Submitted to the Annals of Applied Statistics

COMPUTATIONALLY EFFICIENT WHOLE-GENOME SIGNAL REGION DETECTION FOR QUANTITATIVE AND BINARY TRAITS

BY FAN WANG[1,a] WEI ZHANG[2,b], AND FANG YAO[2,c]

[1]*Department of Biostatistics, Mailman School of Public Health, Columbia University,* [a]*fw2400@cumc.columbia.edu*

[2]*School of Mathematical Sciences, Center for Statistical Science, Peking University,* [b]*wei_zhang@stu.pku.edu.cn;* [c]*fyao@math.pku.edu.cn*

The identification of genetic signal regions in the human genome is critical for understanding the genetic architecture of complex traits and diseases. Numerous methods based on scan algorithms (i.e. QSCAN, SCANG, SCANG-STAAR) have been developed to allow dynamic window sizes in whole-genome association studies. Beyond scan algorithms, we have recently developed the binary and re-search (BiRS) algorithm, which is more computationally efficient than scan-based methods and exhibits superior statistical power. However, the BiRS algorithm is based on two-sample mean test for binary traits, not accounting for multidimensional covariates or non-binary outcomes. In this work, we propose a new maximal score test based on summary statistics computed from a generalized linear model, which accommodates regression-based statistics and allows testing of both continuous and binary outcomes. We then present a distributed version of the BiRS algorithm (dBiRS) that incorporates this new test, enabling parallel computing of block-wise results by aggregation through a central machine to ensure both detection accuracy and computational efficiency, which has theoretical guarantees for controlling family-wise error rates and false discovery rates while maintaining the power advantages of the original algorithm. Applying dBiRS to detect genetic regions associated with fluid intelligence and prospective memory using whole-exome sequencing data from the UK Biobank, we validate previous findings and identify numerous novel rare variants near newly implicated genes. These discoveries offer valuable insights into the genetic basis of cognitive performance and neurodegenerative disorders, highlighting the potential of dBiRS as a scalable and powerful tool for whole-genome signal region detection.

1. Introduction. Understanding the genetic underpinnings of human diseases and traits remains a central focus of genetic research. Genome-Wide Association Studies (GWAS) have been instrumental in exploring the genetic architecture of complex diseases and traits over the past decade (Visscher et al., 2017). By using array-based technologies, GWAS analyzes millions of single nucleotide polymorphisms (SNPs) across the genome to identify those associated with specific traits or disease outcomes. While GWAS has successfully identified thousands of common genetic variants linked to disease susceptibility, these common variants explain only a small fraction of heritability, which is often referred to as the "missing heritability problem"(Manolio et al., 2009). The majority of genetic variants in the human genome are rare, and these rare variants are believed to contribute significantly to the unexplained heritability. However, the classical GWAS approach which focuses on single-SNP-based analysis has very limited power for analyzing rare variants, as the effect of a single SNP can be too small to be detected. (Liu et al., 2010; Bakshi et al., 2016).

Keywords and phrases: family-wise error rate, signal detection, distributed learning, whole genome association studies.

To address this limitation, whole genome sequencing (WGS) studies are being conducted to identify rare variants associated with disease susceptibility. Progress has been made in developing set-based association methods, which test multiple variants jointly by aggregating their effects within defined genomic regions. These methods include burden tests (Morgenthaler and Thilly, 2007; Madsen and Browning, 2009), the Sequence Kernel Association Test (SKAT) (Wu et al., 2011), and STAAR which incorporates variant functional annotations to enhance detection power (Li et al., 2020). The STAAR-O test extends the STAAR framework as an omnibus test by combining multiple annotation-weighted methods into a single unified test (Li et al., 2020). A common challenge in these approaches is defining regions for variant sets, especially in non-coding or intergenic regions without clear functional boundaries. STAAR addresses this issue by applying gene-centric analysis to well-defined gene-associated regions and fixed sliding window-based analysis to regions without clear boundaries. Alternative approaches include the scan statistic (Naus, 1982), which systematically searches the human genome using a fixed window size. Mean-based scan statistic methods are later proposed for DNA copy number analysis, allowing the use of multiple window sizes in settings closely related to change-point detection problems (Jeng, Cai and Li, 2010; Olshen et al., 2004; Zhang et al., 2010). Other notable frameworks, such as those based on the knockoff (i.e.KnockoffScreen) (He et al., 2021a,b), conducts genome-wide set-based analyses to identify signal regions while mitigating the effects of correlation confounding. However, fixed-window approaches can result in a loss of power because the sizes of signal regions can vary across the genome.

Further advancements have been made to enable signal region detection with dynamic window sizes. Li et al. (2019) introduced SCANG which combines the scan algorithm proposed by Jeng, Cai and Li (2010) with burden tests, SKAT, and the omnibus test to continuously scan the genome. However, this approach lacks comprehensive theoretical and empirical analysis of false discoveries and shows limited power when functional annotations are not incorporated (Li et al., 2022). Building on the scan algorithm, Li, Liu and Lin (2020) developed the quadratic scan statistic (QSCAN), which aggregates information across intervals of varying sizes. QSCAN provides theoretical guarantees for controlling false discoveries and has demonstrated strong detection performance, particularly when signal regions contain both causal and neutral variants. However, scan-based methods require calculating test statistics for many candidate intervals and applying a fixed threshold for selection, which is computationally intensive and tends to be conservative. SCANG-STAAR (Li et al., 2022) extends SCANG's dynamic window scanning by incorporating multiple functional annotations via STAAR to boost the power of detecting rare variant associations. While functional annotations can enhance power and interpretability, their effectiveness relies heavily on the quality and relevance of the annotations, which may introduce biases, computational challenges, and increased resource demand.

Beyond the scan algorithm, Zhang, Wang and Yao (2025) proposed the binary and research (BiRS) algorithm for detecting signal regions with dynamic window sizes in whole genome sequencing (WGS) studies. The BiRS algorithm iteratively splits identified regions until a minimum size is reached, providing theoretical guarantees for FWER and FDR, which is shown more computationally efficient than scan-based algorithms and demonstrates superior power compared to QSCAN and KnockoffScreen. Impressively, BiRS surpasses SCANG-STAAR in detecting moderate or weak signals without requiring functional annotations. The combination of computational speed, increased detection power and theoretical rigor make BiRS as a significant advancement in signal region detection for WGS studies. The BiRS algorithm was combined with the DCF two-sample mean test (Xue and Yao, 2020) for direct application to binary traits. However, it has not yet been extended to accounting for multi-dimensional covariates or testing for non-binary outcomes.

In this paper, we generalize the BiRS algorithm to a distributed version (dBiRS) to combine with a new infinity-norm score test based on the summary statistics obtained from a generalized linear model. As the proposed test statistic fundamentally differs from the DCF two-sample mean test, this extension is non-trivial and operates in two main stages, combining results from local and central machines to support parallel computing and maintain computational efficiency, similar to distributed learning frameworks (Cai and Wei, 2022). In the first stage, the regression-based BiRS is applied within genomic blocks, where local machines detect signal regions by calculating the maximal score test statistics and thresholds within each block. Instead of simply aggregating block-wise results using a global threshold, the detected signal regions, along with their corresponding test statistics and thresholds, are commuted to a central machine in the second stage. Then the central machine evaluates the significance of each block by performing a second layer of BiRS based on block-wise test statistics. Finally, a new threshold is constructed by multiplier bootstrap to reassess the signal regions within the significant blocks. We have shown theoretically that this dBiRS algorithm is able to properly control size and consistently identify true signal regions under more general alternative structures and in the presence of model misspecification under the alternative hypothesis. Simulated studies also demonstrate that dBiRS is more accurate and robust than the state-of-the-art KnockoffScreen and scan procedures.

Finally, we apply the BiRS algorithm to analyze whole exome sequence (WES) data from the UK Biobank, aiming to identify signal regions for intelligence and prospective memory. The dBiRS algorthm identified signal regions involving 327 genes, including 84 of which were previously reported to be assocaited with intelligence. Notable discoveries include both common and rare variants linked to cognitive functions, such as those in COL16A1, CRTC2, and PTPRF. Variants in genes like CRTC2, BRWD1, and TOP2B were also identified, highlighting their roles in endothelial function, immune system regulation, and neuronal development, respectively. In addition, the analysis revealed rare variants in 22 genes not previously associated with intelligence. Some of these genes are linked to Alzheimer's disease, brain connectivity, and neuronal development, providing novel insights into the genetic basis of cognitive traits. For prospective memory, the study identified eight novel genes, including OMA1 and CNGB3, which are associated with neuroimaging measurements and cognitive functions.

