

## Sequence analysis

# Application of qualifying variants for genomic analysis

Dylan Lawless<sup>1,\*</sup>, Ali Saadat<sup>2</sup>, Mariam Ait Oumelloul<sup>2</sup>, Simon Boutry<sup>2</sup>, Veronika Stadler<sup>1</sup>, Sabine Österle<sup>3</sup>, Jan Armida<sup>3</sup>, David Haerry<sup>4</sup>, D. Sean Froese<sup>5</sup>, Luregn J. Schlapbach<sup>1</sup>, Jacques Fellay<sup>2,6</sup>

<sup>1</sup>Department of Intensive Care and Neonatology, University Children's Hospital Zürich, University of Zürich, Zurich, 8008, Switzerland

<sup>2</sup>Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, 1015, Switzerland

<sup>3</sup>Swiss Personalized Health Network, SIB Swiss Institute of Bioinformatics, Basel, 4051, Switzerland

<sup>4</sup>Positive Council, Zurich, 8031, Switzerland

<sup>5</sup>Division of Metabolism and Children's Research Center, University Children's Hospital Zürich, University of Zurich, Zurich, 8008, Switzerland

<sup>6</sup>Precision Medicine Unit, Biomedical Data Science Center, Lausanne University Hospital and University of Lausanne, Lausanne, 1015, Switzerland

\*Corresponding author. Department of Intensive Care and Neonatology, University Children's Hospital Zürich, University of Zürich, Zurich, 8008, Switzerland.  
E-mail: dylan.lawless@kispi.uzh.ch.

Associate Editor: Can Alkan

## Abstract

**Motivation:** Qualifying variants (QVs) are genomic alterations selected by defined criteria within analysis pipelines. Although crucial for both research and clinical diagnostics, QVs are often seen as simple filters rather than dynamic elements that influence the entire workflow. In practice these rules are embedded within pipelines, which hinders transparency, audit, and reuse across tools. A unified, portable specification for QV criteria is needed.

**Results:** Our aim is to embed the concept of a "QV" into the genomic analysis vernacular, moving beyond its treatment as a single filtering step. By decoupling QV criteria from pipeline variables and code, the framework enables clearer discussion, application, and reuse. It provides a flexible reference model for integrating QVs into analysis pipelines, improving reproducibility, interpretability, and interdisciplinary communication. Validation across diverse applications confirmed that QV based workflows match conventional methods while offering greater clarity and scalability.

**Availability and implementation:** The source code and data are accessible at the Zenodo repository <https://doi.org/10.5281/zenodo.17414191>. Manuscript files are available at <https://github.com/DylanLawless/qvApp2025lawless>. The QV framework is available under the MIT licence, and the dataset will be maintained for at least two years following publication.

## 1 Introduction

Qualifying variants (QVs) are genomic alterations selected by specific criteria within genome processing pipelines, serving as dynamic elements essential for both research and clinical diagnostics. QVs are not merely static filters applied at a single step in an analysis pipeline; rather, they are dynamic, multifaceted elements that permeate the entire workflow, from initial data quality control to final result interpretation. This nuanced perspective underscores that QVs play an integral role in shaping the fidelity and reproducibility of genomic analyses, enabling the iterative refinement of data and facilitating the integration of diverse analytical strategies throughout the pipeline.

Often, QV selection adheres to established variant classification and reporting standards (Richards *et al.* 2015, Li *et al.* 2017, Li and Wang 2017, Riggs *et al.* 2020, Tavtigian *et al.* 2020) and standardized workflows (Anderson *et al.* 2010, Pedersen *et al.* 2021, Uffelmann *et al.* 2021). However, a unified framework for QVs is lacking, despite the recognized benefits of similar initiatives, such as Polygenic Risk Score

(PRS) reporting standards (Lambert *et al.* 2021, Wand *et al.* 2021). Tools such as *vcfexpress* (Pedersen and Quinlan 2025) enable flexible filtering and formatting of Variant Call Format (VCF) files using user-defined expressions. Treating QV criteria as an external parameter layer complements these tools by externalizing their thresholds and logic. This approach improves reproducibility across distributed computing environments (Bal *et al.* 1989) and integrates seamlessly with workflow managers like Snakemake (Mölder *et al.* 2021) or Nextflow (Di Tommaso *et al.* 2017).

