

# The case for standardizing gene nomenclature in vertebrates


<https://doi.org/10.1038/s41586-022-05633-w>

Received: 14 September 2021

Accepted: 6 December 2022

Published online: 15 February 2023

Open access

 Check for updates

Fiona M. McCarthy<sup>1,8</sup>, Tamsin E. M. Jones<sup>2,8</sup>, Anne E. Kwitek<sup>3</sup>, Cynthia L. Smith<sup>4</sup>, Peter D. Vize<sup>5</sup>, Monte Westerfield<sup>6</sup> & Elspeth A. Bruford<sup>2,7</sup>✉

ARISING FROM C. Theofanopoulou et al. *Nature* <https://doi.org/10.1038/s41586-020-03040-7> (2021)

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 applied across species. Theofanopoulou et al.<sup>1</sup> propose a new nomenclature for the genes that encode oxytocin and arginine vasopressin and their receptors. Rather than changing to a different nomenclature, we suggest minor updates to the current approved nomenclature of these vertebrate genes to better reflect their evolutionary history. We call on authors, journal editors and reviewers to help support communication and indexing of gene-related publications by working with existing gene nomenclature committees and ensuring that standardized gene nomenclature is routinely used.

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; aliases (or synonyms) used in published literature are also often recorded to facilitate disambiguation, indexing and text mining. In vertebrates, gene nomenclature committees focus on species that represent key groups, including mammals<sup>2–4</sup>, birds<sup>5</sup>, fish<sup>6</sup> and amphibians<sup>7</sup>, and coordinate their efforts to ensure that approved gene names are assigned consistently across species. Historically, mammalian and avian gene symbols are in upper case, with the exception of rodents, which have gene symbols in title case; by contrast, *Xenopus* and zebrafish gene symbols are in lower case. These case conventions were originally established to help researchers distinguish references to genes in different model organisms<sup>8</sup>, and the standardized nomenclature is widely disseminated through all of 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.<sup>1</sup>, and many genes are named on the basis of their homologues in yeast, flies and other non-vertebrates. Gene nomenclature groups work closely with community experts<sup>9</sup>, 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 of Theofanopoulou et al. is to create a universal nomenclature system that can be consistently used across vertebrates. However, such a system is already established by the existing vertebrate nomenclature authorities.

Theofanopoulou et al. propose a new nomenclature for the genes that encode oxytocin, arginine vasopressin and their receptors on the basis of their evolutionary analysis of these genes in the context of newly sequenced, high-quality genomes generated by the Vertebrate Genomes Project (VGP)<sup>10</sup>. Although we share their desire to ensure that gene nomenclature reflects evolutionary relationships, we believe that the existing approved nomenclature, first established in vertebrates 30 years ago, already largely represents these relationships (Table 1). Instead, only minor updates are needed in some species to better reflect the orthology and paralogy between these genes (Supplementary Information). We consider many factors when making nomenclature decisions: structure and function of genes and gene products; evolutionary history (including consideration of gene synteny); current and historical nomenclature usage; utility of nomenclature as search terms (including avoiding symbol clashes with other genes across the tree of life); levels of support for nomenclature updates in the research community; and concordance with nomenclature guidelines in several model systems (see Supplementary Information). In addition, the current remit of the HUGO Gene Nomenclature Committee (HGNC) includes a commitment to move towards gene-symbol stability in humans<sup>2</sup>—especially for genes that are clinically relevant, which include the genes encoding oxytocin and arginine vasopressin and their receptors. Confusion about gene nomenclature in the medical literature could have serious negative consequences for patient safety<sup>11</sup>.

