Final Project: Bibliometric Analysis of *Astyanax mexicanus* Research Papers

By: Serena Amro Gazze

Introduction

I had a great time learning all about R throughout this course and appreciated the opportunity to apply my new R skills in real-world data science projects. So, for my final project, I decided to test out my new skills that I learned from Chapters 22 and 23 of your textbook *Data Science for the Liberal Arts* by conducting a bibliometric case study.

Specifically, I conducted a case study where I analyze a bibliometric network created from a collection of research papers on the *Astyanax mexicanus*, or Mexican cavefish. I decided to do this case study because my research lab at the FAU Wilkes Honors College uses the *Astyanax mexicanus* as our model species in our studies on evolutionary neuroscience, and I wanted to learn more about the research community that also studies cavefish.

The Astyanax cavefish is a fish that has adapted to life in complete darkness from living in underwater caves, leading to unique evolutionary changes such as loss of eyesight and pigmentation. This project allowed me to dive deeper into the scholarly world surrounding the fish species that I actively study in lab.

Load Libraries

```r
# Load required libraries
suppressPackageStartupMessages({
  library(tidyverse)
  library(bibliometrix)
  library(igraph)
  library(ggraph)
  library(qgraph)
  library(tidygraph)
  library(kableExtra)
  library(reshape2)
  library(pheatmap)
})
```

Load and Convert Bibliographic Data

All of these research papers were downloaded from the Web of Science database using the search query “Astyanax mexicanus” in a plain text format on 4/20/2025.

```r
# Define path to the folder containing the Web of Science file
dataDir <- "data/Astyanax2025"

# Retrieve the full file path of file in directory
filenames <- list.files(dataDir, pattern = "*.txt", full.names = TRUE)

# Convert the file into a bibliographic data frame
biblioDF <- data.frame(as.list(filenames)) %>%
  convert2df()
```

```
##
## Converting your wos collection into a bibliographic dataframe
##
## Done!
##
##
## Generating affiliation field tag AU_UN from C1:  Done!
```

View Cited References from First Paper

```r
# Display the cited references from the first paper in dataset
biblioDF %>%
  # Convert row names into a column named "source"
  rownames_to_column("source") %>%
  # Display only paper name and its cited references
  select(source, CR) %>%
  head(1) %>%
  kable(caption = "Cited references in first paper") %>%
  kable_styling()
```

Cited references in first paper

source	CR

source	CR
COGHILL LM, 2014, MOL PHYLOGENET EVOL	ANDREWS S., 2012, BABRAHAM BIOINFORMAT, P56;ARBOGAST BS, 2001, J BIOGEOGR, V28, P819, DOI 10.1046/J.1365-2699.2001.00594.X;AVISE JC, 1972, EVOLUTION, V26, P1, DOI 10.1111/J.1558-5646.1972.TB00170.X;AVISE JC, 1989, EVOLUTION, V43, P1192, DOI 10.1111/J.1558-5646.1989.TB02568.X;AVISE JC, 1998, P ROY SOC B-BIOL SCI, V265, P1707, DOI 10.1098/RSPB.1998.0492;BOLLBACK JP, 2006, BMC BIOINFORMATICS, V7, DOI 10.1186/1471-2105-7-88;BRADIC M, 2012, BMC EVOL BIOL, V12, DOI 10.1186/1471-2148-12-9;CATCHEN J, 2013, MOL ECOL, V22, P3124, DOI 10.1111/MEC.12354;DARRIBA D, 2012, NAT METHODS, V9, P772, DOI 10.1038/NMETH.2109;DOWLING TE, 2002, MOL BIOL EVOL, V19, P446, DOI 10.1093/OXFORDJOURNALS.MOLBEV.A004100;EDGAR RC, 2004, NUCLEIC ACIDS RES, V32, P1792, DOI 10.1093/NAR/GKH340;EMERSON KJ, 2010, P NATL ACAD SCI USA, V107, P16196, DOI 10.1073/PNAS.1006538107;ESQUIVEL-BOBADILLA S., 2011, THESIS CTR INVESTIGA;GALLO ND, 2012, PLOS ONE, V7, DOI 10.1371/JOURNAL.PONE.0041443;GOMPERT Z, 2010, MOL ECOL, V19, P2455, DOI 10.1111/J.1365-294X.2010.04666.X;GROSS JB, 2013, HEREDITY, V111, P122, DOI 10.1038/HDY.2013.26;GROSS JB, 2013, PLOS ONE, V8, DOI 10.1371/JOURNAL.PONE.0055659;GROSS JB, 2012, BMC EVOL BIOL, V12, DOI 10.1186/1471-2148-12-105;GROSS JB, 2009, PLOS GENET, V5, DOI 10.1371/JOURNAL.PGEN.1000326;GUINDON S, 2003, SYST BIOL, V52, P696, DOI 10.1080/10635150390235520;HAILER F, 2012, SCIENCE, V336, P344, DOI 10.1126/SCIENCE.1216424;HAUSDORF B, 2011, MOL PHYLOGENET EVOL, V60, P89, DOI 10.1016/J.YMPEV.2011.03.009;HICKERSON MJ, 2010, MOL PHYLOGENET EVOL, V54, P291, DOI 10.1016/J.YMPEV.2009.09.016;HOHENLOHE PA, 2010, PLOS GENET, V6, DOI 10.1371/JOURNAL.PGEN.1000862;HOLSINGER KE, 2010, MOL ECOL, V19, P2361, DOI 10.1111/J.1365-294X.2010.04667.X;HUELSENBECK JP, 2003, SYST BIOL, V52, P641, DOI 10.1080/10635150390235467;HULSEY CD, 2013, ECOL EVOL, V3, P2262, DOI 10.1002/ECE3.633;JOBB G, 2004, BMC EVOL BIOL, V4, DOI 10.1186/1471-2148-4-18;KOLACZKOWSKI B, 2008, MOL BIOL EVOL, V25, P1054, DOI 10.1093/MOLBEV/MSN042;LANGECKER T.G., 1991, ICHTHYOLOGICAL EXPLORATION OF FRESHWATERS, V2, P209;LEAVITT DH, 2007, MOL ECOL, V16, P4455, DOI 10.1111/J.1365-294X.2007.03496.X;LEWIS PO, 2001, SYST BIOL, V50, P913, DOI 10.1080/106351501753462876;LUIKART G, 2003, NAT REV GENET, V4, P981, DOI 10.1038/NRG1226;MADDISON WP, 2011, MESQUITE MODULAR SYS;ORNELAS-GARCIA CLAUDIA PATRICIA, 2008, BMC EVOL BIOL, V8, P340, DOI 10.1186/1471-2148-8-340;POTTIN K, 2011, DEVELOPMENT, V138, P2467, DOI 10.1242/DEV.054106;PROTAS ME, 2006, NAT GENET, V38, P107, DOI 10.1038/NG1700;PROTAS M, 2007, CURR BIOL, V17, P452, DOI 10.1016/J.CUB.2007.01.051;RASQUIN P, 1951, J EXP ZOOL, V117, P317, DOI 10.1002/JEZ.1401170206;SHIMODAIRA H, 2002, SYST BIOL, V51, P492, DOI 10.1080/10635150290069913;STRECKER U, 2004, MOL PHYLOGENET EVOL, V33, P469, DOI 10.1016/J.YMPEV.2004.07.001;STRECKER U, 2012, MOL PHYLOGENET EVOL, V62, P62, DOI 10.1016/J.YMPEV.2011.09.005;SWOFFORD D., 1993, PAUP: PHYLOGENETIC ANALYSIS USING PARSIMONY;TEYKE T, 1990, BRAIN BEHAV EVOLUT, V35, P23, DOI 10.1159/000115853;VONEIDA TJ, 1976, J COMP NEUROL, V165, P89, DOI 10.1002/CNE.901650108;YOSHIZAWA M, 2012, EVOLUTION, V66, P2975, DOI 10.1111/J.1558-5646.2012.01651.X

Viewing the cited references of the first paper in my data frame shows me that the file uploaded correctly and how the references are formatted in the data frame to use for my later code.