QV selection criteria vary by application. In Genome-Wide Association Study (GWAS), thresholds favor common variants, yielding datasets with over 500 000 variants per subject, whereas rare disease analyses use stringent filters producing fewer than 1000 variants, often limited to known genes or pathogenic loci. Although targeted filtering is valuable (Cirulli *et al.* 2015, Povysil *et al.* 2019), no unified approach exists. In practice, QV sets range from broad quality control filters to specific disease panels, and their definition is critical for reproducibility and accurate reporting,

influencing results as much as the pipeline itself (Olson *et al.* 2023).

As Whole Genome Sequencing (WGS) becomes standard for large cohorts (Lee *et al.* 2018, Jansen *et al.* 2019), the integration of diverse QV protocols is critical for data cleaning and analysis. During sequencing analysis several layers can be responsible for triggering QV protocols, including pre-existing metadata, technical Quality Control (QC) results, and post-calling annotations, highlighting the need for a clear, unified approach.

We introduce the QV as a standalone entity, independent from other pipeline variables. Structured human- and machine-readable criteria, aligned with FAIR principles (Wilkinson *et al.* 2016), facilitate integration across databases (Touré *et al.* 2023, van der Horst *et al.* 2023). We advocate for the use of standard vocabularies, unique identifiers, and flexible file formats to support this integration.

Building on this framework, we propose an openly documented registry model for QV files that assigns a unique `qv_set_id` and records a SHA-256 checksum for each release, enabling direct retrieval and verification for audit and re-analysis. Our accompanying HTML-based QV builder converts simple `key=value` statements into structured YAML and can be embedded in public, private, or commercial websites to simplify the authoring of consistent criteria (Zenodo repository). The framework is designed to support

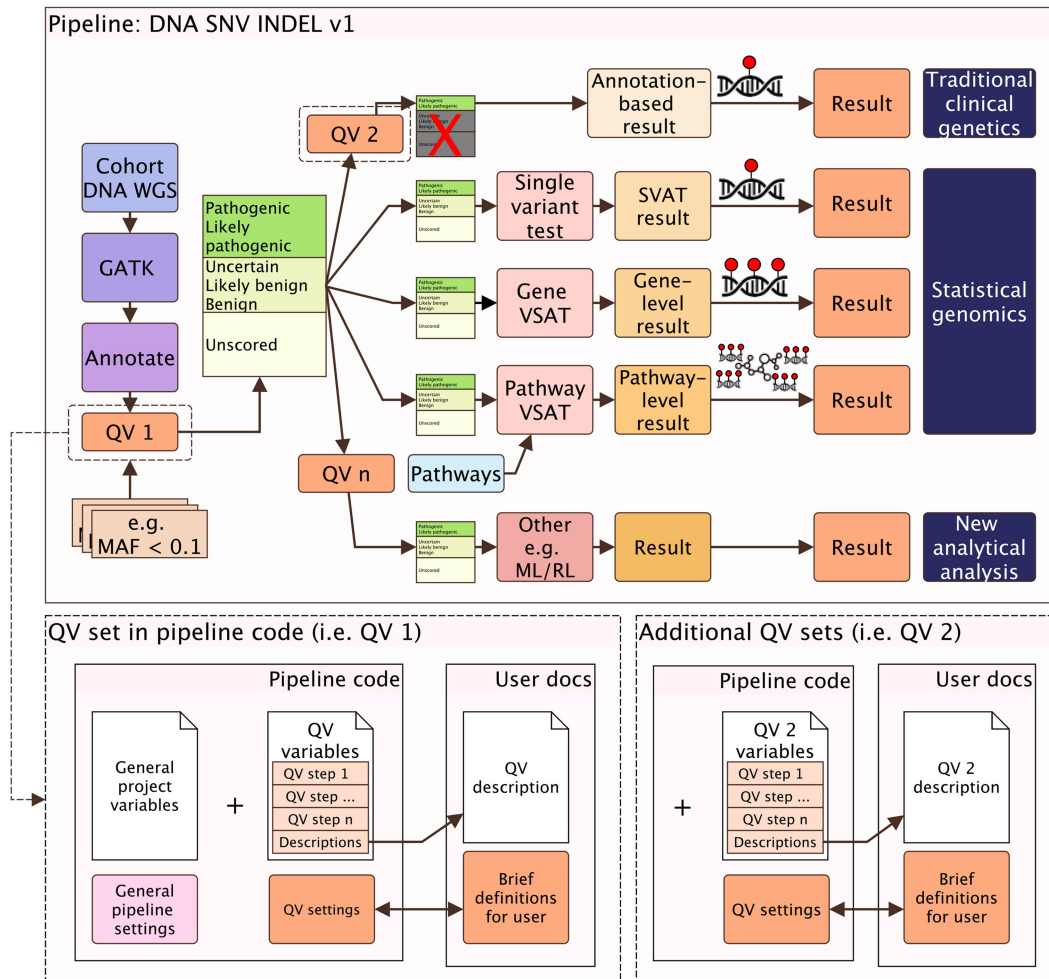
the emergence of a shared, widely adopted registry over time.

## 2 Methods

### 2.1 Implementation

The QV file provides a structured, human- and machine-readable definition of variant qualifying criteria. It is composed of five logical components that define its structure and metadata. It is portable across tools, transparent in content, and verifiable through unique identifiers and checksums. Each file is a lightweight YAML or JSON document specifying the variables and thresholds used in analysis. It can be read programmatically at runtime, e.g. using `yq` in shell-based workflows or `yaml::read_yaml()` in R, providing the same parameters that would otherwise be embedded within pipeline configurations, as illustrated in Fig. 1. The output is identical to that of the native workflow, with the added benefit of an explicit, versioned, and shareable configuration file.

- 1. **Meta:** Descriptive metadata including `qv_set_id`, title, version, author list, creation date, and tags. These fields ensure traceability and version control across analyses.
- 2. **Filters:** Simple rule-based statements that apply inclusion or exclusion logic based on variable thresholds (e.g.



**Figure 1.** Summary of the QV application for a WGS pipeline. QV1 and QV2 are applied as sequential protocol steps. In this example, QV2 differs from QV1 by retaining only likely/pathogenic variants (indicated by a X). The QV file loaded by the analysis pipeline comprises a description field (optional) and a variables field (mandatory). The QV criteria may be distributed across multiple pipeline steps.

minimum allele frequency or coverage depth). Filters can also restrict the analysis to defined genomic regions, such as a target gene panel or BED file.

- **3. Criteria:** Compound logic blocks that combine one or more conditions into interpretable rules, corresponding to concepts such as ACMG criteria or study-specific thresholds.
- **4. Notes:** Optional free-text annotations providing context, assumptions, or technical caveats.
- **5. Descriptions (optional):** Plain-language fields, such as `description_patient` and `description_ppie`, that can record patient preferences or public involvement input. These complement the technical definitions without affecting computational logic.

### 2.1.1 Example QV structure

We include an HTML-based QV builder that can be embedded in research or commercial platforms to simplify the creation of consistent, versioned criteria files (available via Zenodo repository). A minimal QV YAML file is shown in Box 1, equivalent to the configuration generated by this builder. QV files are composed of `key=value` statements, ensuring that all filtering and interpretation rules are explicit, versioned, and reproducible. In simple terms, Box 1 specifies that only variants overlapping a curated disease gene panel are retained and that variants classified as pathogenic or likely pathogenic are prioritized. It also records patient context and patient-public

#### Box 1: `qv_disease_panel_example.yaml`

```
meta:
  qv_set_id: qv_disease_panel_v1_20250828
  version: 1.0.0
  title: Disease panel filter
filters:
  region_include:
    description: >
      Restrict to curated disease gene panel
    logic: keep_if
    field: OVERLAP(targets.disease_panel.bed)
    operator: '> ='
    value: 1
criteria:
  pathogenic:
    description: >
      Variant classified as pathogenic or likely pathogenic
logic: and
conditions:
  - group: any_of: start
  - {field: CLASS, operator: '==', value: P}
  - {field: CLASS, operator: '==', value: LP}
  - group: any_of: end
meta:
  description_patient: >
    We have a strong family history of early heart attacks.
  description_ppie: >
    The PPIE group reviewed the criteria and approved them
    on 2025-08-15.
notes:
  - Gene panel file defines the target regions.
  - Additional quality filters may be added as needed.
```

involvement notes, thereby linking the technical filtering logic with its clinical and ethical rationale.

### 2.1.2 FAIR mapping and patient involvement

Each QV file includes a persistent identifier (`qv_set_id`) that links criteria across analyses and databases. The framework aligns with the Findable, Accessible, Interoperable, and Reusable (FAIR) principles of findability, accessibility, interoperability, and reusability (Wilkinson *et al.* 2016). Findability is achieved through unique identifiers; accessibility through open, human- and machine-readable YAML or JSON files; interoperability through standardized syntax (i.e. `key=value`) and semantic mappings such as Resource Description Framework (RDF) or Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) (Touré *et al.* 2023, van der Horst *et al.* 2023); and reusability through embedded metadata, checksum verification, and versioned registry records.

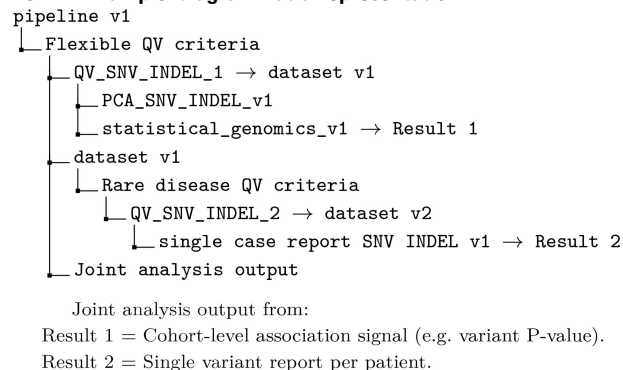
Optional metadata fields such as `description_patient` and `description_ppie` allow patient input and Public and Patient Involvement and Engagement (PPIE) feedback to be recorded in a manner appropriate to the study or application, with patient notes provided through consent-linked forms and PPIE groups offering structured review or approval of criteria within the same FAIR-compliant file.

### 2.1.3 Example QVs in WGS analysis

A typical WGS pipeline applies several QV sets sequentially, as the genetic cause of disease may stem from different variant types such as SNVs, CNVs, or structural variants. Each pass filters data for its purpose, producing both cohort-level and single-patient results within one reproducible framework (Auwera and O'Connor 2020, Li *et al.* 2025). As illustrated in Fig. 1 and Box 2, the description can be written as:

“A cohort of patient WGS data was analyzed to identify genetic determinants for phenotype X. A flexible QV set was applied using the `pipeline v1`, which implements the `QV_SNV_INDEL_1` criteria to produce the prepared dataset (`dataset v1`). This dataset was analyzed alongside other modules (e.g. `PCA_SNV_INDEL_v1` and `statistical_genomics_v1`) to derive a cohort-level association signal (Result 1). It was then re-filtered with stricter `QV_SNV_INDEL_2` criteria to identify known causal

#### Box 2: Example diagrammatic representation



variants, yielding (dataset v2) and single-patient reports (Result 2).”

## 2.2 Usage in a rare disease cohort validation study

We validated the QV framework on an in-house rare disease cohort of 940 individuals using Whole Exome Sequencing (WES) comparing a conventional manual implementation with a QV-based YAML configuration. The analysis targeted rare variants [Minor Allele Frequency (MAF)  $< 0.01$ ] in known disease genes from the Genomics England “Primary immunodeficiency or monogenic inflammatory bowel disease” panel, retrieved via PanelAppRex (Quant Group *et al.* 2025). This yielded 6026 candidate variants annotated with 376 information sources, prepared in R using the GuRu variant interpretation tool and imported from gVCF files processed by Variant Effect Predictor (VEP).

We applied the first eight American College of Medical Genetics and Genomics (ACMG) criteria for pathogenicity scoring (Richards *et al.* 2015), six of which were relevant to this cohort. The manual pipeline encoded each criterion directly, while the QV workflow read the same definitions from a YAML file. The YAML criteria included ACMG\_PS1 (known pathogenic amino acid change), ACMG\_PS3 (supporting functional evidence), ACMG\_PS5 (compound heterozygosity), and frequency- and segregation-based criteria (PM2, PM3). Criteria PS2 and PS4 were not applicable in this cohort.

## 2.3 Usage in a GWAS validation study

We next applied the QV criteria framework to a GWAS using HapMap3 Phase 3 (R3) consensus genotypes on 1397 individuals (Fairley *et al.* 2020). Again, two pipelines were executed with identical inputs and parameters: one hard-coded and one driven by the QV file. This QV set defined common GWAS thresholds: restriction to autosomal, biallelic SNPs; minimum sample call rate of 95%; variant call rate of 95%; minor allele frequency  $\geq 1\%$ ; and Hardy–Weinberg equilibrium  $p \geq 1 \times 10^{-6}$ . After quality control, variants were LD-pruned and principal components (PC1–PC10) were computed, with sex included as an additional covariate. Logistic regression under an additive model was then performed with a binary simulated phenotype using PLINK. The outputs of the two pipelines were captured and compared across each main PLINK stage. Manhattan plots, Principal Component Analysis (PCA) plots, and md5 checksums were used to confirm exact reproducibility between the hard-coded and QV-driven analyses.

For benchmarking, Message-Digest Algorithm 5 (MD5) checksums were uniquely reported for the GWAS study because PLINK output files are exactly reproducible between runs. In contrast, VCF files used in the other validation studies include variable header fields such as BCFtools view command with a timestamp, which changes with each run and alters the MD5 value. For those cases, we instead report variant count and content.

## 2.4 Usage in a WGS validation study with GIAB and Exomiser

We next applied the QV framework to a WGS trio analysis using the Genome In A Bottle Chinese Trio (HG005-HG007, PRJNA200694, GRCh38 v4.2.1) of the National Institute of Standards and Technology (Wagner *et al.* 2022). Two pipeline phases were executed with identical inputs and parameters: one hard-coded and one driven by the QV file. Both phases

applied identical QC and study filters and included a gene-panel style analysis using the pediatric disorders panel [panel 486; 3853 genes (Quant Group *et al.* 2025)]. The upstream processing used BCFtools for region restriction using BED overlap, site-level thresholds on QUAL and INFO/DP (using computed site depth from per-sample FORMAT/DP when absent), and per-sample thresholds on FORMAT/DP and FORMAT/GQ with exclusion of missing genotypes. Composite criteria were applied to require either all samples to pass or at least one sample to pass. The downstream filtered trio VCF was analyzed with Exomiser using the same trio .ped input and without using Human Phenotype Ontology (HPO) terms.

## 3 Results

### 3.1 Validation rare disease cohort case study

We validated the QV framework using WES analysis with ACMG-based criteria on a rare disease cohort of 940 individuals, comparing a conventional pipeline with parameters defined internally (QV manual) to the new external YAML-based implementation (QV yaml). As shown in Fig. 1, available as supplementary data at *Bioinformatics* online, the outputs from both methods were identical, demonstrating a 100% match. This confirmed that our framework of a standalone, shareable, QV criteria file can be imported and applied programmatically with equivalent accuracy, providing a reproducible resource that is adaptable across different pipelines and programming environments.

### 3.2 Validation in a common variant GWAS

To demonstrate the integration of the QV framework with established best practices in GWAS (Uffelmann *et al.* 2021), we validated it in a standard HapMap3 Phase 3 GWAS by again running two equivalent analyses: a conventional pipeline with parameters defined internally and a YAML-based implementation that externalized all settings. As shown in Fig. 2, available as supplementary data at *Bioinformatics* online, the Manhattan and Principal Component Analysis (PCA) plots were identical between the two methods, and the MD5 checksums of all PLINK outputs matched exactly. These results confirm that QV parameterization reproduces the original workflow precisely while improving clarity, transparency, and reusability.

### 3.3 Validation in a WGS study with GIAB and Exomiser

To demonstrate the ease and benefit of using QV parameterization in established WGS analysis pipelines, we conducted a trio validation study using the Genome In A Bottle Chinese Trio (HG005-HG007, GRCh38 v4.2.1) and the Exomiser tool for variant annotation and interpretation (Cipriani *et al.* 2020). Two equivalent analyses were run: one with hard-coded thresholds and one using an external QV YAML file specifying the same parameters. Both applied identical QC and study filters and restricted analysis to the PanelAppRex pediatric disorders panel (3853 genes). Results were identical: variant counts matched at each step, and Exomiser outputs produced the same candidate genes and variants. Figure 3, available as supplementary data at *Bioinformatics* online shows this agreement. This validation confirms that a shareable QV file reproduces the full variant interpretation workflow exactly, while aligning with established variant effect



predictors and interpretation tools (Cipriani *et al.* 2020, Holtgrewe *et al.* 2020, Riccio *et al.* 2024). Benchmarking showed that QV files introduce no computational overhead and scale equivalently to conventional implementations (Supplemental benchmark, Fig. 4, available as supplementary data at *Bioinformatics* online).

## 4 Implications

### 4.1 General applicability and reproducibility

Across validation studies, the QV framework reproduced conventional workflows in which parameters are embedded within scripts, while externalizing those same variables into a portable, shareable format. The framework itself performs no filtering, calling, annotation, or interpretation, but provides a machine-readable layer for defining and reusing the qualifying variables that underpin these analyses. It complements tools such as GATK and BCFtools for processing, Ensembl VEP, SnpEff, FAVOR, and WGSa for variant effect prediction (Riccio *et al.* 2024), and Exomiser and VarFish for interpretation (Cipriani *et al.* 2020, Holtgrewe *et al.* 2020), by making their analytic criteria explicit.

### 4.2 Scalability and interoperability with genomic tools

The validation studies, covering clinical interpretation, genome-wide association analysis, and WGS trio interpretation, demonstrate that the QV framework generalizes across distinct genomic contexts without altering analytical outcomes or adding computational overhead. The format further allows users to define, combine, and extend their own QV sets using simple declarative syntax, providing a scalable approach for reproducible genomics.

### 4.3 Traceability and confirmation of applied clinical standards

Each QV file includes a persistent identifier and checksum that can be stored in Electronic Health Record (EHR) or laboratory systems such as EPIC, Cerner, Clinisys, or REDCap. This links each patient's analysis (including any associated PPIE input) to the exact QV set used, enabling transparent, auditable, and FAIR-compliant reporting. A clinician or molecular pathologist viewing a result in EPIC or Cerner can access the linked `qv_set_id` to verify the applied standards and filtering criteria. Automated genomic reports should include these details by default, ensuring full traceability without requiring access to the pipeline. For example, if a patient asks whether their genome was screened for breast cancer due to variants in *BRCA1* or *BRCA2*, the EHR-linked report referencing “qv acmg sf v3.3 20250828.json” confirms that the ACMG secondary findings guideline (v3.3) (Miller *et al.* 2023) was applied, including its defined gene set, thresholds, version, and standard.

## 5 Summary

This paper introduces a framework for integrating qualifying variants into genomic analysis pipelines, enhancing reproducibility, interpretability and the seamless translation of research findings into clinical practice.

## Acknowledgements

We would like to thank all the patients and families who have been providing advice on SwissPedHealth and its projects, as well as the clinical and research teams at the participating institutions.

## Author contributions

Dylan Lawless (Formal analysis [equal], Methodology [equal], Supervision [equal], Writing—original draft [equal]), Ali Saadat (Methodology [equal], Writing—original draft [equal]), Mariam Ait Oumelloul (Methodology [equal], Writing—original draft [equal]), Simon Boutry (Methodology [equal], Writing—original draft [equal]), Veronika Stadler (Methodology [equal], Writing—original draft [equal]), Sabine Österle (Methodology [equal], Writing—original draft [equal]), Jan Armida (Methodology [equal], Writing—original draft [equal]), David Haerry (Methodology [equal], Writing—original draft [equal]), D. Sean Froese (Methodology [equal], Writing—original draft [equal]), Luregn J. Schlapbach (Funding acquisition [equal], Methodology [equal], Supervision [equal], Writing—original draft [equal]), and Jacques Fellay (Funding acquisition [equal], Methodology [equal], Supervision [equal], Writing—original draft [equal]).

## Supplementary data

Supplementary data is available at *Bioinformatics* online.

Conflict of interest: None declared.

## Funding

This work was supported through the grant Swiss National Science Foundation 320030\_201060/1, and NDS-2021-911 (SwissPedHealth) from the Swiss Personalized Health Network and the Strategic Focal Area “Personalized Health and Related Technologies” of the ETH Domain (Swiss Federal Institutes of Technology).

## Ethics statement

The projects were approved by the respective ethics committees of all participating centers (Cantonal Ethics Committee Bern, approval number KEK-029/11) and the study was conducted in accordance with the Declaration of Helsinki.

## Data availability

The source code and data are accessible at the Zenodo repository <https://doi.org/10.5281/zenodo.17414191>. Manuscript files are available at <https://github.com/DylanLawless/qvApp2025lawless>. The QV framework is available under the MIT licence.

## References

- Anderson CA, Pettersson FH, Clarke GM *et al.* Data quality control in genetic case-control association studies. *Nat Protoc* 2010;5: 1564–73. <https://doi.org/10.1038/nprot.2010.116>
- Auwera GVD, O'Connor BD. *Genomics in the Cloud: using Docker, GATK, and WDL in Terra*, 1st edn. Beijing, Boston, Farnham, Sebastopol, Tokyo: O'Reilly, 2020.

