Genome Biology

DATABASE

Open Access

MaveDB 2024: a curated community database with over seven million variant effects from multiplexed functional assays

Alan F. Rubin[1,2*], Jeremy Stone[3], Aisha Haley Bianchi[4], Benjamin J. Capodanno[3], Estelle Y. Da[1], Mafalda Dias[5,6], Daniel Esposito[1], Jonathan Frazer[5,6], Yunfan Fu[1,2], Sally B. Grindstaff[3], Matthew R. Harrington[4], Iris Li[1], Abbye E. McEwen[3,4,7], Joseph K. Min[4], Nick Moore[1], Olivia G. Moscatelli[2,8], Jesslyn Ong[8,9], Polina V. Polunina[10], Joshua E. Rollins[11], Nathan J. Rollins[12], Ashley E. Snyder[3], Amy Tam[13], Matthew J. Wakefield[1,2,14], Shenyi Sunny Ye[4], Lea M. Starita[3,4], Vanessa L. Bryant[2,8,15], Debora S. Marks[13,16*] and Douglas M. Fowler[3,4,17*]

*Correspondence: alan.rubin@wehi.edu.au; debbie@hms.harvard.edu; dfowler@uw.edu

[1] Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Australia
[3] Brotman Baty Institute for Precision Medicine, Seattle, USA
[13] Department of Systems Biology, Harvard Medical School, Boston, USA
Full list of author information is available at the end of the article

Abstract

Multiplexed assays of variant effect (MAVEs) are a critical tool for researchers and clinicians to understand genetic variants. Here we describe the 2024 update to MaveDB (https://www.mavedb.org/) with four key improvements to the MAVE community's database of record: more available data including over 7 million variant effect measurements, an improved data model supporting assays such as saturation genome editing, new built-in exploration and visualization tools, and powerful APIs for data federation and streamlined submission and access. Together these changes support MaveDB's role as a hub for the analysis and dissemination of MAVEs now and into the future.

Keywords: Multiplexed assays of variant effect, MAVEs, Deep mutational scanning, DMS, Variant classification, Functional genomics

Background

Variation within genomes produces interindividual differences governing a multitude of traits, including many implicated in disease. As DNA sequencing continues to become less expensive and more widely deployed, new human genetic variants are being observed at a staggering pace. Among 800,000 individuals in gnomAD v4 [1], approximately 786 million small variants comprising single nucleotide changes and small deletions/insertions have been identified, of which 16 million are missense variants (i.e., single amino acid changes). In contrast, only 1 million missense variants have been annotated in ClinVar [2] and 88% are currently variants of uncertain significance that cannot be used for clinical decision-making. Understanding how these observed variants, as

well as others we will encounter as more individuals are sequenced, impact molecular, cellular, and organismal phenotypes represents a central challenge for genomics [3].

In the past, genetic variants would be tested for functional effects in bespoke assays singly or in relatively low numbers, but more recent technologies have enabled multiplexed assays of variant effect (MAVEs) [4, 5]. In a MAVE, the functional effects of thousands or tens of thousands of variants of a DNA regulatory region, coding gene, untranslated region, or other functional element are simultaneously experimentally determined. To achieve this scale, a large library of variants is made and tested in a pooled fashion, using high-throughput DNA sequencing to read out variant effects (for a detailed description see [6–8]).

The result of a MAVE is a comprehensive variant effect map, which contains the experimentally measured effects of most or all of the possible single nucleotide or missense variants, and may include small insertions and deletions. Variant effect maps have proven exceptionally useful. For example, in genes where germline variants can increase disease risk, variant effect maps can help resolve a large proportion of clinical variants of uncertain significance [9, 10]. Variant effect maps can also be used to probe protein sequence/function relationships [11–21], assist in protein design [22], reveal protein structure [23, 24], elucidate regulatory DNA and gene function by interrogating noncoding sequences [25–28], and train or evaluate variant effect predictors [29–32].

Efforts are now underway to scale up MAVEs to cover a significant fraction of the human genome [33, 34], but realizing their potential requires improved discoverability. In 2019, we created MaveDB [35], a public, open source repository for submitting, sharing, and accessing MAVE data and associated metadata in a standardized, searchable format through an easy-to-use web interface. However, the original version of MaveDB suffered from four key limitations. First, it contained only a small fraction of the data available at the time. Second, data from new multiplexed assay methods such as saturation genome editing [19, 36, 37] were not compatible with the original MaveDB data model. Third, the ability to explore datasets was limited and visualizing data required external tools. Finally, MaveDB was not designed with federation across genomic data resources in mind.

