Application of qualifying variants for genomic analysis

Dylan Lawless*[1], Ali Saadat[2], Mariam Ait Oumelloul[2], Simon Boutry[2], Veronika Stadler[1], Sabine Österle[3], Jan Armida[3], David Haerry[4], D. Sean Froese[5], Luregn J. Schlapbach[1], and Jacques Fellay[2]

[1]Department of Intensive Care and Neonatology, University Children's Hospital Zürich, University of Zürich, Switzerland.
[2]Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Switzerland.
[3]SPHN Data Coordination Center, SIB Swiss Institute of Bioinformatics, Basel, Switzerland.
[4]Positive Council, Zürich, Switzerland.
[5]Division of Metabolism and Children's Research Center, University Children's Hospital Zürich, University of Zurich, Zurich, Switzerland.

October 15, 2025

*Addresses for correspondence: Dylan.Lawless@kispi.uzh.ch

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. While best practices follow variant classification standards and standardised workflows, a unified framework to integrate and optimise QVs for advanced applications is missing.

Results:

Our aim is to embed the concept of a "QV" into the genomic analysis vernacular, moving beyond a single filtering step. By decoupling QV criteria from other pipeline variables and code, our approach facilitates easier discussion and application. Our framework, with its new terminology and reference model, offers a flexible approach for integrating QVs into analysis pipelines, thereby enhancing reproducibility, interpretability, and interdisciplinary communication. A validation case study implementing ACMG criteria in a disease cohort shows that our approach matches conventional methods while offering improved clarity and scalability.

Availability:

The source code and data are accessible at https://github.com/DylanLawless/qv2025lawless. The QV file used in this work is available from https://doi.org/10.5281/zenodo.15105594 (`qv_acmg_svnindel_criteria_20250225.yaml`). The QV framework is available under the MIT licence, and the dataset will be maintained for at least two years following publication.

Acronyms

1 Introduction

Qualifying variant (QV)s 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 (1–5) and standardised workflows (6–8). However a unified framework for QVs is lacking, despite the recognised benefits of similar initiatives, such as Polygenic Risk Score (PRS) reporting standards (9; 10). For instance, tools like vcfexpress (11) enable flexible, rapid filtering and formatting of VCF files using user-defined expressions. The application of independently defined QV criteria would complement such tools. This role is particularly important for reproducibility across distributed computing environments (12) and would also integrate with workflow managers such as Snakemake (13) or Nextflow (14), streamlining genomic processing tasks.

The criteria for QV selection vary by application. For example, Genome Wide Association Study (GWAS) may focus on common variants, while clinical analyses usually target rare or known pathogenic variants. Previous studies have demonstrated the utility of QVs (15; 16), yet no common approach exists. Here, we detail four typical applications of QV sets:

1. **QV passing Quality Control (QC) only**: Generates large datasets (e.g. $>$ 500,000 variants per subject) for GWAS or initial Whole Genome Sequencing (WGS) pre-processing.

2. **Flexible QV**: Balances between QC and false positives, yielding intermediate datasets (e.g. fewer than 100,000 variants per subject) for uses such as rare variant association testing.

3. **QV for rare disease**: Applies stringent filtering to produce smaller datasets (e.g. $<$ 1,000 variants per subject), targeting known genes or single causal variants.

4. **Known disease panel QV set**: Focuses on well-established gene panels with pathogenic variants (e.g. the American College of Medical Genetics and Genomics (ACMG) Secondary Findings (SF) set) for clinical reporting (17).

These examples illustrate a few common applications without providing an exhaustive classification of all possible QV uses. The careful selection and categorisation of QVs are thus critical for accurate reporting and reproducibility, sometimes even more so than the choice of the analysis pipeline itself (18).

As WGS becomes standard for large cohorts (19; 20), 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 QC results, and post-calling annotations, highlighting the need for a clear, unified approach.

We propose treating the QV as a standalone entity, independent from other pipeline variables. We suggest structured human- and machine-readable criteria, aligned with FAIR principles (21) to facilitate integration across databases (22; 23). We advocate for the use of standard vocabularies, unique identifiers, and flexible file formats to support this integration.

2 Methods

2.1 Implementation

Implementation configurations and roles within analysis pipelines include, for example: theoretical pipelining of QV sets, establishing public or standardised QV sets for specific analytical scenarios, and recognition that QVs are integral throughout the analysis pipeline rather than confined to a single end-stage. We introduce a simple framework for the effective use of QV protocols, comprising four components as illustrated in **Figure 1 (A)**:

- **1. Variables**: The criteria variables sourced as part of the pipeline (see **Box 2**).
- **2a. Technical description**: An optional narrative detailing each step within the overall QV set (see **Box 2**).

