Current Topics in Genetics
Vol. 5, 2012

Mini-Review

A convention for naming molluscan genes

Christopher J. Bayne*
Department of Zoology, Oregon State University, Corvallis, Oregon 97331, USA

ABSTRACT

As the genomes of increasing numbers of organisms are sequenced, the value of new sequence data is enhanced when the process that is used to assign names to the sequenced genes is clear and consensual. The optimum time to erect a naming system for any taxonomic group is early in the expansion of the genome databases. In spite of being one of the major metazoan phyla, the Mollusca (~93,000 living species) are latecomers to the field of genomics. As molluscs include species that are used as food, others that transmit human diseases, or serve as biomedical models, or play important roles in both terrestrial and aquatic ecosystems, or are admired for their esthetic qualities, a marked increase in genomic data is anticipated. As the first sequenced molluscan genomes (a snail *Biomphalaria glabrata*, and a bivalve *Crassostrea gigas*) will have been released in 2012, a convention is proposed to establish a clear process and a rational naming system. The molluscs comprise one major grouping within the metazoan Lophotrochozoa, so the convention may prove to be of value for a wider array of related taxa.

KEYWORDS: mollusc, Lophotrochozoa, annelid, Platyhelminthes, gene name

INTRODUCTION

For each major taxon within the universe of living organisms, the process of naming genes is both facilitated and improved by the adoption of a suitable convention. In this context, we have considered the needs of stakeholders with interests in molluscan genomes and genetics. Although the phylum Mollusca is enormous (~93,000 known living species), and includes several species that are either economically (e.g. shellfish as food and as indicators of environmental health) or medically (e.g. vectors of human disease) important or are used as models for basic research (e.g. squid, *Octopus* and *Aplysia*), they are late comers to the field of genomics. The history of life reveals the molluscs' lack of close evolutionary relationships to the taxa whose genomes have been thoroughly sequenced and assembled at high quality: Bacteria, Archaea, Deuterostomia such as vertebrates, agnathans and tunicates, and Ecdysozoa such as insects and nematodes. Molluscs are members of another major metazoan grouping in which the underived state of early embryogenesis includes a swimming larva with a band of cilia (the trochophore), a blastopore that persists as the mouth, and determinate cleavage (the fates of the blastomeres are determined early). The major metazoan phyla with these characteristics are the Platyhelminthes, Mollusca and Annelida, and they comprise the Lophotrochozoa.

In the absence of a convention for naming genes, several names can be and often have been assigned for a single genetic locus, illustrated for example by *ATP-binding cassette, sub-family B (MDR/TAP), member 1 (ABCB1), multidrug resistance 1 (mdr1)*, and *glycoprotein P (pgy1)* [1]. This confuses rather than clarifies, and is unfortunate. The purpose of the present paper is to move towards the adoption of a gene-naming system for the Mollusca that provides informative names, and that will be readily adopted.

The first completed genome sequence for a mollusc was released in 2012 [2], and a

*baynec@science.oregonstate.edu

second will soon follow (WUGSC and http://biology.unm.edu/biomphalaria-genome/index.html), so it is timely to adopt a rational system of gene naming for the parent phylum. Drawing on a document that originated with The Council of Scientific Editors and titled: "Resources for Genetic and Cytogenetic Nomenclature" [3], and on guidelines from the human gene-naming authority [4], I herein propose guidelines for consideration by the community of people interested in molluscs, of which several species are currently the subjects of genomic and transcriptomic projects including *Biomphalaria glabrata* [5], *Lottia gigantea* [6] other gastropods, and bivalves such as oysters [1], mussels and others. In addition, the project to DNA Bar Code the universe of life [7] has rapidly increased the number of species for which sequence data on small regions [notably a 648-bp region of the mitochondrial cytochrome c oxidase subunit I (COI) gene] are being added to public databases. Our thinking has benefitted from the work of the Locus. Reference. Genomic authors [8] who base their recommendations on NCBI's RefSeqGene project [9].

Those who have gone before have been preemptive in adopting conventions that yield systems of nomenclature that are unique to one or a few species. Obviously, such practices will not be sustainable as the number of sequenced genomes increases. Accordingly, from the list of species for which conventions already exist, we have selected the zebrafish as a model for our system of nomenclature, basing this on the view that the zebrafish community is large, has shown itself to be thoughtful and deliberate, and is (like us) concerned with an unconventional model organism [10]. Invertebrates whose user groups have created conventions for naming genes include a widely known nematode (*Coenorhabditis elegans*) and a dipteran insect (*Drosophila melanogaster*). Even though these are invertebrates, for our purposes they are not considered to be more suitable models than the vertebrate zebrafish for the following reasons: as ecdysozoans they may be as distantly related to the molluscs as are the deuterostome vertebrates (e.g. zebrafish); the manner in which (some of) their genes have been named is arcane, and often the names are not informative of any aspect of biology. Finally, if a universal system of gene nomenclature is ever developed, it will likely more closely resemble the extant vertebrate (probably human) systems thereby requiring less change for a molluscan system that is modeled after a chordate.

Paraphrasing from the Zebrafish Book, Chapter 7, "*It is very important that all members of the community adopt one set of conventions in order to minimize confusion and maximize the usefulness of the nomenclature and the ease with which everyone can follow the field*".

Desirable features of a system of genetic nomenclature for molluscs, with a note on proteins

Genes

The name should reflect the (putative) function of the product, if known or suspected (for example based on sequence similarity to a gene that has been identified with certainty and whose function(s) is/are known in another organism). An abbreviation should be provided, with an optimum of three letters and a maximum of 6. The letters should be all lower case and italicized. The name need not seek to identify the species; this should be obvious from the context. Table 1 provides examples.

