

neurogenes synteny project

1.

Paper notes - neural genes presence/absences

Moroz et al., 2014 [1]

- Table 34S: -structure of figure shows that it is from bilaterian perspective (closer to bilateria more rectangles filled in)
- remember there are also random absences in cnidaria; the mirror of bilateria, but all the proteins are characterized from bilateria
- no examples where ctenophores or sponges don't have something present in fungi, capsaspora, monosiga
- Suppl Table 12as - it's possible didn't use Amphimedon genome, but Amphimedon was covered by the other papers and I crossreferenced them.

Mentioned in text:

Not in ctenophores:

neurogenin

NeuroD

Achaete-scute

REST

HOX

Otx

-not that much overlap in genes looked at by Riesgo vs Moroz

Ryan et al., 2013 [2]

netrin, slit, unc-5 (axon guidance) not in Mnemiopsis or Amphimedon

-used genomes, since based on Alie and Manuel 2010

Supplementary Table S17: Presence and absence of post-synaptic genes - pretty much Alie and Manuel 2010

Supplementary Table S19: Presence and absence of Dopamine / Norepinephrine /Epinephrine Biosynthetic Pathway components

-are seqs of the animals in S17 genomes? Unsure, but all animals in table have genomes (and the Mle seqs are from the genome)

-AMPA iGluR and NMDA iGluR included as iGluR

Alie and Manuel, 2010 [3]

-used genomes

-Ryan built on Fig. 1. Cross ref with current data to make sure have everything.

-Only use Monosiga, Trichoplax, Amphimedon, Nematostella, Hydra, Homo

Capitella (3 absences), Drosophila (2 absences), Homo very similar with few differences

Unicellular animals mostly missing everything (except B-cat and PMCA). Start with Monosiga which has more things. B-cat and PMCA are ancient - interesting?

AMPA and NMDAR collapsed into iGluR in table; presence of one of these trumped absence of the other
PKC alpha-beta-gamma = PKC on table

Table

Table abbreviations

DBH - dopamine-B-hydroxylase

DDC - DOPA decarboxylase

TH - tyrosine hydroxylase

TPH - tryptophan hydroxylase

PAH - phenylalanine hydroxylase

GAD - glutamate decarboxylase

Qdpr - quinoid dihydropteridine reductase,

Slc18A2 = Homo sapiens solute carrier family 18 member 2,

Pnmt = phenylethanolamine N-methyltransferase

Missing domains

Piccolo - Pleurobrachia - missing ZF (Moroz et al., 2014)

Erbin - Pleurobrachia - missing PDZ (Moroz et al., 2014)

Species names written to the broadest level - eg. *Monosiga brevicollis* in Riesgo et al but only *Monosiga* in Moroz, so put *Monosiga* only

Many entries have NA but if combine:

Salpingoeca + *Monosiga* = Choanoflagellida

Pleurobrachia + Mnemiopsis = Ctenophora

Amphimedon + Oscarella = Porifera

Nematostella + Hydra = Cnidaria

Get only 4 entries that have an NA.

(What about 0/1s (conflicting info?) >> decided to transform 0/1s into NA

Loss_Status: P1C0: present in Porifera, absent in Ctenophora - 6 instances

C1P0: present in Ctenophora, absent in Porifera - 3 instances

T0: absent in Trichoplax but present in Ctenophora or Porifera - 5 instances

Second Iteration:

-There is only 2 instances where Capsaspora has a 1 while choanoflagellates have 0: GABAR and DDC. Don't use column in second iteration

-'Fungi' is very vaguely defined - don't use column in second iteration

-Stopped at Delta catenin -Make new table where all 0/1s or missing_domains (i.e. not 0,1,NA) into NA

Via R Create a new table where species for Ctenophora, Porifera combined: <https://stackoverflow.com/questions/14563531/combine-column-to-remove-nas>

Synten programs/papers

ghost locus hypothesis:

Ramos et al., 2012 [4]

<https://www.sciencedirect.com/science/article/pii/S0960982212009888>

-first ghost locus paper - parahox, Amphimedon + Trichoplax

Fortunato et al, 2014 [5] - sycon Parahox

<https://www.nature.com/articles/nature13881>

-second ghost locus paper?

