Introduction to paxtoolsr

Augustin Luna
27 June, 2016

Research Fellow
Department of Biostatistics and Computational Biology
Dana-Farber Cancer Institute

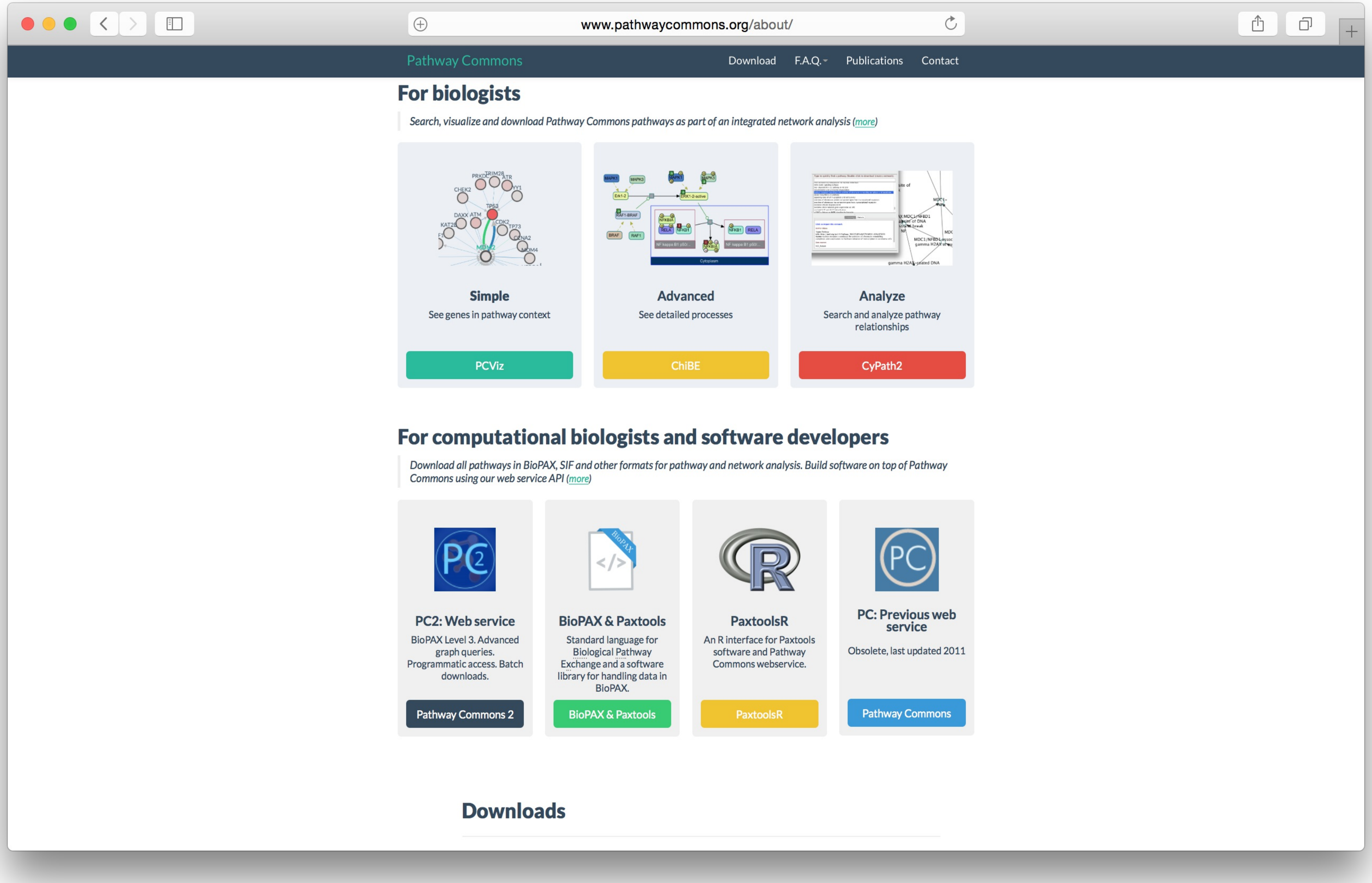
What is Pathway Commons?

