Bayesian Aggregation of Multiple Annotations Enhances Rare Variant Association Testing

Antonio Nappi[1,2], Na Cai[1,2,3], and Francesco Paolo Casale[1,2,4]

[1] Institute of AI for Health, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
[2] Helmholtz Pioneer Campus, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
[3] TUM School of Medicine and Health, Technical University of Munich and Klinikum Rechts der Isar, Munich, Germany
[4] School of Computation, Information and Technology, Technical University of Munich, Garching, Germany

Abstract. Gene-based rare variant association tests (RVATs) are essential for uncovering disease mechanisms and identifying potential drug targets. However, existing RVAT methods often rely on rigid models or analyze single annotations in isolation, lacking flexible frameworks that integrate multiple variant annotations. To address this, we introduce BayesRVAT, a Bayesian framework for RVAT that models variant effects as a function of multiple annotations. BayesRVAT accommodates diverse genetic architectures by enabling the specification of priors on variant effects and estimates gene-trait-specific burden scores through variational inference, yielding well-calibrated P values via a novel approximate likelihood ratio test. We benchmark the framework as a burden test in simulations and in real data applications from the UK Biobank, demonstrating that BayesRVAT outperforms other state-of-the-art burden test strategies. Our results reveal novel biologically meaningful associations, underscoring BayesRVAT's potential for accelerating genetic discoveries.

1 Introduction

Gene-based rare variant association studies (RVATs) focus on rare variants likely to impact protein function, providing a more interpretable path to understanding disease biology compared to common variant association studies [8, 31]. Traditionally, RVATs have been performed using burden tests, which aggregate likely pathogenic variants based on functional annotations into a gene burden score and then regress these scores against trait values across individuals in a formal gene-level association test [2, 18, 19, 29, 25, 5]. Recent burden test models aim to capture the concept of an *allelic series* [31, 9], where progressively larger functional impacts of rare variants within a gene lead to correspondingly stronger phenotypic effects [32]. However, these approaches are often restricted to a limited set of consequence annotations or impose fixed aggregation schemes across multiple genes and traits, potentially missing gene- and trait-specific nuances.

To address these limitations, we present BayesRVAT, a Bayesian framework for RVAT that enables, for the first time, the flexible aggregation of multiple variant annotations without enforcing a universally fixed scheme across genes and traits. BayesRVAT introduces priors on aggregation parameters and infers a posterior distribution for each analyzed gene and trait. Additionally, by modeling uncertainty in the aggregation parameters, BayesRVAT produces well-calibrated association P values using a novel approximate likelihood ratio test. Through simulations and analysis of twelve blood traits from the UK Biobank (UKB), we demonstrate that BayesRVAT outperforms other burden test strategies, showcasing the value of our framework in uncovering novel genetic associations.

2 Related work

Burden tests. While our Bayesian RVAT framework is general, we here specialize it to perform burden tests. Burden tests aggregate the effects of rare variants into a gene burden score to improve statistical power. This score is then regressed against a trait of interest in a formal test [8, 25, 24, 35, 40, 36]. While traditional burden tests have been limited to few annotations such as minor allele frequency (MAF) and variant consequences [14, 25], BayesRVAT can integrate multiple annotations leveraging a Bayesian framework.

Allelic series. BayesRVAT can model gene-level effects as a function of multiple rare variants and their functional annotations, effectively capturing allelic series. The term "allelic series" refers to a collection of variants within a gene that exhibit a gradation of phenotypic effects based on their severity, suggesting a dose-response relationship between gene functionality and the resulting phenotype [39, 37]. As allelic series enable the assessment of the feasibility of pharmacological modulation [37, 10, 31], methods that can accurately capture these relationships are of significant interest. For example, COAST models allelic series by weighting variants based on the expected deleteriousness of few functional consequences [31]. Conversely, DeepRVAT uses a data-driven approach to

learn a trait-gene-agnostic aggregation function from multiple annotations using neural networks [9]. In contrast to these methods, BayesRVAT can handle larger sets of variant annotations without enforcing a universally fixed scheme across genes and traits.

Variance component models. In contrast to burden tests, which assume a uniform effect direction across all variants, variance component approaches allow for both deleterious and protective effects by employing random effect models. The most widely used variance component test for rare variants is SKAT [48]. Given the complementary strengths of burden and variance component tests [3], omnibus tests that combine both, such as SKAT-O [23], have become increasingly popular[34, 26, 53]. In this work, we demonstrate that BayesRVAT integrates smoothly within omnibus test procedures, maintaining its power advantages over other integrated burden tests.

Bayesian inference. BayesRVAT performs Bayesian inference on parameters modeling variant effects as a function of multiple annotations. Given the intractability of exact posterior computation, we use black-box variational inference [41], which reformulates the inference problem as an optimization task, directly optimizing a variational distribution to approximate the true posterior using gradient-based methods [4, 43, 21, 11, 45]. While Bayesian methods for RVAT have been previously explored [50, 46, 28, 51], BayesRVAT is the first unified Bayesian framework that can incorporate multiple genetic architectures and variant annotations.

