Introduction

With the advent of technologies that accelerate whole genome sequencing (WGS), genetic studies are not anymore limited on well-explored model organisms. Through the identification of evolution-promoted homology among newly sequenced organisms, the conservation of protein domains is elucidated. The discovery of such regions allows for a more accurate functional and structural determination of the studied molecules. Additionally, by examining the conserved motifs of the regulatory elements of orthologous genes, a more rigid understanding of the regulatory mechanisms is gained. In this work, two such organisms are analyzed, belonging to the fungi taxon, Neurospora crassa, a type of read bread mold, and Puccinia gramitis, also known as stem rust. Proteomics analysis is initially performed, giving rise to co-orthologous groups (COGs), followed by the comparative analysis of a certain identified COG, represented by a PIKK/FRAP protein kinase of the Puccinia gramitis species, with NCBI RefSeq identifier XP_003321789.2, and its relation with other, similar, fungi species orthologs. Subsequently, functional regions of that group are examined, making use of Simpson diversity criterion. Finally, the focus is shifted towards the regulatory regions of the corresponding genes of this COG orthologs, and highly conserved motifs are discovered near their transcription start site (TSS), with these being eventually looked up in JASPAR database, for functional and structural characteristics.

Methods/Data Retrieval

FASTA file of all known proteins from the Puccinia gramitis (PG) and Neurospora crassa (NC) species were retrieved from NCBI RefSeq database, with the search terms '"Puccinia graminis"[Organism] AND refseq[Filter]' and '"Neurospora Crassa"[Organism]) AND refseq[Filter]' respectively. To identify COGs, blastp was performed in 4 different ways,in an one-to-self and one-to-another fashion, after having created the appropriate databases from the retrieved proteins. Python scripts, based mainly on BioPython and Pandas modules, were later used to find the best bidirectional hits (BBHs) and determine COGs. For the selected COG,which displayed the highest blastp score, homologs from 23 other species were retrieved, through the usage of blastp online search platform. Multiple sequence alignment and Bootstrap Phylogenetic trees creation were later applied, using ClustalW2 suite (Larkin et al., 2007). Furthermore, internal transcribed spacers (ITS) were queried in NCBI for the species under consideration, in order to construct the species tree. Both trees were constructed using 1000 bootstrap iterations and visualized and annotated using iTOL web interface. Afterwards, a python script was used to measure Simpson diversity of the aligned amino acid (a/a) sequences, and find conserved contiguous, ungapped and of consensus threshold 0.7, regions of minimum size 8, by issuing an upper threshold of 0.3. The identified conserved sequences were examined using batch conserved domain search (Batch CD-Search) utility provided by NCBI (Lu et al., 2020). Subsequently, regulatory regions of the genes corresponding to the studied orthologs were collected, using Biopython and Entrez interface. meme-chip tool from MEME suite (Machanick et al., 2011) was finally used, to identify motifs using meme and stream, locate the motifs relative position to the TSS using centrimo, and compare the findings with the JASPAR 2022 fungi database (Montragon et al., 2021). The found matches were then traced back to UniProt database and biological process enrichment analysis was performed using Panther (Huaiyu et al., 2020).

Results

COGs Collection

Out of the 15835 PG and 5980 NC proteins collected from the NCBI, 3976 PG proteins were found to be orthologous to 3289 NC ones. All the resulting orthologs can be accessed in the file report.xlsx, in the "orthologs" sheet.

PG Proteins	NC Proteins
XP_003321789.2	XP_958289.2,XP_960307.2,XP_960570.2
XP_003333296.2	XP_956489.2,XP_964396.3
XP_003307362.1	XP_957191.3,XP_958141.3
XP_003326732.2	XP_964712.3,XP_011394277.1,XP_011393666.1,XP_960771.2,XP_011394072.1,XP_964346.3
XP_003322017.1	XP_011394745.1,XP_011394746.1
XP_003889740.1	XP_011393274.1,XP_011393275.1
XP_003324647.2	XP_011394236.1,XP_011394235.1
XP_003330716.1	XP_965342.2,XP_962160.3
XP_003337444.1	XP_011393197.1,XP_011393196.1,XP_964993.2
XP_003326781.2	XP_961904.3,XP_963175.2

Table 1 The COGs of NC corresponding to a single PG protein with the highest scores.