Major revisions to approved nomenclature are considered when the benefits clearly outweigh the downsides. Benefits can include the correction of incorrect or misleading gene nomenclature, better representation of evolutionary relationships, standardizing nomenclature throughout a gene family and providing nomenclature that can be used across all vertebrate species. Theofanopoulou et al. argue that the nomenclature of the oxytocin and arginine vasopressin genes and their receptors merits an update for all of these benefits. We believe that the existing approved nomenclature does not merit major revision as it is widely used, is not incorrect or misleading in the vast majority of vertebrate species, largely represents evolutionary relationships (with only minor additions needed to represent subclades in the AVPR2 subfamily) and has long been standardized across species (see Supplementary Information). The drawback is the introduction of additional identifiers in databases and the literature, which increases the risk of confusion to researchers and readers. Unfortunately, the potential for confusion has already been exemplified in a recent paper by Ocampo Daza et al.<sup>12</sup>, who disagreed with Theofanopoulou et al.'s assignment

<sup>1</sup>The Chicken Gene Nomenclature Committee (CGNC), School of Animal and Comparative Biomedical Sciences, University of Arizona, Tucson, AZ, USA. <sup>2</sup>HUGO Gene Nomenclature Committee (HGNC), European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK. <sup>3</sup>Rat Genome Database, Medical College of Wisconsin, Milwaukee, WI, USA. <sup>4</sup>Mouse Genome Database, The Jackson Laboratory, Bar Harbor, ME, USA. <sup>5</sup>Xenbase, Departments of Biological Sciences and Computer Science, University of Calgary, Calgary, Alberta, Canada. <sup>6</sup>ZFIN, Institute of Neuroscience, University of Oregon, Eugene, OR, USA. <sup>7</sup>Department of Haematology, University of Cambridge School of Clinical Medicine, Cambridge, UK. <sup>8</sup>These authors contributed equally: Fiona M. McCarthy, Tamsin E.M. Jones. <sup>✉</sup>e-mail: [elspeth@genenames.org](mailto:elspeth@genenames.org)

**Table 1 | Comparison of approved and proposed symbols for the oxytocin and arginine vasopressin ligand and receptor genes**

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* or AVPR2L	VTR2A

Newly approved symbols are indicated with \*.

of ABC suffixes in the AVPR2/VTR2 subfamily and therefore used the same symbols to refer to different genes.

Theofanopoulou et al. argue that their study acts as a model for the revision of gene nomenclature in the context of large-scale vertebrate-sequencing projects, including the VGP. Their stated intent is to completely revise vertebrate gene nomenclature, including that of human genes, to fully reflect evolutionary histories that are revealed by large-scale sequencing projects. We are concerned that the authors may not fully appreciate the level of disruption that would be caused by major revisions to gene nomenclature on this scale. It is worth noting that the gene family analysed in their study is relatively simple with regard to its evolutionary history, and to perform such an analysis for every vertebrate gene family is an inconceivably large task. Given that over 40 years and millions of dollars of public funding have been invested in the current standardized nomenclature projects, we propose that an overhaul of the entire system would not be a prudent use of the limited resources we have in genomics.

Requiring scientists to consistently use approved nomenclature avoids confusion and supports search indexing. Although 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, and citing the approved gene symbol and its associated gene ID should be compulsory in all journals. The instructions provided by *Nature* to authors state that authors can “use their preferred terminology” for genes and proteins, which enables authors to publish novel nomenclature without first checking with the relevant authority. If all journals—and especially influential ones such as *Nature*—were to insist that authors consult with nomenclature committees when suggesting updates, much confusion could potentially be avoided. Unequivocally communicating about genes will facilitate research and development in all biological and clinical fields.

We assert that the changes suggested by Theofanopoulou et al. to the official vertebrate gene nomenclature would cause considerable confusion with little perceivable benefit.

Our analysis of their study (Supplementary Information) 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. Theofanopoulou et al. call for collaboration between the gene nomenclature committees

and genomic initiatives, which we have always wholeheartedly supported. We continue to encourage researchers and communities to collaborate with the gene nomenclature committees when proposing nomenclature updates.

## Reporting summary

Further information on experimental design is available in the Nature Portfolio Reporting Summary linked to this article.

