Knowledge multiplies when shared — when calling things by their right name: improving the validation and exchange of genetic data in research and diagnostics

Fokkema, I.F.A.C.

CHAPTER 6

VariantValidator: Standardising variant naming in literature to increase diagnostic rates

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.

Adapted from: *Standardizing variant naming in literature with VariantValidator to increase diagnostic rates.*
Published in: *Nature Genetics 2024; 56:2284–2286*.
DOI: 10.1038/s41588-024-01938-w.

6.1 Abstract

Accurate naming of genetic variants in biomedical journals and publicly accessible databases is essential in the process of identifying clinical data that interprets their clinical and functional consequences. In partnership with the Human Genome Organization (HUGO), we advocate the integration of VariantValidator into journal publishing and database submission systems, aiming to improve the quality of shared genetic data and ultimately improve patient outcomes.

6.2 Introduction

Rare diseases, as classified by the European Union, affect <1 in 2,000 individuals, but with over 8,000 rare genetic diseases recognised, they affect ∼10% of all births globally.[1] Identification and curation of genomic variants are fundamental to the diagnosis and clinical management of individuals. Databases, such as ClinVar[2] and Leiden Open Variation Database (LOVD),[3] offer insights into genetic variants (see also Chapter 4). Clinical scientists rely on these resources to identify documented evidence presented in the literature and reach a diagnosis, but most variants are not described according to the de facto naming standard developed by the Human Genome Variation Society (HGVS)[4,5] (`hgvs-nomenclature.org`). Substandard 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 (UI) termed VariantValidator (`variantvalidator.org`)[7] to assist users (researchers, students and trainers, clinicians, and database curators who generate and utilise genetic data) navigate the HGVS Nomenclature. VariantValidator provides correctly formatted descriptions in the context of all relevant reference sequences (genome, transcript, protein), automatically projecting between genome builds GRCh37 and GRCh38. Additionally, VariantValidator automatically interconverts 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) format (termed pseudo-VCF). Since 2018, VariantValidator has been used to standardise descriptions of genetic variants in clinical reports, literature, and databases, and has been and embedded into our clinical bioinformatics educational programmes for healthcare scientists. The VariantValidator code is hosted on GitHub and our live services are deployed on virtual LAMP (Linux, Apache, MySQL, Python) servers hosted at the University of Leicester, UK, and LEMP (Linux, EnginX, MySQL, Python) servers at the University of Manchester, UK.

6.3 Methods, Results, and Discussion

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 (Figure 6.1). 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 6.1). 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.