- **2b. Public and Patient Involvement and Engagement (PPIE) description**: An optional narrative providing a patient-focused interpretation of the protocol, incorporating preferences and priorities.
- **3. QV set ID**: A unique identifier that links analysis records.
- **4. Source code**: The implementation of the variables file within the pipeline code, for example through custom scripts or workflow managers.

We propose the QV set ID as a unique identifier linking variant sets used in analyses. This facilitates integration into databases, by representing data in formats such as Resource Description Framework (RDF) schemas (23), and allows for features including Secure Hash Algorithm 256 (SHA-256) hash functions, Universally Unique Identifier (UUID)s, semantic combinations, Internationalised Resource Identifier (IRI) incorporation, registry-based allocation, and standard mapping such as Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT). The results can be used alongside other genomic-specific concepts spanning from sample processing to the sequencing run (22).

This framework efficiently manages QV-specific variables (e.g. allele frequency thresholds) separately from general pipeline settings, ensuring clarity and specificity. Its versatile format supports applications across genomic analyses and by linking the QV set ID to both results and raw data sources in a database for downstream interpretation and reporting.

2.2 Example application of qualifying variants in WGS analysis

Multiple QV protocols can be combined to generate progressively filtered datasets tailored to specific analytical needs. Often, different QV sets are applied sequentially, with the final outcomes merged to address distinct objectives. For instance, a comprehensive analysis pipeline might integrate:

- `QV SNV/INDEL` Single Nucleotide Variant / Insertion Deletion (SNV/INDEL),
- `QV CNV` Copy Number Variant (CNV),
- `QV structural variation`,
- `QV rare disease known`, and

- `QV statistical association QC`.

The final analysis yields (1) a joint cohort disease association (e.g. variant P-values) and (2) individual single-case results (e.g. clinical genetics diagnosis for a patient) (24; 25). As an example, in **Figure 1 (A)** we focus on a SNV/INDEL pipeline employing two QV sets: `QV SNV INDEL 1` for flexible cohort-level filtering, and `QV SNV INDEL 2` for stricter filtering in subsequent single-case analysis. The pipeline is illustrated in **Box 1**, and can be summarised as follows:

"A cohort of patient WGS data was analysed to identify genetic determinants for phenotype X. Initially, a flexible QV set was applied using the `pipeline DNA SNV INDEL v1`, which implements the `QV SNV INDEL 1` criteria to produce the prepared dataset (`dataset DNA SNV INDEL v1`). This dataset was then analysed alongside other modules (e.g. `PCA SNV INDEL v1` and `statistical genomics v1`) to derive a cohort-level association signal (Result 1). Next, the same prepared dataset was re-filtered with the stricter `QV SNV INDEL 2` criteria to identify known causal variants for each patient, yielding the final dataset (`dataset DNA SNV INDEL v2`) and resulting in individual case reports (Result 2)."

Box 1: Example diagrammatic representation

```
pipeline DNA SNV INDEL v1
└─ Flexible QV criteria
   ├─ QV SNV INDEL 1 → dataset DNA SNV INDEL v1
   │  ├─ PCA SNV INDEL v1
   │  └─ statistical genomics v1 → Result 1
   ├─ dataset DNA SNV INDEL v1
   │  └─ Rare disease QV criteria
   │     └─ QV SNV INDEL 2 → dataset DNA SNV INDEL v2
   │        └─ single case report SNV INDEL v1 → Result 2
   └─ Joint analysis output
```

Joint analysis output from:
Result 1 = Cohort-level association signal (e.g. variant P-value).
Result 2 = Single variant report per patient.

2.3 Usage in a rare disease Validation Study

In a validation study, we demonstrate the use of our QV criteria framework compared to the conventional manual approach. This analysis was performed on an in-house rare disease cohort of 940 individuals, which had been pre-processed for QC. We used genome-wide set of variants which was filtering to target rare varaints (Minor Allele Frequency (MAF) < 0.01) restricted to known disease genes based on the Genomics England panel "Primary immunodeficiency or monogenic inflammatory bowel disease," retrieved using our PanelAppRex R repository (https://github.com/DylanLawless/PanelAppRex) (26). This provided us with a prepared dataset of 6026 candidate variants annotated with 376 information sources. The dataset was prepared in R using GuRu, our variant interpretation tool that consolidates all annotation sources and scores variants as candidate causal, and was imported from gVCF format as output by Variant Effect Predictor (VEP).

