Conceptualising qualifying variants for genomic analysis

Dylan Lawless*[1] and Ali Saadat[2]

[1]Department of Intensive Care and Neonatology and Children's Research Centre, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland.
[2]Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Switzerland.

February 25, 2025

Acronyms

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

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

Qualifying variant (QV)s are genomic alterations selected through specific criteria during genomic processing pipelines. These variants are essential for downstream analysis in genetic research and clinical diagnostics. This paper explores the application and conceptualisation of QVs not merely as filtering criteria but as a dynamic concept crucial for various genomic analysis scenarios. Generally, the selection of QVs follows well-established best practices in variant classification and reporting standards (1–5), as well as standardized workflows (6–8). However, a standard guide for QV themselves remains missing. Polygenic Risk Score (PRS) reporting standards have been developed to encourage their application and translation as well as open cataloguing for reproducibility and systematic evaluation (9; 10). We propose an equivalent is beneficial for QV.

The choice of QV thresholds often depends on the specific context of the research or clinical needs. For instance, Genome Wide Association Test (GWAS) might prioritise common variants, Variant Set Association Test (VSAT) might prioritise rare variant collapse, and clinical genetic reports may focus on rare or novel variants. Therefore, QVs are categorised by the extent and nature of the filtering or quality

control they undergo, tailored to the research or clinical requirements. **Figure 1** demonstrates a typical WGS and VSAT analysis pipeline, showing QVs as sequential and potentially piped protocol steps.

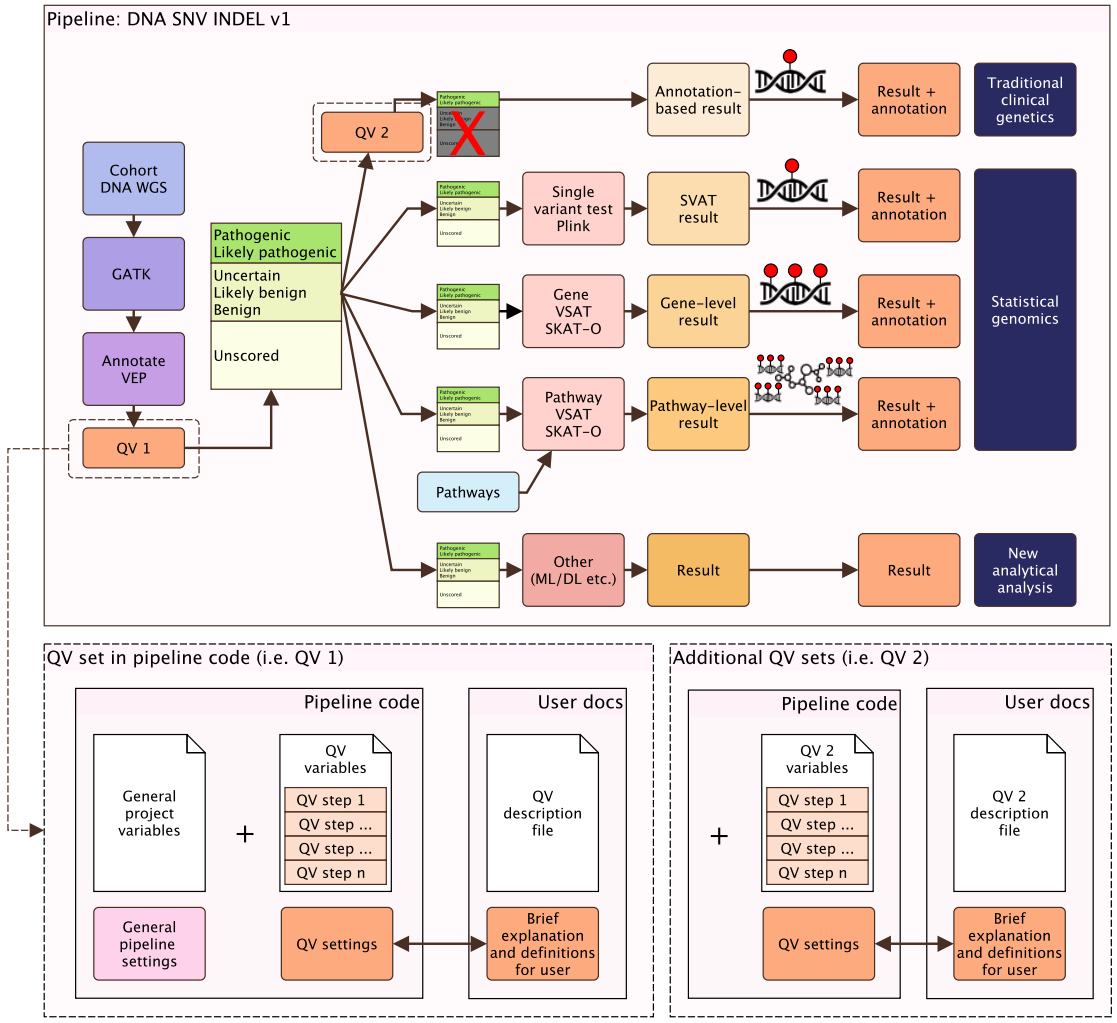

Figure 1: Summary of the example application design DNA SNV INDEL v1 pipeline. QV1 and QV2 are shown as sequential and potentially piped protocol steps. QV files used in a pipeline. The description file contains brief summary of how or why a step is used in the QV set (not mandatory). The variables file contains the necessary values of a QV setting (mandatory). These variables are loaded by the analysis pipeline. This illustration highlights a single stage in the QV1 set (i.e. step 10 in our example WGS analysis pipeline where the GATK VQSR method is applied). The full pipeline illustrated in the top panel simplifies this process under the QV1 icon.

The common approach to representing QV steps are illustrated in **figure 2**. This style simplifies the variant filtering process where each layer may arise from different stages of a pipeline. The raw omic data can be processed into a multi-use analysis-

ready format, as illustrated in **figure 3**. Initial QV steps can include Quality Control (QC) filtering. After variant annotation, further QV steps can be applied on this new information as shown in **figure 3**.