1. Theofanopoulou, C., Gedman, G., Cahill, J. A., Boeckx, C. & Jarvis, E. D. Universal nomenclature for oxytocin–vasotocin ligand and receptor families. *Nature* **592**, 747–755 (2021).
2. Bruford, E. A. et al. Guidelines for human gene nomenclature. *Nat. Genet.* **52**, 754–758 (2020).
3. Smith, J. R. et al. The year of the rat: the rat genome database at 20: a multi-species knowledgebase and analysis platform. *Nucleic Acids Res.* **48**, D731–D742 (2020).
4. Blake, J. A. et al. Mouse Genome Database (MGD): knowledgebase for mouse–human comparative biology. *Nucleic Acids Res.* **49**, D981–D987 (2021).
5. Burt, D. W. et al. The chicken gene nomenclature committee report. *BMC Genomics* **10**, S5 (2009).
6. Howe, D. G. et al. The Zebrafish Information Network: major gene page and home page updates. *Nucleic Acids Res.* **49**, D1058–D1064 (2021).
7. James-Zorn, C. et al. Xenbase: core features, data acquisition, and data processing. *Genesis* **53**, 486–497 (2015).
8. Bruford, E. A. Highlights of the ‘Gene Nomenclature Across Species’ meeting. *Hum. Genomics* **4**, 213–217 (2010).
9. Olender, T., Jones, T. E. M., Bruford, E. A. & Lancet, D. A unified nomenclature for vertebrate olfactory receptors. *BMC Evol. Biol.* **20**, 42 (2020).
10. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
11. Braschi, B., Seal, R. L., Tweedie, S., Jones, T. E. M. & Bruford, E. A. The risks of using unapproved gene symbols. *Am. J. Hum. Genet.* **108**, 1813–1816 (2021).
12. Ocampo Daza, D., Bergqvist, C. A. & Larhammar, D. The evolution of oxytocin and vasotocin receptor genes in jawed vertebrates: a clear case for gene duplications through ancestral whole-genome duplications. *Front. Endocrinol.* **12**, 792644 (2021).

**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. A.E.K. is supported by National Institutes of Health (NIH) grants from the NHGRI (U24HG010859) and the National Heart Lung and Blood Institute (NHLBI; RO1HL064541). C.L.S. is supported by NHGRI grant U24HG000330. P.D.V. is supported by NIH grants P41 HD095831 and P41 HD064556. M.W. is supported by NHGRI grants U24HG002659 and U24HG010859. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or any of the funders.

**Author contributions** Study conceptualization: E.A.B. Analysis of gene nomenclature and evolutionary relationships: T.E.M.J. and E.A.B. Manuscript writing and reviewing: initial writing by F.M.M., T.E.M.J. and E.A.B., with input from all authors.

**Competing interests** The authors declare no competing interests.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41586-022-05633-w>.

**Correspondence and requests for materials** should be addressed to Elspeth A. Bruford.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |     |           |
|-----|-----------|
| n/a | Confirmed |
|-----|-----------|
- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
  - A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
  - The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
  - A description of all covariates tested
  - A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
  - A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
  - For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
  - For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
  - For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
  - Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All data analysed in this study are referenced by unique database IDs in the Supplementary Data and are publicly available in the resources <https://www.genenames.org> (HGNC), <https://vertebrate.genenames.org> (VGNC), <http://www.informatics.jax.org> (MGI), <https://rdg.mcw.edu> (RGD), <http://birdgenenames.org> (CGNC), <http://www.xenbase.org> (XenBase), <https://zfin.org> (ZFIN) and <https://www.ncbi.nlm.nih.gov/> (NCBI).

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<input type="text" value="we were not analyzing a sample set"/>
Data exclusions	<input type="text" value="there was no data exclusion"/>
Replication	<input type="text" value="for the maximum likelihood phylogeny we used UltraFast Bootstrap 4 with 1000 replicates, values &lt;95 are not well supported"/>
Randomization	<input type="text" value="we were not analyzing a sample set so this is not applicable"/>
Blinding	<input type="text" value="we were not analyzing a sample set so this is not applicable"/>

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

# Reply to: The case for standardizing gene nomenclature in vertebrates

<https://doi.org/10.1038/s41586-022-05634-9>

Constantina Theofanopoulou<sup>1,2</sup>✉ & Erich D. Jarvis<sup>1,2,3</sup>✉

Published online: 15 February 2023

REPLYING TO F. M. McCarthy et al. *Nature* <https://doi.org/10.1038/s41586-022-05633-w> (2023)

