmt"/>
        </clielement>
        <clielement optionIdentifier="-min">
            <mapping referenceName="gsea.minSize"/>
        </clielement>
        <clielement optionIdentifier="-max">
            <mapping referenceName="gsea.maxSize"/>
        </clielement>
        <clielement optionIdentifier="-p">
            <mapping referenceName="gsea.nperm"/>
        </clielement>
    </cli>
</tool>
```

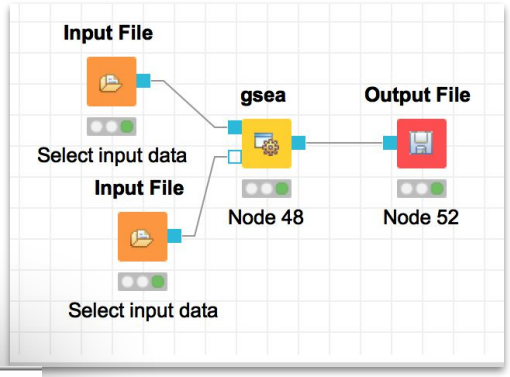


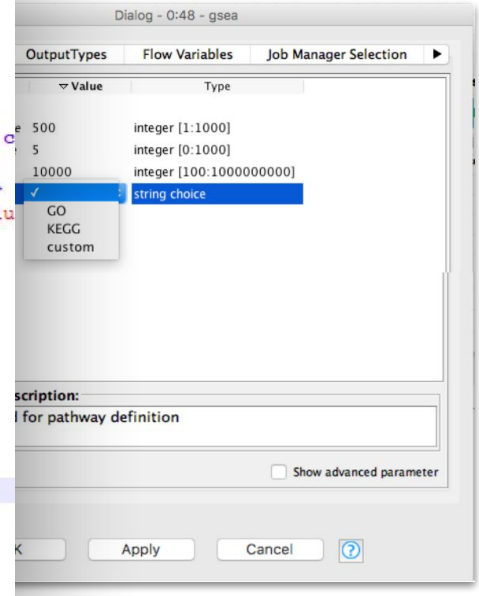
# Example

```
# ------------------------------------------------------------------------
#' Script to run GSEA analysis
#'
#' @author Ines Assum
#' @param input string // .RDS input file: named numerical vector (gene symbols)
#' @param output string // file path to save results as .RDS
#' @param gmt string // Gene set definition (KEGG / GO / custom [mygmt])
#' @param minSize integer // min size of GSs to be considered, default: 15
#' @param maxSize integer // max size of GSs to be considered, default: 500
#' @param nperm integer // number of permutations, default: 10000
#' @param mygmt string // custom .gmt file // not necessary
# ------------------------------------------------------------------------