The rest of the article is organized as follows. In Section 2, we introduce the testing procedure under GLM and describe the proposed dBiRS algorithm. We conduct comprehensive simulations in Section 3 to demonstrate that the proposed method enjoys preferable numerical performance compared with existing approaches. In Section 4, we apply the dBiRS algorithm to conduct Whole Exome Sequencing (WES) analyses on intelligence and prospective memory, aiming to deepen our understanding of cognitive aging and uncover genetic factors contributing to the risk of neurodegenerative disorders. We conclude the article with a discussion in Section 5, while the theoretical properties of the proposed algorithm, including size control and detection consistency, technical assumptions, lemmas, and proofs of the theoretical results, are provided in the Supplementary Material (Wang, Zhang and Yao, 2025).

2. Maximal score test and distributed detection algorithm.

2.1. *Maximal score test based on summary statistic.* Suppose there are n observations in the study. For the i-th observation, Y_i represents the outcome, $X_i = (X_{i1}, \ldots, X_{iq})^\top$ is a vector containing q covariates, and $G_i = (G_{i1}, \ldots, G_{ip})^\top$ is the genotype vector with p variants. Let $Y = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$, and $\boldsymbol{G} = (G_1, \ldots, G_n)^\top$. Conditional on X_i and G_i, we assume that Y_i belongs to an exponential family with the density $f(Y_i) = \exp\left\{\frac{Y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(Y_i, \phi)\right\}$, where $a_i(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions, and θ_i and ϕ

are the canonical parameter and dispersion parameter, respectively, which indicates that Y following a generalized linear model (GLM):

$$g(\eta) = \boldsymbol{X}\gamma + \boldsymbol{G}\beta, \tag{1}$$

where $\eta = \mathbb{E}(Y \mid \boldsymbol{X}, \boldsymbol{G})$ and $g(\cdot)$ is a monotone link function.

Under the global null model where no genetic effect is present across the genome (i.e. $\beta = 0$), the GLM in (1) simplifies to $g(\eta) = \boldsymbol{X}\gamma$. Let $\hat{\eta}_0 = g^{-1}(\boldsymbol{X}\hat{\gamma})$, where $\hat{\gamma}$ is the maximum likelihood estimator (MLE) of γ under the global null model. The variance of Y_i is $\operatorname{var}(Y_i) = a_i(\phi)v(\eta_i)$, where $v(\eta_i) = b''(\theta_i)$ is a variance function. We define $\boldsymbol{\Lambda} = \operatorname{diag}\{a_1(\phi)v(\eta_{01}), \ldots, a_n(\phi)v(\eta_{0n})\}$ and let $\boldsymbol{P} = \boldsymbol{\Lambda}^{-1} - \boldsymbol{\Lambda}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}$.

In genome-wide association studies (GWAS) and whole-genome sequencing (WGS) studies, the test statistic for the j-th variant is constructed using the working marginal model $g(\eta_i) = X_i^T\alpha + G_{ij}\beta_j$, where we regress Y_i on each variant G_{ij}, adjusting for the covariates X_i. The marginal score test statistic for β_j of the j-th variant is given by

$$U_j = G_{\cdot j}^T(Y - \hat{\eta}_0)/\sqrt{n}. \tag{2}$$

Marginal score statistics U_j's are often made available in public databases or provided by investigators to facilitate meta-analysis across multiple cohorts. Let $U = (U_1, \ldots, U_p)^\top$, under the global null model (i.e. $\beta = 0$), $U \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{G}^\top P\boldsymbol{G}/n$. Let $\hat{\boldsymbol{\Lambda}} = \operatorname{diag}\left\{a_1(\hat{\phi})v(\hat{\eta}_{01}), \ldots, a_n(\hat{\phi})v(\hat{\eta}_{0n})\right\}$, where $\hat{\phi}$ is the MLE of ϕ under the global null hypothesis and $\hat{\eta}_{0i}$ is the i-th coordinate of $\hat{\eta}_0$, $i = 1, \ldots, n$. Let $\hat{\boldsymbol{P}} = \hat{\boldsymbol{\Lambda}}^{-1} - \hat{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\hat{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\hat{\boldsymbol{\Lambda}}^{-1}$. In practice, Σ can be well approximated by $\boldsymbol{G}^\top\hat{P}\boldsymbol{G}/n$.

To test for the global null hypothesis, our proposed test statistic is defined as the maxima of marginal scores:

$$T = \|U\|_\infty = \left\|\boldsymbol{G}^\top(Y - \hat{\eta}_0)/\sqrt{n}\right\|_\infty,$$

where $\|x\|_\infty = \max\{|x_1|, \ldots, |x_p|\}$ for $x \in \mathbb{R}^p$. The null hypothesis is rejected at a specified significance level α if $T > c(\alpha)$, where $c(\alpha)$ is a predefined threshold, specifically,

$$c(\alpha) = \inf\{t \in \mathbb{R} : \mathbb{P}(T \le t) \ge 1 - \alpha\}.$$

Computing the threshold $c(\alpha)$ requires the eigenvalues of $\hat{\Sigma}$, which are computationally intensive to calculate in practice when p is large. To address this challenge, we propose an efficient Gaussian multiplier bootstrap procedure to approximate $c(\alpha)$. We first generate N normal random vectors of dimension $n \times 1$, denoted as $e_1, e_2, \ldots, e_N$. We then calculate pseudo scores $U^{e_b} = \boldsymbol{G}^\top\hat{P}^{1/2}e_b/\sqrt{n}$ for $b = 1, \ldots, N$. The threshold $c(\alpha)$ is approximated by the $100(1-\alpha)$-th percentile of pseudo scores:

$$\hat{c}(\alpha) = f_\alpha\left(\|U^{e_1}\|_\infty, \ldots, \|U^{e_N}\|_\infty\right),$$

where $f_\alpha(x_1, x_2, \ldots, x_N)$ represents the $100(1-\alpha)$-th percentile of the set $\{x_1, x_2, \ldots, x_N\}$.

2.2. *Detection with maximal score test.* In this subsection, we introduce our signal region detection algorithm combined with the above test statistic T. We first introduce some key concepts for signal regions used throughout the paper. Under the alternative hypothesis that there exists signal regions, we define a signal point with index j given that $\beta_j \neq 0$, $1 \le j \le p$. Given the polygenic nature of the genome, a genetic region I may contain consecutive signal points. We assume that a signal region satisfies the continuity property if $\beta_I \neq 0$, where the "$\neq$" means that no pair of elements is equal. Furthermore, we assume that the signal region

I satisfies the separability assumption: for any area that contains I, the edges of I are signal points and there are no signal points next to I. Lastly, we denote the true signal regions by $I_1^*, \ldots, I_\ell^*$ (if exist) and let $\mathcal{I} = \{I_1^*, \ldots, I_\ell^*\}$.

Our goal is to determine whether signal regions exist and, if so, to identify their locations. Specifically, we aim first to test:

$$H_0 : \mathcal{I} = \emptyset \quad \text{v.s.} \quad H_1 : \mathcal{I} \neq \emptyset, \tag{3}$$

if H_0 is rejected, we proceed to detect each signal region in $\mathcal{I}$. Following Zhang, Wang and Yao (2025), the algorithm first compares T with $\hat{c}(\alpha)$ to determine whether signal regions exit. If $T > \hat{c}(\alpha)$, we conduct a binary search procedure that utilizes a sequence of dynamic thresholds generated by the bootstrap vectors to find the specific locations of the signal regions. For an index set $I = \{i_1, \ldots, i_r\}$, we define $U(I) = (U_{i_1}, U_{i_2}, \ldots, U_{i_r})$ and similar for $U^{e_b}(I)$, $b = 1, \ldots, N$. The detailed binary search procedure is as follows.

