DATABASE
The Journal of Biological Databases and Curation

A nomenclature for echinoderm genes

Thomas R. Beatman [1,2,*], **Katherine M. Buckley**[3], **Gregory A. Cary**[4], **Veronica F. Hinman**[1,2] and **Charles A. Ettensohn**[1,2]

[1]Department of Biological Sciences, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
[2]Echinobase, #646 Mellon Institute, 4400 Fifth Ave, Pittsburgh, PA 15213, USA
[3]Department of Biological Sciences, Auburn University, 101 Rouse Life Sciences, Auburn, AL 36849, USA
[4]The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

*Corresponding author: Tel: +330-217-2337; Fax: +412-268-7129; Email: tbeatman@cmu.edu

Abstract

Echinoderm embryos and larvae are prominent experimental model systems for studying developmental mechanisms. High-quality, assembled, annotated genome sequences are now available for several echinoderm species, including representatives from most classes. The increased availability of these data necessitates the development of a nomenclature that assigns universally interpretable gene symbols to echinoderm genes to facilitate cross-species comparisons of gene functions, both within echinoderms and across other phyla. This paper describes the implementation of an improved set of echinoderm gene nomenclature guidelines that both communicates meaningful orthology information in protein-coding gene symbols and names and establishes continuity with nomenclatures developed for major vertebrate model organisms, including humans. Differences between the echinoderm gene nomenclature guidelines and vertebrate guidelines are examined and explained. This nomenclature incorporates novel solutions to allow for several types of orthologous relationships, including the single echinoderm genes with multiple vertebrate co-orthologs that result from whole-genome-duplication events. The current version of the Echinoderm Gene Nomenclature Guidelines can be found at https://www.echinobase.org/gene/static/geneNomenclature.jsp

Database URL: https://www.echinobase.org/

Background

Echinoderms have served as important experimental model systems in biology for more than a century, particularly in the field of developmental biology (1–5). Recently, echinoderm embryos and larvae have emerged as a powerful system in which to delineate the gene regulatory networks that operate during embryogenesis (6–9). As a basal lineage of deuterostomes, and because echinoderms exhibit a striking diversity of developmental programs, studies in echinoderms have been invaluable in understanding the evolution of developmental processes (10–16).

Experimental studies using echinoderms have been facilitated, in part, by a wealth of recently available sequencing data. The sequencing of echinoderm genomes began with the Sea Urchin Genome Sequencing Consortium (RRID:SCR_002841) (17), which resulted in genome assembly from the purple sea urchin (*Strongylocentrotus purpuratus*). As technology has advanced, this assembly has been improved several times and is now joined by genome sequences from many other organisms within this phylum. Echinoderm genome assemblies are complemented by a wealth of transcriptome data from several life stages and tissue types (18, 19). These sequence-based resources are a critically important component of modern echinoderm research

(20–23). Gene-based information is organized and accessible via Echinobase, the major public repository of echinoderm genomic data (RRID:SCR_013732) (24).

The existing strategy for annotating echinoderm genes involves assigning genes identifiers referred to as 'names', rather than 'gene symbols', which are more commonly used by other model organism databases (MODs). Proper gene names, which could provide clarification of identity and function, are listed as 'synonyms' but are not consistently present for echinoderm genes. As a result, legacy echinoderm gene names are not easily translatable across MODs. This inconsistency raises the risk of researchers incorrectly conflating echinoderm genes with one another or with genes in other organisms based on nomenclature rather than true evolutionary relationships. Establishing coherence in the naming of orthologous genes is a powerful rationale for aligning nomenclatures between different MODs (25). In echinoderms, this problem is further complicated by a lack of clarification and codified standards underlying the present nomenclature. In addition, the high degree of polymorphism in some echinoderm species has led to the existence of numerous artifactual duplicates of individual genes in assemblies (26). Finally, as the amount of sequencing data expands, it becomes increasingly important to ensure that gene symbols and names are both human and

machine readable to facilitate their implementation in Echinobase. Standardizing nomenclature according to orthology with established vertebrate genomes will facilitate comparative biological research across deuterostomes.

