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 28, 2025

Acronyms

ACAT Aggregated Cauchy Association Test . . . 27

ACMG American College of Medical Genetics and Genomics . . . 6

AF Allele Frequency . . . 11

AD Autosomal Dominant

AR Autosomal Recessive

GATK Genome Analysis Took Kit . . . 9

GWAS Genome Wide Association Test . . . 3

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

INDEL	Insertion / Deletion	4
IRI	internationalised resource identifier	19
MAF	Minor Allele Frequency	15
PPI	Patient and Public Involvement	26
PRS	Polygenic Risk Score	3
QC	Quality Control	4
QV	Qualifying variant	3
RDF	Resource Description Framework	8
$\mathbf{QV_{ax}}$	Axiomatic Variants	30
SF	Secondary Findings	6
SKAT	sequence kernel association test	30
SNV	Single Nucleotide Variant	4
SNP	Single Nucleotide Polymorphism	
SPHN	Swiss Personalized Health Network	8
VQSR	Variant Quality Score Recalibration	9
VSAT	Variant Set Association Test	3
VUS	Variants of Unknown Significance	6
WGS	Whole Genome Sequencing	4

Abstract

Qualifying variants (QV) are specific genomic alterations chosen through defined criteria in processing pipelines, and are essential for analyses in genetic research and clinical diagnostics. This paper reframes QVs not merely as simple filtering criteria but as a dynamic, multifaceted concept crucial for varied genomic analysis scenarios. We argue that standardising and optimising QVs for advanced, multi-stage use - rather than confining them to simplistic, single-stage filters - can significantly advance omics research and open new theoretical avenues. Although typically viewed as tools to exclude benign or unrelated variants, QVs actually involve complex, distributed steps throughout the analysis pipeline. We propose a redefinition of QVs by outlining several common sets and demonstrating their roles within analysis pipelines, thereby elucidating their integration and standardisation for specific analytical contexts. By introducing new terminology and a standard reference model, we aim to enhance understanding and communication about QVs, thus improving methodological discussions across disciplines. Finally, we present a validation case study demonstrating implementation of ACMG criteria in a disease cohort of 940 subjects with exome sequence data.

1 Introduction

1.1 Use and application of qualifying variants

Qualifying variant (QV)s are genomic alterations selected by specific criteria within processing pipelines, and they are essential for downstream analyses in genetic research and clinical diagnostics. This paper explores the application and conceptualisation of QVs not merely as simple filters, but as dynamic elements that are integral throughout genomic analysis pipelines. Typically, the selection of QVs follows established best practices in variant classification and reporting standards (1–5), as well as standardised workflows (6–8). Nonetheless, a standard guide for the QV concept is currently lacking. Analogous to the development of Polygenic Risk Score (PRS) reporting standards, which promote reproducibility and systematic evaluation (9; 10), a similar approach for QVs is both necessary and beneficial.

The thresholds for QV selection are tailored to the specific requirements of each study. For example, Genome Wide Association Test (GWAS) may prioritise common variants, Variant Set Association Test (VSAT) may require rare variant collapse, while clinical genetic reports often focus on rare or novel variants. Thus, QVs are typically classified by the nature and extent of filtering or quality control they undergo. **Figure 1** illustrates a typical Whole Genome Sequencing (WGS) and VSAT analysis pipeline for Single Nucleotide Variant (SNV)-Insertion / Deletion (INDEL), where QV steps are arranged sequentially and may be piped together within the protocol.

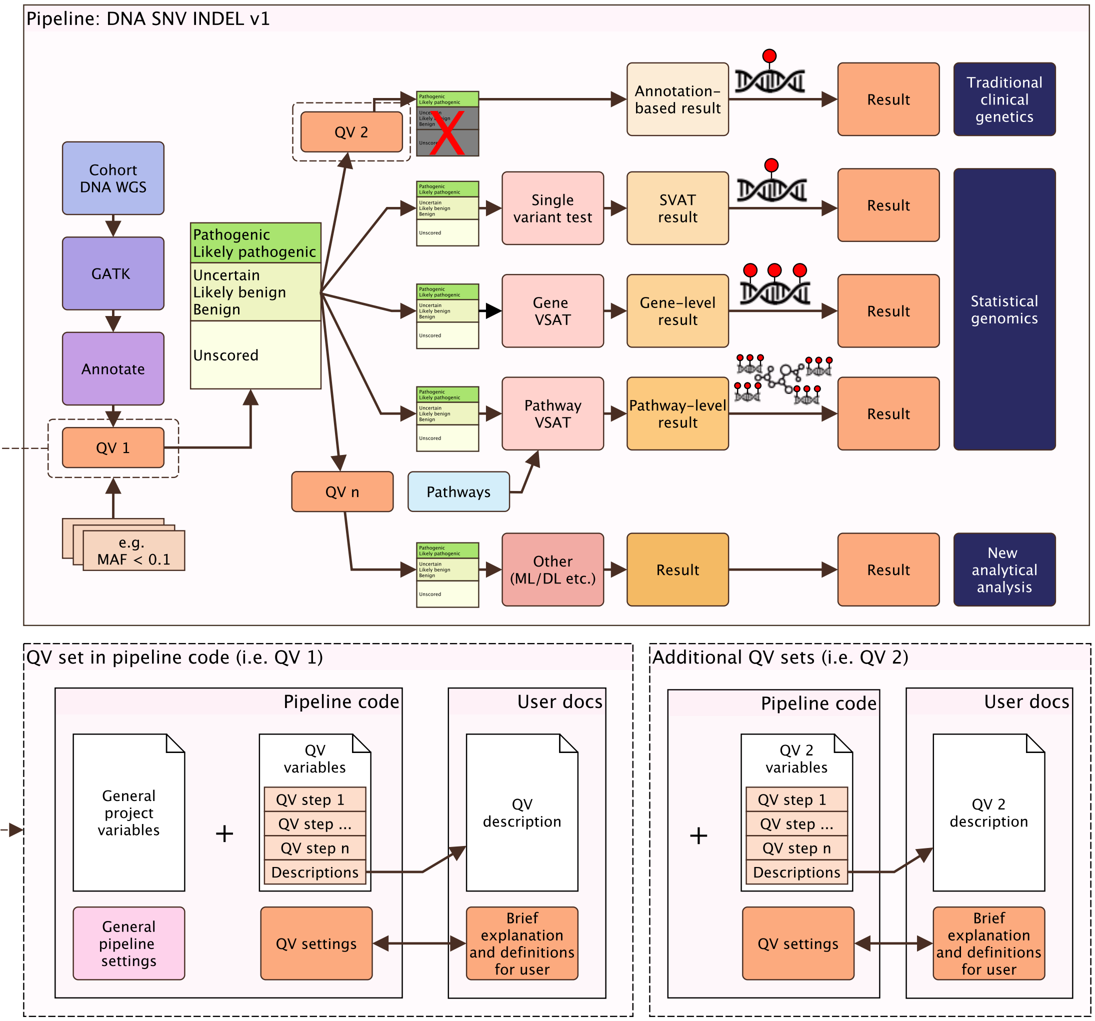


Figure 1: Summary of the example application design for the DNA SNV INDEL v1 pipeline. QV1 and QV2 are shown as sequential and potentially piped protocol steps. The description file (non-mandatory) and the variables file (mandatory) form part of the QV files that are loaded by the analysis pipeline. 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.

The typical representation of QV steps is shown in **figure 2**. This figure summarises the common steps in the variant filtering process such as Quality Control (QC) filtering of raw omic data, removing background noise such as high frequency variants compared to a reference population, and filtering on variant effect metrics like pathogenicity scores. The transition from raw data to annotated variants, as illustrated in **figure 3**, underscores how initial QV steps (e.g. QC) are complemented by additional filtering based on new annotation data.

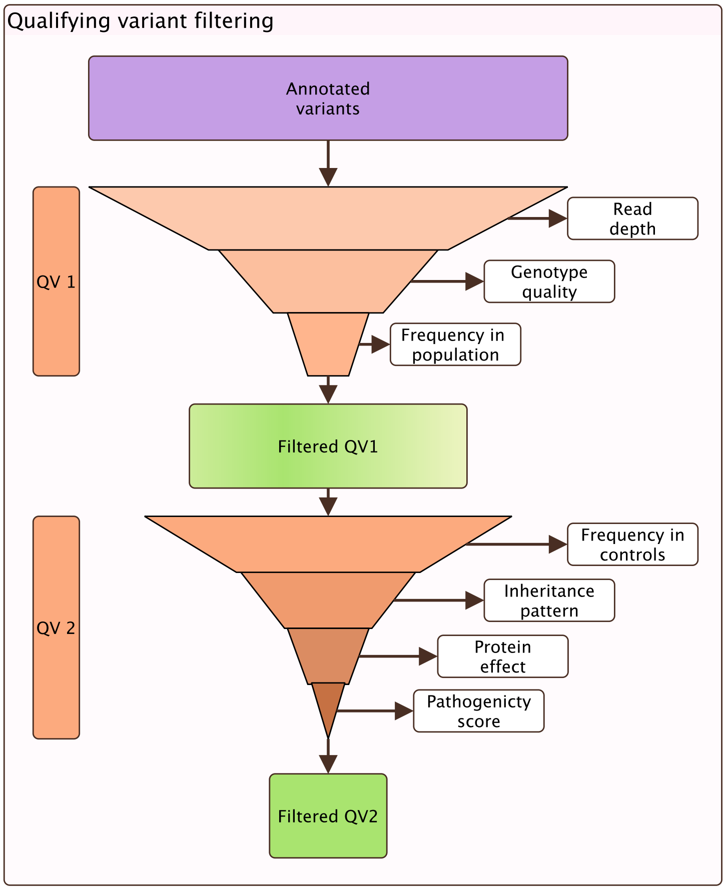


Figure 2: Illustration of the qualifying variant workflow. Each layer of filtering can arise from different stages of the pipeline. In reality, we observe that each layer of the filters comes from disparate stages of a pipeline.