Orthology Analysis

The identified NC COG of the PG XP_003321789.2 was considered during the orthology analysis (first row of **Table 1**). This protein has been theoretically determined as a kinase, possibly partaking of phosphatidylinositol-3 kinases and FRAP/TOR kinases, whose activity is crucial for cell proliferation and signaling (Choi et al., 2002). 23 other species from the fungi kingdom were considered, for each of which an orthologous protein has been identified, with high degree of similarity. The constructed bootstrap tree, as computed from ClustalW2 algorithm, after multiple sequence alignment, when compared with the species phylogenetic tree, led to the identification of the number of duplication events that led to the existence of the NC COG (**Figures 1** and **2**). Duplication and speciation events were recognized and manually annotated on the bootstrap tree. Specifically, two duplication events have been identified, one out of which appears to be unique for NC, and has led to the existence of XP_958289.2 and XP_960307.2 proteins. One of the duplication products was lost during evolution, while the ancestor of X_960570.2 was affected by a variety of speciation events, that led to the existence of this protein in a variety of organisms, including PG. The large branch length of XP_958289.2 and XP_960307.2 infers that they accumulated a high amount of mutations, under the regimen of asymmetric divergence, with possible functional differentiation.

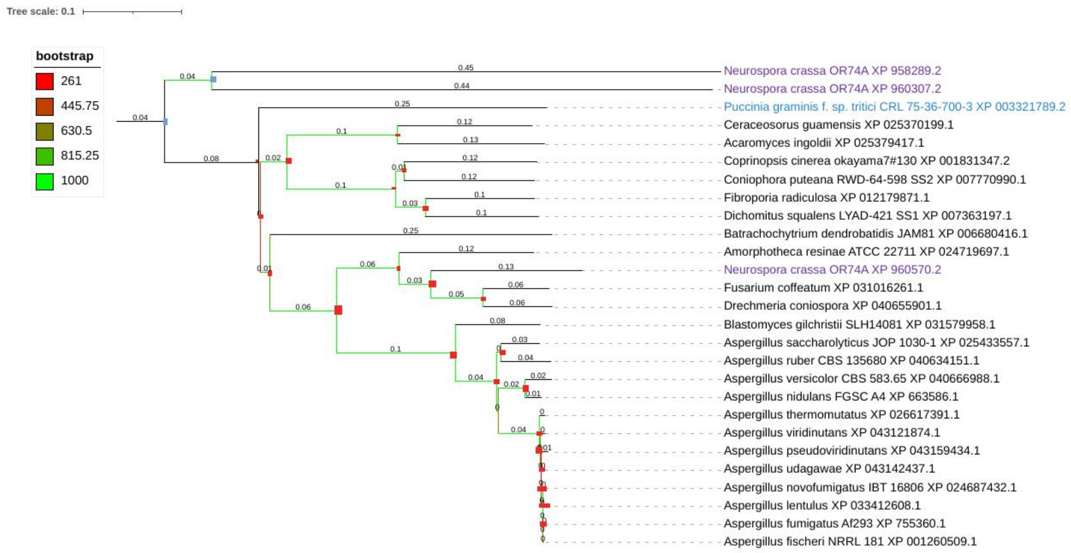
Tree scale: 0.1
bootstrap
🟥 261
🟫 445.75
🟫 630.5
🟩 815.25
🟩 1000
Neurospora crassa OR74A XP 958289.2
Neurospora crassa OR74A XP 960307.2
Puccinia graminis f. sp. tritici CRL 75-36-700-3 XP 003321789.2
Ceraceosorus guamensis XP 025370199.1
Acaromyces ingoldii XP 025379417.1
Coprinopsis cinerea okayama7#130 XP 001831347.2
Coniophora puteana RWD-64-598 SS2 XP 007770990.1
Fibroporia radiculosa XP 012179871.1
Dichomitus squalens LYAD-421 SS1 XP 007363197.1
Batrachochytrium dendrobatidis JAM81 XP 006680416.1
Amorphotheca resinae ATCC 22711 XP 024719697.1
Neurospora crassa OR74A XP 960570.2
Fusarium coffeatum XP 031016261.1
Drechmeria coniospora XP 040655901.1
Blastomyces gilchristii SLH14081 XP 031579958.1
Aspergillus saccharolyticus JOP 1030-1 XP 025433557.1
Aspergillus ruber CBS 135680 XP 040634151.1
Aspergillus versicolor CBS 583.65 XP 040666988.1
Aspergillus nidulans FGSC A4 XP 663586.1
Aspergillus thermomutatus XP 026617391.1
Aspergillus viridinutans XP 043121874.1
Aspergillus pseudoviridinutans XP 043159434.1
Aspergillus udagawae XP 043142437.1
Aspergillus novofumigatus IBT 16806 XP 024687432.1
Aspergillus lentulus XP 033412608.1
Aspergillus fumigatus Af293 XP 755360.1
Aspergillus fischeri NRRL 181 XP 001260509.1