3 Method

3.1 The RVAT framework

Gene-level RVATs aggregate selected rare variants within a gene to assess their collective association with a trait of interest. Formally, given trait values $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$ for N individuals, genotype matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T \in \mathbb{R}^{N\times S}$ for S rare variants in the gene under investigation, annotation matrix $\boldsymbol{A} \in \mathbb{R}^{S\times L}$ comprising L annotations for these S variants, and covariate matrix $\boldsymbol{F} \in \mathbb{R}^{N\times K}$ for K factors (e.g., age and sex), the model for gene-level RVAT can be written as

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{F}\boldsymbol{\alpha} + g(\boldsymbol{X}, \boldsymbol{A})\beta, \sigma_n^2 \boldsymbol{I}\right). \tag{1}$$

Here $\boldsymbol{\alpha} \in \mathbb{R}^{K\times 1}$ are the covariate effects, $g(\boldsymbol{X}, \boldsymbol{A}) = [g(\boldsymbol{x}_1, \boldsymbol{A}), \ldots, g(\boldsymbol{x}_N, \boldsymbol{A})]^T \in \mathbb{R}^{N\times 1}$ is the vector of gene burden scores obtained by aggregating variants based on their annotations through the aggregation function g, β is the effect size of the gene burden score, and σ_n^2 is the residual variance. Within this framework, the association between the gene burden score and the phenotype can be assessed by testing whether $\beta \neq 0$. Several widely-used RVAT models can be expressed in this form, each making different assumptions about the aggregation function g: burden tests use sum- or max-pooling on selected variants[8, 25, 24, 35, 40, 36],

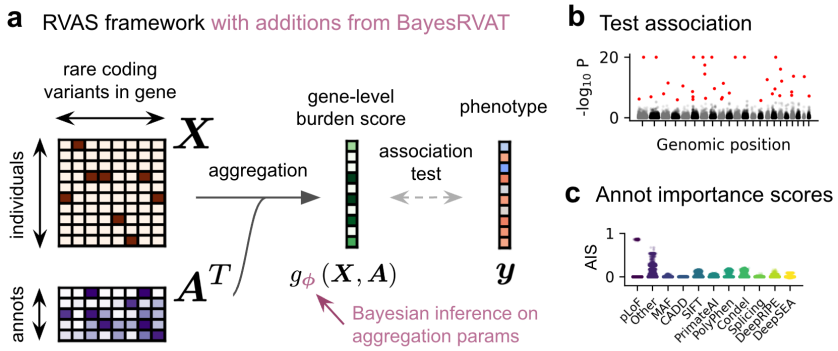

Fig. 1: **Overview of the BayesRVAT framework.** (**a**) In rare variant association studies (RVAT), rare variants $\boldsymbol{X}$ and their annotations $\boldsymbol{A}$ are aggregated into a gene burden score, which is tested for association with the phenotype $\boldsymbol{y}$. BayesRVAT explicitly introduces aggregation function $g_{\boldsymbol{\phi}}(\boldsymbol{X}, \boldsymbol{A})$ and a prior over aggregation parameters $\boldsymbol{\phi}$. (**b**) BayesRVAT enables scalable gene burden testing accounting for multiple annotations. (**c**) It also provides annotation importance scores (AIS) for each analyzed gene-trait pair.

variance component models consider additive random variant effects [48], while DeepRVAT employs a pretrained neural network function [9]. In contrast, in BayesRVAT, g is a flexible function of multiple variant annotations with Bayesian priors on its parameters.

3.2 A Bayesian framework for RVAT

In BayesRVAT, we parameterize the aggregation function $g_{\boldsymbol{\phi}}$ with parameters $\boldsymbol{\phi}$ and introduce a prior distribution $p(\boldsymbol{\phi})$ that incorporates our prior beliefs on how to aggregate variants into a burden score based on their annotations (**Figure 1a**). Introducing compact notations for input data $\mathcal{D} = \{\boldsymbol{F}, \boldsymbol{X}, \boldsymbol{A}\}$ and model parameters $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \beta, \sigma_n^2\}$, the model marginal likelihood can be written as

$$p(\boldsymbol{y} \mid \mathcal{D}, \boldsymbol{\theta}) = \int p(\boldsymbol{y} \mid \mathcal{D}, \boldsymbol{\theta}, \boldsymbol{\phi})\, p(\boldsymbol{\phi})\, \mathrm{d}\boldsymbol{\phi} \tag{2}$$

$$= \int \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{F}\boldsymbol{\alpha} + g_{\boldsymbol{\phi}}(\boldsymbol{X}, \boldsymbol{A})\beta, \sigma_n^2 \boldsymbol{I}\right) p(\boldsymbol{\phi})\, \mathrm{d}\boldsymbol{\phi}. \tag{3}$$

We note that our Bayesian framework allows the data for each gene and trait to update these prior beliefs, effectively adapting the posterior on aggregation parameters $\boldsymbol{\phi}$ to the specific gene/trait pair being analyzed.

Optimization. The optimization of model parameters $\boldsymbol{\theta}$ by maximum likelihood is intractable for a general aggregation function $g_{\boldsymbol{\phi}}(\boldsymbol{X}, \boldsymbol{A})$ due to the integral over $\boldsymbol{\phi}$. To address this, we use black-box variational inference [41], which approximates the true posterior $p(\boldsymbol{\phi} \mid \boldsymbol{y}, \mathcal{D}, \boldsymbol{\theta})$ with a simpler variational distribution

