PLANNED ANALYSES FOR TESTING EXISTENCE OF SPONGE PARAHOX

Principal Investigator: Joseph Ryan

Draft or Version Number: v. 1.4

28 June 2017

LIST OF ABBREVIATIONS

SOWH	Swofford-Olsen-Waddell-Hillis
AU	Approximately unbiased
ML	Maximum likelihood
NJ	Neighbor-joining

1 INTRODUCTION: BACKGROUND INFORMATION AND SCIENTIFIC RATIONALE

1.1 BACKGROUND INFORMATION

It has been proposed that the calcareous sponges *Sycon ciliatum* and *Leucosolenia complicata* have an ortholog of the ParaHox gene Cdx (Fortunato et al., 2014).

1.2 RATIONALE

This document serves as an *a priori* established protocol laying out our planned phylogenetic experiments to test whether the *Sycon ciliatum* and *Leucosolenia complicata* homeobox genes reported to be Cdx orthologs are truly *bona fide* orthologs of Cdx (or any other Hox/ParaHox gene).

1.3 OBJECTIVES

We will test the sensitivity of results acquired in Fortunato et al. (2014) to methods, models, and taxon sampling. In addition, we will apply the Swofford-Olsen-Waddell-Hillis (SOWH) and the approximately unbiased (AU) tests to evaluate the robustness of all topologies to relative alternative hypotheses.

2 STUDY DESIGN AND ENDPOINTS

1- Repeat the phylogenetic analyses performed in the Fortunato et al. (2014) using the same software, and models on the two data matrices presented in the paper.

   a) Prottest3.0 to determine the best suitable model of sequence evolution (Fortunato et al. (2014) found LG+G to be the best model).
   b) Neighbour-joining (NJ) analysis in Phylip v.3.696, maximum-likelihood (ML) analysis in PhyML v.3.0, and Bayesian analysis in MrBayes v.3.1.2.

2- Apply these additional phylogenetic methods on the Fortunato et al. (2014) data matrix

   a) To test for sensitivity to the model we will conduct NJ analyses using the protdist and neighbor programs in Phylip v.3.696 with all models available in protdist (i.e., JTT, PMB, PAM and Kimura); all other settings being default.
   b) To test for sensitivity to method and model we will conduct ML analyses using RAxML v. 8.2.10 with the following models: PROTGAMMALG, PROTGAMMAJTT, PROTGAMMAWAG, and PROTGAMMAAUTO with 100 fast bootstraps. We will use 5 starting parsimony trees

and 5 random starting trees for each run. We anticipate PROTGAMMAAUTO will choose LG based on previous homeodomain analyses, in which case the PROTGAMMAAUTO would be redundant and not considered.

c) To test for sensitivity to method and model we will run the following Bayesian analyses using MrBayes v3.2.6 with the following execution blocks (to test if there is sensitivity to the model and the method):

i. `prset aamodelpr = fixed(LG); lset rates = gamma;`
ii. `prset aamodelpr = fixed(WAG); lset rates = gamma;`
iii. `prset aamodelpr = fixed(JTT); lset rates = gamma;`
iv. `prset aamodelpr = mixed; lset rates = gamma;`

Each of the above blocks will include the following additional commands:

```
mcmcp ngen=10000000 samplefreq=10000 mcmcdiagn=yes stoprule=yes
stopval=.01 nruns=2 nchains=5 savebrlens=yes;
mcmc;
sumt filename=FILE.nex nRuns=2 Relbrunin=YES BurninFrac=.25
Contype=Allcompat;
```

v. convergence of all paired runs will be considered achieved when the average standard deviation of split frequencies equals 0.01 (stopval command)

3- To test for the sensitivity to taxon sampling, we will create a custom dataset that includes the 60-amino acid sequences from HomeoDB of the subclasses HOXL and NKL from human, beetle, amphioxus, and fruitfly. In addition, to these we will add the HOXL and NKL datasets from *Capitella teleta, Crassostrea gigas*, and *Nematostella vectensis* from Zwarycz et al. (2016). Lastly, we will include the putative Cdx genes from *Sycon ciliatum* and *Leucosolenia complicata*.

After running an initial FastTree analysis (`FastTree -fastest -lg ANTP_matrix_v2.phy >out.tre`) to help in our initial construction of constraints for hypothesis testing, we noted two very long branches one of which formed a clade within the NK clade with the putative sponge Cdx genes. After examining the sequences corresponding to these long branches we noticed that 43 out of 60 residues from one of these sequences (Cg_DlxB) were gaps. The other Ct_ HOXLHD20, which was the one that grouped with the putative sponge Cdx genes, also was more than 50% gaps as well. With such little sequence information it is likely that we will not be able to accurately place these sequences, and they could potentially influence the position of the genes we are testing. In order to avoid this, we removed all sequences from our custom dataset that included 10 or more gaps.