Figure 1 Bootstrap tree of analyzed proteins. The rectangles on each node show whether a speciation (red) or a duplication (blue) event has most likely occurred

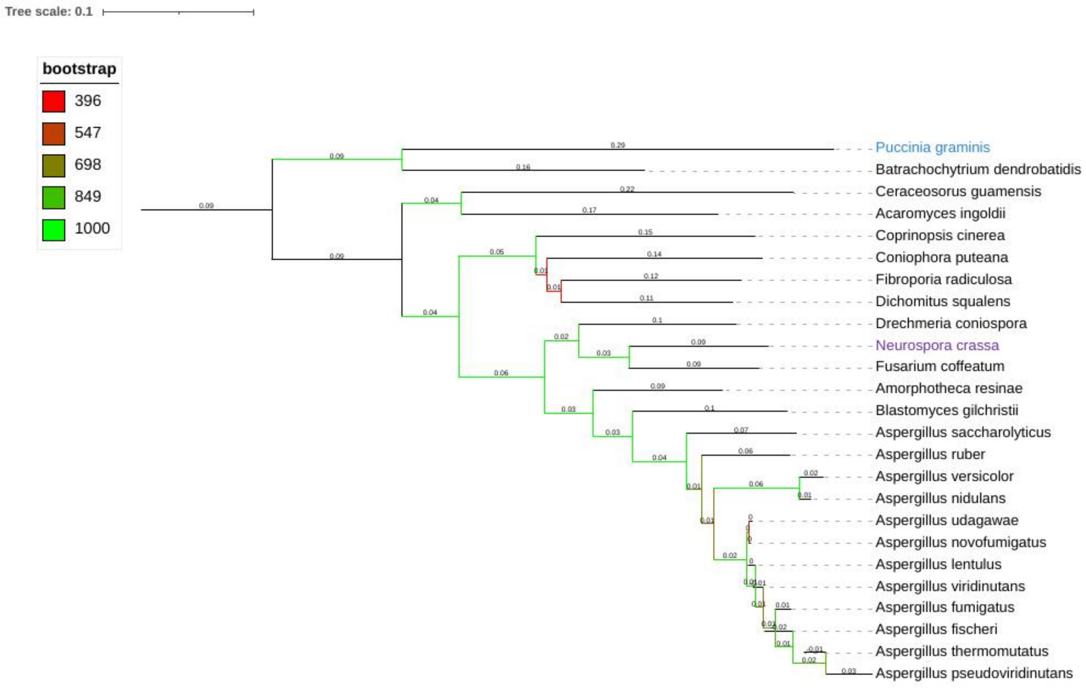

Figure 2 Species tree, related to PG and NC, as produced by interspaced transcribed spacers comparisons.