Open access

 Check for updates

Here we reply to points raised by McCarthy et al. in the accompanying Comment<sup>1</sup> concerning our proposal<sup>2</sup> for an evolution-based and universal vertebrate nomenclature for the oxytocin and vasotocin ligand and receptor families, and the principles considered for homology-based gene nomenclatures. We strengthen our claims with additional evidence and propose evidence-based criteria for homologous gene nomenclature, in the following order of reliability: synteny, phylogenetic inference, sequence identity and gene function. We believe that the time is ripe for gene nomenclature committees and initiatives generating high-quality assemblies to join forces in a universal gene nomenclature committee.

Our proposed universal gene nomenclature (that is, naming) for the oxytocin and vasotocin ligands and receptors<sup>2</sup> was based on several criteria, including gene synteny, phylogeny, identity and function, and provides a case study that is applicable across gene families. McCarthy et al.<sup>1</sup> argue that a standardized system of nomenclature already exists, “first established in vertebrates 30 years ago”, and that only minor changes are needed in this gene family, with a focus on tradition, name stability, phylogeny, identity and gene function, and with the order of priority of evidence determined on a case-by-case basis. We disagree with both of these claims, especially because determining gene orthology was not fully possible until the recent availability of high-quality genomes. Below, we discuss the principles that we suggest should be applied across gene families and future initiatives. In the Supplementary Information, we respond to the gene-specific claims made by McCarthy et al.<sup>1</sup>

In our study<sup>2</sup>, for each of the oxytocin and vasotocin ligands and receptors, we listed two to six commonly used aliases (Table 1 in Theofanopoulou et al.<sup>2</sup>). Many of these reflect incorrect orthologies or paralogies, indicating that there was not a universally used standard before our study, nor one that sufficiently portrayed gene orthology. We view the vertebrate-wide gene nomenclature that McCarthy et al.<sup>1</sup> present as “approved” in their Table 1 as newly proposed. They adopted the most common gene names for mammals, revised some on the basis of our study and others, and applied them to all other vertebrates where possible (Supplementary Note 1). None of the other aliases were listed, which makes the translation of findings across species and the literature difficult. Furthermore, in their newly proposed nomenclature, tradition overrides orthology and paralogy. For example, they maintain very different names for the genes oxytocin and vasotocin that do not echo their paralogy (that is, oxytocin and arginine vasopressin); and for species that do not have the arginine amino acid, they change the name to another alias (vasopressin), but still abbreviate it to *AVP*. We think that allowing tradition and stability to override naming rules of orthology and paralogy could lead to confusion.