At the first step, we split the genome into two segments, denoted as $I_{11} = \{1, 2, \ldots, \lfloor p/2 \rfloor\}$ and $I_{12} = \{\lfloor p/2 \rfloor + 1, \ldots, p\}$ with $\lfloor \cdot \rfloor$ taking the value of the largest integer below. Let $I_1 = I_{11} \cup I_{12}$. The test statistic for the kth segment is $T_{1k} = \|U(I_{1k})\|_\infty$, for $k = 1, 2$, and the threshold for the two tests is calculated by $\hat{c}_1(\alpha) = f_\alpha(\|U^{e_1}(I_1)\|_\infty, \ldots, \|U^{e_N}(I_1)\|_\infty)$. If $T_{1k} > \hat{c}_1(\alpha)$, then I_{1k} is considered to possibly contain signal regions and is subjected to further binary search procedures. Otherwise, it is concluded that there are no signals within I_{1k}.

Suppose we further conduct the binary search procedure for regions that may contain signals. During the j-th binary search, there are n_j segments to be tested ($1 \leq n_j \leq 2^j$), denoted as $I_{j_1}, \ldots, I_{jn_j}$, with $I_j = \cup_{k=1}^{n_j} I_{j_k}$. The critical value for all n_j tests is calculated as

$$\hat{c}_j(\alpha) = f_\alpha(\|U^{e_1}(I_j)\|_\infty, \ldots, \|U^{e_N}(I_j)\|_\infty).$$

If $T_{jk} > \hat{c}_j(\alpha)$ for $k = 1, \ldots, n_j$, we perform binary segmentation on I_{jk} to conduct tests in the next iteration of binary search procedure. The segmentation stops if the length of I_{jk} is sufficiently small such that $|I_{jk}| \leq 2^s$, where s is a truncation parameter. This binary search procedure is repeated iteratively until no significant segments remain or all segments become sufficiently small. The detected signal regions are denoted as $\hat{I}_{11}, \ldots, \hat{I}_{1l_1}$. We want to emphasis that this binary search step utilizes a sequence of dynamic critical values, which enables the detection of more signals compared to procedures with fixed thresholds. Specifically, since $U^{e_b}(I_j)$ is a sub-vector of U^{e_b}, it follows that $\|U^{e_b}(I_j)\|_\infty \leq \|U^{e_b}\|_\infty$, for $b = 1, \ldots, N$. This implies $\hat{c}(\alpha) \geq \hat{c}_1(\alpha) \geq \cdots \geq \hat{c}_j(\alpha) \geq \cdots$, which increase the power to detect weaker signals as the binary search progresses.

We next implement a re-search step to detect signals that may have been missed. We substitute the variables within $\hat{I}_{11}, \ldots, \hat{I}_{1l_1}$ with zeros, and repeat the global test and binary search. This process is iterated until the global test confirms that no signals remain or all segments are sufficiently small. We finally rearrange the detected signal regions based on their locations in the genome, denoted as $\hat{I}_1, \ldots, \hat{I}_\ell$. The complete detection procedure is referred to as sBiRS and is summarized in Algorithm 1.

2.3. *Detection over multiple blocks: a distributed algorithm.* In whole genome association studies, the number of variants in the genome is about 10^7. Implementing this searching algorithm requires the use of bootstrap vectors throughout the procedure, which results in excessive memory usage. Additionally, it is impractical to load the entire genotype matrix at once when working with individual-level data. A straightforward solution is to divide the genome into K blocks and detect signal regions within each block using a significant level of α/K (i.e., the Bonferroni correction), but this approach is conservative in most cases. Alternatively, detection procedures can be performed with fixed thresholds by applying global

Algorithm 1: BiRS with summary statistics (sBiRS)

Input: the vector of marginal scores U; the bootstrap vectors $U^{e_1}, \ldots, U^{e_N}$; truncation parameter s; significant level α;
Calculate $T = \|U\|_\infty$ and $\hat{c}(\alpha) = f_\alpha(\|U^{e_1}\|_\infty, \ldots, \|U^{e_N}\|_\infty)$;
if $T \le \hat{c}(\alpha)$ **then**
  Output: There are no signal regions.
end
if $T > \hat{c}(\alpha)$ **then**
  Conduct the binary search procedure based on U and $U^{e_1}, \ldots, U^{e_N}$ with truncation parameter s;
  Replace the elements of U and $U^{e_1}, \ldots, U^{e_N}$ in estimated signal regions from the last step as zeros;
  Recalculate T and $\hat{c}(\alpha)$;
  while $T > \hat{c}(\alpha)$ **do**
    Conduct the binary search procedure based on the new U and $U^{e_1}, \ldots, U^{e_N}$;
    Replace the elements of U and $U^{e_1}, \ldots, U^{e_N}$ in estimated signal regions from the last step as zeros;
    Recalculate T and $\hat{c}(\alpha)$;
  end
  Rearrange the estimated signal regions to be continuous and separated;
  Output: Estimated signal regions $\hat{I}_1, \ldots, \hat{I}_{\hat{\ell}}$.
end

thresholds to control the detection results within each block. While this method avoids conservativeness, it reduces the power of sBiRS due to the use of higher critical values compared to dynamic thresholds which are adjusted based on prior testing results. This issue is analogous to the power reduction phenomenon observed in distributed learning (Cai and Wei, 2022), where large-scale learning tasks are performed across multiple nodes. To address this problem, we introduce a distributed version of our detection algorithm in this subsection. This approach preserves the advantages of dynamic thresholds while maintaining control over family-wise error rates (FWERs) and false discovery rates (FDRs).

We divide the entire region into K blocks, denoted as $B_1, \ldots, B_K$. For block k ($k = 1, \ldots, K$), the vector of marginal scores is represented by $U(B_k)$, and the corresponding bootstrap vectors are $U^{e_1}(B_k), \ldots, U^{e_N}(B_k)$. Within each block, a local machine applies the BiRS algorithm, resulting in a collection of detected signal regions $\hat{\mathcal{I}}_k = \left\{\hat{I}_1^{(k)}, \ldots, \hat{I}_{\ell_k}^{(k)}\right\}$ for block k, $k = 1, \ldots, K$. The union of the detected signal regions in block k is denoted as $\hat{I}^{(k)} = \cup_{i=1}^{\ell_k} \hat{I}_i^{(k)}$. The corresponding collection of maximum marginal scores is defined as:

$$\mathcal{T}_k = \left\{T(B_k), T(\hat{I}_1^{(k)}), \ldots, T(\hat{I}_{\ell_k}^{(k)})\right\} = \left\{\|U(B_k)\|_\infty, \|U(\hat{I}_1^{(k)})\|_\infty, \ldots, \|U(\hat{I}_{\ell_k}^{(k)})\|_\infty\right\}.$$

For variables related to the bootstrap vectors, we define the following quantities for block k:

$$M(B_k) = (M_1(B_k), \ldots, M_N(B_k))^\top = (\|U^{e_1}(B_k)\|_\infty, \ldots, \|U^{e_N}(B_k)\|_\infty)^\top,$$

$$L(B_k) = (L_1(B_k), \ldots, L_N(B_k))^\top = \left(\|U^{e_1}(\hat{I}^{(k)})\|_\infty, \ldots, \|U^{e_N}(\hat{I}^{(k)})\|_\infty\right)^\top.$$

Finally, we transfer the block results

$$\mathcal{R}_k = \left\{\hat{\mathcal{I}}_k, \mathcal{T}_k, M(B_k), L(B_k)\right\}, \quad k = 1, \ldots, K,$$

to the central machine for further analysis.

In the central machine, we treat these blocks as potential signal points and apply the sBiRS procedure to identify which blocks are significant. Precisely, we define

$$\tilde{U} = (T(B_1), \ldots, T(B_K))^\top,$$

where each component represents the maximum marginal score within each block. The corresponding bootstrap vectors for b-th boostrap are defined as

$$\tilde{U}^{e_b} = (M_b(B_1), \ldots, M_b(B_K))^\top, b = 1, \ldots N.$$

We run the sBiRS algorithm with $\tilde{U}, \tilde{U}^{e_1}, \ldots, \tilde{U}^{e_N}$, truncation parameter $s = 0$ and significant level α. Suppose the significant blocks identified by sBiRS are $B_{j_1}, \ldots, B_{j_{\hat{k}}}$, where $j_1, \ldots, j_{\hat{k}}$ are indices of significant blocks. We remove insignificant blocks, even if there are signal regions detected by the local machine, and then perform a control procedure for the detected signal regions within the significant blocks. For the b-th boostrap vector, we define

$$\tilde{L}^{e_b} = \big(L_b(B_{j_1}), \ldots, L_b(B_{j_{\hat{k}}})\big)^\top = \Big(\|U^{e_b}(\hat{I}^{(j_1)})\|_\infty, \ldots, \|U^{e_b}(\hat{I}^{(j_{\hat{k}})})\|_\infty\Big)^\top.$$

Based on the $\tilde{L}^{e_b}$'s, we calculate new critical values $\hat{c}_{\min}(\alpha)$ for the test statistics in each signal region within the significant blocks:

$$\hat{c}_{\min}(\alpha) = f_\alpha(\tilde{L}^{e_1}, \ldots, \tilde{L}^{e_N}).$$

For the r-th block and i-th signal region within the block ($r = j_1, \ldots, j_{\hat{k}}$, $i = 1, \ldots, \ell_r$), the signal region $\hat{I}_i^{(r)}$ is significant if $T_n(\hat{I}_i^{(r)}) > \hat{c}_{\min}(\alpha)$. This distributed algorithm is summarized in Algorithm 2.