Discovery of Conserved Regions

Simpson diversity was computed per a/a and is being shown in **figure 3**. The 10 identified conserved sequences are displayed in **table 2** . Batch CD-Search was able to identify 5 out of these to belong to a superfamily cluster with E-value less than 0.05. The superfamily consensus was PKC_like, mainly composed of the catalytic domains of serine/threonine-specific and tyrosine-specific protein kinases. The remaining superfamilies were FRB_dom, referring to FKBP12-rapamycin binding domain, and FAT, which contains FRAP, ATM and TRRAP protein binding domains.

From the annotation it is evident that these regions are mainly binding sites to proteins on which the orthologs of XP_003321789.2 act upon, directly related to and essential for the proteins functionality.

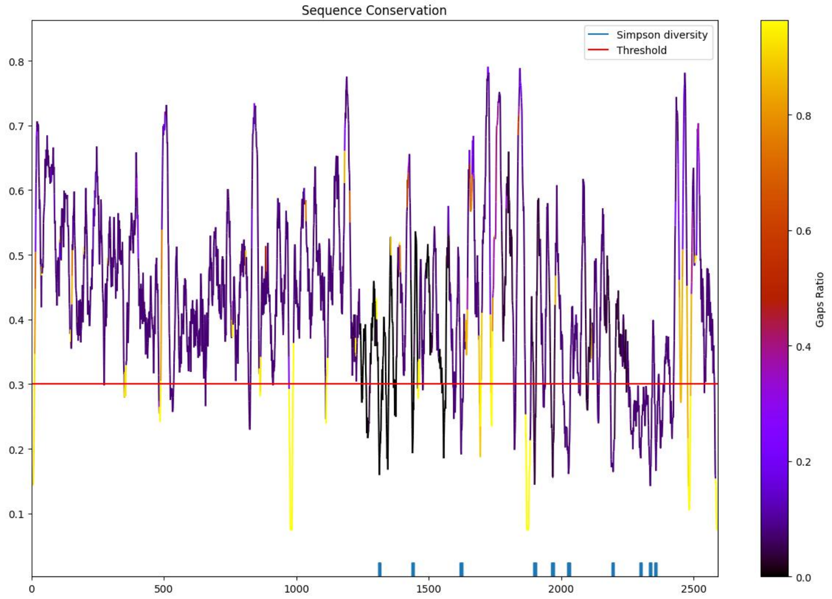

Figure 3 Conservation regions extraction, using a threshold value of 0.3, with the relative locations of the found conserved sub-sequences displayed in the rug plot (blue).

Start position	End position	Sequence	E-Value	Superfamily accession	Superfamily
1309	1318	LLNLAEFMEH			
1436	1445	MRCLHALGEW			
1618	1628	WIKFANLCRKS	0,00806645	cl26693	FAT
1895	1906	LQDTLRLLTLWF			
1963	1974	HPQALVYPLTVA	1,76805	cl34875	TEL1
2023	2035	WHEGLEEASRLYF	0,00452806	cl38290	FRB_dom
2191	2200	RQDERVMQLF	0,00409988	cl21453	PKc_like
2296	2305	DLYRVLWLKS	0,173268	cl21453	PKc_like
2331	2342	ILGLGDRHPSNL	2,99731E-05	cl21453	PKc_like
2353	2361	HIDFGDCFE	0,000137208	cl21453	PKc_like

Table 2 Conserved domains retrieved using Simpson diversity criterion.

Conservation of Promoter Regions

26 motifs were identified in the interval -3kb to 3kb from the TSS as displayed in **table 3** . Tomtom identified 9 of them to be highly similar to transcription factors binding sites, as listed in UniProt. Biological Process enrichment analysis on the found proteins using Panther and Saccharomyces cerevisiae as the closest model organism showed that they are mainly related to the regulation of transcription of RNA polymerase II (**table 4**). The distribution of 3 motifs, one of which matched a Jaspar Matrix entry, with high proximity to TSS is shown in **figure 4**.