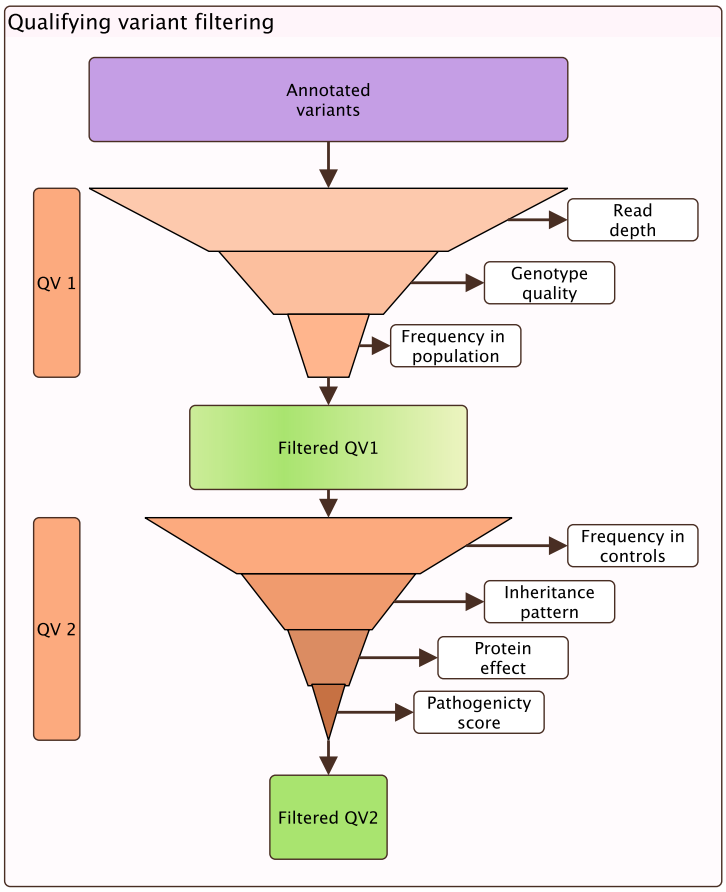

Figure 2: Illustration of the qualifying variant workflow. This figure summarises the conceptualised variant filtering step. This style is relatively common. In reality, we observe that each layer of the filters comes from disparate stages of a pipeline.

Povysil et al. (11) previously provided a tangible example of QV for variant collapsing analyses for complex traits. Cirulli et al. (12) reported one of the first variant collapse analysis and introduced the QV concept. However, we have no a standardised framework for presenting QV themselves. We detail four typical applications of QV sets:

1. **QV passing QC only**: Generates large datasets, e.g. 500,000 variants per subject, used in GWAS or Whole Genome Sequencing (WGS) pre-processing.

2. **Flexible QV**: Balances between QC and false positives. For instance, fewer than 100,000 variants per subject in preparation for rare variant association testing.

3. **QV for rare disease**: Produces smaller datasets after stringent filtering, e.g. 1,000 variants per subject, such as pre-processing to target known genes or a single causal variant in single-case genetic reports.

4. **Known disease panel QV set**: A well known gene panel with pathogenic variants, e.g. the American College of Medical Genetics and Genomics (ACMG) Secondary Findings (SF) set, recommended for clinical reporting (13).

Two exemplary applications of QVs are in clinical genetics reporting and GWAS. In clinical genetics single-case analysis, QVs may be selected from a list of disease-causing genes identified by an expert panel. Variants within these genes can be categorised based on their potential pathogenicity into variants of unknown significance (VUS), or as known, candidate, or causal variants pending further analysis. In GWAS, QVs generally refer to consensus variants that have undergone standard QC procedures to ensure their statistical suitability for the main analysis. The rigorous selection and categorisation of QVs in genetic research and diagnostics are crucial for accurately reporting and reproducing such studies, underscoring the importance of QV criteria, which can sometimes be more critical than the choice of analysis pipeline itself (14).

2 Background problem and proposed solution

As study sizes surpass the 1,000,000 subjects milestone (15; 16), the shift towards WGS over genotyping has become standard. This transition enables the inclusion of rare variants in GWAS and VSAT, allowing for more comprehensive analyses of complex traits (17; 18). QV protocols are essential in data cleaning and preparation, serving as a critical step in ensuring the integrity of data analysis. While often grouped under the single term "QV" for simplicity, the processes involved actually span various stages of a pipeline and originate from diverse steps or sources. **Figure 4** shows the structural framework of a variant's features that may trigger specific QV protocols, highlighting both pre-existing metadata and annotations added post-variant calling.

Moreover, complex analyses often necessitate multiple processing streams that merge into a cohesive analysis. This multifaceted approach to QV becomes apparent in multi-component analyses, which require the integration of two or more data sets. A standardised QV format will allow for the use of various QV sets, each based on

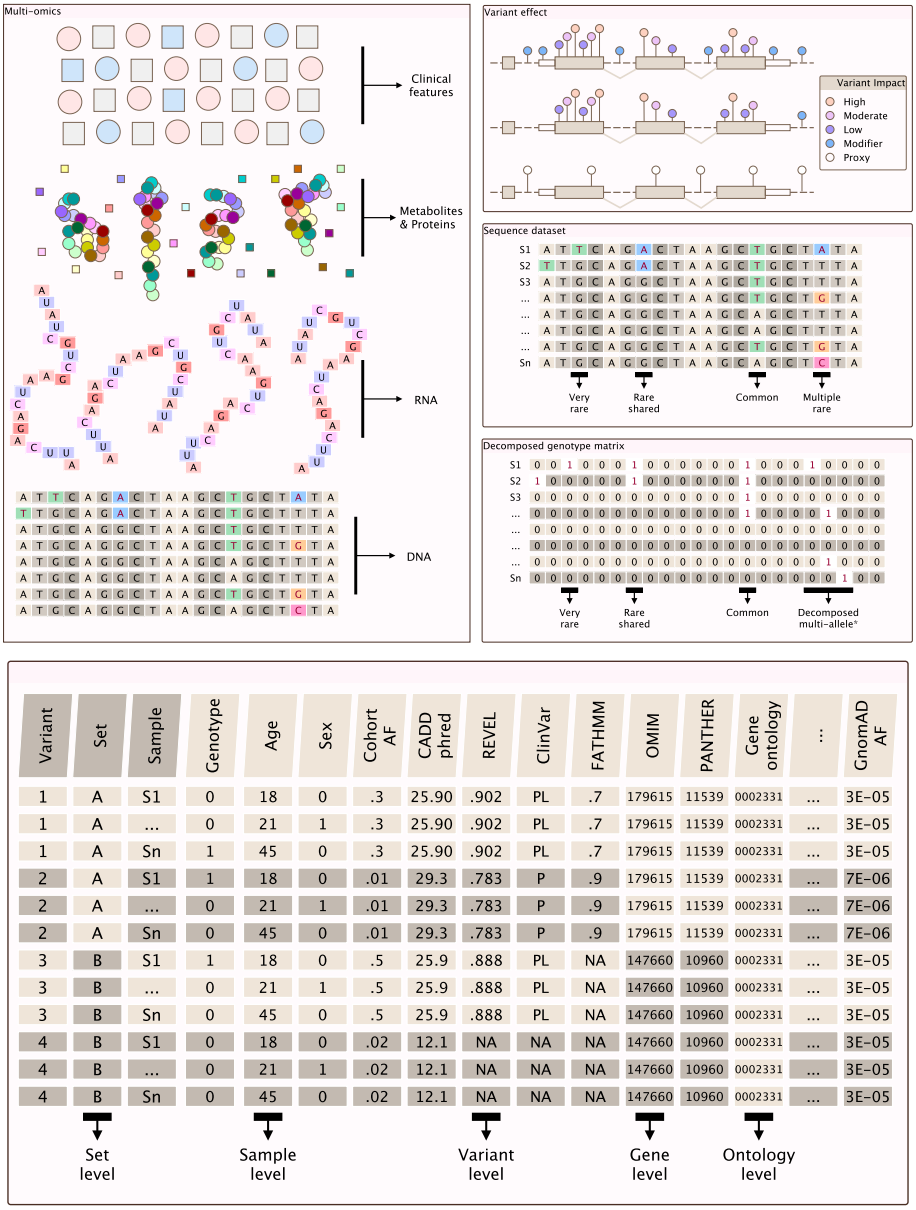

Figure 3: Top: From raw omics to data matrix. We focus on DNA variants in QV but the same concept applies to other datasets. Bottom: The initial variant detection pipeline generlly requires QC and filtering rules that are the first QV steps. Once complete, annotation of variants can follow. Further QV steps can be run based on these new annotations.