With the release of an improved, chromosome-scale assembly and annotation of the *S. purpuratus* genome (Spur_5.0), as well as major improvements to Echinobase, the web resource hosting genome-related information, the development of an improved gene nomenclature is timely. Accordingly, an Echinobase Gene Nomenclature Committee (EGNC) was formed to generate a standardized echinoderm nomenclature. The revised nomenclature was developed using the nomenclature guidelines built for *Xenopus* species (27) as a scaffold. A core goal of this nomenclature system is to incorporate robust orthology relationships between echinoderm and human genes when assigning gene symbols and names. Here, we describe these new nomenclature guidelines for echinoderm genes and illustrate their utility in integrating biological information obtained from echinoderms with related information from vertebrate models.

Criteria for echinoderm genomes

As a product of Echinobase, the nomenclature pipeline has been developed for use with genomes supported by the MOD. Genomes are considered for hosting on Echinobase and processing through our nomenclature pipeline once they have been processed, approved and annotated by NCBI's RefSeq eukaryotic genome annotation pipeline (28, 29). This serves as an external validator of genome quality and systematically produces provisional gene identifiers for Echinobase. The order of genome integration on Echinobase is then prioritized to improve breadth of taxonomic diversity before expanding depth.

Nomenclature for echinoderm gene names and symbols

The revised Echinobase gene nomenclature guidelines have transitioned from solely having gene symbols (historically identified on Echinobase as 'gene names'), to having both full gene names and short symbols (together referred to herein as 'gene identifiers'). This strategy is modeled after the common features of the nomenclatures used for the human (30), mouse and rat (31), chicken (32), *Xenopus* (33) and zebrafish MODs (34). Furthermore, gene identifiers are formatted similarly to vertebrate gene names and symbols: gene names and symbols are presented in lower case and italicized, and Greek letters and Roman numerals are converted to Latin and Arabic equivalents, respectively (35). As gene pages are updated in accordance with the guidelines described here, all legacy names, symbols and aliases are relegated to the synonym lists for each gene page. When users may wish to refer to a gene by its previous symbol, we recommend referring to it by appending the previous symbol in parentheses following the current symbol. The synonym list is fully integrated into Echinobase's gene search tool, enabling users to locate genes by historically associated identifiers when needed. Up to date mappings of current gene identifiers to NCBI locus IDs will be provided both through Echinobase's jbrowse tracks and within Echinobase's FTP site. While an increasing number of human gene identifiers are being made stable (30, 36), to account for possible changes the up to date gene identifiers of human genes

will be annually cross-referenced to keep our nomenclature in alignment.

Orthology pipeline

To facilitate intra- and interphylum comparisons, echinoderm genes are assigned identifiers on the basis of orthology. To assess orthology, we use the previously described orthology pipeline developed by Echinobase to build orthology maps (37), whose approach is based on the DRSC Integrative Ortholog Prediction Tool (DIOPT) (38) and the HGNC Comparison of Orthology Predictions tool (39), integrating output from several algorithms to build consensus on orthology relationships. Genes are determined to be orthologs if three or more tools used in the Echinobase orthology pipeline support an orthologous relationship, in accordance with Alliance of Genome Resources (RRID:SCR_015850) standards (37, 40). Any orthologous relationships that meet this threshold are used to inform the nomenclature. At present, the orthology pipeline is currently composed of six tools: InParanoid v4.1 (RRID:SCR_006801) (41, 42), ProteinOrtho v6 (43), SwiftOrtho (RRID:SCR_017122) (44), FastOrtho (45), OMA v2.4.1 (RRID:SCR_011978) (46), and OrthoFinder v2.4 (RRID:SCR_017118) (47). Selection of tools is informed by local accessibility/functionality by Echinobase bioinformaticians, resulting in the prioritizing of tools that have publicly available documentation or tools whose providers were able to run our data internally and provide Echinobase with outputs. Anticipated expansion of the orthology pipeline will eventually increase the number of metrics to 12 or more tools to allow for a DIOPT-like analysis (37, 38). The completion of Echinobase's orthology pipeline with a full suite of tools will precede the implementation of more complex orthologous relationships (e.g. one:many and many:one echinoderm:human orthologs) in our nomenclature as described below.

