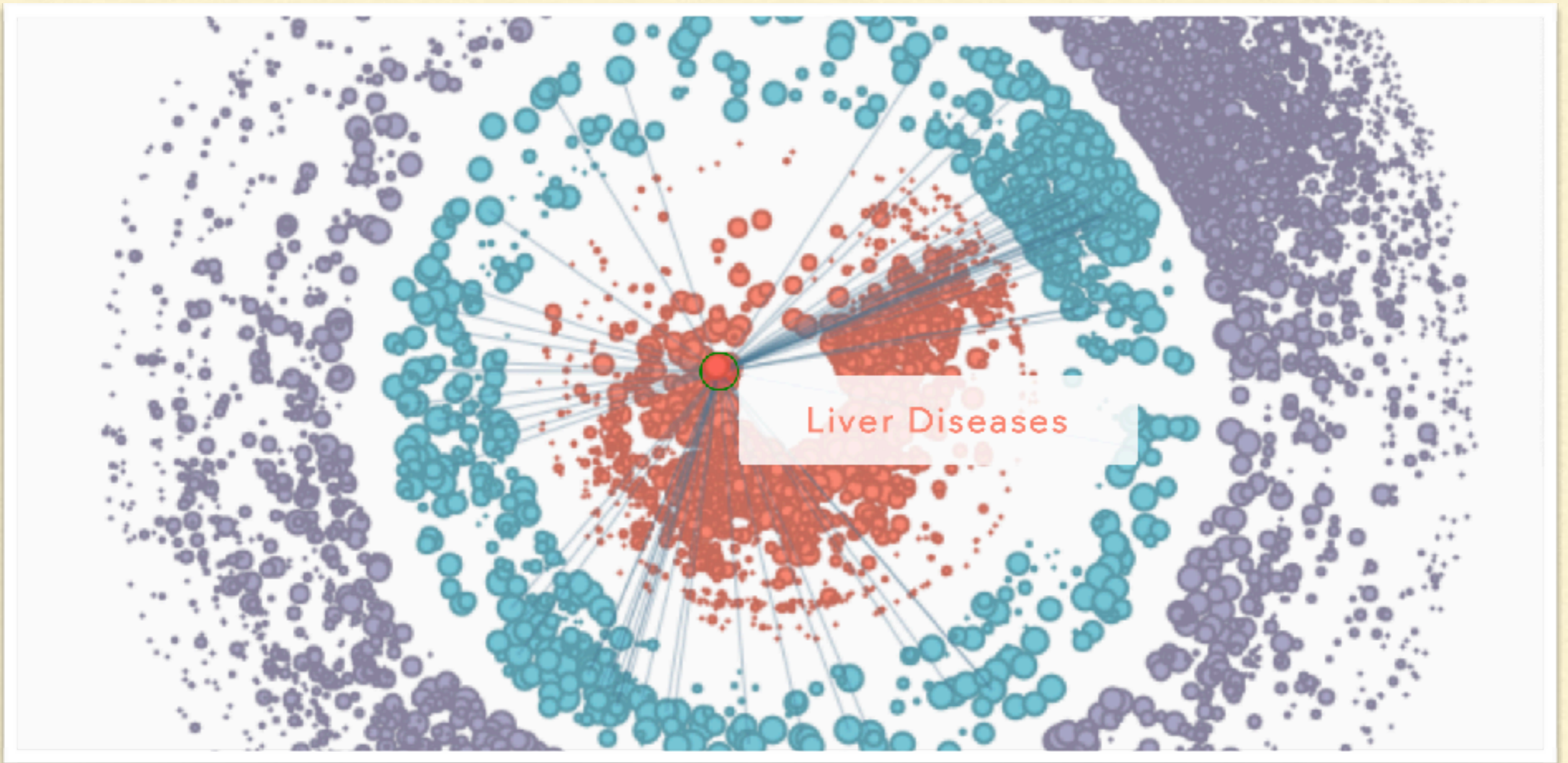

DISEASE-MIRNA RELATIONSHIPS

Alexander Liang & Matt Laws Under Dr. Nestoras Karathanasis



Studio[®]



INTRODUCTION

GOAL

- Text-mining miR-Disease Relations
- Existing Models
 - HMDD (Manual)
 - miR2Disease (Manual)
 - miR2Cancer (Incomplete)
- Our Goal
 - Automatic, Efficient

The PubMed logo features the word "PubMed" in a blue, sans-serif font. The letter "M" is stylized, appearing as a white shape with a blue outline that resembles a book or a speech bubble.The logo for the Human microRNA Disease Database is presented on a light gray background with a subtle grid pattern. The text "Human microRNA Disease Database" is written in a bold, black, sans-serif font, with the words "microRNA", "Disease", and "Database" highlighted in red.The logo for Epi miRBase features the word "Epi" in a large, gray, serif font. Below it, the word "miRBase" is written in a black, serif font. The letter "i" in "miRBase" is stylized, with a blue and green patterned background.