Run Bibliometric Analysis

```r
# Perform a bibliometric analysis on the full dataset
twoModeStats <- biblioAnalysis(biblioDF)

# Generate a summary from the bibliometric stats
twoModeSummary <- twoModeStats |> summary()
```

```
## 
## 
## MAIN INFORMATION ABOUT DATA
## 
##  Timespan                              1990 : 2025 
##  Sources (Journals, Books, etc)        33 
##  Documents                             50 
##  Annual Growth Rate %                  -1.15 
##  Document Average Age                  8.48 
##  Average citations per doc             21.64 
##  Average citations per year per doc    2.106 
##  References                            1850 
##  
## DOCUMENT TYPES                     
##  article                         49 
##  article; proceedings paper      1 
##  
## DOCUMENT CONTENTS
##  Keywords Plus (ID)                    200 
##  Author's Keywords (DE)                172 
##  
## AUTHORS
##  Authors                               184 
##  Author Appearances                    256 
##  Authors of single-authored docs       1 
##  
## AUTHORS COLLABORATION
##  Single-authored docs                  1 
##  Documents per Author                  0.272 
##  Co-Authors per Doc                    5.12 
##  International co-authorships %        28 
##  
## 
## Annual Scientific Production
## 
##  Year    Articles
##     1990        3
##     2002        1
##     2003        1
##     2006        1
##     2007        1
##     2010        1
##     2011        1
##     2012        1
##     2013        2
##     2014        1
##     2015        1
##     2016        1
##     2018       10
##     2019        3
##     2020        3
##     2021        3
##     2022        4
##     2023        4
##     2024        6
##     2025        2
##
## Annual Percentage Growth Rate -1.15
##
##
## Most Productive Authors
##
##    Authors        Articles Authors        Articles Fractionalized
## 1      ROHNER N         13     ROHNER N                     2.368
## 2      JEFFERY WR        5     JEFFERY WR                   1.333
## 3      KOWALKO JE        5     SOARES D                     1.083
## 4      KRISHNAN J        5     TEYKE T                      1.000
## 5      PEUSS R           5     KOWALKO JE                   0.887
## 6      ESPINASA L        4     RÉTAUX S                     0.811
## 7      TABIN CJ          4     ESPINASA L                   0.754
## 8      BISWAS T          3     PEUSS R                      0.750
## 9      HASSAN H          3     MA L                         0.750
## 10     MA L              3     KRISHNAN J                   0.719
##
##
## Top manuscripts per citations
##
##                           Paper                                            DOI  TC TCperYear
NTC
## 1  DOWLING TE, 2002, MOL BIOL EVOL       10.1093/oxfordjournals.molbev.a004100 142      5.92
1.000
## 2  TEYKE T, 1990, BRAIN BEHAV EVOLUT     10.1159/000115853                     124      3.44
2.906
## 3  O'QUIN KE, 2013, PLOS ONE             10.1371/journal.pone.0057281           76      5.85
1.048
## 4  KLAASSEN H, 2018, DEV BIOL            10.1016/j.ydbio.2018.03.014            71      8.88
2.795
## 5  BIBLIOWICZ J, 2013, EVODEVO           10.1186/2041-9139-4-25                 69      5.31
0.952
## 6  COGHILL LM, 2014, MOL PHYLOGENET EVOL 10.1016/j.ympev.2014.06.029            58      4.83
1.000
## 7  PATTON P, 2010, J COMP PHYSIOL A      10.1007/s00359-010-0567-8              58      3.62
1.000
## 8  FRANZ-ODENDAAL TA, 2006, EVOL DEV     10.1111/j.1525-142X.2006.05078.x       53      2.65
1.000
## 9  XIONG SL, 2018, DEV BIOL              10.1016/j.ydbio.2018.06.003            52      6.50
2.047
## 10 STAHL BA, 2019, DEV DYNAM             10.1002/dvdy.32                        32      4.57
2.043
##
##
## Corresponding Author's Countries
##
##          Country Articles   Freq SCP MCP MCP_Ratio
## 1 USA                  32 0.6531  23   9     0.281
## 2 MEXICO                7 0.1429   4   3     0.429
## 3 CANADA                3 0.0612   3   0     0.000
## 4 FRANCE                2 0.0408   1   1     0.500
## 5 GERMANY               2 0.0408   2   0     0.000
## 6 UNITED KINGDOM        2 0.0408   2   0     0.000
## 7 CROATIA               1 0.0204   0   1     1.000
## 
## 
## SCP: Single Country Publications
## 
## MCP: Multiple Country Publications
## 
## 
## Total Citations per Country
## 
##     Country      Total Citations Average Article Citations
## 1 USA                        668                      20.9
## 2 MEXICO                     128                      18.3
## 3 FRANCE                      75                      37.5
## 4 CANADA                      56                      18.7
## 5 UNITED KINGDOM              27                      13.5
## 6 GERMANY                      4                       2.0
## 7 CROATIA                      0                       0.0
## 
## 
## Most Relevant Sources
## 
##                                                                   Sources         Articles
## 1  DEVELOPMENTAL BIOLOGY                                                                 8
## 2  JOURNAL OF EXPERIMENTAL ZOOLOGY PART B-MOLECULAR AND DEVELOPMENTAL EVOLUTION         4
## 3  JOVE-JOURNAL OF VISUALIZED EXPERIMENTS                                                4
## 4  PEERJ                                                                                 3
## 5  INTERNATIONALE REVUE DER GESAMTEN HYDROBIOLOGIE                                       2
## 6  SUBTERRANEAN BIOLOGY                                                                  2
## 7  ACTA CARSOLOGICA                                                                      1
## 8  ANNALS OF THE NEW YORK ACADEMY OF SCIENCES                                            1
## 9  BEHAVIORAL ECOLOGY AND SOCIOBIOLOGY                                                   1
## 10 BIOLOGICAL RHYTHM RESEARCH                                                            1
## 
## 
## Most Relevant Keywords
## 
##    Author Keywords (DE)      Articles Keywords-Plus (ID)     Articles
## 1         ASTYANAX MEXICANUS       28   EVOLUTION                  21
## 2         CAVEFISH                 26   BLIND CAVEFISH             12
## 3         ASTYANAX                 10   CAVEFISH                   12
## 4         >                         7   FISH                       12
## 5         EVOLUTION                 6   ADAPTATION                  9
## 6         ADAPTATION                4   TELEOSTEI                   8
## 7         DEVELOPMENT               3   CONVERGENCE                 7
## 8         TROGLOMORPHY              3   EYE DEGENERATION            7
## 9         ALARM SUBSTANCE           2   EXPRESSION                  6
## 10        CAVE                      2   REGRESSIVE EVOLUTION        6
```

As a researcher in the *Astyanax mexicanus* field, I have read many cavefish papers. But, running this bibliometric analysis gave me a clear overview of the entire *Astyanax mexicanus* research landscape.

Seeing the publication timeline and annual article count helped me understand when Astyanax research began gaining traction and becoming more widely studied.

One of the most exciting parts of the analysis for me was the most productive authors section because I am working with Dr. Johanna Kowalko on my current research project. So, I was amazed to see that she is such a big name in this field, being the 3rd most productive author.

Viewing the published countries section showed me how collaborative and widespread the *Astyanax mexicanus* research community is, reaching from the USA to France to Croatia.

Finally, examining the most relevant keywords was valuable for me as a researcher because it gave me a better sense of which terms to use in future paper searches to efficiently find the most relevant *Astyanax mexicanus* research papers.

Improving Growth Rate Measurement using Linear Regression

```
# Make articles per year data frame
annual_counts <- biblioDF %>%
  count(PY, name = "Articles") %>%
  complete(PY = full_seq(range(PY), 1),
           fill = list(Articles = 0)) %>%
  arrange(PY)

# Fit a linear regression model
growth_model <- lm(Articles ~ PY, data = annual_counts)

# View full model summary
summary(growth_model)
```

```
##
## Call:
## lm(formula = Articles ~ PY, data = annual_counts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4917 -0.9291 -0.1924  0.2562  7.3922
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -231.65680   55.29176  -4.190 0.000187 ***
## PY             0.11609    0.02754   4.215 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.717 on 34 degrees of freedom
## Multiple R-squared:  0.3432, Adjusted R-squared:  0.3239
## F-statistic: 17.77 on 1 and 34 DF,  p-value: 0.0001742
```

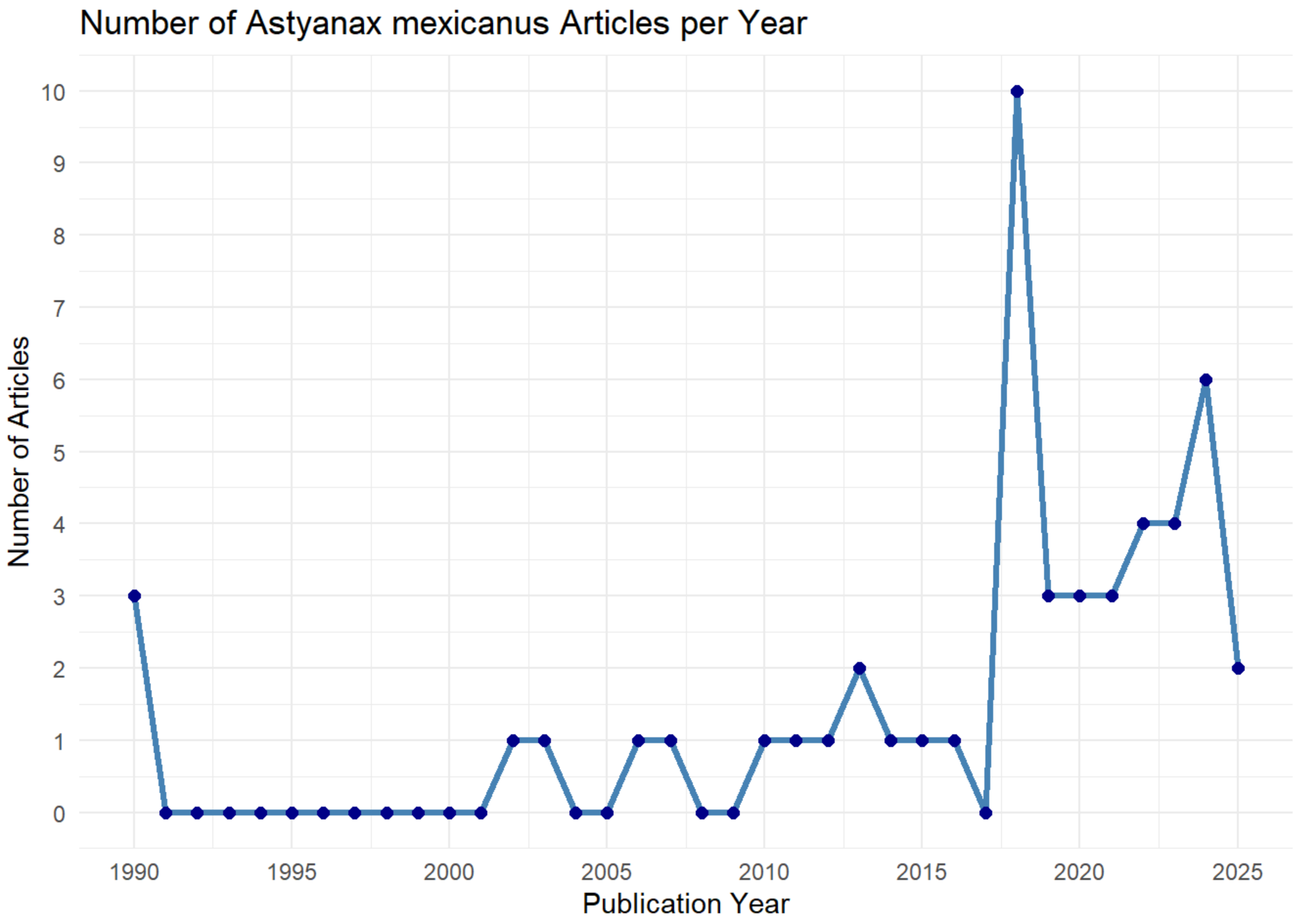
```
# Extract just the slope (coefficient for PY)
growth_slope <- coef(growth_model)["PY"]
cat("Average annual growth rate (slope):", round(growth_slope, 2), "articles per year")
```

```
## Average annual growth rate (slope): 0.12 articles per year
```

While examining the Bibliometric Analysis summary, I noticed that the growth rate statistic (Annual Percentage Growth Rate) used was not very informative because it calculates the annual growth rate solely based on the number of publications in the first and last years divided by the total time span. To better understand the long-term publication trends, I used linear regression to model the relationship between publication year and the number of research articles published. The regression output produced a statistically significant ($p < 0.001$) slope of 0.116, indicating that the field has experienced a consistent average increase of approximately 0.12 articles per year. Unlike APGR, this slope accounts for all data points across the timeline and provides a more stable and accurate picture of the overall trajectory of Astyanax research publications.