Since the significance level for the detection procedure within each block was set to α, the family-wise error rate (FWER) across multiple tests on the final estimated signal regions may not be controlled. Therefore, it is necessary to construct the threshold $\hat{c}_{\min}(\alpha)$ to ensure FWER control by filtering out less significant signal regions. This control procedure also helps manage the proportion of false discoveries, leading to consistent detection results, as demonstrated in the proof of Theorem 3 in the Supplementary Material (Wang, Zhang and Yao, 2025).

In practice, the dimensions of blocks will be adaptively determined based on data sizes and computational resources. For instance, if the memory resource of each local machine (core) is m Gigabyte (GB), then we recommend the dimension of each block $10^8 m/n$, which results in $np/10^8 m$ blocks. Because signal regions are usually sparsely distributed across the genome, most blocks are unlikely to contain signals. Applying sBiRS to determine the significance of blocks will sequentially remove most of the blocks and result in a reduction of dynamic critical values, which helps find more significant blocks with relatively weaker signals. Moreover, this procedure is equivalent to running sBiRS with truncation $\log_2(\max_k p_k) \le s \le \log_2(\min_k p_k) + 1$, where p_k represents the dimension of block B_k for $k = 1, \ldots, K$.

2.4. *Theoretical guarantees.* For the flow of exposition, we provide a description of the theoretical results here and defer the technical presentation of the family-wise error rate and the detection accuracy in the Supplementary Material (Wang, Zhang and Yao, 2025). For size analysis, after assuming some mild conditions to the exponential family, the dBiRS algorithm asymptotically controls the family-wise error rate of test problem (3). The method allows the dimension to grow at an exponential rate relative to the sample size. For detection accuracy analysis, we prove that the dBiRS algorithm overcomes the power reduction phenomenon and maintains the facilitated properties of the BiRS algorithm under GLM model. Specifically, when there are model misspecifications under the alternative hypothesis, the dBiRS

Algorithm 2: Distributed BiRS (dBiRS)

Input: the vector of marginal scores U; the bootstrap vectors $U^{e_1}, \ldots, U^{e_N}$; truncation parameter s; significant level α; number of blocks K;

Divide the genome into K blocks $B_1, \ldots, B_K$;

for $k = 1, \ldots, K$ **do**

  Run sBiRS$(U(B_k), U^{e_1}(B_k), \ldots, U^{e_N}(B_k), s)$ to derive the collection of detection results $\mathcal{R}_k$ in block B_k;

end

Let $\tilde{U} = (T(B_1), \ldots, T(B_K))^\top$ and $\tilde{U}^{e_b} = (M_b(B_1), \ldots, M_b(B_K))^\top$, $b = 1, \ldots, N$;

Run sBiRS$(\tilde{U}, \tilde{U}^{e_1}, \ldots, \tilde{U}^{e_N}, 0)$ to get the significant blocks $j_1, \ldots, j_{\hat{k}}$, $\hat{k} \leq K$;

Let $\hat{c}_{\min}(\alpha) = f_\alpha(\tilde{L}^{e_1}, \ldots, \tilde{L}^{e_N})$, where $\tilde{L}^{e_b} = (L_b(B_{j_1}), \ldots, L_b(B_{j_{\hat{k}}})^\top$;

for $r = j_1, j_2, \ldots j_{\hat{k}}$ **do**

  for $i = 1, \ldots, \ell_r$ **do**

    if $T_n(\hat{I}_i^{(r)}) > \hat{c}_{\min}(\alpha)$ **then**

      Take $\hat{I}_i^{(r)}$ as one of the estimated signal regions;

    end

  end

end

Rearrange these estimated signal regions to be continuous and separable;

Output: Estimated signal regions $\hat{I}_1, \ldots, \hat{I}_{\hat{\ell}}$.

algorithm can still achieve detection consistency, even when the signal strength of the signal regions decays at a certain rate. In contrast, the Q-SCAN method requires balanced signal strengths across regions. Moreover, the dBiRS algorithm relaxes the M-dependence assumption to a "weak" dependence assumption, which permits long-range correlation (LD) and is more suitable to genetic association studies. Additionally, with an appropriate block-splitting strategy, the dBiRS algorithm imposes fewer restrictions on the lengths of signal regions, which enables consistent detection of both shorter or longer signal regions compared to Q-SCAN. See Theorems 1-3 in the Supplementary Material (Wang, Zhang and Yao, 2025) for more details.

3. Simulation Study. We conduct simulation studies to compare the proposed dBiRS method with the state-of-art Q-SCAN and KnockoffScreen procedure. We generate sequence data of European ancestry from 10,000 chromosomes across 5-megabase (Mb) regions using the calibrated coalescent model (Sabeti and Schaffner, 2014), and the total number of variants is 349,640. We evaluate the family-wise error rate (FWER), false discovery rate (FDR), signal region detection rate (DR), and true positive rate (TPR) for both continuous and binary phenotypes.

For the evaluation of FWER, the continuous phenotypes are generated using the model:

$$Y = 0.5X_1 + 0.5X_2 + \varepsilon,$$

where X_1 is a continuous covariate sampled from a standard normal distribution, $X_1 \sim \mathcal{N}(0,1)$, and X_2 is a dichotomous covariate that takes values 0 and 1 with equal probability. The random noise ε is generated from a standard normal distribution, $\varepsilon \sim \mathcal{N}(0,1)$. The dichotomous phenotypes are generated using the following logistic regression model:

$$\text{logit}\left\{\mathbb{P}(Y=1)\right\} = 0.5X_1 + 0.5X_2,$$

where $\mathbb{P}(Y=1) = (\mathbb{P}(Y_1=1), \ldots, \mathbb{P}(Y_n=1))^\top$. We perform dBiRS and Q-SCAN analyses based on 1,000 Monte Carlo runs under the linear and logistic models to compute FWERs.

For both dBiRS and Q-SCAN, the number of bootstrap iterations is set to 1,000. We do not calculate the FWER for KnockoffScreen because it only controls the FDR (He et al., 2021a). The empirical FWERs estimated for dBiRS and Q-SCAN are presented in Table 1 for significance levels $\alpha = 0.01$ and $\alpha = 0.05$. The FWERs of Q-SCAN and dBiRS are accurate at both significance levels, demonstrating that both methods effectively control the FWER.

TABLE 1
FWERs of dBiRS and Q-SCAN in continuous and dichotomous phenotypes

	continuous		dichotomous	
α	0.01	0.05	0.01	0.05
dBiRS	0.011	0.053	0.011	0.051
Q-SCAN	0.011	0.051	0.009	0.048

We next evaluate the detection accuracy of dBiRS and compare its performance with Q-SCAN and KnockoffScreen. We randomly select four 5kb causal windows, denoted as I_i, $i = 1, \ldots, 4$. We generate continuous and dichotomous phenotypes by

$$Y = 0.5X_1 + 0.5X_2 + \boldsymbol{G}_{I_1}\beta_{I_1} + \cdots + \boldsymbol{G}_{I_4}\beta_{I_4} + \varepsilon,$$

and

$$\text{logit}\left\{\mathbb{P}(Y = 1)\right\} = 0.5X_1 + 0.5X_2 + \boldsymbol{G}_{I_1}\beta_{I_1} + \cdots + \boldsymbol{G}_{I_4}\beta_{I_4},$$

where $\boldsymbol{G}_{I_1}, \ldots, \boldsymbol{G}_{I_4}$ represent the genotypes of the causal windows, and $\beta_{I_1}, \ldots, \beta_{I_4}$ are the corresponding effect sizes. In each causal window, 10% of the variants are randomly designated as causal, and each causal variant is assigned an effect size as a decreasing function of the minor allele frequency (MAF), i.e., $|\beta| = c|\log_{10}(\text{MAF})|$. The parameter c is set to $c \in \{0.12, 0.15\}$ for continuous outcomes and $c \in \{0.25, 0.30\}$ for dichotomous outcomes. The sign of β is randomly assigned with 50% of the values being positive and 50% negative.