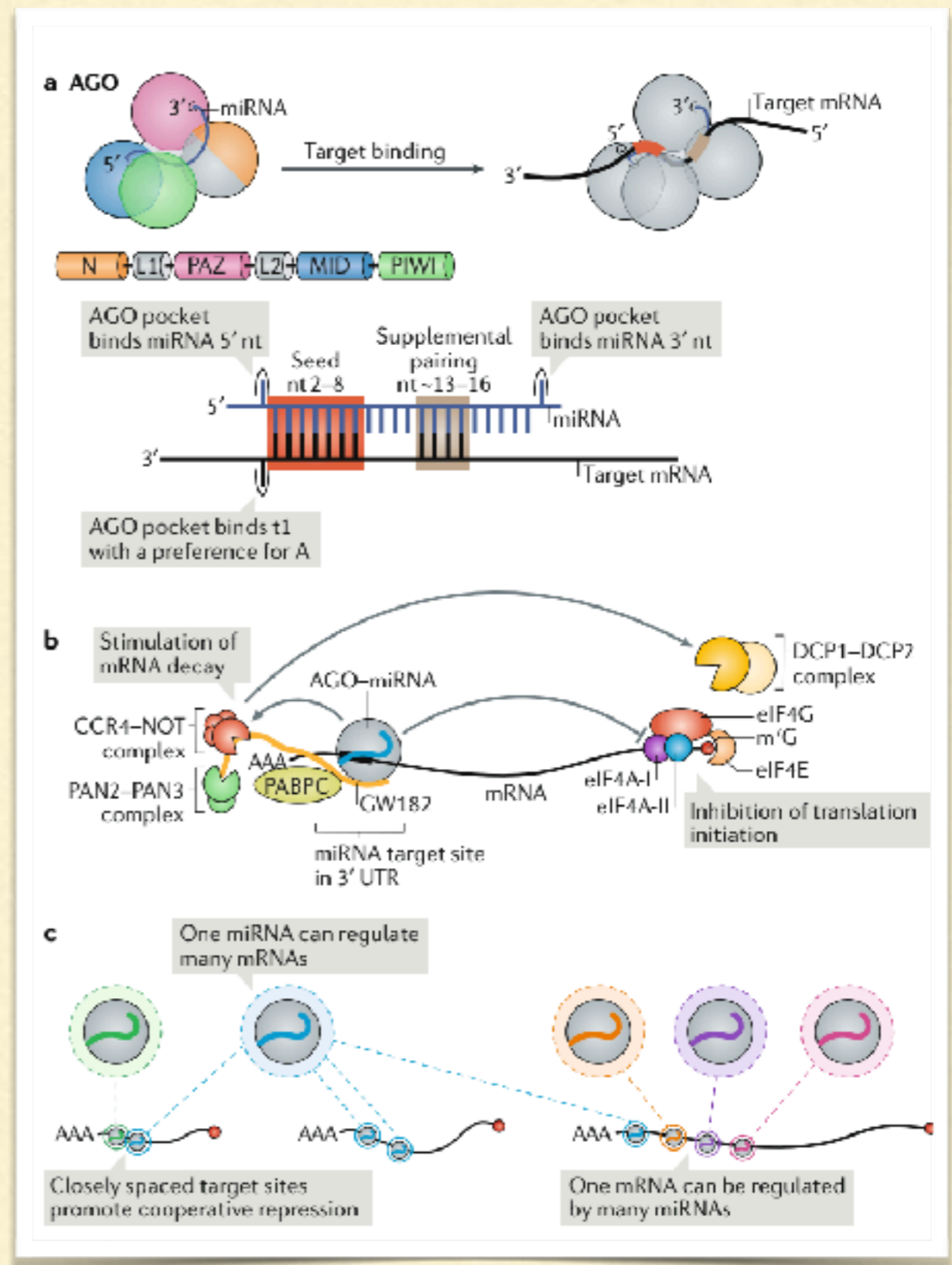
PROCESS

- Understanding the Field
- Learning R
- Reading Literature
- Extracting miRNA
- Finding Relation Sentences
- Evaluating Our Findings
- Cleaning Code



