The Case for Standardising Gene Nomenclature across Vertebrates

Fiona M. McCarthy[1]*, Tamsin E.M. Jones[2]*, Anne E. Kwitek[3], Cynthia L. Smith[4], Peter D. Vize[5], Monte Westerfield[6], Elspeth A. Bruford[2,7]**

***These authors contributed equally to this work**

1 The Chicken Gene Nomenclature Committee (CGNC), School of Animal and Comparative Biomedical Sciences, University of Arizona, AZ, USA

2 HUGO Gene Nomenclature Committee (HGNC), European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK

3 Rat Genome Database, Medical College of Wisconsin, Milwaukee, WI, USA

4 Mouse Genome Database, The Jackson Laboratory, Bar Harbor, ME, USA

5 Xenbase, Departments of Biological Sciences and Computer Science, University of Calgary, Calgary, Alberta, Canada

6 ZFIN, Institute of Neuroscience, University of Oregon, Eugene, OR, USA

7 Department of Haematology, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge, UK

**** To whom correspondence should be addressed**

ABSTRACT

Standardized gene nomenclature supports unambiguous communication and identification of the scientific literature associated with genes. To support the increasing number of annotated genomes that are now available for comparative studies, gene nomenclature authorities coordinate the assignment of approved gene names that can be readily propagated across species without losing their sense of meaning. Theofanopoulou *et al* (Theofanopoulou et al. 2021) propose nomenclature changes to the genes encoding oxytocin and arginine vasopressin and their receptors which would hinder comparative studies and literature identification. Instead, we propose minor updates to the current approved nomenclature of these vertebrate genes to better reflect their evolutionary history, without confusing the literature that already exists around these well-studied genes. We encourage authors to work with nomenclature committees to ensure any novel gene names fit current guidelines so that their publications can be readily indexed and made accessible. Moreover, we call on journal editors and reviewers to help support communication and

indexing of gene-related publications by ensuring that standardized gene nomenclature is routinely used.

Standardized gene nomenclature. Standardized gene nomenclature provides a common language for the biomedical community, and beyond. Gene nomenclature refers to both the full gene name and the unique gene symbol; often aliases (or synonyms) used in published literature are also recorded. Projects to sequence multiple genomes (e.g. Rhie et al. 2021; Lewin et al. 2018) are expanding our ability to include more than just a handful of species as a proxy for all organisms, and it is increasingly important that standardized gene nomenclature is available as a point of reference for these new genomes. To support this in vertebrates, gene nomenclature committees focus on species that represent key classes within vertebrates, including mammals (Bruford et al. 2020; Smith et al. 2020; Blake et al. 2021), birds (Burt et al. 2009), fish (Howe et al. 2021), and amphibians (James-Zorn et al. 2015), and coordinate their efforts to ensure that approved gene names are assigned consistently across representative vertebrates. This standardized nomenclature is widely disseminated through all the major genomic resources and model organism databases. Notably, this approach takes into account genetic and evolutionary similarities in addition to function, exactly as proposed by Theofanopoulou et al. (Theofanopoulou et al. 2021). Gene nomenclature groups work closely with community experts, researchers, clinicians, bioinformaticians and biocurators to ensure that the approved gene names and symbols are informative, non-redundant and broadly applicable across diverse biological fields of study. One rationale cited for the newly proposed nomenclature system of Theofanopoulou et al is to create a universal nomenclature system that can be consistently used across vertebrates. We argue that such a system is already established by the existing vertebrate nomenclature authorities (Table 1).

Standardized nomenclature for oxytocin and arginine vasopressin. Oxytocin is a well-studied peptide hormone and neuropeptide with a large body of published scientific literature. The human gene name ‘oxytocin/neurophysin I prepropeptide’ with the symbol *OXT* (HGNC:8528) represents the full length protein which is post-translationally cleaved to produce oxytocin and neurophysin I, the oxytocin carrier protein (Brownstein, Russell, and Gainer 1980). An important feature of gene symbols is that they should be specific search terms. The “OT” symbol proposed by Theofanopoulou et al. returns over 12,000 PubMed results, many which are not related to oxytocin (or genes), making it a poor search term.

