Conceptualising 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], Jacques Fellay[2], and Luregn J. Schlapbach[1]

[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]Personalized Health Informatics Group, SIB Swiss Institute of Bioinformatics, Basel, Switzerland.

March 25, 2025

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

Abstract

Motivation:

Qualifying variants (QVs) are genomic alterations selected using defined criteria within genomic processing pipelines. Although essential for both genetic research and clinical diagnostics, QVs are typically regarded as simple filters rather than dynamic, multifaceted components that influence the entire analytical workflow. Existing best practices adhere to variant classification standards and standardised workflows, yet a unified framework to integrate and optimise QVs for advanced, multi-stage applications is lacking.

Results:

We propose a redefinition of QVs by outlining several common QV sets and demonstrating their roles within analysis pipelines. By introducing new terminology and a standard reference model, our framework enables systematic integration and standardisation of QVs, thereby enhancing reproducibility, interpretability, and interdisciplinary communication. A validation case study, implementing ACMG criteria in a disease cohort demonstrates that our standardised approach achieves results identical to conventional methods while offering improved clarity and scalability.

Availability:

The source code and data are accessible at https://github.com/DylanLawless/qv2025lawless. 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 6

FAIR Findability, Accessibility, Interoperability, and Reusability 5

GWAS Genome Wide Association Test 4

MAF Minor Allele Frequency 8

PRS Polygenic Risk Score 4

QC Quality Control 4

QV Qualifying variant 4

SF Secondary Findings 5

SNV/INDEL Single Nucleotide Variant / Insertion Deletion 6

WGS Whole Genome Sequencing 4

1 Introduction

Qualifying variants (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 lua expressions. By providing filtering criteria in a standardised QV format, our approach complements such tools

The criteria for QV selection vary by application. For example, Genome Wide Association Test (GWAS) may focus on common variants, while clinical analyses usually target rare or known pathogenic variant. **Figure 1** illustrates a typical Whole Genome Sequencing (WGS) and variant filtering pipeline, where QV steps—from initial quality control to subsequent annotation-based filtering—are integrated.

Previous studies have demonstrated the utility of QVs (12; 13), yet no standardised framework 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 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 rare variant association testing.

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

4. **Known disease panel QV set**: Utilises 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 (14).

These examples illustrate a small 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 (15).

As WGS becomes standard for large cohorts (16; 17), 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, standardised vocabulary. Our framework offers structured, human- and machine-readable definitions that adhere to the principles for Findability, Accessibility, Interoperability, and Reusability (FAIR) (18), thereby promoting integration across databases. We implement standard vocabularies, unique identifiers, and formats like YAML to support this integration. By explicitly documenting variant filtering criteria and making QV data available in accessible formats, our framework builds trust and supports meaningful patient and public involvement (19). This transparency ensures that both experts and lay persons receive information in a format suited to their needs, improving diagnostic traceability and accelerating the translation of genetic research into clinical practice.

2 Methods

2.1 Implementation

By introducing a new vocabulary and a standard reference model for QVs, we aim to clarify the concept and improve communication and methodological discussion across disciplines for more advanced tasks. We note a few brief configurations and roles within analysis pipelines:

1. Theoretical pipelining of QV sets.
2. Establishment of standardised QV sets for specific analytical scenarios.
3. 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. This framework comprises three components, as illustrated in **Figure 1**:

1. **Variables**: The criteria variables that are sourced as part of the pipeline (see **Box 2**).
2. **Description**: An optional, recommend, narrative of each step within the overall QV set (see **Box 2**).
3. **Source code**: The implementation of the variables file within the pipeline code.

This framework efficiently manages QV-specific variables (e.g. allele frequency thresholds) separately from general pipeline settings, ensuring both clarity and specificity. The resulting format is versatile, supporting applications across WGS analysis, where it controls filtering and annotation, as well as downstream result interpretation and reporting, by linking the QV set ID to both results and raw data sources in a database.