potentially different filters and variables, yet provides a common foundation to ensure consistency and validity across disparate data streams

Unsurprisingly, the term QV is often ambiguously used across different contexts within genomic studies, necessitating a clear definition for each application. Moreover, while QVs are typically perceived as a set of filters and algorithms to remove benign or

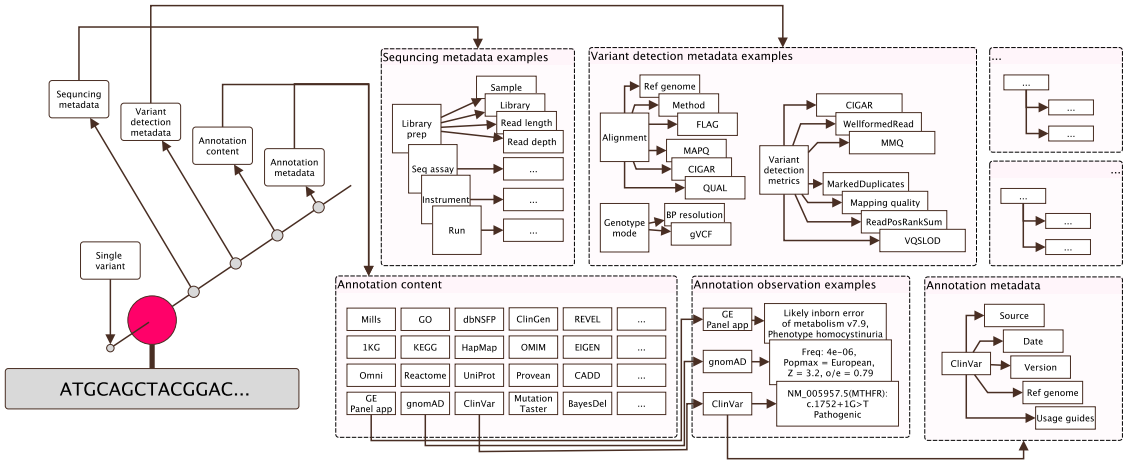

Figure 4: This illustration shows the structural framework of an annotated variant in relation to the features used for qualification. For every individual variant, a number features are capable of triggering QV protocols. The diagram highlights a select group of these features, showcasing both pre-existing metadata established before data generation and annotations applied after variant calling.

unrelated variants, they actually encompass many complex steps distributed throughout the entire analysis pipeline, and not necessarily confined to a single step. This dispersion of QV steps challenges the conventional view and highlights the need for a flexible definition that not only encompasses their common uses but also acknowledges their implementation across multiple stages of genomic analysis.

The complex, multi-step nature of QVs often goes unrecognised by those outside the field of bioinformatics. This makes it challenging to share knowledge across disciplines for more advanced tasks and underscoring the importance of a clear and comprehensive understanding of QV protocols.

By introducing a new vocabulary and a standard reference model for QVs, we aim to clarify the concept and improve the communication and methodological discussions around QVs. We therefore define and exemplify several common sets of QVs, illustrating their potential configurations and roles within analysis pipelines:

1. We demonstrate the theoretical pipelining of QV sets.

2. We outline how standardised QV sets can be established for specific analytical scenarios.

3. We highlight that QVs are integral throughout the analysis pipeline, not merely as an end-stage addition but as essential components distributed across the process.

Accordingly, the QV framework should furnish structured definitions that are both human and machine-readable to standardise the selection and interpretation of variants across diverse genomic studies. The methodology promotes efficient and precise variant detection and interpretation, which are essential for advancing both research and clinical diagnostics. Adhering to these structured analysis criteria aligns with the FAIR principles of findability, accessibility, interoperability, and reusability (19).

3 Advanced applications and case study

3.1 How and why QV are used

Here we provide an in-depth look at specific scenarios where QVs have inadvertently been crucial, such as in GWAS, WGS, and clinical genetics. We note that application of QVs would be appropriate in large-scale studies and rare disease research, such as sophisticated risk models that integrate clinical and genomic data, which enhance predictive accuracy in large, well-defined cohorts (20–22). Likewise, QV protocols can aid reproducibility across studies, especially in isolated or specific populations. This includes the analysis of complex signals within well-defined cohorts which can significantly aid rare disease studies (23).

3.2 Example application of qualifying variants in WGS analysis

Several QV protocols can be piped together to create increasingly filtered datasets to match the needs at a certain stage of analysis. It is also typical that different analyses from QVs sets are used and the final results from each step are merged to cover multiple scenarios. For example, a complex analysis pipeline might use all of `QV SNV/INDEL` + `QV CNV` + `QV structural variation` + `QV rare disease known` + `QV statistical association QC`, merged for a thorough multi-part analysis to reach the final combination of (1) a joint cohort disease association and (2) several single case analysis results. We propose an example focusing on a SNV/INDEL pipeline using two QV sets named `QV SNV INDEL v1` and `QV SNV INDEL v2`. The QV sets are illustrated in **box 1** and would be described in an analysis pipeline as follows:

"A cohort of patient WGS data was analysed to identify genetic determinants for the clinical diagnosis of phenotype X. This pipeline was concerned with WGS germline

short variant discovery (SNVs + Indels) and interpretation. First, a flexible QV set (v1) was used for cohort-level statistical genomics and second a rare disease QV set (v2) was used for single-case analysis. (**1**) Data was processed with the `pipeline DNA SNV INDEL v1` pipeline, which implements (**a**) `QV SNV INDEL v1` criteria, resulting in the prepared dataset `dataset DNA SNV INDEL v1`. (**b**) The dataset was subsequently analysed in combination with other modules including `PCA SNV INDEL v1` and `statistical genomics v1` to complete statistical analysis on a joint cohort. (**2**) Next, the prepared dataset (from step 1a) `Dataset DNA SNV INDEL v1` was processed further with more strict filtering using `QV SNV INDEL v2` to identify previously known causal genetic variants for each patient based on disease-gene panel and curated evidence sources, resulting in `Dataset DNA SNV INDEL v2` and final interpretation in `single case report SNV INDEL v1`."

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 for single patient.

4 Methodological framework

4.1 Three steps to a QV protocol

We introduce a simple framework for the effective use of QV protocols. We use three steps to fulfil the needs of a pipeline as illustrated in **figure 1**:

1. **Variables**: The criteria variables responsible which are sourced as part of a pipeline, as shown in **example box 2**.