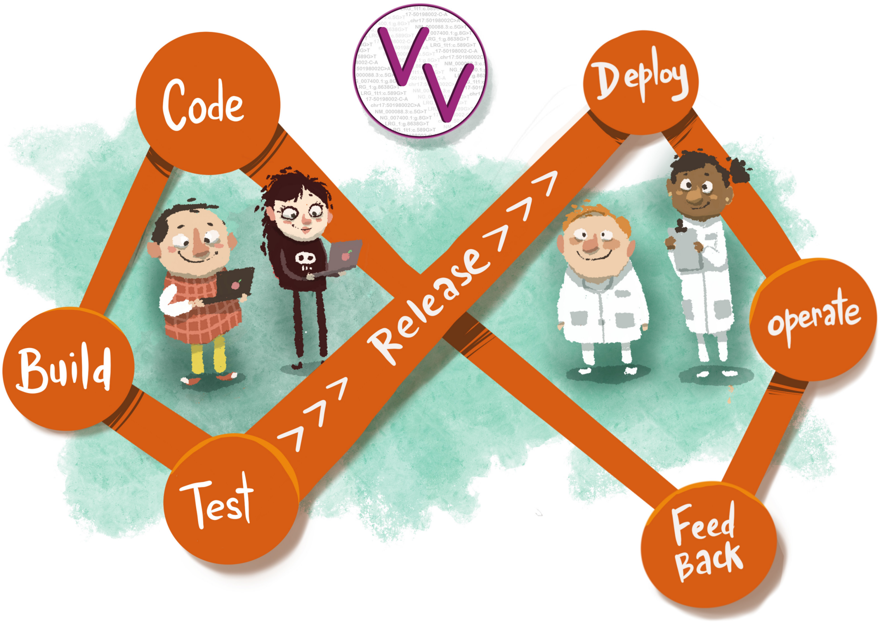

Figure 6.1: **Community of Practice-driven development.** The success of VariantValidator was driven by engagement with our community of users, following an agile development model. We provide platforms that are used in clinical practices, education, and research. Our users test new releases, identify issues, and consider improvements that would assist their professional practice. The users contact us via our dedicated support page (`variantvalidator.org/help/contact/`) or via GitHub issues (`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.

Responding to technological demands, we made VariantValidator compatible with integration into omics platforms. The core VariantValidator engine can be installed as a Python Library (`github.com/openvar/variantValidator`) and we also developed a Python module termed VariantFormatter (`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 will integrate Ensembl transcript reference sequences across our entire tool set by the end of 2024. For rapid deployment, we developed a REST (Representational State Transfer) API (Application Programming Interface), allowing programmatic access to VariantValidator's capabilities without the need for local installation. The API is widely used in the UK National Health Service, across Europe, and in the

Table 6.1: **Summary of key upgrades to VariantValidator functionality since 2018.**

Enhancement	**Example**

Integration of Ensembl data
The *SHANK3* (HGNC:14294) gene MANE Select transcript has 2 additional exons not found in the GRCh37/8 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)`

Filter by transcript Allows users to limit the outputs when submitting genomic (g.) or pseudo-VCF variant descriptions.	Options for `select_transcripts`: **mane_select**: MANE Select transcript **mane**: MANE Select and MANE Plus Clinical **all**: All transcripts at their latest version **raw**: All transcripts at all available versions or, one or more specific transcript ID(s).

Alignment gap-aware projection
Identifies variants aligned within or proximal to imperfect alignments between genome and transcript. Also see row 1 of this table.
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` due to this error in the genomic reference sequence. This projects to `NM_014249.4:c.951=` when correctly handled, not `NM_014249.4:c.951del`, as returned by Mutalyzer. Additionally, if the genomic variant is not seen, we can miss frame-shifting variation (see column 2).
	Input:
	`NC_000015.9:g.72105928_72105929=`
	VariantValidator
	Warning: NM_014249.4 contains 1 fewer bases between c.951_952 than NC_000015.9.
	Correct output:
	`NM_014249.4:c.951dup`
	`NP_055064.1:p.(Thr318HisfsTer23)`
	Mutalyzer
	No warning provided.
	Incorrect projection to the transcript:
	`NM_014249.4:c.=`
	`NP_055064.1:p.=`

RNA variant description (r.) formatting
Following the HGVS Nomenclature standard, variants on RNA level describe changes to mRNA molecules, after the cell's splicing machinery has removed the transcript's introns. As such, the HGVS Nomenclature rule to shift a variant as 3-prime as possible has to be implemented differently than on DNA level. Using RNA as input is useful for users who know the result of a variant on the RNA level but want to obtain the protein variant description for this change. When given RNA input, the generated protein change is no longer regarded a prediction, but an observation. As such, the parentheses are left out from the generated protein description, following HGVS Nomenclature rules.
	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`

US. This API (**`rest.variantvalidator.org`**, **`github.com/openvar/rest_variantValidator`**) interfaces with both VariantValidator and VariantFormatter returning data in standardised formats. We also developed a specialised endpoint to support the LOVD3 suite of variation databases, which can also be optimised for direct integration into genomics workflows. This highly customizable version of the VariantFormatter endpoint allows VariantValidator to replace Mutalyzer,[8] the first software application for variant nomenclature validation, predating the deployment of VariantValidator, as the LOVD variant-description gateway tool. VariantValidator's **`gene2transcripts_v2`** endpoint 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 UI 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[9] and other MANE transcripts via an intuitive table.