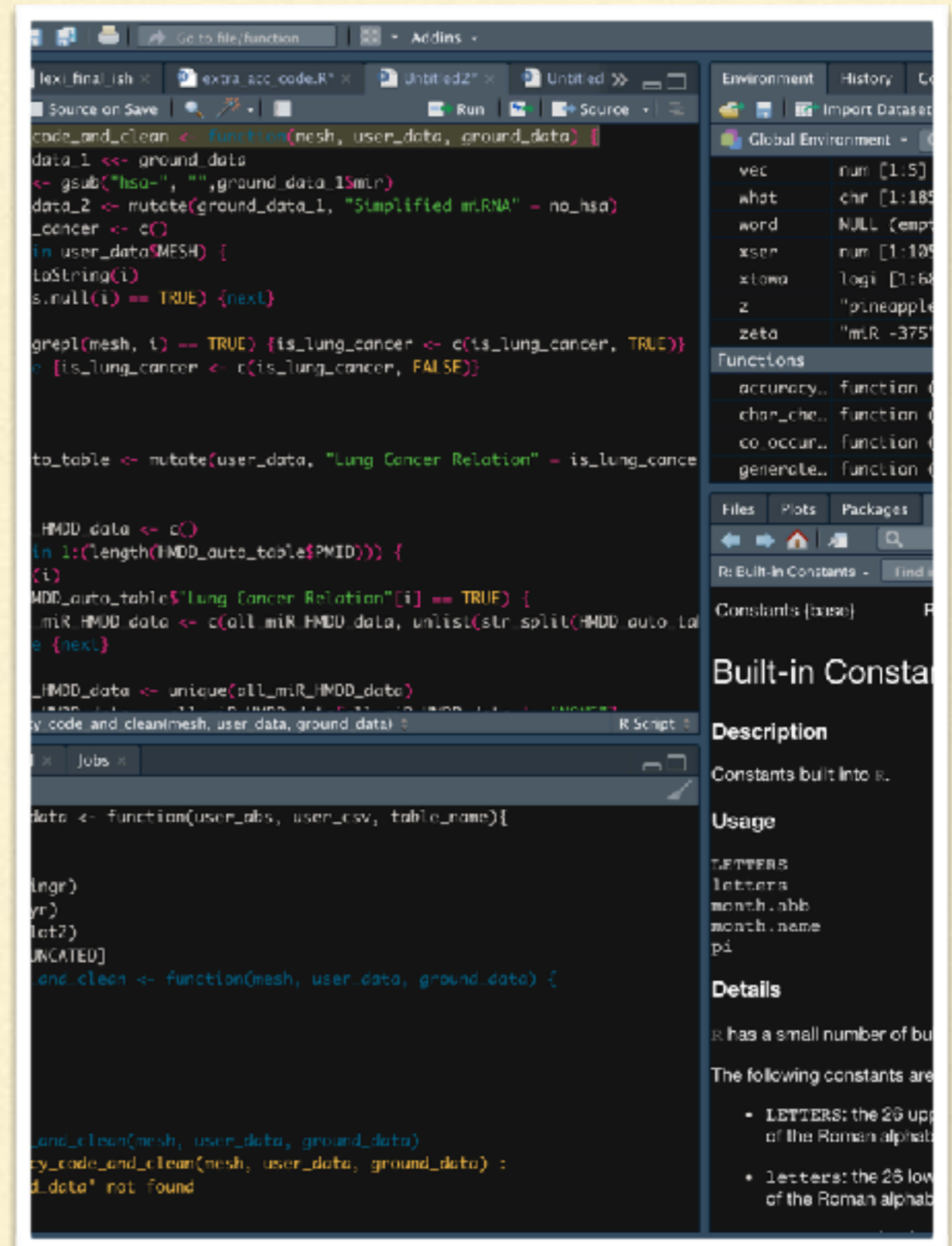
MIRNA

- Non-Coding Strand
- Roughly 18-28 Nucleotides Long
- Controls Gene Expression
- Dynamic Relationship With mRNAs



R Language

- Functional Oriented Language
- Data Analysis Program
- Useful Packages
- Widely Used for Bioinformatics



```
tmpFormat = 14 #Replace string by value
str(key)) tempString = tempString.replace("czDataType
str(int(value*pow(10,14-tmpFormat)))) tempString = temp
elif(typeOfFID == "BUFFER"): s = value dataCal =
tempString.replace("czFieldID",str(key)) tempStri
elif(typeOfFID == "ASCII_STRING"): s = value dataC
tempString = tempString.replace("czData",
if "date value=" in line and flagCheckRichnam
prop(1) if "</Message>" in line: myEvent = "RT_CHA
+onlyFilename+"\n" if typeOfFile == "RT_CHA
If not os.path.exists(path): os
```

PROGRAMMING

PROGRAMMING

- pubmed.mineR package
- miRNA
- Disease/MeSH ID
- Relation
- Organism, Country, PMID

miRNA	Disease	Relationship
miR-21	lung cancer	CONCLUSIONS: MiR-21 expression levels in
miR-21	cancer	CONCLUSIONS: MiR-21 expression levels in
MiR-21	lung cancer	CONCLUSIONS: MiR-21 expression levels in
MiR-21	cancer	CONCLUSIONS: MiR-21 expression levels in
miR-21	lung carcinoma	NA
miR-21	cancer	In summary, our results suggest that miR-2
miR-21	lung cancer	In summary, our results suggest that miR-2
miR-24	lung carcinoma	NA
miR-24	cancer	In summary, our results suggest that miR-2
miR-24	lung cancer	In summary, our results suggest that miR-2
miR-30d	lung carcinoma	NA
miR-30d	cancer	In summary, our results suggest that miR-2
miR-30d	lung cancer	In summary, our results suggest that miR-2
miR-205	lung carcinoma	NA
miR-205	cancer	In summary, our results suggest that miR-2
miR-205	lung cancer	In summary, our results suggest that miR-2
miR-21	tumor	Our results suggest that tumor miR-21, mi
miR-21	lung cancer	NA
miR-21	cancer	While the level of serum miR-21 was increas
miR-21	NSCLC	Our results suggest that tumor miR-21, mi
miR-21	lymph node metastasis	Overexpression of serum miR-21 was stron
miR-200c	tumor	In addition, this study, for the first time, ide
miR-200c	lung cancer	NA
miR-200c	cancer	While the level of serum miR-21 was increas
miR-200c	NSCLC	In addition, this study, for the first time, ide
miR-141	tumor	Our results suggest that tumor miR-21, mi
miR-141	tumor	While the level of serum miR-21 was increas