The arginine vasopressin gene (with the symbol *AVP,* HGNC:894) encodes a preprotein that is cleaved to form arginine vasopressin, neurophysin II and copeptin (Brownstein, Russell, and Gainer 1980; Land et al. 1982). Because of the action of these peptide hormones as antidiuretics and vasoconstrictors, this gene is well studied with the body of literature that has now settled on a common gene name for vertebrate orthologs. Theofanopoulou *et al* suggest that vasotocin is commonly used, however a PubMed search returns only 2,557 results for ‘vasotocin’, compared to 47,716 results for ‘vasopressin’. Furthermore, the approved name ‘arginine vasopressin’ refers to a highly conserved arginine in the AVP peptide product, which is present in the vast majority of sequenced vertebrates.

The existing approved OXT and AVP symbols are in use across vertebrates and are specific search terms. The high level of conservation of these two genes across vertebrates is already reflected in their consistent approved gene nomenclature, and changing this to the proposed two letter symbols (Table 1) would only result in confusion and hinder literature searches. A detailed discussion of the current and proposed nomenclature of vertebrate oxytocin and vasopressin receptor genes is included in the Supplementary Data.

Standardized nomenclature for oxytocin and arginine vasopressin receptors. Theofanopoulou *et al.* have confirmed the existence of six distinct clades of the oxytocin/vasopressin receptor family in vertebrates and proposed a novel nomenclature system for these clades. While we share their desire to ensure gene nomenclature reflects evolutionary relationships, we disagree that there is a need to revise all of the currently approved gene symbols to achieve this, as the existing approved nomenclature is already largely representing these relationships (Table 1). Instead, only minor updates are needed in some species to better reflect the orthology and paralogy between these genes. Gene symbol stability is especially important for genes that are linked to human health, and the oxytocin and vasopressin ligands and receptors all fall into this category, with hundreds of papers using the current approved nomenclature.

Table 1. Comparison of approved and proposed symbols for the oxytocin and vasopressin ligand and receptor genes. Newly approved symbols are indicated with *.

Approved symbol from joint nomenclature committees	**Theofanopoulou *et al.* proposed symbol**
OXT	OT
AVP	VT
OXTR	OTR
AVPR1A	VTR1A
AVPR1B	VTR1B
AVPR2 (aliased as AVPR2A)*	VTR2C

*AVPR2B**	VTR2B
AVPR2C / AVPR2L*	VTR2A

We also disagree with the stated order of gene divergence presented by Theofanopoulou *et al* for the AVPR2 clade. The phylogenies presented show that the AVPR2 gene first diverged from the common ancestor of the AVPR2C and AVPR2B genes prior to the duplication that gave rise to the AVPR2C and AVPR2B clades. This suggests that, despite its absence in sharks, AVPR2 may have been present in the common ancestor of vertebrates and was subsequently lost in some lineages, including sharks, conflicting with the conclusions reached by Theofanopoulou *et al*.

We will retain the current mammalian symbol for *AVPR2* and transfer this symbol to its orthologs. We therefore propose to use the same root symbol for its paralogs, appending the letters B and C, as shown in Table 1. We will also retain the current approved symbol for *OXTR*. A detailed discussion of the current and proposed nomenclature of vertebrate oxytocin and vasopressin receptor genes is included in Supplementary Data.

Another difference between our analysis and that of Theofanopoulou *et al.* is our finding that teleost *avpr2l* genes (current nomenclature) are not clearly orthologous to *AVPR2C* genes in other taxa, despite their partial shared synteny. Reciprocal BLAST searches and phylogenetic analysis (Figure S1) do not group *avpr2l* with *AVPR2C* genes. Theofanopoulou *et al.* state "our phylogenetic sequence analyses revealed tree topologies with almost 1:1 consistency to our synteny-defined relationships (Fig. 4)". Interestingly, Fig. 4a in Theofanopoulou *et al.* does not include the zebrafish (and other teleost) *avpr2l* genes, although it does include all other zebrafish oxtr* and avpr* genes. Due to this uncertainty about the lineage of *avpr2l*, we will leave this nomenclature unchanged in zebrafish.

The importance of applying standardized gene nomenclature in scientific journals.
Requiring scientists to consistently use approved nomenclature avoids confusion and supports search indexing. While an increasing number of scientific journals mandate the use of standardized gene nomenclature, this requirement is not always clearly stated or strictly enforced for authors – at least quoting the standardized gene symbol and the associated gene ID should be compulsory in all journals. Nature’s instructions to authors states that authors can "use their preferred terminology" for genes and proteins, which enables authors to publish novel nomenclature without first checking with the relevant nomenclature authority. If all journals, and especially influential ones such as Nature, would insist authors consult with nomenclature committees when suggesting updates much confusion could potentially be avoided. Unequivocally communicating about genes facilitates research and development in all biological and clinical fields.