We evaluate the DR,TPR and h kilobase (kb) FDR for three methods based on 100 Monte Carlo runs across different values of c (Li, Liu and Lin, 2022; He et al., 2021a). Specifically, we denote the true signal regions as $I_1^*, \ldots, I_\ell^*$ and the estimated signal regions as $\hat{I}_1, \ldots, \hat{I}_{\hat{\ell}}$. Let $I^* = \cup_{i=1}^{\ell} I_i^*$ represent the union of the true signal regions, and $\hat{I} = \cup_{i=1}^{\hat{\ell}} \hat{I}_i$ represent the union of the estimated signal regions. The DR and TPR are defined as

$$\text{DR} = \frac{1}{\ell}\mathbb{E}\sum_{i=1}^{\ell} \mathbf{1}\left\{\hat{I} \cap I_\ell^* \neq \emptyset\right\}, \quad \text{TPR} = \mathbb{E}\left(\left|I^* \cap \hat{I}\right| / |I^*|\right).$$

Following He et al. (2021a), we define the FDR(h) for $h \in \{25, 50, 75\}$ as

$$\text{FDR}(h) = \mathbb{E}\left(\left|\left\{j : j \in \hat{I}, d(j, \hat{I}) \geq h\right\}\right| / \left|\hat{I}\right|\right)$$

to account for the spill-over effect of the LD structure, where $d(j, I^*) = \min_{i \in I^*} |i - j|$ is the minimum distance between point j and true signal regions. The DR represents the proportion of true signal regions that are detected and measures the ability to identify signal regions. In contrast, the TPR and FDR(h) evaluate how well the true signal regions are recovered by assessing the similarity or distance between the true and identified signal regions. The DR, TPR, and FDR(h) of the three methods for continuous and dichotomous phenotypes are presented in Table 2. Additionally, we provide the standard deviations (SDs) for DR, TPR, and FDR(h) in Table 3.

TABLE 2
Detection results (DR, TPR, FDR(h)) for continuous and dichotomous phenotypes, $h = 25, 50, 75$.

	DR	TPR	FDR(25)	FDR(50)	FDR(75)
			continuous, $c = 0.12$		
dBiRS	**0.990**	**0.830**	**0.192**	**0.049**	**0.023**
Q-SCAN	0.980	0.783	0.234	0.077	0.034
KnockoffScreen	0.980	0.227	0.475	0.379	0.287
			continuous, $c = 0.15$		
dBiRS	**1.000**	**0.881**	**0.285**	**0.094**	**0.037**
Q-SCAN	**1.000**	0.837	0.326	0.113	0.040
KnockoffScreen	**1.000**	0.505	0.605	0.516	0.408
			dichotomous, $c = 0.25$		
dBiRS	**0.990**	**0.793**	**0.183**	**0.049**	**0.020**
Q-SCAN	0.930	0.682	0.203	0.077	0.039
KnockoffScreen	0.960	0.184	0.495	0.416	0.237
			dichotomous, $c = 0.30$		
dBiRS	**1.000**	**0.853**	**0.235**	**0.058**	**0.019**
Q-SCAN	0.980	0.766	0.277	0.082	0.041
KnockoffScreen	0.990	0.361	0.533	0.440	0.294

TABLE 3
Standard deviations for DR, TPR, FDR(h) in continuous and dichotomous phenotypes, $h = 25, 50, 75$.

	sd(DR)	sd(TPR)	sd(FDR(25))	sd(FDR(50))	sd(FDR(75))
			continuous, $c = 0.12$		
dBiRS	**0.050**	**0.065**	**0.051**	**0.034**	**0.027**
Q-SCAN	0.069	0.105	0.083	0.056	0.052
KnockoffScreen	0.069	0.178	0.254	0.249	0.203
			continuous, $c = 0.15$		
dBiRS	**0.000**	**0.037**	**0.046**	**0.039**	**0.031**
Q-SCAN	**0.000**	0.072	0.073	0.050	0.042
KnockoffScreen	**0.000**	0.220	0.111	0.091	0.074
			dichotomous, $c = 0.25$		
dBiRS	**0.05**	**0.07**	**0.088**	**0.040**	**0.031**
Q-SCAN	0.115	0.126	0.096	0.066	0.050
KnockoffScreen	0.093	0.119	0.210	0.214	0.188
			dichotomous, $c = 0.30$		
dBiRS	**0.000**	**0.039**	**0.079**	**0.037**	**0.024**
Q-SCAN	0.069	0.102	0.107	0.068	0.044
KnockoffScreen	0.050	0.168	0.131	0.121	0.123

Table 2 shows that dBiRS achieves the highest DRs and TPRs with the lowest FDRs across all signal strengths c and for all values of the distance parameter $h = 25, 50, 75$ in both continuous and dichotomous phenotypes. As the value of h increases, the FDR of all methods decreases, and the FDR of dBiRS and Q-SCAN falls below 0.05 when $h = 75$ kb. Q-SCAN outperforms KnockoffScreen in recovering true signal regions by achieving higher TPR and lower FDR. However, it identifies a smaller proportion of true signal regions (i.e., a lower DR) compared to KnockoffScreen. KnockoffScreen employs LD-pruning with a 0.75 correlation threshold to remove highly correlated variants before conducting the analysis. It identifies a larger number of signal regions, which increases the chance of overlap with true signal regions and thereby boosting the DR. However, the signal regions identified by KnockoffScreen have lower coverage of true signal regions due to both the LD-pruning process and a higher number of false discoveries. Additionally, Table 3 shows that dBiRS has the lowest standard deviations in all cases, indicating that dBiRS is the most stable method.

To further illustrate the recovery of true signal regions across different methods, we present the selection probabilities for all variants, where the selection probability of a variant j is defined as the proportion of replications in which it is identified as a signal variant. We present the selection probabilities under the continuous phenotype and the dichotomous phenotype in Figure 1 and Figure 2, respectively.

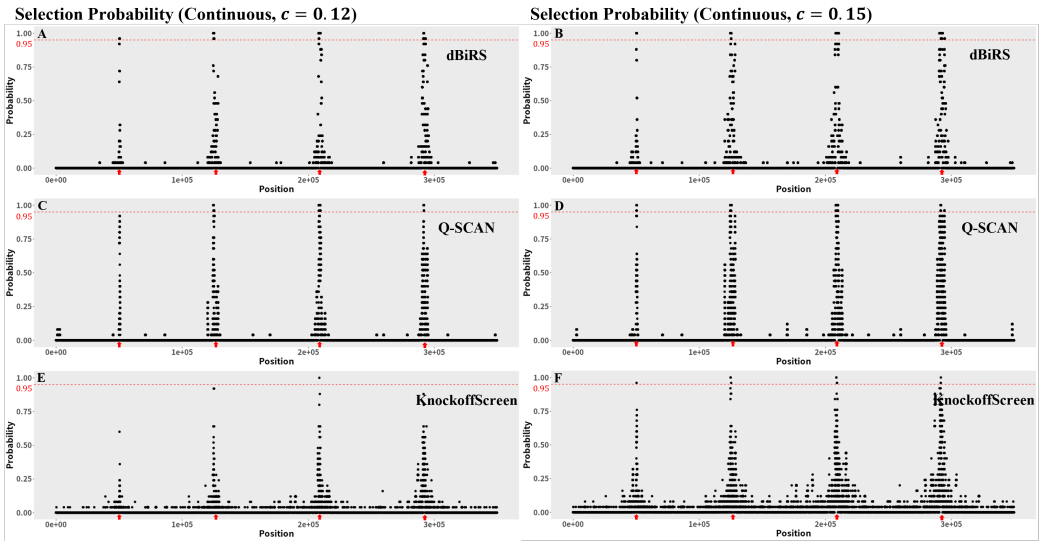

Fig 1: The selection probabilities of dBiRS, Q-SCAN and KnockoffScreen under dichotomous phenotype. The left panel is the selection probabilities when $c = 0.25$ and the right panel is the selection probabilities when $c = 0.30$. The red arrows refer to the locations of four true signal regions.