MIRNA EXTRACTION

```
contextSearch(subsetabs(liverCancer,a), c("miRNA","mir","m
if (file.exists("companion.txt")==FALSE) {
  next #if no miRNA found, next
} else {
  rnasearch <- read_file("companion.txt")
  rnasearch <- strsplit(rnasearch," ")[[1]]
  mir <- grep("miR", ignore.case = TRUE, rnasearch, value
  mir <- paste(c(mir,grep("let", ignore.case = TRUE, rnase
```

- Extract Sentences: contextSearch()
- Extract miRs: grep()
- Clean miR List

Slashed

mir-21/mir-22

mir-21/22

Attached

mir-21-overexpressed

Numberless

mir, miRNA, antimir, oncomir

Special characters

miR-21, (miR-21)

RELATION EXTRACTION

```
for (i in 1:length(mir)) { #for loop within for
  for (j in 1:length(diseases)) { #for loop al
    unlink("testco_occurrence.txt")
    #search for co-occurrence between one disea
    diseaseTerm <- substr(x=diseases[j], start
                          stop=which(strsplit(
#extract MESH of disease, sometimes is "No
if (grep(pattern = "[[:digit:]]", disease
  MESH <- "No Data"
} else {
  MESH <- substr(x=diseases[j], start=whic
                 stop=nchar(diseases[j]))
}
co_occurrence_fn(mir[i], subsetabs(LiverCa
```

- Nested for loops
- Extract Sentences:
co_occurrence_fn()
- Filter Sentences:
 - read_lines()
 - strsplit()
 - grep()

RESULTS

PMID	Disease	MESH	miRNA	Relation	Organism	Country
30256056	chronic hepatitis	D056487	miR-34a	So both miR-34a and miR-183 were suit...	Human	India
30256056	Cirrhosis	D005355	miR-34a	So both miR-34a and miR-183 were suit...	Human	India
30256056	chronic hepatitis	D056487	miR-183	So both miR-34a and miR-183 were suit...	Human	India
30256056	Cirrhosis	D005355	miR-183	So both miR-34a and miR-183 were suit...	Human	India
30127924	hepatocellular carcinoma	D006528	miR-122	Exosomal microRNAs (miRNAs) have bee...	Human	Japan
30127924	HCC	D006528	miR-122	Exosomal microRNAs (miRNAs) have bee...	Human	Japan
30127924	HCC	D006528	miR-122	Taken together, our results demonstrate ...	Human	Japan
30127924	tumor	D009369	miR-122	The expression levels of exosomal miR-...	Human	Japan
30127924	liver cirrhosis	D008103	miR-122	According to the median relative expres...	Human	Japan
30127924	liver cirrhosis	D008103	miR-122	Taken together, our results demonstrate ...	Human	Japan
30127924	hepatocellular carcinoma	D006528	miR-21	Exosomal microRNAs (miRNAs) have bee...	Human	Japan
30127924	HCC	D006528	miR-21	Exosomal microRNAs (miRNAs) have bee...	Human	Japan
30065664	Hepatocellular Carcinoma	D006528	mir-21	Extracellular Vesicle-Associated mir-21 ...	Human	China/Au...

- Comprehensive Data Frame
- Relation & Non-Relation Versions
- Applicable to Multiple Diseases
- Time-efficient