Motif	Hit E-value	MEME tool	Jaspar motif hit	UniProt ID	UniProt function
TYGAGGTYTTCGRATACGYYMTSGAYAAS	3,5E-54	MEME	MA0404.1 (MA0404.1.TBS1)	P38114	Involved in tolerance to thiabendazole.
TYCATATCGMTTTCRGTGACTGCWTCGAG	2,3E-53	MEME			
GTTYATHCAYGATCCCHTSATCAAYYRGMS	2,2E-50	MEME			
GAYKAYGACAAYCTBACYCTSATGCARAA	1,3E-49	MEME	MA0295.1 (MA0295.1.FHL1)	P39521	Controls the pre-rRNA processing machinery in conjunction with IFH1. Presumably acts as a transcriptional regulator of genes specifically involved in that process. IFH1 convert FHL1 from a repressor to an activator.
TGGCAYGARCTSTGGCABGAAGGYYTGGA	6,4E-49	MEME	MA0335.1 (MA0335.1.MET4)	P32389	Positive trans-acting factor capable of stimulating the transcription of the MET genes from the methionine biosynthetic pathway. MET4, MET28 and CBF1 are required for full induction of MET25 and MET16 gene transcription. MET4 controls as well the derepression of MET6. Required for the transcription of genes necessary for sulfur amino acid biosynthesis. Involved in the transcription activation of MET28 and MET30. Required for MET3 gene expression via assembly of the MET4-MET28-MET31 and MET4-MET28-MET32 complexes. Involved in response to cadmium and arsenic. Cadmium-activated MET4 also induces glutathione biosynthesis.
GGAGAACHTMTGTCAGCAYTGSAWTGGAT	2,2E-48	MEME			
CAAGGATKCCSTYATGGCYGTYCTGGAAGC	1,3E-46	MEME	MA0305.1 (MA0305.1.GCR2)	Q01722	Transcriptional activator required for the expression of glycolytic genes. Enhances the CT box-dependent transcriptional activation of a RAP1-GCR1 complex. Required for GCR1 phosphorylation.
TTRACTCSCATGYTGACHTTCGCCATGGA	3,3E-46	MEME	MA0368.1 (MA0368.1.RIM101)	P33400	Transcription factor that mediates regulation of both acid- and alkaline-expressed genes in response to ambient pH. At alkaline ambient pH, activates transcription of alkaline-expressed genes (including RIM101 itself), mainly by repressing transcriptional repressors of those genes, and represses transcription of acid-expressed genes. Required for meiosis, sporulation and invasive growth.
TYGAACRKCGBMYHATGCTGCAGATGGCG	1E-45	MEME	MA0355.1 (MA0355.1.PHD1)	P36093	Putative transcription factor that functions in pseudohyphal growth.
GCTAYWTTCTGGGYCTGGGTGACCGWCAYC	3,4E-45	MEME			
AYATGTGGATCAAATTTGCCAACCTCTGC	1E-44	MEME			
GCCTGGCAYDCSTGGGCKYTGGCCAACTT	8,3E-44	MEME			
CATYATGRAYAKTRTGCGDCAGCACAGYG	4E-43	MEME			
GAYCTCAACMAAGCTTGGGATYTHTACTA	4,6E-42	MEME			
GGHGAYCAYDAYRTRGAGGGCATGTTTGC	5,1E-41	MEME	MA0406.1 (MA0406.1.TEC1)	P18412	TEC1 is involved in the activation of TY1 and TY1-mediated gene expression. It is not involved in mating or sporulation processes.
RGATSTGTWYCRHGWTCTGTGGCWCAARA	9,3E-41	MEME	MA0332.1 (MA0332.1.MET28)	P40573	Acts as an accessory factor in the activation of sulfur amino acids metabolism genes. Possesses no intrinsic transcription activation abilities. Binds to the MET16 promoter as a complex with MET4 and CBF1. Enhances the DNA-binding activity of CBF1.
YGATCRDGTSGACAARCTKCTCGCVCARGC	3,5E-40	MEME			
TKCAGGTGSTTGCTCGGGTCAWGGAGAAGC	4,9E-40	MEME			
CSTACCAGAGTGGSARACCRATCATBMGGA	1,5E-37	MEME			
GGTCTKMTGGGMTGGGTCBVCAACAGYGA	4,2E-34	MEME			
AGTTCATTCACGA	2,1E-12	STREME			
ACTACGACTA	5,5E-10	STREME			
CGAAGCAGTCACCGAA	0,000000002	STREME			
CTGCATGACWCKCTCRTC	2,8E-09	STREME			
ATCAACAACCAGCTC	0,00000001	STREME	MA0317.1 (MA0317.1.HCM1)	P25364	Transcription factor regulating the cell cycle specific transcription of a spindle pole body (SPB) calmodulin binding protein SPC110. Required for full induction of SPC110 transcription in late G1. Binds to DNA consensus sequence 5'-[AT]AA[TC]AAACAA[AT]-3'. Dosage dependent suppressor of calmodulin mutants which have specific defects in SPB assembly.
WHWTCGTRTAWAGTGAYAND	0,0015	CENTRIMO			