- Website: http://www.pathwaycommons.org/
- An aggregation of public pathway database information
- Provides data in multiple formats
  - Biological Pathway Exchange (BioPAX) Format
  - Simple Interaction Format (SIF)
  - Gene sets as Gene Matrix Transposed (GMT) Format
- Provides infrastructure for searching the aggregated pathway data

Biological Pathway Exchange (BioPAX) Format

- BioPAX: http://biopax.org/
- Community-wide effort to represent biological pathways
  - Pathways are collections of interactions that biologists have found useful to group together for organizational, historic, biophysical or other reasons
- Types
  - Metabolic pathways
  - Signaling pathways
  - Protein-protein interactions
  - Gene regulatory pathways
- Advanced tutorial on BioPAX
  - https://github.com/cannin/biopaxTutorial

Pathway Commons Homepage



For biologists

Search, visualize and download Pathway Commons pathways as part of an integrated network analysis (more)

Simple

See genes in pathway context

Advanced

See detailed processes

Analyze

Search and analyze pathway relationships

For computational biologists and software developers

Download all pathways in BioPAX, SIF and other formats for pathway and network analysis. Build software on top of Pathway Commons using our web service API (more)

PC2: Web service

BioPAX Level 3. Advanced graph queries. Programmatic access. Batch downloads.

BioPAX & Paxtools

Standard language for Biological Pathway Exchange and a software library for handling data in BioPAX.

PaxtoolsR

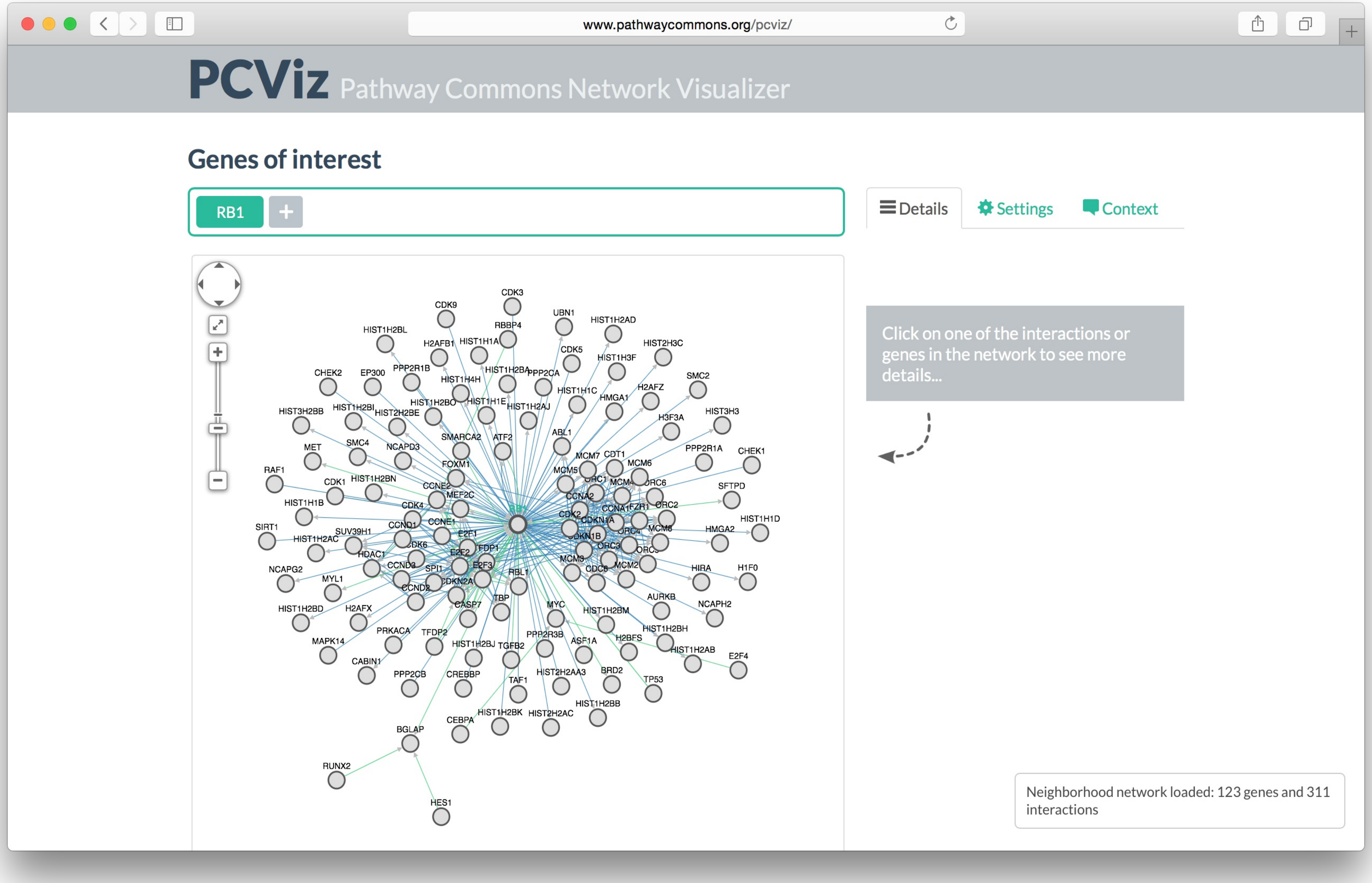
An R interface for Paxtools software and Pathway Commons webservice.