$q_{\boldsymbol{\psi}}(\boldsymbol{\phi})$ parameterized by $\boldsymbol{\psi}$. Within this framework, we optimize both the model parameters $\boldsymbol{\theta}$ and the variational parameters $\boldsymbol{\psi}$ by maximizing the Evidence Lower Bound (ELBO):

$$\text{ELBO}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \mathbb{E}_{q_{\boldsymbol{\psi}}(\boldsymbol{\phi})} \left[\log p(\boldsymbol{y} \mid \mathcal{D}, \boldsymbol{\theta}, \boldsymbol{\phi}) - \log \frac{q_{\boldsymbol{\psi}}(\boldsymbol{\phi})}{p(\boldsymbol{\phi})} \right], \tag{4}$$

which provides a lower bound on the log marginal likelihood of the data. We assume a mean-field Gaussian variational posterior for $q_{\boldsymbol{\psi}}(\boldsymbol{\phi})$ and approximate the expectation using Monte Carlo sampling. To optimize the ELBO, we use gradient descent with the reparameterization trick to propagate gradients through the Monte Carlo estimator [43, 21]. By maximizing the ELBO, we jointly estimate the values of $\hat{\boldsymbol{\theta}}$ and the variational parameters $\hat{\boldsymbol{\psi}}$, providing approximations of the maximum likelihood estimator of $\boldsymbol{\theta}$ and the exact posterior distribution, respectively.

Association testing. Within the BayesRVAT framework in Eq (3), we can assess associations between gene burden scores and trait values by testing the hypothesis $\beta \neq 0$ (**Figure 1b**). As the likelihood ratio test statistic is intractable due to the integral over $\boldsymbol{\phi}$ in the alternative hypothesis, we introduce an approximate Likelihood Ratio Test statistic. Briefly, we replace the intractable log marginal likelihood under the alternative hypothesis with the importance-weighted variational evidence lower bound (IW-ELBO) [6]:

$$\text{IW-ELBO}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}) = \mathbb{E}_{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K \sim q_{\hat{\boldsymbol{\psi}}}(\boldsymbol{\phi})} \left[\log \left(\frac{1}{K} \sum_{k=1}^{K} \frac{p(\boldsymbol{y} \mid \mathcal{D}, \hat{\boldsymbol{\theta}}, \boldsymbol{\phi}_k)\, p(\boldsymbol{\phi}_k)}{q_{\hat{\boldsymbol{\psi}}}(\boldsymbol{\phi}_k)} \right) \right], \tag{5}$$

which is computed using the approximate maximum likelihood estimators $\hat{\boldsymbol{\theta}}$ and the variational posterior $q_{\hat{\boldsymbol{\psi}}}(\boldsymbol{\phi})$, obtained by optimizing the ELBO. The IW-ELBO is a tighter bound on the log marginal likelihood compared to the standard ELBO, leading to more accurate P values. We note that this approximation yields conservative P values as it replaces the log marginal likelihood under the alternative hypothesis with its lower bound, while the log marginal likelihood under the null hypothesis can be computed exactly. Consequently, this leads to lower test statistics than an exact likelihood ratio test, and thus, conservative P values.

Annotation Importance Scores. Similar to sensitivity analysis [15], we can evaluate the importance of a set of annotations by comparing the gene burden scores computed using all annotations (denoted as $\boldsymbol{A}_1$) with the scores obtained by setting a subset of annotations to their median values (denoted as $\boldsymbol{A}_0$). Specifically, we compute the expected value of the difference between these two burden scores:

$$\boldsymbol{s} = \mathbb{E}_{\boldsymbol{\phi} \sim q_{\hat{\boldsymbol{\psi}}}(\boldsymbol{\phi})} [g_{\boldsymbol{\phi}}(\boldsymbol{X}, \boldsymbol{A}_1) - g_{\boldsymbol{\phi}}(\boldsymbol{X}, \boldsymbol{A}_0)]. \tag{6}$$

The result $\boldsymbol{s} \in \mathbb{R}^{N \times 1}$ quantifies how much each individual's gene burden is influenced by the annotations under investigation. We refer to these scores as Annotation Importance Scores (AIS, **Figure 1c**).

3.3 Specialization to Bayesian burden test

While the introduced framework is general and can be applied to a variety of RVAT models by choosing different forms of g, we here specialize it to the gene burden test, where we aggregate variant effects based on their functional annotations rather than individual variant effects.

Choice of aggregation function. After preprocessing annotations $\boldsymbol{A}$ to ensure that higher values correspond to more deleterious effects (**Supplementary Information**), we assume linear variant effects in $\boldsymbol{A}$ and use an additive model with saturation to collapse the contributions of multiple variants into a single gene burden score:

$$g_{\boldsymbol{\phi}}(\boldsymbol{X}, \boldsymbol{A}) = \text{sigmoid}(\boldsymbol{X}(\boldsymbol{A}\boldsymbol{\phi}) - b_0). \tag{7}$$

Here, b_0 is a bias term that ensures individuals carrying no rare variants receive a burden score close to zero. The sigmoid function introduces a saturation mechanism, reflecting the biological intuition that once gene function is sufficiently impaired, additional variant effects do not further increase the burden. This choice of g specializes BayesRVAT to a burden test, as the parameters $\boldsymbol{\phi}$ capture the effects of different annotations rather than individual variants, which is a key characteristic of burden tests.

Choice of Priors. We set priors on the parameters $\boldsymbol{\phi}$ to reflect biological knowledge about the expected effects of different annotations (**Supplementary Figure 1**). For example, we apply strong priors to pLoF variants, ensuring that carriers are highly likely to receive a gene burden score close to one. In contrast, for other non-synonymous variants, we use weaker priors with greater variability, accounting for the uncertainty on their effects. For functional, regulatory and splicing annotation scores, we apply priors that allow moderate, positive adjustments to the burden score. Full details on the choice of priors are provided in **Supplementary Information**.