Decision regarding nonhuman-vertebrate orthologs and nonvertebrate orthologs

Some consideration was given to expanding the orthology-derived nomenclature to allow for naming echinoderm genes on the basis of orthology with nonhuman orthologs. Three of the orthology tools outputs used in the current Echinobase orthology pipeline are provided from DIOPT [Inparanoid (41, 42), OMA (46) and Orthofinder (47)] that uses protein models derived from the *S. purpuratus* genome and those of numerous major model organisms (37, 38). Of the *S. purpuratus* proteins within this subset of orthology tools that meet the three-criteria threshold (i.e. 'nameable genes', 8066 in total; all of these nameable genes were one:one orthologies), 70.79% (5710 proteins) have human orthologs. This reflects 20.8% of all 27 447 *S. purpuratus* protein-coding genes. Further 17.3% of the nameable *S. purpuratus* genes (1392 total; 5.07% of all *S. purpuratus* protein-coding genes) lack orthologs in humans but do have other one:one vertebrate orthologs. For the time being, given this low fraction of genes and the limited coverage of the orthology pipeline at this time, we have maintained the current nomenclature pipeline with a focus on intraphylum and echinoderm:human orthologies; this strategy may be re-visited in the future once the scope of the orthology relationships is expanded for analysis of additional vertebrate species in the Echinobase orthology pipeline tools. *S. purpuratus* genes with only invertebrate

or non-animal orthologs comprise 12% of nameable genes from these three tools (964 genes; 3.5% of all *S. purpuratus* protein-coding genes). Because these systems' nomenclatures include features that are incompatible with ours and those of vertebrate MODs (primarily common use of characters that conflict with our machine-readability criteria), these systems are currently excluded from informing gene nomenclature.

Nomenclature for echinoderm genes with single human orthologs (one:one)

The nomenclature pipeline begins with echinoderm genomes processed through the NCBI annotation pipeline, which provides initial annotation. Annotated sequences are then processed through Echinobase's orthology pipeline (37) and subsequently through the nomenclature pipeline (Figure 1). In the most straightforward cases, the orthology pipeline identifies a single echinoderm gene that is orthologous to a single human gene. If an echinoderm gene is orthologous to a single human ortholog, it is assigned the human gene identifier. The current suite of orthology pipeline outputs is being used to generate one:one orthology-derived gene identifiers, reflecting the low likelihood that further tools will cause shifts in nomenclature but identifiers will be updated in accordance with updates to the orthology pipeline as the need arises. The assignment of more complex orthology relationships between echinoderms and human (i.e. one:many, many:one, and many:many) then go through further processing.

Nomenclature for echinoderm genes orthologous to multiple human genes (one:many)

Many of the primary differences between echinoderm and vertebrate genome sequences are consequences of the two rounds of whole genome duplication that occurred at the base of the vertebrate clade (48–50). Thus, echinoderm genomes often contain a single gene or smaller subset of genes than are present in vertebrate genomes. This requires a robust set of guidelines for naming echinoderm genes with multiple human orthologs. In the event of multiple orthologs generated by differing numbers of orthology metrics, the approach taken by the EGNC is to assign identifiers to such genes after the ortholog that is supported by the most orthology tools. In the event that multiple human orthologs sharing the same number of metrics are all members of the same gene family (have the same stem symbol) and no members of that family show orthology to another echinoderm species' gene, the echinoderm gene will be named with the stem, rather than a specific family members identifiers. When this is not the case, curators will examine such sets to determine if relevant information can resolve ties. If this is not possible the first ortholog alphanumerically is selected. This approach is a tradeoff between the loss of orthology information embedded in the gene identifiers and maintaining machine readability and ease of use of identifiers by users. More details of orthology information will be provided on each gene page on Echinobase to supplement informational constraints in the gene identifiers proper.

Nomenclature for sets of echinoderm paralogs that are orthologous to one or more human orthologs (many:one or many:many)

In contrast to the one:many relationships described above, echinoderm genomes are characterized by many lineage-specific gene family expansions that result in many:one or many:many orthologies with vertebrate genes. This is particularly true for rapidly evolving immune genes and genes involved in making echinoderm skeletons (51, 52). For these cases, genomes are assessed to identify pseudoduplicates, defined here as highly similar genes that may or may not be assembly artifacts, which are treated differently in this nomenclature than paralogs which fall outside of this classification.