Ferrier 2015 [6] (review)

<https://academic.oup.com/bfg/article/15/5/333/1741867>

Reviews:

Liu et al 2018 [7]

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2026-4>

Liu Results:

- Pos control (default params): comparison of genome with itself: SynChro (most synteny), DAGchainer/i-ADHoRe, MCScanX -Fragmentation: break 2 species genomes [C. elegans, C. briggsae, S. rattis, S. stercoralis] (diff gene density) into diff fragment lengths. pos corr between error rate and level of fragmentation except for Satsuma. For anchor-based programs, fragmentation has biggest effect on MCScanX and lowest on SynChro.
- Find synteny between sister species: Satsuma worst (diff with alignments/nt homology), Anchor-based programs found much higher synteny - can't access Add. file 2 Table S1 that describes this.
- Fragmented species and compared to its own genome: MCScanX worst. The rest of anchor-based progs performed similarly. SyncChro often better or second best. Degree of error differed per species.
- high gene density, less error (more anchors)
- fragmented assemblies lead to erroneous GO terms.
- reference-guided assembly methods rely on assumption of synteny = synteny errors.

Bottom line:

- most synteny programs designed with assumption working on complete high quality genomes. -SynChro best for fragmented assemblies in general, imo.
- Satsuma does not have a positive correlation with error rate and level of fragmentation, but performed poorly comparing two closely related species due to failure of alignment - go with anchor-based.
- need N50 of at least 200 kb and gene density 290 genes/Mb, or N50 1 Mb and gene density 200 genes/Mb for error rate < 5%. Higher N50 for genomes with less gene density or many paralogues/expansion of gene families.
- annotation quality really doesn't matter just need high genome assembly contiguity.
- more paralogues less synteny finding?? may have misunderstood.

To do:

- Check to see whether genomes assembled via ref to related species -> leads to false synteny due to assumption of shared synteny
- Check N50s of my genomes

Scaffold N50s of my genomes

Mnemiopsis (from orig [2]): 187 kb

Pleurobachia (Suppl Table 5S from [1]): 20.607 kb

Amphimedon orig, Ensembl (https://metazoa.ensembl.org/Amphimedon_queenslandica/Info/Annotation/): 120 kb >> likely improved, looking for Aqu2.1 (or is this only a re-annotation not reassembly?)

- Slightly confused with scaffold N50 stats in [8] suppl Table S2.3.2

Oscarella carmela (Suppl from [9]): 5.897 kb!! :O

SYNTENY ANALYSIS PLAN

Test programs against the same genome - concentrate on the four cteno and por genomes (?).

Determine whether genes of interest are actually found within blocks of synteny (in other animals). Characterize these blocks of synteny.

Examine Cteno and Por genomes.

-ancestral absence: no synteny

-secondary loss: synteny without gene of interest.

First run attempt with Synchro

-Aqu with Aqu: A vs B

.gbff from Genbank = “.dat” for Synchro

f.dat lat file from embl = “.embl” for Synchro

Found .dat aka .embl here: https://www.ebi.ac.uk/ena/data/view/GCA_000090795.1

Found .gbff aka genbank .dat here: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/090/795/GCA_000090795.1_v1.0

genbank: it’s possible .dat is outdated. .gbff has replaced .gbk and Synchro. Yeah, C. elegans gbff doesn’t work.

Big wall: Can’t get file types to work.

ALTERNATE

Ramos method:

Identify all orthologous genes on scaffold = curated list of homologous genes (eg. 27 in Ramos)

Classify as belonging to chromosome in comparison animal where gene of interest is found or somewhere else.

If many paralogues, two tests: single location of each of the human orthologues, or collapsed location of human paralogues if all located at same locus

Do a bunch of stuff to calculate probabilities.

Do Exact binomial test and Fisher’s test to test whether the number of correct neighbours is statistically significantly higher than expected by chance.

-unless they wrote a script it seems very manual and time intensive.

Data/downloads

Pleurobachia genome: https://www.ncbi.nlm.nih.gov/assembly/GCA_000695325.1

P.bachei_draft_genome_v.1.1

Organism: *Pleurobrachia bachei* (ctenophores)

Submitter: University of Florida

Date: 2014/05/21

Assembly level: Scaffold

Genome representation: full

RefSeq category: representative genome

GenBank assembly accession: GCA_000695325.1 (latest)

RefSeq assembly accession: n/a

IDs: 180401 [UID] 1073948 [GenBank]

-paper said deposited at Moroz’s website; links don’t work on website.

