Planned analyses for investigating Innexin clusters

Jennifer Ortiz, Melissa DeBiasse, and Joseph Ryan

v.1.3

12 August 2020

LIST OF ABBREVIATIONS

ML	maximum likelihood
TPM	Transcripts per million
MC	Monte carlo

1 INTRODUCTION: BACKGROUND INFORMATION AND SCIENTIFIC RATIONALE

1.1 BACKGROUND INFORMATION

Invertebrate gap junctions are formed by innexins. These genes have not been characterized in detail in the ctenophore *Mnemiopsis leidyi*. Preliminary BLAST analyses suggest that *M. leidyi* has a cluster of four unrelated innexins that show correlated expression.

1.2 RATIONALE

A detailed comparative phylogenetic analysis of innexins will allow us to better understand the function of these genes in ctenophores and role of clustering in *M. leidyi* innexins.

1.3 HYPOTHESES

The innexin cluster in *M. leidyi* is ancient and was present in the last common ancestor of *Beroe* and *Mnemiopsis*.

1.4 OBJECTIVES

Phylogenetically classify innexin genes in *M. leidyi*, *Beroe ovata*, and *Pleurobrachia bachei*.

2 STUDY DESIGN AND ENDPOINTS

2.1 Identify putative innexins within genomes

We will use the protein sequence of innexin 2 (isoform D) from *Drosophila melanogaster* (Abascal and Zardoya 2012) as a query sequence to identify innexins in the genome of the following: *H. vulgaris*, *P. bachei*, *M. leidyi*, and *B. ovata* genomes. Protein models with BLAST E-values below 1e-3 will be retained.

2.1.1 Building a database

We will search putative innexin genes against Pfam to identify the protein domain. If partial innexins occur, we will search for the complete sequence in unfiltered gene models and gene transcripts. Innexin predictions will be extended to the full length of the domain if found to be incomplete by 5 or fewer amino acids at the N or C terminus of the domain.

2.1.2 Filter duplicate copies of *B. ovata* innexin gene models

Our current genome assembly of the diploid *B. ovata* includes scaffolds from both haplotypes, with some regions of the genome represented by one collapsed sequence and the others represented by both haplotypes. To be sure that we only include one representation of each *B. ovata* innexin, we will build an ML tree of the innexins found in *B. ovata* and use this tree to identify and remove any duplicate innexins present on phased haplotigs scaffolds.

```
iqtree-omp -s [Bova_innexins.mafft-gb] -nt AUTO -bb 1000 -m TEST -pre [output
prefix] > iq.out 2> iq.err
```

2.2 Alignment of putative innexins

We will align putative innexin sequences with MAFFT using default parameters.

```
mafft [fasta_file] > [fasta_file].mafft
```

2.3 Innexin gene trees

We will infer the phylogenetic relationships among innexins by estimating gene trees.

2.3.1 IQTREE

```
iqtree-omp -s [infile.mafft-gb] -nt AUTO -bb 1000 -m TEST -pre [output
prefix] > iq.out 2> iq.err
```

2.4 RAxML with 25 starting parsimony trees and 25 random starting trees;

```
raxmlHPC-SSE3.PTHREADS -f a -T 25 -p [random_number] -# 25 -m PROTGAMMA[best-
fit_model] -s [alignment_file] -n [name]_mp
```

```
raxmlHPC-SSE3.PTHREDAS -f a -T 25 -d -p [random_number] -# 25 -m
PROTGAMMA[best-fit_model] -s [alignment_file] -n [name]_rt
```

2.5 Mr. Bayes

We will run the following Bayesian analyses using MrBayes v3.2.6 with the following execution block (best model will be determined from IQ-TREE run):

```
prset aamodelpr = fixed(BEST_MODEL); lset rates = gamma;
```

2.6 Choose best tree to be the main figure by comparing likelihood scores of each generated tree as calculated by RAxML (We will use RAxML -? to calculate likelihood score of trees from IQ-Tree and Mr. Bayes). All other trees will be reported in supplement.

```
raxmlHPC -p 12345 -m PROTGAMMA[best-fit_model] -s [aln file] -n [name]
```

2.6 Identifying potential pairing of Innexin gene expression using developmental timecourse expression data (Levin et al. 2016)

2.6.1 We will generate a similarity metric by subtracting the log of medians between each time point and then summing these differences. Because timepoints that share 0 expression will be exactly the same while high expression values that are coregulated will be affected by noise and therefore highly unlikely to be the same, we will set any differences less than 0.1 to 0.1, which is the smallest log(median) difference between all of the timepoints between Ml_ctINXB and Ml_ctINXD (both of these are highly expressed and there is multiple evidence suggesting that they are coregulated).

3 WORK COMPLETED SO FAR WITH DATES

27 June 2019: We have identified the putative innexin cluster using BLAST.

30 June 2019: *B.ovata* innexin genes identified. Used Pfam to identify complete protein domains for the four species.

16 July 2019: *B.ovata* innexin genes trimmed down, removing duplicates. All trees were rerun, and best tree was determined.

22 July 2019: MEME and TOMTOM runs completed and possible regulation factors identified.

4 LITERATURE REFERENCES

Abascal F, Zardoya R. Evolutionary analyses of gap junction protein families. *BBA.* 2012

Kuznetsov *et al*. Gap junctions in nematodes. *Russian Journal of Nematology*. 2016

Ryan *et al*. The Genome of the Ctenophore Mnemiopsis leidyi and Its Implications for Cell Type Evolution. *Science*. 2013

Sebé-Pedros *et al.* Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nature Ecology & Evolution*. 2018

Slivko-Koltchik G *et al.* Evolution of Pannexin/Innexin Gap Junction Protein Family. *Homo* 3: 70.

5 PHYLOTOCOL AMENDMENT HISTORY

Version	Date	Significant Revisions
V1.1	16 July 2019	To better understand the *M. leidyi* gene cluster, using Sebé-Pedros *et al* (2018) single cell data, we were able to determine the cell types that the innexin cluster genes were located in. TPM counts for embryogenesis time course for *M. leidyi* were used to see expression of the gene cluster through the first twenty hours post fertilization. We will also be using MEME and TOMTOM, a motif database and motif identifier, to determine possible regulation factors for the gene cluster.
V1.2	1 October 2019	Expanded the database to include a fourth ctenophore species: *Hormiphora californiensis*. This was done to further explore the possibility of gene loss. Putative innexin genes were identified in the same manner as B. ovata and ran through the same ML and Bayesian tree commands as originally determined.

Version	Date	Significant Revisions
V1.1	16 July 2019	To better understand the *M. leidyi* gene cluster, using Sebé-Pedros *et al* (2018) single cell data, we were able to determine the cell types that the innexin cluster genes were located in. TPM counts for embryogenesis time course for *M. leidyi* were used to see expression of the gene cluster through the first twenty hours post fertilization. We will also be using MEME and TOMTOM, a motif database and motif identifier, to determine possible regulation factors for the gene cluster.
V1.3	12 August 2020	After graphing ctINXA, ctINXB, ctINXC, and ctINXD, it looks as though these might be coregulated during development. We have added section 2.6 to lay out some criteria before testing for correlation.