Pseudoduplicates

As echinoderms are broadcast spawners that produce highly outbred diploid crosses, individual echinoderms' genome sequences are highly polymorphic[817](53). Consequently, assembling a single haplotype is a challenge and can result in under-collapsed heterozygosity in which several alleles are retained in genome assemblies (27, 54), resulting in groups of genes with highly similar sequences. It is expected that with each new assembly, the occurrence of false duplicates will be reduced and these false expansions will be identified and collapsed. To identify false duplicates, we perform a BLAST of all gene models (including introns and 1 kb up- and downstream of the genic sequence) in each genome against themselves and then extract those that match on 90% or more identity along 90% or more of the longer sequence's length. This approach is relatively conservative compared to extant methods for identifying false duplicates and collapsing heterozygosity post-assembly (55) and serves to identify genes whose high similarity will confound orthology assignments. Given this likelihood, we refer to clusters of multiple genes with highly similar sequences as 'pseudoduplicates'.

Genes identified as pseudoduplicates should be assigned the human gene identifiers appended with a decimal point followed by a 'letter' (e.g. gene.a, gene.b, etc.). This provides each individual gene with a distinct identity and the associated orthology information while indicating to users that the sequence belongs to a cluster of pseudoduplicates and may require additional confirmation as a true paralog. Suffixes appended to pseudoduplicates in individual echinoderm species will be independent of one another and do not reflect any specific orthology information.

Paralogs

For sets of paralogous genes that have many:one relationships with humans and are not classified as pseudoduplicates, gene identifiers are matched to the single human ortholog and then appended with a '.#' suffix. This approach conveys both orthology and membership in a gene subfamily. When echinoderm paralogs have multiple human orthologs, the above rule is conjoined with the one:many rules, using whichever human ortholog matches on the most metrics, or, preceding any pertinent determining information, the first alphanumerically if multiple human genes tie for most metrics as the stem for the suffix. Whenever able, orthologs across species will be given the same suffixes.

Nomenclature for echinoderm genes that lack human orthologs

There remains a set of echinoderm genes (i.e. those without human orthologs) that is not addressed by the preceding
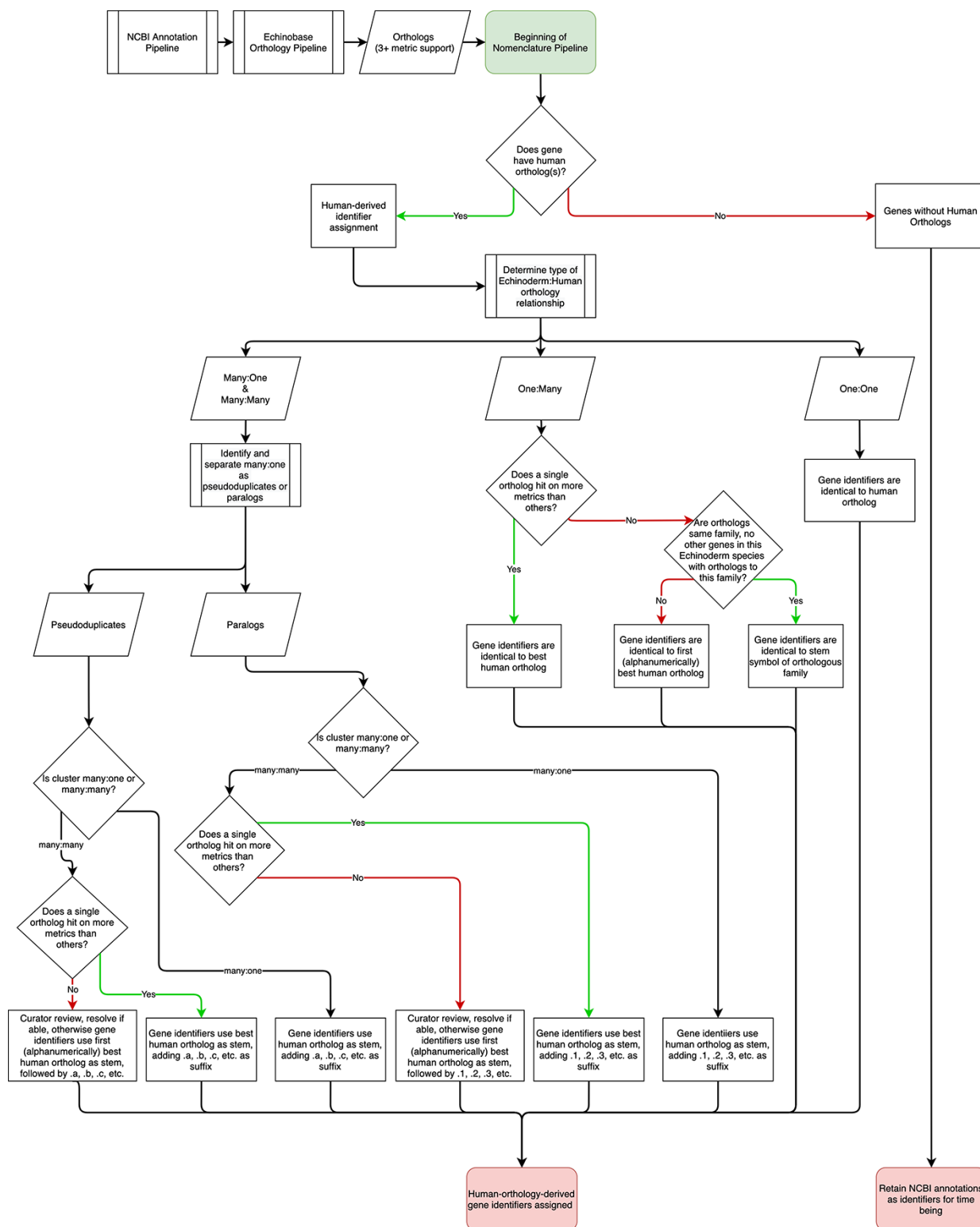
Figure 1. Flowchart describing the Echinobase nomenclature pipeline. Following processing by NCBI and the Echinobase orthology pipeline, gene identifiers are assigned depending on orthology relationships to humans.