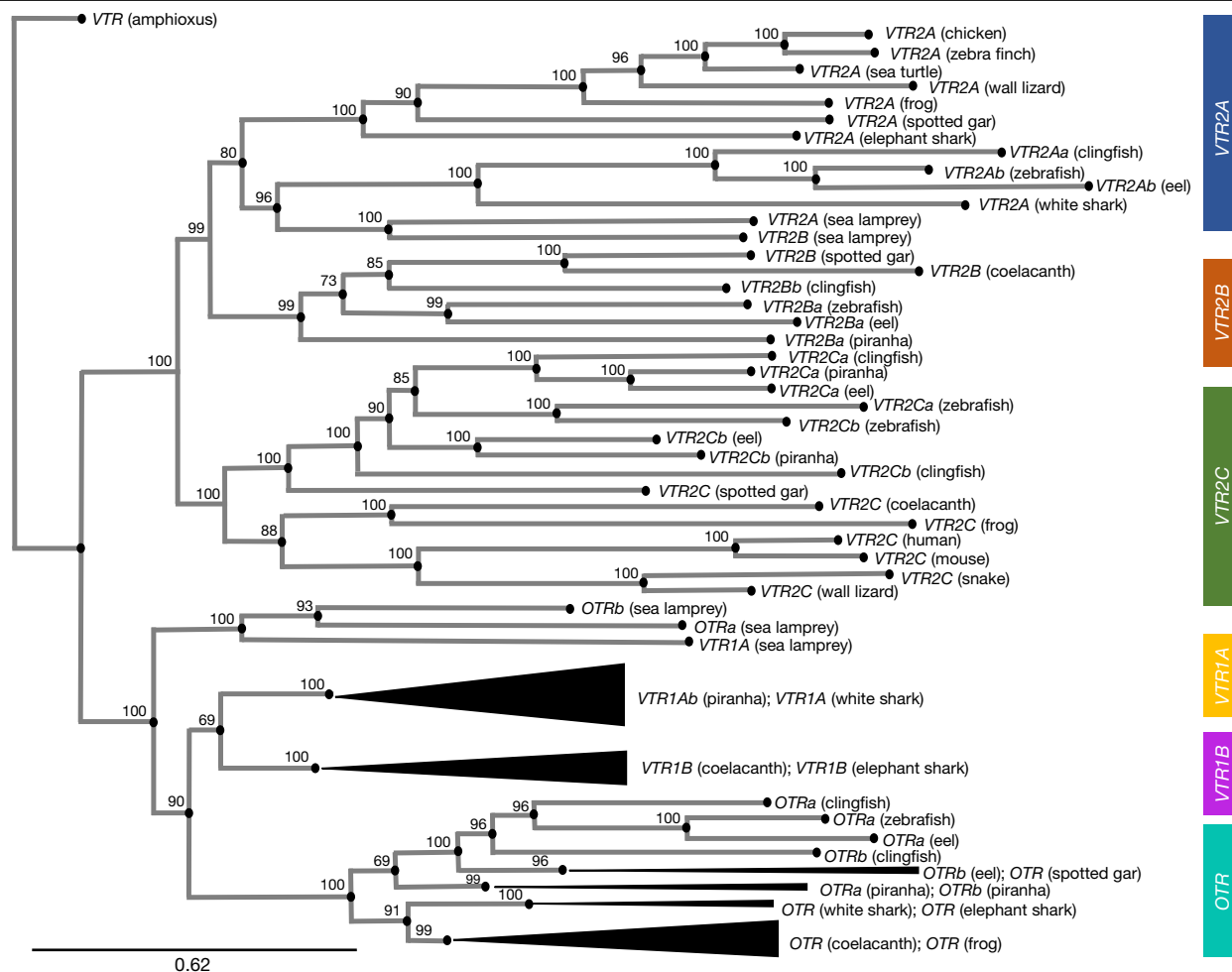
However, we believe it is possible to consider both tradition and orthology/paralogy. For example, because vasotocin is the evolutionarily older gene, with oxytocin resulting from a local duplication of it<sup>2</sup>, if we were strict with evolutionary naming, we would have renamed vasotocin to ‘vasotocin 1’ and oxytocin to ‘vasotocin 2’. But to conserve some continuity with traditional use, we proposed the already used ‘vasotocin’ for vasopressin, to mirror the ending of ‘oxytocin’. In forming this proposal, we consulted with experts, whom we acknowledged<sup>2</sup>, and with the leaders of the Ensembl annotation team.

Valuing accuracy over tradition comes with some downsides. Perhaps the greatest would be the effort required to ensure continuity between previous publications and annotated genomes with the new nomenclature. To mitigate this, we suggest a translation table from old to revised gene names (for example, Table 1 in Theofanopoulou et al.<sup>2</sup>), which would be available in platforms like the National Center for Biotechnology Information. Current committees already use such tables, but their practices of establishing nomenclature changes are either different than the ones we propose or not consistent with each other (Supplementary Notes 2–4).

McCarthy et al.<sup>1</sup> also criticize our proposed two-letter symbols for oxytocin and vasotocin (*OT* and *VT*), in that they give broader results in a literature search compared to three-letter symbols (such as *OXT* and *AVP*). We agree and further argue that three-letter symbols could still reflect an evolution-based nomenclature; for example, *OTC* (oxytocin) and *VTC* (vasotocin). We also suggest that gene-symbol consistency across species should be adopted in their letter capitalization. The landscape at present, in which only some mammalian and avian gene symbols are upper case, mouse and rat symbols are lower case except for an initial upper-case letter and amphibian and fish species are all lower case, does not depict the real orthology of these genes, and perpetuates anthropocentric practices. In our universal nomenclature proposal<sup>2</sup>, we suggest that gene symbols should be upper case across species.