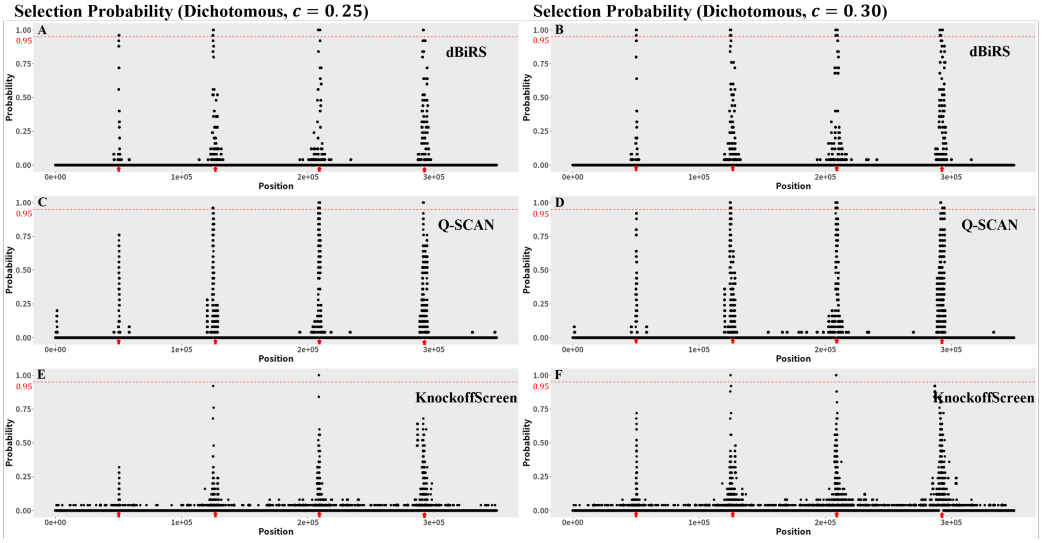

Fig 2: The selection probabilities of dBiRS, Q-SCAN and KnockoffScreen under dichotomous phenotype. The left panel is the selection probabilities when $c = 0.25$ and the right panel is the selection probabilities when $c = 0.30$. The red arrows refer to the locations of four true signal regions.

While all procedures exhibit relatively high selection probabilities around the four true signal regions, dBiRS demonstrates the greatest accuracy and stability, achieving selection probabilities exceeding 0.95 in all true signal regions across all settings. When signals are relatively weak (i.e., $c = 0.12$ for the continuous trait in Figure 1A and $c = 0.25$ for the binary trait in Figure 2A), Q-SCAN shows relatively low selection probabilities in the leftmost true signal region (i.e., selection probability < 0.95) and identifies some variants in non-signal regions at the beginning of the genome. This suggests that the Q-SCAN procedure is more prone to producing misleading results across replications, which is consistent with results that

show a higher FDR compared to dBiRS. KnockoffScreen demonstrates the lowest selection probabilities in the four true signal regions and identifies only one variant with a selection probability greater than 0.95 (Figure 1A and Figure 2A). When the signal is stronger (i.e., $c = 0.15$ for the continuous trait in Figure 1B and $c = 0.30$ for the binary trait in Figure 2B), KnockoffScreen exhibits substantially higher selection probabilities in non-signal regions compared to both Q-SCAN and dBiRS (Figure 1B). The spread of selection probabilities across non-signal regions highlights a lack of stability for KnockoffScreen and indicates its tendency to identify false positives for spurious variants. Based on these results, we conclude that the dBiRS method outperforms the other procedures in terms of both detection accuracy and robustness across different settings.

4. Application. Extensive research has identified common genetic variants associated with cognitive health, but the roles of rare genetic variants, which often have more subtle and complex effects, remain poorly understood. In this study, we apply the dBiRS algorithm to conduct Whole Exome Sequencing (WES) analyses on the core cognitive phenotypes: fluid intelligence (Field ID 20016) and prospective memory (Field ID 20018), using data from the UK Biobank (https://biobank.ctsu.ox.ac.uk/crystal/index.cgi). The WES data used in this study is derived from the final exome release in PLINK format (Field ID 23158) from the UK Biobank. Fluid intelligence reflects problem-solving and reasoning abilities, while prospective memory measures the capacity to remember and execute planned actions, a critical daily-life function that often declines with age. These phenotypes are closely linked to aging and neurodegeneration. By analyzing these traits, we aim to uncover rare variants that may confer either risk or resilience to cognitive functions, providing insights into the genetic and biological mechanisms underlying cognitive impairment. This approach has the potential to deepen our understanding of cognitive aging and uncover genetic factors contributing to the risk of neurodegenerative disorders such as Alzheimer's disease (AD).

We implemented a standard quality control procedure to ensure the integrity of the dataset before analysis. First, we excluded individuals flagged as outliers by the UK Biobank based on genotyping missingness rates or heterogeneity, as well as those whose genotypically inferred sex did not align with their self-reported sex. To address population stratification, we utilized principal component analysis provided by the UK Biobank and excluded individuals identified as non-European. Specifically, we removed individuals whose values for either of the first two principal components deviated by more than five standard deviations from the mean. We also excluded participants who self-reported an ethnicity other than European. Furthermore, individuals with more than 5% missing genotype data across variants passing UK Biobank's quality control were removed. For variant-level quality control, we retained only biallelic autosomal variants assayed by both genotyping arrays used by the UK Biobank. Variants that failed UK Biobank quality control in any genotyping batch were excluded. Additionally, we removed variants with a Hardy-Weinberg equilibrium (HWE) p-value below 10^{-50} or with a minor allele count (MAC) ≤ 1. This stringent filtering resulted in 13,681,006 variants across 22 chromosomes in the WES dataset, 154,785 samples for fluid intelligence analysis, and 155,448 samples for prospective memory analysis. After the quality control procedure, the dBiRS analysis was performed using summary statistics derived from a generalized linear model. The model adjusted for covariates, including sex, age, assessment center, and the top five genomic principal components provided by the UK Biobank.

Figure 3 illustrates the distribution of functional consequences of variants identified by dBiRS, with a substantial number of variants located in exonic, intronic, and UTR regions. Variants in the exonic region may directly alter protein structure and function. Variants in the 3' UTR (UTR3) are likely involved in post-transcriptional regulation, including mRNA

stability and microRNA binding. Intronic variants could disrupt splicing mechanisms or long-range regulatory elements, further impacting gene function. While exonic variants are commonly the focus of WES studies, our findings also underscore the important role of rare variants in non-coding and regulatory regions, such as UTRs and intronic regions, which have historically been understudied in the context of intelligence and prospective memory.

For the study on intelligence, dBiRS identified signal regions encompassing 327 genes, including 84 genome-wide significant genes previously reported for their association with intelligence (Sollis et al., 2023). Additionally, there are 11 genome-wide significant genes overlapping with those associated with general cognitive function, 30 genes overlapping with those implicated in schizophrenia, 65 genes associated with educational attainment, and 35 genes overlapping with autism spectrum disorder (Sollis et al., 2023). Based on the GWAS Catalog, 33 genome-wide significant common variants were also identified in our WES study (Sollis et al., 2023). Notable examples include SNP rs2271928 in COL16A1, SNP rs11264680 in CRTC2, and SNP rs539096 in PTPRF, among others.

The dBiRS method identifies numerous additional rare variant associations in the exonic, intronic, and 3' UTR regions of genes that are previously reported to be associated with intelligence or cognitive performance. For example, a previous study highlights the significance of CRTC2 in endothelial function, which is essential for maintaining vascular physiology in the brain (Kanki et al., 2020). While prior genome-wide studies have identified common variants in CRTC2 associated with cognitive traits (Savage et al., 2018), the dBiRS method uncovers 4 rare variants in the exonic region of CRTC2, which may directly alter protein function, 3 rare variants in the intronic region that could influence gene regulation, and 4 rare variants in the 3' UTR region, potentially affecting mRNA stability and translation efficiency. Another example is BRWD1 that is previously identified as a significant gene associated with intelligence (Savage et al., 2018). In our study, we identified rare variants in its exonic, intronic, and 3' UTR regions. Research has shown that BRWD1 plays a role in establishing epigenetic states during B cell development, which is essential for proper immune function (Mandal et al., 2018). Further studies are needed to clarify its specific contributions to neural development and cognition.