2. **Description**: The description of each step as part of an overall QV set, as shown in **example box 3**.

3. **Source code**: The variables file can be sourced in pipeline code, as shown in **example box 4**.

The practical application of our framework efficiently manages QV-specific variables, such as Allele Frequency (AF) thresholds, distinct from general pipeline settings. These variables can be imported and handled separately within the workflow to maintain clarity and specificity. We first provide a detailed example using Variant Quality Score Recalibration (VQSR) to show a full step in application of this method in a real-world genomic analysis scenario. Later, we also illustrate this approach using workflow managers like Snakemake or Nextflow, demonstrating the separation and integration of these variables to streamline genomic processing tasks.

Individual steps in QV criteria can have multiple types. For organisation in our protocols we suggest simple labels such as "QC" and "filter". (1) Filtering thresholds such as AF (e.g. >0.1 in cohort, <0.1 in gnomAD). These might be directly applied in place to remove all affected variants. (2) Multiple steps with annotation labels such as QC flags which do not remove variants but allow for downstream dissensions which which depend on multiple QV criteria. Thus, in a QC protocol a single step might run and filter all variants from criteria (1 "filter") and another filtering step might depend on several combined criteria (2 "QC") which were added upstream.

4.2 Detailed example QV variables

As a detailed example, we select the step **[QC]** `vqsr` from the example QV set `QV SNV INDEL v1` to illustrate the variables sourced during the pipeline. The following code snippet shows the from variables sourced during VQSR. First we set the mandatory QV varaints in **box 2**. Second, we provide the optional description in **box 3**. Third, we apply the variables in the source code **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 description file (highlighting VQSR steps)

1. [**QC**] `fastp` The tool fastp is used for ...
2. [**QC**] `collectwgsmetrics` BAMs that fail are ...
3. [**QC**] `rmdup_merge` is used to mark ...
4. [**QC**] `haplotype_caller` used `-ERC GVCF` mode for ...
5. [**QC**] `QC_summary_stats` is used to log QC ...
6. [**QC**] `vqsr` employs Variant Quality Score Recalibration (VQSR) using GATK. Includes the use of key metrics such as Quality by Depth (QD), Mapping Quality (MQ), and Read Position Rank Sum Test (ReadPosRankSum) to filter variants. The setting for SNPs are:
   - **VQSR SNP Mode - HapMap:** known=false, training=true, truth=true, prior=15.0. Used as a high-confidence reference set for training the recalibration model.
   - **VQSR SNP Mode - Omni:** known=false, training=true, truth=false, prior=12.0. Provides additional training data derived from Omni genotyping arrays.
   - **VQSR SNP Mode - 1000G:** known=false, training=true, truth=false, prior=10.0. Utilizes data from the 1000 Genomes Project to inform the model on common SNP variations.

- **VQSR Annotations - QD, MQ, MQRankSum, ReadPosRankSum, FS, SOR:** Annotations are metrics used to predict the likelihood of a variant being a true genetic variation versus a sequencing artifact. They include quality by depth, mapping quality, mapping quality rank sum test, read position rank sum test, Fisher's exact test for strand bias, and symmetric odds ratio of strand bias.
- **VQSR Truth Sensitivity Filter Level:** 99.7. Specifies the percentage of true variants to retain at a given VQSLOD score threshold, set here to capture 99.7% of true variants.

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
```

4.3 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

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

Box 6: Example worflow manager - 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

```
# 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}
    ""
```

5 Examples of real-world QV applications

5.1 Discovery research

Greene et al. (24) developed what can effectively be considered a prime example of QV standardisation with their "Rareservoir," a relational database schema optimised for handling genetic data from rare disease studies. This database optimises storage and query speed by focusing on rare variants - those with a Minor Allele Frequency (MAF) below 0.1% - thus reducing data size by about 99%. Variants are stored as 64-bit integers ("RSVR IDs"), organised by genomic position to support efficient location-based queries. Additional data include MAFs from gnomAD, CADD pathogenicity scores, and impact predictions per the Sequence Ontology. Genetic impacts are encoded into a 64-bit integer ("CSQ ID"), where each bit corresponds to a specific gene function impact, ranked by severity. This structure enables precise database queries. By employing the Bayesian genetic association method BeviMed, initially described by Greene et al. (25), the study effectively inferred associations between genes and

rare disease phenotypes, demonstrating the capacity to handle and analyze complex genetic datasets effectively. The protocol by Greene et al. (24) can be produced in the QV format, easing the interpretation and reproducibility.

We are applying the QV framework in "SwissPedHealth". This is a joint paediatric national data stream where one of the objectives is to investigate rare/unknown disease using a multiomic approach, including WGS, RNAseq, proteomics, and clinical data (26). We optimise the processing pipelines by using consensus raw datasets (e.g. WGS in patients and families), which is annotated and filtered to several QV levels. The resulting pre-processed datasets are suitable for a range of analysis including GWAS, VSAT, single-case clinical genetics reports, machine learning, and joint multiomic analysis.

National-scale projects reuse data many times and naturally require the use of QV, currently in the traditional format; mixed throughout pipelines without explicit distinction. The Genomics England 100,000 Genomes Project performs central automated analysis with interpretation and clinical reporting (27). This data is then used in many subsequent research projects. One significant application of QV protocols is in genomics-based newborn screening, which has been gaining traction as a pivotal healthcare innovation (28).

5.2 Rapid diagnostics

Screening for known disease is by default, based on searching for a known QV set. The use of a formalised standard would allow for consistency and reliability for stakeholders. Genome analysis in neurodevelopmental disorders in 465 families found causal variants in 36% of 489 affected individuals (29). The DDD study involved over 13,500 families; approximately 41% of probands received a genetic diagnosis (30). Genomic lifespan association in Iceland was reported by Jensson et al. (31), which included 57,933 participants, identifying 2,306 individuals with actionable genotypes linked to a decrease in median lifespan by approximately three years for carriers.

We propose that the use of standardised QV can aid in developing rapid genomic diagnostics. In UAE, rapid whole-genome sequencing - (32) allowed for rapid whole-genome sequencing with a turnaround of 37 hours on average. Meng et al. (33) reported on 278 critically ill infants; molecular diagnosis achieved in 36.7% of cases, with higher rates (50.8%) in critical trio exome cases. They reported an impacted medical management in 52.0% of diagnosed cases. Lunke et al. (34) reported in the national scale multi-omics for rare diseases, involving 290 critically ill infants and

children; diagnostic yield from WGS initially at 47%, increased to 54% with extended analysis. Altered critical care management occurred for 77% of diagnosed cases.

5.3 Complex variant calls

Additionally, using the QV framework for different analysis types, such as SNV-INDEL, copy-number variants, and structural variants, means that simple QV IDs can be used for database reporting. Researchers can then easily query whether or not additional analysis is expected to reveal more findings. For instance, Wojcik et al. (35) used genome sequencing in 822 families with rare monogenic diseases, achieving a diagnostic yield of 29.3%. They focused on broader genomic coverage including structural and non-coding variants, identifying causative variants in 8.2% of cases previously undetected by exome sequencing.

5.4 Secondary findings

The ACMG SF v3.2 list, exemplifies a well-defined and impactful application of QV standards in genomic medicine (13). This list specifies gene-phenotype pairs recommended for reporting as secondary findings during clinical exome and genome sequencing. Such standardisation not only streamlines the identification of clinically actionable genetic information but also enhances the consistency and quality of genomic data interpretation across different settings. The key limitation is that this dataset is relatively unstructured meaning that each implementation requires extensive work and care on the part of front-line analysts. 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 approach to curating the SF list ensures that it remains a reliable resource for opportunistic screening in a clinical context.

Table 1 lists the first two entries (transposed) from the ACMG SF list showcasing specific genes associated with cardiovascular phenotypes. We then represent this data in a standardised QV format in **box 8 - 9**, which can be used in any variant filtering program as demonstrated. In bioinformatics pipelines, specifying QV sets consistently, such as ACMG SF v3.2 allows patients to receive the most relevant and up-to-date information regarding their genetic health risks, without missing out on simple checks due to the burden of manually implementing new QV.

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
```