Mnemiopsis leidyi: https://www.ncbi.nlm.nih.gov/assembly/GCA_000226015.1/

MneLei_Aug2011

Organism name: *Mnemiopsis leidyi* (sea walnut)

BioSample: SAMN02953801 BioProject: PRJNA64405

Submitter: National Human Genome Research Institute, National Institutes of Health

Date: 2011/09/19

Assembly level: Scaffold

Genome representation: full

RefSeq category: representative genome

GenBank assembly accession: GCA_000226015.1 (latest)

RefSeq assembly accession: n/a

RefSeq assembly and GenBank assembly identical: n/a

WGS Project: AGCP01

Assembly method: Phusion v. 1.02

Genome coverage: 12xSequencing technology: 454 GS-FLX Titanium; Illumina GA IIX

IDs: 304208 [UID] 304208 [GenBank]

Amphimedon queenslandica: v1.0: https://www.ncbi.nlm.nih.gov/assembly/GCF_000090795.1

Organism: *Amphimedon queenslandica* (sponges)

Submitter: US DOE Joint Genome Institute (JGI-PGF)

Date: 2010/05/28

Assembly level: Scaffold

Genome representation: full

RefSeq category: representative genome

GenBank assembly accession: GCA_000090795.1 (latest)

RefSeq assembly accession: GCF_000090795.1 (latest)

IDs: 293608 [UID] 111438 [GenBank] 293608 [RefSeq]

Oscarella carmela: <http://www.compagen.org/datasets.html> OCAR not *Oscarella* sp.

-added .fna and gzipped

-*Oscarella carmela* (this assembly) renamed *Oscarella pearsei*; *Oscarella* sp. in Compagen redescribed *Oscarella carmela*. See <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0183002> - old *O. carmela* papers based on two species (*carmela* + *pearsei*). Both in compagen, but although there is press release compagen didn't change their names in db. *Can't find annotation file.*

Papers+Links

Riesgo et al., 2014

<https://academic.oup.com/mbe/article/31/5/1102/993377>

Neural genes Fig

<https://academic.oup.com/view-large/figure/74385341/msu057f3p.jpeg>

Moroz et al., 2014

<https://www.nature.com/articles/nature13400>

Table 34S: neural genes

<https://media.nature.com/original/nature-assets/nature/journal/v510/n7503/extref/nature13400-s1.pdf>

Ryan et al, 2013

<http://science.sciencemag.org/content/342/6164/1242592>

Suppl Mat:

<http://science.sciencemag.org/content/sci/suppl/2013/12/11/342.6164.1242592.DC1/Ryan.SM.pdf>

Alie and Manuel, 2010

<https://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-10-34>

Srivastava et al., 2010

Suppl.S8.9 - neural genes <https://media.nature.com/original/nature-assets/nature/journal/v466/n7307/extref/nature09201-s1.pdf>

Nichols et al., 2012

<https://www.pnas.org/content/109/32/13046>

Suppl

<https://www.pnas.org/content/pnas/suppl/2012/07/25/1120685109.DCSupplemental/sapp.pdf>

2. Should we be expecting these genes in these animals?

Why should they use similar genes?

How misguided is this approach? What is the true question implied by this approach?

References

1. Leonid L. Moroz MRC Kevin M. Kocot. The ctenophore genome and the evolutionary origins of neural systems. *Nature*. Nature Publishing Group; 2014;510: 109–114. doi:10.1038/nature13400
2. Joseph F. Ryan CES Kevin Pang. The genome of the ctenophore *mnemiopsis leidyi* and its implications for cell type evolution. *Science*. American Association for the Advancement of Science; 2013;342: 1242592. doi:10.1126/science.1242592
3. Alié A, Manuel M. The backbone of the post-synaptic density originated in a unicellular ancestor of choanoflagellates and metazoans. *BMC Evolutionary Biology*. 2010;10: 34. doi:10.1186/1471-2148-10-34
4. Ramos OM, Barker D, Ferrier DE. Ghost loci imply hox and parahox existence in the last common ancestor of animals. *Current biology*. Elsevier; 2012;22: 1951–1956.
5. Fortunato SA, Adamski M, Ramos OM, Leininger S, Liu J, Ferrier DE, et al. Calcisponges have a parahox gene and dynamic expression of dispersed nk homeobox genes. *Nature*. Nature Publishing Group; 2014;514: 620.
6. Ferrier DE. The origin of the hox/parahox genes, the ghost locus hypothesis and the complexity of the first animal. *Briefings in functional genomics*. Oxford University Press; 2015;15: 333–341.
7. Liu D, Hunt M, Tsai IJ. Inferring synteny between genome assemblies: A systematic evaluation. *BMC bioinformatics*. BioMed Central; 2018;19: 26.
8. Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier ME, Mitros T, et al. The amphimedon *queenslandica* genome and the evolution of animal complexity. *Nature*. Nature Publishing Group; 2010;466: 720.
9. Nichols SA, Roberts BW, Richter DJ, Fairclough SR, King N. Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/ β -catenin complex. *Proceedings of the National Academy of Sciences*. National Acad Sciences; 2012;109: 13046–13051.