To address those limitations, firstly we have expanded the database content by extensively curating multiplexed assay results and encouraging community contributions, constituting a six-fold increase in the total number of variant effect measurements in the database and an over 30-fold increase in the number of datasets compared to the original publication. As of November 2024, MaveDB contained over 7 million variant effect measurements and 1884 datasets. We have also implemented numerous technical advances and data model improvements. This includes refining and formalizing our variant representation with an emphasis on compliance with established standards like HGVS [38], allowing us to support more diverse types of variants and associated experimental designs, while also improving compatibility with emerging standards like the GA4GH Variant Representation Specification (VRS) [39] that will simplify mapping datasets to reference genomic coordinates. We have updated our data model by adding a new type of record for imputation or the combination of results across multiple assays. We also invested in an improved interface for searching and filtering datasets, as well as adding new automatically generated visualizations. Lastly, we further improved the user

experience by adding API-based user uploads aimed at researchers who are submitting large or complex datasets, or engaging in MAVE data production at scale.

Construction and content

MaveDB is designed to store and distribute multiplexed variant functional data, including scores and associated metadata. Minimally, this consists of a collection of variant effect scores that describe the functional consequences of the nucleotide or amino acid variants, as well as information about the target sequence. The metadata typically includes descriptions of the experimental and data analysis methods and references to information in other databases, such as DNA sequencing reads. Most datasets in MaveDB are from published papers, although this is not required for inclusion.

When the original MaveDB manuscript was published in 2019, only 54 datasets from published MAVEs were included. Thus, we launched a concerted effort to deposit datasets that were not yet included in MaveDB, adding 1228 new datasets containing a total of 3.7 million variant effect measurements. Thanks to this curation and contributions from the community, as of November 2024 MaveDB contained 1884 datasets encompassing 7 million variant effect measurements across diverse targets (Fig. 1).

Our curation team spanned three sites: WEHI and the University of Melbourne in Melbourne, Australia; University of Washington in Seattle, USA; and Harvard University in Boston, USA. We developed a robust process for summarizing heterogeneous experimental results, including training materials, much of which has been incorporated into updated MaveDB documentation available on the website. Key information was extracted from publications and synthesized into a title, short description, abstract, and methods as metadata for each record. Accession numbers for raw sequence data and target sequence identifiers for each dataset were also included. Each curated entry was peer reviewed by at least one other team member to ensure all relevant information was present and accurate before submission to the database. In addition to writing the free text sections and organizing associated metadata, our curation team also formatted scores and related values from published supplemental data.

To make it easier for users to discover MAVE data from publications, in addition to PubMed identifiers, we updated our data model to support bioRxiv and medRxiv preprints and Crossref DOIs. We also store structured metadata for each of these references, including journal or preprint server and all author names, and allow users to search and filter based on this information. MaveDB also now distinguishes between a primary reference, which describes the data contained in the record, and secondary references, which describe methods, key reagents, or software used to generate the data.

MaveDB has a hierarchical structure populated by score set, experiment, and experiment set records. Score set records contain the variant effect scores and associated data columns, such as variance estimates and variant counts, details about the experimental target sequence, and a description of the score calculations. Scores are required, but any number of additional numeric columns can be named by the submitter. Experiment records summarize the assay that was performed and can group multiple score sets, preventing double-counting of assays when raw data is reanalyzed and improving discoverability for users. Experiment set records do not have any data or metadata themselves, but group related experiments, such as multiple assays performed on a single target

