

Conceptualising qualifying variants for genomic analysis

Dylan Lawless^{*1}, Consortium Members¹, and Luregn J. Schlapbach^{†1}

¹Department of Intensive Care and Neonatology and Children's Research Centre, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland.

²Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Switzerland.

February 15, 2025

Contents

1	List of Acronyms	2
2	Introduction	4
3	Background	8
3.1	The problem and proposed solution	8
4	Advanced applications and case studies	9
4.1	Example application of qualifying variants in WGS analysis	9
5	Methodological framework	10
6	Standardisation of Qualifying variant (QV) advances theoretical domains	13
6.1	Applications in multiblock data fusion	14
6.2	Protocol development and standardisation needs	15
7	Challenges and innovations in data integration	15
7.1	Future directions and implications	16
7.2	Notation typical to GWAS, VSAT, and other statistical applications	16

^{*}Addresses for correspondence: Dylan.Lawless@uzh.ch

[†]Addresses for correspondence: email@epfl.ch

1 List of Acronyms

QV Qualifying variant	1
VSAT Variant Set Association Test	5
GWAS Genome Wide Association Test	4
PRS Polygenic Risk Score	4
WGS Whole Genome Sequencing	8

Abstract

QVs represent specific genomic alterations selected through defined criteria throughout processing pipelines, essential for downstream analyses in genetic research and clinical diagnostics. Here we explore QVs not just as simple filtering criteria but as a dynamic, multifaceted concept crucial across various genomic analysis scenarios. 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. Moreover, QVs, typically seen as a set of filters and algorithms to exclude benign or unrelated variants, more often encompass complex steps distributed throughout the analysis pipeline. We redefine QVs by illustrating several common sets and their roles within analysis pipelines, demonstrating their theoretical pipelining and standardisation for specific analytical scenarios. By introducing a new vocabulary and a standard reference model, we aim to improve understanding and communication around QVs, enhancing methodological discussions across disciplines.

2 Introduction

QVs are genomic alterations selected through specific criteria after the primary stages of routine 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 are based on well-established best practices in variant classification and reporting standards (1–4), established work-flows (5–7). Polygenic Risk Score (PRS) reporting standards to have been developed to encourage their application and translation as well as open cataloguing for reproducibility and systematic evaluation (8; 9). However, a standard guide for QV themselves remain missing.

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. Povysil et al. (10) previously provided a tangible example of QV for variant collapsing analyses for complex traits. Cirulli et al. (11) reported one of the first variant collapse analysis and introduced the QV concept. However, a standardised framework for presenting QV themselves remains missing. We detail three typical applications of QV sets:

1. **QV passing quality control (QC) only:** Generates large datasets, typically over 500,000 variants per subject, used primarily in GWAS.
2. **QV for rare disease:** Produces smaller datasets after stringent filtering, around 10,000 variants per subject, useful in single-case genetic reports.
3. **Flexible QV:** Balances between quality control and false positives, yielding datasets of fewer than 100,000 variants per subject for rare variant association testing.

Two critical 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 quality control procedures to ensure their statistical suitability for the main analysis. Rigorous QV selection and categorisation in genetic research and diagnostics to accurately report and reproduce such

studies, particularly since the the QV criteria may be more important than the choice of analysis pipeline.

Figure 1 demonstrates a typical WGS and Variant Set Association Test (VSAT) analysis pipeline, showing QVs as sequential and potentially piped protocol steps. 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 QC filtering. After variant annotation, further QV steps can be applied on this new information as shown in **figure 4**. **Figure 5** 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.