components of the nomenclature pipeline. In vertebrate nomenclatures, this is typically addressed by leaving these genes with their NCBI annotation provided identifiers (typically these are formatted with a symbol comprising of LOC followed by the entrez ID of the gene and a name that is based on its predicted protein product), with the opportunity for new identifiers to be generated in conjunction with authors of papers studying such genes (30). This approach is retained here, which allows for the reimplementation of established legacy identifiers provided they meet our nomenclature

guidelines, the generation of new identifiers following new studies or an expansion of the Echinobase nomenclature pipeline to additional species (see above), in coordination with the EGNC coordinator.

Future directions

With assembled genome sequences from *S. purpuratus* and *Acanthaster planci* (OKI-Apl_1.0 (56)) incorporated into Echinobase and assemblies for *Anneissia japonica (ASM1163010v1), Patiria miniata (Pmin_3.0)* and *Lytechinus variegatus (Lvar_3.0* (57)) soon to be supported, intercommunicability of gene identities between species will spur comparative analysis of gene structure and function. 'Following the assembly of the complete orthology pipeline (12+ tools) more complex (one:many, many:one, many:many) echinoderm:human orthologies will be used to inform associated gene identities'. Planned future additions to the Echinoderm Gene Nomenclature Guidelines include developing nomenclatures for non-coding RNAs, regulatory DNA elements (e.g. enhancers), and gene-related reagents (morpholino antisense oligonucleotides, antibodies, guide RNAs, etc.), in alignment with human gene nomenclature standards whenever possible. As the resources provided by Echinobase expand in conjunction with this revised nomenclature, the usability and functionality of this resource will continue to improve in the coming years.

Acknowledgements

Author's contributions

T.R.B. wrote the first draft. All authors edited the manuscript, provided constructive feedback and approved the final version. T.R.B., K.M.B., and G.A.C. took part in the elaboration of the nomenclature.

References

1. Angerer,L.M. and Angerer,R.C. (2003) 4 Patterning the sea urchin embryo: gene regulatory networks, signaling pathways, and cellular interactions. *Curr. Top. Dev. Biol.*, **53**, 159–198.

2. Wilt,F.H. (2005) Developmental biology meets materials science: morphogenesis of biomineralized structures. *Dev. Biol.*, **280**, 15–25.

3. Ernst,S.G. (2011) Offerings from an Urchin. *Dev. Biol.*, **358**, 285–294.

4. Adomako-Ankomah,A. and Ettensohn,C.A. (2014) Growth factors and early mesoderm morphogenesis: insights from the sea urchin embryo. *Genesis*, **52**, 158–172.

5. McClay,D.R. (2016) Sea urchin morphogenesis. *Curr. Top. Dev. Biol.*, **117**, 15–29.

6. Davidson,E.H. (2009) Network design principles from the sea urchin embryo. *Curr. Opin. Genet. Dev.*, **19**, 535–540.

7. Ettensohn,C.A. (2013) Encoding anatomy: developmental gene regulatory networks and morphogenesis. *Genesis*, **51**, 383–409.