3.4 Implementation Details

We implemented BayesRVAT using PyTorch, leveraging its automatic differentiation capabilities for optimization. For optimizing the ELBO, we initialized the variational parameters $\boldsymbol{\psi}$ to match the prior $p(\boldsymbol{\phi})$, randomly initialized the effect sizes of covariates $\boldsymbol{\alpha}$ and the gene-level effect β to ≈ 0, and the residual variance σ_n^2 to one. We used the Adam optimizer with a learning rate of 1×10^{-2}, operating in full-batch mode. To approximate the expectation in the ELBO, we used a single Monte Carlo sample per gradient step. The optimization was performed for a maximum of 5,000 gradient steps; however, we terminated the optimization after 1,000 steps if the P value exceeded 0.05, indicating a lack of signal. For estimating the IW-ELBO to compute P values during association testing, we approximated the outer expectation by averaging 64 Monte Carlo estimates, each using $K = 8$ importance samples. Notably, the accuracy of the

resulting P values can be adjusted by varying the number of importance samples (**Supplementary Figure 6**). Finally, for calculating the annotation importance scores, we computed the expectations of burden score differences over 64 Monte Carlo samples. In real data analyses, traits were rank-inverse transformed to a unit Gaussian distribution—a standard practice in genetic analyses[38, 30].

4 Experiments

4.1 Dataset preprocessing and variant annotations.

All experiments were performed using the UKB cohort [7], based on the latest release of the whole-exome sequencing (WES) data. Individual and variant quality control (QC) was conducted following the protocols described in the GeneBass study [19]. For variant annotations, we used the same set of annotations considered in [9], which we ultimately collapsed into 25 annotations: three based on variant consequences(pLoF [5], missense and other non-synonymous variants, using VEP [33]), MAF, five functional impact scores (CADD [42], SIFT [22], PolyPhen-2 [1], PrimateAI [44], and Condel [13]) , two splicing impact predictions (SpliceAI [16] delta score and the AbSpliceDNA [47] score), eight RNA-binding protein binding propensity scores (delta scores from DeepRiPE [12]), and six regulatory annotations (principal components of DeepSEA [52] delta embeddings). We used the phred scale to encode each annotation, ensuring that higher values corresponded to higher pathogenicity[26]. The processed dataset consisted of 329,087 unrelated European individuals, 5,845,828 variants with MAF $\leq 0.1\%$, 16,458 genes, and 25 variant annotations. Full details on the individual and variant quality control (QC) and the preprocessing of the variant annotations can be found in **Supplementary Information**.

4.2 Methods Considered

We compared BayesRVAT with several commonly used burden score strategies:

- **pLoF**: A burden test based on the sum of pLoF variants;
- **ACAT-Conseq**: Burden tests were performed separately for the sum of pLoF, missense, and other non-synonymous variants, and the results were aggregated using the Aggregated Cauchy Association Test (ACAT) [27]. This approach is similar to the allelic series burden test model described in [31];
- **ACAT-MultiAnnot**: Burden tests were performed across all consequence categories and continuous annotations used in BayesRVAT, resulting in a total of 25 tests per gene. These results were then aggregated using the ACAT method. This strategy is similar to the burden test implemented in [26].

All burden tests were implemented as likelihood ratio tests within a linear model framework, adjusting for age, sex, the top twenty genetic principal components and WES batch effect covariates (**Supplementary Information**).

[5] pLoF is defined as any of the following variant consequences: splice donor, frameshift, splice acceptor, stop gained, stop lost or start lost.

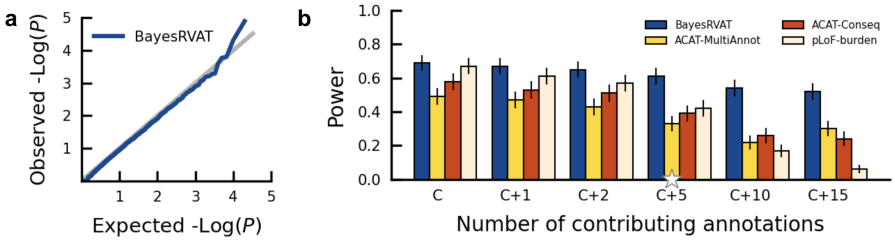

Fig. 2: **Evaluation of calibration and power in BayesRVAT using synthetic data.** (**a**) QQ plot assessing the calibration of P values from BayesRVAT on synthetic data generated under the null model with no genetic effects. (**b**) Statistical power comparison between BayesRVAT, pLoF burden test, ACAT-Conseq and ACAT-MultiAnnot across varying numbers of contributing continuous annotations: simulating only effects from pLoF and missense consequences (C), and considering additional effects from 1, 2, 5, 10, and 15 continuous annotations. Power is measured at the exome-wide significance threshold of $P < 2.5 \times 10^{-6}$, computed over 100 replicates for each scenario.

4.3 Simulations

Simulation Setup. We used synthetic data to assess both the calibration and power of BayesRVAT under various simulated conditions, based on 100,000 individuals from the processed UKB cohort. To evaluate power, we simulated additive genetic effects from a gene burden test using the additive model with saturation in Eq 7. We varied key parameters such as sample size, variance explained by the burden, and the number of continuous annotations contributing to the burden score. Power was estimated at the exome-wide significance threshold of $P < 2.5 \times 10^{-6}$, with 100 replicates performed for each simulation configuration. To assess calibration, we simulated phenotypes under a null model with no genetic effects. Full details on the simulation procedures are provided in **Supplementary Information**.

