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 17, 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 the Zenodo repository (`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

ACMG American College of Medical Genetics and Genomics 5

CNV Copy Number Variant 8

GWAS Genome Wide Association Study 4

HPO Human Phenotype Ontology 11

MAF Minor Allele Frequency 9

PCA Principal Component Analysis 10

PRS Polygenic Risk Score 4

QC Quality Control 4

QV Qualifying variant 4

RDF Resource Description Framework 8

SF Secondary Findings 5

SNV/INDEL Single Nucleotide Variant / Insertion Deletion 8

SNOMED CT Systematized Nomenclature of Medicine-Clinical Terms . . . 8

VCF Variant Call Format 11

VEP Variant Effect Predictor 10

WGS Whole Genome Sequencing 4

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. Treating QV criteria as an external, identifiable parameter layer complements such tools by externalising the thresholds and logic they consume. This role is particularly important for reproducibility across distributed computing environments (12) and integrates cleanly 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 introduce the QV as a standalone entity, independent from other pipeline variables. Structured human- and machine-readable criteria, aligned with FAIR principles (21), facilitate integration across databases (22; 23). 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). While no central database is released here, 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 designed to be portable across tools, transparent in content, and verifiable through unique identifiers and checksums. Each file is composed of five

logical components that define its structure and metadata, as illustrated in **Figure 1 (A)**.

- **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 (for example, 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.

Example QV structure

A minimal QV YAML is shown below, equivalent to the configuration produced by the QV builder.

