Standardizing variant naming in literature with VariantValidator to increase diagnostic rates

Freeman, P.J.; Wagstaff, J.F.; Fokkema, I.F.A.C.; Cutting, G.R.; Rehm, H.L.; Davies, A.C.; ... ; Dalgleish, R.

https://doi.org/10.1038/s41588-024-01938-w

Standardizing variant naming in literature with VariantValidator to increase diagnostic rates

Peter J. Freeman, John F. Wagstaff, Ivo F. A. C. Fokkema, Garry R. Cutting, Heidi L. Rehm, Angela C. Davies, Johan T. den Dunnen, Liam J. Gretton & Raymond Dalgleish

Accurate naming of genetic variants is essential to identify clinical data that interpret the consequences of such variants. In partnership with the Human Genome Organization, we advocate for integration of VariantValidator in publishing of journals and databases, to improve the quality of shared genetic data and ultimately patient outcomes.

Rare diseases, as classified by the European Union, affect fewer than 1 in 2,000 individuals, but with over 8,000 rare genetic diseases recognized, they affect approximately 10% of all births globally[1]. Identification and curation of genomic variants are fundamental to diagnosis and clinical management of individuals. Databases, such as ClinVar[2] and Leiden Open Variation Database (LOVD)[3], offer insights into genetic variants. Clinical scientists rely on these resources to identify documented evidence presented in the literature and reach a diagnosis, but most variants are not accurately described according to the de facto naming standard developed by the Human Genome Variation Society (HGVS)[4,5] (https://hgvs-nomenclature.org/stable/). Sub-standard naming renders variants (and subsequently data associated with them) unidentifiable, creating a data-flow bottleneck from journal to database that is a contributing factor to slowing the diagnostic process, resulting in poor patient outcomes through missed diagnoses[6].

To address this issue, we created an open-source web-based user interface termed VariantValidator[7] (https://variantvalidator.org/) to assist users (researchers, students and trainers, clinicians and database curators who generate and utilize genetic data) in navigating the HGVS nomenclature. VariantValidator provides correctly formatted descriptions in the context of all relevant reference sequences (genome, transcript and protein), automatically projecting between genome builds GRCh37 and GRCh38. Additionally, VariantValidator automatically inter-converts between the HGVS format and genomic coordinate-based variant descriptions derived from (and adhering to the variant naming standards of) the variant call format (VCF; termed pseudo-VCF). Since 2018, VariantValidator has been used to standardize descriptions of genetic variants in clinical reports, literature and databases, and has been embedded into our clinical bioinformatics educational programs for healthcare scientists. VariantValidator is developed in GitHub and our live services are deployed on virtual LAMP (Linux, Apache, MySQL, programming language) servers hosted at the University of Leicester, UK, and LEMP (Linux, EnginX, MySQL, programming language) servers at the University of Manchester, UK.

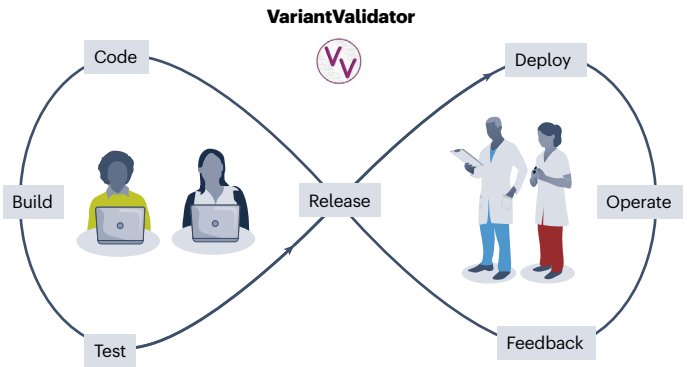

Fig. 1 | Community of practice-driven development. The success of the VariantValidator was driven by engagement with our community of users, following an agile development model. We provide platforms that are used in clinical, education and research practice by our users. These users test new releases, identify issues and consider improvements that would assist their professional practice. The users contact us via our dedicated support page (https://variantvalidator.org/help/contact/) or via GitHub issues (https://github.com/openvar/variantValidator/issues), and we involve them directly in the planning and acceptance of new resources or bug fixing methodology to ensure their exact needs are met. Therefore, VariantValidator keeps pace with the fast-moving discipline of genomic medicine.

Based on user feedback, we improved the functionality of VariantValidator, introducing a range of tools for validating variant descriptions with greater accuracy than any similar platform (Fig. 1). A key focus of user-driven iterative improvements is strict compliance with evolving HGVS nomenclature standards. For example, we increased support for additional HGVS formats, including RNA (r.) descriptions, which are not generally provided by other nomenclature validation tools (Table 1). VariantValidator is regularly updated to handle more complex HGVS nomenclature formats, and users can request the addition of new formats by contacting us directly or adding a feature request to our GitHub pages.

Responding to technological demands, we made VariantValidator compatible with integration into omics platforms. The core VariantValidator engine can be installed as a Python library (https://github.com/openvar/variantValidator), and we also developed a Python module termed VariantFormatter (https://github.com/openvar/variantFormatter), designed for direct integration into genomics workflows. VariantFormatter uses both custom and native VariantValidator functions to validate genomic variant descriptions (pseudo-VCF and HGVS) and map them to transcript (c.) and protein (p.) variant descriptions in the context of both RefSeq and Ensembl reference sequences. We also integrated Ensembl transcript reference sequences across our

Table 1 | Summary of key upgrades to VariantValidator functionality since 2018

Enhancement	Guidance	Example	Status
Integration of Ensembl data	https://hgvs-nomenclature.org/stable/background/refseq/ The *SHANK3* (HGNC:14294) gene MANE Select transcript has two additional exons not found in the GRCh37/GRCh38 reference sequences. This gives a positional discrepancy when mapping from genome to the MANE Select transcript (RefSeq) in comparison to the Ensembl canonical transcript (which is an exact match with the genomic reference sequences)	NC_000022.11:g.50720669A>C mapped to the MANE Select (RefSeq) and Ensembl Select transcripts NM_001372044.2:c.3061A>C NP_001358973.1:p.(Lys1021Gln) ENST00000445220.5:c.2815A>C ENSP00000489407.1:p.(Lys939Gln)	Deployed across all relevant services
Filter by transcript	Allows users to limit the outputs when submitting genomic (g.) or pseudo-VCF variant descriptions	Options (select_transcripts field) mane_select: MANE Select transcript mane: MANE Select and MANE Plus Clinical specific transcript ID(s) all: All transcripts at their latest version raw: All transcripts at all available versions	Deployed across all relevant services
Alignment gap aware projection[9]	Identifies variants aligned within or proximal to imperfect alignments between genome and transcript (additionally, see row 1) Aligned to GRCh37, the *NR2E3* (HGNC:7974) MANE Select transcript contains 1 base fewer than the genomic reference sequence NC_000015.9. During genomics analysis, we would expect to see NC_000015.9:g.72105933del owing to this error in the genomic reference sequence. This projects to NM_014249.4:c.951= when correctly accounted handled, not c.951+1del (which is not HGVS compliant), as returned by Mutalyzer[11]. Additionally, if the genomic variant is not seen, we can miss frame-shifting variation (see column 3)	VariantValidator NC_000015.9:g.72105928_72105929= Warning: NM_014249.4 contains 1 fewer bases between c.951_952 than NC_000015.9 Corrected output NM_014249.4:c.951dup NP_057430.1:p.(Thr318HisfsTer23) Mutalyzer NC_000015.9:g.72105928_72105929= No warning provided Incorrect projection to the transcript NM_016346.4:c.947_948= NP_057430.1:p.=	Deployed across all relevant services
RNA variant description (r.) formatting	https://hgvs-nomenclature.org/stable/recommendations/RNA/deletion/	Input NM_000089.4:r.1033_1035del Correct c. following the 3-prime rule NM_000089.4:c.1035_1035+2del NP_000080.2:p.? Correct r. following the 3-prime rule NM_000089.4:r.1034_1036del NP_000080.2:p.Val345del	Deployed in VariantValidator. Not applicable to VariantFormatter

entire tool set over the summer of 2024. For rapid deployment, we developed a representational state transfer (REST) application programming interface (API), allowing programmatic access to VariantValidator capabilities without the need for local installation. This is widely used in the UK National Health Service and across Europe and the USA. This API (https://rest.variantvalidator.org/; https://github.com/openvar/rest_variantValidator) interfaces with both VariantValidator and VariantFormatter, returning data in standardized formats. We also developed a specialized endpoint to support the LOVD[3] suite of variation databases, which can also be optimized for direct integration into genomics workflows. This highly customizable version of the VariantFormatter endpoint allows VariantValidator to replace Mutalyzer, the first software application for variant nomenclature validation, predating the deployment of VariantValidator, as the LOVD variant-description gateway tool. The gene2transcripts_v2 endpoint of VariantValidator returns all transcripts and alignment data for submitted Human Gene Nomenclature Committee (HGNC) gene symbols or geneIDs, enabling the creation of gene panel bed files. We equipped the VariantValidator web user interface with a simplified version of gene2transcripts_v2, allowing users to see which transcripts we support and to identify MANE (matched annotation from NCBI and EMBL-EBI) Select[8] and other MANE transcripts via an intuitive table.

We co-developed reference sequence guidelines with the Human Genome Organization (HUGO) HGVS Variant Nomenclature Committee (https://www.hugo-international.org/standards/) and reformatted the Universal Transcript Archive (UTA)[9] database that underpins VariantValidator to ensure strict adherence. For example, our version of UTA (VVTA) ensures reference sequence IDs are complete and correctly versioned and coding transcripts minimally comprise a complete coding sequence. We also improved handling of differing exon structures for single transcripts that may occur when the transcript is mapped to different genomic ALT assemblies and patches, as well as major genome builds. Additionally, our transcript alignments are faithfully derived from the official published versions provided by RefSeq and Ensembl.

In parallel to software deployment, changes in publishing standards with respect to the use of accurate and complete DNA variant naming are required. To this end, the VariantValidator team joined a Human Genome Organization Reporting of Sequence Variants (HUGO RSV) committee, working to encourage standards compliance in variant reporting. This committee, comprising editors, editorial staff and bioinformatics experts, focuses on the need to improve variant reporting in journals, and published guidance recommending that authors use validation software before publication[10]. Despite these efforts, ongoing research with the *Genetics in Medicine* journal shows that

>95% of submitted manuscripts need correction of variant descriptions before publication, and <2% of manuscripts contain the complete set of descriptions the HGVS nomenclature requires for comprehensive and accurate naming.

To ensure further adoption of HGVS standards in publications, the committee enjoined a global multi-organization working group led by the American College of Medical Genetics and Genomics (ACMG) to establish a professional practice standard related to reporting and sharing of interpreted genomic variation. To support the work of the committee and the professional standard, we developed a VariantValidator web-API that generates a structured dataset containing accurately formatted variant descriptions, which can be submitted as supplementary material in manuscripts. The dataset will be represented in a human-readable table as well as a computer-readable format such as the Java Script Object Notation (JSON) format, which will assist identification by machine learning algorithms. This allows authors to describe variation in formats recognizable within their profession, while ensuring that structured variant evidence in manuscripts is findable. Concurrently, we are working with the LOVD team who developed an HGVS syntax validator based on analysis of common mistakes users make when applying HGVS nomenclature. Their tool validates descriptions on the syntax level only, so can support a larger part of the HGVS nomenclature than sequence-level validators such as VariantValidator. We aim to integrate the LOVD syntax validator by the end of 2024, allowing recognition of additional common mistakes and suggestions of what the user most likely intended and assistance with corrections. Steered by the HUGO RSV committee and the professional-standards working group, this iterative practice-driven development strategy is being used to drive improvements in the quality of shared genetic data and ensure a positive impact in terms of improved patient outcomes. To achieve our ultimate goal, the HUGO committee and professional-standard working group are advocating that publishing groups provide direct interfacing of our web-API within editorial management systems, establishing direct integration of the capabilities of VariantValidator. Such interaction will enforce accurate variant descriptions and standardized diagnostic data within the supplementary sections of manuscripts. VariantValidator would autoformat these data, taking the onus off authors and editors for the majority of reported variants (leaving only complex cases requiring manual curation). Data would be made discoverable via an accessible platform that provides human and artificial intelligence (AI) searching, and links the data to their origin manuscript via a DOI. Additionally, the data will be submitted to ClinVar and LOVD. This intervention would allow researchers and AI solutions to rapidly and accurately identify literature containing variants of interest. Ultimately, the aim is to make diagnostic evidence contained in biomedical journals discoverable, accessible and interpretable, thereby in time speeding up diagnostic rates.

In conclusion, many clinical disciplines rely heavily on genetic diagnostic data published in clinical literature, yet the standards enforced by publishers are insufficient to ensure accurate representation of these data. Through collaboration between technology providers, editorial standards committees, professional standards working groups, journals and publishers, we are paving the way toward accurate representation of diagnostic genomic data in both literature and databases.

Peter J. Freeman [1,2] ✉, John F. Wagstaff [1,2], Ivo F. A. C. Fokkema [3], Garry R. Cutting [4], Heidi L. Rehm [5], Angela C. Davies [1], Johan T. den Dunnen[3], Liam J. Gretton [6] & Raymond Dalgleish [1,2]

[1]Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK. [2]Department of Genetics, Genomics and Cancer Sciences, University of Leicester, Leicester, UK. [3]Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. [4]Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [5]Center for Genomic Medicine, Massachusetts General Hospital, Cambridge Street, Boston, MA, USA. [6]Digital Services, University of Leicester, Leicester, UK.
✉e-mail: peter.j.freeman@manchester.ac.uk

Published online: 2 October 2024

References

1. Haendel, M. et al. *Nat. Rev. Drug Discov.* **19**, 77–78 (2020).
2. Landrum, M. J. et al. *Nucleic Acids Res.* **48**, D835–D844 (2020).
3. Fokkema, I. F. A. C. et al. *Eur. J. Hum. Genet.* **29**, 1796–1803 (2021).
4. Antonarakis, S. E. *Hum. Mutat.* **11**, 1–3 (1998).
5. den Dunnen, J. T. et al. *Hum. Mutat.* **37**, 564–569 (2016).
6. Salgado, D., Bellgard, M. I., Desvignes, J.-P. & Béroud, C. *Hum. Mutat.* **37**, 1272–1282 (2016).
7. Freeman, P. J., Hart, R. K., Gretton, L. J., Brookes, A. J. & Dalgleish, R. *Hum. Mutat.* **39**, 61–68 (2018).
8. Morales, J. et al. *Nature* **604**, 310–315 (2022).
9. Wang, M. et al. *Hum. Mutat.* **39**, 1803–1813 (2018).
10. Higgins, J. et al. *Hum. Mutat.* **42**, 3–7 (2021).
11. Lefter, M. et al. *Bioinformatics* **37**, 2811–2817 (2021).

Acknowledgements

P.F. has been awarded funding from UKRI IAA Commercial Development Fund (R130510) and Wellcome Trust TPA (R126087). We would like to acknowledge the contributions of all those who have submitted code to the VariantValidator project (https://github.com/openvar/variantValidator/graphs/contributors) and our users who have requested new features and alerted us to bugs and issues. Their contributions have driven the success of VariantValidator. We thank Rosaria for an original artistic illustration of Fig. 1 that can be found in our information documentation at https://github.com/openvar/variantValidator/blob/master/README.md.

Competing interests

P.F. has received honoraria from the American College of Molecular Genetics (ACMG). All other authors declare no competing interests.

Additional information

Peer review information *Nature Genetics* thanks Peter Robinson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.