Results. BayesRVAT showed well-calibrated P values when simulating under the null model with no genetic effects (**Figure 2a**). In power assessments, BayesRVAT maintained robustness across various numbers of contributing annotations, showing resilience to the inclusion of non-informative annotations (**Figure 2b**). Furthermore, as the number of causal annotations increased, BayesRVAT consistently outperformed alternative methods, sustaining higher power in more complex genetic architectures with many contributing annotations (**Figure 2b**). The superior performance of BayesRVAT was maintained across varying sample sizes and levels of variance explained by genetics (**Supplementary Figure 2**).

4.4 Analysis of blood biomarkers

Setup. We applied BayesRVAT to analyze twelve key blood traits from the UKB cohort (**Supplementary Information**). We compared the performance of BayesRVAT with the pLoF, ACAT-Conseq and ACAT-MultiAnnot burden tests.

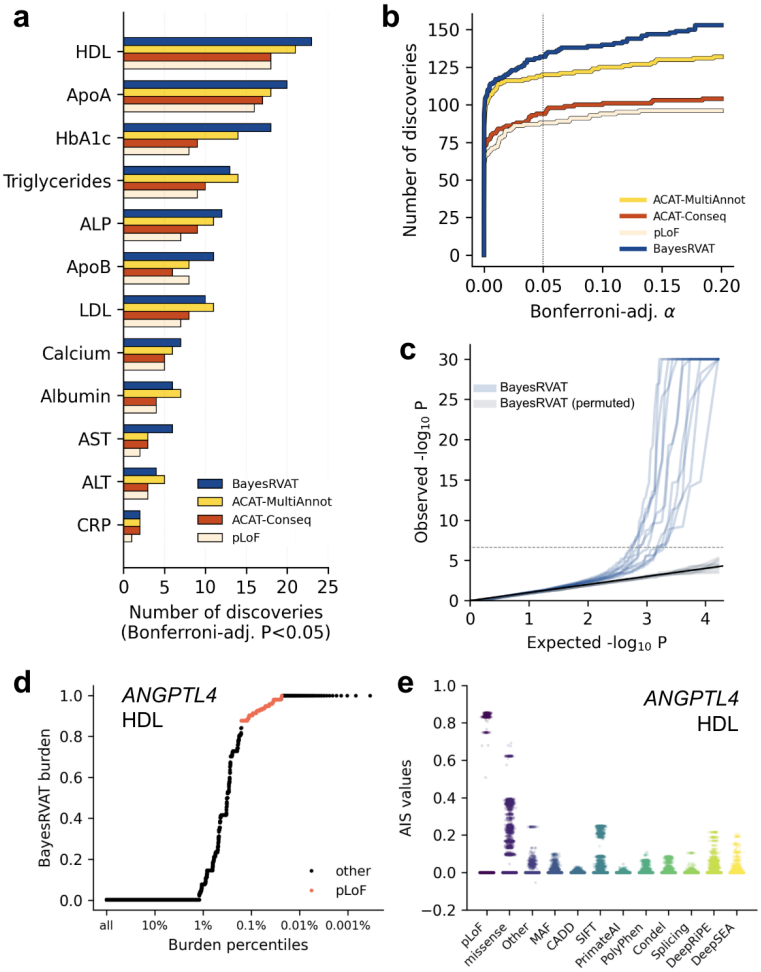

Fig. 3: **Analysis of blood biomarkers in the UK Biobank.** (**a**) Number of significant gene-trait associations (Bonferroni-adjusted $P < 0.05$) discovered by BayesRVAT, ACAT-MultiAnnot, ACAT-Conseq, and pLoF burden tests for each analyzed blood trait. (**b**) Cumulative number of discoveries at varying Bonferroni-adjusted significance thresholds α. (**c**) QQ plot showing the distribution of P values from BayesRVAT in real data and under a null with permuted genotype data, confirming well-calibrated P values. (**d**) Burden scores learned by BayesRVAT for *ANGPTL4* and HDL cholesterol across burden percentiles, showing individuals carrying pLoF mutations in red. (**e**) Annotation importance scores (AIS) from BayesRVAT for the association between *ANGPTL4* and HDL cholesterol, which highlights contributions from missense, SIFT, DeepRiPE, and DeepSEA annotations.

Results. BayesRVAT identified a greater number of significant gene-trait associations compared to other methods (132 for BayesRVAT vs 120 for ACAT-

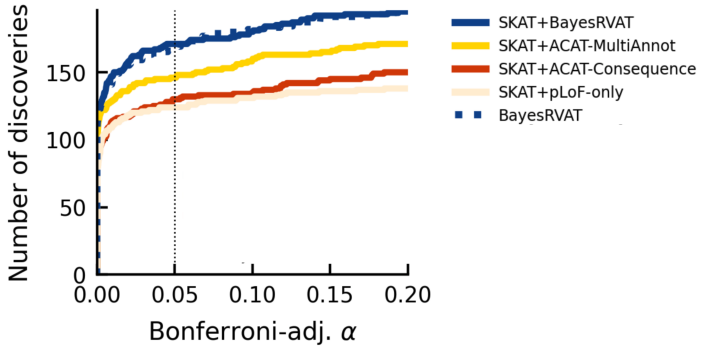

Fig. 4: **Integration of BayesRVAT and other burden tests with variance component models.** Comparison of cumulative significant gene-trait discoveries (Bonferroni-adjusted $P < 0.05$) across various methods: BayesRVAT + SKAT, ACAT-MultiAnnot + SKAT, ACAT-Conseq + SKAT, and pLoF + SKAT. We also report BayesRVAT without SKAT integration, highlighting its superior performance compared to other optimal tests even without SKAT integration.