6 Challenges and innovations

6.1 Avoiding pitfalls

In the pursuit of advancing omics research through multiblock data, we recognize the imperative need to standardise and optimise the use of QV. This need mirrors the simple pitfalls in the analysis of repeated measures, where combining repeated measurements without appropriate controls can lead to misleading conclusions (36). So we must approach the integration of complex QV layers with rigor.

In multi-omic integration, where data from various layers such as DNA, RNA, and protein are fused, the naive merging of data without considering the unique source and nature of each data block can similarly mislead. The warning from Bland and Altman (36) about repeat data, or Simpson's paradox, where aggregated data can obscure real relationships, underscore the necessity for sophisticated statistical frameworks that acknowledge and adjust for the intricacies of source-specific variations. Once acknowledged, these features can be addressed potentially with existing methods (37–39). Increasingly deep phenotyping and precision medicine with omic data are reshaping data integration strategies. Standardised database formats are thus critical for genomics and QV should not be an afterthought (40–42).

6.2 Applications in simple independent tests

An example of a multi-part analysis with sets QV sets 1, 2, and 3, is illustrated in **Figure 5**. For simplicity, we can assume that each set represents one GWAS. The outcomes of each test can subsequently be combined with statistical methods such as the Aggregated Cauchy Association Test (ACAT) (43; 44). It becomes possible to aggregate and compare these results from separate tests. This process combines p-values across different analyses or variant sets, accounting for the directions and magnitudes of the effects. It not only enhances the power to detect significant associations, especially when variants have heterogeneous effects but also simplifies the interpretation of aggregated genomic data. By employing ACAT, we can synthesize findings from multiple QV filters applied to the same genomic dataset, leading to a more comprehensive understanding of the genetic architecture of traits under study. This method is useful in scenarios where variants across different QV sets may contribute in varying degrees to the phenotype, allowing for a nuanced analysis that respects the complexity of genomic data. The combination method must be adapted

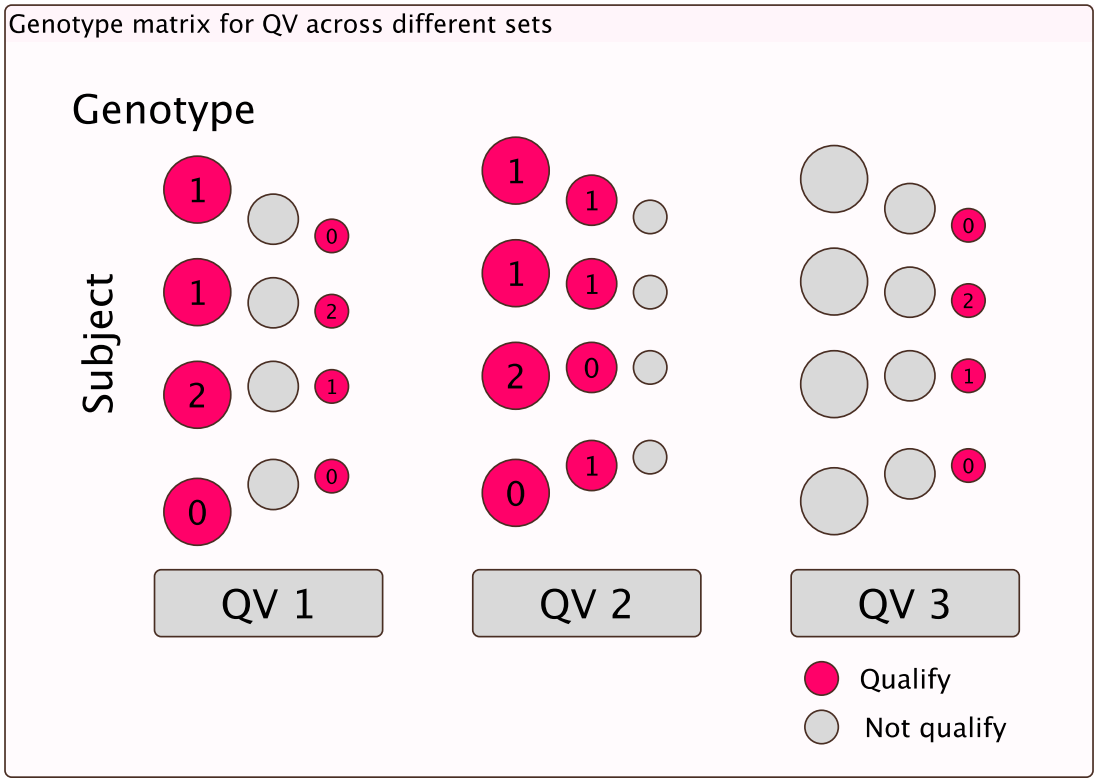

Figure 5: Genotype matrices for three different layers of QV analysed in genetic studies. Each matrix represents a specific set, showing the genotypes of three individuals for three SNPs (SNP1, SNP2, SNP3). In the QV1 layer, SNP1 and SNP3 qualify as a QV (highlighted in red). The genotypes are for QV are coded as 0 for homozygous reference, 1 for heterozygous, and 2 for homozygous alternative.

in scenarios where variants (or collapsed variant sets) do not overlap. We have previously provided methods to plot the results of such analysis with archipelago (cite pre-print instead of package).

6.3 Applications in complex data and multiblock fusion

Multiblock data fusion is an emerging yet nascent field in statistics and machine learning which is championed by multi-omics. The interplay between statistical theory and machine learning unveils profound opportunities for advancing our understanding of complex biological systems. This approach harnesses the power of diverse data types through sophisticated fusion techniques that integrate multiple blocks of omics data - be it DNA, RNA, protein, or clinical data - into a coherent analytical framework. Such integration not only enhances the resolution at which we understand disease

mechanisms but also refines our predictive capabilities across different scales of biological organisation. By applying advanced statistical models researchers can uncover nuanced relationships within and between datasets that were previously obscured. Kong et al. (45) and Howe et al. (46) have previously shown how complex signals can exist within single datasets.