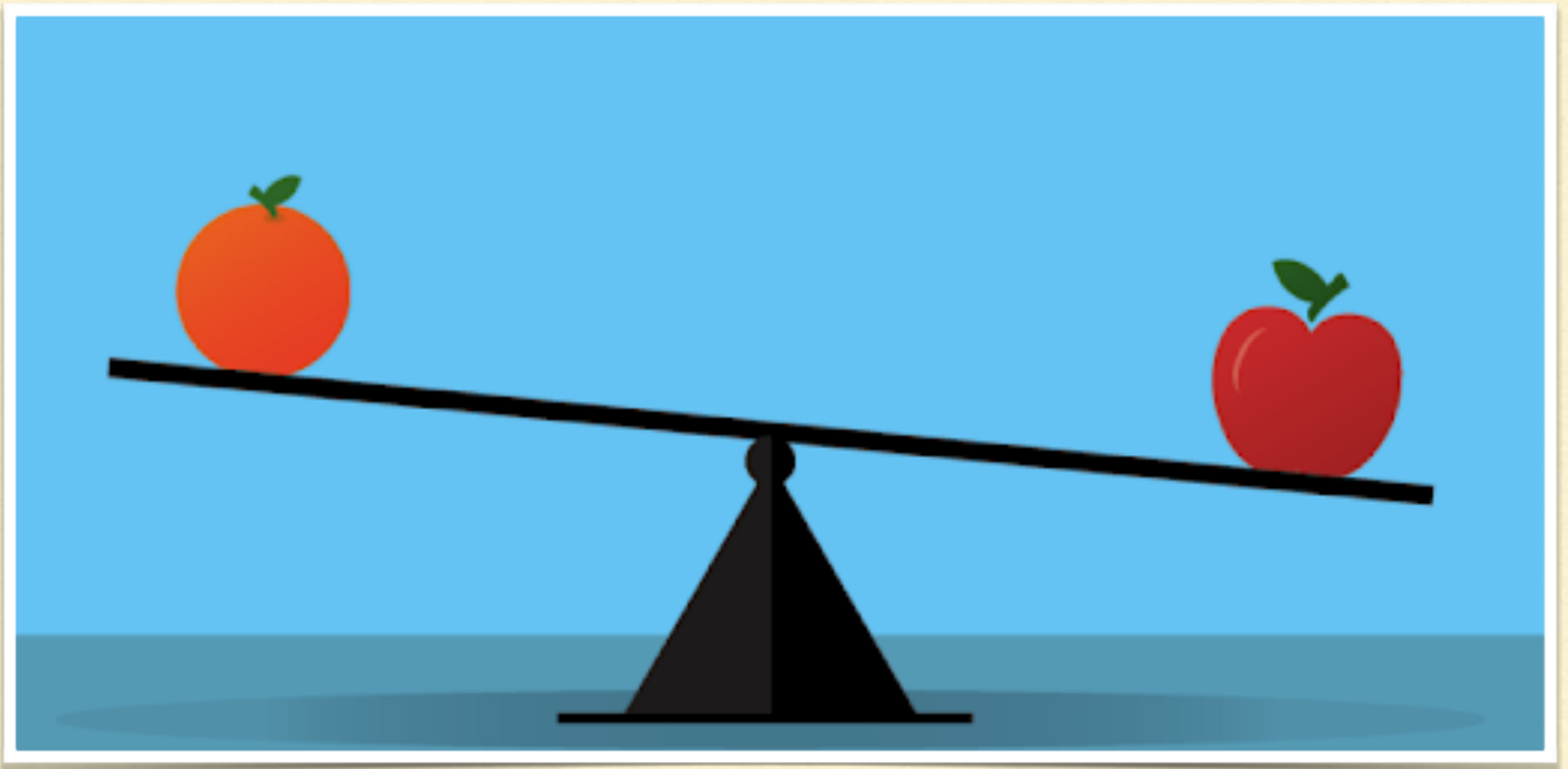
DIFFICULTIES

- Finding the correct package
- Filtering out smaller problems
- Combining major parts of code
- Formatting miRNAs to match HMDD

```
nic/ ↵
rattle")
to '/usr/local/lib/R/3.2/site-library'
(ied)
ependency 'RGtk2'

red % Xferd Average Speed Time Time Time Current
      Dload Upload Total Spent Left Speed
  0 0 0 0 ---:---:---:---:---:---:---: 0 0 (
19k 43 1188k 0 0 796k 0 0:00:03 0:00:01 0:00:02
:-- 1355k
red % Xferd Average Speed Time Time Time Current
      Dload Upload Total Spent Left Speed
  0 0 0 0 ---:---:---:---:---:---: 0 0 (
12k 13 339k 0 0 228k 0 0:00:11 0:00:01 0:00:10
08 241k 34 2602k 34 908k 0 0 252k 0 0:00:10 0:
0:00:04 0:00:04 311k 76 2602k 76 1980k 0 0 350k
0 0:00:06 0:00:06 ---:---:---: 481k
' package 'RGtk2' ...
ccessfully unpacked and MD5 sums checked
fig... /usr/local/bin/pkg-config
is at least version 0.9.0... yes
CTION... no
IO
( version 2.8.0 required
failed for package 'RGtk2'
l/lib/R/3.2/site-library/RGtk2'
ackages :
ackage 'RGtk2' had non-zero exit status
'gtk2' is not available for package 'rattle'