Table 3 Discovered promoter regions motifs (-3kb to 3kb from TSS) from XP_003321789.2 CDS homologs , with those identified by TomTom annotated with the function reported by UniProt.

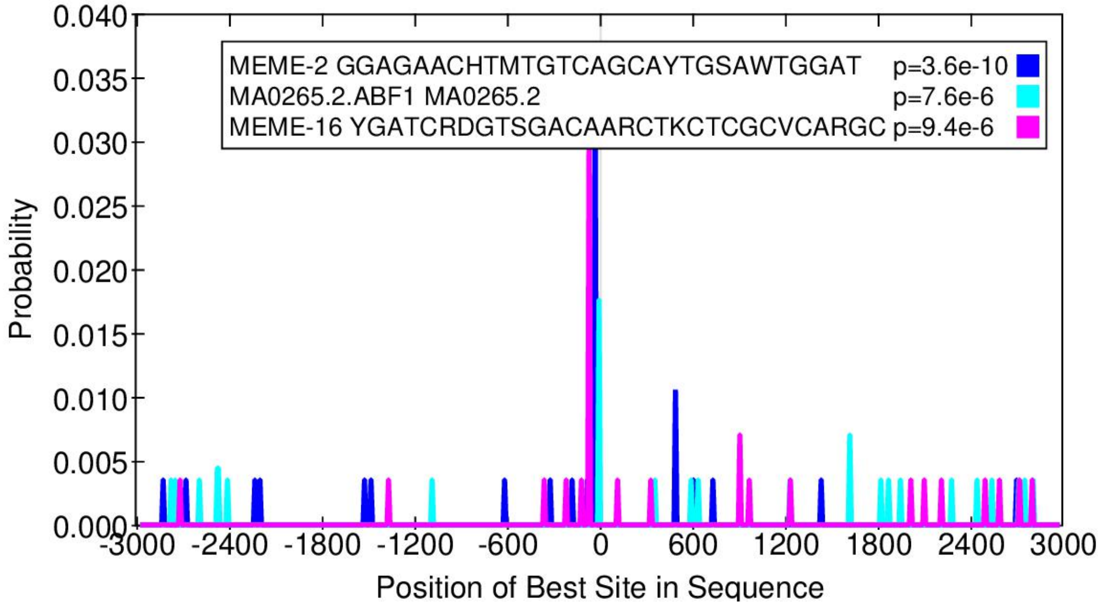

Figure 4 Motifs identified by Centrimo to be centrally enriched in the XP_003321789.2 homologs sequences.

Biological Process	FDR
Positive regulation of transcription by RNA polymerase II	1.52E-05
Positive regulation of pseudohyphal growth	1.52E-02
Cysteine biosynthetic process	1.48E-02

Table 4 Biological process enriched GO terms as identified by Panther, using the recognized proteins, matched to the motifs, identified using meme-chip on the XP_003321789.2 homologs sequences.

