Computational Statistics and Data Analysis 219 (2026) 108343

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Efficient genome-wide association studies via low-rank approximatio

Zhongyuan Chen[a,*], Han Liang[b], Peng Wei[c]

[a] School of Mathematics, Nanjing University, 22 Hankou Road, Nanjing, Jiangsu, 210093, China
[b] Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center, 1400 Pressler St, Houston, TX, 77030, USA
[c] Department of Biostatistics, MD Anderson Cancer Center, 1400 Pressler St, Houston, TX, 77030, USA

ARTICLE INFO

Keywords:
Genome-wide association studies
aSPU tests
Low-rank matrix approximations
Computational cost
Parameter selection
Representative SNPs.

ABSTRACT

In genome-wide association studies (GWASs), gene-based large-scale score tests such as the adaptive sum of powered score (aSPU) test have some attractive features (such as nice data adaptivity and effective information aggregation) that lead to high statistical powers. However, these test methods are computationally expensive due to the need of a large number of matrix-vector multiplications. To address such a limitation, a series of effective and efficient strategies for association studies is proposed based on low-rank approximations. A low-rank approximation to a SNP matrix, constructed using fast randomized SVDs, yields an aSPU-LR test method that significantly reduces the cost of aSPU tests while ensuring the reliability. Furthermore, leveraging the low-rank approximation, a procedure is developed to quickly select certain effective parameters in the aSPU tests that give valuable insights into the association patterns. Additionally, a fast pivoting approach is designed using the low-rank approximation so as to quickly identify representative SNPs which help capture major genetic information of the data. The efficiency and the effectiveness of the proposed strategies are demonstrated through extensive simulations and real data studies. For a large-scale International Cancer Genome Consortium (ICGC) dataset, the proposed aSPU-LR test rapidly identifies associations between germline variations and somatic mutations across multiple cancer types, offering a significant speed advantage over the original aSPU test while maintaining similar accuracy. The low-rank approximations are further used to quickly extract useful SNP information about the germline variations and identify the patterns of their associations with the somatic mutations.

1. Introduction

Genome-wide association studies (GWASs) provide powerful tools for uncovering associations between genetic variants (such as single nucleotide polymorphisms or SNPs) and specific traits or diseases. For example, GWASs have recently been used to identify interactions between inherited germline variations (SNPs) and somatic mutations (traits acquired during life), which help to understand tumor progression, predict cancer risks, and develop personalized cancer therapy (Barfield et al., 2022; Carter et al., 2017; Mamidi et al., 2019; Ramroop et al., 2019; Vali-Pour et al., 2022; Vosoughi et al., 2020). Such studies become practical with the availability of whole-exome and whole-genome sequencing data through consortia such as the International Cancer Genome Consortium (ICGC) (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020) and The Cancer Genome Atlas (TCGA). On the other hand, traditional GWASs often rely on low-power statistical methods (such as Fisher's exact test) which perform SNP-by-SNP analyses and fail to fully leverage the complexity of genetic/genomic data.

* Corresponding author.
E-mail address: zychenstat@nju.edu.cn (Z. Chen).

https://doi.org/10.1016/j.csda.2026.108343
Received 3 May 2025; Received in revised form 16 January 2026; Accepted 16 January 2026
Available online 18 January 2026

In recent years, large-sample score test methods such as the data-adaptive sum of powered score (aSPU) test (Pan et al., 2014) were proposed. The aSPU method aggregates SNPs into genes and enhances the statistical power by combining a family of sum of powered score (SPU) tests, utilizing the benefits of some existing tests such as the burden test, the sum of squared score (SSU) test (Pan, 2009), and the univariate minP (UminP) test. It is particularly effective in scenarios where the number of nonassociated variants is large (Deng et al., 2022). The aSPU method was recently successfully adopted to study interactions between germline variations and somatic mutations in human cancers (Chen et al., 2023).

Despite its advantages, the aSPU test is computationally intensive. As noted by Deng et al. (2022), the aSPU test is time consuming since it requires heavy computations to estimate small p-values. It was also indicated by Chen et al. (2019) that aSPU is subject to much greater computational loads than either the burden test or SKAT (sequence kernel association test) (Wu et al., 2011) in the generalized linear mixed models framework due to large numbers of permutations for calculating p-values with large-scale datasets.

In fact, the main reason for the high computational cost of aSPU is as follows. It needs to calculate score vectors which involve multiplications of a full SNP matrix X with vectors (see (8) below). Many such multiplications are used since the accurate computation of p-values needs a large number of permutations (in the orders of 10^3, 10^4, or even up to 10^{11}) (Deng et al., 2022), each requiring a score vector (see (4) below). This leads to high computational burden, especially for large-scale datasets like the ICGC. The high computational costs potentially limit the practical usage of the aSPU test and other similar tests that involve many score vector calculations and permutations. Therefore, it is desirable to improve the efficiency of the methods while maintaining the effectiveness.

In this paper, we propose a series of efficient strategies to perform association studies and to further extract other valuable information. We significantly reduce the cost of aSPU by accelerating the matrix-vector multiplications through low-rank approximations. The low-rank approximations also make it feasible to quickly extract other information.

1. We construct low-rank approximations to SNP matrices via randomized singular value decompositions (SVDs) so as to quickly perform aSPU tests while ensuring the reliability.
2. Based on the low-rank approximations, we design a fast procedure to identify certain parameters in aSPU that give useful insights into the association patterns.
3. We also design a fast pivoting strategy based on the low-rank approximations so as to quickly identify representative SNPs from the full SNP matrix, which is useful to capture essential genetic information with little extra cost.

The details are as follows.

First, we have noticed that, in practical data analysis, SNP matrices X can often be approximated by low-rank forms with reasonable accuracy. This is also consistent with recent studies on the compressibility of large data matrices (Udell and Townsend, 2019). We utilize randomized SVDs to compress the original SNP matrix X into a low-rank form $\tilde{X} = QR$ with modest rank and reasonable accuracy. The low-rank form QR enables to quickly perform matrix-vector multiplications, substantially reducing the computational cost of aSPU. This results in a test method, denoted as aSPU-LR. Compared to the vast number of full matrix-vector multiplications (due to permutations) in aSPU, it only needs a small number (roughly proportional to the approximation rank r) of full matrix-vector multiplications in randomized SVDs to get QR. All the remaining multiplications are in low-rank forms and have very small costs. An appropriately chosen rank r guarantees reasonable approximation accuracy of QR so that aSPU-LR can maintain statistical powers and type I error rates similar to those of aSPU.

Next, low-rank approximations further provide an opportunity to quickly reveal potential association patterns. The data adaptivity of the aSPU test, introduced by Pan et al. (2014), specifically relies on a range of different γ parameters in the score tests. Different parameters are used to accommodate different possible association patterns in the data. To roughly decide such effective parameters, we keep track of the calculations of certain minimum p-values in aSPU-LR through a set of data subsets simulated from the problems under consideration. It can be seen that even with very small ranks in low-rank approximations, this strategy can identify effective γ parameters which are very close to those γ parameters directly identified based on aSPU. The effective parameters may also be used to further accelerate association studies.

Additionally, low-rank approximations can also be used to extract valuable genetic information (representative SNPs) from the data matrix X. We design a fast procedure to transform $\tilde{X}$ into a low-rank approximation $\hat{X} = \hat{Q}\hat{R}$ to the centered SNP matrix. $\hat{R}$ approximately preserves the variance information of the SNPs and can further be used to quickly perform column pivoting to get a set of pivot indices. Such pivot indices can be used to identify a set of representative SNPs. Thus, this method not only functions as a dimension reduction strategy like the principal component analysis (PCA), but also provides potential biomarker information.

We show the efficiency and the effectiveness of the strategies through extensive simulation studies and large-scale ICGC case studies. Our results demonstrate that aSPU-LR has significantly lower computation time than aSPU while achieving similar accuracy in the tests. Even for modest problem sizes like in Table 6, we have observed a speedup of over 20 times. Using our proposed parameter selection procedure with highly compressed low-rank approximations, we also show that effective parameters can be reliably selected and can be used to infer certain association patterns. With the selected parameters, the accuracy of aSPU is similar to that of aSPU with a full set of parameters. Furthermore, our fast pivoting approach effectively selects important SNPs from the SNP matrices in both simulations and the ICGC dataset.

This work not only accelerates aSPU tests but also paves the path of using low-rank approximations to perform efficient association studies. It can benefit many test methods that involve expensive matrix operations. For instance, it not only benefits gene-based tests but can also be used to accelerate methods like in Pan et al. (2015) for pathways-based association tests. Additionally, our techniques

are not limited to GWASs, as the low-rank approximations potentially work for many different types of data. Thus, we expect them to be much more widely applicable.

The remainder of the article is organized as follows. In Section 2, we present the aSPU-LR test following a brief review of the aSPU test. Section 3 gives a fast procedure to select effective parameters via aSPU-LR. In Section 4, we give a fast pivoting approach to select representative SNPs. Sections 5 and 6 respectively give simulation results and ICGC real data studies. Section 7 includes some discussions and concludes this paper.

2. Fast association tests via low-rank approximations

In this section, we first briefly review the aSPU test (Pan et al., 2014), introduce some notation, and point out the efficiency limitation of aSPU. Then, we adopt low-rank approximations to compress SNP matrices and accelerate aSPU while maintaining the accuracy of the test, which leads to a test method aSPU-LR. Our discussions are mainly in terms of some data from GWASs, but the techniques are more widely applicable, provided that relevant data matrices can be approximated by low-rank forms.

2.1. Review: aSPU test

The aSPU test was originally designed for the case-control study in GWASs to investigate associations between gene-based SNPs and a binary trait such as an indicator of a disease (Pan et al., 2014). The test was extended by Chen et al. (2023) to study associations between germline variations and somatic mutations in human cancers.

For n subjects, let a SNP matrix X and a binary outcome vector y respectively be

$$X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T \quad \text{and} \quad y = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^T, \tag{1}$$

where

- X is an $n \times p$ matrix with n the number of subjects and p the number of SNPs in a gene,
- $X_i^T = \begin{bmatrix} X_{i1} & \cdots & X_{ip} \end{bmatrix}$ is the i-th row of X with X_{ij} the genotype score 0, 1, or 2 (number of the minor allele) at SNP j for subject i, and
- $y_i = 0$ or 1, indicating the binary outcome/response such as the somatic mutation of a gene.

Let $L(\boldsymbol{\beta}; X)$ be the likelihood function with $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \cdots & \beta_p \end{bmatrix}^T$. The score vector $u = \begin{bmatrix} u_1 & \cdots & u_p \end{bmatrix}^T$ associated with L is given by $u_i(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_i} \log L(\boldsymbol{\beta}; X)$. Based on the logistic regression model

$$\text{logit}[\Pr(y_i = 1)] = \beta_0 + X_i^T \boldsymbol{\beta}, \tag{2}$$

the null hypothesis of the test is $H_0 : \boldsymbol{\beta} = 0$, meaning there is no association between any SNP and the binary response. In standard score tests, the score vector looks like

$$u = \sum_{i=1}^{n} X_i (y_i - \bar{y}), \tag{3}$$

where $\bar{y}$ is the sample mean of all y_i's.

In the aSPU association test, the SPU statistic $T_{SPU}(\gamma) = \sum_{j=1}^{p} (u_j)^\gamma$ is first calculated, where γ is a parameter from a candidate set S of integers (including infinity). Then, the p-value is estimated using resampling methods such as permutations (Churchill and Doerge, 1994). Specifically, y is permuted to obtain a set of trait vectors $y^{(k)}$, $k = 1, 2, \ldots, N$ and the corresponding score vectors $u^{(k)}$ and SPU statistics $T_{SPU}^{(k)}(\gamma)$ are computed:

$$u^{(k)} = \sum_{i=1}^{n} X_i (y_i^{(k)} - \bar{y}) \quad \text{and} \quad T_{SPU}^{(k)}(\gamma) = \sum_{j=1}^{p} (u_j^{(k)})^\gamma. \tag{4}$$

Then, the p-value is approximately computed as

$$P_{SPU}(\gamma) = \frac{1}{N+1} \left(1 + \sum_{k=1}^{N} \mathbb{1}\left(|T_{SPU}^{(k)}(\gamma)| \geq |T_{SPU}(\gamma)| \right) \right), \tag{5}$$

where $\mathbb{1}(\cdot)$ is the indicator function.

By including different γ values in S, the aSPU test accommodates diverse association patterns, such as same-direction associations, opposite-direction associations, and single-SNP-dominated associations. This flexibility makes it especially useful when the underlying association patterns are unknown. Thus, it enhances the statistical power by combining the benefits of test methods such as the sum test and the adaptive SSU test (Pan and Shen, 2011).

Following the minimum p method (Tippett, 1931), we take the minimum p-value among the $P_{SPU}(\gamma)$ values for all $\gamma \in S$ as the new test statistic:

$$T_{aSPU} = \min_{\gamma \in S} P_{SPU}(\gamma). \tag{6}$$

T_{aSPU} is not a genuine p-value any more. Its p-value is estimated as

$$P^{(k)}(\gamma) = \frac{1}{N}\left(1 + \sum_{j\neq k} \mathbb{I}(T^{(j)}_{SPU}(\gamma) \geq T^{(k)}_{SPU}(\gamma))\right).$$

Note that it is not necessary to use double permutations to compute $P^{(k)}(\gamma)$ (Pan et al., 2014). Instead, the values $T^{(j)}_{SPU}(\gamma)$ may be sorted in advance to immediately deduce $P^{(k)}(\gamma)$ for all k's.

Then, let

$$T^{(k)}_{aSPU} = \min_{\gamma\in S} P^{(k)}(\gamma).$$

Finally, the p-value of the aSPU test is

$$P_{aSPU} = \frac{1}{N+1}\left(1 + \sum_{k=1}^{N} \mathbb{I}\left(T^{(k)}_{aSPU} \leq T_{aSPU}\right)\right). \tag{7}$$

2.2. Efficiency issue of aSPU

The score vectors u in (3) and $u^{(k)}$ in (4) are written as

$$u = X^T(y - \bar{y}) \quad \text{and} \quad u^{(k)} = X^T(y^{(k)} - \bar{y}). \tag{8}$$

Since u and $u^{(k)}$ are calculated through full matrix-vector multiplications, they are expensive to compute when X is large, i.e., when there is a large number of SNPs in a gene and there are many subjects. Moreover, in order to accurately calculate the p-values of the association tests, a large number (N) of permutations is needed. Each permutation requires the calculation of $u^{(k)}$ in (8) and thus one matrix-vector multiplication. Therefore, the matrix-vector multiplication cost is the major computational burden in aSPU. Specifically, when X is dense, each multiplication costs about $2np$ floating point operations (flops) and the total cost for computing u and all the N vectors $u^{(k)}$ is then roughly

$$2(N+1)np. \tag{9}$$

For large N, n, p, this cost is very high. Therefore, it is desirable to improve the efficiency of the aSPU test.

2.3. aSPU-LR: Fast aSPU via low-rank approximations

We now show how to accelerate aSPU while maintaining the accuracy of the test.

2.3.1. Fast SNP matrix-vector multiplications via low-rank approximations

In practice, when the SNP matrix X is perturbed slightly, the impact on association studies is small. More precisely, if X is approximated by another matrix $\tilde{X}$ with modest approximation accuracy, we can still effectively perform association tests. An efficient and effective mathematical way to obtain such an approximation is through low-rank approximations. That is, to reach reasonable accuracy, it only needs to keep a small number (r) of leading singular values $\sigma_1, \ldots, \sigma_r$ of X. The resulting low-rank approximation $\tilde{X}$ then has relative approximation accuracy $\frac{\sigma_{r+1}}{\sigma_1}$ (Golub and Van Loan, 2013), where σ_{r+1} is the $(r+1)$th largest singular value. See, e.g., Figs. 1 and 4 later for some examples of singular value plots.

Specifically, for an $n \times p$ SNP matrix X, we let $\tilde{X}$ be a low-rank approximation as follows:

$$X \approx \tilde{X} = QR \quad \left(\boxed{}\right), \tag{10}$$

where Q is $n \times r$ and R is $r \times p$ with $r \ll \min\{n, p\}$. That is, r is the rank of $\tilde{X}$. Then, the score vector u in (8) is approximated by

$$\tilde{u} = \tilde{X}^T(y - \bar{y}) = R^T[Q^T(y - \bar{y})] \quad \left(\boxed{} \cdot \boxed{} \cdot \boxed{}\right). \tag{11}$$

Note that $Q^T(y - \bar{y})$ is computed first. Similarly, $u^{(k)}$ in (8) is approximated by

$$\tilde{u}^{(k)} \equiv \tilde{X}^T(y^{(k)} - \bar{y}) = R^T[Q^T(y^{(k)} - \bar{y})].$$

The cost to compute $\tilde{u}$ or $\tilde{u}^{(k)}$ is roughly $2r(p+n)$, which is a significant reduction from the cost $2pn$ for computing u or $u^{(k)}$.

Throughout all the permutations, the low-rank approximation helps gain significant savings in the multiplication cost. The resulting aSPU test, denoted as aSPU-LR for convenience, has total matrix-vector multiplication cost

$$2(N+1)r(p+n), \tag{12}$$

which is in sharp contrast with (9) for small r.

Note that aSPU-LR maintains the test accuracy by controlling the low-rank approximation accuracy. In fact, based on the low-rank approximation, the approximation accuracy of the score vector u is as follows:

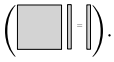
$$\begin{aligned}\|u - \tilde{u}\|_2 &= \|(X^T - \tilde{X}^T)(y - \bar{y})\|_2 \\ &\leq \|X - \tilde{X}\|_2 \|y - \bar{y}\|_2 \leq \sqrt{n}\|X - \tilde{X}\|_2.\end{aligned}$$

(Note that the entries of y are 0 or 1.) Thus, if $\|X - \tilde{X}\|_2$ is small, the approximate vector $\tilde{u}$ is close to the exact score vector u.

2.3.2. Randomized SVDs for quickly finding low-rank approximations

To obtain the low-rank approximation in (10) (or to compress X), truncated SVDs can be used, but are often inefficient. We instead use randomized SVDs (Halko et al., 2011; Liberty et al., 2007), which has a major benefit in that the dominant cost is just for about r matrix-vector multiplications. The main idea is as follows.

First, we choose a $p \times \tilde{r}$ Gaussian random matrix Z, where $\tilde{r} = r + \alpha$ with α a small integer (called over-sampling size). Then, we compute the product

$$W = XZ \quad \left(\text{  } \right). \tag{13}$$

With a small target rank r, W is a tall and skinny matrix and it is convenient to find a low-rank approximation, say, $W \approx Q\tilde{R}$, where Q has orthonormal columns. Using the formulation of Halko et al. (2011), the low-rank approximation $\tilde{X}$ in (10) can be expressed as:

$$\tilde{X} = QR \quad \text{with} \quad R = Q^T X. \tag{14}$$

The randomized SVD can be made highly reliable for low-rank approximations with controlled accuracy. With high probability, the approximation quality is close to that of truncated SVDs. The over-sampling size α decides the probability and needs not to be very large.

In fact, the failure probability behaves like a multiple of $\frac{1}{\alpha^\alpha}$. In addition, power iterations are often used to further improve the reliability (Halko et al., 2011). That is, W in (13) can be replaced by

$$W = (XX^T)^s XZ, \tag{15}$$

where the power s is a small integer like 1 or 2. This is useful when the singular values of X do not decay fast enough.

Note that the cost of matrix-vector multiplications for computing W and R is a one-time cost. The number of such multiplications is negligible as compared with that in the permutations. The details are given next.

2.3.3. Efficiency benefit of aSPU-LR over aspu

In aSPU-LR, the randomized SVD is first used to obtain the low-rank approximation $\tilde{X}$ in (14). With power iterations in (15), it needs to multiply X with $\tilde{r}$ vectors for $s+1$ times and to multiply X^T with $\tilde{r}$ vectors for $s+1$ times. To get R in (14), it needs to multiply X with another r vectors. The total multiplication cost is about

$$2(2s+1)\tilde{r}np + 2rnp, \tag{16}$$

which is a one-time cost and is much smaller than that in (9) for small values of s and α and a modest value r. In the entire aSPU-LR test, the total cost for matrix-vector multiplications includes (16) and (12), and is much smaller than (9).

The dominant costs of aSPU-LR are shown in Table 1 and are compared with those of the original aSPU. The table also includes additional costs that are less significant, as indicated in the $O(\cdot)$ notation. An example of the cost is to orthogonalize the intermediate matrix-vector products for the purpose of stability.

Table 1
Dominant computational costs (flops) related to score vectors in aSPU and aSPU-LR tests, where n is the sample size, p is the number of SNPs, k is the number of permutations, b is the number of parameters in S, r is the rank of $\tilde{X}$ that approximate X, $\tilde{r}$ is r plus the small over-sampling size α, and s is the power in power iterations for randomized SVDs.

	aSPU	aSPU-LR
Algorithm step		Randomized SVD
Cost		$2(2s+1)\tilde{r}np + 2rnp + O(s\tilde{r}^2(n+p))$
Algorithm step	$u = X^T(y - \bar{y})$	$\tilde{u} = R^T(Q^T(y - \bar{y}))$
Cost	$2np$	$2r(n+p)$
Algorithm step	$T_{SPU}(\gamma) = \sum_{j=1}^{p} (u_j)^\gamma, \gamma \in S$	$T_{SPU}(\gamma) = \sum_{j=1}^{p} (\tilde{u}_j)^\gamma, \gamma \in S$
Cost	$O(bp)$	$O(bp)$
Algorithm step	$u^{(k)} = X^T y^{(k)}, k = 1:N$	$\tilde{u}^{(k)} = R^T(Q^T y^{(k)}), k = 1:N$
Cost	$2npN$	$2r(n+p)N$
Algorithm step	$T_{SPU}^{(k)}(\gamma) = \sum_{j=1}^{p} (u_j^{(k)})^\gamma, k = 1:N$	$T_{SPU}^{(k)}(\gamma) = \sum_{j=1}^{p} (\tilde{u}_j^{(k)})^\gamma, k = 1:N$
Cost	$O(bpN)$	$O(bpN)$

3. Fast selection of effective parameters

The low-rank approximation to X provides another opportunity to help with association studies. That is, we can further design a procedure to quickly select γ parameters that are likely the most effective in aSPU. This procedure potentially sheds lights on association patterns and gives new possibilities to improve the efficiency of association tests.

3.1. Motivation: γ parameters and association patterns

As mentioned in Section 2.1, one of the major advantages of the aSPU test is that it accommodates different unknown patterns of associations by adopting a whole range of γ values. On the other hand, for specific problems, some of the parameters γ's may be more effective, depending on the actual association patterns. Here, we can view a certain γ to be more effective if this γ tends to be more likely to produce the minimum in the observed statistic (6).

As illustrated in Pan et al. (2014), some examples that relate the γ parameters to the association patterns are as follows.

- If $\gamma = 1$ is very effective in the aSPU test, then most SNP-outcome association directions are likely the same. In this case, aSPU is approximately equal to the burden test.
- If $\gamma = 2$ is very effective in the aSPU test, then the SNP-outcome associations likely involve different directions. In this case, the aSPU test is approximately equal to variance-component tests such as the SKAT test.
- If more odd γ's than even γ's are effective in aSPU test, then most of the SNP-outcome association directions tend to be the same. Otherwise, the directions tend to the different.
- If large γ's are effective, then the association signals tend to be sparse, meaning few SNPs are significantly associated with the outcome. The extreme case is $\gamma = \infty$.

Since aSPU empirically employs multiple parameters γ, it is not exactly known which γ's are more effective and hence the association patterns are unclear. This motivates us to find those γ's that are likely effective in aSPU so as to gain more insights into the association patterns. The availability of the low-rank approximation to X makes it possible to quickly do so.

3.2. Fast selection of effective γ parameters

Given a type of problems or a scenario, our ideas to select effective γ parameters from the candidate set S are outlined in the following scheme.

- In aSPU, each γ corresponds to a p-value $P_{SPU}(\gamma)$ in (5). The minimum p-value among the $P_{SPU}(\gamma)$ values is taken as the new observed test statistic as in (6). Thus, it is reasonable to expect that, among all the γ's, the γ that produces the $\min_{\gamma\in S} P_{SPU}(\gamma)$ is an effective parameter.
- In order to reliably detect the most effective γ's for the entire scenario or type of problems under consideration, we aim to test multiple (M) datasets from the problem type or scenario. Among the set S, the γ's that are effective for most of the datasets are likely the most effective for the entire problem type or scenario.
- Since we are dealing with the given SNP matrix X and binary outcome vector y, we simulate M datasets (SNP matrices $\dot{X}$ and binary outcome vectors $\dot{y}$) for the problem type or scenario.
  - In simulation studies, we generate multiple simulations $\dot{X}$ and $\dot{y}$ following the setup of the simulation scenario.
  - In real data analysis, we adopt a random sampling strategy to generate M data subsets from the original dataset. That is, we randomly pick a certain percentage of samples from the original dataset.

  For each simulation i, when the aSPU statistic as in (6) is obtained from the SPU p-value (5) (denoted $P^{\langle i\rangle}_{SPU}(\gamma)$ here corresponding to i), let

  $$\gamma^{\langle i\rangle}_* = \arg\min_{\gamma\in S} P^{\langle i\rangle}_{SPU}(\gamma).$$

  At this point, we count the number of simulation cases for which each γ serves as $\gamma^{\langle i\rangle}_*$:

  $$\rho_\gamma = \sum_{i=1}^{M} \mathbb{1}\Big(P^{\langle i\rangle}_{SPU}(\gamma) = P^{\langle i\rangle}_{SPU}(\gamma^{\langle i\rangle}_*)\Big), \tag{17}$$

  where $\mathbb{1}(\cdot)$ is again the indicator function. Then, the γ's corresponding to the highest values ρ_γ are likely the most effective parameters in aSPU.
- It is quite expensive to use full SNP matrices for matrix-vector multiplications. Instead, we may accelerate the multiplications by low-rank approximations like in aSPU-LR.
  - In the simulations, we replace each simulated matrix $\dot{X}$ by a low-rank approximation $\tilde{X}_{\tilde{r}}$ with a very small rank $\tilde{r}$:

    $$\tilde{X}_{\tilde{r}} = Q_{\tilde{r}} R_{\tilde{r}}, \tag{18}$$

    which enables fast matrix-vector multiplications. For the purpose of parameter detection, $\tilde{r}$ may be set to be even smaller than the rank r previously used to find $\tilde{X}$. $\tilde{X}_{\tilde{r}}$ is used in aSPU-LR following the procedure above to detect effective γ parameters.

– In real data analysis, since $\dot{X}$ is sampled from the original SNP matrix X, we can simply take a highly compressed low-rank approximation $\tilde{X}_{\tilde{r}} \approx X$ with a very small rank $\tilde{r}$ and then only sample the basis matrix $Q_{\tilde{r}}$ in (18). That is, only one randomized SVD is needed to get (18). Then, the randomized sampling is just performed on the skinny matrix $Q_{\tilde{r}}$. This directly yields a low-rank approximation to each sampled $\dot{X}$ matrix, which is then used in aSPU-LR following the procedure above to detect effective γ parameters. The whole process is highly efficient.