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.
```

The HTML-based QV builder generates such files interactively from `key=value` statements, supporting both technical and human-readable documentation. The framework manages QV-specific variables (for example, allele frequency thresholds or gene panel boundaries) separately from general pipeline settings, ensuring clarity and reproducibility.

FAIR mapping

Each QV file carries a persistent `qv_set_id` that uniquely links variant criteria across analyses and databases. Identifiers can be represented in semantic or registry formats such as Resource Description Framework (RDF) schemas (23) or mapped to controlled vocabularies including Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) for audit and interoperability across systems (22). The framework was designed to align with the FAIR principles of findability, accessibility, interoperability, and reusability (21). Findability is ensured through unique identifiers; accessibility through open, human- and machine-readable YAML files; interoperability through standardised `key=value` syntax and semantic compatibility; and reusability through embedded metadata, checksum verification, and versioned registry records.

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 2**, 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 2: 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 cohort validation study

In a validation study, we demonstrated 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 variants (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" Zenodo repository). 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.

2.4 Usage in a GWAS validation study

We next applied the QV criteria framework to a GWAS using HapMap3 Phase 3 (R3) consensus genotypes (27). 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, 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.

2.5 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) (28). 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 paediatric disorders panel (panel 486; 3,853 genes, (26)). The upstream processing used BCFtools for region restriction by BED overlap, site-level thresholds on QUAL and INFO/DP (with computed site depth from per-sample FORMAT/DP if 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 Variant Call Format (VCF) was analysed with Exomiser using the same trio `.ped` input and without Human Phenotype Ontology (HPO) terms.

3 Results

3.1 Validation rare disease cohort case study