These methods allow for a detailed exploration of how different biological signals interact, offering a richer, more comprehensive view of the genomic landscape. As these techniques evolve, they promise to break new ground in predictive modeling and theoretical biology, providing insights that are as profound as they are essential for precision medicine and personalised health interventions.

We contend that the term "QV", when standardised and optimised for advanced multi-stage use rather than simplistic, single-stage filters, not only advances omics research but also opens up unexplored theoretical domains. This includes a multi-dimension analysis of a single data source through exploring new concepts; for example, such jointly analysing probative variants (potentially axiomatically-causal with missing evidence), associational, causal, and counterfactual queries, in combination with traditional analyses that integrate other omic markers like RNA and protein abundance. Sophisticated QV applications that combine various sets of QVs on a single data source may prepare the correct joint dataset for such complex analyses. The resulting mixed-up mixed model requires new frameworks.

By deploying a variety of QV protocols simultaneously on a single dataset, we orchestrate a multi-dimensional analysis that spans the full spectrum of genomic inquiry. This integrated approach allows for the combination of various QV protocols tailored to the specifics of the dataset, engaging different types of data analyses that can range from genetic variations to complex disease markers and beyond. The integration of these diverse analytical layers facilitates a comprehensive examination of genetic factors on both individual and cohort levels, promoting understanding that could propel genetic insights. This complex interplay between multiple QV sets catalyses the advancement of new theories in multi-omic research.

6.4 Protocol development and standardisation needs

These complex approaches requires a clear protocol for merging data across different layers, ensuring that each contributes meaningfully to the unified model without conflating their distinct signals. Therefore, a standardised definition and reporting style for QV are crucial for the rapid development of new theories, especially in scenarios

where data may not be publicly available, and codebases are complex. The nuanced and widespread steps of QV across lengthy pipelines will benefit from being reported explicitly as a protocol with a detailed list of definitions and variables, building on our demonstrated examples for one such set, QV1.

7 Future directions and implications

7.1 Integration strategies

We consider the impact of new publishing formats like Registered Reports on the field of genomics, promoting transparency and reproducibility (47). This kind of approach will be crucial as we develop increasingly sophisticated machine learning and artificial intelligence models capable of integrating vast multi-omic datasets. The potential for these models to unravel complex biological phenomena is immense, yet the challenge remains in assembling sufficient training data. Particularly in the realm of rare diseases, the raw data from human cases potentially do not meet the extensive needs of these advanced models. The embeddings or feature representations derived from raw data may be insufficient for training robust models; however, properly formatted and curated QVs may enrich these representations, enhancing the potential for accurate model training. If so, the accurate and strategic application of QVs becomes essential. By effectively identifying key data through refined QV protocols, researchers can enhance the accuracy and efficacy of predictive models, opening up new avenues for significant biological discoveries.

The need for advanced QV protocols that can effectively manage such complexity is critical, particularly in the development of statistical methods designed to navigate the intricate relationships within and across diverse omic data blocks. A standardised and nuanced application of QVs, detailed through explicit protocols and definitions, is fundamental for the evolution of new analytical frameworks. Therefore, we advocate for a more refined and comprehensive use of QVs, advancing beyond traditional single-stage filters to meet the sophisticated demands of modern multi-omic research.

7.2 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, based on their genotypic and covariate data, β_0 is the intercept, β_k are the coefficients for the covariates, x_{ik} represents the covariate values for the i-th individual, β_{geno} is the coefficient for the genetic effect, G_i is the genotype of the i-th individual, coded as 0, 1, or 2 (representing the number of minor alleles).

The following version shows the explicit GWAS model with QV notation:

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

where: β_1 adjusts for sex (1 if male, 0 if female), β_2 adjusts for the log-transformed age in days, β_3 to β_{12} correspond to the first ten principal components, adjusting for population stratification, β_{13+k} represents the effects of the genotype on the phenotype for each additional qualifying variant set k, $G_{\text{QV}_{i,v,k}}$ denotes the genotype of the i-th individual for the v-th variant in the k-th QV set.

Likewise, SKAT and its optimal unified version, SKAT-O, are now popular methods for gene-based association tests that accommodate multiple variants within a gene or variant set while accounting for their potentially differing directions and magnitudes of effects (48; 49). The logistic regression model for SKAT, taking into account the specific variants from the QV set, can be described as follows:

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

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, $G_{\text{QV}_{i,v}}$ 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 $G_{\text{QV}} W G_{\text{QV}}^{\top}$, 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.

7.3 Conceptual framework and statistical representation

In GWAS, the transition from empirically testable variants (QV1) to theoretical Axiomatic Variants (QV_{ax}) marks a pivotal stage in genetic research. QV_{ax} comprises genetic variants that ideally conform to fundamental genetic principles and are thus considered correct by genetic doctrine. However, due to technological constraints and gaps in genetic understanding, QV_{ax} remains largely theoretical and unverifiable empirically. In contrast, QV1 includes those variants from QV_{ax} that survive rigorous empirical filtering, applying standard GWAS pre-processing criteria such as –geno, –maf, –hwe, and –mind, aimed at ensuring the quality and relevance of data by removing variants based on missing genotype data, minor allele frequency, Hardy-Weinberg equilibrium deviations, and individual missing data thresholds.

It is important to emphasise the distinction we are considering: we are dealing with unobserved or unknown variants, rather than variants of unknown significance, in the Bayesian sense. The mathematical representation of the relationship between QV_{ax} and QV1 is crucial for understanding the impact of this transition. Firstly, the intersection operation:

$$\text{TP} = QV_{\text{ax}} \cap QV1,$$

identifies true positive variants, which are both theoretically ideal and empirically robust, thus successfully passing the GWAS filtering criteria. Secondly, the set difference operation:

$$\text{FN} = QV_{\text{ax}} \setminus QV1,$$

calculates false negatives, representing the axiomatic variants that were erroneously excluded by the empirical filters, potentially omitting key genetic signals. Lastly, the quantification of unknowns:

$$\text{Unknowns} = |QV_{\text{ax}}| - |\text{TP}|,$$

provides a measure of the magnitude of theoretical variants that remain untested or unconfirmed after processing, emphasizing the potential loss of valuable genetic information.

This structured approach not only clarifies the dynamics between the axiomatic and filtered variants but also underscores the trade-offs involved in WGS pre-processing. By balancing data quality against the risk of overlooking significant genetic contributors, this analytical framework aids us in navigating the complexities of genetic data preparation and evaluation.

We explore these applications in detail elsewhere (cite Bayesian framework paper), but we briefly summarise the ideas here. In the context of GWAS, Bayesian statistics offers a potent framework for integrating theoretical and empirical knowledge. This approach leverages prior biochemical and genetic data to refine our understanding of the landscape of genetic variants, particularly those beyond current empirical capabilities.