3.3. Implications of the effective γ parameter selection

The fast selection of effective γ parameters can potentially benefit association studies from multiple aspects.

- As mentioned in Section 3.1, the identified effective γ parameters can potentially reflect the association patterns.
- In practical applications to a problem type or scenario, once selected effective γ's are identified, they may be directly used for individual association tests. This avoids the need to go through all empirical parameters in S and can reduce the computational cost for evaluating score vectors.
- Also, additional savings may be achieved for other aSPU related procedures. For example, in a recent method that uses importance sampling in aSPU, different procedures are designed for different γ values (Deng et al., 2022). If some effective γ values can be preselected, then it avoids the need to implement all the procedures for different γ's, which can significantly reduce the complexity of the implementation and improve the efficiency of the tests.

4. Fast selection of representative SNPs

The existence of low-rank approximations to the SNP matrix further suggests the possibility of finding representative SNP columns that capture the major genetic information of the full SNP matrix X. On the other hand, these columns are unknown from $\tilde{X}$ in (10) or from other dimension reduction methods like the PCA or truncated SVD. Here, we show how to select representative SNPs from the full SNP matrix X with little extra cost. That is, the available low-rank approximations $\tilde{X}$ can help quickly extract valuable SNP information from the data.

4.1. Pivoting for SNP selection

To select representative SNPs, one idea is to apply pivoting to the columns of the centered SNP matrix $\hat{X}$, which is X except with the sample means subtracted from the columns. That is, let $\bar{x}_j$ be the sample mean of column j of X and subtract $\bar{x}_j$ from the column. This can be put into the following matrix form:

$$\hat{X} = X - \mathbf{1}\bar{x}^T, \tag{19}$$

where $\mathbf{1}$ is a length-n vector with each entry 1 and $\bar{x} = \begin{bmatrix}\bar{x}_1 & \cdots & \bar{x}_p\end{bmatrix}^T$. Since the column norms approximately reflect the variances of the corresponding SNPs, we can apply column pivoting to $\hat{X}$ so as to identify important columns.

In matrix computations, this procedure may be accomplished through careful column permutations to produce a so-called interpolative decomposition (Liberty et al., 2007) or structure-preserving rank-revealing factorization (Xia et al., 2012) as follows:

$$\hat{X}\Pi \approx \hat{X}_1 \begin{bmatrix} I & E \end{bmatrix}, \tag{20}$$

where Π is a permutation matrix, $\hat{X}_1$ corresponds to selected columns of $\hat{X}$, I is the identity matrix, and E is a matrix computed through an appropriate procedure to ensure the accuracy and stability of the low-rank approximation. The key component of this approximation is a pivot selection process to permute the columns of $\hat{X}$ so as to find a set of pivot indices that can be used to pick $\hat{X}_1$ from $\hat{X}$. A reliable way to obtain the approximation (20) is through the so-called strong rank-revealing factorizations (Gu and Eisenstat, 1996), which are quite expensive.

4.2. Fast pivoting

In practice, it usually suffices to perform pivot selection by using more efficient alternatives to strong rank-revealing factorizations. One convenient and frequently used pivoting strategy is the Gram-Schmidt process with column pivoting (Golub and Van Loan, 2013). The process involves the following main steps, explained within the context of the centered SNP matrix $\hat{X}$.

1. Find the column, say, $\hat{x}_1$ of $\hat{X}$ with the largest 2-norm. Statistically, this means that $\hat{x}_1$ is the column with the largest sample variance since the columns of $\hat{X}$ have been centered. Record the column index of $\hat{x}_1$ in $\hat{X}$ as the first pivot.
2. Project the remaining columns of $\hat{X}$ into the direction given by $\hat{x}_1$ and remove the projected information from those columns. Mathematically, this is to replace the j-th column $\hat{x}_j$ by

$$\hat{x}_j - \frac{\hat{x}_j^T \hat{x}_1}{\|\hat{x}_1\|_2^2}\hat{x}_1.$$

3. For these remaining modified columns $\hat{x}_j$, apply the above procedure again to continue with the pivoting.

Applying the Gram-Schmidt process to the full matrix $\hat{X}$ costs $O(rnp)$. We can actually reduce this to as little as $O(r(n+p)+r^2p)$ since we already have the low-rank approximation (14) for X. Following (19) and (10), we have

$$\hat{X} = X - \mathbf{1}\bar{x}^T \approx QR - \mathbf{1}\bar{x}^T = \begin{bmatrix} Q & \mathbf{1} \end{bmatrix} \begin{bmatrix} R \\ -\bar{x}^T \end{bmatrix}. \tag{21}$$

Here, we seek to quickly perform the pivot selection. For this purpose, we would like to make sure $\begin{bmatrix} Q & \mathbf{1} \end{bmatrix}$ has orthonormal columns. That is, we only need to apply one Gram-Schmidt step to orthonormalize the last column. That is, let

$$\tilde{q} = \mathbf{1} - Q(Q^T\mathbf{1}), \quad q = \frac{\tilde{q}}{\|\tilde{q}\|_2}.$$

Then,

$$\begin{bmatrix} Q & \mathbf{1} \end{bmatrix} = \begin{bmatrix} Q & \tilde{q} + Q(Q^T\mathbf{1}) \end{bmatrix} = \begin{bmatrix} Q & \tilde{q} \end{bmatrix} \begin{bmatrix} I & Q^T\mathbf{1} \\ & 1 \end{bmatrix} = \begin{bmatrix} Q & q \end{bmatrix} \begin{bmatrix} I & Q^T\mathbf{1} \\ & \|\tilde{q}\|_2 \end{bmatrix}.$$

(21) now becomes

$$\hat{X} \approx \begin{bmatrix} Q & q \end{bmatrix} \begin{bmatrix} I & Q^T\mathbf{1} \\ & \|\tilde{q}\|_2 \end{bmatrix} \begin{bmatrix} R \\ -\bar{x}^T \end{bmatrix}.$$

Thus, we obtain

$$\hat{X} \approx \hat{Q}\hat{R} \quad \text{with} \quad \hat{Q} = \begin{bmatrix} Q & q \end{bmatrix}, \quad \hat{R} = \begin{bmatrix} R - (Q^T\mathbf{1})\bar{x}^T \\ -\|\tilde{q}\|_2\bar{x}^T \end{bmatrix}. \tag{22}$$

At this point, since $\hat{Q}$ has orthonormal columns, the j-th column $\hat{x}_j$ of $\hat{X}$ therefore satisfies

$$\|\hat{x}_j\|_2 \approx \|\hat{Q}\hat{R}_j\|_2 \approx \|\hat{R}_j\|_2,$$

where $\hat{R}_j$ is the j-th column of $\hat{R}$. Thus, the norms of the columns of $\hat{R}$ approximate those of the corresponding columns of $\hat{X}$. On the other hand, $\hat{R}$ has only r rows and it is very efficient to apply Gram-Schmidt with column pivoting to select representative indices. Accordingly, we can approximately identify the SNPs that are likely the most representative. To enhance the reliability of the pivoting process, we may also try to find an interpolative decomposition like in (20) for $\hat{R}$.

4.3. Significance of representative SNP selection

We now say a few words about the implications of this SNP selection process.

- For highly correlated data $\hat{X}$, most columns are approximately linear combinations of some basis columns. The process essentially chooses the most linearly independent columns.
- Similarly to the PCA, our method reduces the dimension of $\hat{X}$ and can uncover underlying structures in high-dimensional datasets. However, as compared with the PCA, our method enjoys an advantage. It can quickly and directly select biomarkers (i.e., important SNPs). From the perspective of practical applications, it is beneficial to directly extract patients' biological profiles so as to improve the effectiveness of personalized medicine and reduce healthcare costs.
- For large-scale high-dimensional $\hat{X}$, we may first use this fast pivoting process to select the most important SNPs and then put the selected SNPs in association tests (e.g. fisher's exact SNP-based tests) to study the interactions between individual SNPs and the binary outcome. Depending on the quality of the identified SNPs, this may further improve the efficiency.

Note that the SNP selection process does not involve the outcome information y. The real data studies in Section 6 show that we can still use the selected representative SNPs to gain some biology insights without y. For such cases, the representative SNPs are likely driver SNPs that can interact with many outcome traits and hence may master downstream biology regulation mechanisms such as transcriptions, protein functions, and clinical outcomes. In other words, those driver SNPs may be key contributors to traits or diseases.