run_fgsea <- function(input, output, gmt, minSize, maxSize, nperm, mygmt){
  library(fgsea)
  pathways <- gmtPathways(gmt)
  GSEA <- fgsea(pathways,
                rank,
                nperm,
```



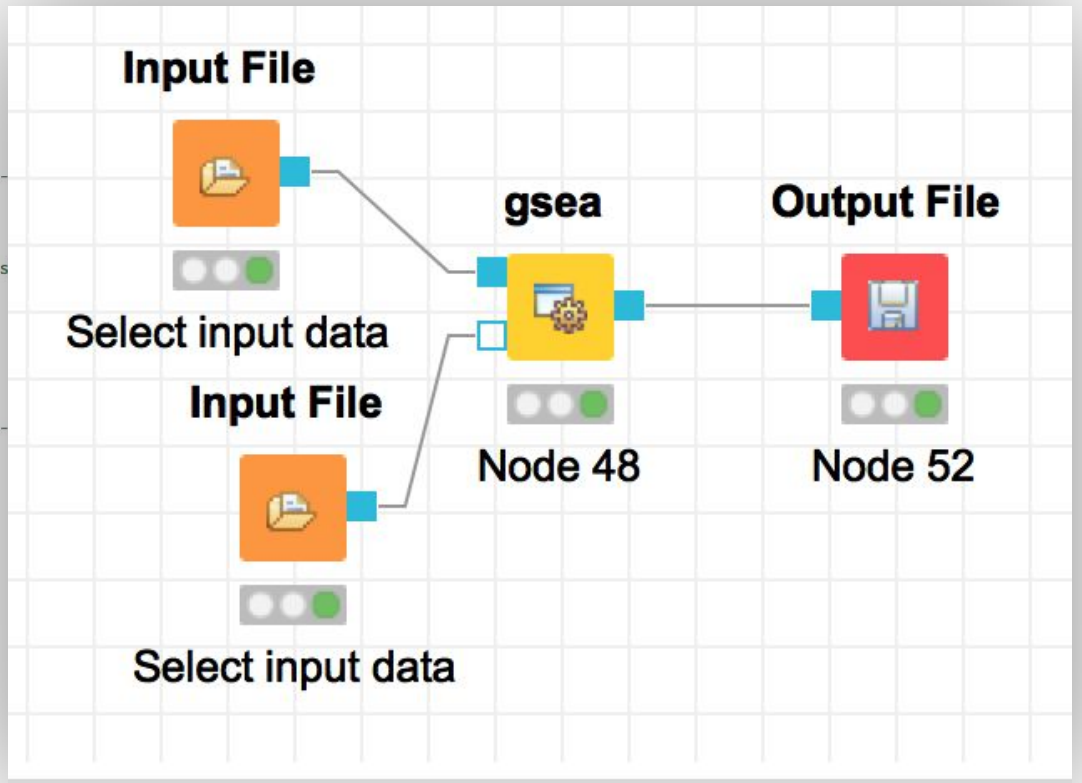

```
<tool name="gsea">
    <description>Runs GSEA using fgsea R package.</description>
    <manual>Detailed description goes here.</manual>
    <executableName>run_fgsea.R</executableName>
    <PARAMETERS >
        <NODE name="gsea" description="Node the runs GSEA">
            <ITEM description="GMT file used for pathway definition" name="gmt" restrictions="GO,KEGG,c
            <ITEM description="min Size" name="minSize" restrictions="0:1000" type="int" value="15"/>
            <ITEM description="max Size" name="maxSize" restrictions="1:1000" type="int" value="500"/>
            <ITEM description="Permutations" name="nperm" restrictions="100:1000000000" type="int" valu
        </NODE>
    </PARAMETERS>
    <cli>
        <clielement optionIdentifier="-gmt">
            <mapping referenceName="gsea.gmt"/>
        </clielement>
        <clielement optionIdentifier="-min">
            <mapping referenceName="gsea.minSize"/>
        </clielement>
        <clielement optionIdentifier="-max">
            <mapping referenceName="gsea.maxSize"/>
        </clielement>
        <clielement optionIdentifier="-p">
            <mapping referenceName="gsea.nperm"/>
        </clielement>
    </cli>
</tool>
```



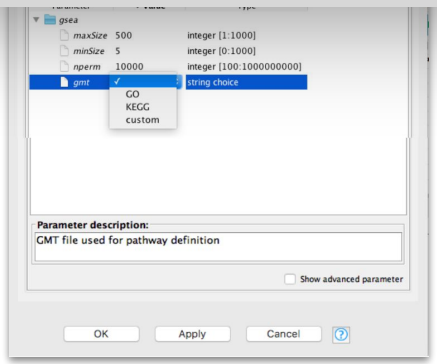
# Example

```
# ------------------------------------------------------------------------------
#' Script to run GSEA analysis
#'
#' @author Ines Assum
#' @param input string // .RDS input file: named numerical vector (gene symbols
#' @param output string // file path to save results as .RDS
#' @param gmt string // Gene set definition (KEGG / GO / custom [mygmt])
#' @param minSize integer // min size of GSs to be considered, default: 15
#' @param maxSize integer // max size of GSs to be considered, default: 500
#' @param nperm integer // number of permutations, default: 10000
#' @param mygmt string // custom .gmt file // not necessary
# ------------------------------------------------------------------------------

run_fgsea <- function(input, output, gmt, minSize, maxSize, nperm, mygmt){
  library(fgsea)
  pathways <- gmtPathways(gmt)
  GSEA <- fgsea(pathways,
                rank,
                nperm,
                minSize=minSize,
                maxSize=maxSize)
}
```