We validated the QV framework using ACMG-based criteria in 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 **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

To demonstrate the integration of the QV framework with established best practices in GWAS (29), 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 externalised all settings. As shown in **Figure 2**, the Manhattan and PCA plots were identical between the two methods, and the `md5` checksums of all PLINK outputs matched exactly. These results confirm that QV parameterisation 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 parameterisation 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 (30). Two equivalent analyses were performed: one using a conventional pipeline with parameters defined internally and another using an external YAML configuration that specified the same thresholds. Both applied identical QC and study filters and restricted analysis to the PanelAppRex paediatric disorders panel, comprising 3,853 genes. Outputs were identical across phases: upstream file-level variant counts matched at every filtering step, and Exomiser annotation and filtering metrics yielded the same candidate genes and reported variants. **Figure 3** illustrates this agreement: panel **(A)** summarises upstream processing counts by file, panel **(B)** compares Exomiser metrics, and panel **(C)** shows the key variant fields for the two variants identified. 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 (30–32).

3.4 Computational benchmark

No significant runtime difference is expected between the traditional and QV-based pipelines, as both read equivalent variables from different sources. In the WGS trio pipeline (validation study 3), the pre-processing execution times of variant filtering,QC, gene panel selection were nearly identical (16-17 seconds, median difference ~0.5 seconds favouring the QV yaml) (**Figure 4**). An incidental 5 second delay was observed from Singularity container initialisation when launching the `yq` utility to read the

YAML file (step 0). This behaviour is specific to our implementation environment and not inherent to the QV framework.

3.5 Implications

3.5.1 General applicability and reproducibility