We selected the first eight ACMG criteria for assigning pathogenicity scores to variants (1); six of these were relevant for this cohort. First, the analysis was performed manually by hard-coding each criterion in the pipeline script, reflecting a typical workflow. Second, the same criteria were imported from the QV YAML file for the new framework approach, using the file "qv acmg svnindel criteria 20250225.yaml" (see **Box 2** or https://doi.org/10.5281/zenodo.15105594). The outputs from both methods were captured and compared.

Additional details of the YAML criteria in this QV set included definitions for `ACMG_PS1` (identifying previously established pathogenic amino acid changes), `ACMG_PS3` (supporting functional studies with matching inheritance patterns), and `ACMG_PS5` (covering compound heterozygosity with high-impact variants). The criteria for `ACMG_PM2` and `ACMG_PM3` assess variant frequency and in trans occurrences, respectively, while `PS2` and `PS4` were not applicable to this cohort.

Box 2: qv_files/acmg_criteria.yaml

```
qv_set_id: qv_acmg_svnindel
criteria:
  ACMG_PVS1:
    conditions:
      - field: IMPACT
        operator: '=='
        value: HIGH
      - group: any_of:start
      - field: genotype
        operator: '=='
        value: 2
      - field: Inheritance
        operator: '=='
        value: AD
      - field: comp_het_flag
        operator: '=='
        value: 1
      - group: any_of:end
    description: 'Null variants (IMPACT == HIGH) in genes where
    loss of function causes disease. Includes homozygous variants,
    compound heterozygous cases, or dominant inheritance.
    Warning to phase check compound heterozygosity.'
    logic: and

...
shasum -a 256 acmg_criteria.yaml | fold -w 32
d91fde41a5fff48631adecba38773d61
9ae8cd5cff9b9b42ef7f5efbd6bbfcdf
acmg_criteria.yaml
```

2.4 Usage in a GWAS validation study

We next applied the QV criteria framework to a genome-wide association study using HapMap3 Phase 3 (R3) consensus genotypes. The analysis was performed twice: first as a conventional hard coded pipeline (Phase 1), and second using an externalised QV

YAML configuration file (Phase 2). Both pipelines used the same data and processing stages, ensuring comparability.

The hard coded implementation embedded all filtering and analysis thresholds directly in the bash scripts, while the QV-based implementation sourced them from the file "qv_gwas_common_v1_20250826.yaml" (see https://doi.org/10.5281/zenodo.XXXXXXX). This QV set defines 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 PLINK.

The outputs of the two pipelines were captured and compared across each main PLINK stage: step1 (autosomal biallelic SNPs), step2 (sample call rate), step3 (variant call rate), step4 (minor allele frequency), step5 (Hardy–Weinberg equilibrium), and step6 (association results). Manhattan plots, PCA plots, and `md5` checksums were used to confirm exact reproducibility between the hard coded and QV-driven analyses.

2.5 Usage in a WGS validation study with GIAB and Exomiser

We validated the QV approach on the Genome In A Bottle Chinese Trio (HG005-HG007, PRJNA200694, GRCh38 v4.2.1) using a two-phase trio pipeline that differs only in how parameters are provided: Phase 1 hard coded thresholds in scripts, while Phase 2 externalised the same thresholds in a QV YAML file. Both phases applied identical QC and study filters and additionally demonstrated gene-panel style analysis using the PanelAppRex paediatric disorders panel (panel_486; 3,853 genes). The upstream configuration demonstrated a range of QV settings used, including region restriction by BED overlap, site-level thresholds on QUAL and INFO/DP with computed site depth from per-sample FORMAT/DP if absent, per-sample thresholds on FORMAT/DP and FORMAT/GQ with exclusion of missing GT, and composite criteria that require either all samples to pass or at least one sample to pass. Upstream processing used `bcftools` for region filtering, tag filling, site depth range, and per-sample filters. The filtered trio **vcf!** (**vcf!**) was analysed with Exomiser using a trio `.ped` and no **hpo!** (**hpo!**).

3 Results

3.1 Validation rare disease case study

We validated our QV protocol using ACMG-based criteria for a rare disease cohort of 940 individuals. We then conducted the variant classification using two approaches: a conventional manual method with hard-coded criteria, and our new YAML-based implementation. As shown in **Figure 1 (B)**, 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

we dont need redundant repetition of values and steps 1-6. just leave it in methods.