Plot Article Count Per Year

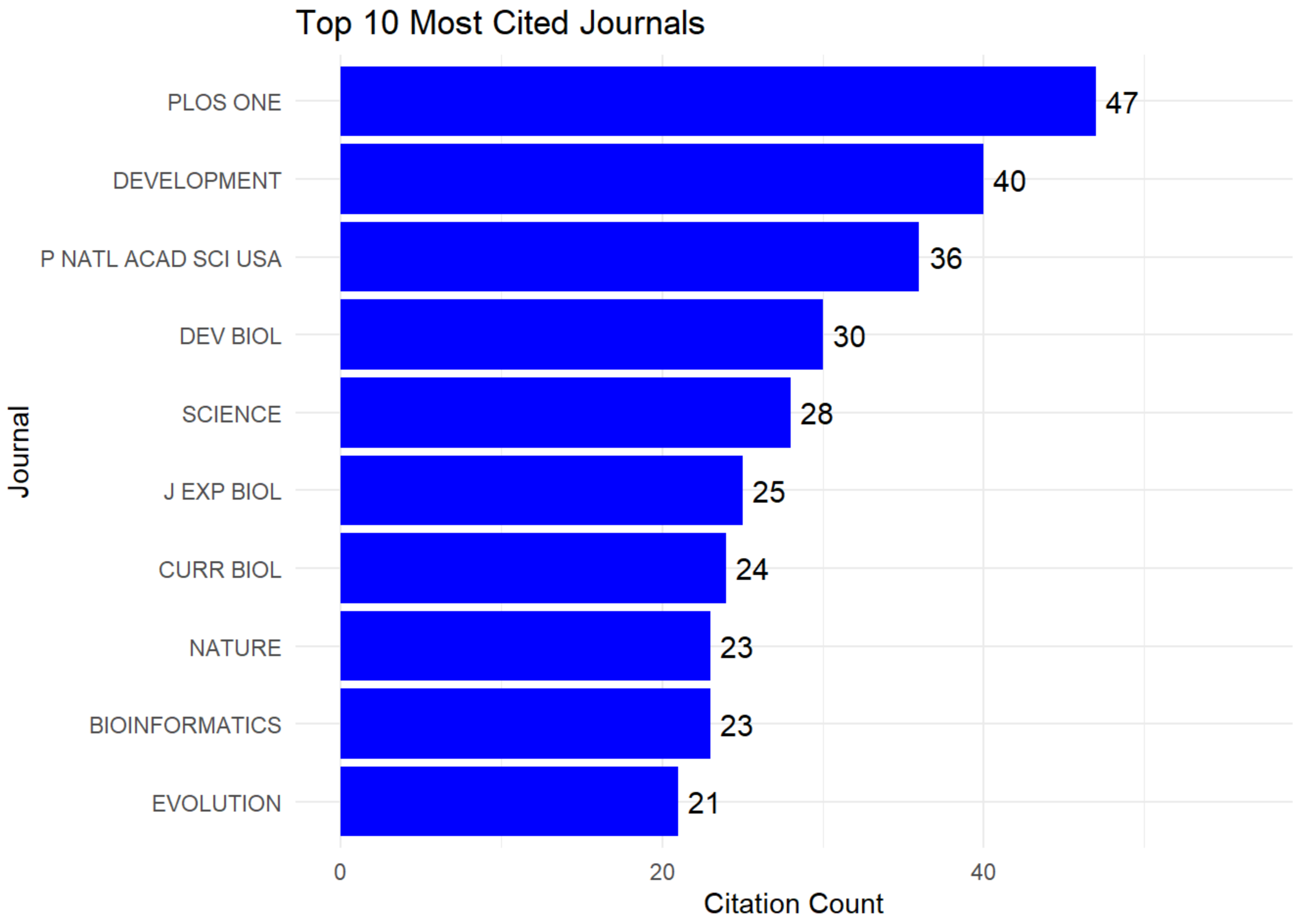
```
# Create articles per year line graph
ggplot(annual_counts, aes(x = PY, y = Articles)) +
  geom_line(color = "steelblue", linewidth = 1.2) +
  geom_point(color = "darkblue", size = 2) +
  labs(
    title = "Number of Astyanax mexicanus Articles per Year",
    x = "Publication Year",
    y = "Number of Articles"
  ) +
  scale_y_continuous(breaks = seq(0, max(annual_counts$Articles), by = 1)) +
  scale_x_continuous(breaks = seq(1990, 2025, by = 5)) +
  theme_minimal()
```



This line graph visually supports the linear regression findings by illustrating the year-by-year publication trends in Astyanax research. As you can see, there was minimal activity throughout the 1990s and early 2000s, followed by a gradual increase and a sharp spike in 2018. After this spike, the number of publications slightly declined but remained higher than in earlier decades, with little year-to-year fluctuations. This trend suggests a surge of interest around 2018, followed by a possible plateau or shift in research focus.

Identify and Visualize Most Cited Journals

```r
# Extract most cited journals from bibliographic data
sourceJournals <- biblioDF %>%
  citations(field = "article", sep = ";") %>%
  pluck(3) %>%
  as_tibble() %>%
  rename(CitedJournal = 1)

# Count and sort top 10 most cited journals
top10Journals <- sourceJournals %>%
  count(CitedJournal) %>%
  arrange(desc(n)) %>%
  head(10)

# Create top 10 most cited journals bar graph
ggplot(top10Journals, aes(x = reorder(CitedJournal, n), y = n)) +
  geom_col(fill = "blue") +
  geom_text(aes(label = n), hjust = -0.3, size = 4) +
  coord_flip() +
  labs(
    title = "Top 10 Most Cited Journals",
    x = "Journal",
    y = "Citation Count"
  ) +
  theme_minimal() +
  expand_limits(y = max(top10Journals$n) * 1.2)
```



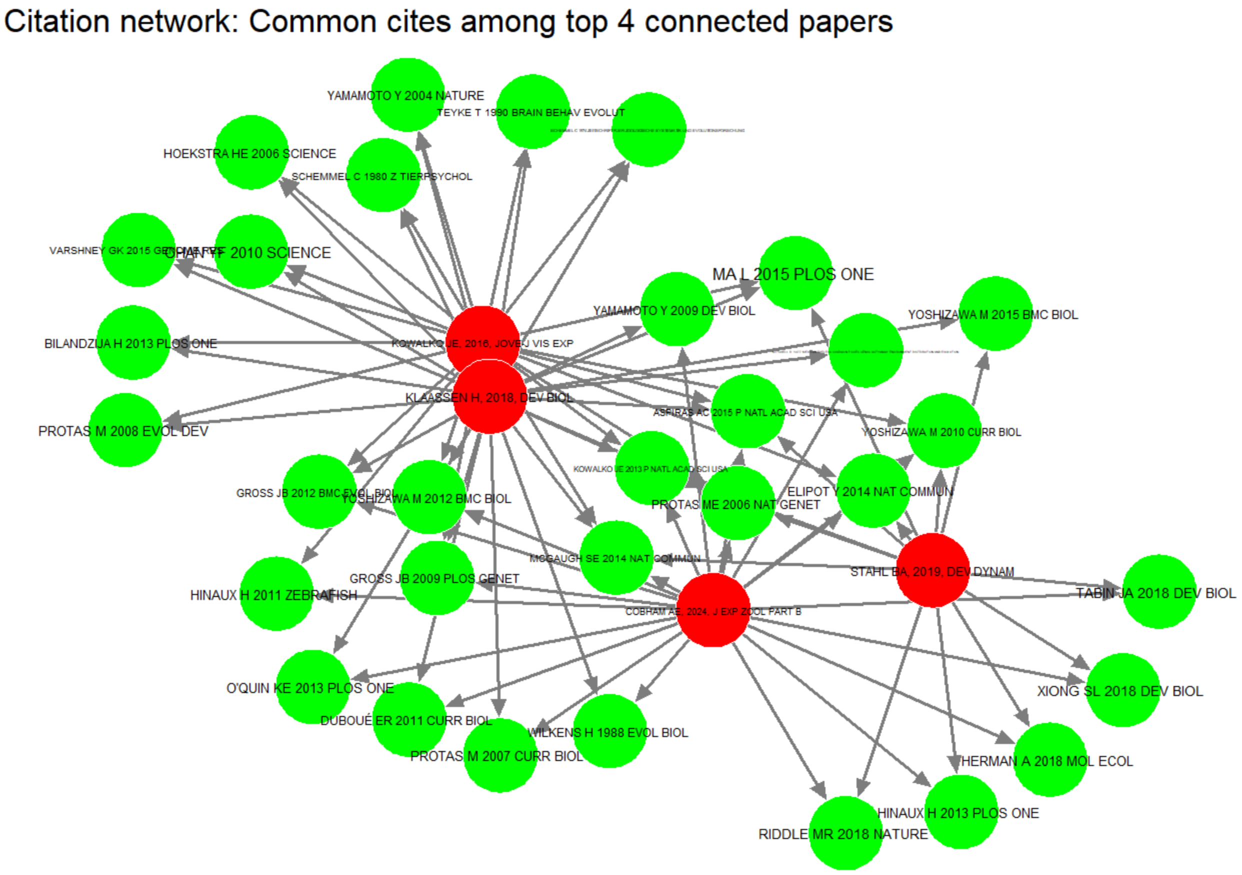
Analyzing the most cited journals helped me understand which publications are central to the field. As a student researcher who's still learning where to look for the best and most reputable research, this list is incredibly valuable. Journals like PLOS ONE and PNAS stand out not only for their high citation counts but also for their interdisciplinary reach. This list helps me set goals of where to publish in the future or where I should focus my reading for the best papers.