l/lib/R/3.2/site-library/rattle'
ackages :
ackage 'rattle' had non-zero exit status

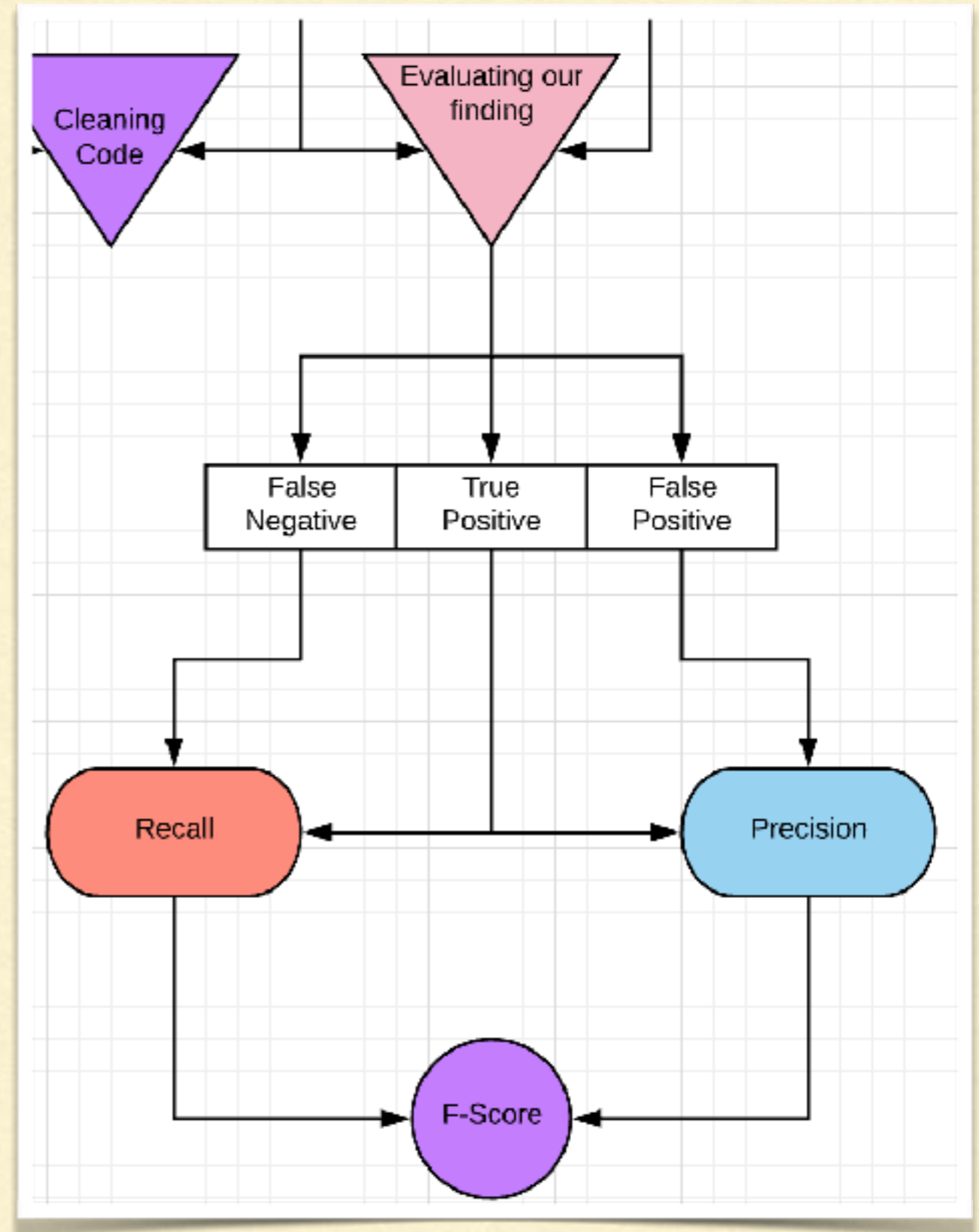
: packages are in
v/folders/cz/9kx_fjf17492w0m696hfs_80000gp/T/RtmpwWspk/downloader
```



EVALUATION

METHODOLOGY

- Evaluate the significance of the findings
- Compare to existing databases
- Conclude if the data is reliable enough for use



COMPARISON

- Baseline: HMDD Abstracts
- Official **mirList** and our **mymirList**
- Expected Systematic Error

Examples of TP, FN, FP

TP: “mir-21” in both mirList & mymirList

FN: “mir-26a” only in mirList

FP: “let-7e” only in mymirList

```
for (check in 1:length(mirList))
  if (any(grepl(pattern = mirList[check], mymirList)))
    TP = TP+1
    tpositive <- append(tpositive, check)
  next
}
if (all(grepl(pattern = mirList, mymirList)))
  FN=FN+1
  fnegative <- append(fnegative, check)
next
}
}
FP=length(mymirList)-TP
```

STATISTICS

	TP	FN	FP	RECALL	PRECISION	FSCORE
Alex Liver	71	10	46	0.877	0.607	0.717
Matt Liver	50	39	18	0.562	0.735	0.634
Alex Lung	38	14	6	0.731	0.864	0.792
Matt Lung	31	25	12	0.554	0.721	0.626