Discussion

In this work, two fungi species, which exhibit significant evolutionary differences, as it can be observed in the species tree (**figure 2**), had their proteomes compared and analyzed, with a focus on the recognition of similarities. The noticeable amount of imbalanced COGs discovered implies that there was a differential set of duplication events in a broad genomic region, which could be further examined by a synteny analysis, where the 18 PG chromosome pairs could be compared to the 7 NC chromosome pairs in a pairwise manner, which was not possible using GoGe online tool, as only the contigs and supercontigs were available. The low number of orthologs identified, in comparison to the total amount of protein-coding genes, further highlights the differences between the two species. The isolated COG, to which orthologs from other related organisms have been retrieved, displayed a number of conserved regions, related to the function of it as a member of kinases superfamily. The structure of the phylogenetic species tree, up to the species level, agrees with the structure of the constructed tree, implying that it was central in the species differentiation. This observation gives fertile ground to assess the essentiality of this protein through wet-lab ablation experiments. By studying the conserved promoter regions and identifying retained motifs, such studies can be further improved, by inhibiting the protein function in-vivo, with designed molecules that could bind to them and arrest the protein production during cell cycle. Such a fact could be considered useful, to assess the FRB_dom related protein binding region, as it is related to G1 progression (Villela-Bach et. al., 1999). The related biological processes to the promoters show a strong connection with positive regulation, implying that negative regulation is an undesired effect for this protein, providing further evidence of its essentiality.

Code Access

The code used, as well as the additional files required for this work can be accessed in the GitHub repository: https://github.com/VasLem/comparative_genomics_21 .

References

Ray, Paul D., and Rebecca C. Fry. "The Cell: The Fundamental Unit in Systems Biology." *Systems Biology in Toxicology and Environmental Health*, Academic Press, Jan. 2015, pp. 11–42, doi:10.1016/B978-0-12-801564-3.00002-X.

Choi, Jae H., et al. "The FKBP12-Rapamycin-Associated Protein (FRAP) Is a CLIP-170 Kinase." *EMBO Reports*, vol. 3, no. 10, John Wiley & Sons, Ltd, Oct. 2002, pp. 988–94, doi:https://doi.org/10.1093/embo-reports/kvf197.

Castro-Mondragon, Jaime A., et al. "JASPAR 2022: The 9th Release of the Open-Access Database of Transcription Factor Binding Profiles." *Nucleic Acids Research*, Nov. 2021, p. gkab1113, doi:10.1093/nar/gkab1113.

Lu, Shennan, et al. "CDD/SPARCLE: The Conserved Domain Database in 2020." *Nucleic Acids Research*, vol. 48, no. D1, Jan. 2020, pp. D265–68, doi:10.1093/nar/gkz991.

Machanick, Philip, and Timothy L. Bailey. "MEME-ChIP: Motif Analysis of Large DNA Datasets." *Bioinformatics*, vol. 27, no. 12, June 2011, pp. 1696–97, doi:10.1093/bioinformatics/btr189.

Larkin, M. A., et al. "Clustal W and Clustal X Version 2.0." *Bioinformatics*, vol. 23, no. 21, Nov. 2007, pp. 2947–48, doi:10.1093/bioinformatics/btm404.

Mi, Huaiyu, et al. "PANTHER Version 16: A Revised Family Classification, Tree-Based Classification Tool, Enhancer Regions and Extensive API." *Nucleic Acids Research*, vol. 49, no. D1, Jan. 2021, pp. D394–403, doi:10.1093/nar/gkaa1106.

Vilella-Bach, M., et al. "The FKBP12-Rapamycin-Binding Domain Is Required for FKBP12-Rapamycin-Associated Protein Kinase Activity and G1 Progression." *The Journal of Biological Chemistry*, vol. 274, no. 7, Feb. 1999, pp. 4266–72, doi:10.1074/jbc.274.7.4266.