Conceptualising Qualifying Variants for Genomic Analysis

Dylan Lawless*[1], Consortium Members[1], and Luregn J. Schlapbach†[1]

[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.

January 17, 2025

Contents

*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 . . . 4

GWAS Genome Wide Association Test . . . 3

PRS Polygenic Risk Score . . . 3

WGS Whole Genome Sequencing . . . 5

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) provide a tangible example of QV for variant collapsing analyses for complex traits. 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 **Figure 3** 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.

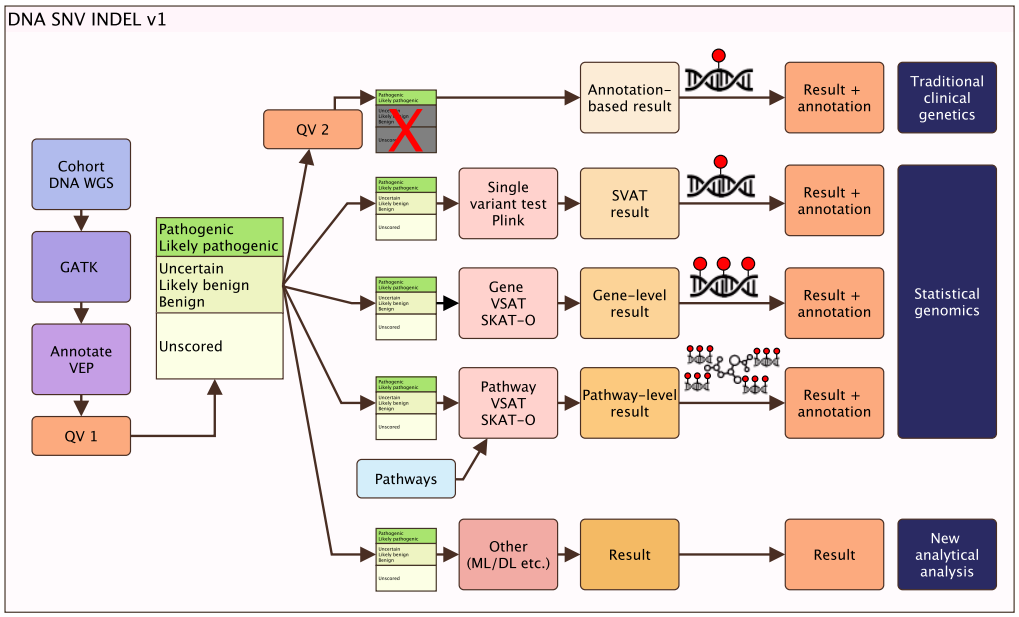

Figure 1: Summary of the example application design DNA SNV INDEL v1 pipeline. QV1 and QV2 are shown as sequential and potentially piped protocol steps.

3 Background

3.1 The problem and proposed solution

Study sizes are beginning to reach above 1,000,000 subjects (11; 12). 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 (13; 14). 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

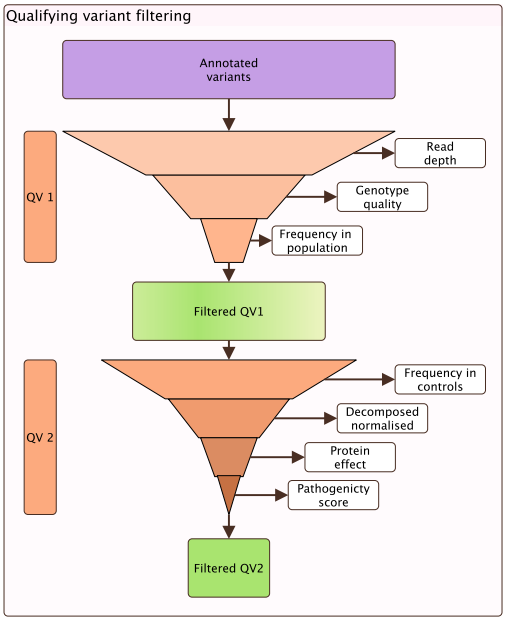

Figure 2: Illustration of the qualifying variant workflow. This figure summarises the conceptualised variant filtering step. This style is relatively common. In reality, we observe that each layer of the filters comes from disparate stages of a pipeline.

analysis. This multifaceted concept of QV naturally appears in these multicomponent analysis since two or more sets will be required.