PC: Previous web service

Obsolete, last updated 2011

Downloads

Pathway Commons Visualizer



Click on one of the interactions or genes in the network to see more details...

Neighborhood network loaded: 123 genes and 311 interactions

Pathway Commons Data sets

Database	Interaction Count
Reactome	11924
NCI PID	16017
PhosphoSitePlus	13642
HumanCyc	7024
HPRD	40618
PantherDB	5282
DIP	7102
BioGRID	244843

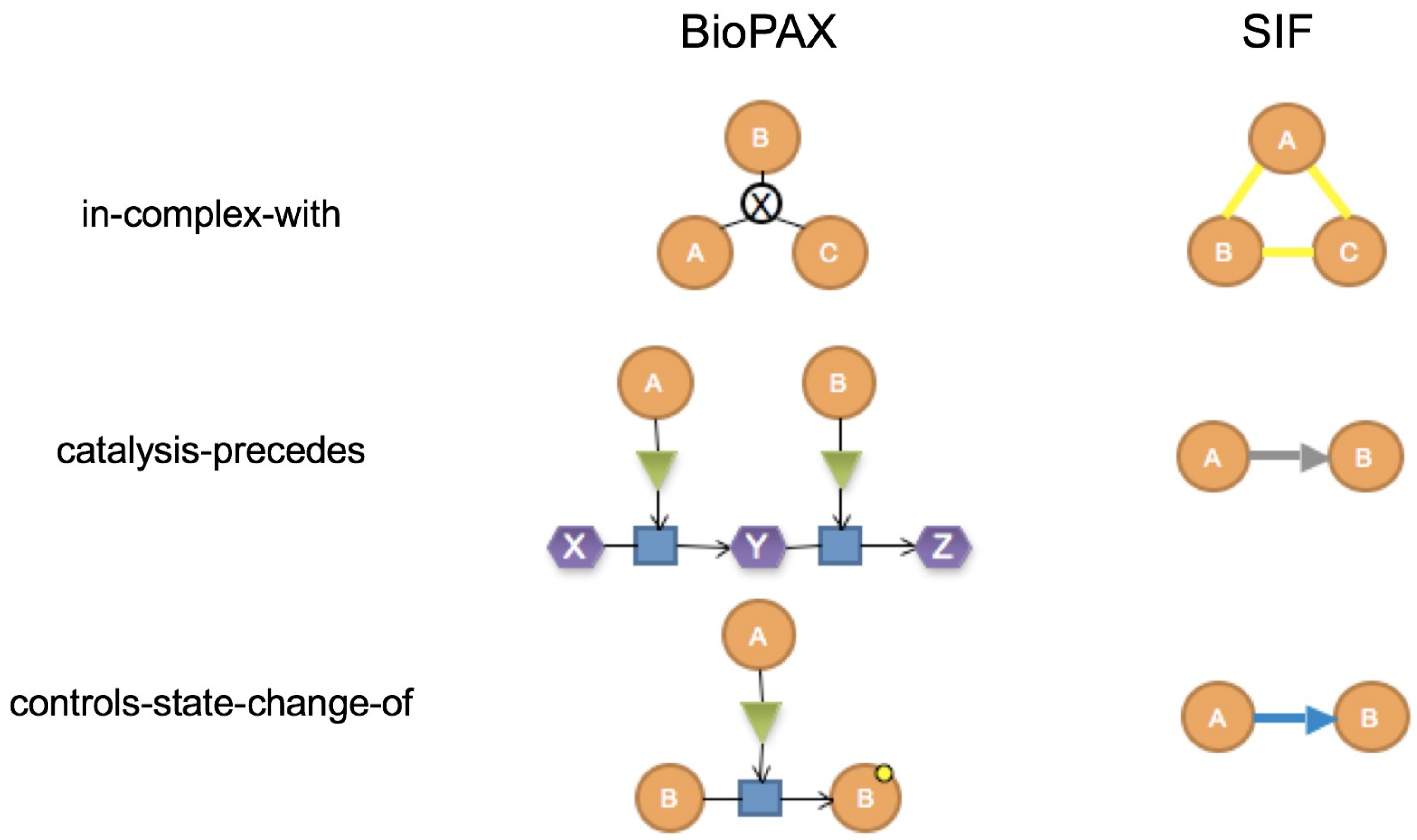
Database	Interaction Count
InAct	98347
BIND	35566
TRANSFAC	261624
mirTarBase	51214
DrugBank	19159
Recon X	10910
CTD	313174
KEGG	4472

Simple Interaction Format (SIF)

- An edgelist with interaction type: 3 columns
  - PARTICIPANT_A, INTERACTION_TYPE, PARTICPANT_B
- Expected representation for many network analyses
- Extracted using graph queries that detect biologically interesting interaction patterns in Pathway Commons data
  - Complexes, metabolic, modification, control interactions
  - Generates binary interactions and integrates them across databases

SIF Interaction Types

- Complete list of interaction types in Google Docs
- Examples of conversions from BioPAX to SIF



14 Interaction Types Total

Gene Set (GMT) Format

Gene Set	Description	Gene 1	Gene 2	Gene 3	...
KEGG_GLYCOLYSIS_GLUCONEOGENESIS	KEGG	GCK	PGK2	PGK1	...
REACTOME_SIGNALING_BY_EGFR_IN_CANCER	Reactome	AKT3	ADAM10	SPRY1	...

What is paxtoolsr?