MultiAnnot, 94 for ACAT-Conseq, and 88 for pLoF; Bonferroni-adjusted P < 0.05; **Figure 3a-b**; **Supplementary Figure 3**), while demonstrating well-calibrated P values under the null, obtained permuting genotype data (**Figure 3c**). Interestingly, BayesRVAT consistently outperformed the ACAT-MultiAnnot burden test (**Supplementary Figure 3**), except in cases where strong deviations from the allelic series assumptions occurred—such as when other annotations had stronger effects than pLoF variants (**Supplementary Figure 4**). Among the associations uniquely identified by BayesRVAT, several showed strong biological relevance. For instance, BayesRVAT uniquely detected an association between *ANGPTL4* and HDL cholesterol (BayesRVAT-P $< 5 \times 10^{-7}$ vs ACAT-MultiAnnot-P $> 5 \times 10^{-6}$ vs pLoF-P $> 5 \times 10^{-4}$). *ANGPTL4* is a key regulator of lipid metabolism, inhibiting lipoprotein lipase, which affects triglyceride breakdown and HDL cholesterol levels, with certain variants associated with increased HDL and cardiovascular protection [49]. In this case, BayesRVAT's burden score assigned higher weight to annotations beyond loss-of-function (pLoF) mutations (**Figure 3d**), with AIS scores indicating contributions from missense, SIFT, DeepRiPE, and DeepSEA annotations (**Figure 3e**). Additional interesting hits are presented in **Supplementary Figure 5**.

4.5 Optimal test integrating BayesRVAT burden with variance component tests

To further evaluate the performance of BayesRVAT, we integrated it into an optimal test that combines both burden and the SKAT variance component model. Following recent methodologies [31, 26, 9], we used ACAT to combine the burden and SKAT test results. This approach is equivalent to the optimal SKAT test SKAT-O initially introduced in [23]. We compared the performance of

BayesRVAT + SKAT against other optimal tests integrating alternative burden models with SKAT, including pLoF, ACAT-Conseq, and ACAT-MultiAnnot. Even without SKAT integration, BayesRVAT outperformed all other methods (164 for BayesRVAT vs 145 for ACAT-MultiAnnot + SKAT, 126 for ACAT-Conseq + SKAT, and 124 for pLoF + SKAT; Bonferroni-adjusted $\alpha = 0.05$; **Figure 4**). When integrated with SKAT, BayesRVAT showed a modest further improvement, identifying 4 additional significant associations (Bonferroni-adjusted $\alpha = 0.05$; **Figure 4**).

5 Discussion

In this work, we introduced BayesRVAT, a new Bayesian model for RVAT. By leveraging variational inference and an approximate likelihood ratio test for association testing, BayesRVAT provides a flexible framework that can accommodate multiple aggregation functions and prior assumption. We specialized BayesRVAT to a Bayesian burden test and demonstrated its utility through synthetic and real data analyses. In real data, BayesRVAT identified multiple biologically meaningful hits missed by other approaches. For example, BayesRVAT uniquely identified associations such as *ANGPTL4* with HDL cholesterol (**Figure 3d-e**), consistent with *ANGPTL4*'s established role in lipid metabolism [49]. It also detected associations between *EPB42* and glycated hemoglobin (**Supplementary Figure 5**), in line with recent findings linking *EPB42* variants to glycemic traits [20], and between *NPC1L1* and apolipoprotein B (**Supplementary Figure 5**), reflecting its impact on lipid transport and metabolism [17]. Additionally, when integrated with variance component models (e.g., SKAT), BayesRVAT continued to outperform alternative models, demonstrating broad applicability.

In ongoing work, we are exploring priors that model individual variant's effects rather than effects from variant annotations, an approach that provides flexibility similar to omnibus methods integrating burden and variance component tests [34, 26, 23]. We are also using BayesRVAT to investigate the benefits and limitations of learning posteriors across genes and traits, inspired by recent advances in RVAT methodologies [9]. Lastly, we aim to leverage BayesRVAT to study traits where traditional RVATs have been less successful, such as psychiatric conditions and imaging-derived phenotypes.

Use of artificial intelligence. In the preparation of this manuscript, we utilized the large language model GPT-4o (https://chat.openai.com/) for editing assistance, including language polishing and clarification of text. Although this tool assisted in refining the manuscript's language, it was not used to generate contributions to the original research, data analysis, or interpretation of results. All final content decisions and responsibilities rest with the authors.

Acknowledgements. We thank Eva Holtkamp and Shubhankar Londhe for critical insights on our blood trait analysis. This research has been conducted using the UK Biobank Resource (Application Number 87065). F.P.C. and A.N. were funded by the Free State of Bavaria's Hightech Agenda through the Institute of AI for Health (AIH).

A.N. is supported by the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS".

Author contributions. A.N. and F.P.C. implemented the methods. A.N. and F.P.C. analyzed the data. A.N., N.C., and F.P.C. interpreted the results. A.N., N.C., and F.P.C. wrote the manuscript. F.P.C. conceived the study with support from N.C. F.P.C. and N.C. supervised the work.