As study sizes surpass the 1,000,000 subjects milestone (11; 12), 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 (13; 14). 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

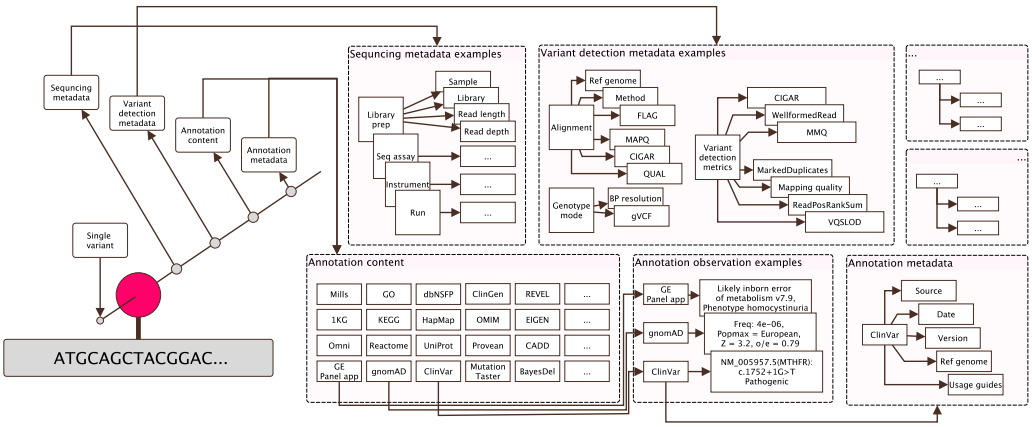

Figure 3: This illustration shows the structural framework of an annotated variant in relation to the features used for qualification. For every individual variant, a number features are capable of triggering QV protocols. The diagram highlights a select group of these features, showcasing both pre-existing metadata established before data generation and annotations applied after variant calling.

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. (15–17).

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 (18).

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 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`."

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.

5 Methodological innovations and framework

Introduction of new statistical methods and frameworks needed for the effective use of QVs. Detailed protocol descriptions and variable examples.

5.1 Qualifying variant protocol

We use two levels to handle QV protocols.

1. **Description**: The description of each step as part of an overall QV set. An example is shown in**section 5.2**.

2. **Variables** The variables responsible which are sourced as part of a pipeline. An example is shown in**section 5.3**.

5.2 Qualifying variant protocol description example

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.

1. [**QC**] `01_fastp` The tool fastp is used for QC. FASTQ that fail are investigated or removed. See fastp for more.

2. [**QC**] `03b_collectwgsmetrics` BAMs that fail are investigated or removed. See metrics_CollectWgsMetrics for more.

3. [**QC**] `05_rmdup_merge` is used to mark optical duplicates. See GATK Duplicates for more.

4. [**QC**] `07_haplotype_caller` used `-ERC GVCF` mode. This does not remove variants but unlike `BP_RESOLUTION`, `GVCF` mode condenses non-variant blocks which could be misunderstood later as missing if not recognised by the user, for example in a genotype matrix which has been merged with other cohorts. See our VCF and gVCF documentation for more.

5. [**QC**] `07c_qc_summary_stats` is used to log QC. This implements `bcftools stats` and subsequently the bcftools `plot-vcfstats` using `python -m venv envQCplot` with matplotlib. Subjects fail are investigated or removed. See `metrics_bcftoolsstats` documentation for more.

6. [**QC**] `10_vqsr` employs Variant Quality Score Recalibration (VQSR) using GATK. The method 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. Resources like HapMap and Omni SNP chip array data train the recalibration model, which assigns a VQSLOD score to each variant indicating the likelihood of its authenticity. This step refines variant filtering to enhance accuracy in genomic research.

7. [**QC**] see `10b_qc_summary_stats` for plink logs. If any value fails to meet the threshold, it is either removed or investigated.

8. [**QC**] **Optional** CollectVariantCallingMetrics. An example is shown in our documents on "How to filter variants with VQSR" and "CollectVariantCallingMetrics-Picard".