- Bal HE, Steiner JG, Tanenbaum AS. Programming languages for distributed computing systems. *ACM Comput Surv* 1989;21:261–322. <https://dl.acm.org/doi/10.1145/72551.72552>
- Cipriani V, Pontikos N, Arno G *et al.* An improved phenotype-driven tool for rare Mendelian variant prioritization: benchmarking exomiser on real patient whole-exome data. *Genes (Basel)* 2020;11:460. <https://doi.org/10.3390/genes11040460>
- Cirulli ET, Lasseigne BN, Petrovski S *et al.*; FALS Sequencing Consortium. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science (1979)* 2015; 347:1436–41.
- Di Tommaso P, Chatzou M, Floden EW *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;35:316–9. <https://doi.org/10.1038/nbt.3820>
- Fairley S, Lowy-Gallego E, Perry E *et al.* The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res* 2020;48:D941–7. <https://doi.org/10.1093/nar/gkz836>
- Holtgrewe M, Stolpe O, Nieminen M *et al.* VarFish: comprehensive DNA variant analysis for diagnostics and research. *Nucleic Acids Res* 2020;48:W162–9. <https://doi.org/10.1093/nar/gkaa241>
- Jansen PR, Watanabe K, Stringer S *et al.*; 23andMe Research Team. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat Genet* 2019; 51:394–403.
- Lambert SA, Gil L, Jupp S *et al.* The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat Genet* 2021;53:420–5.
- Lee JJ, Wedow R, Okbay A *et al.* Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. *Nat Genet* 2018;50:1112.
- Li MM, Datto M, Duncavage EJ *et al.* Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the association for molecular pathology, American Society of Clinical Oncology, and College of American Pathologists. *J Mol Diagn* 2017;19:4–23.
- Li Q, Wang K. Intervar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am J Hum Genet* 2017;100:267–80.
- Li X, Chen H, Selvaraj MS *et al.* A statistical framework for multi-trait rare variant analysis in large-scale whole-genome sequencing studies. *Nat Comput Sci* 2025;5:125–43.
- Miller DT, Lee K, Abul-Husn NS *et al.*; ACMG Secondary Findings Working Group. ACMG sf v3. 2 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med* 2023;25:100866.
- Mölder F, Jablonski KP, Letcher B *et al.* Sustainable data analysis with Snakemake. *F1000Res* 2021;10:33. <https://doi.org/10.12688/f1000research.29032.1>
- Olson ND, Wagner J, Dwarshuis N *et al.* Variant calling and benchmarking in an era of complete human genome sequences. *Nat Rev Genet* 2023;24:464–83.
- Pedersen BS, Brown JM, Dashnow H *et al.* Effective variant filtering and expected candidate variant yield in studies of rare human disease. *NPJ Genom Med* 2021;6:60–8.
- Pedersen BS, Quinlan AR. Vcfexpress: flexible, rapid user-expressions to filter and format VCFs. *Bioinformatics* 2025;41:btaf097. <https://doi.org/10.1093/bioinformatics/btaf097>
- Povysil G, Petrovski S, Hostyk J *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat Rev Genet* 2019;20:747–59. <https://doi.org/10.1038/s41576-019-0177-4>
- Quant Group, Boutry S, Saadat A *et al.* PanelAppRex aggregates disease gene panels and facilitates sophisticated search. medrxiv. <https://doi.org/10.1101/2025.03.20.25324319>, 2025.
- Riccio C, Jansen ML, Guo L *et al.* Variant effect predictors: a systematic review and practical guide. *Hum Genet* 2024;143:625–34. <https://doi.org/10.1007/s00439-024-02670-5>
- Richards S, Aziz N, Bale S *et al.*; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–24.
- Riggs ER, Andersen EF, Cherry AM *et al.* Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMGE) and the Clinical Genome Resource (CLINGEN). *Genet Med* 2020;22:245–57.
- Tavtigian SV, Harrison SM, Boucher KM *et al.* Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. *Hum Mutat* 2020;41:1734–7.
- Touré V, Krauss P, Gnodtke K *et al.* Fairification of health-related data using semantic web technologies in the SWISS personalized health network. *Sci Data* 2023;10:127.
- Uffelmann E, Huang QQ, Munung NS *et al.* Genome-wide association studies. *Nat Rev Methods Primers* 2021;1:59. <https://doi.org/10.1038/s43586-021-00056-9>
- van der Horst E, Unni D, Kopmels F *et al.* Bridging clinical and genomic knowledge: an extension of the SPHN RDF schema for seamless integration and fairification of omics data. 2023. <https://doi.org/10.20944/preprints202312.0373.v1>
- Wagner J, Olson ND, Harris L *et al.* Benchmarking challenging small variants with linked and long reads. *Cell Genom* 2022;2:100128. <https://doi.org/10.1016/j.xgen.2022.100128>
- Wand H, Lambert SA, Tamburro C *et al.* Improving reporting standards for polygenic scores in risk prediction studies. *Nature* 2021; 591:211–9.
- Wilkinson MD, Dumontier M, Aalbersberg IJJ *et al.* The fair guiding principles for scientific data management and stewardship. *Sci Data* 2016;3:160018–9.