We validated the QV approach in a standard HapMap3 Phase 3 GWAS by running two equivalent analyses: a hard coded pipeline and a QV YAML driven pipeline that externalised all quality control and analysis thresholds. Both used autosomal biallelic Single Nucleotide Polymorphism (SNP)s with sample call rate $\geq 95\%$, variant call rate $\geq 95\%$, minor allele frequency $\geq 1\%$, and Hardy–Weinberg equilibrium $p \geq 1 \times 10^{-6}$, followed by LD pruning and inclusion of sex and PC1-PC10 as covariates in logistic regression. As shown in **Figure 2** panels A and B, the Manhattan and PCA plots were indistinguishable between methods. Panel C shows identical `md5` checksums for the main PLINK outputs across stages: step 1 (autosomal biallelic SNPs), step 2 (sample call rate), step 3 (variant call rate), step 4 (minor allele frequency), step 5 (Hardy–Weinberg equilibrium), and step 6 (association results). These results demonstrate exact reproducibility of the hard coded workflow using a shareable QV criteria file, confirming that QV parameterisation preserves analytic intent while improving clarity and reuse.

3.3 Validation in a WGS study with GIAB and Exomiser

We conducted the trio analysis using a standard filtering approach and the popular Exomiser tool for variant annotation and interpretation pipeline. Again we tested the conventional manual hard-coded criteria to our new YAML-based implementation

Both phases produced identical outputs. Upstream file-level counts matched at each step and Exomiser annotation and filtering metrics were the same in both phases, resolving to the same candidate set and reported variants. **Figure 3 (A)** summarises upstream processing counts by file, **(B)** shows equality of Exomiser metrics, and **(C)** presents the key variant fields used for reporting. Together these results confirm that a shareable QV file reproduces a hard coded workflow exactly while improving clarity and reuse on top of existing popular methods.

3.4 Implications

In the validation study, we applied ACMG criteria for variant interpretation. In clinical genetics, for instance, the resulting output can be used to retrieve candidate pathogenic variants using ACMG scoring methods (1; 5). Application of additional QV sets, such as the widely used ACMG SF set for clinical reporting (17), can be used to confirm any secondary findings.

In a clinical setting it is necessary to bridge the gap between technical detail and lay understanding. By explicitly documenting variant qualifying criteria and making QV data accessible, our framework builds trust and supports meaningful PPIE (27). The QV file adapts by integrating the main criteria variables with optionally dedicated fields for both technical description and PPIE description. This approach captures the analysis intent defined by the QV set creator and embeds patient preferences from the start.

For example, patient preferences recorded in the PPIE description can be automatically incorporated into a genetic report without additional interpretation, ensuring clarity and consistency throughout the analysis. This transparency guarantees that both experts and laypersons receive information in a format suited to their needs, thereby improving diagnostic traceability and accelerating the translation of genetic research into clinical practice.