- Website and Tutorial (Vignette):
  - https://bioconductor.org/packages/release/bioc/html/paxtoolsr.html
- Publication:
  - http://www.ncbi.nlm.nih.gov/pubmed/26685306
- Read and write
  - Biological Pathway Exchange (BioPAX)
  - Binary Simple Interaction Format (SIF)
  - Extended SIF: Includes additional information about SIF network
  - Gene Set (GMT)
  - Systems Biology Graphical Notation Markup Language (SBGN-ML)
- Search and summarize local BioPAX files
- Search Pathway Commons

Downloading and Reading Pathway Commons Data

- Load library

```
library(paxtoolsr)
```

- List possible downloads

```
downloadPc2()
```

- Download databases

```
# Single databases
geneSets <- downloadPc2("PathwayCommons.8.Reactome.GSEA.hgnc.gmt.gz", version="8")
sif <- downloadPc2("PathwayCommons.8.kegg.EXTENDED_BINARY_SIF.hgnc.txt.gz",
version="8")

# All databases
geneSets <- downloadPc2("PathwayCommons.8.All.GSEA.hgnc.gmt.gz", version="8")
```

Filtering Pathway Commons Data

```
sif <- filterSif(sif$edges, ids=c("GPI"))

nrow(sif)
```

```
[1] 26
```

```
colnames(sif)
```

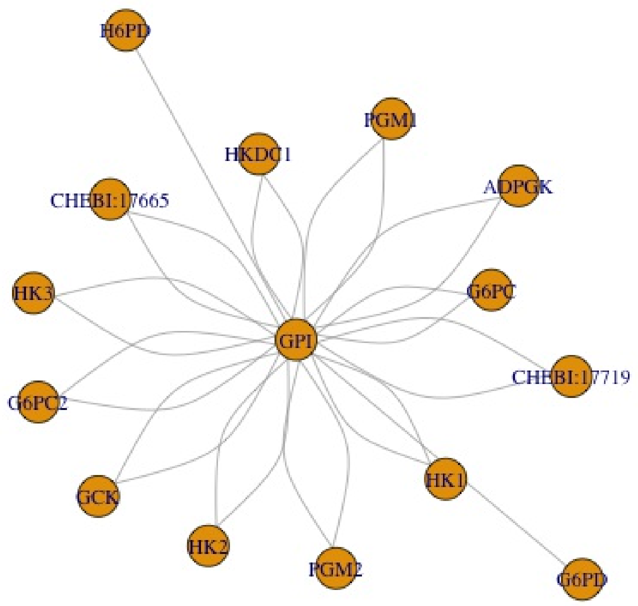
```
[1] "PARTICIPANT_A"           "INTERACTION_TYPE"
[3] "PARTICIPANT_B"           "INTERACTION_DATA_SOURCE"
[5] "INTERACTION_PUBMED_ID"   "PATHWAY_NAMES"
[7] "MEDIATOR_IDS"
```

```
head(sif[, 1:3, with=FALSE], 2)
```

```
   PARTICIPANT_A       INTERACTION_TYPE PARTICIPANT_B
1:           GPI     catalysis-precedes         ADPGK
2:           GPI controls-production-of   CHEBI:17665
```

Visualize Pathway Commons Data

```r
library(igraph); library(data.table) # SIF files read as data.table for speed

setDF(sif) # Convert data.table to data.frame

# graph.edgelist requires a matrix
g <- graph.edgelist(as.matrix(sif[, c(1, 3)]), directed = FALSE)
plot(g, layout = layout.fruchterman.reingold)
```



ID Conversion Using the Chemical Translation Service

```r
library(webchem)

cts_convert('16-hydroxypalmitate', 'Chemical Name',
'ChEBI')
```

```
$`16-hydroxypalmitate`
[1] "CHEBI:55328" "CHEBI:55329"
```

Get Metabolite Interactions (1)

- Load Example Metabolite ChEBI IDs

```r
metab <- read.table("example_chebi.txt", sep="\t",
header=TRUE, quote="", comment.char="",
stringsAsFactors=FALSE)
```