```
<tool name="gsea">
    <description>Runs GSEA using fgsea R package.</description>
    <manual>Detailed description goes here.</manual>
    <executableName>run_fgsea.R</executableName>
    <PARAMETERS >
        <NODE name="gsea" description="Node the runs GSEA">
            <ITEM description="GMT file used for pathway definition" name="gmt" restrictions="GO,KEGG,c
            <ITEM description="min Size" name="minSize" restrictions="0:1000" type="int" value="15"/>
            <ITEM description="max Size" name="maxSize" restrictions="1:1000" type="int" value="500"/>
            <ITEM description="Permutations" name="nperm" restrictions="100:1000000000" type="int" valu
        </NODE>
    </PARAMETERS>
    <cli>
        <clielement optionIdentifier="-gmt">
            <mapping referenceName="gsea.gmt"/>
        </clielement>
        <clielement optionIdentifier="-min">
            <mapping referenceName="gsea.minSize"/>
        </clielement>
        <clielement optionIdentifier="-max">
            <mapping referenceName="gsea.maxSize"/>
        </clielement>
        <clielement optionIdentifier="-p">
            <mapping referenceName="gsea.nperm"/>
        </clielement>
    </cli>
</tool>
```

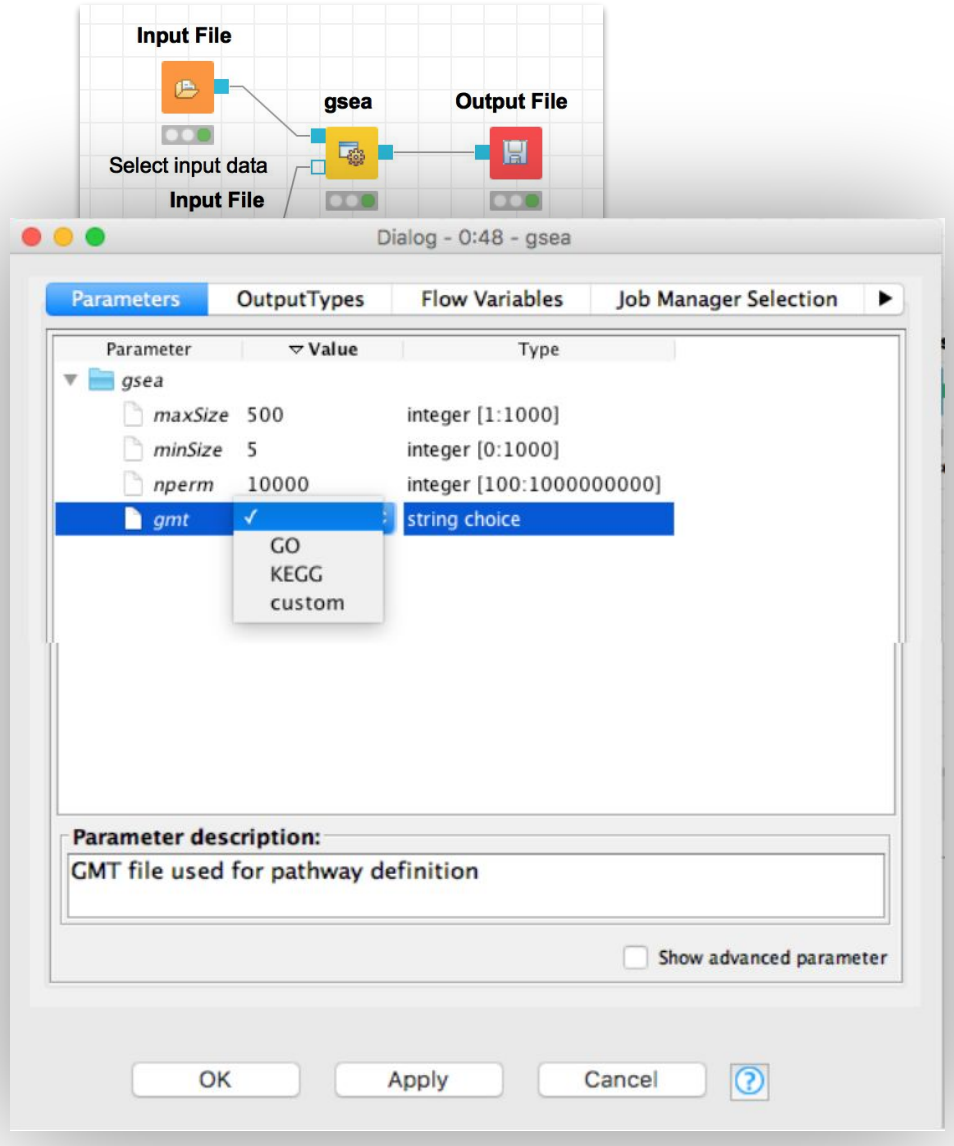


# Example