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:

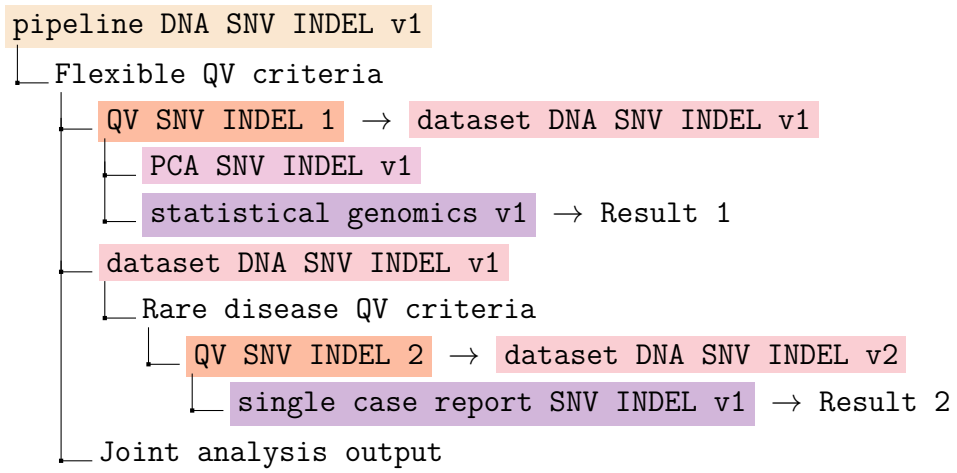
- `QV SNV/INDEL` Single Nucleotide Variant / Insertion Deletion (SNV/INDEL),
- `QV CNV` Copy Number Variant (CNV),
- `QV structural variation`,
- `QV rare disease known`, and
- `QV statistical association QC`.

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) (20; 21). As an example, 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



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 Validation Study

In the following validation study, we demonstrate that standardised QV criteria achieve a 100% match in criterion application when compared to the conventional manual approach. This analysis was performed on a rare disease cohort of 940 individuals (lawless spss 2025), which had been pre-processed for QC and filtered using a minimal QV test set, as described previously. Initially, we implemented an ACMG variant classification protocol (1) manually. We then re-implemented the same pro-

tocol using the new standardised QV criteria in YAML format. Our findings confirm that both methods produce identical results.

For ease of reporting, this example was restricted to chromosome 1, which contained 596 QV after strict filtering (Minor Allele Frequency (MAF) < 0.01) and was limited 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).

The annotation interpretation dataset was prepared in R using GuRu, our variant interpretation tool that consolidates all annotation sources and scores variants as candidate causal. The dataset, imported from gVCF format (output by VEP), consisted of 596 variant rows and 377 annotation columns.

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 standardised approach. The outputs from both methods were captured and compared, as shown in **Figure 1**. The QV criteria were provided in YAML format in the file `qv_files/acmg_criteria.yaml` (see **Box 2**).

Individual steps within the QV criteria can be further classified for organisational purposes using simple labels such as "QC" and "filter". For example, filtering thresholds (e.g. allele frequency > 0.1 in a cohort, < 0.1 in gnomAD) may be applied directly to exclude variants, while annotation-based criteria (e.g. QC flags) might not remove variants outright but instead inform downstream analyses that integrate multiple QV filters.

Box 2: qv_files/acmg_criteria.yaml

```
ACMG_PVS1:
  description: >
    Null variants (IMPACT = HIGH) in genes where
    loss-of-function causes disease.
    Includes homozygous variants, dominant inheritance,
    and compound heterozygous cases.
    Compound heterozygosity is considered when both
    variants are HIGH impact. WARNING: Not phase checked.
  logic: "or"
  conditions:
    - condition:
        field: IMPACT
        value: "HIGH"
        operator: "=="
...
shasum -a 256 acmg_criteria.yaml | fold -w 32
d91fde41a5fff48631adecba38773d61
9ae8cd5cff9b9b42ef7f5efbd6bbfcdf
acmg_criteria.yaml
```

3 Results

3.1 Validation Case Study

We validated our ACMG-based QV protocol using a rare disease cohort of 940 individuals (Lawless SPSs 2025). For ease of demonstrating here, were pre-processed with stringent filtering (MAF < 0.01) to isolate 596 variants on chromosome 1 in known disease genes (as defined by the Genomics England panel from our PanelAppRex repository). We then conducted the variant classification using two approaches: a conventional manual method with hard-coded criteria, and our new YAML-based implementation Annotation data were processed in R with GuRu, our variant interpretation tool, where key annotations (377 columns) were used to assign pathogenicity scores based on the first eight ACMG criteria (six applicable to this cohort). As shown in **Figure** 1, the outputs from both methods were identical, demonstrating a 100%

match. This confirms that our standardised, shareable QV criteria can be imported and applied programmatically with equivalent accuracy, providing a reproducible resource that is adaptable across different pipelines and programming environments.