Across the validation studies, the QV framework achieved exact reproducibility of conventional workflows in which parameters are embedded directly within pipeline scripts, while externalising those same variables into a portable, shareable format. This confirms that the QV framework itself does not perform variant filtering, calling, annotation, or interpretation; instead, it provides a formal, machine-readable layer for defining and reusing the qualifying variables that underlie these analyses. It complements existing processing software such as GATK and BCFtools, variant effect predictors such as Ensembl VEP, SnpEff, FAVOR, and WGSA (31), and interpretation platforms such as Exomiser and VarFish (30; 32), by making their operational criteria explicit, auditable, and reproducible.

3.5.2 Scalability and interoperability with genomic tools

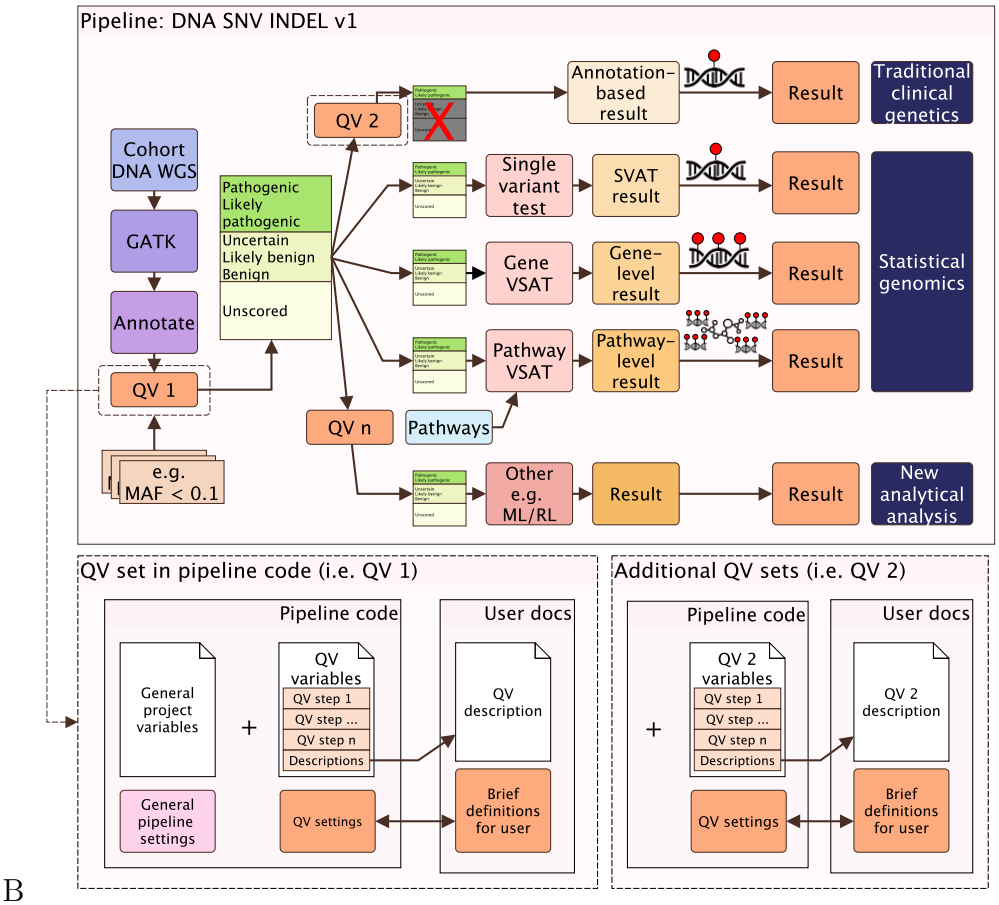
The validation studies, covering clinical interpretation, genome-wide association analysis, and WGS trio interpretation, demonstrate that the QV framework generalises across distinct genomic contexts without altering analytical outcomes or adding computational overhead. The format further allows users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

3.5.3 Traceability and confirmation of applied clinical standards

Each QV file includes a persistent `qv_set_id` and checksum that can be integrated into clinical databases such as REDCap or EPIC, enabling automatic linkage to patient records and inclusion in reports for transparent, auditable genomic interpretation. This structure ensures traceable provenance of variant interpretation and supports the FAIR principles of findability, accessibility, interoperability, and reusability. A key feature of the QV framework is that it enables direct confirmation of which clinical standards were applied in an analysis without requiring access to the underlying pipeline code. A common concern among patients and even technical collaborators

is having a clear record of what was actually tested, since readers often have specific questions that are not easily answered by conventional reports. For example, a patient might wish to know whether their genome was screened for variants in the well-known hereditary breast cancer genes *BRCA1* and *BRCA2*. By referencing the file `qv_acmg_sf_v3_3_20250828.json`, the report can confirm that the ACMG secondary findings guideline (v3.3) (17) was applied, including the defined set of genes and criteria for reporting pathogenic and likely pathogenic variants. This ensures that both analysts and patients can verify the applied standards and scope of analysis without ambiguity. Multiple QV sets can be combined in one analysis to enable reporting tailored to research, clinical, or commercial needs.

A



B

