GPCR identification in parasitic worm genome assemblies

Nicolas Wheeler (njwheeler@wisc.edu)
Zamanian Lab
Tim Day
Iowa State University
Mostafa Zamanian
University of Wisconsin-Madison

Method Article

Keywords: genome sequencing, assembly, parasites, transmembrane protein, GPCRs, genomes, assemblies, parasitic worms, helminths, nematodes, flatworms, platyhelminths, Nematoda, Platyhelminthes

Posted Date: May 17th, 2018

DOI: https://doi.org/10.1038/protex.2018.061

Abstract

G protein-couple receptors are often a primary point of interest in parasitic worms. Thus, after the sequencing of the genome of one of these worms, GPCR identification and annotation is often a priority. Here, we describe a computational pipeline for the identification of orthologous GPCRs in a set of over 50 newly sequenced genomes from parasitic worms.

Introduction

G protein-coupled receptors \(GPCRs) are a superfamily of plasma membrane receptors that have diverse functions in parasitic worms \(helminths), including neuromuscular signaling, chemosensation, and development. As GPCRs are the most popular class of drug targets in humans, they have been implicated as possible next-generation targets for anthelmintics. Thus, identification and annotation of GPCRs in helminth genomes is often an immediate priority. Here, we present a computational pipeline for GPCR identification and annotation that leverages comparative genomics performed in the recent release of over 50 helminth genomes. The most robust methods for GPCR identification often use structural information alone as a first-pass identification of putative GPCRs. GPCRs have a canonical structure that includes seven transmembrane regions, and as such they are readily identified in genomic open-reading frames using pan-genome transmembrane domain predictions. However, this is computationally intensive for a single genome, let alone >100 genomes. It is also not optimized for highly fragmented genomes that have confounded assembly by high concentrations of repetitive regions and high AT content – features that are often found in helminth genomes. Thus, we devised an alternative strategy that leveraged the lucrative comparative genomic data included at WormBase ParaSite. Using previously identified helminth and free-living worm GPCRs as seeds, we developed a homologous family-centric approach for identifying conserved GPCR families and filtering out false-positives \(Figure 1 See figure in Figures section.).

Reagents

Compara clusters at www.parasite.wormbase.org

Equipment

Software: mafft \(https://mafft.cbrc.jp/alignment/software/) trimal \(https://github.com/scapella/trimal) HHsuite \(http://www.soeding.genzentrum.lmu.de/software-and-servers-2/) NCBI blast command line \(ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/)

Procedure

1. Acquire all previously identified GPCR gene IDs from C. elegans, B. malayi, O. volvulus. S. mansoni, and S. mediterranea. 2. Using the gene ID seeds from \(1), use the WormBase ParaSite API to pull out all

orthologues in the database. Pull the orthologous gene IDs and their sequences and organize into gene families based on the original seed. 3. Use MAFFT \(Katoh & Standley 2013) \(–auto) to individually align the sequences of each family. 4. Use trimal \(Capella-Gutiérrez 2009) \(-automated1) to trim remove columns that contain ambiguous sites or poorly aligned characters. 5. Use HHsuite \(hmmbuild) to create profile Hidden Markov Models \(HMMs) for each trimmed alignment. 6. Acquire curated HMM databases from UniProt, SCOPUS, Pfam, and the PDB. 7. Use HHsuite \(Soding 2005) \(hmmsearch) to search each family HMM against the four databases and retain the best-hit for each search \(i.e. resulting in four best-hits for each family). 8. Manually peruse the best-hits for each family and remove families that contain <2 GPCR-like best-hits – these are the most likely false positives. 9. Take two or three representative sequences from each family and use blastp \(Camacho et al. 2009) to search against the NCBI non-redundant protein database. Cross-reference the annotations of significant hits with GPCRdb \(Pandy-Szekeres et al. 2018) to categorize the GPCR family as Class A \(rhodopsin-like), Class B \(secretin/adhesion-like), Class C \(glutamate-like), or Class F \(frizzled-like). 10. Append manually identified GPCR families and families identified via a separate analysis of synapomorphic families to the final list of putative GPCR families and their members.

Troubleshooting

As with an in silico method, there are caveats to this approach. First, it is unlikely to capture some of the interesting clade- and class-specific families, as well as those that are more highly diverged or don't have representatives among the original set of seeds. Second, the approach relies heavily on the trustworthiness of the initial homology classifications, and so any assignment mistakes upstream will trickle down into GPCR classification as well. Despite these, this approach provides a conservative set of confidently called GPCRs, which can be used for further optimization for gene annotation in helminth genomes.

Anticipated Results

A list of GPCR families, along with their individual members, with categorization into GPCRdb classes and subclassifications of Class A families into aminergic, peptidergic, chemosensory, and other receptors.

References

1. Katoh, Kazutaka, and Daron M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." Molecular Biology and Evolution 30 \(4): 772–80. 2. Wit, Janneke, and John S. Gilleard. 2017. "Resequencing Helminth Genomes for Population and Genetic Studies." Trends in Parasitology 33 \(5): 388–99. 3. Howe, Kevin L., Bruce J. Bolt, Myriam Shafie, Paul Kersey, and Matthew Berriman. 2017. "WormBase ParaSite - a Comprehensive Resource for Helminth Genomics." Molecular and Biochemical Parasitology 215 \(July): 2–10. 4. Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. "trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses." Bioinformatics 25 \(15): 1972–73. 5. Söding,

Johannes. 2005. “Protein Homology Detection by HMM-HMM Comparison.” Bioinformatics 21 \(7): 951–60. 6. Pándy-Szekeres, Gáspár, Christian Munk, Tsonko M. Tsonkov, Stefan Mordalski, Kasper Harpsøe, Alexander S. Hauser, Andrzej J. Bojarski, and David E. Gloriam. 2018. “GPCRdb in 2018: Adding GPCR Structure Models and Ligands.” Nucleic Acids Research 46 \(D1): D440–46. 7. Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. “BLAST+: Architecture and Applications.” BMC Bioinformatics 10 \(December): 421.

Figures

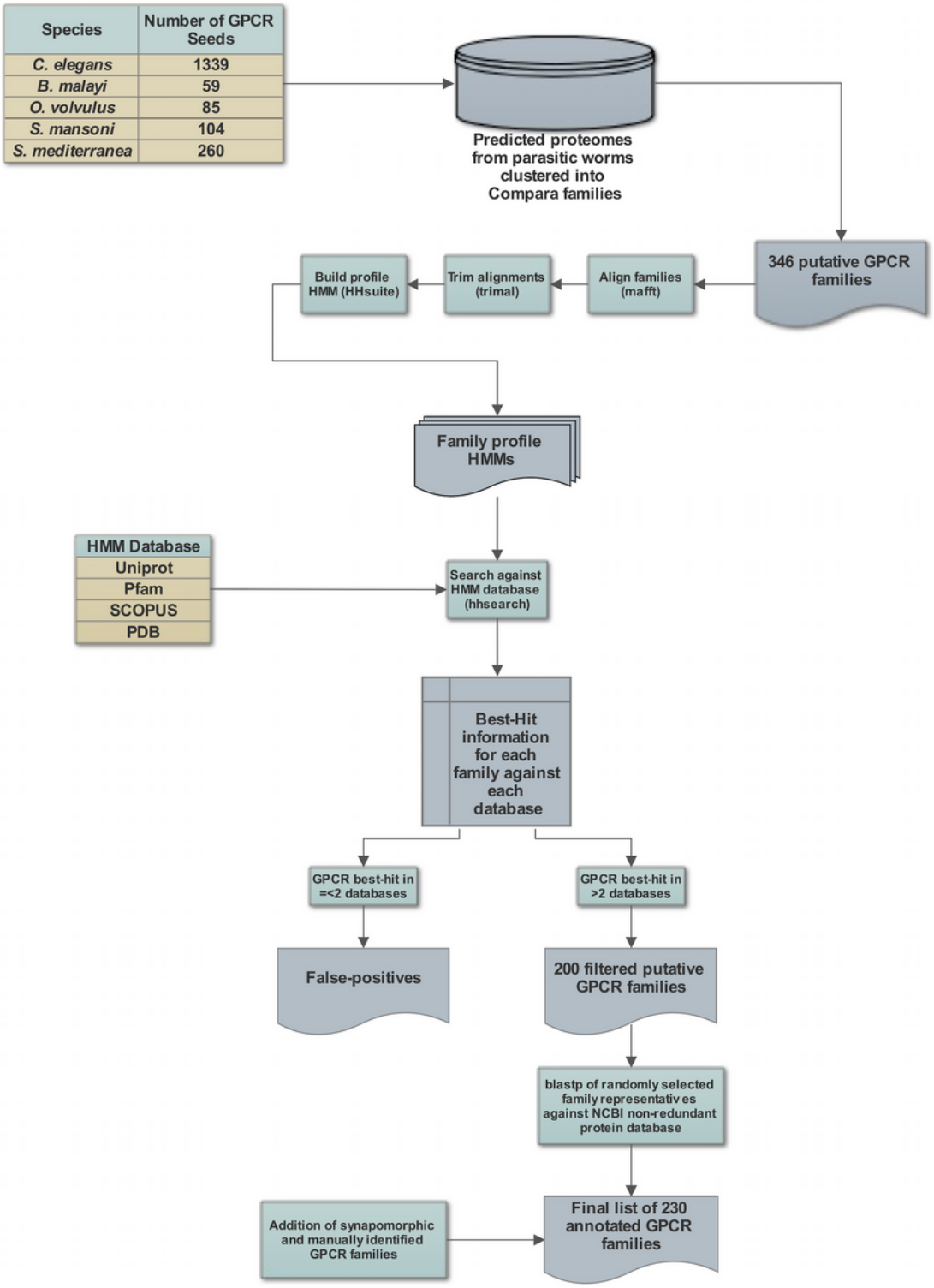

Figure 1

Computational pipeline for identifying GPCRs in the genomes of parasitic worms.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement0.pdf