5. Simulation studies

We conducted extensive simulation studies to assess the performance of the proposed strategies. We used the studies of the interactions between germline variations and somatic mutations like in Chen et al. (2023) as an example. Following similar setups as in Chen et al. (2023) and Pan et al. (2015), we simulated the SNP matrices X and a single-gene somatic mutation vector y. Some details are as follows. To simulate X, we first generated a latent vector $x = \begin{bmatrix} x_1 & \cdots & x_p \end{bmatrix}^T$ from a multivariate distribution based on the autoregressive covariance structure, where the entry of the correlation matrix corresponding to x_i and x_j is $\xi^{|i-j|}$ with ξ following a uniform distribution $\mathcal{U}(0.8, 0.9)$. We then generated two haplotypes independently based on a value of minor allele frequency (MAF) that was randomly selected between 0.05 and 0.4. Next, we combined these two independent haplotypes to get a germline SNP matrix X. To simulate y, we followed the logistic regression model (2), with a 10% background mutation probability. In (2), all $\beta_j = 0, j = 1, \ldots, p$ for the null hypothesis and $\beta_j = \log\text{OR} \neq 0$ for some SNPs within a gene for the alternative hypothesis.

The following notation will be used in the presentation of the results.

- The patient data involves n simulated samples in the cohort study.
- The number of SNPs in one gene is p.
- The number of causal SNPs in one gene is c.
- N permutations are used in the p-value computations.
- M simulation runs are used to evaluate statistical powers and type I error rates.

5.1. Comparison of aSPU and aSPU-LR tests

To demonstrate the computational efficiency and the effectiveness of the aSPU-LR test, we compared it with the original aSPU test in terms of their computation time, statistical powers, and type I error rates under 9 cases. As usual, the test significance level is 0.05. The aSPU code is based on the R package aSPU (Kwak et al., 2021) but with some efficiency improvements. For example, the permutations of y are done collectively outside the main test routine. The code was then further used to design the aSPU-LR routines. All the tests were done on a laptop with AMD Ryzen 7 5825U 2.0GHz CPU and 64GB memory.

We first inspected the singular values of the SNP matrices X of size $n \times p$ with $n = 500$ and $p = 100$ or 500. For one set of simulations, Fig. 1 shows reasonably decaying singular values for SNP matrices X under different cases. In particular, if we take a small amount

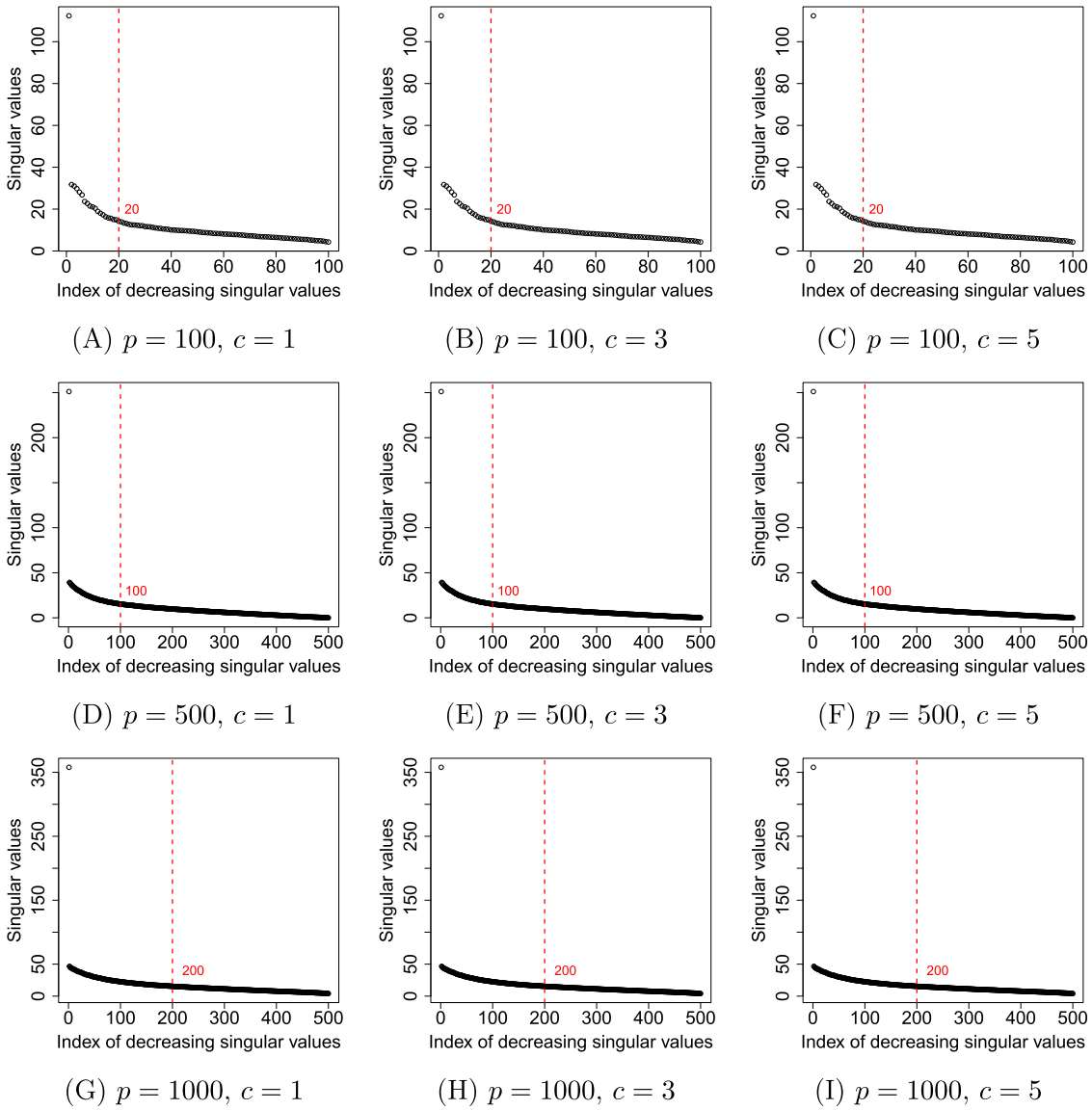


Fig. 1. Singular values of SNP matrices X under 9 cases with different choices of p (number of SNPs) and c (number of causal SNPs), where each red dashed vertical line indicates the numerical rank we used for the approximation of X. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of dominate singular values with rank r as

$$r = \left\lfloor \frac{1}{5} \min\{n, p\} \right\rfloor, \tag{23}$$

then the approximation matrix $\tilde{X}$ can reach about one digit of relative accuracy. When $n \ll p$ (e.g., $n = 500$ and $p = 1000$), the singular values decay more slowly than in the case $n \geq p$ (see Fig. 1). Consequently, more singular values are required to approximate X with comparable accuracy. Thus, we increased the rank to

$$r = \left\lfloor \frac{2}{5} \min\{n, p\} \right\rfloor. \tag{24}$$

In fact, the current setup has high linkage disequilibrium, meaning the SNPs in a gene are highly correlated. Thus, in aSPU-LR, we constructed low-rank approximations $\tilde{X}$ in (10) with rank r in (23) or (24). Throughout the tests in this paper, the power $s = 2$ was used for power iterations in randomized SVDs for constructing low-rank approximations.

For X with different sizes and with one causal SNP, Table 2 shows the computation time of aSPU and aSPU-LR for 100 simulations. (Similar speedup was observed with more causal SNPs.) The candidate set of γ parameters is

$$S = \{1, 2, \ldots, 8, \infty\}, \tag{25}$$

which is often sufficient in practice (Pan et al., 2014). It is clear that, even with small matrix sizes, aSPU-LR is already much faster than aSPU. The results in Table 2 show speedups of about 2 to 9 times. Such speedups grow with the dimensions of X. Also, when the number of permutations increases, the speedup gets more significant. The timing of aSPU-LR also includes that for randomized SVDs, which only accounts for a small portion of the total time. For example, for 100 simulations, when $p = 100$ and $n = 500$, the total randomized SVD time is 0.36 seconds, while the total time of aSPU-LR with 10^4 permutations is 24.24 seconds. When $p = 500, n = 4000$, the total randomized SVD time is 45.45 seconds while the total time of aSPU-LR with 10^4 permutations is 519.85 seconds. As previously noted, when $n \ll p$ (e.g., $n = 500$ and $p = 1000$), the singular value decay is less significant as compared with the $n \geq p$ cases. Then, in order to get comparable approximation accuracy of X (and thus comparable statistical powers and type I error rates), a higher rank as in (24) is used. Accordingly, aSPU-LR has slightly higher computational costs. Nevertheless, aSPU-LR still gains reasonable time savings compared to aSPU even when the sample size is small ($n = 500$) (see the last row of Table 2).

Table 2
Computation time (in seconds) for aSPU and aSPU-LR, where each result includes the time for 100 simulations. The number of causal SNPs in one gene is $c = 1$.

		$N = 10^3$		$N = 10^4$	
p	n	aSPU	aSPU-LR	aSPU	aSPU-LR
100	500	10.14	2.94	96.39	24.24
	1000	22.33	4.52	222.98	35.23
	2000	45.68	6.96	451.84	57.46
	4000	91.19	12.82	922.86	108.44
500	500	57.60	18.87	566.46	125.00
	1000	113.72	29.14	1119.54	178.03
	2000	221.06	48.73	2228.45	289.21
	4000	464.13	92.98	4662.39	519.85
1000	500	113.26	61.31	1144.47	191.41

With greatly improved efficiency, aSPU-LR still maintains high reliability. Fig. 2 shows the statistical powers of aSPU-LR and aSPU. The two tests exhibit very similar powers. As the number of causal SNPs increases, the powers of both aSPU and aSPU-LR increase. When the number of causal SNPs is 5, the powers of the tests are nearly indistinguishable across different $\log \mathrm{OR}$ situations. Note that when $n \ll p$ (e.g., $n = 500$ and $p = 1000$), the statistical powers of aSPU and aSPU-LR (with the numerical rank $r = 200$) are still very similar.