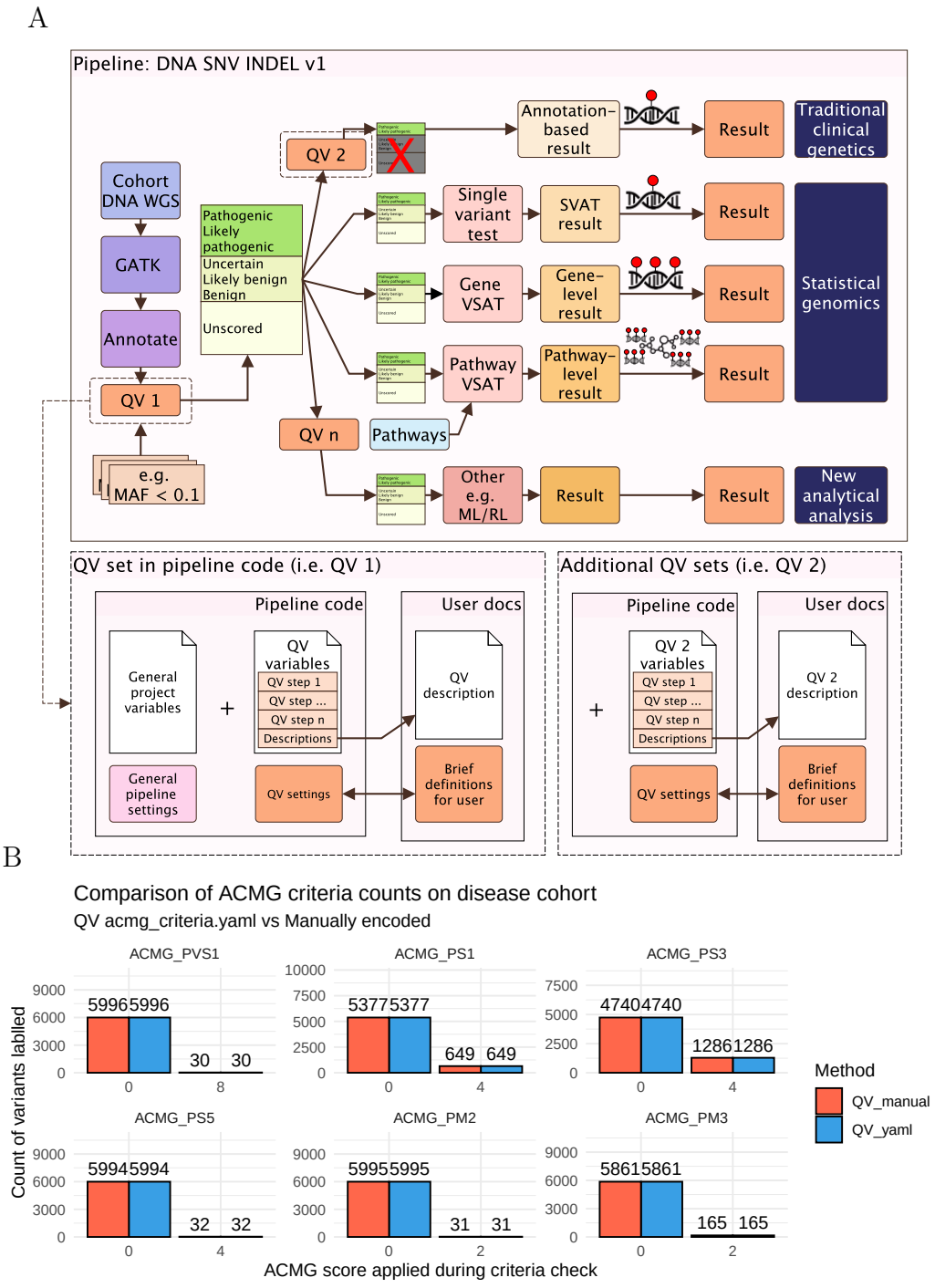


Figure 1: Summary of the QV application for a WGS pipeline. In panel (A), QV1 and QV2 are presented as sequentially piped protocol steps. In this example, QV2 differs from QV1 by retaining only likely/pathogenic variants (indicated by a red X). The QV file loaded by the analysis pipeline comprise a description field (optional) and a variables field (mandatory). The QV criteria may be spread throughout the pipeline. (B) Validation case study using an ACMG criteria subset, demonstrating a 100% match between manually encoded and standalone YAML-based QV (`qv_files/acmg_criteria.yaml`) for assigning pathogenicity scores.