9. [**QC**] **Optional** Other Picard metrics which are not in used by default. See "Picard metric definitions" for more.

10. [**QC**] `11_genotype_refinement` This step uses the Genotype Refinement workflow to enhance the precision of genotype calls. The process includes: (1) `CalculateGenotypePosteriors`: Refines genotype probabilities using family data and/or population allele frequencies, primarily from gnomAD, to improve initial likelihoods from variant callers. This method is particularly effective in trio studies, reducing false positives and enhancing genotype accuracy. (2) `VariantFiltration`: Applies filters on genotype quality scores (e.g. `GQ < 20`) to flag lower confidence genotypes. It refines the quality of variant calls by annotating individual genotypes based on specified criteria, thus isolating high-confidence calls.

11. [**QC**] The same stats as `10b_qc_summary_stats` for plink logs are run in step `11_genotype_refinement`.

12. [**Filter**] `12_pre_annotation_processing`. This step involves filtering variants using `bcftools filter` and further processing with GATK `SelectVariants`. (1) `QUAL`$\geq$ `30`: Ensures high confidence score in the variant call. (2) `INFO/DP`$\geq$`20`: Required total depth of quality base calls, supporting the variant's presence. (3) `FORMAT/DP`$\geq$`10`: Ensures a minimum depth per sample reads per genotype, confirming sufficient data support. (4) `FORMAT/GQ`$\geq$`20`: Ensures the genotype quality for each sample, reflecting confidence in genotype assignment. (5) Then GATK `SelectVariants` is subsequently applied with `--exclude-filtered`, `--exclude-non-varia` and `--remove-unused-alternates`.

13. [**QC**] `12_pre_annotation_processing` also incorporates a second independent variant processing technique using the vt tool to enhance data quality for downstream analysis. This includes `vt decompose` which splits multiallelic variants into separate observations, simplifying complexity and reducing potential errors in subsequent analyses, and `vt normalization` which adjusts variant representations to conform to a consensus format, ensuring that each variant is represented parsimoniously as described in Tan et. al (2015). This process makes the representation of variants as concise as possible without reducing any alleles to length zero.

14. [**Filter**] [**QC**] `13_pre_annotation_MAF` runs data preparation steps including a filter with `vcftools --max-maf` using the source variable (e.g.`MAF_value="0.4"`). This step is simply for reducing the data size to remove highly common variants. We do not expect more than 40% of the cohort to share a relevant variant.

5.3 Qualifying variant protocol variables example

We select the step [**QC**] `10_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. Table 1 shows the details about the sourced variables used during VQSR.

!Show the variable.sh snippet here.

Make a yed map to show the pipeline again, this time adding the explicit example of the VQSR variable settings from the variables.sh file. This show graphic summary icon for other QV steps throughout the pipeline. Compare it to figure 1 where the summary of the pipeline simply illustrate QV v1 as a single step but in this figure we see that it is actually spread throughout the pipeline by necessity.

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 (19; 20).

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

Table 1: Description of the VQSR variable settings used as part of the step [**QC**] `10_vqsr` in the example QV set `QV SNV INDEL v1`.

VQSR Settings	**Explanation**
SNP Mode	
- **HapMap:** known=false, training=true, truth=true, prior=15.0	Used as a high-confidence reference set for training the recalibration model.
Omni: known=false, training=true, truth=false, prior=12.0	Provides additional training data derived from Omni genotyping arrays.
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.
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.
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.
INDEL Mode	
Mills: known=false, training=true, truth=true, prior=12.0	Utilizes the Mills and 1000G gold standard indel dataset for high-accuracy recalibration of indels.
dbSNP: known=true, training=false, truth=false, prior=2.0	Includes known indel sites from the dbSNP database to enhance the detection and filtering process.
Annotations: QD, MQRankSum, ReadPosRankSum, FS, SOR	Same as for SNPs, these annotations are critical for assessing the likelihood of indels being true genetic variations rather than errors.
Truth Sensitivity Filter Level: 95	This setting defines the percentage of true indels to retain, aiming to capture 95% of true indels at the specified VQSLOD threshold.

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. T he 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 (21–23).

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. (24)

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.

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 (acmge 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] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.