Build and Visualize a Small Citation Network: A Network of the Top 4 Source Papers with the Most

Connected Citations

```
# Create a full bibliographic coupling matrix
fullCouplingMatrix <- biblioNetwork(biblioDF, analysis = "coupling", network = "references", sep = ";") %>%
  # Normalize coupling matrix using the Salton index
  normalizeSimilarity(type = "salton") %>%
  as.matrix()

# Remove self-loops
diag(fullCouplingMatrix) <- 0

# Identify top 4 most connected papers based on total shared references
top4_names <- names(sort(rowSums(fullCouplingMatrix), decreasing = TRUE))[1:4]

# Add an ID column to bibliographic data and filter to keep only the top 4 source papers
someSources <- biblioDF %>%
  mutate(ID = rownames(fullCouplingMatrix)) %>%
  filter(ID %in% top4_names)

# Build co-citation edge list for top 4 connected papers
smallBiblioDF <- someSources %>%
  # Generate a document-by-reference matrix
  cocMatrix(Field = "CR", sep = ";") %>%
  as.matrix() %>%
  # Melt the matrix to long format
  reshape2::melt() %>%
  # Keep only meaningful connections
  filter(value > 0) %>%
  # Rename columns for clarity
  rename(source = 1, citation = 2) %>%
  # Drop the raw count
  select(-value) %>%
  # Ensure all identifiers are character strings
  mutate(source = as.character(source), citation = as.character(citation)) %>%
  # Remove self-links where a paper cites itself
  filter(source != citation) %>%
  # Group by citation to count how many papers share that reference
  group_by(citation) %>%
  # Keep only references cited by more than one paper
  filter(n() > 1) %>%
  ungroup()

# Identify all unique source paper IDs in the network (red nodes)
source_nodes <- unique(smallBiblioDF$source)

# Identify all unique citation targets in the network (green nodes)
citation_nodes <- unique(smallBiblioDF$citation)

# Combine all nodes involved in the network
all_nodes <- unique(c(smallBiblioDF$source, smallBiblioDF$citation))
```

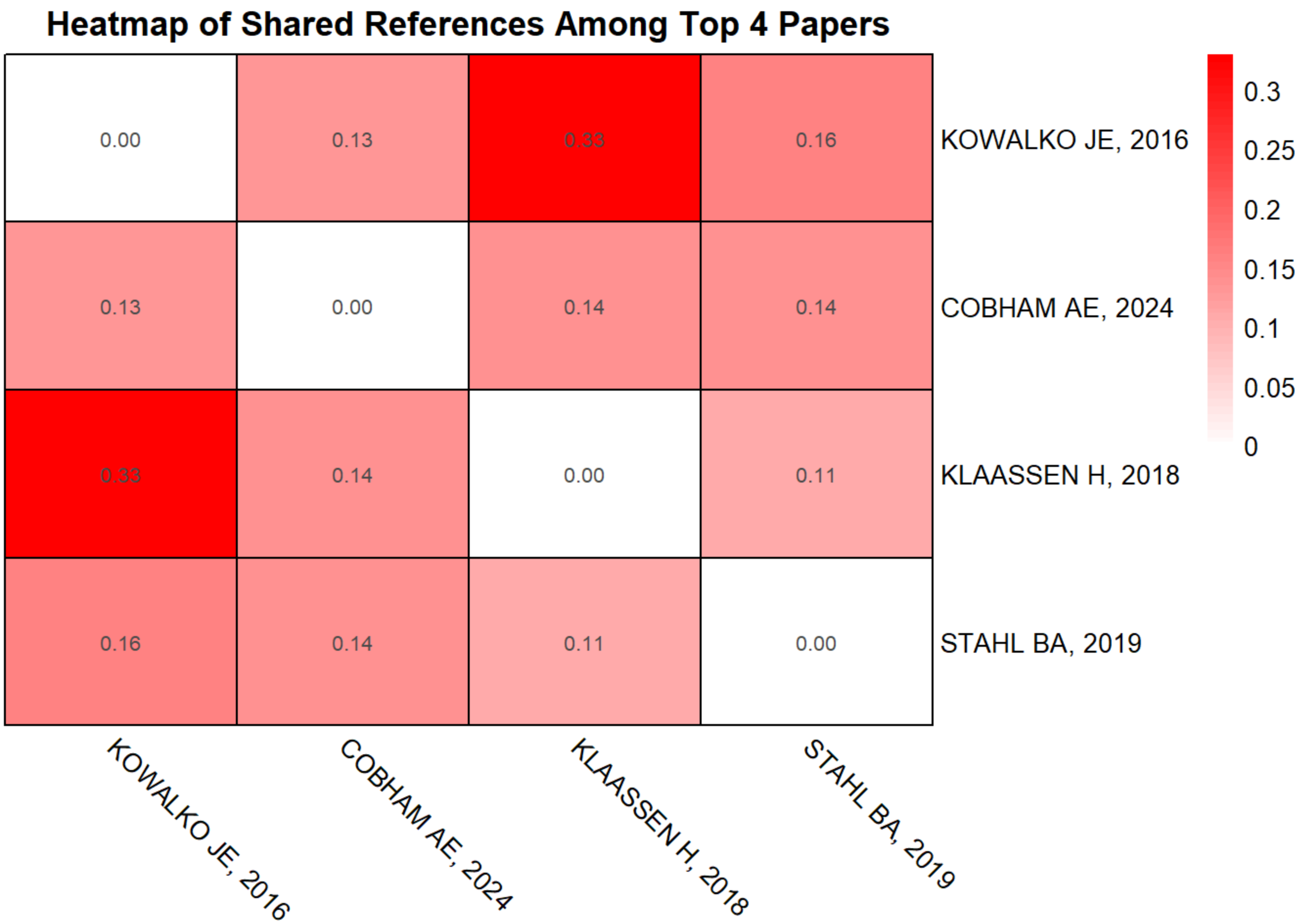
```
# Assign colors: red for sources, green for citations
node_colors <- ifelse(all_nodes %in% source_nodes, "red", "green")

# Build adjacency matrix for citation links
adj_matrix <- matrix(0, nrow = length(all_nodes), ncol = length(all_nodes), dimnames = list(all_nodes, all_nodes))

# Fill the adjacency matrix with citation links
for (i in seq_len(nrow(smallBiblioDF))) {
  from <- smallBiblioDF$source[i]
  to <- smallBiblioDF$citation[i]
  # Create directed edge from source to citation
  adj_matrix[from, to] <- 1
}

# Plot network using qgraph
qgraph::qgraph(
  adj_matrix,
  directed = TRUE,
  labels = rownames(adj_matrix),
  color = node_colors,
  label.cex = 3,
  edge.width = 2,
  border.color = "white",
  title = "Citation network: Common cites among top 4 connected papers"
)
```


Citation network: Common cites among top 4 connected papers
YAMAMOTO Y 2004 NATURE
TEYKE T 1990 BRAIN BEHAV EVOLUT
SCHEMMEL C 1974 ZEITSCHRIFT FUER ZOOLOGISCHE SYSTEMATIK UND EVOLUTIONSFORSCHUNG
HOEKSTRA HE 2006 SCIENCE
SCHEMMEL C 1980 Z TIERPSYCHOL
VARSHNEY GK 2015 GENOME RES
CHAN YF 2010 SCIENCE
MA L 2015 PLOS ONE
YAMAMOTO Y 2009 DEV BIOL
YOSHIZAWA M 2015 BMC BIOL
BILANDZIJA H 2013 PLOS ONE
KOWALKO JE, 2016, JOVE-J VIS EXP
KLAASSEN H, 2018, DEV BIOL
ASPIRAS AC 2015 P NATL ACAD SCI USA
YOSHIZAWA M 2010 CURR BIOL
PROTAS M 2008 EVOL DEV
KOWALKO JE 2013 P NATL ACAD SCI USA
GROSS JB 2012 BMC EVOL BIOL
YOSHIZAWA M 2012 BMC BIOL
PROTAS ME 2006 NAT GENET
ELIPOT Y 2014 NAT COMMUN
MCGAUGH SE 2014 NAT COMMUN
GROSS JB 2009 PLOS GENET
STAHL BA, 2019, DEV DYNAM
HINAUX H 2011 ZEBRAFISH
TABIN JA 2018 DEV BIOL
COBHAM AE, 2024, J EXP ZOOL PART B
O'QUIN KE 2013 PLOS ONE
XIONG SL 2018 DEV BIOL
DUBOUÉ ER 2011 CURR BIOL
WILKENS H 1988 EVOL BIOL
PROTAS M 2007 CURR BIOL
HERMAN A 2018 MOL ECOL
HINAUX H 2013 PLOS ONE
RIDDLE MR 2018 NATURE

This graph shows a directed citation network constructed from the top four *Astyanax mexicanus* research papers with the most shared references, identified using bibliographic coupling. In the visualization, red nodes represent the four most connected source papers, while green nodes indicate the commonly cited articles. Arrows point from each red node to the green nodes it cites, illustrating the flow of knowledge and highlighting frequently shared references. The density of connections and overlap among the cited works reveals which studies are foundational across multiple top papers. By focusing on only the most interconnected source papers, this network emphasizes the core papers shaping the field and provides a clearer understanding of the intellectual structure of *Astyanax mexicanus* research.