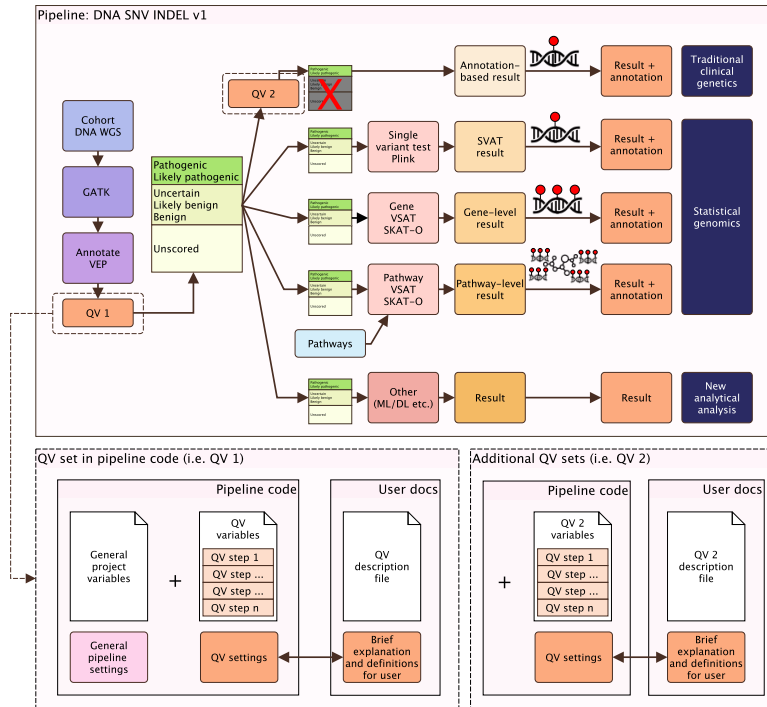


Figure 1: Summary of the example application design DNA SNV INDEL v1 pipeline. QV1 and QV2 are shown as sequential and potentially piped protocol steps. QV files used in a pipeline. The description file contains brief summary of how or why a step is used in the QV set (not mandatory). The variables file 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.

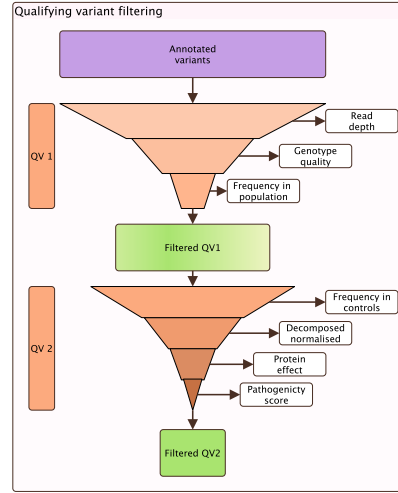


Figure 2: Illustration of the qualifying variant workflow. This figure summarises the conceptualised variant filtering step. This style is relatively common. In reality, we observe that each layer of the filters comes from disparate stages of a pipeline.