While calculating the powers of the tests, we also computed the type I error rates. Since type I error rates in genetic association studies may vary depending on the value of c in aggregation methods (see, e.g., Chen et al. (2023)), different c values were tested so as to perform a comprehensive comparison between aSPU and aSPU-LR. Table 3 shows similar type I error rates around the significance level 0.05 for the two tests. When $n \ll p$ (e.g., $n = 500, p = 1000$), the type I error rates of aSPU and aSPU-LR (with the numerical rank $r = 200$) are still very similar. These results confirm the efficiency and the reliability of aSPU-LR based on low-rank approximations under different scenarios.

5.2. Fast selection of effective parameters

We then used the procedure in Section 3.2 to select effective γ parameters from a candidate set

$$S = \{1, 2, \ldots, 32, \infty\}. \tag{26}$$

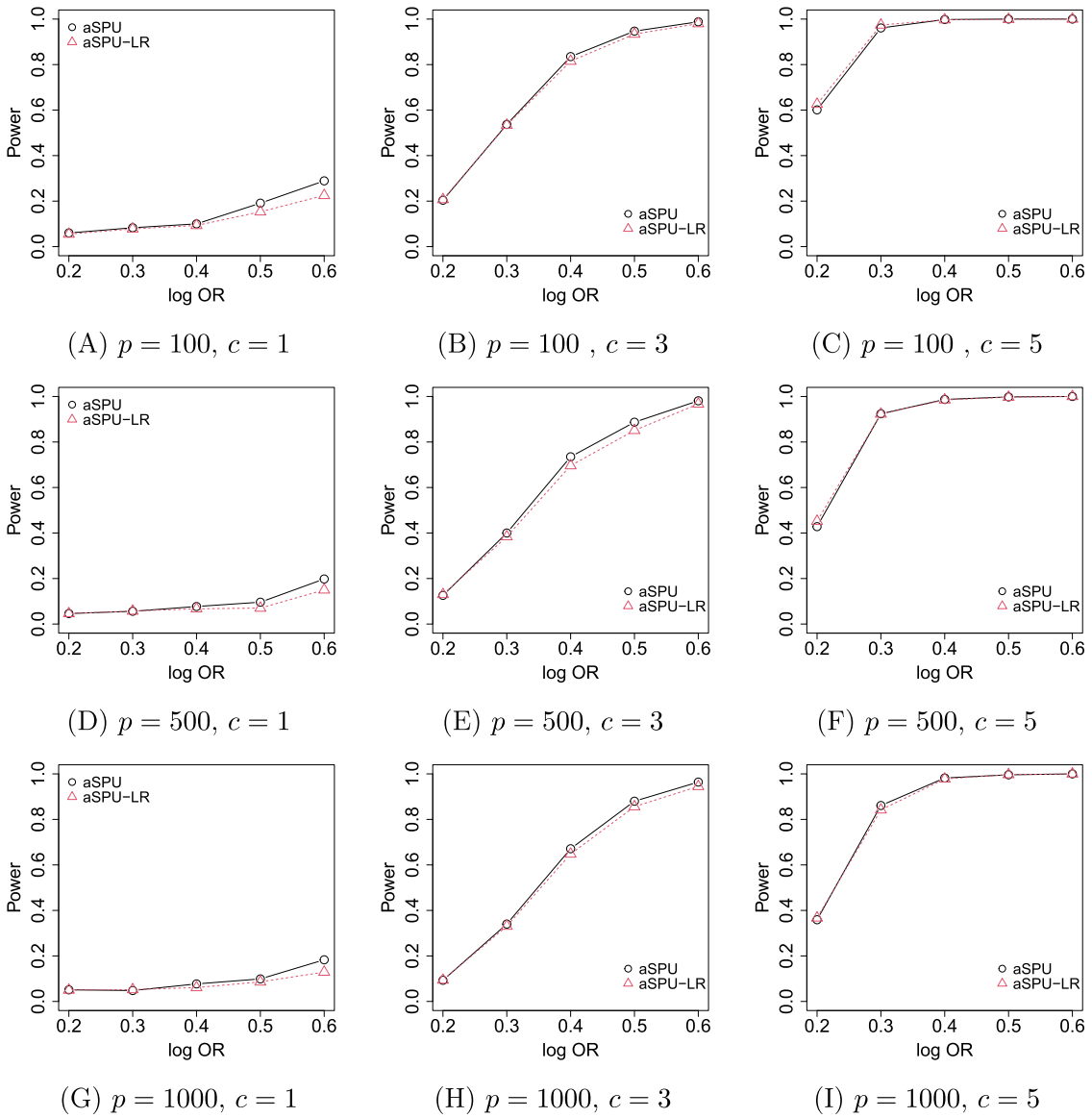


Fig. 2. Statistical powers of aSPU and aSPU-LR under 9 cases with different choices of p (number of SNPs) and c (number of causal SNPs). The sample size is $n = 500$. To calculate the p-values, the number of permutations for both aSPU and aSPU-LR is $N = 10^3$.

Table 3
Type I error rates for aSPU and aSPU-LR for the 9 test cases, where the sample size is $n = 500$ and p is the number of SNPs. To calculate the p-values, the number of permutations for both aSPU and aSPU-LR is $N = 10^3$.

p	c	aSPU	aSPU-LR	p	c	aSPU	aSPU-LR	p	c	aSPU	aSPU-LR
	1	0.055	0.053		1	0.059	0.054		1	0.063	0.064
100	3	0.055	0.053	500	3	0.059	0.054	1000	3	0.063	0.064
	5	0.055	0.053		5	0.059	0.054		5	0.063	0.064

To demonstrate how the selected effective γ parameters may reflect the association patterns, we simulated different scenarios. Suppose there are $n = 500$ patients and $p = 100$ or 1000 SNPs in a gene. The following scenarios were considered based on different causal SNP information:

(i) $c = 5$ causal SNPs, with $\log \mathrm{OR}$ values $2, 2, 2, 2, 2$, respectively, and $p = 100$;
(ii) $c = 5$ causal SNPs, with $\log \mathrm{OR}$ values $2, -2, 2, -2, 2$, respectively, and $p = 100$;

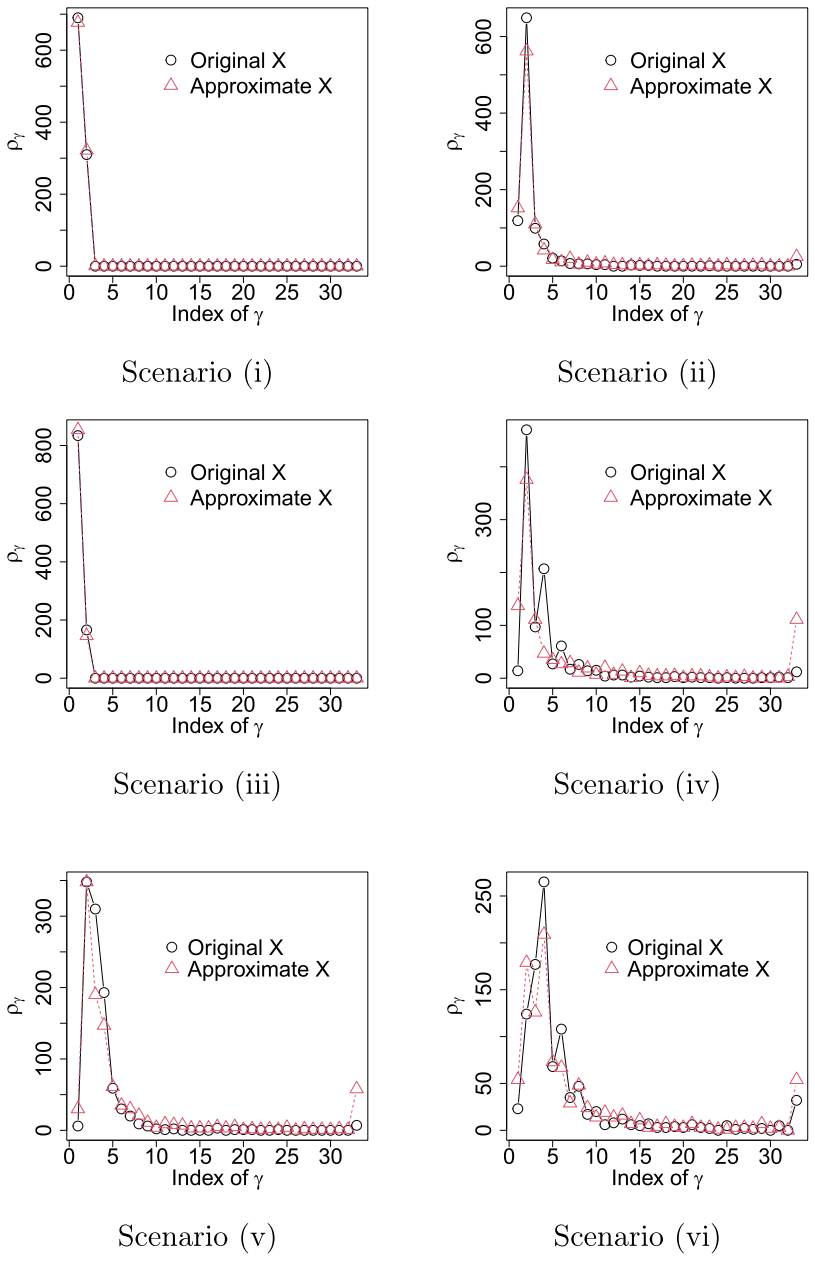


Fig. 3. ρ_γ in (17) based on the original matrix X and on the approximation matrix $\tilde{X}_{\tilde{r}}$, where the horizontal axis is for the indices of γ in the set S in (26) (with the index 33 for ∞). The numerical rank is $r = 100$ for the scenarios (i)-(iv) and $r = 200$ for the scenario (vi). The number of permutations for both aSPU and aSPU-LR is $N = 10^3$.

(iii) $c = 10$ causal SNPs, with $\log \mathrm{OR}$ values $2, 2, 2, 2, 2, 2, 2, 2, 2, 2$, respectively, and $p = 100$;
(iv) $c = 10$ causal SNPs, with $\log \mathrm{OR}$ values $2, -2, 2, -2, 2, -2, 2, -2, 2, -2$, respectively, and $p = 100$.
(v) $c = 5$ causal SNPs, with $\log \mathrm{OR}$ values $2, -2, 2, -2, 2$, respectively, and $p = 1000$;
(vi) $c = 10$ causal SNPs, with $\log \mathrm{OR}$ values $2, -2, 2, -2, 2, -2, 2, -2, 2, -2$, respectively, and $p = 1000$.

In the implementation of the procedure in Section 3.2, $N = 1000$ permutations were used in p-value calculations. For each $\gamma \in S$, to calculate ρ_γ in (17), we used $M = 1000$ simulations.

The simulation results are shown in Fig. 3. We plotted ρ_γ in (17) against the indices of γ in the set S in (26). To show that we can use highly compressed low-rank approximations for quick parameter selection, we used a very small rank $\tilde{r} = 10$ (when $p = 100$) or $\tilde{r} = 100$ (when $p = 1000$) to obtain $\tilde{X}_{\tilde{r}}$ in (18) which was then used in aSPU-LR to compute ρ_γ. We compared the ρ_γ values obtained with aSPU-LR (using $\tilde{X}_{\tilde{r}}$) and those with aSPU (using the original matrices X). Clearly, the ρ_γ patterns from the two strategies are very similar. The most significant ρ_γ values match very well.

This test indicates that highly compressed low-rank approximations are sufficient to quickly identify effective γ parameters. Accordingly, this gives valuable insights into the association patterns of SNPs. For example, in Fig. 3, for the scenarios (i) and (iii), $\gamma = 1$ corresponds to the highest ρ_γ value that is much larger than with other γ's. Thus, $\gamma = 1$ is viewed as the most effective parameter, which further indicates that the association directions between most SNPs in the germline gene and the somatic mutation are likely the same. This is consistent with the setup at the beginning of this subsection. Similarly, for the scenarios (ii), (iv), (v), and (vi), $\gamma = 2$ is the most effective parameter, which further indicates likely different association directions.

When $n \ll p$ (e.g., $n = 500$ and $p = 1000$), the performance of the low-rank approximations and the effective parameter selection is as follows. In general, when we still use a small numerical rank r as in eq. (23), the approximation accuracy of the SNP matrix X seems worse than that in the case of $n \geq p$. Especially, as the number of causal SNPs increases (scenario (vi)), the approximation accuracy of the SNP matrix X decreases. However, with an increased rank r as in eq. (24), the approximation accuracy becomes reasonably good (see scenario (vi) in Fig. 3). At the same time, the increased rank does not significantly increase the computational time. This is consistent with the performance of aSPU-LR in Section 5.1.

In general, when $n \ll p$, it is challenging to use low-rank forms to accurately approximate the data matrix and to identify key features among so many SNPs. It may also be harder to identify γ parameters that can precisely indicate the association patterns of SNPs. For example, we have observed that, with $n = 500$, $p = 1000$, and $\log \mathrm{OR} = 1$ for all causal SNPs, the effective parameter we identified is $\gamma = 2$. This is inconsistent with the true effect direction. This means that when $n \ll p$, it is challenging to identify the effective γ parameters that can precisely indicate the association patterns of SNPs. This needs further investigations in future research.

To further validate that the identified γ parameters are the key parameters for the accuracy of aSPU, we calculated the statistical powers and type I error rates of aSPU using only those identified γ parameters, and compared with those using the full candidate set S in (26). The aSPU tests with the selected effective parameters and with full S show similar statistical powers and type I error rates. See Table 4. For the scenario (vi) ($n = 500, p = 1000, c = 10$), the statistical powers of both tests (with the selective parameters and with the full set S) are relatively lower than those of tests in other settings. In fact, when $n \ll p$, the increasing number of causal SNPs reduces the signal-to-noise ratio (Loh et al., 2015). A high number of causal SNPs when $n \ll p$ also increases the model complexity and uncertainty. In particular, it increases the risk of overfitting to the data (Ling et al., 2021) and the signals are confounded when the multiple causal SNPs are in linkage disequilibrium (Faye et al., 2013).

Table 4
Statistical powers and type I error rates of aSPU tests with the full candidate set S in (26) and with selected parameters, where scenarios (i) and (iii) use $\gamma = 1$, (ii) and (iv) use $\gamma = 2$, and (v) and (vi) use $\gamma \in S = \{2, 3, 4\}$. The sample size is $n = 500$. The scenarios (i)–(iv) use the number of SNPs $p = 100$ and the scenarios (v)-(vi) use $p = 1000$. To calculate the p-values, the number of permutations for both aSPU and aSPU-LR is $N = 10^3$.

	Power		Type I error	
Scenario	Full S	Selected γ	Full S	Selected γ
(i)	1.000	0.975	0.043	0.047
(ii)	0.959	0.915	0.043	0.055
(iii)	1.000	0.995	0.043	0.047
(iv)	0.896	0.828	0.043	0.055
(v)	0.939	0.930	0.056	0.048
(vi)	0.757	0.739	0.056	0.048

5.3. Identification of representative SNPs

Based on low-rank approximations, we also used the procedure in Section 4.2 to identify representative SNPs. The low-rank approximation $\tilde{X}$ with rank r given in (23) was used to obtain a low-rank approximation (22) for the centered matrix $\hat{X}$ in (19). The Gram-Schmidt process with column pivoting (available from the R software routine `qr`) was then applied to $\hat{R}$ to pick representative SNPs. The sample size is $n = 500$. The total number of SNPs in a gene is $p = 100$, 500, or 1000 with $c = 5$ or 10 causal SNPs. The $\log \mathrm{OR}$ value is 0.8 for each causal SNP, meaning the effect size of each causal SNP is $e^{0.8} \approx 2.23$.

To validate the identified representative SNPs, we performed Fisher's exact tests between each SNP of a germline gene and the somatic mutation of another gene. We did 1000 simulations. For each simulation, we identified two lists: one containing the significant SNPs from Fisher's exact tests, and the other containing the representative SNPs selected by our fast pivoting strategy. We calculated three counts as follows.

- Count-1: the total number of occurrences among the 1000 simulations when the two lists have overlaps, i.e., the number of simulations when at least one representative SNP from our fast pivoting strategy is in the list of significant SNPs from Fisher's exact tests.
- Count-2: the total number of occurrences among the 1000 simulations when the most significant SNP from Fisher's exact test is within the list of representative SNPs from fast pivoting.
- Count-3: the number of overlapping SNPs between the two lists occurring on average over 1000 simulations.

Table 5
Among 1000 simulations, the three counts as defined above for indicating overlaps between significant SNPs from Fisher's exact test and representative SNPs from our fast pivoting strategy.

p	c	Count-1	Count-2	Count-3
100	5	982	356	1.165
100	10	1000	356	2.199
500	5	966	339	1.186
500	10	1000	386	2.194
1000	5	946	350	1.196
1000	10	1000	365	2.242

See Table 5 for the three counts. They demonstrate a high degree of overlap between the two lists. Count-1 is very close to 1000, indicating that our strategy almost always identifies some significant SNPs (within a germline gene) that are associated with the somatic mutation of another gene. Count-2 is also reasonably large, meaning our method selects the most significant SNP with a reasonable probability (33.9% to 38.6% in our simulation settings). Count-3 indicates that on average our new strategy identifies at least one significant SNP that overlaps with the significant SNPs identified by Fisher's exact tests, even if a large amount of data (e.g., $n = 500$ and $p = 1000$) has been compressed through the low-rank approximations. Overall, with a small numerical rank r, our fast pivoting strategy has essentially extracted some useful information on important SNPs that are likely associated with somatic mutations of many genes. Such SNPs are possibly driver SNPs that are potentially key contributors to a disease.