We co-developed reference sequence guidelines with the HGVS Variant Nomenclature Committee (HVNC, `hgvs-nomenclature.org/hvnc/`), a committee authorised by the Human Genome Organization (HUGO), and reformatted the Universal Transcript Archive (UTA)[10] database that underpins VariantValidator to ensure strict adherence. For example, our version of UTA (termed VVTA) ensures reference sequence IDs are complete and correctly versioned and coding transcripts minimally comprise a complete coding sequence. We also improved the handling of differing exon structures for single transcripts, which 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 HUGO's Reporting of Sequence Variants committee (HUGO RSV), 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 has published guidance recommending that authors use validation software before publication.[11] 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 for 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 JavaScript Object Notation (JSON) format, which will assist identification by machine learning algorithms. This dual publication allows authors to describe variation in formats recognisable within their profession while ensuring that structured variant evidence in manuscripts is findable.

Concurrently, we are working with the LOVD team, which developed an HGVS syntax validator based on an analysis of common mistakes users make when applying HGVS Nomenclature (see also Chapter 7). Their tool validates descriptions on the syntax level only, so it can support a larger part of the HGVS Nomenclature than sequence-level validators such as VariantValidator which are mostly limited to descriptions of variants with a well-defined sequence change. We aim to integrate the LOVD syntax validator by the end of 2024, allowing recognition of additional common mistakes, suggestions of what the user most likely intended, and assistance with corrections.

Steered by the HUGO RSV committee and the professional-standards working group, our 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 RSV committee and professional-standard working group advocate that Publishing Groups provide direct interfacing of our web API within Editorial Management systems, establishing direct integration of VariantValidator's capabilities. Such interaction will enforce accurate variant descriptions and standardised diagnostic data within the supplementary sections of manuscripts. VariantValidator would autoformat this 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 allows humans and AI alike to search through it, and links the data to its 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 Findable, Accessible, Interoperable, and Reusable (FAIR) for humans and machines alike, thereby speeding up diagnostic rates.

6.4 Conclusions

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 this data. Through collaboration between technology providers, editorial standards committees, professional standards working groups, journals, and publishers, we are paving the way towards accurate representation of diagnostic genomic data in both literature and databases.

6.5 Author contributions

PF* is the VariantValidator PI and lead developer, and primary author of this manuscript.
JW is the technical lead of VariantValidator and contributed sections of the manuscript.
IF* is the LOVD PI and its lead developer and contributes code that enhances VariantValidator

(in collaboration). IF revised and provided sections of the manuscript.
GC* reviewed and edited the manuscript.
HR* reviewed and edited the manuscript.
AD is the lead on embedding VariantValidator into clinical education programmes at The University of Manchester and contributed to the manuscript.
JdD* contributed to the manuscript. LG manages the IT infrastructure which hosts VariantValidator and provided technical solutions for web deployments.
RD* is the former PI of VariantValidator and contributed to the manuscript.

*Where marked, we also acknowledge contributions to VariantValidator by stakeholders in professional committees who have steered developments of our web pages designed to facilitate accurate reporting of sequence variation descriptions in biomedical journals and steered the scope of this manuscript.

6.6 Acknowledgements

PF has been awarded funding from the 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 (`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 the figure illustration.

6.7 Competing interests

PF has received honoraria from the American College of Molecular Genetics (ACMG). The other authors do not report competing interests.

6.8 References

[1] Melissa Haendel, Nicole Vasilevsky, Deepak Unni, Cristian Bologa, Nomi Harris, Heidi Rehm, Ada Hamosh, Gareth Baynam, Tudor Groza, Julie McMurry, Hugh Dawkins, Ana Rath, Courtney Thaxon, Giovanni Bocci, Marcin P. Joachimiak, Sebastian Köhler, Peter N. Robinson, Chris Mungall, and Tudor I. Oprea. *How many rare diseases are there?* Nature Reviews Drug Discovery 2020; 19 (2) 77–78.