We start by defining a prior distribution $P(\theta)$ based on established knowledge about DNA mutation rates and variant frequencies, reflecting our initial beliefs about the genetic variant distribution. Given a dataset D from GWAS pre-processing (QV1), the likelihood function $P(D|\theta)$ assesses the probability of observing the data under various genetic configurations dictated by θ.

The **posterior distribution** $P(\theta|D)$, derived from Bayes' theorem,

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)},$$

where $P(D)$ serves as a normalizing constant, updates our beliefs in light of new data. This posterior distribution integrates both the prior information and the empirical data from GWAS, providing a nuanced estimate of the distribution of genetic variants. The "unknowns" in our study, representing genetic variants not observed but theoretically possible within QV_ax, are quantified as follows:

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

where $\Theta_{\text{unobserved}}$ encompasses all parameter values corresponding to unobserved variants. This integral effectively measures the total probability of variants that are conceivable but not detected in the empirical dataset QV1.

With this short demonstration of potential future directions, we conclude our exploration of the methodological framework and practical applications.

8 Enhancing semantic interoperability

The Swiss Personalized Health Network (SPHN) promotes data sharing based on the FAIR principles, supported through the SPHN RDF Schema to enhance semantic interoperability, particularly for clinical routine data (19). The 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 (50). This extension includes critical concepts such as the sequencing instrument and QC metrics, which are necessary in ensuring the integrity and reproducibility of genomic analyses. To further integrate omics data within clinical frameworks, we have developed additional concepts, such as omic analysis results, that enable the reporting of outcomes directly tied to clinical care.

The QV framework allows for the explicit recording of QV sets used in analyses. This is particularly beneficial as it provides 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. Such a feature ensures that users can query and confirm the use of specific QV sets, like the ACMG SF, without the need to delve into the specifics of source protocols. This not only streamlines the verification process but also enhances the transparency and traceability of genomic analyses within the SPHN framework.

Therefore, to enhance reproducibility and traceability in omics research, we propose the QV Set ID (`qualifying_variant_set_id`). This identifier crucially links the variant sets used in analyses, facilitating precise and consistent replication of research methodologies. Implementing unique identifiers for qualifying variant sets is essential to ensure the reproducibility of omics analyses. These identifiers must be unique, consistent, and align with existing data management standards, integrating seamlessly into RDF schemas that incorporate standards like SNOMED CT.

As a community, we have yet to agree on the consensus sharing method. We provide several examples for generating unique identifiers:

1. **Hash functions:** SHA-256 to generate a unique hash of the set's characteristics, ensuring a unique and reliable identifier.
2. **UUIDs:** Employ randomly generated UUIDs which provide high uniqueness across systems.
3. **Semantic combination:** Create identifiers by combining relevant semantic elements like project ID and data release version in a structured format.

4. **IRI incorporation:** Develop internationalised resource identifiers (IRI) that provide traceability and integrate neatly into linked data frameworks.

5. **Registry-based allocation:** Use a centralised registry to manage identifier assignment and ensure consistency.

6. **Integrating standards:** Use standards such as SNOMED CT with local identifiers to form comprehensive, domain-specific identifiers.

We demonstrate an example analysis plan (or result database entry) in **box 10**. This lists the pipeline used, three hypothetical internal QV sets (`qv1`, `qv2`, `qv3`) and one well-known public shared QV set `acmg_sf_v3.2` where the sha256 can be confirmed. Anyone reviewing the analysis results can be sure that QV criteria `acmg_sf_v3.2` has been included in the protocol.

Box 10: Example implementation

pipeline: `pipeline DNA SNV INDEL v1`
qualifying_variant_set_id: `qv1_20250201.yaml`
qualifying_variant_set_id: `qv2_20250201.yaml`
qualifying_variant_set_id: `qv3_20250201.yaml`
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, facilitating precise and reproducible research across various projects.

9 Validation case study

We demonstrate that 100% of criteria were correctly applied using the standardised QV criteria compared with the typical manual version in our case study example. In

this process we applied an ACMG variant classification protocol (1) using a standardised QV criteria in YAML format. We used a rare disease cohort of 940 individuals, pre-processed for QC and a minimal QV test set, as previously described (lawless spss 2025). For ease of reporting, this example was restricted to chromosome 1, which contained 596 qualifying variants after strict filtering (MAF < 0.01) and limited to known disease genes based on the Genomics England panel "Primary immunodeficiency or monogenic inflammatory bowel disease," retrieved using our PanelAppRex R repository (GitHub link). We prepared this annotation interpretation dataset in R with GuRu, our variant interpretation tool, which consolidates all annotation sources and scores variants. The annotated dataset was imported from gVCF format (output by VEP) and stored as a table of 596 variant rows with 377 annotation columns. A brief selection of the annotations used used for QV is shown in **figure 6**.

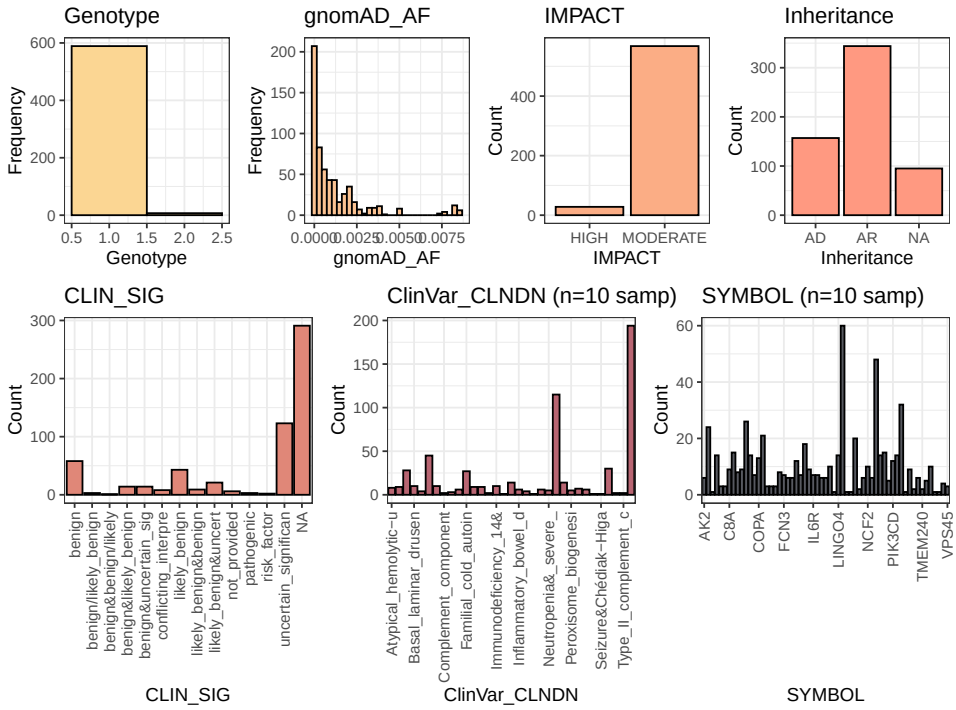

Figure 6: Example of the annotated dataset describing variant calls from the disease cohort on chromosome 1. The subset shown highlights key variables used in quality control and variant interpretation, resulting in 596 variants with 377 annotation fields per variant for 940 subjects. Axis which contain many term labels are down sampled to list every tenth label (n=10 samp).