References

[1] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. Nature methods, 7(4):248–249, 2010.

[2] Joshua D Backman, Alexander H Li, Anthony Marcketta, Dylan Sun, Joelle Mbatchou, Michael D Kessler, Christian Benner, Daren Liu, Adam E Locke, Suganthi Balasubramanian, et al. Exome sequencing and analysis of 454,787 uk biobank participants. Nature, 599(7886):628–634, 2021.

[3] Saonli Basu and Wei Pan. Comparison of statistical tests for disease association with rare variants. Genetic epidemiology, 35(7):606–619, 2011.

[4] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. Journal of the American statistical Association, 112(518):859–877, 2017.

[5] Nadav Brandes, Nathan Linial, and Michal Linial. Pwas: proteome-wide association study—linking genes and phenotypes by functional variation in proteins. Genome biology, 21(1):173, 2020.

[6] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. arXiv preprint arXiv:1509.00519, 2015.

[7] 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.

[8] Elizabeth T Cirulli, Simon White, Robert W Read, Gai Elhanan, William J Metcalf, Francisco Tanudjaja, Donna M Fath, Efren Sandoval, Magnus Isaksson, Karen A Schlauch, et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nature communications, 11(1):542, 2020.

[9] Brian Clarke, Eva Holtkamp, Hakime Öztürk, Marcel Mück, Magnus Wahlberg, Kayla Meyer, Felix Munzlinger, Felix Brechtmann, Florian R Hölzlwimmer, Jonas Lindner, et al. Integration of variant annotations using deep set networks boosts rare variant association testing. Nature Genetics, pages 1–10, 2024.

[10] Calliope A Dendrou, Adrian Cortes, Lydia Shipman, Hayley G Evans, Kathrine E Attfield, Luke Jostins, Thomas Barber, Gurman Kaur, Subita Balaram Kuttikkatte, Oliver A Leach, et al. Resolving tyk2 locus genotype-to-phenotype differences in autoimmunity. Science translational medicine, 8(363):363ra149–363ra149, 2016.

[11] Jan P Engelmann, Alessandro Palma, Jakub M Tomczak, Fabian Theis, and Francesco Paolo Casale. Mixed models with multiple instance learning. In International Conference on Artificial Intelligence and Statistics, pages 3664–3672. PMLR, 2024.

[12] Mahsa Ghanbari and Uwe Ohler. Deep neural networks for interpreting rna-binding protein target preferences. Genome research, 30(2):214–226, 2020.

[13] Abel González-Pérez and Nuria López-Bigas. Improving the assessment of the outcome of nonsynonymous snvs with a consensus deleteriousness score, condel. The American Journal of Human Genetics, 88(4):440–449, 2011.

[14] Fang Han and Wei Pan. A data-adaptive sum test for disease association with multiple common or rare variants. Human heredity, 70(1):42–54, 2010.

[15] Bertrand Iooss and Paul Lemaître. A review on global sensitivity analysis methods. Uncertainty management in simulation-optimization of complex systems: algorithms and applications, pages 101–122, 2015.

[16] Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, et al. Predicting splicing from primary sequence with deep learning. Cell, 176(3):535–548, 2019.

[17] Lin Jia, Jenna L Betters, and Liqing Yu. Niemann-pick c1-like 1 (npc1l1) protein in intestinal and hepatic cholesterol transport. Annual review of physiology, 73(1):239–259, 2011.

[18] Sean J Jurgens, Seung Hoan Choi, Valerie N Morrill, Mark Chaffin, James P Pirruccello, Jennifer L Halford, Lu-Chen Weng, Victor Nauffal, Carolina Roselli, Amelia W Hall, et al. Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the uk biobank. Nature genetics, 54(3):240–250, 2022.

[19] Konrad J Karczewski, Matthew Solomonson, Katherine R Chao, Julia K Goodrich, Grace Tiao, Wenhan Lu, Bridget M Riley-Gillis, Ellen A Tsai, Hye In Kim, Xiuwen Zheng, et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 uk biobank exomes. Cell Genomics, 2(9), 2022.

[20] Young Jin Kim, Sanghoon Moon, Mi Yeong Hwang, Sohee Han, Hye-Mi Jang, Jinhwa Kong, Dong Mun Shin, Kyungheon Yoon, Sung Min Kim, Jong-Eun Lee, et al. The contribution of common and rare genetic variants to variation in metabolic traits in 288,137 east asians. Nature communications, 13(1):6642, 2022.

[21] Diederik P Kingma. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[22] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. Nature protocols, 4(7):1073–1081, 2009.

[23] Seunggeun Lee, Michael C Wu, and Xihong Lin. Optimal tests for rare variant effects in sequencing association studies. Biostatistics, 13(4):762–775, 2012.

[24] Seunggeung Lee, Gonçalo R Abecasis, Michael Boehnke, and Xihong Lin. Rare-variant association analysis: study designs and statistical tests. The American Journal of Human Genetics, 95(1):5–23, 2014.

[25] Bingshan Li and Suzanne M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. The American Journal of Human Genetics, 83(3):311–321, 2008.

[26] 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.

[27] 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.

[28] Benjamin A Logsdon, James Y Dai, Paul L Auer, Jill M Johnsen, Santhi K Ganesh, Nicholas L Smith, James G Wilson, Russell P Tracy, Leslie A Lange, Shuo Jiao, et al. A variational bayes discrete mixture test for rare variant association. Genetic epidemiology, 38(1):21–30, 2014.