[2] Melissa J. Landrum, Shanmuga Chitipiralla, Garth R. Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Kuljeet Kaur, Chunlei Liu, Vitaly Lyoshin, Zenith Maddipatla, Rama Maiti, Joseph Mitchell, Nuala O'Leary, George R. Riley, Wenyao Shi, George Zhou, Valerie Schneider, Donna Maglott, J. Bradley Holmes, and Brandi L. Kattman. *ClinVar: improvements to accessing data*. Nucleic Acids Research 2020; 48 (D1) D835–D844.

[3] Ivo F.A.C. Fokkema, Mark Kroon, Julia A. López Hernández, Daan Asscheman, Ivar Lugtenburg, Jerry Hoogenboom, and Johan T. den Dunnen. *The LOVD3 platform: efficient genome-wide sharing of genetic variants*. European Journal of Human Genetics 2021; 29 (12) 1796–1803.

[4] Stylianos E. Antonarakis and Nomenclature Working Group. *Recommendations for a Nomenclature System for Human Gene Mutations*. Human Mutation 1998; 11 1–3.

[5] Johan T. den Dunnen, Raymond Dalgleish, Donna R. Maglott, Reece K. Hart, Marc S. Greenblatt, Jean McGowan-Jordan, Anne-Francoise Roux, Timothy Smith, Stylianos E. Antonarakis, and Peter E.M. Taschner. *HGVS Recommendations for the Description of Sequence Variants: 2016 Update*. Human Mutation 2016; 37 (6) 564–569.

[6] David Salgado, Matthew I. Bellgard, Jean Pierre Desvignes, and Christophe Béroud. *How to Identify Pathogenic Mutations among All Those Variations: Variant Annotation and Filtration in the Genome Sequencing Era*. Human Mutation 2016; 37 (12) 1272–1282.

[7] Peter J. Freeman, Reece K. Hart, Liam J. Gretton, Anthony J. Brookes, and Raymond Dalgleish. *VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions*. Human Mutation 2018; 39 (1) 61–68.

[8] Mihai Lefter, Jonathan K Vis, Martijn Vermaat, Johan T. den Dunnen, Peter E.M. Taschner, and Jeroen F.J. Laros. *Mutalyzer 2: next generation HGVS nomenclature checker*. Bioinformatics 2021; 37 (18) 2811–2817.

[9] Joannella Morales, Shashikant Pujar, Jane E. Loveland, Alex Astashyn, Ruth Bennett, Andrew Berry, Eric Cox, Claire Davidson, Olga Ermolaeva, Catherine M. Farrell, Reham Fatima, Laurent Gil, Tamara Goldfarb, Jose M. Gonzalez, Diana Haddad, Matthew Hardy, Toby Hunt, John Jackson, Vinita S. Joardar, Michael Kay, Vamsi K. Kodali, Kelly M. McGarvey, Aoife McMahon, Jonathan M. Mudge, Daniel N. Murphy, et al. *A joint NCBI and EMBL-EBI transcript set for clinical genomics and research*. Nature 2022; 604 (7905) 310–315.

[10] Meng Wang, Keith M. Callenberg, Raymond Dalgleish, Alexandre Fedtsov, Naomi K. Fox, Peter J. Freeman, Kevin B. Jacobs, Piotr Kaleta, Andrew J. McMurry, Andreas Prlić, Veena Rajaraman, and Reece K. Hart. *hgvs: A Python package for manipulating sequence variants using HGVS nomenclature: 2018 Update*. Human Mutation 2018; 39 (12) 1803–1813.

[11] Jan Higgins, Raymond Dalgleish, Johan T. den Dunnen, Greg Barsh, Peter J. Freeman, David N. Cooper, Sara Cullinan, Kay E. Davies, Huw Dorkins, Li Gong, Issei Imoto, Teri E. Klein, Bruce Korf, Adya Misra, Mark H. Paalman, Sarah Ratzel, Juergen K.V. Reichardt, Heidi L. Rehm, Katsushi Tokunaga, Karen E. Weck, and Garry R. Cutting. *Verifying nomenclature of DNA variants in submitted manuscripts: Guidance for journals*. Human Mutation 2021; 42 (1) 3–7.