We agree with McCarthy et al.<sup>1</sup> that for name revisions, the benefits should outweigh the risks. We are guided by the belief that “names have a powerful influence on the experiments we do and the way in which we think”<sup>3</sup>, and hence it is important that names do not give rise to false expectations. For example, the binding of oxytocin to the ‘vasopressin’ receptors has been often identified as surprising—something that could be avoided with names that reflect their common origin (-tocin). This knowledge will also be useful for medicine, so that physicians are more aware of drug interactions between the two receptor families. Similarly, in other gene families, McCarthy et al.<sup>1</sup> endorse a nomenclature that differs in orthologous genes with a different function across species. For example, the *CSAD* gene is named ‘cysteine sulfinic acid decarboxylase’ in all species except chickens, in which it is called ‘cysteine acid decarboxylase’. If sequence and/or function changes were routinely

<sup>1</sup>Laboratory of Neurogenetics of Language, Rockefeller University, New York, NY, USA. <sup>2</sup>Hunter College, City University of New York, New York, NY, USA. <sup>3</sup>Howard Hughes Medical Institute, Chevy Chase, MD, USA. ✉e-mail: [ktheofanop@rockefeller.edu](mailto:ktheofanop@rockefeller.edu); [ejarvis@mail.rockefeller.edu](mailto:ejarvis@mail.rockefeller.edu)



**Fig. 1 | Family tree for genes that encode the oxytocin and vasopressin receptors.** Tree topology inferred with the phylogenetic maximum likelihood method on an exon nucleotide alignment (MAFFT), with 1,000 non-parametric bootstrap replicates (IQ-TREE). Bootstrap values are shown as percentages at the branch points. The tree is rooted with the *VTR* gene in amphioxus. The gene names of the current accessions (see Table 1 in Theofanopoulou et al.<sup>2</sup> and

Supplementary Tables 4a–e in Theofanopoulou et al.<sup>2</sup> for a full list of synonyms) were written over according to our revised synteny- and phylogeny-based orthology. All sequences used, FASTA alignment and Newick tree files can be accessed here at <https://github.com/constantinatheo/universalnomenclature/>. Scale bar, 0.62 substitutions. For a discussion on interchanging *VTR2A* and *VTR2C* naming, see Supplementary Note 3.

used to change gene names, then nearly all orthologous genes would have different names across species.

McCarthy et al.<sup>1</sup> decided not to suggest blanket ‘rules’ about which factors should be weighed more heavily than others, as each case will differ depending on the context. In our experience, not all evidence has equal weight. For example, McCarthy et al.<sup>1</sup> did not accept our nomenclature in part due to the lack of sequence-identity resolution (Basic Local Alignment Search Tool (BLAST) analyses). However, sequence-identity percentages do not always provide a solid basis for gene nomenclature, because orthologous syntenic genes can misleadingly have higher sequence identity with a paralogous gene (Supplementary Table 12 in Theofanopoulou et al.<sup>2</sup>). In addition, McCarthy et al.<sup>1</sup> presented an amino acid phylogeny as not being conclusive enough for some of our interpretations. However, we showed<sup>2</sup> that amino acid phylogenies have low bootstrap support on some branches, whereas exonic nucleotide phylogenies yielded a higher resolution that supports our conclusions (Fig. 4 in Theofanopoulou et al.<sup>2</sup>). With more high-quality genome assemblies generated by the Vertebrate Genomes Project (VGP) since our original publication, we ran a new exonic phylogeny that even more strongly supports our conclusions (Fig. 1 and Supplementary Notes 3 and 4).