Create Heatmap Matrix Table of Small Top 4 Network

```r
# Create matrix of shared references from top 4 papers
smallOneModeMatrix <- biblioNetwork(
  someSources,
  analysis = "coupling",
  network = "references",
  sep = ";"
) %>%
  # Normalize the similarity scores using the Salton index
  normalizeSimilarity(type = "salton") %>%
  as.matrix() %>%
  round(2)

# Remove self-loops
diag(smallOneModeMatrix) <- 0

# Store the matrix in a variable for plotting
heatmap_matrix <- as.matrix(smallOneModeMatrix)

# Plot heatmap of the shared reference matrix
pheatmap(
  heatmap_matrix,
  cluster_rows = FALSE,
  cluster_cols = FALSE,
  display_numbers = TRUE,
  main = "Heatmap of Shared References Among Top 4 Papers",
  color = colorRampPalette(c("white", "red"))(100),
  fontsize_row = 10,
  fontsize_col = 10,
  border_color = "black",
  angle_col = 315
)
```

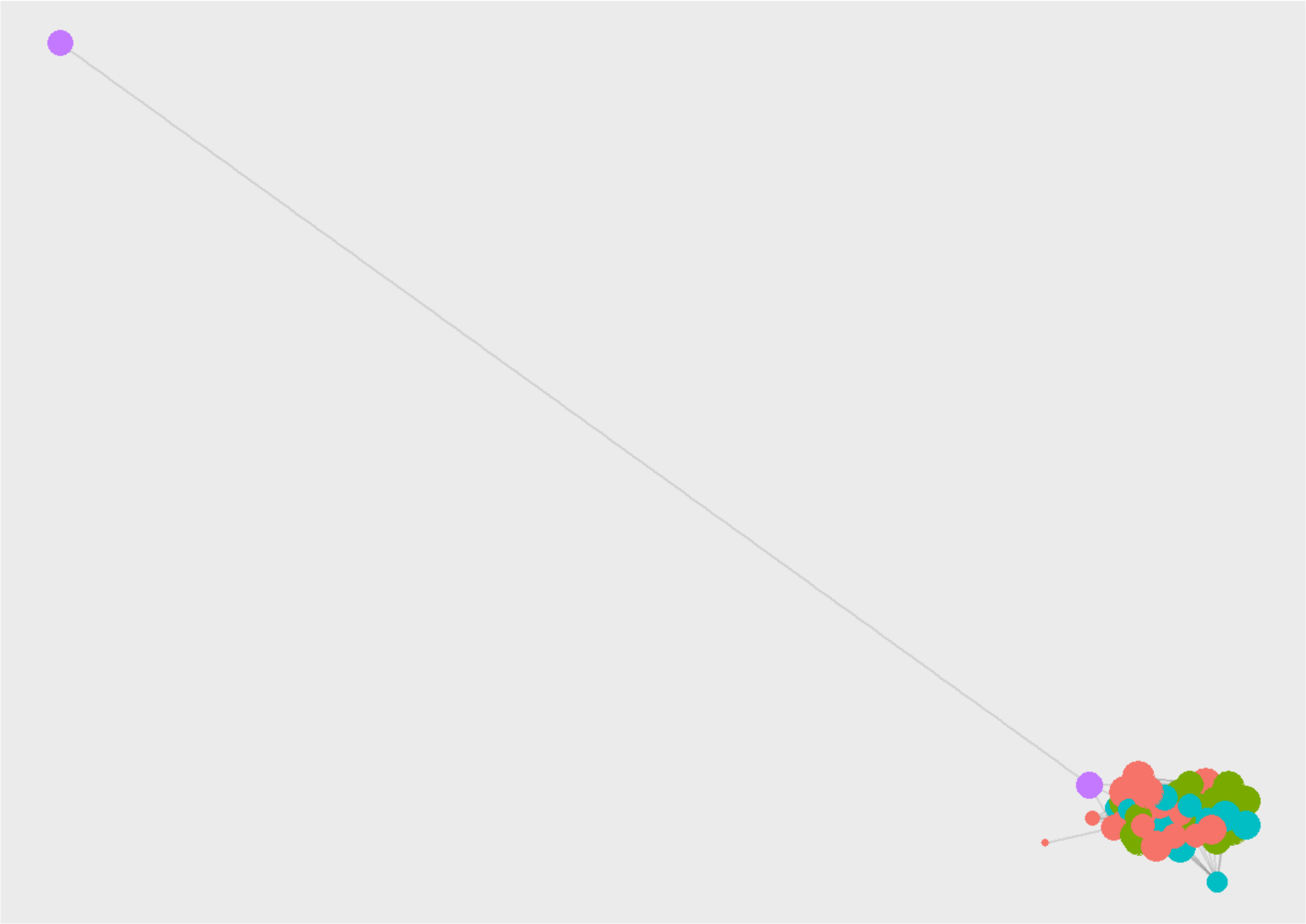


The heatmap provided a clearly understandable way to assess the similarity of the papers based on shared references between the top four most connected *Astyanax mexicanus* research papers. The color gradient from white to red represents increasing similarity in shared references, with darker red cells indicating a higher degree of overlap. Kowalko JE, 2016 and Klaassen H, 2018 have the strongest coupling (0.33), suggesting that these two papers reference a very similar set of foundational studies. On the other hand, other pairings like Klaassen H, 2018 and Stahl BA, 2019 show low overlap (0.11), suggesting that while the two papers share some conceptual or methodological foundations, they likely focus on different aspects of Astyanax mexicanus research. This heatmap helps reveal which papers are conceptually or methodologically aligned based on their reference patterns, offering insights into how closely different studies are linked in the research landscape.

Build and Visualize a Full Citation Network

```r
# Create a full bibliographic coupling matrix using all papers
fullOneModeMatrix <- biblioNetwork(
  biblioDF,
  analysis = "coupling",
  network = "references",
  sep = ";"
) %>%
  # Normalize the similarity scores using the Salton index
  normalizeSimilarity(type = "salton") %>%
  as.matrix() %>%
  round(2)

# Remove self-loops
diag(fullOneModeMatrix) <- 0

# Convert the adjacency matrix to an igraph object for network analysis
oneModeGraph <- fullOneModeMatrix %>%
  graph_from_adjacency_matrix(mode = "undirected", diag = FALSE, weighted = TRUE)

# Convert igraph object to tidygraph and Calculate key network metrics for each node
tidyBibGraph <- oneModeGraph %>%
  as_tbl_graph() %>%
  activate(nodes) %>%
  mutate(
    # PageRank centrality
    centralPR = centrality_pagerank(weights = weight),
    # Rank papers from most to least central
    nodePRRank = rank(-centralPR),
    # Degree centrality (number of connections)
    central0D = centrality_degree(weights = NULL),
    # Community detection using Louvain algorithm
    communityLouv = group_louvain(weights = weight),
    # Convert community labels to factor for plotting
    group = as.factor(group_louvain()),
    # Assign a unique ID to each node
    ID = row_number()
  )

# Extract edges and nodes as data frames
edgeList <- tidyBibGraph %>% activate(edges) %>% as_tibble()
nodeList <- tidyBibGraph %>% activate(nodes) %>% as_tibble()
```

```
# Set max number of nodes to plot and min edge weight to show
nNodesToPlot <- 794
minEdgeWeightToPlot <- 0.00001

# Plot entire network using ggraph
tidyBibGraph %>%
  activate(nodes) %>%
  filter(central0D > 0) %>%
  filter(nodePRRank < nNodesToPlot) %>%
  activate(edges) %>%
  filter(weight > minEdgeWeightToPlot) %>%
  ggraph(layout = 'stress') +
  geom_edge_link(alpha = 0.1) +
  geom_node_point(aes(size = centralPR, color = as.factor(communityLouv))) +
  theme(legend.position = "none")
```



This graph presents a full citation network of all of the *Astyanax mexicanus* research papers using a bibliographic coupling approach, where connections are based on shared references. The layout reveals a tightly clustered main body of research, with papers grouped by community color and sized by PageRank centrality to reflect their influence.

Notably in this graph, a single purple node appears in the upper left, entirely separated from the dense core, linked only by a single thin edge. This outlier indicates a paper that shares few references with the rest of the literature—suggesting it may represent a highly specialized study or explore a novel direction within the field. Identifying and analyzing outliers is important, as they can highlight unique contributions or signal emerging subfields in *Astyanax mexicanus* research.