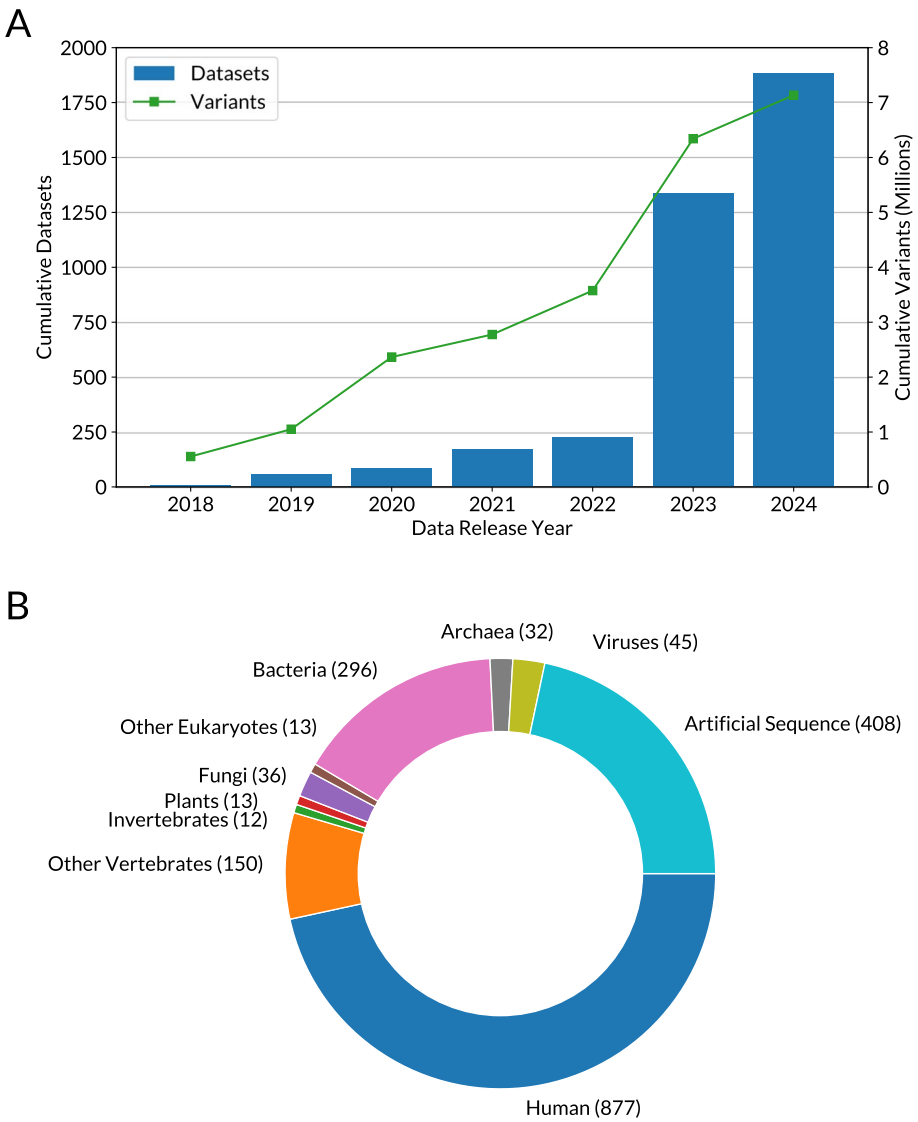

Fig. 1 MaveDB contents as of November 2024. **A** Growth of the database by year. The bars show the cumulative number of datasets and the green line shows the cumulative number of variant effect measurements. **B** Diversity of target sequences. NCBI Taxonomy IDs were assigned and grouped according to the categories shown

and described in the same publication. Note that when counting "datasets" above, we counted experiment records since each describes a unique assay on a target.

To represent scores based on the transformation or combination of existing scores, MaveDB now implements meta-analysis score sets. For example, a dataset that imputes the values of missing scores should be represented as a meta-analysis linked to the pre-imputation score set, ensuring the original scores are preserved and discoverable. Another use case is representing the combination of multiple assay results at the level of the associated scores (Fig. 2).

To improve compatibility with the HGVS Sequence Variant Nomenclature [38], support additional variant types, and enable more robust validation, we implemented MAVE-HGVS, which replaces the previous MaveDB variant representation based on

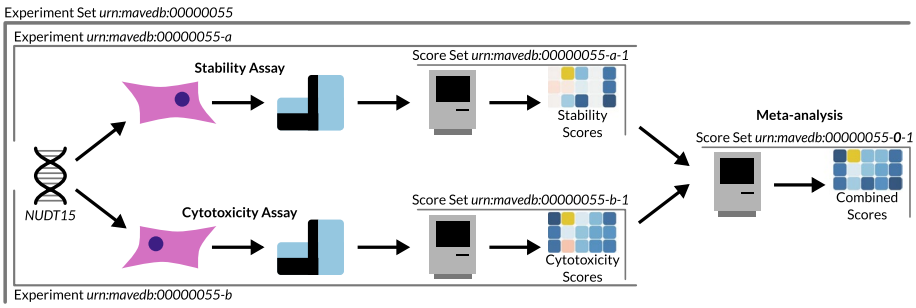

Fig. 2 Example of a meta-analysis score set. The cartoon uses a real-world dataset to illustrate the relationship between experiment sets, experiments, score sets, and meta-analysis score sets. The results from two assays performed on the gene *NUDT15* were combined into a resulting "function score" that summarized performance across both assays [40]