Get Metabolite Interactions (2)

```r
# KEGG
sifKegg <- downloadPc2("PathwayCommons.8.kegg.EXTENDED_BINARY_SIF.hgnc.txt.gz",
version="8")
sif <- sifKegg

paths <- unique(unlist(sif$edges$PATHWAY_NAMES))
purineIdx <- grepl("purine", paths, ignore.case=TRUE)
purinePaths <- paths[purineIdx]

metabFilteredSif <- filterSif(sif$edges, ids=metab$chebi)
tmp <- searchListOfVectors(purinePaths, metabFilteredSif$PATHWAY_NAMES)
purineIdx <- unique(unlist(tmp))

purineOnlySif <- metabFilteredSif[purineIdx]
setDF(purineOnlySif)
purineOnlySif[1:2, 1:6]
```

```
  PARTICIPANT_A          INTERACTION_TYPE PARTICIPANT_B
1   CHEBI:15422 consumption-controlled-by         ADCY3
2   CHEBI:15422           used-to-produce   CHEBI:15996
  INTERACTION_DATA_SOURCE INTERACTION_PUBMED_ID
1                    KEGG                    NA
2                    KEGG                    NA
                          PATHWAY_NAMES
1                     Purine metabolism
2 Metabolic pathways, Purine metabolism
```

```r
tmp <- c(purineOnlySif[, 1], purineOnlySif[, 3])
idx <- which(!grepl("^CHEBI:", tmp))

resKegg <- sort(table(tmp[idx]))
length(resKegg)