Identify Outlier Node

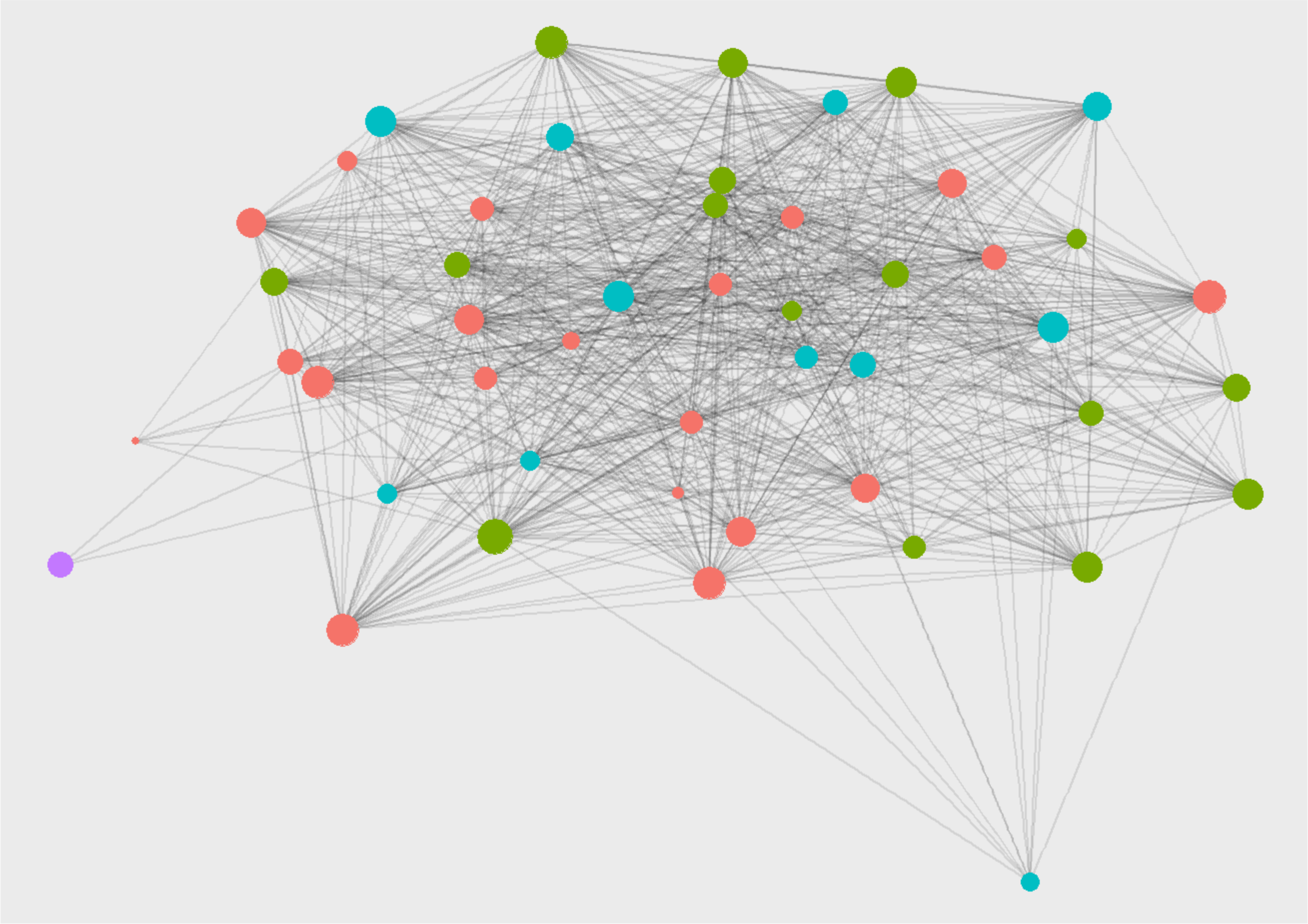
```
# Find the ID of the outlier node with the lowest degree
outlier_node <- nodeList %>%
  filter(central0D == min(central0D)) %>%
  select(ID) %>%
  pull()

# Match this node ID to bibliographic metadata
outlier_info <- biblioDF %>%
  mutate(ID = row_number()) %>%
  filter(ID == outlier_node) %>%
  select(TI, AB)

# Display outlier's title and abstract
print(outlier_info)
```

```
##
TI
## VALDÉZ-MORENO ME, 2003, P BIOL SOC WASH SKULL OSTEOLOGY OF THE CHARACID FISH ASTYANAX MEXICAN
US (TELEOSTEI: CHARACIDAE)
##
AB
## VALDÉZ-MORENO ME, 2003, P BIOL SOC WASH THE SKULL OF THE CHARACID FISH ASTYANAX MEXICANUS IS
DESCRIBED BASED ON TWENTY ALIZARIN-STAINED ADULT SPECIMENS FROM RIO SALADO, RIO CONCHOS, RIO ALA
MO, AND RIO SAN JUAN POPULATIONS, ALL OF THEM RIO GRANDE TRIBUTARIES IN NORTHEASTERN MEXICO. THE
SKULL HAS A CIRCULAR SHAPE IN LATERAL VIEW. THE SECOND INFRAORBITAL IS TRIANGULAR AND NEVER OVER
LAPS THE INFERIOR MARGIN OF THE THIRD INFRAORBITAL. THE THIRD INFRAORBITAL NEVER REACHES THE LAT
EROSENSORY CANAL OF THE PREOPERCULAR BONE. THE SUPRAOCCIPITAL IS SHORT. THE PALATINES, ECTOPTERY
GOIDS AND MESOPTERYGOIDS LACK TEETH. WE DESCRIBE TWO FEATURES NOT REPORTED IN ASTYANAX BEFORE: T
EETH ON THE SECOND SUSPENSORY PHARYNGEAL AND POSTERIOR GILL RAKERS ON THE FOUR GILL ARCHES. DIFF
ERENCES IN SKULL OSTEOLOGY BETWEEN A. MEXICANUS AND OTHER DESCRIBED SPECIES OF ASTYANAX ARE DENO
TED.
```

This outlier node is a 2003 paper by Valdéz-Moreno titled “Skull Osteology of the Characid Fish Astyanax mexicanus.” It examines the skull osteology of Astyanax mexicanus from surface populations in northeastern Mexico. Its unique focus, older publication date, and lack of citations to mainstream cavefish literature likely contribute to its bibliographic isolation, as it does not heavily overlap in references with the more frequently cited, interconnected cavefish citation clusters. However, its distinct position highlights its potential value as a specialized but under-referenced resource.

Visualize Full Network Excluding Outlier

```r
# Set max number of nodes to plot and min edge weight to show
nNodesToPlot <- 794
minEdgeWeightToPlot <- 0.00001

# Visualize the entire network excluding outliers
tidyBibGraph %>%
  # Work with node data
  activate(nodes) %>%
  # Filter out nodes with very few connections
  filter(central0D > 1) %>%
  # Retain only top nodes based on PageRank centrality
  filter(nodePRRank < nNodesToPlot) %>%
  # Switch to edge data
  activate(edges) %>%
  # Filter edges to include only those with meaningful bibliographic coupling
  filter(weight > minEdgeWeightToPlot) %>%
  # Create a graph layout using the 'stress' algorithm
  ggraph(layout = 'stress') +
  # Faint, semi-transparent lines to represent citation connections between nodes
  geom_edge_link(alpha = 0.1) +
  # Plot each paper as a node (sized by PageRank and colored by community)
  geom_node_point(
    aes(
      # Node size reflects PageRank
      size = centralPR,
      # Node color reflects Louvain community group
      color = as.factor(communityLouv)
    )
  ) +
  # Hide the legend
  theme(legend.position = "none")
```



This visualization represents the full bibliographic coupling network after filtering out the most extreme outlier node. The network graph now shows a more cohesive and interconnected structure, highlighting the major clusters of papers, which are color-coded by Louvain community detection, that frequently share references. By removing the outlier, the overall layout is tightened, which enhances the visibility of the internal structures and relationships within the network. Notably in the graph, the purple node still appears in an isolated community, indicating it remains relatively isolated compared to the other research clusters. This helps spotlight another niche paper that might contribute a unique perspective to the Astyanax research field.

Find Paper in Isolated Community

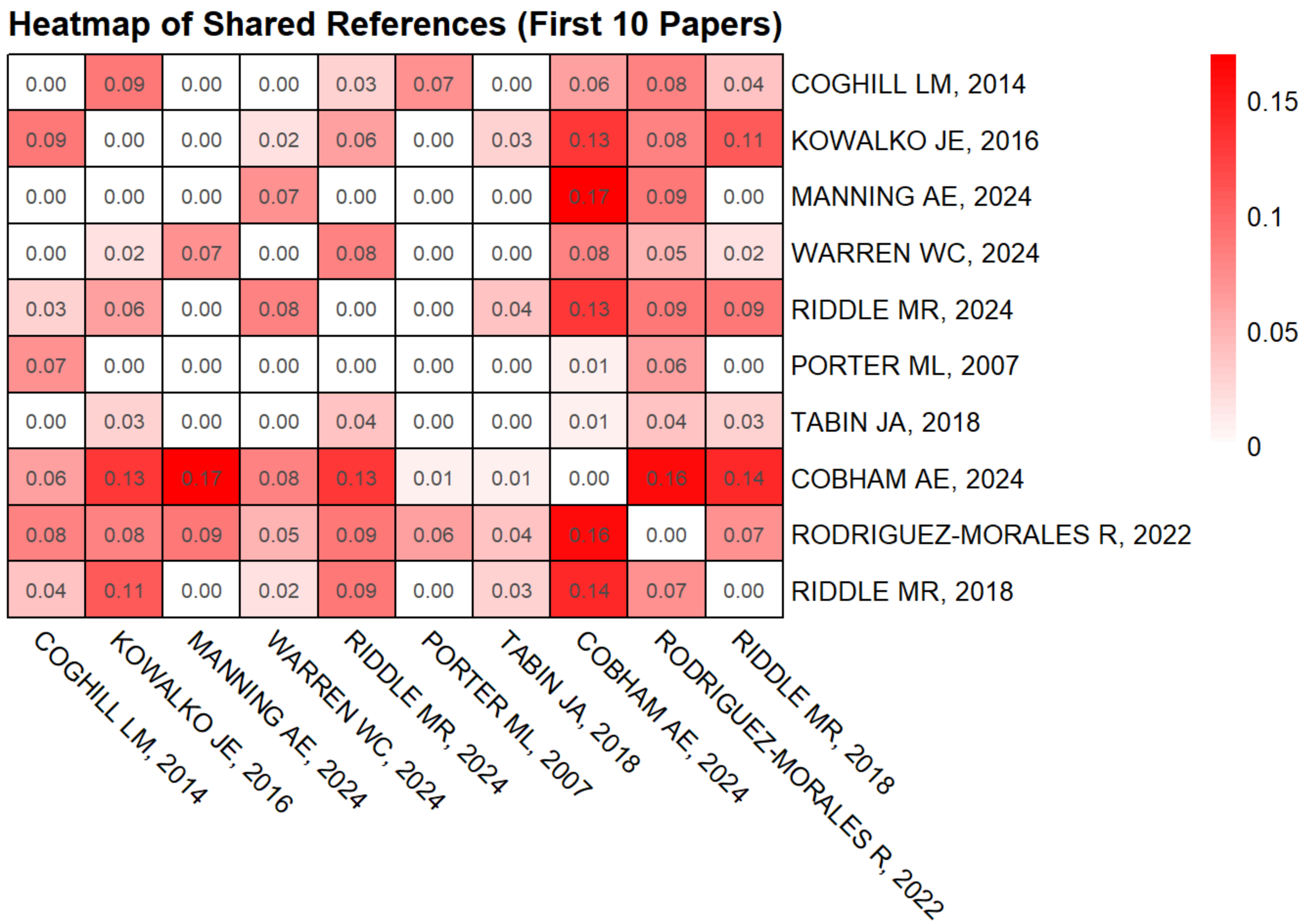
```r
# Find community with the fewest members
rarest_community <- nodeList %>%
  count(communityLouv) %>%
  arrange(n) %>%
  slice(1) %>%
  pull(communityLouv)

# Get node in rarest community
rarest_node <- nodeList %>%
  filter(communityLouv == rarest_community)

# Match metadata from biblioDF
rarest_metadata <- biblioDF %>%
  mutate(ID = row_number()) %>%
  filter(ID %in% rarest_node$ID) %>%
  select(ID, TI, AB, J9, PY)

# Display the Title, Abstract, Journal, and Year of purple node
print(rarest_metadata)
```

```
##                                        ID
## PETERS N, 1990, INT REV GES HYDROBIO   28
## PETERS N, 1990, INT REV GES HYDROBIO-a 29
##
TI
## PETERS N, 1990, INT REV GES HYDROBIO   FINE-STRUCTURE OF THE CLUB CELLS (ALARM SUBSTANCE CELL
S) IN THE EPIDERMIS OF ASTYANAX-MEXICANUS FILIPPI 1853 (CHARACINIDAE, PISCES) AND ITS CAVE FORMS
ANOPTICHTHYS
## PETERS N, 1990, INT REV GES HYDROBIO-a FINE-STRUCTURE OF THE CLUB CELLS (ALARM SUBSTANCE CELL
S) IN THE EPIDERMIS OF ASTYANAX-MEXICANUS FILIPPI 1853 (CHARACINIDAE, PISCES) AND ITS CAVE FORMS
ANOPTICHTHYS
##
AB
## PETERS N, 1990, INT REV GES HYDROBIO   THE CLUB CELLS IN THE EPIDERMIS OF ASTYANAX MEXICANUS
ARE OF THE SAME ULTRASTRUCTURE AS THOSE IN OTHER OSTARIOPHYSAN FISH. THERE ARE ALSO NO ESSENTIAL
DIFFERENCES TO BE FOUND BETWEEN THE CLUB CELLS OF EPIGEAN ASTYANAX MEXICANUS AND ITS CAVE FORMS
"ANOPTICHTHYS". OUR FINDINGS CORRESPOND TO FORMER STATEMENTS IN THAT THE FUNCTION OF THE CLUB CE
LLS PRODUCING AND RELEASING ALARM SUBSTANCES HAS BEEN MAINTAINED IN THE CAVE FORMS.
## PETERS N, 1990, INT REV GES HYDROBIO-a THE CLUB CELLS IN THE EPIDERMIS OF ASTYANAX MEXICANUS
ARE OF THE SAME ULTRASTRUCTURE AS THOSE IN OTHER OSTARIOPHYSAN FISH. THERE ARE ALSO NO ESSENTIAL
DIFFERENCES TO BE FOUND BETWEEN THE CLUB CELLS OF EPIGEAN ASTYANAX MEXICANUS AND ITS CAVE FORMS
"ANOPTICHTHYS". OUR FINDINGS CORRESPOND TO FORMER STATEMENTS IN THAT THE FUNCTION OF THE CLUB CE
LLS PRODUCING AND RELEASING ALARM SUBSTANCES HAS BEEN MAINTAINED IN THE CAVE FORMS.
##                                                           J9   PY
## PETERS N, 1990, INT REV GES HYDROBIO   INT REV GES HYDROBIO 1990
## PETERS N, 1990, INT REV GES HYDROBIO-a INT REV GES HYDROBIO 1990
```

This paper in the isolated community is “Fine-structure of the club cells (alarm substance cells) in the epidermis of Astyanax mexicanus Filippi 1853 and its cave forms Anoptichthys,” written by Peters in 1990. Despite its relevance to evolutionary and physiological studies, it appears in an isolated community within the network, likely because it was completed and published prior to the main wave of modern Astyanax research and doesn’t share references with newer papers in the dataset. Its isolated position highlights how older, foundational work can remain not as relevant in modern citation networks even if it contributes important early findings.

Create Heatmap Matrix Table of Full Network

```
# Subset to first 10 rows and columns
heatmap_matrix <- fullOneModeMatrix[1:10, 1:10]

# Plot heatmap of the first 10x10 shared reference matrix
pheatmap(
  heatmap_matrix,
  cluster_rows = FALSE,
  cluster_cols = FALSE,
  display_numbers = TRUE,
  main = "Heatmap of Shared References (First 10 Papers)",
  color = colorRampPalette(c("white", "red"))(100),
  fontsize_row = 10,
  fontsize_col = 10,
  border_color = "black",
  angle_col = 315
)
```



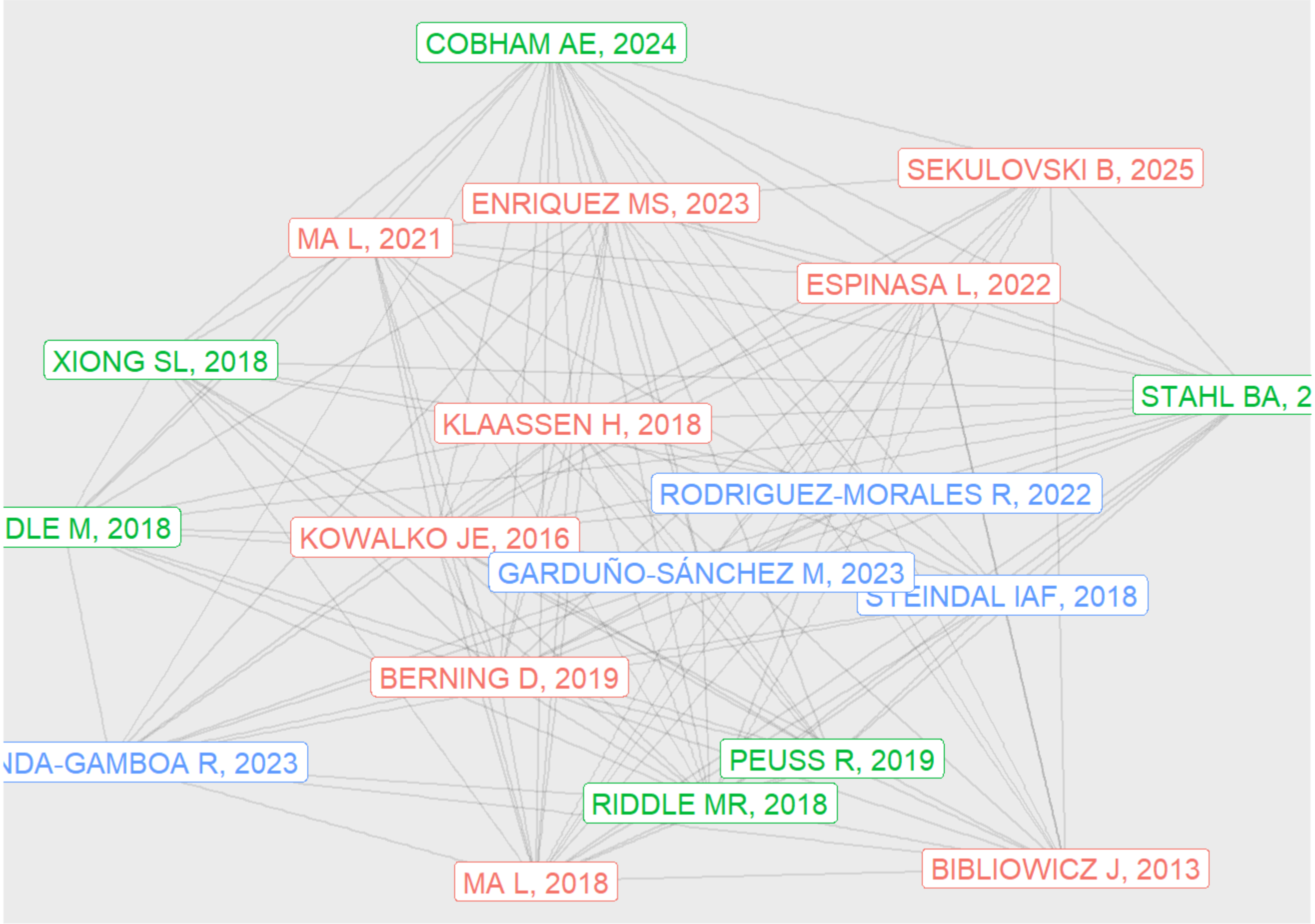
Heatmap of Shared References (First 10 Papers)

	COGHILL LM, 2014	KOWALKO JE, 2016	MANNING AE, 2024	WARREN WC, 2024	RIDDLE MR, 2024	PORTER ML, 2007	TABIN JA, 2018	COBHAM AE, 2024	RODRIGUEZ-MORALES R, 2022	RIDDLE MR, 2018
COGHILL LM, 2014	0.00	0.09	0.00	0.00	0.03	0.07	0.00	0.06	0.08	0.04
KOWALKO JE, 2016	0.09	0.00	0.00	0.02	0.06	0.00	0.03	0.13	0.08	0.11
MANNING AE, 2024	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.17	0.09	0.00
WARREN WC, 2024	0.00	0.02	0.07	0.00	0.08	0.00	0.00	0.08	0.05	0.02
RIDDLE MR, 2024	0.03	0.06	0.00	0.08	0.00	0.00	0.04	0.13	0.09	0.09
PORTER ML, 2007	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.00
TABIN JA, 2018	0.00	0.03	0.00	0.00	0.04	0.00	0.00	0.01	0.04	0.03
COBHAM AE, 2024	0.06	0.13	0.17	0.08	0.13	0.01	0.01	0.00	0.16	0.14
RODRIGUEZ-MORALES R, 2022	0.08	0.08	0.09	0.05	0.09	0.06	0.04	0.16	0.00	0.07
RIDDLE MR, 2018	0.04	0.11	0.00	0.02	0.09	0.00	0.03	0.14	0.07	0.00

This heatmap displays the bibliographic coupling strength based on shared references among the first 10 papers in the full dataset. (This matrix is a 10x10 subset of the full bibliographic coupling matrix because displaying the full matrix was too large, but you can display the full matrix by removing the row and column subsetting.) The darker red cells indicate higher similarity values, meaning those papers cite more of the same sources. For example, MANNING AE, 2024 and COBHAM AE, 2024 share a strong coupling value of 0.17, suggesting close topical or methodological alignment. In contrast, many cells remain at or near 0.00, indicating minimal to no shared references between certain pairs, which reflects the diversity of research focus even within a specific field. This heatmap is useful for visually identifying the clusters of highly interconnected works within the Astyanax field.