We find<sup>2,4</sup> that synteny-based approaches in most cases give the best resolution for gene orthologies and paralogies, and hence for gene nomenclature. Wherever available, we propose using

chromosome-scale genomes that are highly contiguous and have a high base-call accuracy<sup>2</sup>. When synteny is not clear, we suggest that priority is given to nucleotide phylogenetic inference with the same prerequisites for genome quality. In Extended Data Fig. 1 and Supplementary Note 5, we provide specific suggestions and caveats with regard to our recommended practices for synteny and phylogenetic analyses. We propose that a combination of synteny and highly supported phylogeny is the backbone of a universal gene nomenclature.

According to the guidelines for human gene nomenclature<sup>5</sup>, initiatives that aim to revise a nomenclature when the old one is “misleading...are welcomed”. We agree with this practice. However, we believe that the process that is used to approve those revisions should take a different approach to the ones proposed by McCarthy et al.<sup>1</sup> We do not think that journal editors should require “scientists to consistently use approved nomenclature”<sup>1</sup> by a limited committee. Rather, we believe that they should allow new uses in the light of new evidence (see checklist in Extended Data Fig. 1).

Moreover, the current nomenclature committees represent nomenclature focused on only 0.01% of the 70,000 extant vertebrates, with genome assemblies that were much more fragmented, and with traditions that we think need reconsideration. Although several authors of the accompanying Comment by McCarthy et al.<sup>1</sup> are part of a recently formed Vertebrate Gene Nomenclature Committee (VGNC), in their database (<https://vertebrate.genenames.org/>) at the time of

writing (19 November 2022) there is no inclusion of gene aliases used in the literature (versus Table 1 in Theofanopoulou et al.<sup>2</sup>).

The high-quality genomes generated by the VGP (<https://vertebratengenomesproject.org/>) and related initiatives such as the Earth BioGenome Project (<https://www.earthbiogenome.org/>)<sup>6,7</sup> are greatly improving the identification of gene orthology and thereby gene annotation, bringing about an opportunity to establish a universal nomenclature for most genes. Our experience in these initiatives is that existing gene annotation and nomenclature bodies are not yet coordinated or consistent in their approaches. We envisage a universal gene nomenclature committee that involves scientists working on sequencing, assembly, annotation, phylogeny and genome evolution, as well as on the respective lineages and genes for all life.

One possible organizing principle would be to create one committee per major lineage (for example, cyclostomes), group these as subcommittees under one larger committee (for example, all vertebrate species), group all of them under a committee for all species of one of the animal kingdoms (for example, eukaryotic species) and then group all of them under all life. We believe that such an effort would be likely to require changes both to infrastructure (for example, committees and publication policies) and to the way systems operate (for example, high-quality genomes, synteny and phylogenetics).

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

All of the data used can be found in the Supplementary Notes and in the following repository: <https://github.com/constantintheo/universalnomenclature>.

## Code availability

All the code used in this study can be found in the following repository: <https://github.com/constantintheo/universalnomenclature>.

1. McCarthy, F. M. et al. The case for standardizing gene nomenclature in vertebrates. *Nature* <https://doi.org/10.1038/s41586-022-05633-w> (2022).

2. Theofanopoulou, C., Gedman, G., Cahill, J. A., Boeckx, C. & Jarvis, E. D. Universal nomenclature for oxytocin–vasotocin ligand and receptor families. *Nature* **592**, 747–755 (2021).
3. Jarvis, E. D. et al. Avian brains and a new understanding of vertebrate brain evolution. *Nat. Rev. Neurosci.* **6**, 151–159 (2005).
4. Theofanopoulou, C. Reconstructing the evolutionary history of the oxytocin and vasotocin receptor gene family: Insights on whole genome duplication scenarios. *Dev. Biol.* **479**, 99–106 (2021).
5. Wain, H. M. et al. Guidelines for human gene nomenclature. *Genomics* **79**, 464–470 (2002).
6. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
7. Lewin, H. A. et al. Earth BioGenome Project: sequencing life for the future of life. *Proc. Natl Acad. Sci. USA* **115**, 4325–4333 (2018).

**Acknowledgements** We thank all the members, and especially the genome annotation committees, of the Vertebrate Genomes Project, the Earth BioGenome Project and the European Reference Genome Atlas for helping us to appreciate, over the past years, the complexity of the issues we have touched on here. C.T. was supported by funds from the Rockefeller University and E.D.J. by funds from the Howard Hughes Medical Institute and the Rockefeller University.