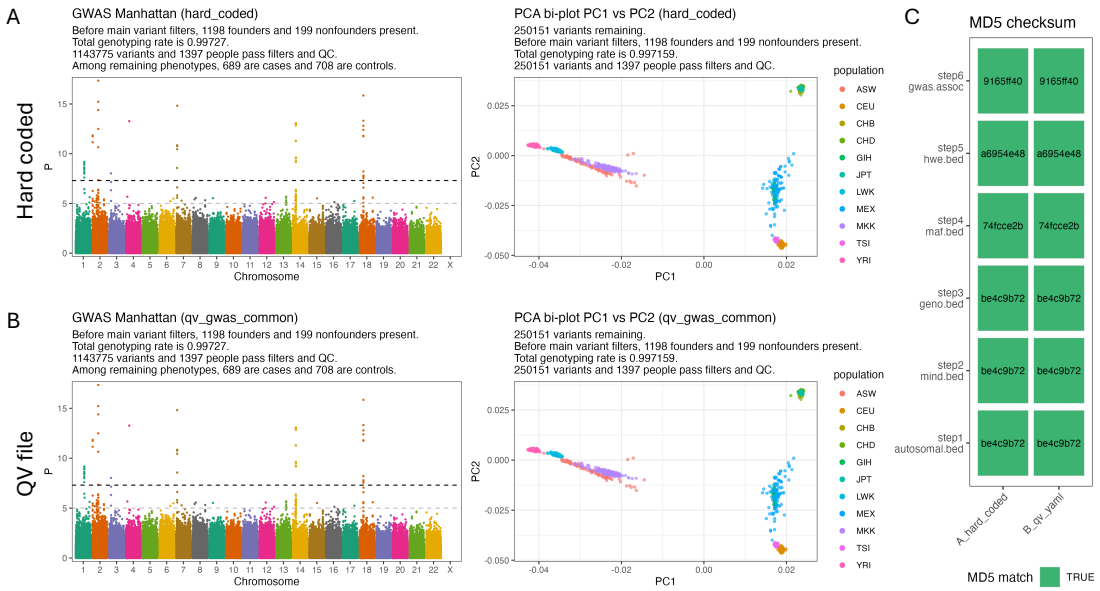
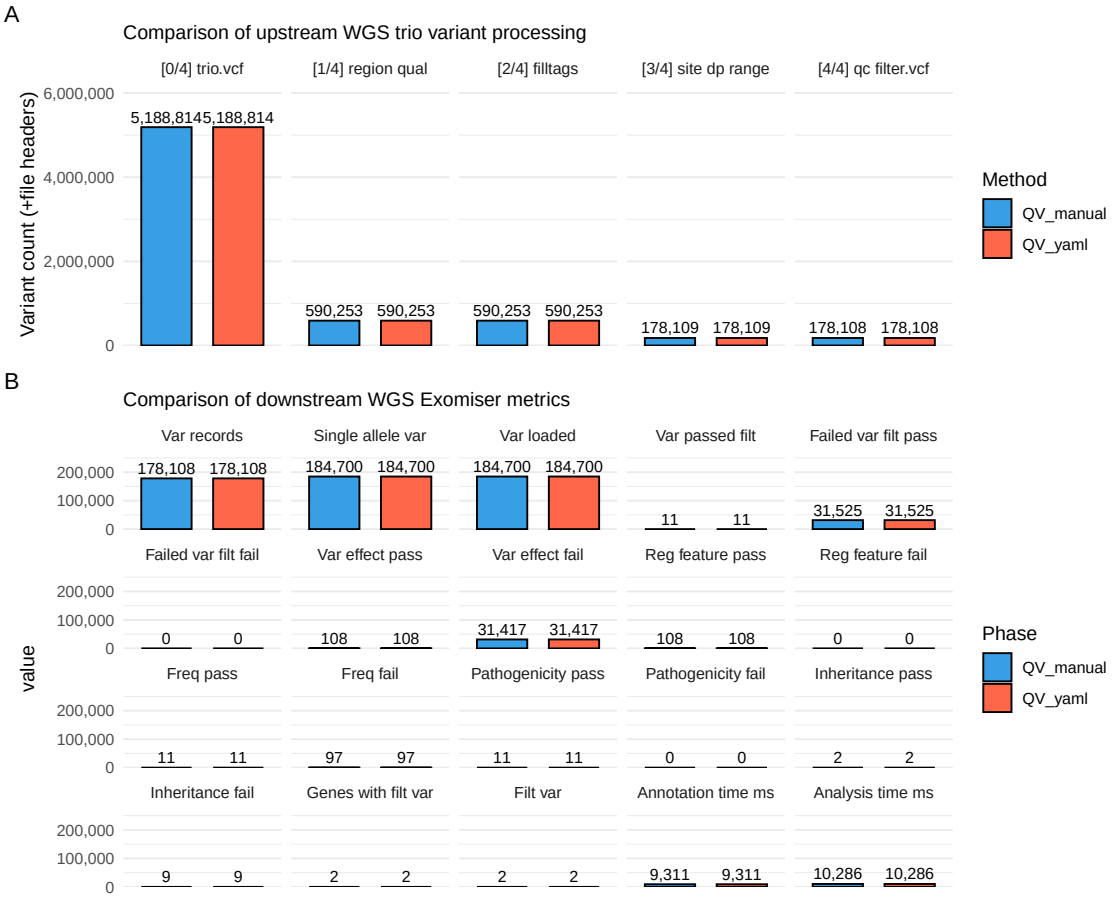


Figure 2: Reproducible GWAS using QV parameterisation. (A) GWAS of simulated binary phenotypes in HapMap3 Phase 3 (R3) using a hard-coded pipeline. Shown are the Manhattan plot of logistic regression results (left) and the correction for population structure with principal component analysis (PC1 vs PC2, right). (B) Identical GWAS using a QV YAML configuration file to supply quality-control and analysis thresholds. Manhattan and PCA results are indistinguishable from panel A. (C) Verification of reproducibility. MD5 checksums of the main PLINK outputs — step1 (autosomal, biallelic SNPs), step2 (sample call rate), step3 (variant call rate), step4 (minor allele frequency), step5 (Hardy–Weinberg equilibrium), and step6 (association results) — are identical between phases A and B. The QV file encodes these thresholds (sample call rate $\geq$95%, variant call rate $\geq$95%, MAF $\geq$1%, HWE p $\geq$1e-6, autosomal biallelic SNPs only) together with covariates (sex and PC1-PC10) and logistic regression settings. This demonstrates that a shareable QV file reproduces hard-coded pipelines exactly while improving transparency and reusability.