Previous work has demonstrated tangible applications of QVs. For instance, Povysil et al. (11) provided an example of QVs in variant collapsing analyses for complex traits, while Cirulli et al. (12) introduced the concept in early studies. Despite these contributions, a standardised framework for presenting QVs is absent. Here, we detail four typical applications of QV sets:

1. **QV passing 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 (13).

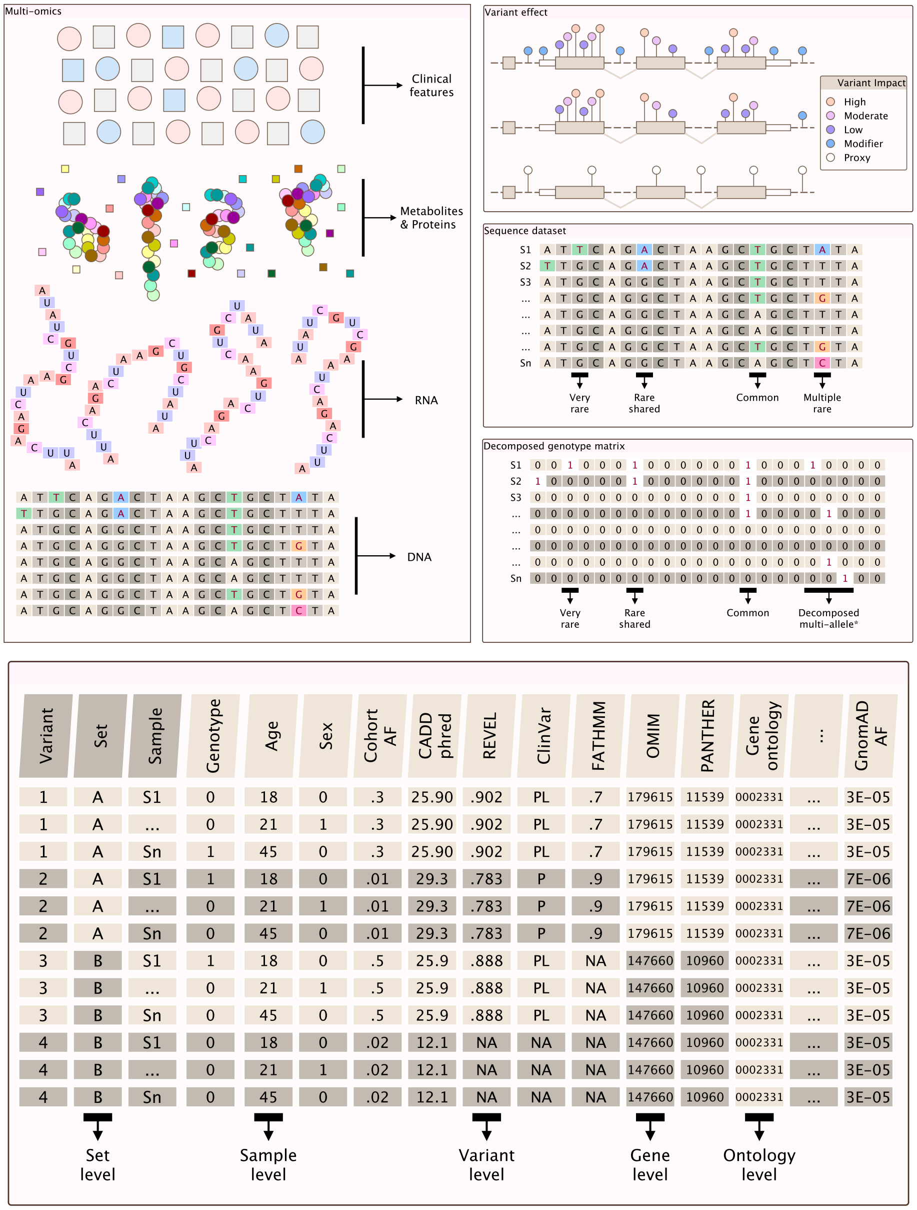
Two exemplary applications of QVs are found in clinical genetics reporting and GWAS. In clinical genetics, single-case analyses may select QVs from disease-causing gene lists provided by expert panels, with variants being categorised as Variants of Unknown Significance (VUS), known, candidate, or causal. In GWAS, QVs typically represent consensus variants that have passed rigorous QC procedures, ensuring their suitability for statistical analyses. 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 (14).

1.2 Background problem and proposed solution

As study sizes now exceed one million subjects (15; 16), the shift from genotyping to WGS is now standard, allowing rare variants to be included in GWAS and VSAT for more comprehensive analyses of complex traits (17; 18). QV protocols play a crucial role in data cleaning and preparation, ensuring the integrity of subsequent analyses. Although the term "QV" is often used as a catch-all descriptor for various filtering steps, in practice it encompasses multiple stages that originate from diverse parts of the pipeline. **Figure 4** illustrates the structural framework of an annotated variant, highlighting the features, both pre-existing metadata and post-calling annotations, that can trigger specific QV protocols. This figure emphasises that each variant is subject to multiple criteria, which may derive from distinct processing steps.

Moreover, complex analyses often require multiple processing streams that are ultimately merged into a cohesive result. A standardised QV format would allow for the use of various QV sets, each based on different filters and variables, while providing a common foundation for consistency across disparate data streams. The ambiguity in the term QV across genomic studies underscores the need for a clear and flexible definition that captures both its common uses and its implementation across multiple stages.

By introducing a new vocabulary and a standard reference model for QVs, we aim to clarify the concept and improve communication and methodological discussion



Variant	Set	Sample	Genotype	Age	Sex	Cohort AF	CADD phred	REVEL	ClinVar	FATHMM	OMIM	PANTHER	Gene ontology	...	GnomAD AF
1	A	S1	0	18	0	.3	25.90	.902	PL	.7	179615	11539	0002331	...	3E-05
1	A	...	0	21	1	.3	25.90	.902	PL	.7	179615	11539	0002331	...	3E-05
1	A	Sn	1	45	0	.3	25.90	.902	PL	.7	179615	11539	0002331	...	3E-05
2	A	S1	1	18	0	.01	29.3	.783	P	.9	179615	11539	0002331	...	7E-06
2	A	...	0	21	1	.01	29.3	.783	P	.9	179615	11539	0002331	...	7E-06
2	A	Sn	0	45	0	.01	29.3	.783	P	.9	179615	11539	0002331	...	7E-06
3	B	S1	1	18	0	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
3	B	...	0	21	1	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
3	B	Sn	0	45	0	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
4	B	S1	0	18	0	.02	12.1	NA	NA	NA	147660	10960	0002331	...	3E-05
4	B	...	0	21	1	.02	12.1	NA	NA	NA	147660	10960	0002331	...	3E-05
4	B	Sn	0	45	0	.02	12.1	NA	NA	NA	147660	10960	0002331	...	3E-05

Figure 3: Top: Transformation from raw omic data to a data matrix. Bottom: Initial variant detection requires QC and filtering rules, which are the first QV steps. Subsequent annotation of variants enables further QV filtering based on new information.

across disciplines for more advanced tasks. We define and exemplify several common QV sets, illustrating their potential configurations and roles within analysis pipelines:

1. Theoretical pipelining of QV sets.

2. Establishment of standardised QV sets for specific analytical scenarios.

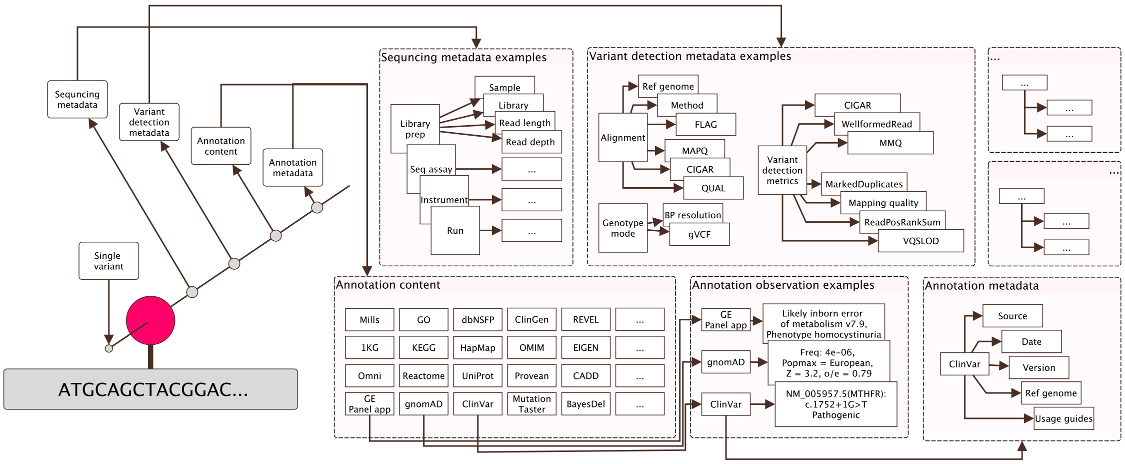


Figure 4: Structural framework of an annotated variant. The diagram highlights selected features, both pre-existing metadata and annotations added after variant calling, that can trigger specific QV protocols.

3. Recognition that QVs are integral throughout the analysis pipeline rather than confined to a single end-stage.

The proposed QV framework provides structured, human- and machine-readable definitions to standardise the selection and interpretation of variants across diverse genomic studies. This methodology promotes efficient and precise variant detection and interpretation, essential for both research and clinical diagnostics. Moreover, the structured criteria adhere to the principles of **fair!** (**fair!**) (19). In **section 2.4** we discuss how they can be versioned and integrated using standard vocabularies (e.g. SNOMED CT, Swiss Personalized Health Network (SPHN) **rfd!** (**rfd!**) Schema), assigned unique identifiers (via SHA-256 hashes, UUIDs, or semantic combinations), and provided in human- and machine-readable formats such as YAML and Resource Description Framework (RDF), ensuring seamless integration across databases and platforms.

2 Results

In the following sections, we first present a high-level view of a complete analysis using our QV framework (**sec 2.1**), illustrating its application in diverse contexts such as a multi-part genomic analysis. We then introduce the underlying methodological framework (**sec 2.2**), followed by an explicit example of one individual step (**sec 2.2.1**) to demonstrate how QV criteria are integrated into a pipeline. This specific example use the Variant Quality Score Recalibration (VQSR) step from the Genome Analysis Took Kit (GATK) WGS workflow (20) - a typical pipeline step.

2.1 Application of Qualifying Variants and an Example Use Case

We examine scenarios where QVs have proved essential, including applications in GWAS, WGS and clinical genetics. In large-scale studies and rare disease research, for example, sophisticated risk models that integrate clinical and genomic data can significantly enhance predictive accuracy in well-defined cohorts (21–23). Similarly, standardised QV protocols support reproducibility across studies, particularly when analysing complex signals in isolated populations (24).

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`,
- `QV CNV`,
- `QV structural variation`,
- `QV rare disease known`, and
- `QV statistical association QC`.

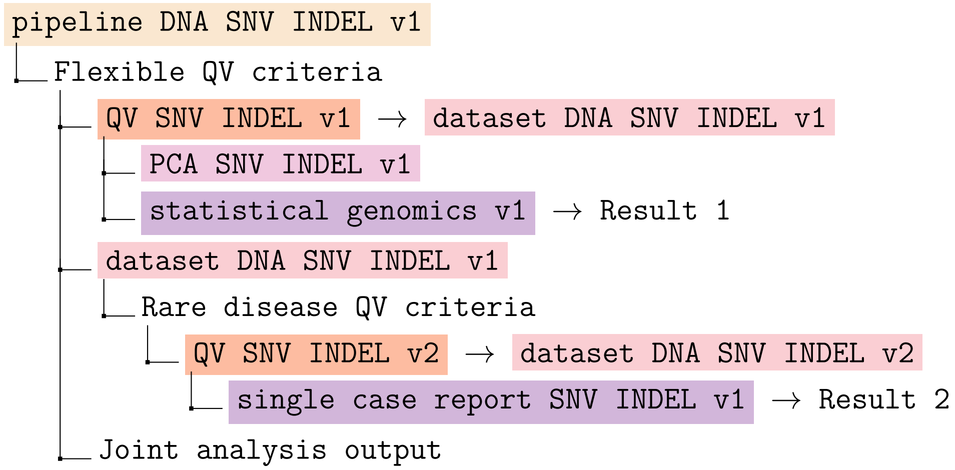
The final analysis yields (1) a joint cohort disease association (e.g. variant P-values) and (2) individual single-case results (e.g. clinical genetics diagnosis for a patient) (20; 25).

As an example, we focus on a SNV/INDEL pipeline employing two QV sets: `QV SNV INDEL v1` for flexible cohort-level filtering, and `QV SNV INDEL v2` 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 (`QV SNV INDEL`

v1) was applied using the pipeline DNA SNV INDEL v1, which implements the QV SNV INDEL v1 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 v2 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 v1 → 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 v2 → 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.

2.2 Methodological framework

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**: A narrative of each step within the overall QV set (see **Box 3**).

3. **Source code**: The implementation of the variables file within the pipeline code (see **Box 4**).

This framework efficiently manages QV-specific variables (e.g. allele frequency thresholds) separately from general pipeline settings, thereby maintaining clarity and specificity. We first present a detailed example using VQSR from the QV set `QV SNV INDEL v1` to illustrate the practical application of this method in a real-world genomic analysis scenario. We later demonstrate how this approach integrates with workflow managers such as Snakemake (26) or Nextflow (27), streamlining genomic processing tasks.

Individual steps within QV criteria may be categorised into different types. For organisational purposes, we recommend using simple labels such as "QC" and "filter". For example: (1) Filtering thresholds (e.g. Allele Frequency (AF) > 0.1 in a cohort, < 0.1 in gnomAD) may be directly applied to exclude affected variants. (2) Multiple steps involving annotation-based criteria (e.g. QC flags) may not remove variants immediately but enable downstream analyses that depend on several QV criteria. In a QC protocol, one step might filter variants based solely on threshold values (criteria 1), while another may combine several upstream QC criteria (criteria 2).

2.2.1 Detailed example of QV variables

As a detailed example, we focus on the `vqsr` step from the QV set `QV SNV INDEL v1`. The process is illustrated in three parts. First, the mandatory QV variables are set (see **Box** 2). Second, an optional description is provided (see **Box** 3). Third, the variables are integrated into the source code (see **Box** 4).

Box 2: Example QV variables - extract from QV1 variables file

```
# VQSR SNP Mode Variables
vqsr_snp_hapmap_known="false"
vqsr_snp_hapmap_training="true"
vqsr_snp_hapmap_truth="true"
vqsr_snp_hapmap_prior="15.0"

vqsr_snp_omni_known="false"
vqsr_snp_omni_training="true"
vqsr_snp_omni_truth="false"
vqsr_snp_omni_prior="12.0"

vqsr_snp_1000g_known="false"
vqsr_snp_1000g_training="true"
```

```
vqsr_snp_1000g_truth="false"
vqsr_snp_1000g_prior="10.0"

vqsr_snp_annotations="QD,MQ,MQRankSum,ReadPosRankSum,FS,SOR"
vqsr_snp_truth_sensitivity="99.7"
```

Box 3: Example QV descriptions - extract from QV1 variables file

1. [**QC**] `fastp`: Performs initial read quality ...
2. [**QC**] `collectwgsmetrics`: Assesses BAM file ...
3. [**QC**] `rmdup_merge`: Marks duplicate reads ...
4. [**QC**] `haplotype_caller`: Generates gVCF using `-ERC GVCF` ...
5. [**QC**] `vqsr`: Performs Variant Quality Score Recalibration (VQSR) using GATK. In this step, SNP mode is applied with three reference resources: **HapMap** is used as a high-confidence reference (training=true, truth=true, prior=15.0), **Omni** provides supplementary training data (training=true, truth=false, prior=12.0), and **1000G** informs on common SNP variation (training=true, truth=false, prior=10.0). Additionally, VQSR uses key annotations (QD, MQ, MQRankSum, ReadPosRankSum, FS, SOR) and applies a truth sensitivity filter of 99.7% to retain high-confidence variants.
6. [**QC**] ...

Box 4: Example code sourcing the variables file

```
#!/bin/bash

# Source master settings (including VQSR) and custom QV1
   ↪ settings
source ./variables_master.sh
source ./variables_qv1.sh

# Run VQSR for SNPs
# 1. Calculate VQSLOD tranches for SNPs using
   ↪ VariantRecalibrator
gatk --java-options "${JAVA_OPTS}" VariantRecalibrator \
-R ${REF} \
-V ${vcf_file} \
--resource:hapmap,known=${vqsr_snp_hapmap_known},training=${
   ↪ vqsr_snp_hapmap_training},truth=${vqsr_snp_hapmap_truth
   ↪ },prior=${vqsr_snp_hapmap_prior} ${hapmap} \
--resource:omni,known=${vqsr_snp_omni_known},training=${
   ↪ vqsr_snp_omni_training},truth=${vqsr_snp_omni_truth},
   ↪ prior=${vqsr_snp_omni_prior} ${omni} \
--resource:1000G,known=${vqsr_snp_1000g_known},training=${
   ↪ vqsr_snp_1000g_training},truth=${vqsr_snp_1000g_truth},
   ↪ prior=${vqsr_snp_1000g_prior} ${thousandG} \
-an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an SOR
   ↪  \
--mode SNP \
-O ${OUTPUT_DIR}/chr${INDEX}_snp1.recal \
--tranches-file ${OUTPUT_DIR}/chr${INDEX}_output_snp1.
   ↪ tranches

# 2. Filter SNPs on VQSLOD using ApplyVQSR

gatk --java-options "${JAVA_OPTS}" ApplyVQSR \
...continued
```

2.2.2 Simple example with a workflow manager

We demonstrate the use of a QV variable file within a workflow manager, such as Snakemake or Nextflow (**box** 5). The setup involves two types of YAML configuration files: one for general pipeline settings and another specifically for QV-related variables (**box** 6). These configurations are integrated into the primary analysis script, typically a Snakefile, ensuring that all parameters required for genomic analyses are systematically managed and applied (**box** 7).

Box 5: Example worflow manager - yaml

```yaml
# qv_config.yaml
min_depth: 10
max_allele_frequency: 0.01
quality_score_threshold: 20
```

Box 6: Example worflow manager - yaml

```yaml
# config.yaml
reference_genome: "path/to/reference/genome.fasta"
annotation_file: "path/to/annotation.gtf"
sample_list: "path/to/samples.txt"
output_dir: "path/to/output"
qv_config: "qv_config.yaml"
```

Box 7: Example worflow manager - python

```python
# Snakefile
configfile: "config.yaml"
qv_settings = read_yaml(config["qv_config"])

rule all:
  input:
    "results/filtered_variants.vcf"

rule filter_variants:
  input:
    "data/raw_variants.vcf"
  output:
    "results/filtered_variants.vcf"
  params:
    depth = qv_settings['min_depth'],
    af = qv_settings['max_allele_frequency'],
    qs = qv_settings['quality_score_threshold']
  shell:
    """
    bcftools filter -i 'DP>{depth} && AF<{af} && \
    QUAL>{qs}' {input} > {output}
```

```
      ""
```

2.3 Examples of real-world QV applications

2.3.1 Discovery research

Greene et al. (28) provide an example of QV standardisation with their “Rareservoir”, a relational database schema optimised for rare disease studies. This database focuses on rare variants (those with a Minor Allele Frequency (MAF) below 0.1%), reducing data size by approximately 99% by storing variants as 64-bit integers (“RSVR IDs”) and organising them by genomic position for efficient querying. Additional data, such as MAFs from gnomAD, CADD pathogenicity scores and impact predictions per the Sequence Ontology, are encoded into a 64-bit integer (“CSQ ID”), where each bit corresponds to a specific gene function impact, ranked by severity. By employing the Bayesian genetic association method BeviMed, initially described by Greene et al. (29), the study effectively inferred associations between genes and rare disease phenotypes, demonstrating the capacity to handle and analyse complex genetic datasets. The protocol of Greene et al. (28) can be reproduced in the QV format, thereby facilitating interpretation and reproducibility.

In our work, we are applying the QV framework within “SwissPedHealth”, a national paediatric data stream aimed at investigating rare or unknown diseases using a multiomic approach including WGS, RNAseq, proteomics, metabolomics, and clinical data (30). The SwissPedHealth Lighthouse project involves approximately 450 paediatric cases with rare, life-threatening conditions, with the goal of improving diagnosis by integrating clinical data with multiomics. By using consensus raw datasets (e.g. WGS in patients and families) that are annotated and filtered to multiple QV levels, we generate pre-processed datasets suitable for a range of analyses, including GWAS, VSAT, single-case clinical genetics reporting, machine learning, and joint multiomic studies.

This approach aligns with practices in large-scale national projects, such as the Genomics England 100,000 Genomes Project, which performs central automated analysis with interpretation and clinical reporting (31). Although such projects currently embed QV protocols throughout their pipelines without explicit standardisation, our method aims to improve consistency and reproducibility. Moreover, the increasing importance of QV protocols in genomics-based newborn screening, a rapidly emerging healthcare innovation (32), highlights the critical role of standardised variant filtering in bridging research and clinical practice.

2.3.2 Rapid diagnostics

Screening for known diseases typically involves searching for a predefined QV set. The adoption of a formalised standard would enhance consistency and reliability for stakeholders. For instance, genome analysis in neurodevelopmental disorders in 465 families identified causal variants in 36% of 489 affected individuals (33), while the DDD study, involving over 13,500 families, achieved a genetic diagnosis in approximately 41% of probands (34). In addition, genomic lifespan association studies in Iceland, which included 57,933 participants, identified 2,306 individuals with actionable genotypes associated with a reduction in median lifespan by around three years (35).

Rapid genomic diagnostics can be further enhanced by the use of standardised QV protocols. Standardisation of QV filtering ensures high-confidence in the analysis protocol, thereby streamlining data interpretation, reproducibility, and meta analysis. For example, in the United Arab Emirates, rapid whole-genome sequencing has been achieved with an average turnaround time of 37 hours (36). Moreover, Meng et al. (37) reported that, among 278 critically ill infants, a molecular diagnosis was achieved in 36.7% of cases, with higher diagnostic rates (50.8%) observed in critical trio exome cases, and a subsequent impact on medical management in 52.0% of diagnosed cases. Similarly, Lunke et al. (38) demonstrated that, in a national-scale multiomic study for rare diseases involving 290 critically ill infants and children, the diagnostic yield from WGS initially stood at 47% but increased to 54% with extended analysis, leading to altered critical care management in 77% of diagnosed cases. With a consistent QV protocol these reports become benchmarks to reproduce and improve upon - even in cases where the underlying software or algorithms are proprietary.

2.3.3 Complex variant calls

Applying the QV framework to diverse analysis types, including SNV-INDEL, copy-number variants, and structural variants, allows simple QV IDs to be used for database reporting. This facilitates querying to determine whether further analysis may reveal additional findings. For example, Wojcik et al. (39) employed genome sequencing in 822 families with rare monogenic diseases, achieving a diagnostic yield of 29.3%. Their broader genomic coverage, which included structural and non-coding variants, identified causative variants in 8.2% of cases that were previously undetected by exome sequencing.

2.3.4 Secondary findings

The ACMG SF v3.2 list exemplifies a set of QV resulting in a widely accepted and impactful guideline in genomic medicine (13). This list specifies gene-phenotype pairs recommended for reporting as secondary findings during clinical exome and genome sequencing. Such standardisation streamlines the identification of clinically actionable genetic information and enhances the consistency and quality of genomic data interpretation across different settings. However, the dataset is relatively unstructured, requiring extensive manual curation by front-line analysts for each implementation. The ACMG SF list is revised annually, reflecting its dynamic nature and the evolving understanding of gene-disease correlations. Each version, such as the current v3.2, includes detailed criteria for the inclusion or exclusion of specific genes, based on rigorous evidence of their association with significant health outcomes. This methodical curation ensures that the list remains a reliable resource for opportunistic screening in clinical contexts.

Table 1 lists the first two transposed entries from the ACMG SF list, showcasing specific genes associated with cardiovascular phenotypes. We subsequently represent this data in a standardised QV format (see **Boxes 8–9**), which can be incorporated into any variant filtering program. In bioinformatics pipelines, consistent specification of QV sets, such as ACMG SF v3.2, enables patients to receive the most relevant and up-to-date information regarding their genetic health risks without the burden of manually implementing new QV standards.

Table 1: The first two entries from the ACMG SF v3.2 list, transposed, for reporting of secondary findings in clinical exome and genome sequencing (13).

Detail	**ACTA2**	**ACTC1**
Disease/Phenotype	Familial thoracic aortic aneurysm	Hypertrophic cardiomyopathy
Gene MIM	102620	102540
Disorder MIM	611788	612098
Phenotype Category	Cardiovascular	Cardiovascular
Inheritance	AD	AD
Variants to report	All P and LP	All P and LP

Box 8: QV configuration for SF - yaml

```
# qv_sf_v3.2_config.yaml
genes:
- gene: "ACTA2"
   inheritance_pattern: "AD"
   variant_class : ["Pathogenic", "Likely Pathogenic"]
- gene: "ACTC1"
   inheritance_pattern: "AD"
   variant_class: ["Pathogenic", "Likely Pathogenic"]
...
```

Box 9: Filtering command for QV SF

```
# Pseudo-code to filter variants for each gene
# in ACMG SF v3.2 list:

Read genes from qv_sf_v3.2_config.yaml

For each gene entry in genes:
  Apply filter command:
    filter -i 'GENE=="{gene['gene']}" &&
    INHERITANCE=="{gene['inheritance_pattern']}" &&
    (VARIANT_CLASSIFICATION in gene['variant_class'])'
    input.vcf > output_{gene['gene']}_qv_sf.vcf
```

ACMG SF is a widely known protocol in clinical genomics. In our SwissPedHealth work is supported by SPHN and uses recommendations from the ELSI Advisory Group (ELSIag) on ethical, legal, and social implications. This group - comprising experts in bioethics, life sciences law, and social sciences, as well as representatives from SAMS, swissethics, and patient advocacy - recommends best practices for reporting actionable genetic findings to research participants. In this context, our internal QV reporting framework parallels the ACMG SF approach, but is specifically tailored to meet the needs of our clinical and research environments.

2.4 Enhancing semantic interoperability

The SPHN promotes data sharing based on the FAIR principles, supported by the SPHN RDF Schema to enhance semantic interoperability, particularly for clinical routine data (19; 40). A recent extension of this schema incorporates genomic data processing, enriching it with detailed genomic-specific concepts that span from sample processing to the sequencing run (41). This extension includes critical elements such as the sequencing instrument and QC metrics, which are essential for ensuring the integrity and reproducibility of genomic analyses. To further integrate omics data within clinical frameworks, we have developed additional concepts (e.g. `https://biomedit.ch/rdf/sphn-schema/sph#OmicsAnalysis`, `https://biomedit.ch/rdf/sphn-schema/sph#OmicsAnalysisResult`, and `https://git.dcc.sib.swiss/sphn-semantic-framework/sphn-schema/`), which enable the direct reporting of outcomes tied to clinical care.

The QV framework allows explicit recording of the QV sets used in analyses, providing a robust mechanism to track and verify the application of specific variant sets. such as those defined by the ACMG SF, independent of internal protocol changes. This feature enables users to query and confirm the use of specific QV sets without needing to examine the underlying source protocols, thereby streamlining verification and enhancing the transparency and traceability of genomic analyses within the SPHN framework.

To enhance reproducibility and traceability in omics research, we propose the QV Set ID (`qualifying_variant_set_id`), which links the variant sets used in analyses and facilitates precise and consistent replication of research methodologies. Unique, consistent identifiers that align with existing data management standards and integrate seamlessly into RDF schemas, such as those incorporating SNOMED CT, are essential. Examples for generating such identifiers include:

1. **Hash functions:** Using SHA-256 to generate a unique hash of the set's characteristics.

2. **UUIDs:** Randomly generated UUIDs, which provide high uniqueness across systems.

3. **Semantic combination:** Creating identifiers by combining relevant semantic elements (e.g. project ID, data provider ID, and data release version) in a structured format.

4. **IRI incorporation:** Developing internationalised resource identifier (IRI)s for traceability and integration into linked data frameworks.

5. **Registry-based allocation:** Using a centralised registry to manage identifier assignment.

6. **Linking standards:** Mapping local identifiers to established international standards (e.g. SNOMED CT) through equivalence classes.

Box 11 provides an example analysis plan (or result database entry), listing the pipeline used, three hypothetical internal QV sets (`qv1, qv2, qv3`) and one public QV set (`acmg_sf_v3.2`). The SHA-256 hash of the `acmg_sf_v3.2` file is provided to verify its integrity:

Box 10: Example implementation of QV Set ID

pipeline: `pipeline DNA SNV INDEL v1`
qualifying_variant_set_id: `qv1_20250201`
qualifying_variant_set_id: `qv2_20260101`
qualifying_variant_set_id: `qv3_20260101`
qualifying_variant_set_id: `acmg_sf_v3.2`

where

```
$ shasum -a 256 acmg_sf_v3.2.tsv | fold -w 32
6ad26a7df2feda3e2d4bfabf4a3cb1ca
4356b098ccc0890a7a17f198a9ab117f
acmg_sf_v3.2.tsv
```

Incorporating `qualifying_variant_set_id` not only enhances transparency but also increases operational efficiency in omics data handling, thus facilitating precise and reproducible research across various projects.

In our work, we consider the QV YAML file an electronic resource that can be stored, accessed, and transferred as a single unit. This aligns with the `sphn:DataFile` concept which is used in the SPHN RDF schema (**Figure 5**). Below is an illustrative RDF /Turtle snippet showing how it might be represented in a RDF :

Box 11: Use of a QV data file in RDF

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix sphn: <https://biomed.it/rdf/sphn-schema/> .
@prefix ex: <http://example.org/> .

ex:qv_acmg_svnindel_criteria_20250225.yaml
    rdf:type sphn:DataFile ;
    sphn:hasFilename "qv_acmg_svnindel_criteria_20250225.yaml
       ↪ " ;
    sphn:hasFileFormat "text/x-yaml" ;
    sphn:hasCreationDateTime "2025-02-25T00:00:00Z"^^xsd:
       ↪ dateTime ;
    sphn:hasSourceSystem ex:SomeSourceSystem ;
    sphn:hasFoundAt "file:///path/to/
       ↪ qv_acmg_svnindel_criteria_20250225.yaml"
```

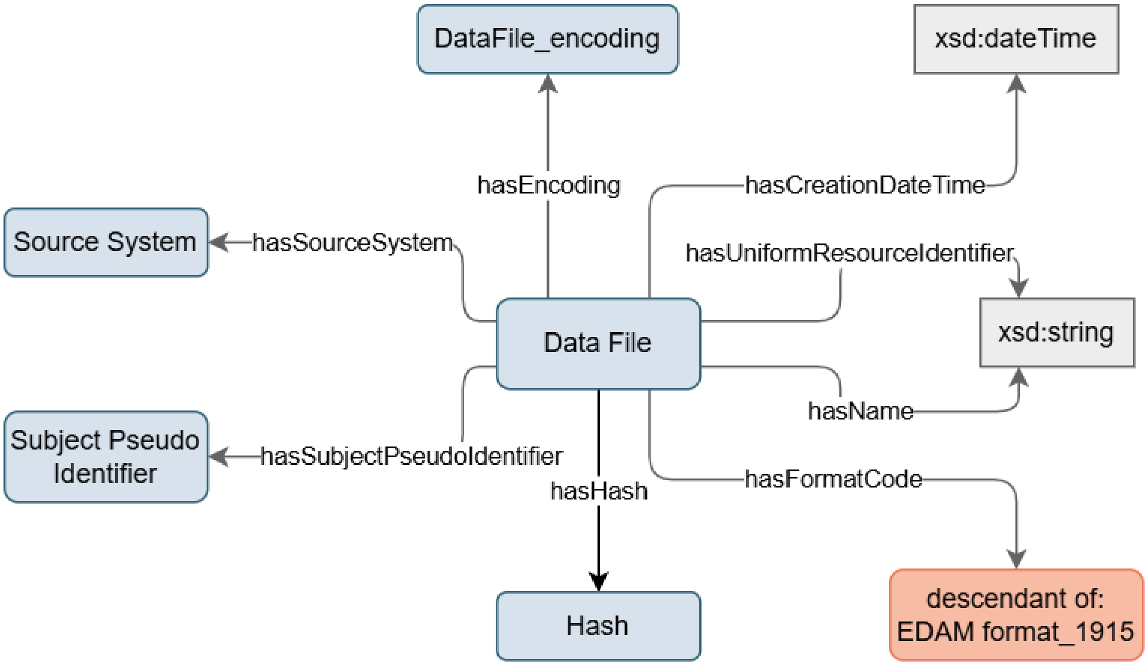


Figure 5: An overview of the `sphn:DataFile` concept in the SPHN RDF schema.

2.5 Validation case study

In the following case 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 protocol 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 (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`) (42).

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. A subset of key annotations used for QV is illustrated in **Figure 6**.

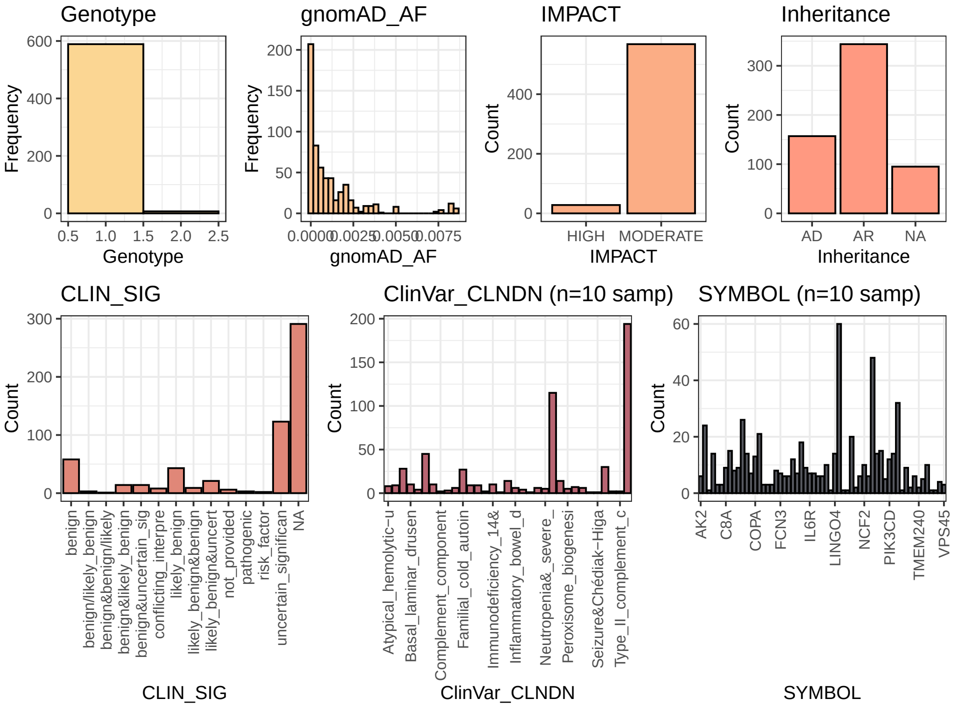


Figure 6: Annotated dataset describing variant calls from a disease cohort on chromosome 1. Key variables used in quality control and variant interpretation are shown. Axis which contain many term labels are down sampled to list every tenth label (n=10 samp). Genotype 1 is heterozygous, 2 is homozygous; gnomAD AF is allele frequency in the gnomAD population database, inheritance patterns AD Autosomal Dominant, AR Autosomal Recessive.

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 7**. The QV criteria were provided in YAML format in the file `qv_files/acmg_criteria.yaml` (see **Box 12**):

Box 12: 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
```

Our results, presented in **Figure 7**, show a 100% match between the manual and YAML-based methods, confirming that the criteria can be imported from YAML and applied programmatically with equivalent accuracy. Although accuracy still depends on the underlying implementation, the QC YAML file serves as a shareable, standalone resource that can be adapted across different pipelines or programming languages, thereby ensuring reproducibility of QV criteria.

The YAML criteria used here include `ACMG_PS1`, described as“the same amino acid change was a previously established pathogenic variant regardless of nucleotide change.” It includes `terms` such as “pathogenic,” applied to the `CLIN_SIG` (clinical significance) annotation field, and uses “or” logic. Additionally, `ACMG_PS3` describes well-established functional studies supporting a damaging effect on the gene product, with a user-defined inheritance pattern matching the genotype, and `ACMG_PS5` covers compound heterozygosity with at least one high-impact variant (per Ensembl VEP definitions).

The `ACMG_PM2` criterion specifies that the variant is absent from controls or present

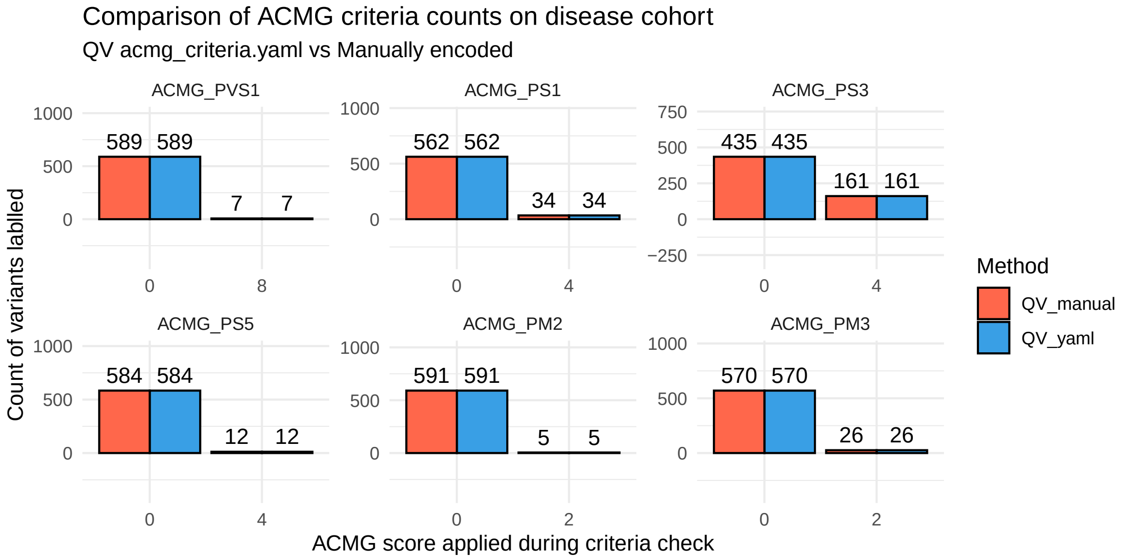


Figure 7: GuRu case study using an ACMG criteria subset, demonstrating a 100% match between manually encoded and standardised YAML-based methods (`qv_files/acmg_criteria.yaml`) for assigning pathogenicity scores.

at extremely low frequency in population databases such as gnomAD. For `ACMG_PM3`, the criterion checks for variants in trans with a pathogenic variant in recessive disorders; some overlap exists with PS5, as our filtering already treats “IMPACT = HIGH” similarly.

We omitted the PS2 criterion (requiring confirmed de novo status with no family history) due to the lack of parental data, and PS4 (indicating a significantly increased prevalence in cases compared with controls) as it was evaluated separately in a case-control analysis for this cohort.

Finally, **Figure** 8 illustrates the final annotation results for the test disease cohort, showing the number of criteria applied per subject and per variant. This facilitates the automatic retrieval of top candidate causal pathogenic variants using ACMG scoring methods (1; 5).

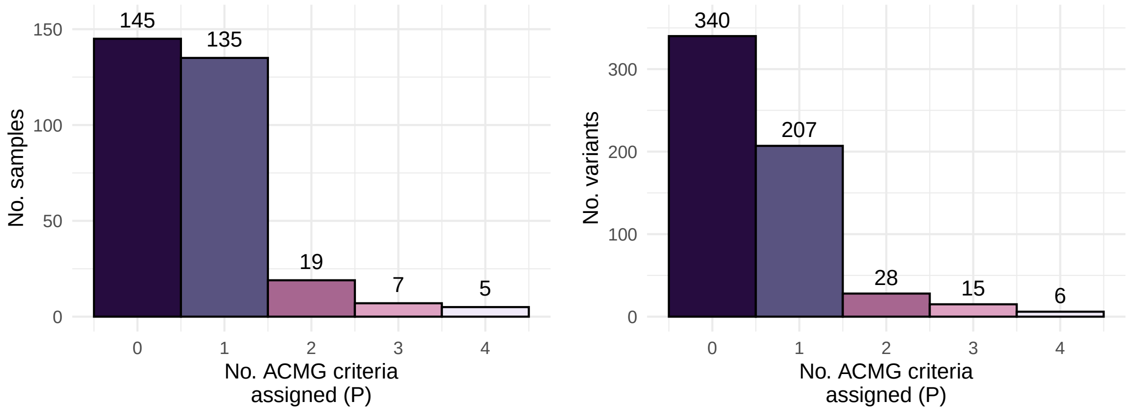


Figure 8: Final annotation interpretation for the test disease cohort, showing the number of criteria applied per subject (left) and per variant (right). This enables the automatic retrieval of the top candidate pathogenic variants.

2.6 Alignment with patient and public involvement needs

A key benefit of our framework is that it provides explicit, citable documentation of which genetic variants are included or excluded in a database or report. This transparency not only bolsters trust by allowing both analysts and patients to clearly see how filtering is performed, but it also facilitates meaningful Patient and Public Involvement (PPI). In essence, standardisation of QV protocols can be seen as an integral part, if not the first step, of an implementation method that rapidly translates new genetic research findings into clinical diagnostics. Given the well-documented time lag between research findings and their clinical application (43), such an approach ensures that patients receive the most relevant and up-to-date information about their individual genetic health risks.

For example, as illustrated in **Section 2.3.4**, both the analyst and the patient can confirm that the ACMG clinical guidelines on secondary findings have been applied. This transparent documentation enhances diagnostic traceability and aids lay understanding of genetic diagnostic processes, thereby increasing patient engagement and informed decision-making.

3 Discussion

3.1 Challenges and innovations

3.1.1 Avoiding pitfalls

In advancing omics research through multiblock data integration, it is imperative to standardise and optimise the use of QV (44). This need parallels the pitfalls seen in repeated measures analysis, where combining repeated measurements without appropriate controls may yield misleading conclusions (45). Similarly, in multi-omic integration, where data from DNA, RNA, and protein are merged, naively combining datasets without accounting for the unique characteristics of each data block can lead to erroneous interpretations. Warnings such as those concerning Simpson's paradox (46–48) underscore the necessity for sophisticated statistical frameworks that adjust for source-specific variations. As deep phenotyping and precision medicine evolve, standardised database formats become critical for genomics, and QV should not be an afterthought (49–51).

3.1.2 Applications in simple independent tests

Consider a multi-part analysis involving QV sets 1, 2, and 3, each representing a distinct GWAS experiment. Statistical methods such as the Aggregated Cauchy Association Test (ACAT) (52; 53) can then combine p-values across these tests, taking into account the direction and magnitude of effects. This aggregation not only increases the power to detect significant associations, especially when variant effects are heterogeneous, but also simplifies the interpretation of aggregated genomic data. Such methods are particularly useful when variants across different QV sets contribute variably to the phenotype (54).

3.1.3 Applications in complex data and multiblock fusion

Multiblock data fusion, an emerging field in statistics and machine learning championed by multi-omics, offers profound opportunities to unravel complex biological systems. By integrating multiple omics data types, DNA, RNA, protein, or clinical data, into a coherent analytical framework, researchers can uncover nuanced interdataset relationships that were previously obscured. Studies by Kong et al. (55) and Howe et al. (56) have demonstrated that complex signals may reside within single

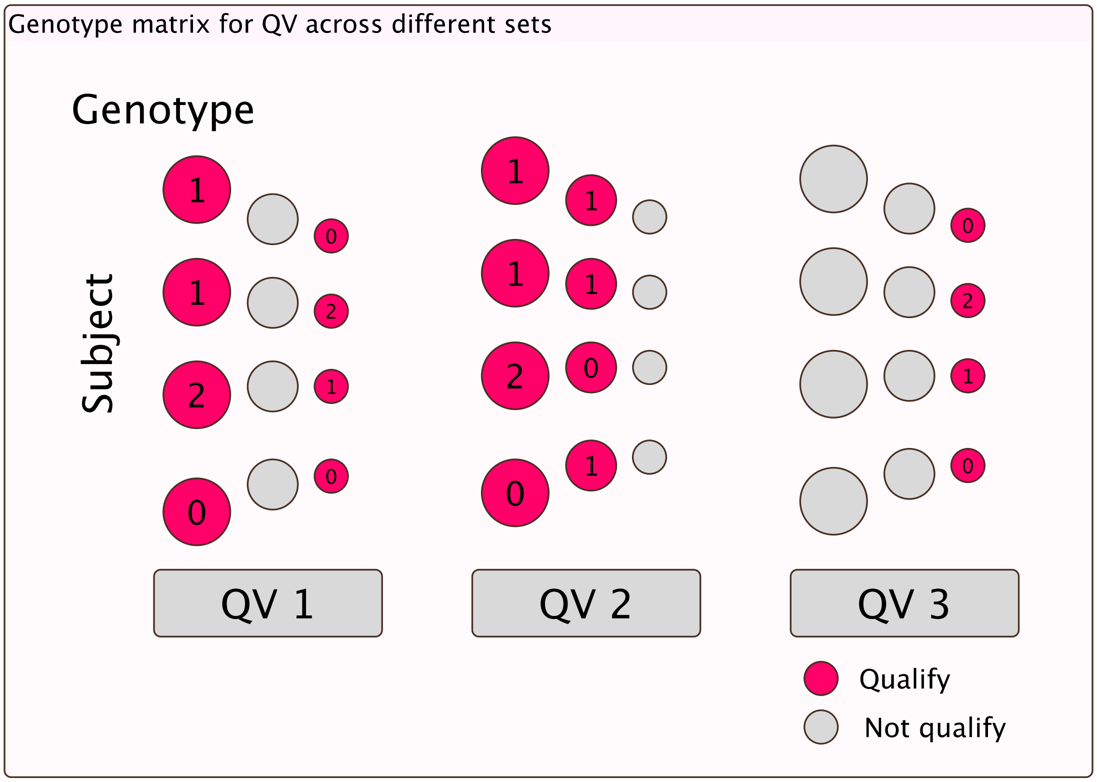


Figure 9: Genotype matrices for three layers of QV analysed in genetic studies. Each matrix represents a specific set, showing the genotypes of three individuals for three SNPs (SNP1, SNP2, SNP3). Genotypes of each SNP are coded as 0 (homozygous reference), 1 (heterozygous), and 2 (homozygous alternative). In the QV1 layer, SNP1 and SNP3 qualify as a QV (highlighted in red).

datasets, underscoring the value of advanced fusion techniques. We contend that standardised and optimised QV protocols not only advance omics research but also open up new theoretical domains. For instance, integrating various QV sets on a single data source may facilitate joint analysis of associational, causal, and counterfactual relationships alongside traditional analyses, ultimately preparing a unified dataset for complex multi-omic investigations.

3.1.4 Protocol development and standardisation needs

The complexity of multi-omic approaches necessitates clear protocols for merging data from different layers, ensuring that each component contributes meaningfully without conflating distinct signals. A standardised definition and reporting style for QV are crucial for rapid theory development, particularly when data are not publicly available and codebases are complex and nuanced. Detailed, explicit protocols that list standardised definitions and variables, such as our demonstrated example for QV1, will enhance reproducibility and foster new analytical frameworks.

3.2 Future directions and implications

3.2.1 Integration strategies

New publishing formats like Registered Reports promote transparency and reproducibility (57), and a similar approach for QV standardisation could expedite the clinical translation of genetic research. As machine learning and artificial intelligence models become increasingly capable of integrating vast multi-omic datasets, accurately formatted and curated QVs will be essential. Particularly in rare diseases, where raw data may be insufficient for robust model training, refined QV protocols can enrich feature representations, thereby improving predictive accuracy (58). The development of advanced QV protocols is critical for statistical methods that navigate complex, interrelated omic data blocks (44). We advocate for a refined, comprehensive use of QVs that goes beyond traditional single-stage filters to meet the sophisticated demands of modern multi-omic research.

3.2.2 Notation typical to GWAS, VSAT, and other statistical applications

We explore the notational use of QV in commonly used applications to demonstrate how the conceptual framework can accelerate adoption in theoretical domains. For example, in GWAS (8) the notation for the logistic regression model for estimating the probability of case status is given by:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum_{k=1}^{n} \beta_k x_{ik} + \beta_{\text{geno}} G_i$$

where p_i is the estimated probability that individual i is a case, β_0 is the intercept, β_k are coefficients for covariates x_{ik}, and β_{geno} is the genetic effect coefficient for the genotype G_i (coded as 0, 1, or 2). An explicit version incorporating QV notation is:

$$\text{logit}(p_i) = \beta_0 + \beta_1\,\text{sex}_i + \beta_2\,\text{log10}(\text{age})_i + \sum_{j=1}^{10} \beta_{2+j}\,\text{PC}_j^{(i)} + \sum_{k=1}^{n} \beta_{13+k}\,G_{\text{QV}_{i,v,k}}$$

where β_{13+k} represents the effect of each additional qualifying variant set k and

$G_{\mathrm{QV}_{i,v,k}}$ denotes the genotype for variant v in the k-th QV set for individual i.

Likewise, sequence kernel association test (SKAT) and its optimal unified version, SKAT-O, are used for association tests that accommodate multiple variants within a set (i.e. gene), accounting for their potentially differing directions and magnitudes of effects (59; 60). The logistic regression model for SKAT, taking into account the specific variants from the QV set, can be described as follows:

$$\log\left(\frac{P}{1-P}\right) = X_i\gamma + G_{\mathrm{QV}_{i,v}}\beta$$

where: P is the disease probability, γ is an $s \times 1$ vector of regression coefficients of covariates, β is an $m \times 1$ vector of regression coefficients for genetic variants, $G_{\mathrm{QV}_{i,v}}$ denotes the genotype values for all variants v in the QV set for individual i. The SKAT statistic is then:

$$Q_S = (y - \hat{\pi})^\top K(y - \hat{\pi})$$

where $\hat{\pi}$ is the vector of the estimated probability of y under the null model, and K is the kernel matrix defined as $G_{\mathrm{QV}} W G_{\mathrm{QV}}^\top$, with W being the diagonal weight matrix for the variants.

With these familiar examples established, we can consider more complex models where other variants outside of the main QV set can be assessed, $QV_{1,\ldots,n}$, which we describe in the next section. These sets can represent different categorisations or stratifications of genetic variants that might be relevant under varying analytical conditions or specific studies.

3.2.3 Conceptual framework and statistical representation

In GWAS, the transition from empirically testable variants (QV1) to theoretical Axiomatic Variants ($\mathrm{QV}_{\mathrm{ax}}$) marks a pivotal stage. The term $\mathrm{QV}_{\mathrm{ax}}$ refers to genetic variants that ideally conform to fundamental genetic principles and are considered correct by genetic doctrine. However, due to technological constraints and gaps in understanding, $\mathrm{QV}_{\mathrm{ax}}$ remains largely theoretical due to sequencing or detection difficulty. In contrast, QV1 consists of variants from $\mathrm{QV}_{\mathrm{ax}}$ that survive rigorous empirical filtering, applying standard GWAS pre-processing criteria (e.g. missing genotype data, MAF, Hardy-Weinberg equilibrium deviations, and individual missing data thresholds) to ensure data quality and relevance.

It is important to emphasise that we refer to unobserved or unknown variants, in the Bayesian sense, rather than VUS. The mathematical relationship between $\mathrm{QV}_{\mathrm{ax}}$

and QV1 can be expressed as follows:

$$\begin{aligned}
\text{TP} &= \text{QV}_{ax} \cap \text{QV1}, \quad \text{(true positives)} \\
\text{FN} &= \text{QV}_{ax} \setminus \text{QV1}, \quad \text{(false negatives)} \\
\text{Unknowns} &= |\text{QV}_{ax}| - |\text{TP}|,
\end{aligned}$$

where TP represents the true positives that are both theoretically ideal and empirically robust, FN represents false negatives (theoretical variants erroneously excluded), and Unknowns quantifies the theoretical variants remaining untested. This structured approach clarifies the trade-offs in WGS pre-processing by balancing data quality against the risk of overlooking significant genetic contributors.

In genomics, Bayesian statistics combines prior knowledge with empirical data to refine our understanding of the genetic variant landscape, particularly for variants beyond current empirical detection. We define a prior distribution $P(\theta)$ based on established data (e.g. mutation rates and population variant frequencies) and combine it with the likelihood $P(D \mid \theta)$ from the pre-processed genomic dataset (QV1) using Bayes' theorem:

$$P(\theta \mid D) = \frac{P(D \mid \theta)\, P(\theta)}{P(D)}.$$

This posterior distribution updates our initial beliefs with the observed data. The "unknowns" - the theoretically possible variants not detected in QV1 - are quantified as:

$$\text{Unknowns} = \int_{\theta \in \Theta_{\text{unobserved}}} P(\theta \mid D)\, d\theta,$$

where $\Theta_{\text{unobserved}}$ represents the parameter space of undetected variants. By accounting for both the observed/unobserved known QV and (estimating) theoretically possible but unknowable QV, we can drastically increase our confidence scores.

This brief demonstration of potential future directions concludes our discussion of the methodological framework and its practical applications.

4 Conclusions

We emphasise the critical importance of QV standardisation in genomics. By proposing a clear framework for integrating QV protocols into analysis pipelines, we demonstrate that systematic handling of these variables enhances reproducibility, accuracy, and efficiency in genetic studies. As genomic technologies and data complexities continue to evolve, robust, scalable, and adaptable QV protocols become ever more essential. Future work should extend these frameworks to accommodate emerging technologies and analytical challenges, thereby improving the fidelity and utility of genomic data interpretation across diverse applications.

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] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[18] Alexander I Young. Solving the missing heritability problem. *PLoS genetics*, 15 (6):e1008222, 2019.

[19] 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.

[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] Fernando Riveros-Mckay, Michael E Weale, Rachel Moore, Saskia Selzam, Eva Krapohl, R Michael Sivley, William A Tarran, Peter Sørensen, Alexander S Lachapelle, Jonathan A Griffiths, et al. Integrated polygenic tool substantially enhances coronary artery disease prediction. *Circulation: Genomic and Precision Medicine*, 14(2):e003304, 2021.

[22] Michael E Weale, Fernando Riveros-Mckay, Saskia Selzam, Priyanka Seth, Rachel Moore, William A Tarran, Eva Gradovich, Carla Giner-Delgado, Duncan Palmer, Daniel Wells, et al. Validation of an integrated risk tool, including

polygenic risk score, for atherosclerotic cardiovascular disease in multiple ethnicities and ancestries. *The American journal of cardiology*, 148:157–164, 2021.

[23] Luanluan Sun, Lisa Pennells, Stephen Kaptoge, Christopher P Nelson, Scott C Ritchie, Gad Abraham, Matthew Arnold, Steven Bell, Thomas Bolton, Stephen Burgess, et al. Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. *PLoS medicine*, 18(1):e1003498, 2021.

[24] Elaine T Lim, Peter Würtz, Aki S Havulinna, Priit Palta, Taru Tukiainen, Karola Rehnström, Tõnu Esko, Reedik Mägi, Michael Inouye, Tuuli Lappalainen, et al. Distribution and medical impact of loss-of-function variants in the finnish founder population. *PLoS genetics*, 10(7):e1004494, 2014.

[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] 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.

[27] 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.

[28] Daniel Greene, Genomics England Research Consortium, Daniela Pirri, Karen Frudd, Ege Sackey, Mohammed Al-Owain, Arnaud PJ Giese, Khushnooda Ramzan, Sehar Riaz, Itaru Yamanaka, et al. Genetic association analysis of 77,539 genomes reveals rare disease etiologies. *Nature Medicine*, 29(3):679–688, 2023.

[29] Daniel Greene, Sylvia Richardson, and Ernest Turro. A fast association test for identifying pathogenic variants involved in rare diseases. *The American Journal of Human Genetics*, 101(1):104–114, 2017.

[30] Rebeca Mozun, Fabiën N Belle, Andrea Agostini, Matthias R Baumgartner, Jacques Fellay, Christopher B Forrest, D Sean Froese, Eric Giannoni, Sandra Goetze, Kathrin Hofmann, et al. Paediatric personalized research network switzerland (swisspedhealth): a joint paediatric national data stream. *BMJ open*, 14(12):e091884, 2024.

[31] Clare Turnbull, Richard H Scott, Ellen Thomas, Louise Jones, Nirupa Murugaesu, Freya Boardman Pretty, Dina Halai, Emma Baple, Clare Craig, Angela Hamblin, et al. The 100 000 genomes project: bringing whole genome sequencing to the nhs. *Bmj*, 361, 2018.

[32] Every baby deserves access to genetic screening. *Nature Medicine*, 30(8):2095–2096, August 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-03227-9. URL https://doi.org/10.1038/s41591-024-03227-9.

[33] Alba Sanchis-Juan, Karyn Megy, Jonathan Stephens, Camila Armirola Ricaurte, Eleanor Dewhurst, Kayyi Low, Courtney E French, Detelina Grozeva, Kathleen Stirrups, Marie Erwood, et al. Genome sequencing and comprehensive rare-variant analysis of 465 families with neurodevelopmental disorders. *The American Journal of Human Genetics*, 110(8):1343–1355, 2023.

[34] Caroline F Wright, Patrick Campbell, Ruth Y Eberhardt, Stuart Aitken, Daniel Perrett, Simon Brent, Petr Danecek, Eugene J Gardner, V Kartik Chundru, Sarah J Lindsay, et al. Genomic diagnosis of rare pediatric disease in the united kingdom and ireland. *New England Journal of Medicine*, 388(17):1559–1571, 2023.

[35] Brynjar O Jensson, Gudny A Arnadottir, Hildigunnur Katrinardottir, Run Fridriksdottir, Hannes Helgason, Asmundur Oddsson, Gardar Sveinbjornsson, Hannes P Eggertsson, Gisli H Halldorsson, Bjarni A Atlason, et al. Actionable genotypes and their association with life span in iceland. *New England Journal of Medicine*, 389(19):1741–1752, 2023.

[36] Ahmad N Abou Tayoun and Alawi Alsheikh-Ali. A rapid whole-genome sequencing service for infants with rare diseases in the united arab emirates. *Nature Medicine*, 29(12):2979–2980, 2023.

[37] Linyan Meng, Mohan Pammi, Anirudh Saronwala, Pilar Magoulas, Andrew Ray Ghazi, Francesco Vetrini, Jing Zhang, Weimin He, Avinash V Dharmadhikari, Chunjing Qu, et al. Use of exome sequencing for infants in intensive care units:

ascertainment of severe single-gene disorders and effect on medical management. *JAMA pediatrics*, 171(12):e173438–e173438, 2017.

[38] Sebastian Lunke, Sophie E Bouffler, Chirag V Patel, Sarah A Sandaradura, Meredith Wilson, Jason Pinner, Matthew F Hunter, Christopher P Barnett, Mathew Wallis, Benjamin Kamien, et al. Integrated multi-omics for rapid rare disease diagnosis on a national scale. *Nature medicine*, 29(7):1681–1691, 2023.

[39] Monica H Wojcik, Gabrielle Lemire, Eva Berger, Maha S Zaki, Mariel Wissmann, Wathone Win, Susan M White, Ben Weisburd, Dagmar Wieczorek, Leigh B Waddell, et al. Genome sequencing for diagnosing rare diseases. *New England Journal of Medicine*, 390(21):1985–1997, 2024.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] Age K. Smilde, Tormod Næs, and Kristian Hovde Liland. *Multiblock Data Fusion in Statistics and Machine Learning: Applications in the Natural and Life Sciences*. Wiley, 1 edition, April 2022. ISBN 978-1-119-60096-1 978-1-119-60097-8. doi: 10.1002/9781119600978. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781119600978.

[45] J Martin Bland and Douglas G Altman. Correlation, regression, and repeated data. *BMJ: British Medical Journal*, 308(6933):896, 1994.

[46] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[47] Sewall Wright. The method of path coefficients. *The annals of mathematical statistics*, 5(3):161–215, 1934.

[48] Judea Pearl. *Causal inference in statistics: a primer.* John Wiley & Sons, 2016.

[49] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[50] The All of Us Research Program Genomics Investigators. Genomic data in the all of us research program. *Nature*, 627(8003):340–346, 2024.

[51] Soichi Ogishima, Satoshi Nagaie, Satoshi Mizuno, Ryosuke Ishiwata, Keita Iida, Kazuro Shimokawa, Takako Takai-Igarashi, Naoki Nakamura, Sachiko Nagase, Tomohiro Nakamura, et al. dbtmm: an integrated database of large-scale cohort, genome and clinical data for the tohoku medical megabank project. *Human Genome Variation*, 8(1):44, 2021.

[52] Yaowu Liu, Sixing Chen, Zilin Li, Alanna C Morrison, Eric Boerwinkle, and Xihong Lin. Acat: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, 104(3):410–421, 2019.

[53] Xihao Li, Zilin Li, Hufeng Zhou, Sheila M Gaynor, Yaowu Liu, Han Chen, Ryan Sun, Rounak Dey, Donna K Arnett, Stella Aslibekyan, et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature genetics*, 52 (9):969–983, 2020.

[54] Dylan Lawless, Ali Saadat, Mariam Ait Oumelloul, and Jacques Fellay. Archipelago method for variant set association test statistics. *medRxiv*, 2025. doi: 10.1101/2025.03.17.25324111. URL https://www.medrxiv.org/content/early/2025/03/17/2025.03.17.25324111.

[55] Augustine Kong, Gudmar Thorleifsson, Michael L Frigge, Bjarni J Vilhjalmsson, Alexander I Young, Thorgeir E Thorgeirsson, Stefania Benonisdottir, Asmundur Oddsson, Bjarni V Halldorsson, Gisli Masson, et al. The nature of nurture: Effects of parental genotypes. *Science*, 359(6374):424–428, 2018.

[56] Laurence J Howe, Michel G Nivard, Tim T Morris, Ailin F Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Penelope A Lind, Teemu Palviainen, Matthijs D van der Zee, et al. Within-sibship gwas improve estimates of direct genetic effects. *BioRxiv*, pages 2021–03, 2021.

[57] Christopher D Chambers, Eva Feredoes, Suresh Daniel Muthukumaraswamy, and Peter Etchells. Instead of" playing the game" it is time to change the rules: Registered reports at aims neuroscience and beyond. *AIMS Neuroscience*, 1(1): 4–17, 2014.

[58] Andrew G Barto, Richard S Sutton, and Charles W Anderson. Looking back on the actor–critic architecture. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1):40–50, 2020.

[59] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[60] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, David C Christiani, Mark M Wurfel, and Xihong Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012.