[29] Bo Eskerod Madsen and Sharon R Browning. A groupwise association test for rare mutations using a weighted sum statistic. PLoS genetics, 5(2): e1000384, 2009.

[30] Zachary R McCaw, Jacqueline M Lane, Richa Saxena, Susan Redline, and Xihong Lin. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. Biometrics, 76(4):1262–1272, 2020.

[31] Zachary R McCaw, Colm O'Dushlaine, Hari Somineni, Michael Bereket, Christoph Klein, Theofanis Karaletsos, Francesco Paolo Casale, Daphne Koller, and Thomas W Soare. An allelic-series rare-variant association test for candidate-gene discovery. The American Journal of Human Genetics, 110(8):1330–1342, 2023.

[32] Barbara McClintock. The relation of homozygous deficiencies to mutations and allelic series in maize. Genetics, 29(5):478, 1944.

[33] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. Genome biology, 17:1–14, 2016.

[34] Remo Monti, Pia Rautenstrauch, Mahsa Ghanbari, Alva Rani James, Matthias Kirchler, Uwe Ohler, Stefan Konigorski, and Christoph Lippert. Identifying interpretable gene-biomarker associations with functionally informed kernel-based tests in 190,000 exomes. Nature communications, 13 (1):5332, 2022.

[35] Stephan Morgenthaler and William G Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 615(1-2):28–56, 2007.

[36] Andrew P Morris and Eleftheria Zeggini. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic epidemiology, 34(2):188–193, 2010.

[37] Kiran Musunuru and Sekar Kathiresan. Genetics of common, complex coronary artery disease. Cell, 177(1):132–145, 2019.

[38] Bo Peng, Robert K Yu, Kevin L DeHoff, and Christopher I Amos. Normalizing a large number of quantitative traits using empirical normal quantile transformation. In BMC proceedings, volume 1, pages 1–5. Springer, 2007.

[39] Robert M Plenge, Edward M Scolnick, and David Altshuler. Validating therapeutic targets through human genetics. Nature reviews Drug discovery, 12(8):581–594, 2013.

[40] Alkes L. Price, Gregory V. Kryukov, Paul I.W. de Bakker, Shaun M. Purcell, Jeff Staples, Lee-Jen Wei, and Shamil R. Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. The American Journal of Human Genetics, 86(6):832–838, 2010. ISSN 0002-9297. https://doi.org/https://doi.org/10.1016/j.ajhg.2010.04.005. URL https://www.sciencedirect.com/science/article/pii/S0002929710002077.

[41] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In Artificial intelligence and statistics, pages 814–822. PMLR, 2014.

[42] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. Cadd: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research, 47(D1):D886–D894, 2019.

[43] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In International conference on machine learning, pages 1278–1286. PMLR, 2014.

[44] Laksshman Sundaram, Hong Gao, Samskruthi Reddy Padigepati, Jeremy F McRae, Yanjun Li, Jack A Kosmicki, Nondas Fritzilas, Jörg Hakenberg, Anindita Dutta, John Shon, et al. Predicting the clinical impact of human mutation with deep neural networks. Nature genetics, 50(8):1161–1170, 2018.

[45] Valentine Svensson, Adam Gayoso, Nir Yosef, and Lior Pachter. Interpretable factor models of single-cell rna-seq via variational autoencoders. Bioinformatics, 36(11):3418–3421, 2020.

[46] Guhan Ram Venkataraman, Christopher DeBoever, Yosuke Tanigawa, Matthew Aguirre, Alexander G Ioannidis, Hakhamanesh Mostafavi, Chris CA Spencer, Timothy Poterba, Carlos D Bustamante, Mark J Daly, et al. Bayesian model comparison for rare-variant association studies. The American Journal of Human Genetics, 108(12):2354–2367, 2021.

[47] Nils Wagner, Muhammed H Çelik, Florian R Hölzlwimmer, Christian Mertes, Holger Prokisch, Vicente A Yépez, and Julien Gagneur. Aberrant splicing prediction across human tissues. Nature genetics, 55(5):861–870, 2023.

[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] Long-Yan Yang, Cai-Guo Yu, Xu-Hong Wang, Sha-Sha Yuan, Li-Jie Zhang, Jia-Nan Lang, Dong Zhao, and Ying-Mei Feng. Angiopoietin-like protein 4 is a high-density lipoprotein (hdl) component for hdl metabolism and function in nondiabetic participants and type-2 diabetic patients. Journal of the American Heart Association, 6(6):e005973, 2017.

[50] Yi Yang, Saonli Basu, and Lin Zhang. A bayesian hierarchically structured prior for rare-variant association testing. Genetic epidemiology, 45(4):413–424, 2021.

[51] Nengjun Yi and Degui Zhi. Bayesian analysis of rare variants in genetic association studies. Genetic epidemiology, 35(1):57–69, 2011.

[52] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. Nature methods, 12(10):931–934, 2015.

[53] Wei Zhou, Wenjian Bi, Zhangchen Zhao, Kushal K Dey, Karthik A Jagadeesh, Konrad J Karczewski, Mark J Daly, Benjamin M Neale, and Seunggeun Lee. Saige-gene+ improves the efficiency and accuracy of set-based rare variant association tests. Nature genetics, 54(10):1466–1469, 2022.