8. Martik,M.L., Lyons,D.C. and McClay,D.R. (2016) Developmental gene regulatory networks in sea urchins and what we can learn from them. *F1000Research*, **5**, 1–9.

9. Peter,I.S. (2017) Regulatory states in the developmental control of gene expression. *Brief. Funct. Genomics*, **16**, 281–287.

10. Raff,R.A. and Snoke Smith,M. (2009) Chapter 7 axis formation and the rapid evolutionary transformation of larval form. *Curr. Top. Dev. Biol.*, **86**, 163–190.

11. Koga,H., Morino,Y. and Wada,H. (2014) The echinoderm larval skeleton as a possible model system for experimental evolutionary biology. *Genesis*, **52**, 186–192.

12. Wessel,G.M. (2016) Germ line mechanics—and unfinished business. *Curr. Top. Dev. Biol.*, **117**, 553–566.

13. Cary,G.A. and Hinman,V.F. (2017) Echinoderm development and evolution in the post-genomic era. *Dev. Biol.*, **427**, 203–211.

14. Shashikant,T., Khor,J.M. and Ettensohn,C.A. (2018) From genome to anatomy: the architecture and evolution of the skeletogenic gene regulatory network of sea urchins and other echinoderms. *Genesis*, **56**, e23253.

15. Annunziata,R., Andrikou,C., Perillo,M. *et al.* (2019) Development and evolution of gut structures: from molecules to function. *Cell Tissue Res.*, **377**, 445–458.

16. Carlisle,J.A. and Swanson,W.J. (2020) Molecular mechanisms and evolution of fertilization proteins. *J. Exp. Zool. B Mol. Dev. Evol.*, 1–14.

17. Sea Urchin Genome Sequencing Consortium (2006) The genome of the sea urchin strongylocentrotus purpuratus. *Science*, **314**, 941.

18. Tu,Q., Cameron,R.A., Worley,K.C. *et al.* (2012) Gene structure in the sea urchin strongylocentrotus purpuratus based on transcriptome analysis. *Genome Res.*, **22**, 2079–2087.

19. Janies,D.A., Witter,Z., Linchangco,G.V. *et al.* (2016) EchinoDB, an application for comparative transcriptomics of deeply-sampled clades of echinoderms. *BMC Bioinform.*, **17**, 48.

20. Lowe,E.K., Cuomo,C. and Arnone,M.I. (2017) Omics approaches to study gene regulatory networks for development in echinoderms. *Brief. Funct. Genomics*, **16**, 299–308.

21. Buckley,K.M. and Ettensohn,C.A. (2019) Techniques for analyzing gene expression using BAC-based reporter constructs. *Methods Cell Biol.*, **151**, 197–218.

22. Cary,G.A., Cameron,R.A. and Hinman,V.F. (2019) Genomic resources for the study of echinoderm development and evolution. *Methods Cell Biol.*, **151**, 65–88.

23. Peter,I.S. (2019) Methods for the experimental and computational analysis of gene regulatory networks in sea urchins. *Methods Cell Biol.*, **151**, 89–113.

24. Cary,G.A., Cameron,R.A. and Hinman,V.F. (2018) EchinoBase: tools for echinoderm genome analyses. *Methods Mol. Biol.*, **1757**, 349–369.

25. Wright,M.W. and Bruford,E.A. (2006) Human and orthologous gene nomenclature. *Gene*, **369**, 1–6.

26. Kelley,D.R. and Salzberg,S.L. (2010) Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome Biol.*, **11**, 1–11.

27. Xenbase (n.d.). Xenbase gene nomenclature guidelines. http://www.xenbase.org/gene/static/geneNomenclature.jsp (21 January 2021, date last accessed).

28. Thibaud-Nissen,F., DiCuccio,M., Hlavina,W. *et al.* (2016) P8008 the NCBI eukaryotic genome annotation pipeline. *J. Anim. Sci.*, **94**, 184.

29. Eukaryotic Genome Annotation at NCBI (n.d.). https://www.ncbi.nlm.nih.gov/genome/annotation_euk/ (8 March 2021, date last accessed).

30. Bruford,E.A., Braschi,B., Denny,P. *et al.* (2020) Guidelines for human gene nomenclature. *Nat. Genet.*, **52**, 754–758.