the Enrich2 [41] output format. While packages exist for parsing HGVS [42, 43], they are intended for use in human genetics and rely on sequence database entries that are not always available for multiplexed assay targets. MAVE-HGVS has a reference Python implementation, mavehgvs, used to validate variants uploaded to the database by ensuring variant strings are correctly formatted and consistent with the score set's target sequence.

To better represent experiments that directly edit the human genome, such as saturation genome editing, we implemented a new way to specify and validate variants. Contributors can now define variants with respect to a transcript accession or a human genome reference, with validation handled by SeqRepo [44] because access to a genome and transcript database is required. This is in contrast to most score set records, which specify their own target sequence and are validated using mavehgvs.

To support current and future developments of the MaveDB platform, particularly API improvements, we have transitioned to a new codebase using FastAPI and Vue.js, replacing the previous codebase that used the Django 1.11 framework. MaveDB now runs as a set of Docker containers orchestrated using Docker Compose, simplifying deployment for the production server as well as for open source developers who wish to contribute to the project. In response to increased usage and demands for greater reliability and future scaling, we have also migrated MaveDB to the cloud using Amazon Web Services.

To promote data federation and the open use of MAVE data globally, we have relicensed nearly all datasets in MaveDB to the Creative Commons CC0 public domain license [45], and now recommend it to submitters. Moving away from the previously recommended but restrictive CC-BY-NC-SA non-commercial license [46] was a result of extensive consultation with maintainers of other biological data repositories as well as the broader MAVE community. This license change combined with the API improvements has allowed us to provide bulk data downloads as described below.

Utility and discussion

Web interface

MaveDB features a purpose-built web interface for users to explore and discover datasets as well as upload newly generated or curated datasets. Since the initial launch, the interface has been completely re-implemented using the Vue JavaScript framework. This

delivers a more responsive and reactive user experience compared to the previous version of MaveDB, which was based on Django's HTML templates.

The score set pages now display automatically generated interactive visualizations for exploration and interpretation, including a score histogram showing the distribution of variant effect scores and a variant effect heatmap (Fig. 3A). The search page has been updated to add categorical filters that encourage exploration of MaveDB data, including publication information such as author or journal (Fig. 3B).

For users who want to contribute data using the web interface, we have overhauled the score set interface to replace the overly-complex single-page form with a guided multi-stage process (Fig. 3C). This simplifies each step of the process and allows for more informative validation and error checking. Guidance for users is now integrated into the form itself, rather than relying entirely on documentation hosted elsewhere on the website.

Improved API support

The previous version of MaveDB only accepted data via a web form, but the server now also supports data deposition through the REST API using the same logic and validation as the web interface to ensure continuity and data integrity. Using the API to deposit programmatically simplifies submission for some complex experimental designs, such as a series of similar assays that measure variant effects with different small molecules.

To facilitate local validation of datasets, we maintain the MaveDB API code as an installable package on PyPI, the Python Package Index. This allows power users to apply the same validators and data models that are running on the server when preparing datasets for submission. We hope that authors of MAVE analysis pipelines will consider adopting the MaveDB API as an output option.

In addition to serving score set data files identical to those downloadable via the web interface, the API also provides structured data and metadata for individual variants. This feature currently only supports access using MaveDB's internal variant identifiers, which we are in the process of mapping to more widely used formats [47].

Bulk data releases

For users who want to access the entirety of MaveDB, we now have an archive of all CC0-licensed data available via Zenodo (see Data availability). It contains a single file in JSON format with all structured metadata for every experiment set, experiment, and score set, as well as a directory of data tables in comma separated value (CSV) format that have the scores and counts for each score set. Archival snapshots increase reproducibility by allowing users to cite a specific version of the database's contents, and we intend to add complete archives biannually in May and November.