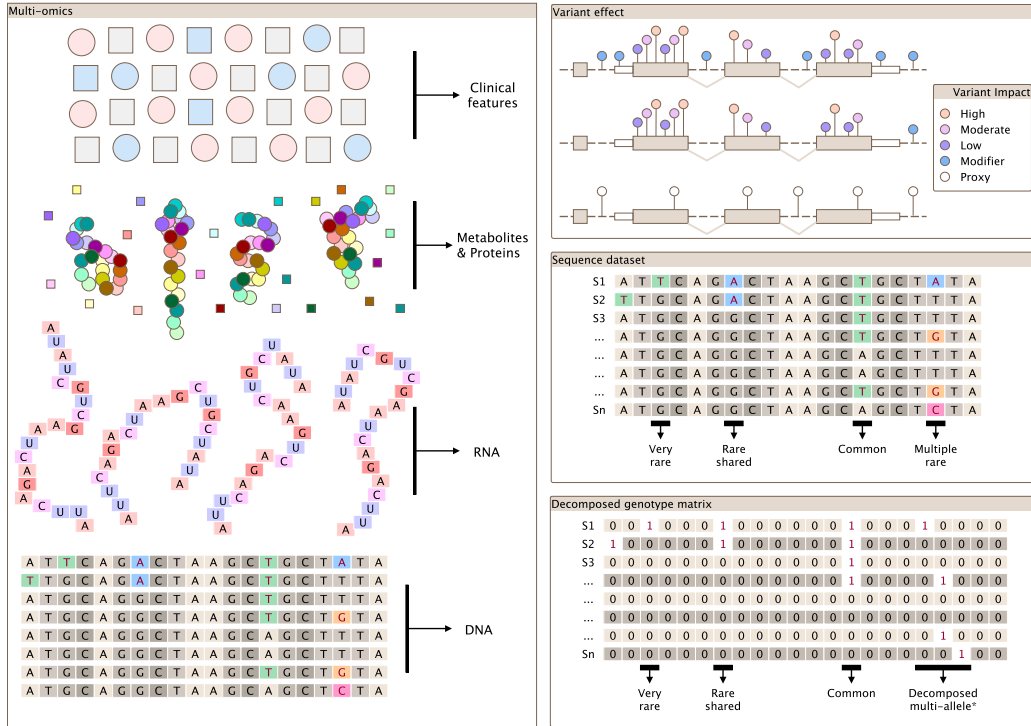






Figure 3: From raw omics to data matrix. We focus on DNA variants in QV but the same concept applies to other datasets.

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


 Set level


 Sample level


 Variant level


 Gene level



 Ontology level

Figure 4: The initial variant detection pipeline generally 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.

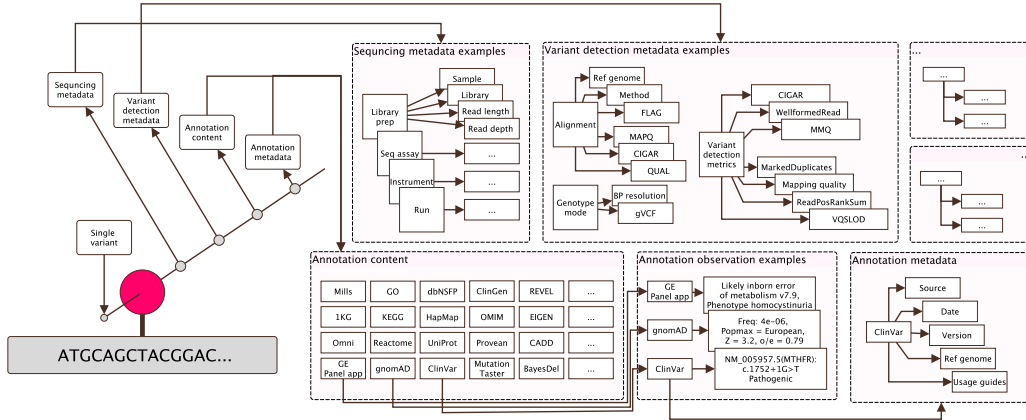


Figure 5: This illustration shows the structural framework of an annotated variant in relation to the features used for qualification. For every individual variant, a number features are capable of triggering QV protocols. The diagram highlights a select group of these features, showcasing both pre-existing metadata established before data generation and annotations applied after variant calling.

3 Background

3.1 The problem and proposed solution

Version1: Study sizes are beginning to reach above 1,000,000 subjects (12; 13). The transitions to WGS instead of genotyping by default means that rare variants can now be used in GWAS and VSAT, allowing for deeper analysis of complex traits (14; 15). QV are a logical necessity of data cleaning and preparation. Labelling a group of procedures under a single umbrella of QV is useful for simplicity. In reality the steps of QV can be separated across a pipeline and result from a mixture of different steps or sources. In addition, complex analysis require multiple different streams of processing that converge into a joint analysis. This multifaceted concept of QV naturally appears in these multicomponent analysis since two or more sets will be required.

Version2: As study sizes surpass the 1,000,000 subjects milestone (12; 13), the shift towards Whole Genome Sequencing (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 (14; 15). 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.