The dBiRS method identified rare variants in 22 novel genes that have not been previously reported in the GWAS Catalog for intelligence. Among these 22 genes, 14 were significant in the gene-based test for general cognitive ability (Davies et al., 2018). Of the remaining 9 genes, SZT2 has been linked to common variants associated with Alzheimer's disease and specific brain regions (Herold et al., 2016); ITIH4 has been associated with schizophrenia (Ripke et al., 2011); and NDUFA6 has been linked to brain connectivity measurements (Zhang et al., 2023). Genes SLC39A8, TOP2B, NKIRAS1, GLTBD1, and TPM3 are new findings that have not been previously reported in relation to cognitive performance measurements. Among these genes, TOP2B plays a critical role in neuronal development by regulating gene expression during brain development (Tiwari et al., 2012). Proper functioning of TOP2B is essential for neuronal connectivity and synaptic plasticity, both of which underlie learning and memory (Madabhushi et al., 2015). Disruptions in TOP2B have been linked to neurodevelopmental disorders, which could implicate it in intelligence (King et al., 2013).

For the study on prospective memory, we identified eight novel genes associated with prospective memory, a key component of cognitive function. Among these, OMA1 stands out due to its established links with multiple neuroimaging measurements, including white matter lesion progression and brain column structure, underscoring its potential role in cognitive processes (Wang et al., 2020). Notably, our analysis uncovered 10 rare variants in the 3' UTR, 6 rare variants in the exonic region, 7 rare variants in the intronic region, and 1 common variant in the 5' UTR of OMA1. These findings suggest that OMA1 may influence prospective memory through diverse mechanisms, such as protein function modulation,

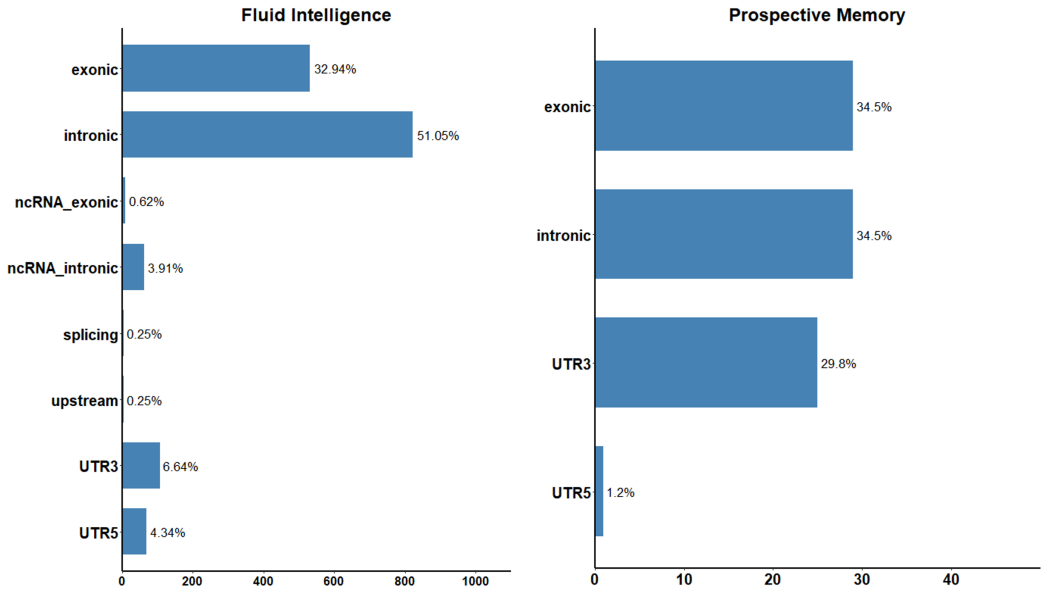

Fig 3: Distribution of the functional consequences of SNPs in signal regions identified by dBiRS.

gene regulation, and mRNA stability. Similarly, we identified variants in the intronic, exonic, and 3' UTR regions of CNGB3, further emphasizing the importance of regulatory regions in prospective memory. CNGB3 is known to be associated with educational attainment (Okbay et al., 2022), cognitive function (Lee et al., 2018) and anxiety (Meier et al., 2019), which supports its potential role in cognitive traits. Additionally, we discovered HIGD1B, previously reported to be associated with educational attainment (Pasman et al., 2022), and SFMBT2 which has been linked to anxiety and stress-related disorders (Hollins and Cairns, 2016). Other novel findings include the genes NT5C3B, NOSTRIN, and APOD, which represent unexplored candidates for prospective memory and related cognitive traits. These discoveries underscore the value of our approach in uncovering novel genetic mechanisms underlying complex cognitive traits and highlight new directions for future research into the genetic architecture of prospective memory.

5. Discussion. We developed a computationally efficient distributed Binary and ReSearch (dBiRS) algorithm by incorporating the infinity norm of regression-based summary statistics to support the analysis of both binary and continuous outcomes for whole-genome and whole-exome sequencing studies. Compared to scan-based algorithms and the knockoffScreen method, dBiRS demonstrates greater power while maintaining strict control over family-wise error rates (FWER) and false discovery rates (FDR). Furthermore, dBiRS enables parallel computing of block-wise results, which are then aggregated through a central machine to ensure both detection accuracy and computational efficiency. Empirical studies further demonstrate its robustness and adaptability, even under conditions of signal decay.

Effectively controlling sample relatedness is a critical step in genetic association studies. Although our dBiRS method is based on the generalized linear model (GLM), which assumes independent samples, it can be readily extended to account for sample relatedness by fitting a null generalized linear mixed model (GLMM) with a genetic-relatedness matrix (GRM) to derive marginal and pseudo scores. Furthermore, the dBiRS framework is flexible and can be adapted to various study designs. It can be extended to accommodate different types of outcomes by calculating summary statistics based on different models, including ordinal

categorical traits (via the proportional odds model), survival data (via the Cox model), and longitudinal data (via mixed-effects models), among others.

Further extensions of dBiRS could explore the integration of functional annotations. While the current framework effectively detects signal regions without annotations, their incorporation may provide additional biological context and improve detection accuracy for rare or weak signals in less characterized regions of the genome. Our framework is also flexible to incorporate various functional annotations by directly adding weights into SNP-level summary statistics, where weights can be estimated using annotation principal components (Zhou et al., 2023), CADD (Kircher et al., 2014) or other methods. Additionally, further validation of the identified associations through experimental studies is necessary to elucidate their functional relevance and causal mechanisms.

Acknowledgments. Fan Wang and Wei Zhang are the co-first authors and contribute equally to this work. Fang Yao is the corresponding author, E-mail: fyao@math.pku.edu.cn. The authors would like to thank the Editor, the AE and reviewers for their constructive comments.

Funding. This research is supported in part by the National Key R&D Program of China (No. 2022YFA1003800), the National Natural Science Foundation of China (No. 12292981, 12288101), the New Cornerstone Science Foundation through Xplorer Prize, the LMAM and the Fundamental Research Funds for the Central Universities, Peking University. This research has been conducted using the UK Biobank Resource under project 79237.

SUPPLEMENTARY MATERIAL

Supplementary Material for "Computationally Efficient Whole-Genome Signal Region Detection for Quantitative and Binary Traits"

In this Supplementary Material, we present theoretical properties, the key lemmas and the proofs of the lemmas and theorems in our work. In addition, all codes and instructions for implementing simulations and real data analysis are included in the file "dBiRS_Code".

REFERENCES

BAKSHI, A., ZHU, Z., VINKHUYZEN, A. A. et al. (2016). Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Scientific reports* **6** 1–9.

CAI, T. T. and WEI, H. (2022). Distributed adaptive Gaussian mean estimation with unknown variance: Interactive protocol helps adaptation. *The Annals of Statistics* **50** 1992 – 2020. https://doi.org/10.1214/21-AOS2167

DAVIES, G., LAM, M., HARRIS, S. E., TRAMPUSH, J. W., LUCIANO, M., HILL, W. D., HAGENAARS, S. P., RITCHIE, S. J., MARIONI, R. E., FAWNS-RITCHIE, C. et al. (2018). Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nature communications* **9** 2098.

HE, Z., LIU, L., WANG, C. et al. (2021a). Identification of putative causal loci in whole-genome sequencing data via knockoff statistics. *Nature communications* **12** 3152.

HE, Z., LE GUEN, Y., LIU, L. et al. (2021b). Genome-wide analysis of common and rare variants via multiple knockoffs at biobank scale, with an application to Alzheimer disease genetics. *The American Journal of Human Genetics* **108** 2336–2353.