```

```
[1] 93
```

Enrichment Analysis with Pathway Commons and CellMiner

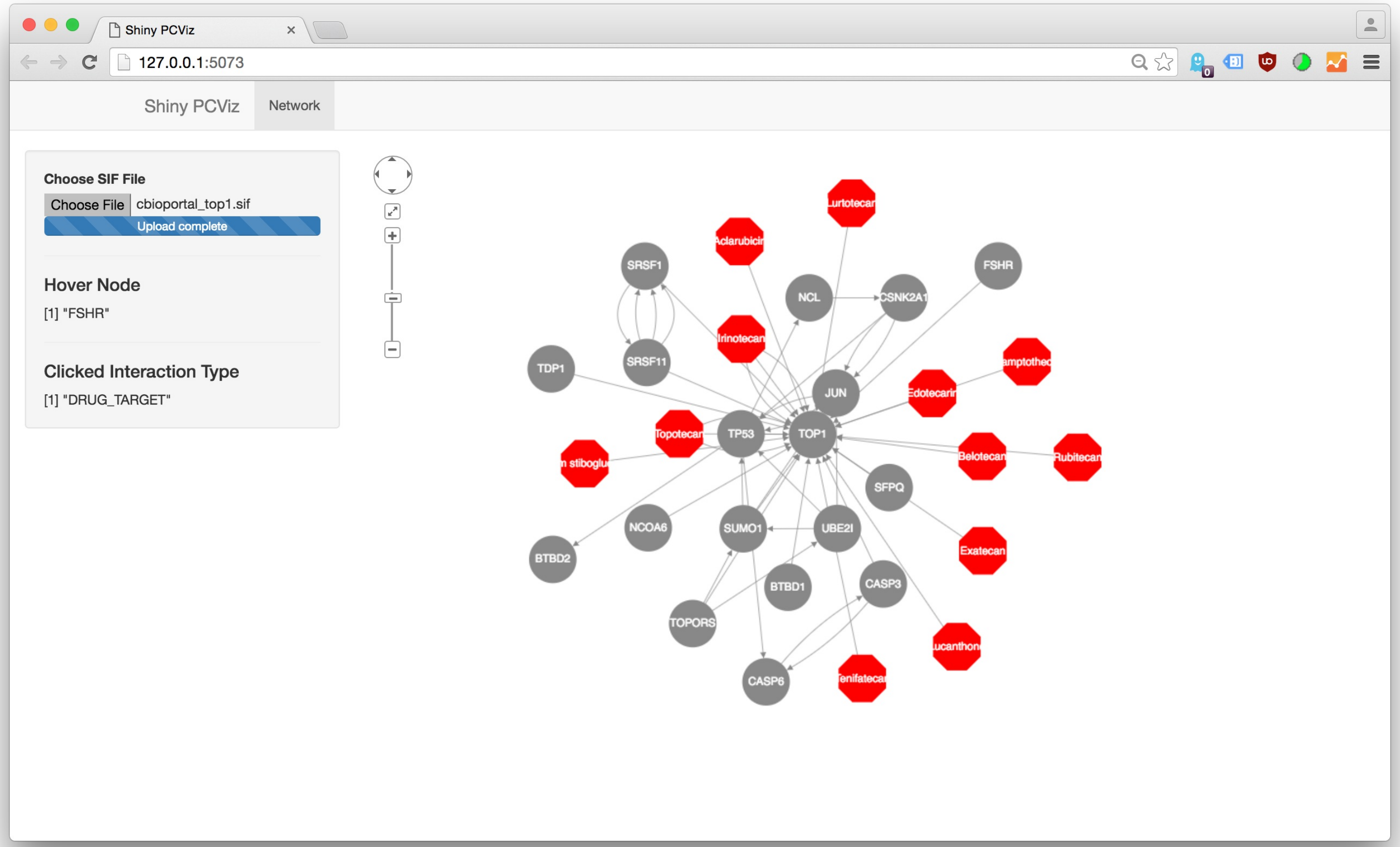
- Example on conducting an enrichment analysis on CellMiner cell line data using gene sets from Pathway Commons

```r
# Load libraries
library(paxtoolsr); library(rcellminer)

# Load data
geneSets <- downloadPc2("PathwayCommons.8.Reactome.GSEA.hgnc.gmt.gz", version="8")
mutData <- getAllFeatureData(rcellminerData::molData)[["mut"]]

hiMutGenes <- head(sort(rowSums(mutData), decreasing=TRUE), 25)

# Initialize variable
pvals <- NULL

for(set in geneSets) {
  #set <- hiMutGenes
  sampleSize <- length(hiMutGenes) # size drawn
  hitInSample <- length(which(hiMutGenes %in% set)) # black drawn
  hitInPop <- length(which(rownames(mutData) %in% set)) # all black
  failInPop <- nrow(mutData)-hitInPop # number of red
  # Calculate over-enrichment for current gene set
  pval <- phyper(hitInSample-1, hitInPop, failInPop, sampleSize, lower.tail= FALSE)
  # Add current result
  pvals <- c(pvals, pval)
}

# Adjust p-values
pvals <- p.adjust(pvals, method="fdr")
length(pvals[pvals < 0.05])
```

```
[1] 0
```