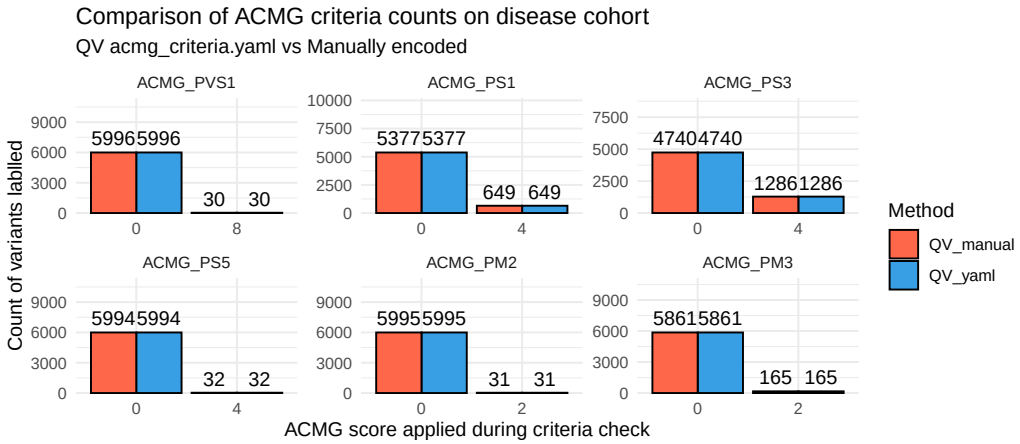


Figure 1: Summary of the QV application for a WGS pipeline. In panel (A), QV1 and QV2 are presented as sequentially piped protocol steps. In this example, QV2 differs from QV1 by retaining only likely/pathogenic variants (indicated by a red X). The QV file loaded by the analysis pipeline comprise a description 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.

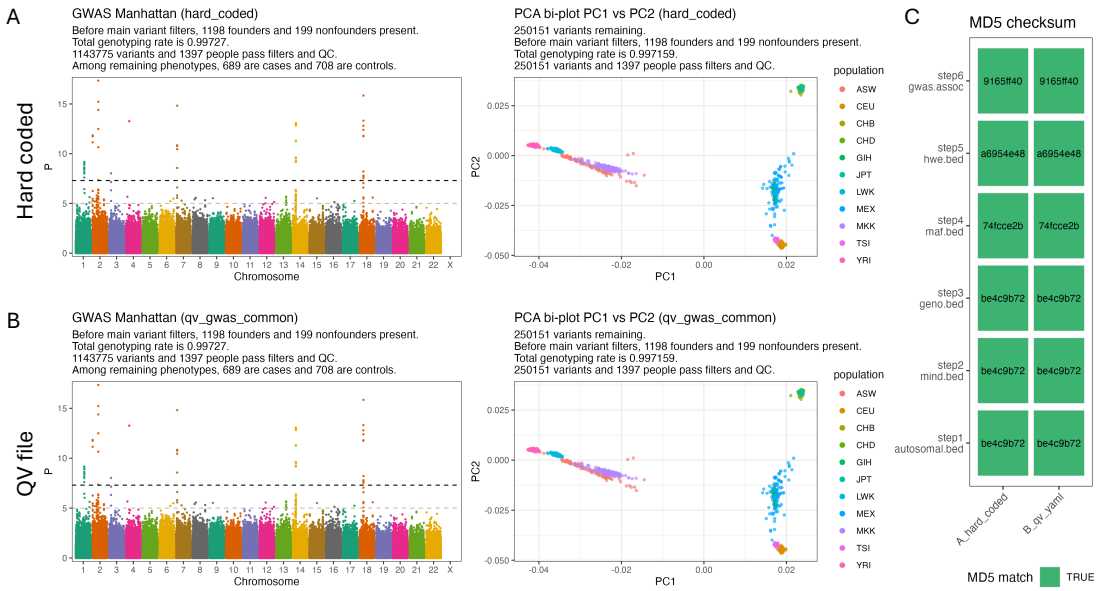
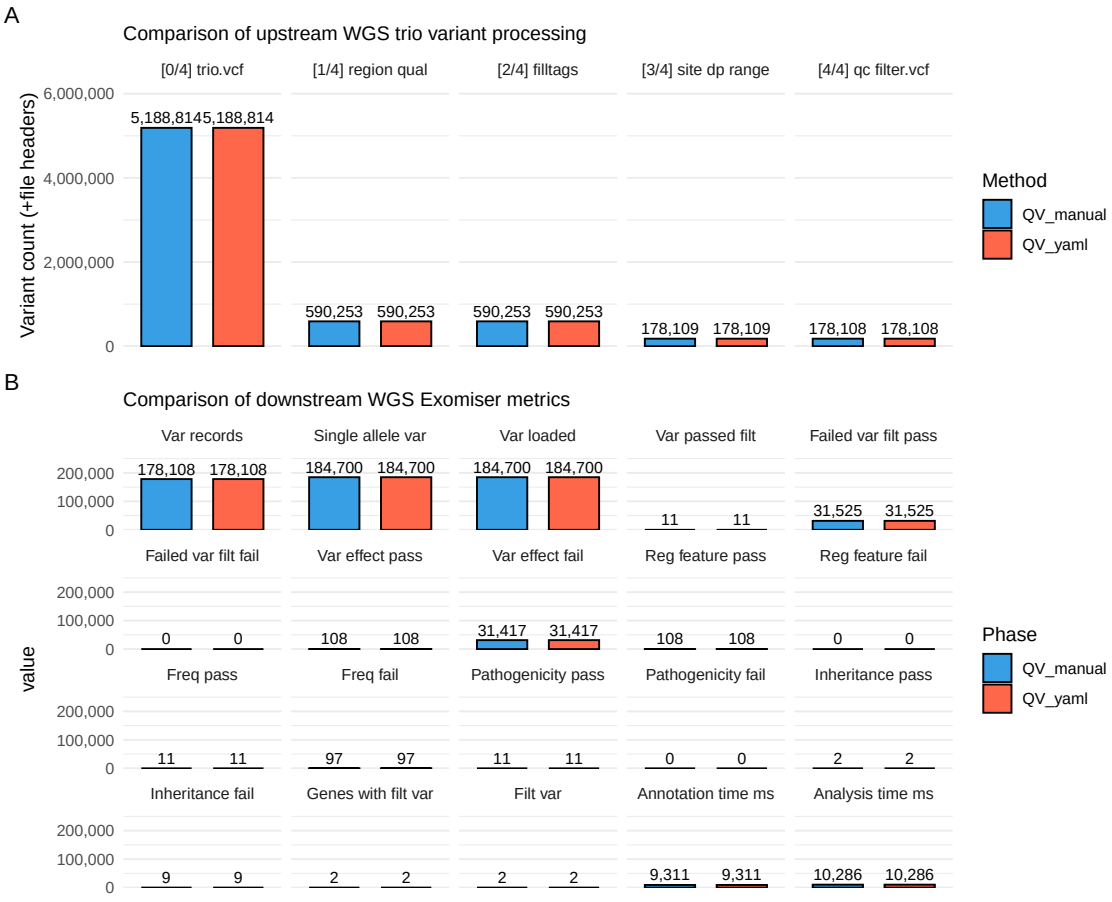


Figure 2: Validation in GWAS using QV parameterisation. (A) GWAS of simulated binary phenotypes in HapMap3 Phase 3 (R3) using a traditional variable embedded pipeline. Shown are the Manhattan plot of logistic regression results (left) and correction for population structure with principal component analysis (PC1 vs PC2, right). (B) Identical GWAS using a QV YAML configuration file. The Manhattan and PCA results are indistinguishable from panel A. (C) Verification of reproducibility. MD5 checksums of the main PLINK outputs are identical between panels A and B. The steps included processing of autosomal biallelic SNPs, sample call rate, variant call rate, minor allele frequency, Hardy–Weinberg equilibrium, and association results. The QV file encoded 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 confirms 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 the trio Exomiser pipeline using QV parameterisation. Panels A, B, and C correspond to upstream processing counts, downstream Exomiser metrics, and final variants detected, respectively. The variant counts in A-C all confirm that intermediate files from both configurations are identical in size. The five sequential preprocessing stages shown in A are: (0) input trio VCF, (1) gene panel region and quality filtering, (2) tag annotation, (3) site-level depth range filtering, and (4) final QC-filtered VCF output. MOI, mode of inheritance; HGVS, Human Genome Variation Society nomenclature.

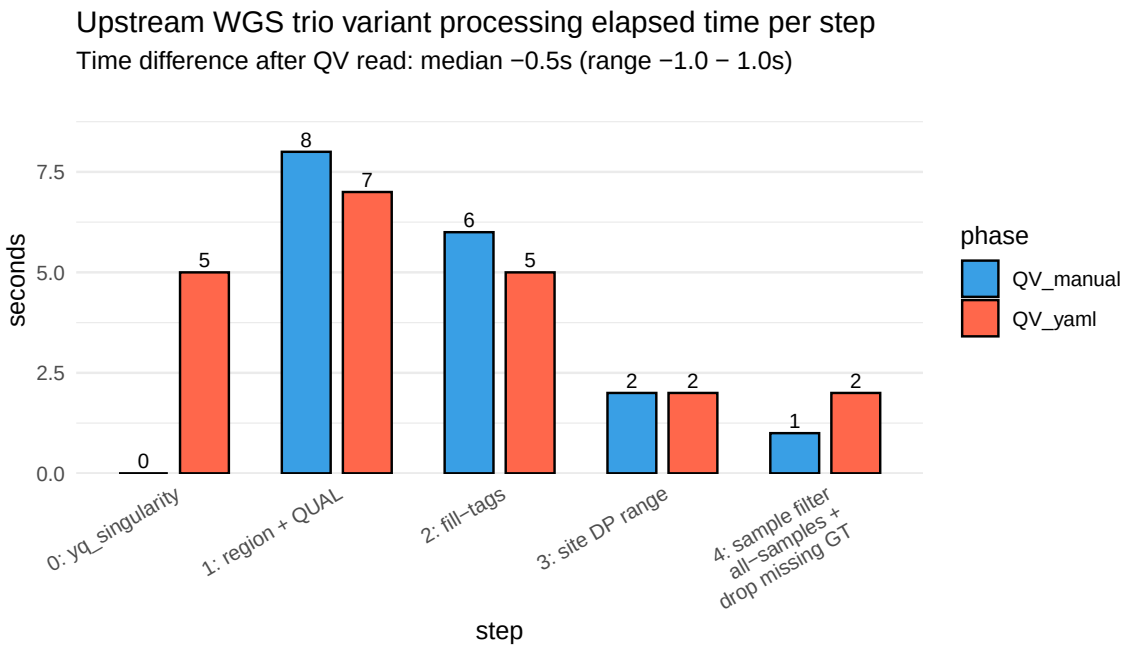


Figure 4: Benchmark of upstream preprocessing times in the WGS trio pipeline comparing QV-based and traditional (manually parameterised) configurations. Stepwise elapsed times were nearly identical across both methods (median difference ~0.5 s), with a fixed 5 s overhead from optional Singularity initialisation of `yq` in the QV pipeline. The four preprocessing steps correspond to: (1) gene panel region and quality filtering of the trio VCF, (2) annotation of variant tags, (3) site-level depth range filtering, and (4) per-sample genotype filtering and exclusion of missing genotypes. All steps used `BCFtools` on VCF preprocessing, as illustrated in **Figure 3 (A)**.