Alleles

These should be the gene name followed sequentially by italicized *A, B, C, D* etc., in capitals

Table 1. Examples illustrating the application of the proposed convention.

Gene name	Abbreviation	Allele	Mutant	Protein name
actin	*act*	*actA*	*actA*^{1}	Actin
guanosine binding protein	*gnb*	*gnbA*	*gnbA*^{1}	Guanosine Binding Protein
superoxide dismutase 1	*sod1*	*sod1A*	*sod1A*^{1}	Superoxide Dismutase 1
glucose-6-phosphate dehydrogenase	*g6pdh*	*g6pdhA*	*g6pdhA*^{1}	Glucose-6-phosphate Dehydrogenase

and placed with no space or hyphen immediately after the gene name.

Mutants

Mutants will be indicated by italicized superscript numeral(s) following the gene/allele name.

Since, strictly speaking, each sequence variant at a given genomic locus may be considered to be an allele, there is need for a criterion for calling a sequence variant a 'mutant' rather than an 'allele'. In cases where there is a demonstrated phenotype ascribable to that variant, the case is clear: the sequence variant is an allele. We suggest that variants with synonymous substitutions (variants in which substituted nucleotides do not encode different amino acids) should be considered mutant alleles, and identified by a superscript numeral after the allele capital letter. Until consensus can be reached, it is left to the discretion of the individual scientist to decide if a variant with one or more non-synonymous substitutions is to be given a unique allele designation without mutant indicated, or be considered a mutant of an already named allele. Designation as an allele will be more strongly warranted when the altered peptide is deemed likely to have (selectively and functionally relevant) phenotypic consequences (altered fitness).

Proteins

The name is the same as the full name of the gene, except it is not italicized, and the first letter is capitalized. In cases where a single name is used for several homologous proteins with the same functions (and encoded at distinct loci), the genes may be numbered in accordance with the protein nomenclature (e.g. Superoxide dismutase 1, Superoxide dismutase 2 and Superoxide dismutase 3).

DISCUSSION

The consensual adoption of a readily understood system of nomenclature will help reduce misunderstanding and confusion. Of course, like taxonomy, a convention is mutable. Indeed, as it is impossible to foresee every situation, complications will arise, and one must remain agreeable to the occasional exception. The purpose of this paper is to encourage both the use of the proposed convention and constructive discussion of ways in which it might be improved. This encouragement comes at an opportune time in the history of molluscan genomics. Geneticists working with the closest relatives of the molluscs (other lophotrochozoans - Annelida, Platyhelminthes and others) may find this convention suitable as they describe genomes in additional species.

ACKNOWLEDGEMENTS AND APOLOGIA

My efforts have been supported by the US NIH, Oregon State University and several fellowships. While colleagues and friends too numerous to name have contributed to my education as far as it has come, all errors and oversights in this work are my responsibility.

REFERENCES

1. http://atlasgeneticsoncology.org//Genes/PGY1ID105.html
2. Zhang, G., Fang, X., Guo X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H., Xiong, Z., Que, H., Xie, Y., Holland, P. W., Paps, J., Zhu, Y., Wu, F., Chen, Y., Wang, J., Peng, C., Meng, J., Yang, L., Liu, J., Wen, B., Zhang, N., Huang, Z., Zhu, Q., Feng, Y., Mount, A., Hedgecock, D., Xu, Z., Liu, Y., Domazet-Lošo, T., Du, Y., Sun, X., Zhang, S., Liu, B., Cheng, P., Jiang, X., Li, J., Fan, D., Wang, W., Fu, W., Wang, T., Wang, B., Zhang, J., Peng, Z., Li, Y., Li, N., Wang, J., Chen, M., He, Y., Tan, F., Song, X., Zheng, Q., Huang, R., Yang, H., Du, X., Chen, L., Yang, M., Gaffney, P. M., Wang, S., Luo, L., She, Z., Ming, Y., Huang, W., Zhang, S., Huang, B., Zhang, Y., Qu, T., Ni, P., Miao, G., Wang, J., Wang, Q., Steinberg, C. E., Wang, H., Li, N., Qian, L., Zhang, G., Li, Y., Yang, H., Liu, X., Wang, J., Yin, Y., and Wang, J. 2012 Oct 4; 490(7418): 49-54. doi: 10.1038/nature11413. Epub. 2012 Sep 19.
3. http://en.wikipedia.org/wiki/Gene_nomenclature
4. http://www.genenames.org/guidelines.html
5. http://www.uniprot.org/uniprot/?query=taxonomy:6526
6. http://genome.jgi-psf.org/Lotgi1/Lotgi1.info.html
7. http://en.wikipedia.org/wiki/DNA_barcoding

8. Dalgleish, R., Flicek, P., Cunningham, F., Astashyn, A., Tully, R.E., Proctor, G., Chen, Y., McLaren, W. M., Larsson, P., Vaughan, B. W., Béroud, C., Dobson, G., Lehväslaiho, H., Taschner, P. E. M., den Dunnen, J. T., Devereau, A., Birney, E., Brookes, A. J., and Maglott, D. R. 2010, Genome Medicine, 2, 24. doi:10.1186/gm145. More generally http://www.lrg-sequence.org/
9. http://www.ncbi.nlm.nih.gov/refseq/rsg/
10. https://wiki.zfin.org/display/general/ZFIN+Zebrafish+Nomenclature+Guidelines