Recommendations for user uploads

With the introduction of meta-analysis score sets, MaveDB's hierarchical data model enables more comprehensive provenance tracking for individual variant measurements from a multiplexed assay. We suggest that users upload minimally transformed scores as standard score sets to MaveDB, and create meta-analysis score sets that describe normalization or imputation steps as applicable. This supports other researchers who

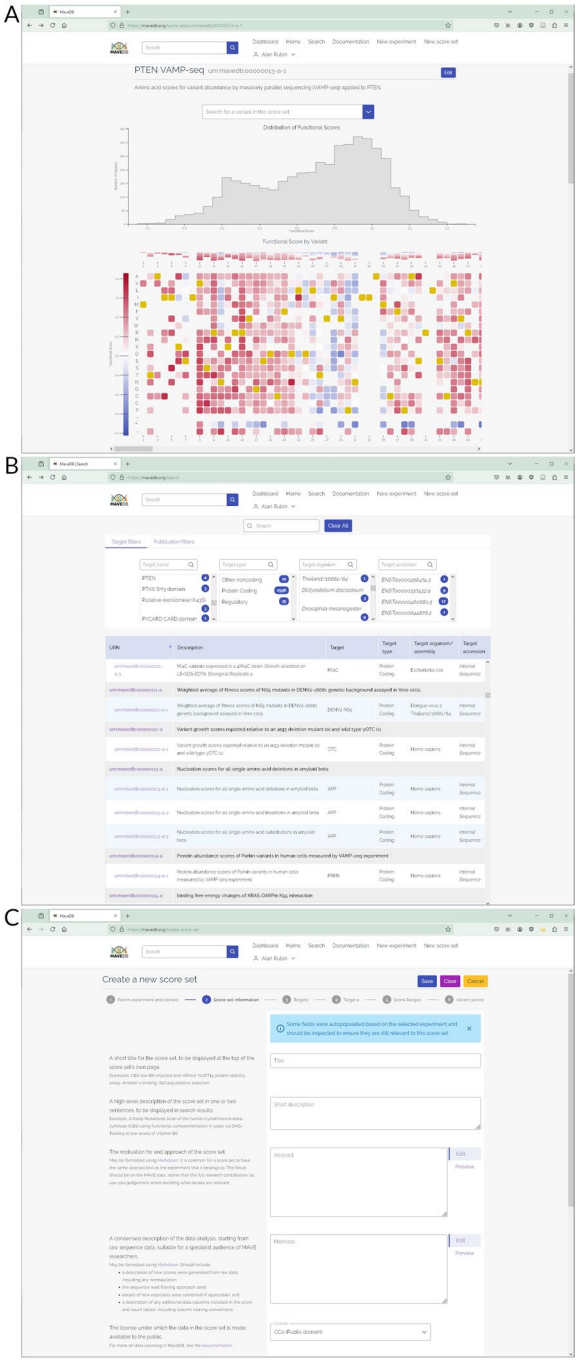

Fig. 3 MaveDB web interface screenshots. **A** Score set visualizations. Score set pages now feature automatically generated visualizations, including a score histogram and variant effect heatmap. For non-coding targets, the heatmap is displayed at the nucleotide level. **B** Search page. The interface includes target sequence-based filters at the top, and listings for each matching experiment and its score sets in the main body of the page. MaveDB also supports filtering on publication information such as author via the "Publication filters" tab. **C** Score set creation. Users contributing score sets via the web form can follow this step-by-step workflow with embedded documentation

want to evaluate their own methods or build models that would be sensitive to data normalization.

MaveDB also accepts optional count data for each variant in addition to scores. We strongly encourage submitters to provide this information as it promotes the development of new statistical models for calculating variant scores.

Users should familiarize themselves with the MaveDB hierarchical structure of score set (including meta-analysis), experiment, and experiment set records described above, and try to follow the convention of one experiment per assay and one experiment set per unique target in a study. We recommend that users include the details specified in the MAVE minimum information standards [48] when preparing their textual metadata.

Conclusions

MAVEs are an important approach for measuring, understanding, and predicting variant effects on a genome-wide scale, but the data must be stored in a stable, standardized fashion along with the metadata required for downstream use. Moreover, MAVE datasets must be readily available and discoverable, and MAVE data must be accessible programmatically. With this 2024 update to MaveDB, we have built on the successes of the initial version of the database and made major strides towards fulfilling these aims.

We made several major improvements to our data model, bolstering our ability to store, standardize, and present heterogeneous MAVE datasets. These changes were made possible by the substantial software engineering effort that went into overhauling the codebase, and we are now better positioned to continue to develop new features like the automatic data visualizations, and respond to innovations in MAVE experimental technologies. Furthermore, we can more easily support specific use cases for MAVE data, including variant effect prediction, drug discovery, and precision medicine.