**Author contributions** For this Matters Arising Reply, the corresponding authors of the original study<sup>2</sup>, C.T. and E.D.J., conceived the ideas for the response. C.T. performed the additional BLAST/BLAT, synteny and phylogenetic analyses and wrote the first draft; E.D.J. supervised the study and helped write the manuscript. The three other co-authors of the original study<sup>2</sup> (G. Gedman, J. A. Cahill and C. Boeckx) did not participate in this Reply, owing to the challenge being on the specific concepts and work that the corresponding authors had developed, but all continue to support the original study.

**Competing interests** The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41586-022-05634-9>.

**Correspondence and requests for materials** should be addressed to Constantina Theofanopoulou or Erich D. Jarvis.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

## Matters arising

### SYNTENY ANALYSES

- High-quality, Chromosomal-level genomes
- Species from all orders of a lineage or class
- Examination of synteny in different gene-windows (e.g., 10, 50, 100) and chromosomal windows
- Caveats: consideration of divergence time when assessing robustness of synteny conservation or when setting thresholds

### PHYLOGENY ANALYSES

- High-quality, Chromosomal-level genomes
- Species from all orders of a lineage or class
- At least two different inferences (exonic sequences, protein coding sequences, or full-length sequences (coding and non-coding sequences))
- Trees rooted with an outgroup-taxa sequence
- Caveats: bootstrap cut-off values, trees based on long vs. short sequences etc.

**Extended Data Fig. 1 | Checklist and caveats.** Suggested checklist and caveats to be considered for synteny (top) and phylogeny (bottom) evidence used to propose gene nomenclature.



## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

We collected our data using gene sequences found in the NCBI (<https://www.ncbi.nlm.nih.gov/>) and Ensembl (release 105; <https://useast.ensembl.org/index.html>). Genome assembly IDs and GenBank assembly accession numbers can be found in Supplementary Table 1.

Data analysis

Our exonic sequences were aligned with MAFFT (v 7) (<https://mafft.cbrc.jp/8/>; default parameters); from this alignment, we generated a Phylogenetic Maximum Likelihood tree using IQTree WebServer (1000 replicates) (v2.2.0), which we visualized via <https://phylo.io/>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We used 17 vertebrate and 1 invertebrate species' genomes, whose IDs and GenBank assembly accession numbers can be found in Supplementary Table 1. All the NCBI/Ensembl/Gene IDs of the genomes and genes we included in the phylogenetic tree can be found in the Suppl.Tables\_TheofanopoulouMattersArising excel document. All the exonic gene sequences, alignment and Newick tree files used for the phylogenetic tree can be found here: <https://github.com/constantinatheo/universalnomenclature>.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences  Behavioural & social sciences  Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the exonic phylogeny we used the longest read-sequences available from species representing 10 vertebrate lineages (1 cyclostome: sea lamprey; 2 sharks: elephant shark and white shark, 1 coelacanth: coelacanth; 1 holost fish: spotted gar; 4 teleost fishes: zebrafish, red bellied piranha, electric eel, and blunt-snouted clingfish; 2 squamata: common wall lizard and Western terrestrial garter snake; 1 turtle: green sea turtle; 1 frog: tropical clawed frog; 2 birds: zebra finch and chicken; and 2 mammals: human and mouse) and 1 invertebrate (amphioxus). No sample size calculation was performed; sample size was determined sufficient in terms of quantity, since all major vertebrate lineages are represented, and quality, since we used high-quality genomes wherever available.
Data exclusions	We did not exclude any genomes of species that would have contributed further to the understanding of the evolution of the OTR-VTR receptors.
Replication	We generated a Phylogenetic Maximum Likelihood tree using IQTree WebServer (1000 replicates). All attempts for replication were successful.
Blinding	Our tests were blind in that we had not assigned specific names to the genes before our synteny analyses showed clearly which gene is orthologous to which (Theofanopoulou et al. 2021).
Randomization	Randomization was not relevant in this study, since it is impossible to randomize the gene sequences found in a species' genome.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging