Stony Brook University

Analysis of the SCARAB-2 Gene

Shane Varghese

Biology 312: Bioinformatics

Dr. Joshua Rest, Ph.D

Alejandro Gil-Gomez

30 Nov. 2022

Introduction

The SCARB-2 gene is part of a diverse supergroup called Scavenger Receptors(SR), these are immune response genes that participate in the phagocytosis of foreign particles. The scavenger proteins were first described by Brown and Goldstein in the 1970s when they identified receptors in macrophages that wouldn't degrade native Low-Density Lipoprotein (LDL). (Mercy et al. 2017) The term SR super group was coined as opposed to typical superfamily due to the unique characteristics of the super group. Each class within the SR super group bears resemblance to the primary sequence but differs vastly from each other. The common features that groups SRs have are their ability to recognize common ligands and to bind to a large variety of microorganisms such as bacteria, viruses, native cells, and many more. Although all SRs have these broad-spectrum binding capabilities, they are all expressed differently within the immune cells (Yap et al. 2015) More recently, they have been classified into 10 families, named SR-A through SR-J. The receptors are typically classified by their sequence domains, although they are also sub-classified based on the different variations within their sequences. (Zani IA et al. 2015) Based on these unique broad-spectrum properties SRs present, scavenger receptors could serve as pattern recognition receptors for innate immunity. With their widespread importance in animal phagocytosis, having an understanding of the evolutionary history of these scavenger proteins can prove to be insightful in learning about the adaptive immune systems within different species. The primary goal of our analysis is to better understand to construct a taxonomically comprehensive evolutionary history of the SCARB-2 and the other SRs within the supergroup. Further looking into if the number of copies in the diverse supergroup of Scavenger Proteins evolve independently in protostomes and deuterostomes, or were these copies already present in the bilaterian common ancestor? We hypothesize that Scavenger Proteins Proteins evolved independently in protostomes and deuterostomes and that more or more would be present within the protostomes class. We believe that all the classes would have conservancy to the early bilaterian ancestors.

Term Paper Methods

Data Collection

We used the National Library of Medicine(NCBI) using Protein BLAST 2.10.1+ (Altschul et al. 1990) to aggregate preliminary data using the specific accession reference sequence tool (accession number: XP_021373098.1) and an e-value cutoff of 1e-35. We decided to use this e-value to have a high sequence identity, and a high alignment score as well as minimize false homologs. Further research was collected, by discovering the gene family; Scavenger Receptor SuperGroup.

Sequence Alignment

Using the filtered BLAST sequences with an e-value less than 1e-35. Now we will use MUSCLE v3.8.31 (Edgar, R.C. 2004), to align all of the sequences with each other along their entire length. Muscle(Edgar, R.C. 2004) is used to create multiple alignments of protein sequences and outputs the highest scores based on alignment accuracy benchmarks. The resulting alignment would be a hypothesis regarding which positions are homologous to each other. alv 1.15.1 was used as a visual aid for our alignment. (Arvestad 2018) This allowed us to see regions where alignments were possible as well as impossible. Using alv we were also able to color coordinate the bases differently. After we obtained information regarding our alignment. After this, we calculated the average percent identity among all sequences in the alignment using the T-Coffee-11.0.8 (Notredame et al. 2000)

Gene Family Phylogeny

We used IQ-Tree multi-core version 2.1.1 (Nguyen et al. 2015) to create an optimal phylogenetic tree based on our sequence alignments. Using IQ-tree will assist with finding the maximum likelihood tree estimates. IQ-tree will test the fit with different substitution models and will create the optimal tree using that best-fit model. We used midpoint rooting, to specify the divergence event. Midpoint will find the longest branch on the tree and find the halfway root. We used gotree v.3.1(Lemoine 2017) to reroot this tree.

Tree Reconciliation and Rearrangement

We used Notung-3.0-beta (Chen et al. 2000) Notung is used to reconcile the gene and species tree. Using this we can estimate duplication and loss events that occurred. This will create a file, with tables with events inferred for different lineages. We used thirdkind v3.0.5(Duchemin et al. 2018) to generate a RecPhyloXML object to view the gene/species tree.

Protein Domain Prediction

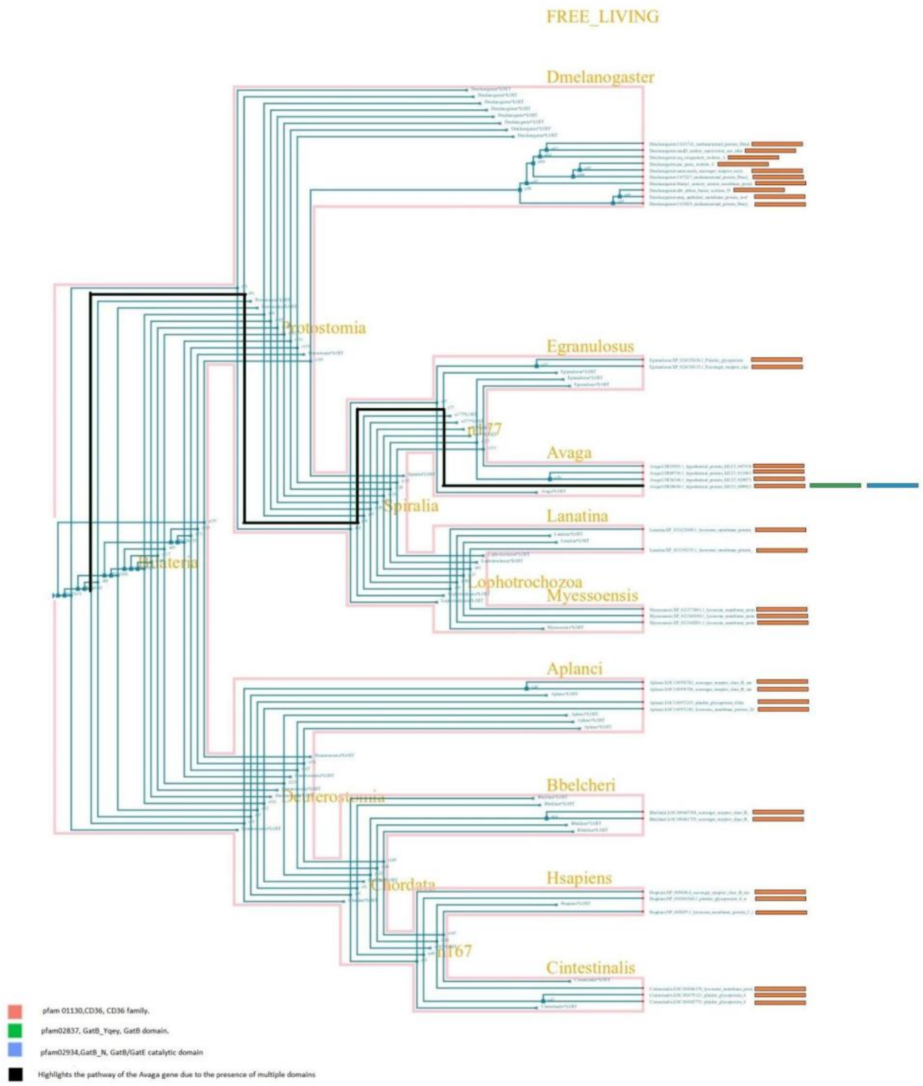
Using RPS Blast 2.10.1+(Altschul, Stephen F., et al 1997), we can identify similar Pfam domains within our protein sequences. We first made a copy of our raw unaligned sequences going back to our data originally found in MUSCLE (Edgar, R.C. 2004) We used an R script to plot supplementary data. On the left side, we used a ggtree package(Yu, Guangchuang, et al 2017) to plot the phylogeny. On the right side, we used a drawProteins package (Brennan, Paul 2018), which organized the pfams domains of the gene. Following this, we downloaded a Pfam(Bateman et al. 2004) database to classify our sequences based on their families and their suggested domains.

Results

The Midpoint-rooted maximum likelihood gene tree with bootstrap supports and contains domain names. In Fig. 1, we can see many of the gene branches are correlated with one domain more specifically CD36. The more interesting point is the presence of two more domains found on the Avaga.UJR28694.1 gene protein. None of our prior datasets presented the presence of a similar domain: GatB/YqeY Domain. To see additional 2 domains specifically on one gene was abnormal, to say the least. SCARB2 and the other scavenger receptors are commonly small genes. In relation to Humans SR genes, we found that 3 different genes were found: Class B messenger, Platelet Glycoprotein 4, and as well Lysomsome_membrane_protein_2(LMP2). In Figure 2, Here we see a Notung reconciliation of the gene and species tree displayed using a thirdkind. Figure 2 better represents various gene speciation and gene duplication events found in the tree. As shown in the Figure 2 diagram, we see a total of 3 domains found in this tree as well. The majority of the gene copies are from the CD36 family, the other 2 found only on the Avaga.UJR28694.1 gene protein. Going back to our original hypothesis we found out that

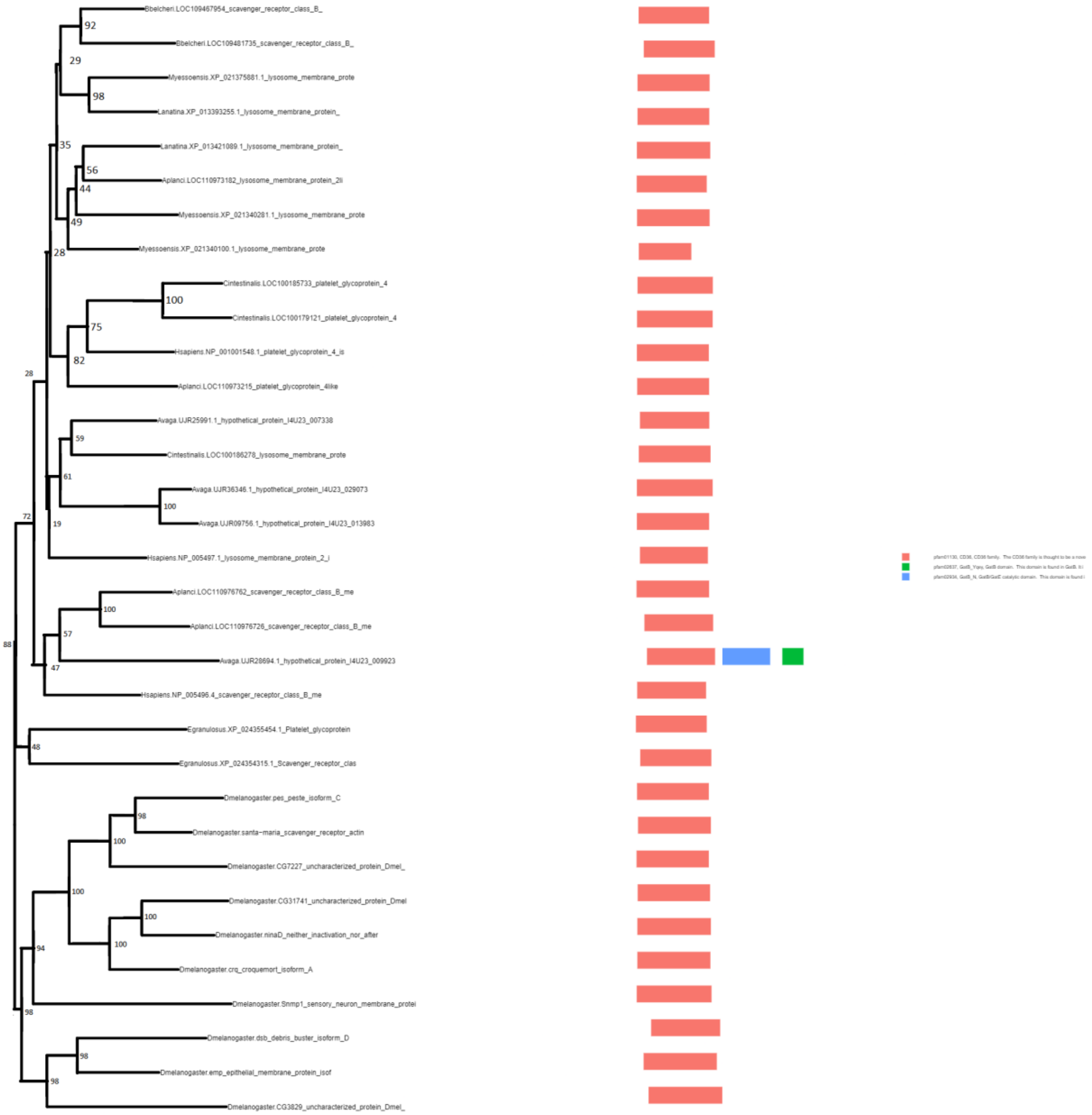
Analyzing the Reconciliation events, we found 25 duplication events within the tree. 9 events occurred in the *D'Melanogaster* and *Protostomia* Bounds. 1 event occurred in the *Aplanci* and *Deuterostomia* Bounds. Looking at the Loss events, we saw a total of 46 loss events. The largest number of losses occurred in the *D'Melanogaster* with a total of 8. The smallest event occurred in the *Bilateria* with a total of 0.

In our original hypothesis, we hypothesized that the scavenger proteins would be evolved independently in protostomes and deuterostomes and that more would be present within the deuterostome class. In the Reconciliation Tree, we determined that there were 9 copies in the protostomes line compared to the 7 copies in the deuterostomes line and that the proteins evolved independently from each other.


FREE_LIVING
Dmelanogaster
Protostomia
Egranulosus
n177
Avaga
Spiralia
Lanatina
Bilateria
Lophotrochozoa
Myessoensis
Aplanci
Deuterostomia
Bbelcheri
Chordata
Hsapiens
n167
Cintestinalis
🟥 pfam 01130,CD36, CD36 family.
🟩 pfam02837, GatB_Yqey, GatB domain.
🟦 pfam02934,GatB_N, GatB/GatE catalytic domain
⬛ Highlights the pathway of the Avaga gene due to the presence of multiple domains

Fig, 1.- Here we have a Midpoint-rooted maximum likelihood gene tree, presenting data from the Bilaterian common ancestry. We can also see Gene loss events throughout the tree, as well as many gene duplications and speciation events. There are 33 Bilaterian gene copies from the CD36 Family. As well as 2 additional domains found on the Avaga.UJR28694.1 gene protein, these are the GatB_Yqey, and GatB/GatE domains. Orange bars were used to present genes within the CD36 family. A Green bar was used to present gene/s within the GatB_Yqey Domain. A Light Blue bar was used to present gene/s within the GatB/GatE catalytic domain. The pathway towards the Avaga gene with the three domains was further bolded black. We can see *DMelanogaster* presented the most gene copies, totaling 10 copies. We can see a three-way tie-in regarding the least amount of gene copies between *Egranulosus, Lanatina, and Bbelcheri*. Each has 2 copies. *Bbelcheri* is the only one descending from Deuterostomes, the other two descended from Protostomes.

Fig, 2.- Here we see a Notung reconciliation of the gene and species tree, displayed using a third kind. Accompanying the tree we see Bootstrap support values. We can also see the different gene events that occur such as duplication, and specifications events. Unfortunately this doesn't present gene loss events. The tree uses the same color convention as Figure 1. Orange bars were


Bbelcheri.LOC109467954_scavenger_receptor_class_B_
92
Bbelcheri.LOC109481735_scavenger_receptor_class_B_
29
Myessoensis.XP_021375881.1_lysosome_membrane_prote
98
Lanatina.XP_013393255.1_lysosome_membrane_protein_
35
Lanatina.XP_013421089.1_lysosome_membrane_protein_
56
Aplanci.LOC110973182_lysosome_membrane_protein_2li
44
Myessoensis.XP_021340281.1_lysosome_membrane_prote
49
Myessoensis.XP_021340100.1_lysosome_membrane_prote
28
Cintestinalis.LOC100185733_platelet_glycoprotein_4
100
Cintestinalis.LOC100179121_platelet_glycoprotein_4
75
Hsapiens.NP_001001548.1_platelet_glycoprotein_4_is
82
Aplanci.LOC110973215_platelet_glycoprotein_4like
28
Avaga.UJR25991.1_hypothetical_protein_I4U23_007338
59
Cintestinalis.LOC100186278_lysosome_membrane_prote
61
Avaga.UJR36346.1_hypothetical_protein_I4U23_029073
100
Avaga.UJR09756.1_hypothetical_protein_I4U23_013983
72
19
Hsapiens.NP_005497.1_lysosome_membrane_protein_2_i
Aplanci.LOC110976762_scavenger_receptor_class_B_me
100
Aplanci.LOC110976726_scavenger_receptor_class_B_me
57
88
Avaga.UJR28694.1_hypothetical_protein_I4U23_009923
47
Hsapiens.NP_005496.4_scavenger_receptor_class_B_me
Egranulosus.XP_024355454.1_Platelet_glycoprotein
48
Egranulosus.XP_024354315.1_Scavenger_receptor_clas
Dmelanogaster.pes_peste_isoform_C
98
Dmelanogaster.santa−maria_scavenger_receptor_actin
100
Dmelanogaster.CG7227_uncharacterized_protein_Dmel_
100
Dmelanogaster.CG31741_uncharacterized_protein_Dmel
100
Dmelanogaster.ninaD_neither_inactivation_nor_after
94
100
Dmelanogaster.crq_croquemort_isoform_A
Dmelanogaster.Snmp1_sensory_neuron_membrane_protei
98
Dmelanogaster.dsb_debris_buster_isoform_D
98
Dmelanogaster.emp_epithelial_membrane_protein_isof
98
Dmelanogaster.CG3829_uncharacterized_protein_Dmel_
🟥 pfam01130, CD36, CD36 family. The CD36 family is thought to be a nove
🟩 pfam02637, GatB_Yqey, GatB domain. This domain is found in GatB. It i
🟦 pfam02934, GatB_N, GatB/GatE catalytic domain. This domain is found i

used to present genes within the CD36 family. A Green bar was used to present gene/s within the GatB_Yqey Domain. A Light Blue bar was used to present gene/s within the GatB/GatE catalytic domain.

Term Paper Discussion

Reconstructing a taxonomically comprehensive evolutionary history of the SCARB-2 and the other scavenger proteins is vital to how these proteins evolved. Scavenger receptor proteins may differ structurally and functionally, but maintain their same properties regarding phagocytosis of foreign particles.

Our objective was to look into the number of copies in the diverse supergroup of Scavenger Proteins that evolve independently in protostomes and deuterostomes, and if the copies were already present in the bilaterian common ancestor. Our first major finding was the presence of the 2 additional domains found on the Avaga.UJR28694.1 gene protein, this also would be a great continuation of this study, as none of the data indicated the origins of these 2 additional domains. We can infer that these additional domains were not present in the last common ancestor of protostomes and deuterostomes and are a recent gene gain event that occurred on this specific ancestor.

Similar studies discussed the application of SR proteins for the treatment of many illnesses and disorders such as inflammatory diseases and cancers. In this study, they researched the SR superfamily presence in the Pathogenesis of Liver Inflammation and Cancer. They discussed the role of scavenger receptor proteins in many of the pathophysiology of inflammatory diseases, such as Alzheimer’s and Chronic Liver Disease (Patten, Daniel A, et al 2022) The authors discussed the further application of scavenger receptor proteins found in humans more specifically liver cells and their involvement in a leading cause of morality. Another study discussed the SR family's versatility regarding Cancer Immunobiology. The Authors believed that due to their large family and how the proteins vastly differ structurally and functionally but maintain similar properties this can aid in advancing Cancer research. They discussed how the receptors are essential regulators regarding tumor behavior and host immune responses to cancers. (Yu, Xiaofei, et al 2015)

Modifications that can be made to adjust the data and explore any more possible domains would be expanding the e-value cutoff to include more species. Issues may occur as expanding the e-value cutoff might introduce false to false homologs. Another possibility for a continuation would be learning more about the GatB_Yqey, and GatB/GatE domains.

Citations

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. J. Molecular Biol., 215(3), pp.403-410.

Altschul, Stephen F., et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic acids research 25.17 (1997): 3389-3402.

Arvestad, L. (2018). alv: a console-based viewer for molecular sequence alignments. *J. Open Source Softw.,* 3(31): 955. https://doi.org/10.21105/joss.00955

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L. and Studholme, D.J., 2004. The Pfam protein families database. Nucleic Acids Res., 32(suppl_1), pp.D138-D141.

Brennan, Paul. "drawProteins: a Bioconductor/R package for reproducible and programmatic generation of protein schematics." F1000Research 7 (2018).

Chen, K., Durand, D. and Farach-Colton, M., 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. J. Computational Bio.l, 7(3-4), pp.429-447.

Duchemin, W., Gence, G., Arigon Chifolleau, A.M., Arvestad, L., Bansal, M.S., Berry, V., Boussau, B., Chevenet, F., Comte, N., Davín, A.A. and Dessimoz, C., 2018. RecPhyloXML: a format for reconciled gene trees. Bioinformatics, 34(21), pp.3646-3652.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792-1797

Lemoine, F. & Wang A. (2017). Gotree. GitHub repository. https://github.com/evolbioinfo/gotree

Mercy R. et al. The Journal of Immunology May 15, 2017, 198 (10) 3775-3789; DOI: 10.4049/jimmunol.1700373

Nguyen, L.-T., H.A. Schmidt, A. von Haeseler, B.Q. Minh (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies.. Mol. Biol. Evol., 32:268-274.

Notredame, C., Higgins, D.G., Heringa, J. (2000). T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. *J. Mol. Biol.*, 302:205-217.

Patten, Daniel A, et al. “Scavenger Receptors: Novel Roles in the Pathogenesis of Liver Inflammation and Cancer.” *Seminars in Liver Disease*, Thieme Medical Publishers, Inc., Feb. 2022, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8893982/.

Yap, Nicholas V. L., et al. “The Evolution of the Scavenger Receptor Cysteine-Rich Domain of the Class A Scavenger Receptors.” *Frontiers*, Frontiers, 1 Jan. 1AD, https://www.frontiersin.org/articles/10.3389/fimmu.2015.00342/full.

Yu, Guangchuang, et al. "ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data." Methods in Ecology and Evolution 8.1 (2017): 28-36.

Yu, Xiaofei, et al. “Scavenger Receptors: Emerging Roles in Cancer Biology and Immunology.” *Advances in Cancer Research*, U.S. National Library of Medicine, 2015, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4631385/.

Zani IA, Stephen SL, Mughal NA, Russell D, Homer-Vanniasinkam S, Wheatcroft SB, Ponnambalam S. Scavenger receptor structure and function in health and disease. Cells. 2015 May 22;4(2):178-201. doi: 10.3390/cells4020178. PMID: 26010753; PMCID: PMC4493455.