Interactive Pathway Commons Applications using rcytoscapejs

- CytoscapeJS for embedding in Shiny applications
- Code Repository: https://github.com/cytoscape/r-cytoscape.js
- Demo in `inst/examples/shinyPCViz`



Choose SIF File

cbioportal_top1.sif

Upload complete

Hover Node

[1] "FSHR"

Clicked Interaction Type

[1] "DRUG_TARGET"

Getting Help

- paxtoolsr: Bioconductor
  - http://bioconductor.org/packages/release/bioc/html/paxtoolsr.html
- paxtoolsr Installation Videos
  - https://youtu.be/lUwP6KncMOo?list=PLpNSl8ajNxXy0fg2YIG5wa5zAV_vh1ULV
- BioPAX Google Group
  - http://groups.google.com/group/biopax
- Pathway Commons Google Group
  - http://groups.google.com/group/pathway-commons-help
- rcytoscapejs
  - https://github.com/cytoscape/r-cytoscape.js

Acknowledgements

- Dana-Farber Cancer Institute
  - Augustin Luna
  - Chris Sander
- Bilkent University
  - Ugur Dogrusoz
- University of Toronto
  - Jeffrey Wong
  - Igor Rodchenkov
  - Gary Bader
- Oregon Health Sciences University
  - Ozgun Babur
  - Emek Demir