Moreover, complex analyses often necessitate multiple processing streams that merge into a cohesive analysis. This multifaceted approach to QV becomes apparent in multicomponent 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 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 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.

4 Advanced applications and case studies

In-depth look at specific scenarios where QVs have been crucial, such as in GWAS and clinical genetics. Examples of successful application of QVs in large-scale studies and rare disease research.

Explore the implications of sophisticated risk models that integrate clinical and genomic data, enhancing predictive accuracy in large, well-defined cohorts. (16–18).

Discuss the unique opportunities and challenges in rare disease research, especially in isolated or specific populations - how we mentioned complex signals but well defined cohort can help in rare diseases (19).

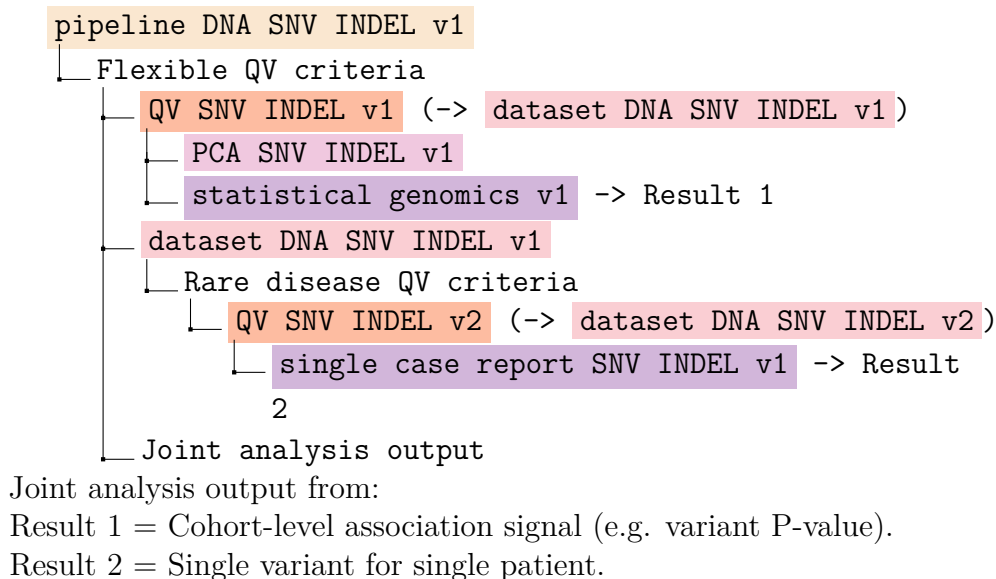
4.1 Example application of qualifying variants in WGS analysis

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) newly identified cohort-level genes associated with disease with (2) single case-level known disease-causing 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 is 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



5 Methodological framework

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 variables responsible which are sourced as part of a pipeline, as shown in **example box 3**.

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**.

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.

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 allele frequency (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.

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

```

6 Standardisation of QV advances theoretical domains

Detailed exploration of the need for and benefits of standardising QVs. Description of common sets of QVs and their roles within analysis pipelines. Discussion on the integration of sophisticated ML/AI models to handle multi-omic datasets.

Discuss the integration of sophisticated ML/AI models to handle diverse and large datasets in the context of genetic studies transitioning to WGS - how complex signals can exist within single datasets (20; 21).

6.1 Applications in multiblock data 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. 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.

6.2 Protocol development and standardisation needs

This approach requires a clear protocol for merging data across different omic layers, ensuring that each contributes meaningfully to the unified model without conflating their distinct signals. As we develop new theories and methods in this space, the precision in defining and reporting QVs becomes crucial, particularly when dealing with non-public data and complex codebases. 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 should be reported explicitly as a protocol with a detailed list of definitions and variables, building on our demonstrated examples for one such set, QV1.

7 Challenges and innovations in data integration

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 - so too must we approach the integration of complex QV layers with rigor (Bland JM, Altman DG. (1994) Correlation, regression and repeated data. 308, 896. <http://www.bmj.com/cgi/content/full/308/6933/896>). 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. Altman and Bland’s warning 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 (Simpson, E. 1951, Wright, 1920, 1934, Pearl, 2016). (Simpson, E. (1951)), The Interpretation of Interaction in Contingency Tables, Journal of the Royal Statistical Society, Series B, 13, 238–241. [406] <https://doi.org/10.1111/j.2517-6161.1951.tb00088.x>) (Wright, Sewall. "The method of path coefficients." The annals of mathematical statistics 5.3 (1934): 161-215. <https://www.jstor.org/stable/2957502>), and Pearl (2016) (Pearl, J., Glymour, M., Jewell, N. P. (2016). Causal inference in statistics: A primer. Wiley.)

Address how deep phenotyping and precision medicine with omic data are reshaping data integration strategies - standardised database formats are critical for genomics and QV should not be an afterthought (22–24).

7.1 Future directions and implications

Discussion on the necessity of sophisticated data integration strategies. Predictions for the future of omics research with the standardized use of refined QVs.

Consider the impact of new publishing formats like Registered Reports on the field of genomics, promoting transparency and reproducibility. (25)

Moreover, this approach is 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

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.

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

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

where: - β_0 is the intercept, - 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} is the effect of the genotype on the phenotype, with $G_{\text{QV}_{i,v}}$ denoting the genotype of the i -th individual for the v -th variant in the QV set, coded as 0, 1, or 2 (representing the number of minor alleles).

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. 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.

Conceptual Framework and Statistical Representation

In genome-wide association studies (GWAS), the transition from theoretical axiomatic variants (QV_ax) to empirically testable variants (QV1) 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.

The mathematical representation of the relationship between QV_ax and $QV1$ is crucial for understanding the impact of this transition. Firstly, the intersection operation:

$$TP = QV_ax \cap QV1,$$

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

$$FN = QV_ax \setminus QV1,$$

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:

$$\text{Unknowns} = |QV_ax| - |TP|,$$

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 GWAS pre-processing. By balancing data quality against the risk of overlooking significant genetic contributors, this analytical framework aids researchers in navigating the complexities of genetic data preparation and evaluation.

We discuss this in detail elsewhere (cite Bayesian framework paper). The ideas are briefly summarised: In the context of genome-wide association studies (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.

8 Conclusions

Summary of the main findings and the importance of QV standardisation. Call to action for the adoption of new methodologies and the continued evolution of QV standards.

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 (acmg) and the clinical genome resource (clingen). *Genetics in Medicine*, 22(2):245–257, 2020.
- [5] 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.
 - [6] 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>.
 - [7] 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>.
 - [8] 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.
 - [9] 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.
 - [10] 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>.
 - [11] 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.
 - [12] 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.
- [13] 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.
 - [14] 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.
 - [15] Alexander I Young. Solving the missing heritability problem. *PLoS genetics*, 15(6):e1008222, 2019.
 - [16] 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.
 - [17] 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.
 - [18] 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.
 - [19] Elaine T Lim, Peter Würtz, Aki S Havulinna, Priit Palta, Taru Tukiainen, Karola Rehnström, Tõnu Esko, Reedik Mägi, Michael Inouye, Tuuli Lapalainen, et al. Distribution and medical impact of loss-of-function variants in the finnish founder population. *PLoS genetics*, 10(7):e1004494, 2014.
 - [20] Augustine Kong, Gudmar Thorleifsson, Michael L Frigge, Bjarni J Vilhjalmsdottir, 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.

- [21] 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.
- [22] 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.
- [23] The All of Us Research Program Genomics Investigators. Genomic data in the all of us research program. *Nature*, 627(8003):340–346, 2024.
- [24] 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.
- [25] 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.