To increase the amount of information available in MaveDB, we launched a massive curation effort involving hundreds of additional datasets, ultimately populating MaveDB with nearly half of all data published in the literature. In addition, we have seen an encouraging level of engagement from the broader MAVE community, with dozens of international researchers contributing their results of their own accord. We hope that our continued investment in the web interface as well as the API will further encourage prospective users to submit their data, and we thank the many members of the community who have already done so.

Acknowledgements

The Atlas of Variant Effects (AVE) Alliance Data Coordination and Dissemination workstream contributed valuable feedback on the design and goals of MaveDB.

Peer review information

Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions

AFR and DMF designed the database. AFR, JKM, NJR, BJC, EYD, DE, SBG, MRH, NM, AES, JS, and PVP wrote the database and associated software. AFR, NJR, AHB, MD, JF, YF, MRH, IL, OM, JO, PVP, JER, MJW, SY, AT, AEM, and DSM curated datasets. AFR, VLB, DSM, and DMF supervised dataset curation. AFR, JS, LMS, DSM, and DMF supervised the software projects. AFR, AHB, OM, SY, and DMF wrote the paper. All authors read and approved the final manuscript.

Funding

This work was supported by the National Institutes of Health (NIH; RM1HG010461 to DMF, UM1HG011969 to LMS and DMF, R01HG013025 to LMS, T32GM007454) and by Chan Zuckerberg Initiative (CZI2018-191853 to DSM). MD and JF were supported by the Spanish Ministry of Science and Innovation (PID2022-140793NA-I00). YF was supported by a Melbourne Research Scholarship. AEM was supported by Early Career Award Alex's Lemonade Stand for Childhood

Cancer and RUNX1 foundation 21-25037, and the Brotman Baty Institute Catalytic Collaborations Grant CC28. PVP was supported by the Freiburg Galaxy Team funded by the German Federal Ministry of Education and Research BMBF grant 031 A538A de.NBI-RBC and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) within the framework of LIBIS/de.NBI Freiburg. The research benefited from support from the Victorian State Government Operational Infrastructure Support and Australian Government NHMRC Independent Research Institute Infrastructure Support. This project received grant funding from the Australian Government.

Data availability

MaveDB source code is available on GitHub [49, 50] and Zenodo [51, 52]. The version of the MaveDB back-end described here is v2024.4.2 and the version of the MaveDB front-end described here is v2024.4.3. MaveDB is distributed under the AGPLv3 license. mavehgvs source code is available on GitHub [53] and Zenodo [54]. The version described here is v0.6.1. mavehgvs is distributed under the 3-Clause BSD license. Notebooks used for generating the panels in Fig. 1 are available on GitHub [55] and Zenodo [56]. The version described here is v0.1.0. The notebooks are distributed under the MIT license. The November 2024 MaveDB bulk data download is available from Zenodo [57]. The dataset depicted in Fig. 2 is available in MaveDB under experiment set urn:mavedb:00000055.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

NJR is employed by Seismic Therapeutics. DSM participates in an advisory role for Dyno Therapeutics, Octant Bio, Jura Bio, Tectonic Therapeutic, and Seismic Therapeutics.

Author details

[1]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Australia. [2]Department of Medical Biology, University of Melbourne, Parkville, Australia. [3]Brotman Baty Institute for Precision Medicine, Seattle, USA. [4]Department of Genome Sciences, University of Washington, Seattle, USA. [5]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. [6]University Pompeu Fabra, Barcelona, Spain. [7]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, USA. [8]Immunology Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Australia. [9]Department of Microbiology and Immunology, University of Melbourne, Parkville, Australia. [10]Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany. [11]Department of Computer Science, The Graduate Center, The City University of New York, New York, USA. [12]Seismic Therapeutics, Watertown, USA. [13]Department of Systems Biology, Harvard Medical School, Boston, USA. [14]Department of Obstetrics, Gynaecology and Newborn Health, University of Melbourne, Parkville, Australia. [15]Department of Clinical Immunology & Allergy, The Royal Melbourne Hospital, Parkville, Australia. [16]Broad Institute of Harvard and MIT, Boston, USA. [17]Department of Bioengineering, University of Washington, Seattle, USA.