We selected the first eight ACMG criteria for assigning pathogenicity scores to variants (1). Of these eight, six were relevant for the cohort. First, we performed this analysis manually by hard coding each criterion in the script, reflecting a typical

analysis scenario. Second, we imported the same criteria from the QC YAML file to represent our standardised approach. We captured the outputs from both scoring methods and compared them, as shown in **figure 7**. The QV criteria in YAML format are found in file `qv_files/acmg_criteria.yaml` and are in the following style in **box 11**:

Box 11: 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 two methods, demonstrating that the criteria can be imported from YAML and applied programmatically to achieve the same outcome. This is not necessarily surprising, as the burden of accuracy remains in the implementation. However, the QC YAML file acts as a shareable, standalone resource that can be adapted for different pipelines or programming languages, ensuring reproducibility for 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` containing "pathogenic," applies to the `CLIN_SIG` (clinical significance) annotation field, and uses "or" logic. We also included `ACMG_PS3`, which describes well-established functional studies supporting a damaging effect on

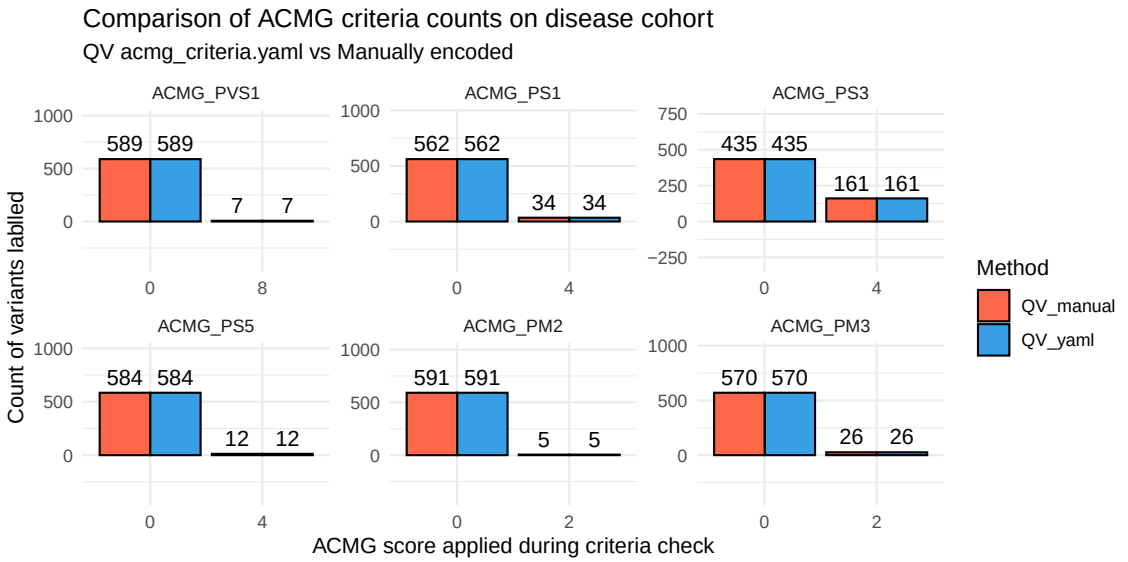

Figure 7: GuRu case study with an ACMG criteria subset, showing a 100% match between manually encoded and YAML-based methods (`qv_files/acmg_criteria.yaml`) for assigning pathogenicity scores.

the gene product, with a user-defined inheritance pattern matching the genotype, and `ACMG_PS5`, which covers compound heterozygosity with at least one high-impact variant (according to Ensembl VEP definitions). The `ACMG_PM2` criterion states that the variant is absent from controls or present at extremely low frequency in gnomAD or other population databases. For `ACMG_PM3`, the criterion checks for variants in trans with a pathogenic variant in recessive disorders, showing some overlap with PS5 because our rare disease cohort filtering already treats "IMPACT = HIGH" as pathogenic.

We skipped the PS2 criterion, which requires confirmed de novo status in a patient with the disease and no family history, because no parental data were available. We also skipped PS4, which measures a significantly increased prevalence in affected individuals compared with controls, because that was evaluated in a separate case-control analysis for this cohort.

Subsequently, we briefly demonstrate why such QV criteria are necessary. In **figure** 8, we show the final annotation results for the test disease cohort by listing the number of criteria applied for both samples (cases) and variants. From this, the top candidate causal pathogenic variants can be automatically retrieved using ACMG scoring methods (1; 5).

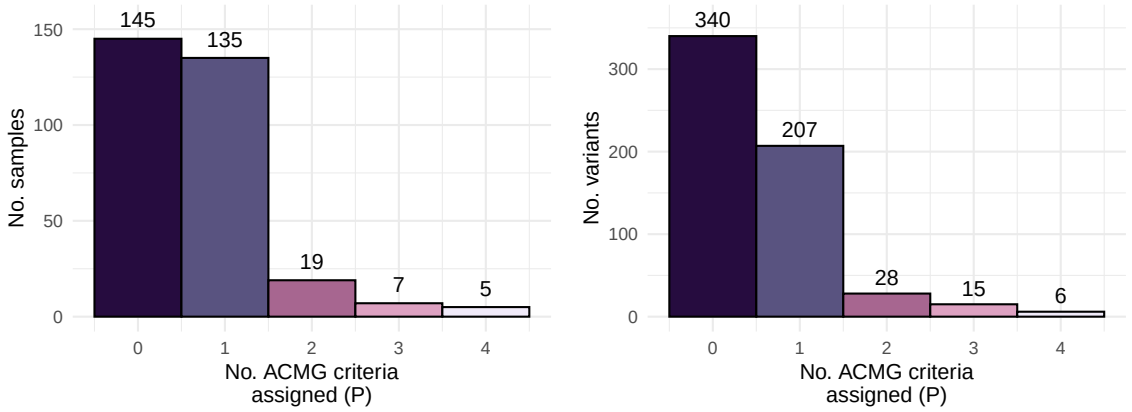

Figure 8: Overview of the final annotation interpretation for the test disease cohort, illustrating the number of criteria applied both for subject samples (left) and for variants (right). This facilitates subsequent retrieval of the top candidate pathogenic variants automatically.

10 Conclusions

We emphasise the critical importance of QV standardisation in genomic research. By proposing a clear framework for incorporating QV protocols into analysis pipelines, we highlight how systematic handling of these variables enhances reproducibility, accuracy, and efficiency in genetic studies. As genomic technologies and data complexities expand, the need for robust, scalable, and adaptable QV protocols becomes ever more pressing. Future work should extend these frameworks to accommodate emerging technologies and analytical challenges, further improving the fidelity and utility of genomic data interpretation across diverse applications.

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] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

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

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

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

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.