31. MGI-Guidelines for Nomenclature of Genes, Genetic Markers, Alleles, & Mutations in Mouse & Rat (n.d.). http://www.informatics.jax.org/mgihome/nomen/gene.shtml (23 March 2021, date last acessed).

32. Burt,D.W., Carrë,W., Fell,M. *et al.* (2009) The chicken gene nomenclature committee report. *BMC Genomics*, **10**, 1–5.

33. James-Zorn,C., Ponferrada,V.G., Burns,K.A. *et al.* (2015) Xenbase: core features, data acquisition, and data processing. *Genesis*, **53**, 486–497.

34. Ruzicka,L., Howe,D.G., Ramachandran,S. *et al.* (2019) The zebrafish information network: new support for non-coding genes, richer gene ontology annotations and the alliance of genome resources. *Nucleic Acids Res.*, **47**, D867–D873.

35. Kusumi,K., Kulathinal,R.J., Abzhanov,A. *et al.* (2011) Developing a community-based genetic nomenclature for anole lizards. *BMC Genomics*, **12**, 1–13.

36. Tweedie,S., Braschi,B., Gray,K. *et al.* (2021) Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res.*, **49**, D939–D946.

37. Foley,S., Ku,C., Arshinoff,B. *et al.* (2021) Integration of 1:1 orthology maps and updated datasets into Echinobase. Database Vol. 2021: article ID baab030.

38. Hu,Y., Flockhart,I., Vinayagam,A. *et al.* (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinform.*, **12**, 1–16.

39. Wright,M.W., Eyre,T.A., Lush,M.J. *et al.* (2005) HCOP: the HGNC comparison of orthology predictions search tool. *Mamm. Genome*, **16**, 827–828.

40. Agapite,J., Albou,L.P., Aleksander,S. *et al.* (2020) Alliance of genome resources portal: unified model organism research platform. *Nucleic Acids Res.*, **48**, D650–D658.

41. Östlund,G., Schmitt,T., Forslund,K. *et al.* (2009) Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196.

42. Remm,M., Storm,C.E.V. and Sonnhammer,E.L.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.

43. Lechner,M., Findeiß,S., Steiner,L. *et al.* (2011) Proteinortho: detection of (Co-)orthologs in large-scale analysis. *BMC Bioinform.*, **12**, 1–9.

44. Hu,X. and Friedberg,I. (2019) SwiftOrtho: a fast, memory-efficient, multiple genome orthology classifier. *GigaScience*, **8**, 1–12.

45. Wattam,A.R., Abraham,D., Dalay,O. *et al.* (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, **42**, D581–D591.

46. Altenhoff,A.M., Glover,N.M., Train,C.M. *et al.* (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.*, **46**, D477–D485.

47. Emms,D.M. and Kelly,S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.*, **20**, 1.

48. Dehal,P. and Boore,J.L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.*, **3**, e314.

49. Putnam,N.H., Butts,T., Ferrier,D.E.K. *et al.* (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**, 1064–1071.

50. Smith,J.J. and Keinath,M.C. (2015) The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. *Genome Res.*, **25**, 1081–1090.

51. Rast,J.P., Smith,L.C., Loza-Coll,M. *et al.* (2006) Genomic insights into the immune system of the sea urchin. *Science*, **314**, 952–956.

52. Rafiq,K., Cheers,M.S. and Ettensohn,C.A. (2012) The genomic regulatory control of skeletal morphogenesis in the sea urchin. *Development*, **139**, 579–590.

53. Britten,R.J., Cetta,A. and Davidson,E.H. (1978) The single-copy DNA sequence polymorphism of the sea urchin strongylocentrotus purpuratus. *Cell*, **15**, 1175–1186.

54. Rhie,A., McCarthy,S.A., Fedrigo,O. *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, **592**, 737–746.

55. Barrière,A., Yang,S.P., Pekarek,E. *et al.* (2009) Detecting heterozygosity in shotgun genome assemblies: lessons from obligately outcrossing nematodes. *Genome Res.*, **19**, 470–480.

56. Hall,M.R., Kocot,K.M., Baughman,K.W. *et al.* (2017) The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef pest. *Nature*, **544**, 231–234.

57. Davidson,P.L., Guo,H., Wang,L. *et al.* (2020) Chromosomal-level genome assembly of the sea urchin Lytechinus variegatus substantially improves functional genomic analyses. *Genome Biol Evol*, **12**, 1080–1086.