Received: 13 July 2024 Accepted: 10 January 2025
Published online: 21 January 2025

References

1. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2024;625:92–100.
2. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46:D1062–7.
3. Fowler DM, Rehm HL. Will variants of uncertain significance still exist in 2030? Am J Hum Genet. 2024;111:5–10.
4. Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, et al. Variant interpretation: functional assays to the rescue. Am J Hum Genet. 2017;101:315–25.
5. Tabet D, Parikh V, Mali P, Roth FP, Claussnitzer M. Scalable functional assays for the interpretation of human genetic variation. Annu Rev Genet. 2022;56:441–65.
6. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11:801–7.
7. Kinney JB, McCandlish DM. Massively parallel assays and quantitative sequence–function relationships. Annu Rev Genomics Hum Genet. 2019;20:99–127.
8. Weile J, Roth FP. Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas. Hum Genet. 2018;137:665–78.
9. Fayer S, Horton C, Dines JN, Rubin AF, Richardson ME, McGoldrick K, et al. Closing the gap: systematic integration of multiplexed functional data resolves variants of uncertain significance in BRCA1, TP53, and PTEN. Am J Hum Genet. 2021;108:2248–58.
10. Scott A, Hernandez F, Chamberlin A, Smith C, Karam R, Kitzman JO. Saturation-scale functional evidence supports clinical variant interpretation in Lynch syndrome. Genome Biol. 2022;23:266.
11. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, et al. High-resolution mapping of protein sequence-function relationships. Nat Methods. 2010;7:741–6.

12. McLaughlin RN Jr, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature. 2012;491:138–42.
13. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. A comprehensive, high-resolution map of a gene's fitness landscape. Mol Biol Evol. 2014;31:1581–92.
14. Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic Acids Res. 2014;42:e112–e112.
15. Mishra P, Flynn JM, Starr TN, Bolon DNA. Systematic mutant analyses elucidate general and client-specific aspects of Hsp90 function. Cell Rep. 2016;15:588–98.
16. Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, et al. Prospective functional classification of all possible missense variants in PPARG. Nat Genet. 2016;48:1570–5.
17. Weile J, Sun S, Cote AG, Knapp J, Verby M, Mellor JC, et al. A framework for exhaustively mapping functional missense variants. Mol Syst Biol. 2017;13:957.
18. Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat Genet. 2018;50:874–82.
19. Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature. 2018;562:217–22.
20. Tsuboyama K, Dauparas J, Chen J, Laine E, Mohseni Behbahani Y, Weinstein JJ, et al. Mega-scale experimental analysis of protein folding stability in biology and design. Nature. 2023;620:434–44.
21. Beltran A, Jiang X, Shen Y, Lehner B. Site-saturation mutagenesis of 500 human protein domains. Nature. 2025.
22. Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, et al. Computational design of ligand-binding proteins with high affinity and selectivity. Nature. 2013;501:212–6.
23. Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, Gauthier NP, Sander C, et al. Inferring protein 3D structure from deep mutation scans. Nat Genet. 2019;51:1170–6.
24. Schmiedel JM, Lehner B. Determining protein structures using deep mutagenesis. Nat Genet. 2019;51:1177–86.
25. Ke S, Anquetil V, Zamalloa JR, Maity A, Yang A, Arias MA, et al. Saturation mutagenesis reveals manifold determinants of exon definition. Genome Res. 2018;28:11–24.
26. Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell RJA, et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat Commun. 2019;10:3583.
27. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012;30:271–7.
28. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, et al. Massively parallel functional dissection of mammalian enhancers *in vivo*. Nat Biotechnol. 2012;30:265–70.
29. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599:91–5.
30. Gray VE, Hause RJ, Luebeck J, Shendure J, Fowler DM. Quantitative missense variant effect prediction using large-scale mutagenesis data. Cell Syst. 2018;6:116–24.e3.
31. Wu Y, Li R, Sun S, Weile J, Roth FP. Improved pathogenicity prediction for rare human missense variants. Am J Hum Genet. 2021;108:1891–906.
32. Notin P, Dias M, Frazer J, Hurtado JM, Gomez AN, Marks D, et al. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. Proceedings of the 39th International Conference on Machine Learning in Proceedings of Machine Learning Research. 2022;162:16990–7017.
33. IGVF Consortium. Deciphering the impact of genomic variation on function. Nature. 2024;633:47–57.
34. Fowler DM, Adams DJ, Gloyn AL, Hahn WC, Marks DS, Muffley LA, et al. An Atlas of Variant Effects to understand the genome at nucleotide resolution. Genome Biol. 2023;24:147.
35. Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, et al. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biol. 2019;20:223.
36. Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. Saturation editing of genomic regions by multiplex homology-directed repair. Nature. 2014;513:120–3.
37. Radford EJ, Tan H-K, Andersson MHL, Stephenson JD, Gardner EJ, Ironfield H, et al. Saturation genome editing of DDX3X clarifies pathogenicity of germline and somatic variation. Nat Commun. 2023;14:7702.
38. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS recommendations for the description of sequence variants: 2016 update. Hum Mutat. 2016;37:564–9.
39. Wagner AH, Babb L, Alterovitz G, Baudis M, Brush M, Cameron DL, et al. The GA4GH Variation Representation Specification: a computational framework for variation representation and federated identification. Cell Genom. 2021;1:100027.
40. Suiter CC, Moriyama T, Matreyek KA, Yang W, Scaletti ER, Nishii R, et al. Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. Proc Natl Acad Sci USA. 2020;117:5394–401.
41. Rubin AF, Gelman H, Lucas N, Bajjalieh SM, Papenfuss AT, Speed TP, et al. A statistical framework for analyzing deep mutational scanning data. Genome Biol. 2017;18:150.
42. Hart RK, Rico R, Hare E, Garcia J, Westbrook J, Fusaro VA. A Python package for parsing, validating, mapping and formatting sequence variants using HGVS nomenclature. Bioinformatics. 2015;31:268–70.
43. Wang M, Callenberg KM, Dalgleish R, Fedtsov A, Fox NK, Freeman PJ, et al. hgvs: a Python package for manipulating sequence variants using HGVS nomenclature: 2018 Update. Hum Mutat. 2018;39:1803–13.
44. Hart RK, Prlić A. SeqRepo: a system for managing local collections of biological sequences. PLoS ONE. 2020;15:e0239883.
45. Creative Commons — CC0 1.0 Universal. Available from: https://creativecommons.org/publicdomain/zero/1.0/
46. Creative Commons — Attribution-NonCommercial-ShareAlike 4.0 International — CC BY-NC-SA 4.0. Available from: https://creativecommons.org/licenses/by-nc-sa/4.0/
47. Arbesfeld JA, Da EY, Kuzma K, Paul A, Farris T, Riehle K, et al. Mapping MAVE data for use in human genomics applications. bioRxiv. 2023;2023.06.20.545702.