4- We will run the following analyses on this custom dataset:

a) NJ analyses using the protdist and neighbor programs in Phylip v.3.696 10 (same models & parameters specified in 2a)
b) ML analyses using RAxML v. 8.2.10 (same models & parameters specified in 2b)
c) Bayesian analyses using MrBayes v3.2.6 (same models & parameters specified in 2c)

5- Hypothesis testing. We will apply the AU test as implemented in CONSEL v 1.20 and the SOWH

test as implemented in sowhat v0.36 on both the original alignment from Fortunato et al (2014) as well as our custom datasets. We will use a single model of evolution (the model chosen automatically by PROTGAMMAAUTO for each dataset—likely PROTGAMMALG) for the SOWH test. We will test the following hypotheses:

a. Fortunato 150-dataset: ((LcoCdx,SciCdx,BflCdx,TcaCad1,TcaCad2),all,other,sequences)
b. Fortunato 150-dataset: ((all,Hox,and,ParaHox),LcoCdx,SciCdx,all,other,sequences)
c. Fortunato 259-dataset: ((LcoCdx,SciCdx,all,Cdx),all,other,sequences)
d. Fortunato 259-dataset: ((all,Hox,and,ParaHox), LcoCdx,SciCdx,all,other,sequences)
e. Custom dataset: ((all,Cdx,plus,LcoCdx,SciCdx),all,other,sequences)
f. Custom dataset: ((all,Hox,and,ParaHox),LcoCdx,SciCdx,all,other,sequences)

3 WORK COMPLETED SO FAR WITH DATES

Date: 5 June 2017 – created new dataset (proposed in 3)

Date: 5 June 2017 – performed RAxML with PROTGAMMAAUTO on new dataset (proposed in 4b)

Date: 8 June 2017 – performed RAxML with PROTGAMMAAUTO on Fortunato set (proposed in 2b)

4 LITERATURE REFERENCES

Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 2005 May 1;21(9):2104-5.

Church SH, Ryan JF, Dunn CW. Automation and Evaluation of the SOWH Test with SOWHAT. Systematic biology. 2015 Nov 1;64(6):1048-58.

Felsenstein J. PHYLIP: phylogenetic inference package, version 3.5 c.

Fortunato SA, Adamski M, Ramos OM, Leininger S, Liu J, Ferrier DE, Adamska M. Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. Nature. 2014 Oct 30;514(7524):620-3.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology. 2010 Mar 29;59(3):307-21.

Rambaut A, Drummond A. Tracer: a program for analysing results from Bayesian MCMC programs such as BEAST & MrBayes. University of Edinburgh, UK. 2003.

Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003 Aug 12;19(12):1572-4.

Shimodaira H, Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 2001 Dec 1;17(12):1246-7.

Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006 Nov 1;22(21):2688-90.

Zwarycz AS, Nossa CW, Putnam NH, Ryan JF. Timing and scope of genomic expansion within Annelida: evidence from homeoboxes in the genome of the earthworm Eisenia fetida. Genome biology and evolution. 2016 Jan 1;8(1):271-81.

Zhou X, Shen XX, Hittinger CT, Rokas A. Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. bioRxiv. 2017 Jan 1:142323.

APPENDIX

Version	Date	Significant Revisions
1.1	15 June 2017	Explicitly mention analyses of both matrices from Fortunato et al. (2014). Removed plan to run GAMMA on NJ analyses—is not standard practice and was not performed in Fortunato et al. (2014). Changed version of phyml—was not able to find version used in Fortunato et al. (2014). Added hypothesis testing for the extended dataset from Fortunato et al. (2014).
1.2	20 June 2017	Removed homeodomains with 10 or more gaps from our custom dataset. After reading results of Zhou et al. (2017), we have chosen to run 5 starting parsimony trees and 5 random starting trees for all RAxML trees.
1.3	28 June 2017	Edited commands for MrBayes so that instead of running a set number of generations, replicate mcmc runs will continue until the average standard deviation of split frequencies equals 0.01
1.4	29 June 2017	After realizing MrBayes would stop at the default 1,000,000 generations even if stopval criteria had not been reached, we edited commands for MrBayes to include 10 million generations, sampling every 10,000 while still maintaining the stopval = 0.01 command.