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] Susan Fairley, Ernesto Lowy-Gallego, Emily Perry, and Paul Flicek. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research*, 48(D1):D941–D947, January 2020. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkz836.

[28] Justin Wagner, Nathan D. Olson, Lindsay Harris, Ziad Khan, Jesse Farek, Medhat Mahmoud, Ana Stankovic, Vladimir Kovacevic, Byunggil Yoo, Neil Miller, Jeffrey A. Rosenfeld, Bohan Ni, Samantha Zarate, Melanie Kirsche, Sergey Aganezov, Michael C. Schatz, Giuseppe Narzisi, Marta Byrska-Bishop, Wayne Clarke, Uday S. Evani, Charles Markello, Kishwar Shafin, Xin Zhou, Arend Sidow, Vikas Bansal, Peter Ebert, Tobias Marschall, Peter Lansdorp, Vincent Hanlon, Carl-Adam Mattsson, Alvaro Martinez Barrio, Ian T. Fiddes, Chunlin Xiao, Arkarachai Fungtammasan, Chen-Shan Chin, Aaron M. Wenger, William J. Rowell, Fritz J. Sedlazeck, Andrew Carroll, Marc Salit, and Justin M. Zook. Benchmarking challenging small variants with linked and long reads. *Cell Genomics*, 2(5):100128, May 2022. ISSN 2666979X. doi: 10.1016/j.xgen.2022.100128.

[29] 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, August 2021. ISSN 2662-8449. doi: 10.1038/s43586-021-00056-9.

[30] Valentina Cipriani, Nikolas Pontikos, Gavin Arno, Panagiotis I. Sergouniotis, Eva Lenassi, Penpitcha Thawong, Daniel Danis, Michel Michaelides, Andrew R. Webster, Anthony T. Moore, Peter N. Robinson, Julius O.B. Jacobsen, and Damian Smedley. An Improved Phenotype-Driven Tool for Rare Mendelian Variant Prioritization: Benchmarking Exomiser on Real Patient Whole-Exome Data. *Genes*, 11(4):460, April 2020. ISSN 2073-4425. doi: 10.3390/genes11040460.

[31] Cristian Riccio, Max L. Jansen, Linlin Guo, and Andreas Ziegler. Variant effect predictors: A systematic review and practical guide. *Human Genetics*, 143(5):625–634, May 2024. ISSN 1432-1203. doi: 10.1007/s00439-024-02670-5.

[32] Manuel Holtgrewe, Oliver Stolpe, Mikko Nieminen, Stefan Mundlos, Alexej Knaus, Uwe Kornak, Dominik Seelow, Lara Segebrecht, Malte Spielmann, Björn Fischer-Zirnsak, Felix Boschann, Ute Scholl, Nadja Ehmke, and Dieter Beule. VarFish: Comprehensive DNA variant analysis for diagnostics and research. *Nucleic Acids Research*, 48(W1):W162–W169, July 2020. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkaa241.

[33] 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.