Additional details of the YAML criteria in this QV set include 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. The final annotation results in this pipeline allow for automated retrieval of top candidate pathogenic variants using ACMG scoring methods (1; 5).

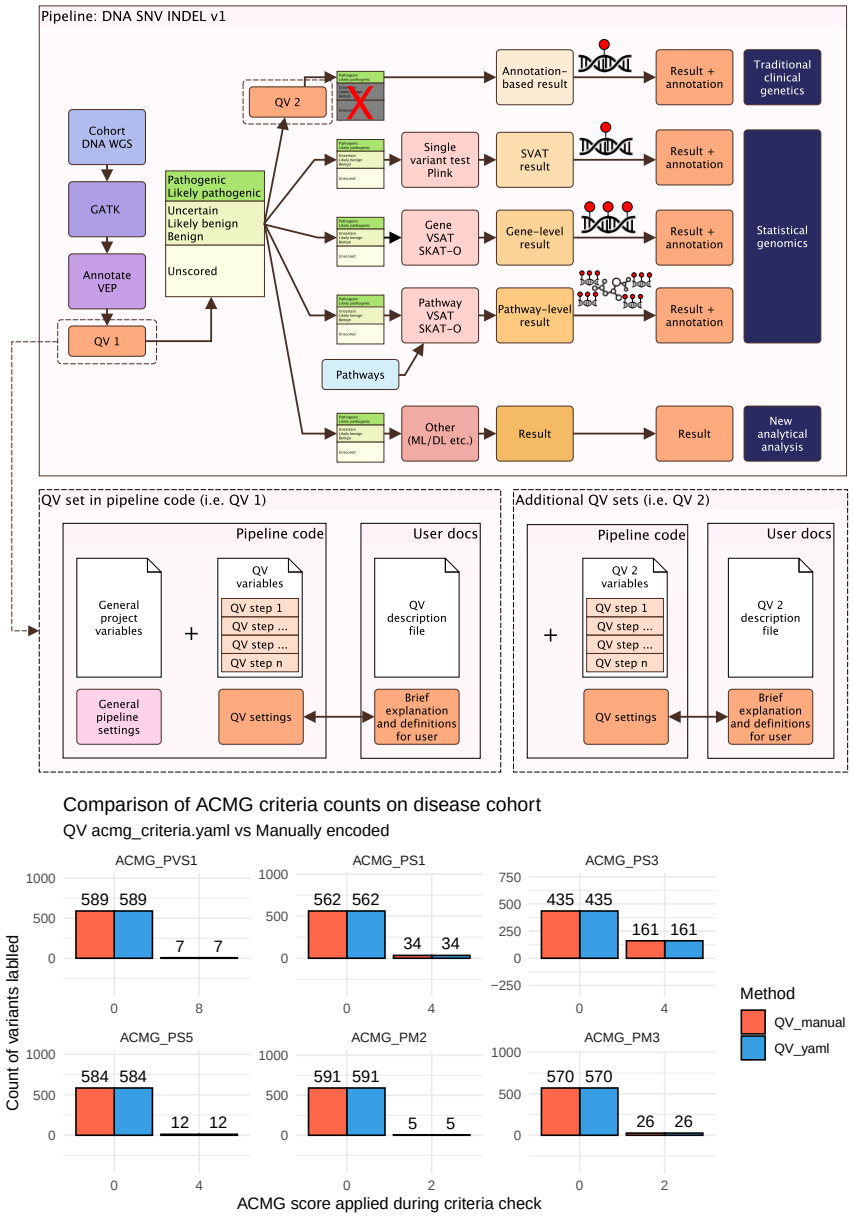


Figure 1: Summary of QV application for a WGS pipeline. Top: QV1 and QV2 are shown as sequential and potentially piped protocol steps. The description file (non-mandatory) and the variables file (mandatory) form part of the QV files that are loaded by the analysis pipeline. This illustration highlights a single stage in the QV1 set (i.e. step 10 where the GATK VQSR method is applied), with the full pipeline simplified under the QV1 icon. Bottol: GuRu case study using an ACMG criteria subset, demonstrating a 100% match between manually encoded and standardised YAML-based methods (`qv_files/acmg_criteria.yaml`) for assigning pathogenicity scores.

4 Conclusions

5 Funding

This project was supported through the grant 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, SÖ, JA contributed to the manuscript. JF, JV, LJS supervised the work and applied for funding.

8 Competing interests

None declared.

9 Collaborators

The SwissPedHealth consortium may be named here for publication and is prepared as a comment in the LaTeX document.

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] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.