48. Claussnitzer M, Parikh VN, Wagner AH, Arbesfeld JA, Bult CJ, Firth HV, et al. Minimum information and guidelines for reporting a multiplexed assay of variant effect. Genome Biol. 2024;25:100.
49. Capodanno BJ, Stone J, Da EY, Grindstaff SB, Harrington MR, Moore N, Syder AE, Rubin AF. mavedb-api. GitHub. https://github.com/VariantEffect/MaveDB-API (2024).
50. Capodanno BJ, Stone J, Da EY, Grindstaff SB, Harrington MR, Polunina PV, Syder AE, Rubin AF. mavedb-ui. GitHub. https://github.com/VariantEffect/MaveDB-UI (2024).
51. Capodanno BJ, Stone J, Da EY, Grindstaff SB, Harrington MR, Moore N, Syder AE, Rubin AF. VariantEffect/mavedb-api: v2024.4.2. Zenodo. https://doi.org/10.5281/zenodo.14201451 (2024).
52. Capodanno BJ, Stone J, Da EY, Grindstaff SB, Harrington MR, Polunina PV, Syder AE, Rubin AF. VariantEffect/mavedb-ui: v2024.4.3. Zenodo. https://doi.org/10.5281/zenodo.14207533 (2024).
53. Rubin AF. mavehgvs. GitHub. https://github.com/VariantEffect/mavehgvs (2023).
54. Rubin AF. mavehgvs. Zenodo. https://doi.org/10.5281/zenodo.8281119 (2023).
55. Rubin AF. MaveDB Analytics. GitHub. https://github.com/afrubin/mavedb-analytics (2024).
56. Rubin AF. afrubin/mavedb-analytics: 0.1.0. Zenodo. https://doi.org/10.5281/zenodo.14172359 (2024).
57. MaveDB contributors. MaveDB. https://doi.org/10.5281/zenodo.14172004 (2024).

Publisher's Note