HEROLD, C., HOOLI, B. V., MULLIN, K., LIU, T., ROEHR, J. T., MATTHEISEN, M., PARRADO, A. R., BERTRAM, L., LANGE, C. and TANZI, R. E. (2016). Family-based association analyses of imputed genotypes reveal genome-wide significant association of Alzheimer's disease with OSBPL6, PTPRG, and PDCL3. *Molecular psychiatry* **21** 1608–1612.

HOLLINS, S. L. and CAIRNS, M. J. (2016). MicroRNA: Small RNA mediators of the brains genomic response to environmental stress. *Progress in neurobiology* **143** 61–81.

JENG, X. J., CAI, T. T. and LI, H. (2010). Optimal Sparse Segment Identification With Application in Copy Number Variation Analysis. *Journal of the American Statistical Association* **105** 1156-1166. PMID: 23543902. https://doi.org/10.1198/jasa.2010.tm10083

KANKI, H., SASAKI, T., MATSUMURA, S., KAWANO, T., TODO, K., OKAZAKI, S., NISHIYAMA, K., TAKEMORI, H. and MOCHIZUKI, H. (2020). CREB coactivator CRTC2 plays a crucial role in endothelial function. *Journal of Neuroscience* **40** 9533–9546.

KING, I. F., YANDAVA, C. N., MABB, A. M., HSIAO, J. S., HUANG, H.-S., PEARSON, B. L., CALABRESE, J. M., STARMER, J., PARKER, J. S., MAGNUSON, T. et al. (2013). Topoisomerases facilitate transcription of long genes linked to autism. *Nature* **501** 58–62.

KIRCHER, M., WITTEN, D. M., JAIN, P., O'ROAK, B. J., COOPER, G. M. and SHENDURE, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46** 310–315.

LEE, J. J., WEDOW, R., OKBAY, A., KONG, E., MAGHZIAN, O., ZACHER, M., NGUYEN-VIET, T. A., BOWERS, P., SIDORENKO, J., KARLSSON LINNÉR, R. et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics* **50** 1112–1121.

LI, Z., LIU, Y. and LIN, X. (2020). Simultaneous Detection of Signal Regions Using Quadratic Scan Statistics With Applications to Whole Genome Association Studies. *Journal of the American Statistical Association* **0** 1-12. https://doi.org/10.1080/01621459.2020.1822849

LI, Z., LIU, Y. and LIN, X. (2022). Simultaneous detection of signal regions using quadratic scan statistics with applications to whole genome association studies. *Journal of the American Statistical Association* **117** 823–834.

LI, Z., LI, X., LIU, Y., SHEN, J., CHEN, H., ZHOU, H., MORRISON, A. C., BOERWINKLE, E. and LIN, X. (2019). Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. *The American Journal of Human Genetics* **104** 802–814.

LI, X., LI, Z., ZHOU, H. et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature genetics* **52** 969–983.

LI, Z., LI, X., ZHOU, H. et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods* **19** 1599-1611.

LIU, J. Z., MCRAE, A. F., NYHOLT, D. R. et al. (2010). A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics* **87** 139–145.

MADABHUSHI, R., GAO, F., PFENNING, A. R., PAN, L., YAMAKAWA, S., SEO, J., RUEDA, R., PHAN, T. X., YAMAKAWA, H., PAO, P.-C. et al. (2015). Activity-induced DNA breaks govern the expression of neuronal early-response genes. *Cell* **161** 1592–1605.

MADSEN, B. E. and BROWNING, S. R. (2009). A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLOS Genetics* **5** 1-11.

MANDAL, M., MAIENSCHEIN-CLINE, M., MAFFUCCI, P., VESELITS, M., KENNEDY, D. E., MCLEAN, K. C., OKOREEH, M. K., KARKI, S., CUNNINGHAM-RUNDLES, C. and CLARK, M. R. (2018). BRWD1 orchestrates epigenetic landscape of late B lymphopoiesis. *Nature communications* **9** 3888.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A. et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.

MEIER, S. M., TRONTTI, K., PURVES, K. L., ALS, T. D., GROVE, J., LAINE, M., PEDERSEN, M. G., BYBJERG-GRAUHOLM, J., BÆKVED-HANSEN, M., SOKOLOWSKA, E. et al. (2019). Genetic variants associated with anxiety and stress-related disorders: a genome-wide association study and mouse-model study. *JAMA psychiatry* **76** 924–932.

MORGENTHALER, S. and THILLY, W. G. (2007). 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** 28-56.

NAUS, J. I. (1982). Approximations for Distributions of Scan Statistics. *Journal of the American Statistical Association* **77** 177-183. https://doi.org/10.1080/01621459.1982.10477783

OKBAY, A., WU, Y., WANG, N., JAYASHANKAR, H., BENNETT, M., NEHZATI, S. M., SIDORENKO, J., KWEON, H., GOLDMAN, G., GJORGJIEVA, T. et al. (2022). Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature genetics* **54** 437–449.

OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557-572. https://doi.org/10.1093/biostatistics/kxh008

PASMAN, J. A., DEMANGE, P. A., GULOKSUZ, S., WILLEMSEN, A., ABDELLAOUI, A., TEN HAVE, M., HOTTENGA, J.-J., BOOMSMA, D. I., DE GEUS, E., BARTELS, M. et al. (2022). Genetic risk for smoking: disentangling interplay between genes and socioeconomic status. *Behavior genetics* **52** 92–107.

RIPKE, S., SANDERS, A., KENDLER, K., LEVINSON, D., SKLAR, P., HOLMANS, P., LIN, D., DUAN, J., OPHOFF, R., ANDREASSEN, O. et al. (2011). Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet* **43** 969–976.

SABETI, I. S. P. C. and SCHAFFNER, S. F. (2014). Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics* **30** 3427–3429.

SAVAGE, J. E., JANSEN, P. R., STRINGER, S., WATANABE, K., BRYOIS, J., DE LEEUW, C. A., NAGEL, M., AWASTHI, S., BARR, P. B., COLEMAN, J. R. et al. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature genetics* **50** 912–919.

SOLLIS, E., MOSAKU, A., ABID, A., BUNIELLO, A., CEREZO, M., GIL, L., GROZA, T., GÜNEŞ, O., HALL, P., HAYHURST, J. et al. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic acids research* **51** D977–D985.

TIWARI, V. K., BURGER, L., NIKOLETOPOULOU, V., DEOGRACIAS, R., THAKURELA, S., WIRBELAUER, C., KAUT, J., TERRANOVA, R., HOERNER, L., MIELKE, C. et al. (2012). Target genes of Topoisomerase IIβ regulate neuronal survival and are defined by their chromatin state. *Proceedings of the National Academy of Sciences* **109** E934–E943.

VISSCHER, P. M., WRAY, N. R., ZHANG, Q. et al. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101** 5-22.

WANG, F., ZHANG, W. and YAO, F. (2025). Supplementary Material for "Computationally Efficient Whole-Genome Signal Region Detection for Quantitative and Binary Traits". https://doi.org/10.1214/xxxx

WANG, H., YANG, J., SCHNEIDER, J. A., DE JAGER, P. L., BENNETT, D. A. and ZHANG, H.-Y. (2020). Genome-wide interaction analysis of pathological hallmarks in Alzheimer's disease. *Neurobiology of aging* **93** 61–68.

WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-Variant Association Testing for Sequencing Data With the Sequence Kernel Association Test. *The American Journal of Human Genetics* **89** 82-93.

XUE, K. and YAO, F. (2020). Distribution and correlation-free two-sample test of high-dimensional means. *The Annals of Statistics* **48** 1304 – 1328. https://doi.org/10.1214/19-AOS1848

ZHANG, W., WANG, F. and YAO, F. (2025). Fast Signal Region Detection with Application to Whole Genome Association Studies. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2025.2464271

ZHANG, N. R., SIEGMUND, D. O., JI, H. and LI, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97** 631–645.

ZHANG, L., PAN, Y., HUANG, G., LIANG, Z., LI, L., ZHANG, M. and ZHANG, Z. (2023). A brain-wide genome-wide association study of candidate quantitative trait loci associated with structural and functional phenotypes of pain sensitivity. *Cerebral Cortex* **33** 7297–7309.

ZHOU, H., ARAPOGLOU, T., LI, X. et al. (2023). FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Research* **51** D1300–D1311.