IMPROVEMENTS

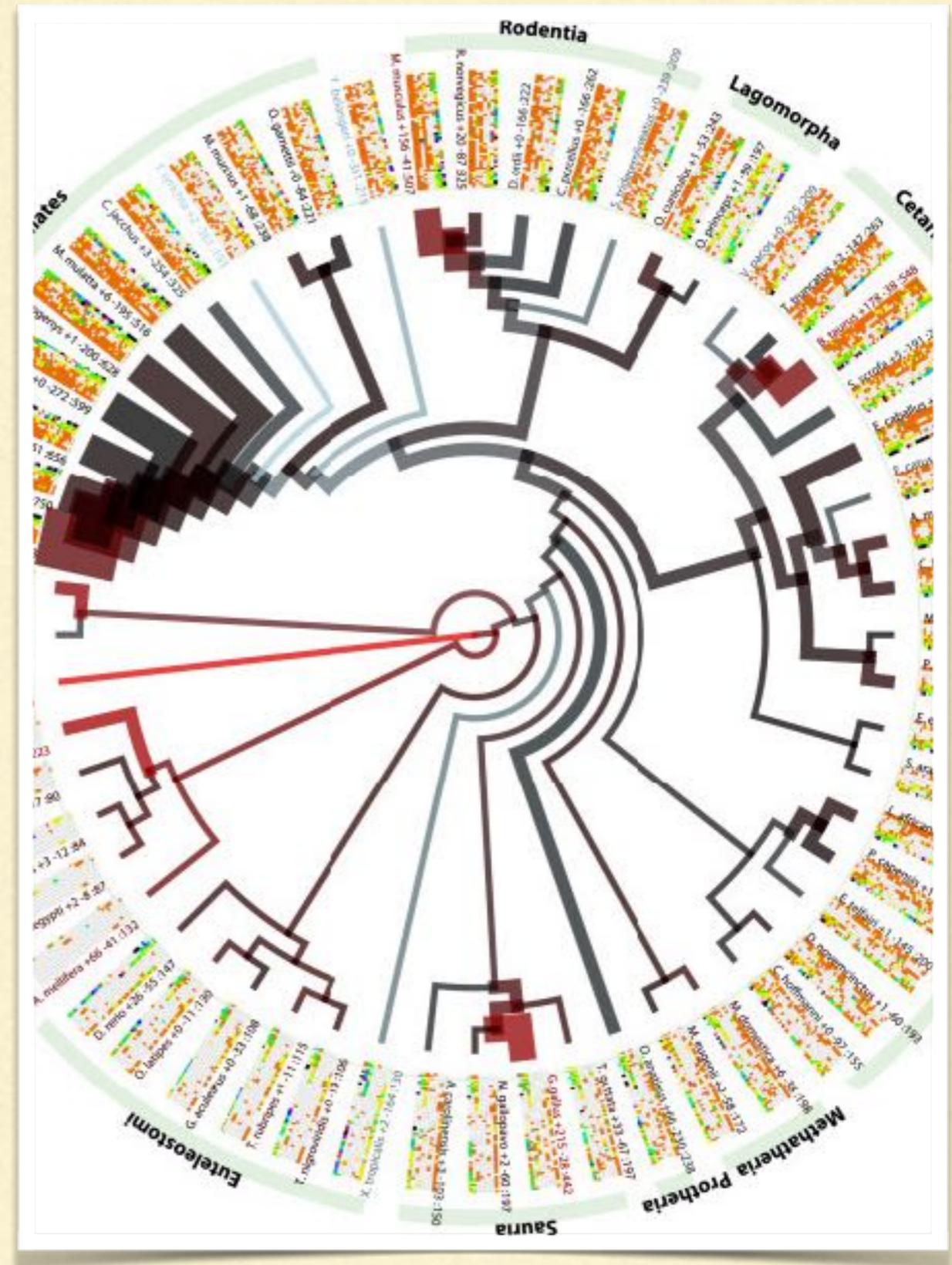
PROGRAMMING

- Difficult miR keywords
 - “miR-29c/DNMTs/miR-34c\
\449a”
 - “miR-106a/b”
- let- miR terms
- Difficult Abstracts
 - Language
- Disease & MeSH ID Matching

```
_check_fn <-  
function(x){allcheck <- c(letters[1:26],  
y1 <- allcheck != 'r'  
lcheck <- allcheck[key1]  
  
w_test <- c()  
rd <- c()  
for (i in x){  
  mirstring_split <- strsplit(i, "")[[1]]  
  for (j in mirstring_split){  
    logic <- j == allcheck  
    if (any(logic) == FALSE) {  
      if (j == "r"){j="R"}  
      else {j=""}}  
    word <- paste(word, j, sep="")  
  }  
  new_test <- append(new_test, word)  
  word <- c()  
}  
w_test <- unique(new_test)  
return(new_test)
```