```
# ------------------------------------------------------------------------------
#' Script to run GSEA analysis
#'
#' @author Ines Assum
#' @param input string // .RDS input file: named numerical vector (gene symbols)
#' @param output string // file path to save results as .RDS
#' @param gmt string // Gene set definition (KEGG / GO / custom [mygmt])
#' @param minSize integer // min size of GSs to be considered, default: 15
#' @param maxSize integer // max size of GSs to be considered, default: 500
#' @param nperm integer // number of permutations, default: 10000
#' @param mygmt string // custom .gmt file // not necessary
# ------------------------------------------------------------------------------

run_fgsea <- function(input, output, gmt, minSize, maxSize, nperm, mygmt){
  library(fgsea)
  pathways <- gmtPathways(gmt)
  GSEA <- fgsea(pathways,
                rank,
                nperm,
                minSize=minSize,
                maxSize=maxSize)
}
```

```
<tool name="gsea">
    <description>Runs GSEA using fgsea R package.</description>
    <manual>Detailed description goes here.</manual>
    <executableName>run_fgsea.R</executableName>
    <PARAMETERS >
        <NODE name="gsea" description="Node the runs GSEA">
            <ITEM description="GMT file used for pathway definition" name="gmt" restrictions="GO,KEGG,c
            <ITEM description="min Size" name="minSize" restrictions="0:1000" type="int" value="15"/>
            <ITEM description="max Size" name="maxSize" restrictions="1:1000" type="int" value="500"/>
            <ITEM description="Permutations" name="nperm" restrictions="100:1000000000" type="int" valu
        </NODE>
    </PARAMETERS>
    <cli>
        <clielement optionIdentifier="-gmt">
            <mapping referenceName="gsea.gmt"/>
        </clielement>
        <clielement optionIdentifier="-min">
            <mapping referenceName="gsea.minSize"/>
        </clielement>
        <clielement optionIdentifier="-max">
            <mapping referenceName="gsea.maxSize"/>
        </clielement>
        <clielement optionIdentifier="-p">
            <mapping referenceName="gsea.nperm"/>
        </clielement>
    </cli>
</tool>
```



# Summary

- Modular R library integrated into easy-to-use KNIME workflow
- Expand functionality with minimal effort
- We will soon make the “EnrichmentNodes” project on GitHub public:
  -> https://github.com/InesAssum/EnrichmentNodes
- Template for exemplary node provided
- Collect and share methods
- Join our slack workspace for discussion and support
  -> Slack.com/KNIME-setup

- Check out Benni’s ImmunoNodes: github.com/FRED-2/ImmunoNodes/

# Acknowledgements

**Epigenereg group**

Matthias Heinig

Ines Assum

Thomas Walzthöni

**ICB (MONA)**

Nikola Müller

Florian Büttner

Andreas Kopf

**GenericKnimeNodes**

Julianus Pfeuffer

**&**

**ICB**

Benjamin Schubert

-> ImmunoNodes