phase	MOI	Effect	HGVS	Genotype
QV_manual	AD	frameshift	RYR2:ENST00000366574.7:c.6493dup:p.(Leu2165Profs*44)	0/1:./.:./.
QV_yaml	AD	frameshift	RYR2:ENST00000366574.7:c.6493dup:p.(Leu2165Profs*44)	0/1:./.:./.
QV_manual	AR	missense	TTN:ENST00000589042.5:c.36982A>G:p.(Thr12328Ala)	1|1:0/1:0/1
QV_yaml	AR	missense	TTN:ENST00000589042.5:c.36982A>G:p.(Thr12328Ala)	1|1:0/1:0/1

Figure 3: Validation of a trio Exomiser pipeline using QV parameterisation. Panels A, B, and C correspond to upstream processing counts, downstream Exomiser metrics, and final variants detected, respectively.

4 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.

5 Funding

This project was supported through the grant Swiss National Science Foundation 320030_201060, 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).

6 Acknowledgements

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.

7 Contributions

DL designed the work and contributed to the manuscript. AS, SB, VS, DH, SÖ, JA contributed to the manuscript. JF, SF, LJS supervised the work, manuscript, and applied for funding.

8 Competing interests

The authors declare no competing interests.

9 Ethics statement

Summary statistics were used from studies which have been previously reported and 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.

References

[1] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al. 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. *Genetics in medicine*, 17 (5):405–423, 2015.

[2] Marilyn M Li, Michael Datto, Eric J Duncavage, Shashikant Kulkarni, Neal I Lindeman, Somak Roy, Apostolia M Tsimberidou, Cindy L Vnencak-Jones, Daynna J Wolff, Anas Younes, 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. *The Journal of molecular diagnostics*, 19(1):4–23, 2017.

[3] Quan Li and Kai Wang. Intervar: clinical interpretation of genetic variants by the 2015 acmg-amp guidelines. *The American Journal of Human Genetics*, 100 (2):267–280, 2017.

[4] Erin Rooney Riggs, Erica F Andersen, Athena M Cherry, Sibel Kantarci, Hutton Kearney, Ankita Patel, Gordana Raca, Deborah I Ritter, Sarah T South, Erik C Thorland, 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). *Genetics in Medicine*, 22(2):245–257, 2020.

[5] Sean V Tavtigian, Steven M Harrison, Kenneth M Boucher, and Leslie G Biesecker. Fitting a naturally scaled point system to the acmg/amp variant classification guidelines. *Human mutation*, 41(10):1734–1737, 2020.

[6] Brent S Pedersen, Joe M Brown, Harriet Dashnow, Amelia D Wallace, Matt Velinder, Martin Tristani-Firouzi, Joshua D Schiffman, Tatiana Tvrdik, Rong Mao, D Hunter Best, et al. Effective variant filtering and expected candidate

variant yield in studies of rare human disease. *NPJ Genomic Medicine*, 6(1):1–8, 2021.

[7] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564–1573, 2010. URL https://doi.org/10.1038/nprot.2010.116.

[8] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina De Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59, 2021. URL https://doi.org/10.1038/s43586-021-00056-9.

[9] Hannah Wand, Samuel A Lambert, Cecelia Tamburro, Michael A Iacocca, Jack W O'Sullivan, Catherine Sillari, Iftikhar J Kullo, Robb Rowley, Jacqueline S Dron, Deanna Brockman, et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature*, 591(7849):211–219, 2021.

[10] Samuel A Lambert, Laurent Gil, Simon Jupp, Scott C Ritchie, Yu Xu, Annalisa Buniello, Aoife McMahon, Gad Abraham, Michael Chapman, Helen Parkinson, et al. The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nature Genetics*, 53(4):420–425, 2021.

[11] Brent S Pedersen and Aaron R Quinlan. Vcfexpress: flexible, rapid user-expressions to filter and format VCFs. *Bioinformatics*, 41(3): btaf097, March 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf097. URL https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btaf097/8051444.