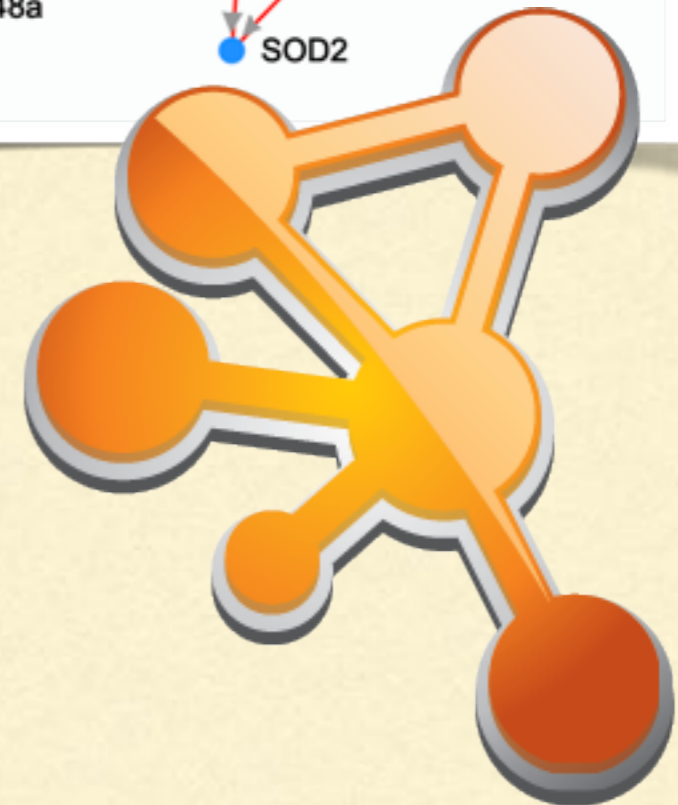
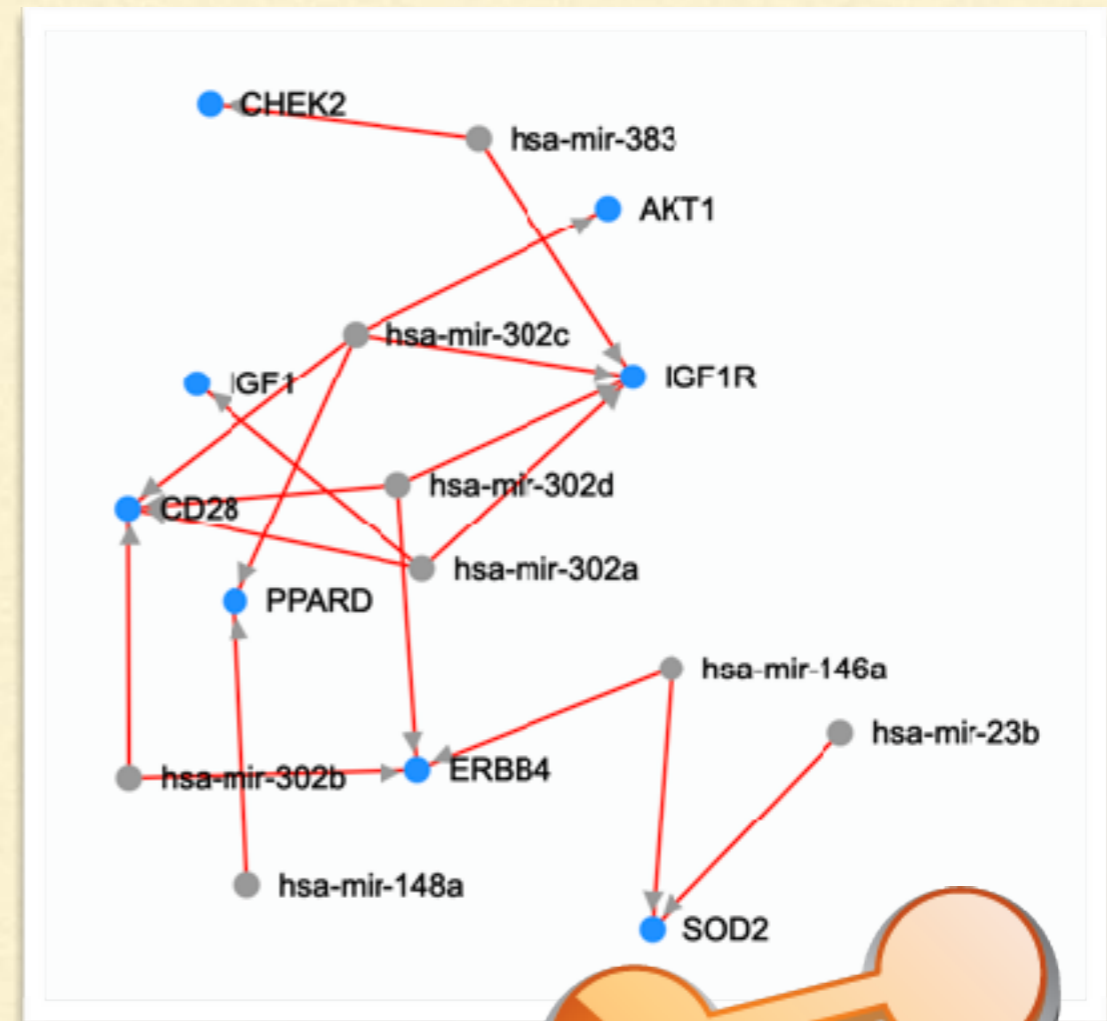
EVALUATION

- Source of False Negative
 - Paper Access
- Crosscheck Databases
- Family miRNA terms
 - “mir-1” & “mir-1-1”
 - “mir-200” & “mir-200a”
 - “mir-26a” & “mir-26a-1”



GENERAL

- Data Visualization via Cytoscape
- Online Database (HMDD)
- Relevant Data Table Features
 - Positive/Negative Relation
 - Relation Extraction Methodology



Thank you Dr. Karathanasis and the rest of the
Jefferson Computational Medicine Center!

–Matt Laws & Alex Liang