Our analysis of the study of Theofanopoulou *et al.* demonstrates how the integration of genomic data from a broader range of species can help us to update and improve an already established nomenclature with only minor modifications. We assert that the changes suggested by Theofanopoulou *et al.* to the official vertebrate gene nomenclature would cause considerable confusion with little perceivable benefit.

Competing Interests Statement

The authors declare no competing interests.

Authors' Contributions

Study conceptualization: EAB

Analysis of gene nomenclature and evolutionary relationships: TEMJ & EAB

Manuscript writing and reviewing: Initial writing by FMM, TEMJ & EAB with input from all authors

Acknowledgements

T.E.M.J and E.A.B. are currently funded by the National Human Genome Research Institute (NHGRI) grant U24HG003345 and Wellcome Trust grant 208349/Z/17/Z. F.M.M. is supported by the US Department of Agriculture National Institute of Food and Agriculture projects ARZT-1370570-R50-131 and ARZT-1370520-R50-130. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health nor any of the funders.

Supplementary data file

Blake, Judith A., Richard Baldarelli, James A. Kadin, Joel E. Richardson, Cynthia L. Smith, Carol J. Bult, and Mouse Genome Database Group. 2021. “Mouse Genome Database (MGD): Knowledgebase for Mouse-Human Comparative Biology.” *Nucleic Acids Research* 49 (D1): D981–87.

Brownstein, M. J., J. T. Russell, and H. Gainer. 1980. “Synthesis, Transport, and Release of Posterior Pituitary Hormones.” *Science* 207 (4429): 373–78.

Bruford, Elspeth A., Bryony Braschi, Paul Denny, Tamsin E. M. Jones, Ruth L. Seal, and Susan Tweedie. 2020. “Guidelines for Human Gene Nomenclature.” *Nature Genetics* 52 (8): 754–58.

Burt, David W., Wilfrid Carrë, Mark Fell, Andy S. Law, Parker B. Antin, Donna R. Maglott, Janet A. Weber, Carl J. Schmidt, Shane C. Burgess, and Fiona M. McCarthy. 2009. “The Chicken Gene Nomenclature Committee Report.” *BMC Genomics* 10 Suppl 2 (July): S5.

Howe, Douglas G., Sridhar Ramachandran, Yvonne M. Bradford, David Fashena, Sabrina Toro, Anne Eagle, Ken Frazer, et al. 2021. “The Zebrafish Information Network: Major Gene Page and Home Page Updates.” *Nucleic Acids Research* 49 (D1): D1058–64.

James-Zorn, Christina, Virgillio G. Ponferrada, Kevin A. Burns, Joshua D. Fortriede, Vaneet S. Lotay, Yu Liu, J. Brad Karpinka, Kamran Karimi, Aaron M. Zorn, and Peter D. Vize. 2015. “Xenbase: Core Features, Data Acquisition, and Data Processing.” *Genesis* 53 (8): 486–97.

Land, H., G. Schütz, H. Schmale, and D. Richter. 1982. “Nucleotide Sequence of Cloned cDNA Encoding Bovine Arginine Vasopressin-Neurophysin II Precursor.” *Nature* 295 (5847): 299–303.

Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, et al. 2018. “Earth BioGenome Project: Sequencing Life for the Future of Life.” *Proceedings of the National Academy of Sciences of the*

doi:10.20944/preprints202109.0485.v1

United States of America 115 (17): 4325–33.

Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. “Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species.” *Nature* 592 (7856): 737–46.

Smith, Jennifer R., G. Thomas Hayman, Shur-Jen Wang, Stanley J. F. Laulederkind, Matthew J. Hoffman, Mary L. Kaldunski, Monika Tutaj, et al. 2020. “The Year of the Rat: The Rat Genome Database at 20: A Multi-Species Knowledgebase and Analysis Platform.” *Nucleic Acids Research* 48 (D1): D731–42.

Theofanopoulou, Constantina, Gregory Gedman, James A. Cahill, Cedric Boeckx, and Erich D. Jarvis. 2021. “Universal Nomenclature for Oxytocin-Vasotocin Ligand and Receptor Families.” *Nature* 592 (7856): 747–55.