[12] Henri E. Bal, Jennifer G. Steiner, and Andrew S. Tanenbaum. Programming languages for distributed computing systems. *ACM Computing Surveys*, 21(3): 261–322, September 1989. ISSN 0360-0300, 1557-7341. doi: 10.1145/72551.72552. URL https://dl.acm.org/doi/10.1145/72551.72552.

[13] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with Snakemake. *F1000Research*, 10:33, January 2021. ISSN 2046-1402. doi:

10.12688/f1000research.29032.1. URL https://f1000research.com/articles/10-33/v1.

[14] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, April 2017. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.3820. URL https://www.nature.com/articles/nbt.3820.

[15] Gundula Povysil, Slavé Petrovski, Joseph Hostyk, Vimla Aggarwal, Andrew S. Allen, and David B. Goldstein. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nature Reviews Genetics*, 20(12):747–759, 2019. doi: 10.1038/s41576-019-0177-4. URL https://doi.org/10.1038/s41576-019-0177-4.

[16] Elizabeth T Cirulli, Brittany N Lasseigne, Slavé Petrovski, Peter C Sapp, Patrick A Dion, Claire S Leblond, Julien Couthouis, Yi-Fan Lu, Quanli Wang, Brian J Krueger, et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science*, 347(6229):1436–1441, 2015.

[17] David T Miller, Kristy Lee, Noura S Abul-Husn, Laura M Amendola, Kyle Brothers, Wendy K Chung, Michael H Gollob, Adam S Gordon, Steven M Harrison, Ray E Hershberger, et al. 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). *Genetics in Medicine*, 25(8): 100866, 2023.

[18] Nathan D Olson, Justin Wagner, Nathan Dwarshuis, Karen H Miga, Fritz J Sedlazeck, Marc Salit, and Justin M Zook. Variant calling and benchmarking in an era of complete human genome sequences. *Nature Reviews Genetics*, 24(7): 464–483, 2023.

[19] James J Lee, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, Richard Karlsson Linnér, et al. Gene discovery and polygenic prediction from a 1.1-million-person gwas of educational attainment. *Nature genetics*, 50(8):1112, 2018.

[20] Philip R Jansen, Kyoko Watanabe, Sven Stringer, Nathan Skene, Julien Bryois, Anke R Hammerschlag, Christiaan A de Leeuw, Jeroen S Benjamins, Ana B

Muñoz-Manchado, Mats Nagel, et al. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nature genetics*, 51(3):394–403, 2019.

[21] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[22] Eelke van der Horst, Deepak Unni, Femke Kopmels, Jan Armida, Vasundra Touré, Wouter Franke, Katrin Crameri, Elisa Cirillo, and Sabine Österle. Bridging clinical and genomic knowledge: An extension of the sphn rdf schema for seamless integration and fairification of omics data. 2023.

[23] Vasundra Touré, Philip Krauss, Kristin Gnodtke, Jascha Buchhorn, Deepak Unni, Petar Horki, Jean Louis Raisaro, Katie Kalt, Daniel Teixeira, Katrin Crameri, et al. Fairification of health-related data using semantic web technologies in the swiss personalized health network. *Scientific Data*, 10(1):127, 2023.

[24] Geraldine Van der Auwera and Brian D. O'Connor. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. O'Reilly, Beijing Boston Farnham Sebastopol Tokyo, first edition edition, 2020. ISBN 978-1-4919-7519-0 978-1-4919-7516-9 978-1-4919-7512-1.

[25] Xihao Li, Han Chen, Margaret Sunitha Selvaraj, Eric Van Buren, Hufeng Zhou, Yuxuan Wang, Ryan Sun, Zachary R McCaw, Zhi Yu, Min-Zhi Jiang, et al. A statistical framework for multi-trait rare variant analysis in large-scale whole-genome sequencing studies. *Nature Computational Science*, pages 1–19, 2025.

[26] Dylan Lawless. PanelAppRex aggregates disease gene panels and facilitates sophisticated search. March 2025. doi: 10.1101/2025.03.20.25324319. URL http://medrxiv.org/lookup/doi/10.1101/2025.03.20.25324319.

[27] Zoë Slote Morris, Steven Wooding, and Jonathan Grant. The answer is 17 years, what is the question: understanding time lags in translational research. *Journal of the Royal Society of Medicine*, 104(12):510–520, December 2011. ISSN 0141-0768, 1758-1095. doi: 10.1258/jrsm.2011.110180. URL https://journals.sagepub.com/doi/10.1258/jrsm.2011.110180.