Visualize Top 20 Central Papers

```
# Set the number of top nodes to plot based on their PageRank ranking
nNodesToPlot <- 20
# Set minimum edge weight (bibliographic coupling strength)
minEdgeWeightToPlot <- 0.05

# Visualize the top 20 central papers in the network
tidyBibGraph %>%
  # Work with node data
  activate(nodes) %>%
  # Retain only the top nodes based on PageRank centrality
  filter(nodePRRank < nNodesToPlot) %>%
  # Switch to edge data
  activate(edges) %>%
  # Filter edges to include only those with meaningful bibliographic coupling
  filter(weight > minEdgeWeightToPlot) %>%
  # Create a graph layout using the 'stress' algorithm
  ggraph(layout = 'stress') +
  # Faint, semi-transparent lines to represent citation connections between nodes
  geom_edge_link(alpha = 0.1) +
  # Plot each paper as a node (sized by PageRank and colored by community)
  geom_node_point(
    aes(
      # Node size reflects PageRank
      size = centralPR,
      # Node color reflects Louvain community group
      color = as.factor(communityLouv)
    )
  ) +
  # Add labels to nodes to display author names and years
  geom_node_label(
    aes(
      label = name,
      color = as.factor(communityLouv)
    )
  ) +
  # Hide the legend
  theme(legend.position = "none")
```



This graph displays the top 20 most central papers in the Astyanax bibliographic network, determined using the PageRank centrality metric. Each paper is represented as a labeled node, with colors indicating their communities based on the Louvain community detection. The edges between nodes reflect shared references, with a minimum coupling weight of 0.05 required for a connection. This visualization highlights not only the most influential papers but also how they cluster into tightly knit research communities, offering insight into the dominant themes and collaborative patterns in the Astyanax research field.

Community 1 (red) contains papers focused on genetic manipulation and developmental biology, including Kowalko et al. (2016) on TALEN genome editing and Klaassen et al. (2018) on pigmentation via CRISPR mutagenesis. Community 2 (green) includes Cobham et al. (2024) (the most central paper) exploring stress resilience in extreme environments, alongside other methodological studies like Stahl et al. (2019) on transgenesis and Riddle et al. (2018) on phenotyping protocols. Community 3 (blue) contains studies with ecological and phylogeographic emphasis, such as Garduño-Sánchez et al. (2023) and Coghill et al. (2014).

Display Metadata for Top Papers

```
# Create metadata table with abstract, title, keywords, journal, and ID
nodeInfo <- biblioDF %>%
  # Add a unique ID to each paper
  mutate(ID = row_number()) %>%
  # Select abstract, title, keywords, journal, and ID
  select(AB, TI, DE, ID, J9) %>%
  # Combine title, abstract, keywords, and ID into a single searchable text field
  mutate(alltext = paste(TI, AB, DE, ID, sep = " ")) %>%
  # Remove the now-redundant individual metadata columns
  select(-AB, -DE, -ID) %>%
  # Reassign a row ID
  mutate(ID = row_number())

# Join the node metadata to the network node list based on ID
allNodeInfo <- nodeList %>%
  left_join(nodeInfo, by = 'ID')

# Filter and display the top 20 papers by PageRank and community
nodeList %>%
  # Retain only papers ranked in the top 20 by PageRank
  filter(nodePRRank < 21) %>%
  # Join with metadata to get titles for these top papers
  left_join(nodeInfo) %>%
  # Select relevant columns for reporting
  select(name, nodePRRank, communityLouv, TI) %>%
  # Order by community and PageRank
  arrange(communityLouv, nodePRRank) %>%
  kable(title = "Top papers in CSS by PR and community") %>%
  kable_styling()
```

```
## Joining with `by = join_by(ID)`
```

name	nodePRRank	communityLouv	TI
KOWALKO JE, 2016	2	1	GENOME EDITING IN ASTYANAX MEXICANUS USING TRANSCRIPTION ACTIVATOR-LIKE EFFECTOR NUCLEASES (TALENS)
MA L, 2018	3	1	MATERNAL GENETIC EFFECTS IN ASTYANAX CAVEFISH DEVELOPMENT
BIBLIOWICZ J, 2013	4	1	DIFFERENCES IN CHEMOSENSORY RESPONSE BETWEEN EYED AND EYELESS ASTYANAX MEXICANUS OF THE RIO SUBTERRANEO CAVE

name	nodePRRank	communityLouv	TI
KLAASSEN H, 2018	5	1	CRISPR MUTAGENESIS CONFIRMS THE ROLE OF OCA2 IN MELANIN PIGMENTATION IN ASTYANAX MEXICANUS
ENRIQUEZ MS, 2023	13	1	EVIDENCE FOR RAPID DIVERGENCE OF SENSORY SYSTEMS BETWEEN TEXAS POPULATIONS OF THE MEXICAN TETRA (ASTYANAX MEXICANUS)
MA L, 2021	14	1	INCREMENTAL TEMPERATURE CHANGES FOR MAXIMAL BREEDING AND SPAWNING IN ASTYANAX MEXICANUS
BERNING D, 2019	16	1	IN-FRAME INDEL MUTATIONS IN THEGENOME OF THE BLIND MEXICAN CAVEFISH, ASTYANAX MEXICANUS
ESPINASA L, 2022	17	1	LATERALITY IN CAVEFISH: LEFT OR RIGHT FORAGING BEHAVIOR IN ASTYANAX MEXICANUS
SEKULOVSKI B, 2025	18	1	MECHANISMS OF SOCIAL BEHAVIOUR IN THE ANTI-SOCIAL BLIND CAVEFISH (ASTYANAX MEXICANUS)
COBHAM AE, 2024	1	2	UNRAVELING STRESS RESILIENCE: INSIGHTS FROM ADAPTATIONS TO EXTREME ENVIRONMENTS BY ASTYANAX MEXICANUS CAVEFISH
STAHL BA, 2019	6	2	STABLE TRANSGENESIS IN ASTYANAX MEXICANUS USING THE TOL2 TRANSPOSASE SYSTEM
RIDDLE M, 2018	7	2	RAISING THE MEXICAN TETRA ASTYANAX MEXICANUS FOR ANALYSIS OF POST-LARVAL PHENOTYPES AND WHOLE-MOUNT IMMUNOHISTOCHEMISTRY
RIDDLE MR, 2018	8	2	MORPHOGENESIS AND MOTILITY OF THE ASTYANAX MEXICANUS GASTROINTESTINAL TRACT
PEUSS R, 2019	9	2	GAMETE COLLECTION AND IN VITRO FERTILIZATION OF ASTYANAX MEXICANUS
XIONG SL, 2018	15	2	EARLY ADIPOGENESIS CONTRIBUTES TO EXCESS FAT ACCUMULATION IN CAVE POPULATIONS OF ASTYANAX MEXICANUS

name	nodePRRank	communityLouv	TI
GARDUÑO-SÁNCHEZ M, 2023	10	3	PHYLOGEOGRAPHIC RELATIONSHIPS AND MORPHOLOGICAL EVOLUTION BETWEEN CAVE AND SURFACE ASTYANAX MEXICANUS POPULATIONS (DE FILIPPI 1853) (ACTINOPTERYGII, CHARACIDAE)
RODRIGUEZ-MORALES R, 2022	11	3	CONVERGENCE ON REDUCED AGGRESSION THROUGH SHARED BEHAVIORAL TRAITS IN MULTIPLE POPULATIONS OF ASTYANAX MEXICANUS
MIRANDA-GAMBOA R, 2023	12	3	A NEW CAVE POPULATION OF ASTYANAX MEXICANUS FROM NORTHERN SIERRA DE EL ABRA, TAMAULIPAS, MEXICO
STEINDAL IAF, 2018	19	3	DEVELOPMENT OF THE ASTYANAX MEXICANUS CIRCADIAN CLOCK AND NON-VISUAL LIGHT RESPONSES
COGHILL LM, 2014	20	3	NEXT GENERATION PHYLOGEOGRAPHY OF CAVE AND SURFACE ASTYANAX MEXICANUS

This metadata of the top 20 central papers shows that the most influential papers span molecular methods, ecological adaptations, and behavioral evolution, reflecting the interdisciplinary nature of Astyanax research. It is specifically important for me as an undergraduate researcher because it guides me towards relevant papers I should be reading to better understand the field and potential collaborators as I further the research in my own lab.

Conclusion

Through this bibliometric case study, I gained valuable insights into both the technical tools of network analysis in R and the scholarly landscape surrounding *Astyanax mexicanus* research. By combining data cleaning, statistical modeling, and network visualization techniques, I was able to uncover trends in publication activity, identify influential authors and journals, and map out how research papers are connected through shared references. Not only did this deepen my understanding of the field I actively study in my research lab, but it also helped me develop a more strategic approach to engaging with the literature, whether I’m identifying underexplored research niches, targeting high-impact journals, or recognizing collaboration patterns. Overall, this project was a fascinating intersection between my interests in neuroscience and data science, and it has equipped me with tools that I will continue using in my academic and research journey! I hope you have a great summer!