6. Real data example

To further demonstrate the efficiency and the effectiveness of the proposed strategies, we then applied them to study interactions between real-world germline variations and somatic mutations across 38 cancer types (Remark 6.1) and extract useful SNP information in those cancers, using the whole genome sequencing ICGC data (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). For data processing, see the reference (Chen et al., 2023, Section 5.1). In the cleaned ICGC data, there are $n = 2561$ samples, 150 germline variation driver genes, and 156 somatic mutation driver genes (Chen et al., 2023, Supplementary material, Section A.4). We only kept the common variants (with MAF larger than 5%).

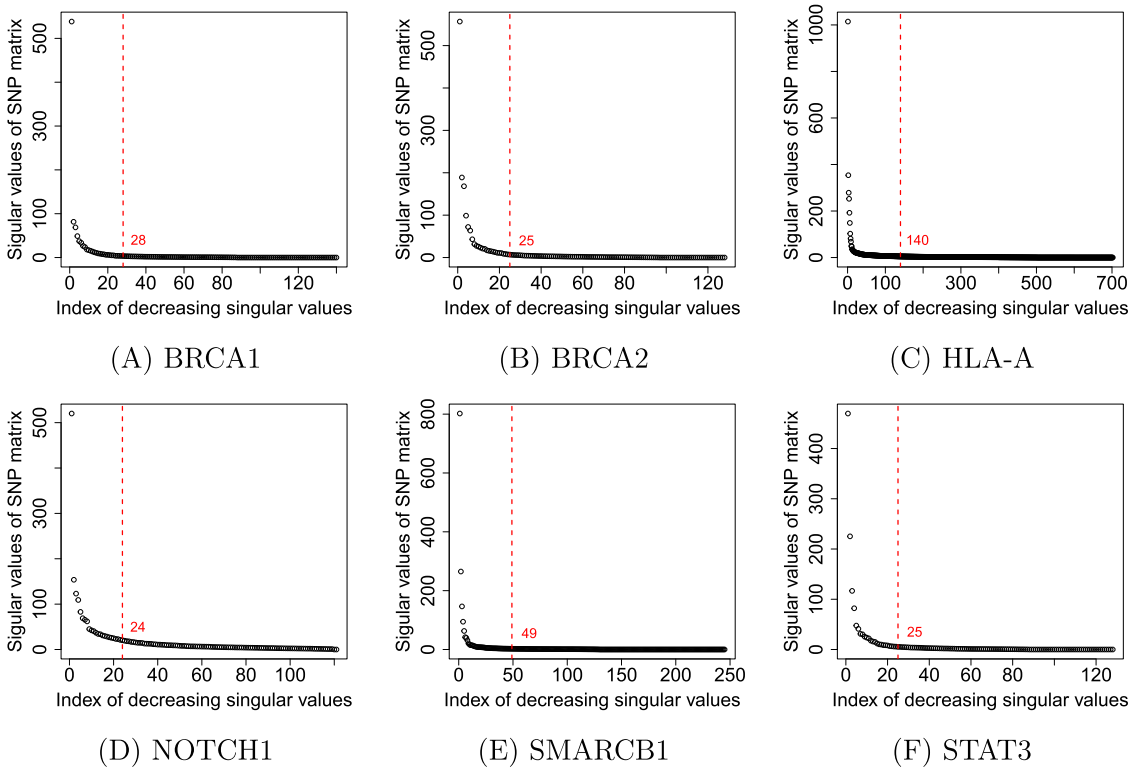


Fig. 4. Singular values of SNP matrices X corresponding to 6 germline variation genes, where the maximum index along each horizontal axis is the total number of SNPs p, where each red dashed vertical line indicates the numerical rank we used for the approximation of X. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6.1. Comparison of aSPU and aSPU-LR tests

Similarly to simulation studies, we inspected the compressibility of the SNP matrices by calculating the singular values. From what we observed, as long as the SNP matrices are not too small (on such occasions, low-rank compression would not be beneficial for accelerating computations), the SNP matrices are all reasonably compressible. Fig. 4 shows the singular values of SNP matrices X from 6 germline variation genes with modest numbers of SNPs p. Again, each plot shows that the number of dominant singular values is reasonably small. Low-rank approximations of X with rank r in (23) can reach up to 2 digits of relative accuracy. This indicates the feasibility of applying low-rank approximations to the SNP matrices in association studies. In fact, this can be further supported by Fig. 5 for the plots of the correlation matrices, which demonstrate high correlations among SNPs within the genes.

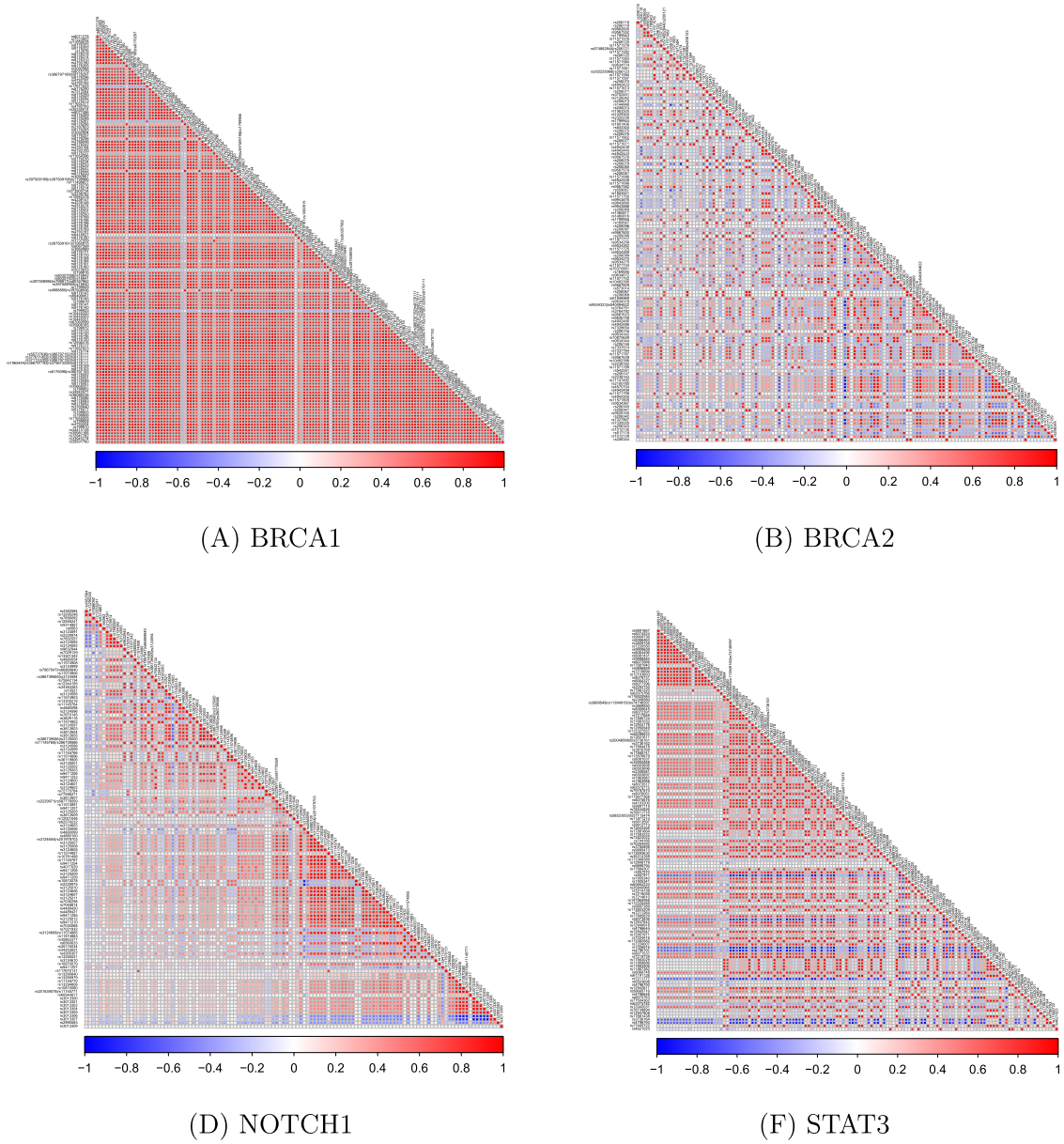


Fig. 5. Plots of correlation matrices for some germline variation genes in Fig. 4, where the cases for (C) and (E) in Fig. 4 are not shown since the matrix sizes are much larger and the plots would be too dense to show meaningful information.

We then used aSPU and aSPU-LR to perform association tests between each of the 6 germline variation genes in Fig. 4 and each of the 156 somatic mutation genes. S in (25) was used for the candidate set of γ parameters in both test methods. The aSPU-LR test uses low-rank approximations $\tilde{X}$ obtained via randomized SVDs with rank r in (23). The total computation time for association tests between one germline gene and all the somatic mutation genes is reported in Table 6. Clearly, aSPU-LR achieves significant efficiency

Table 6
Computation time (in seconds) for aSPU and aSPU-LR, where each result includes the total computation time for association tests between one germline gene and all the 156 somatic mutation genes.

		$N = 10^3$		$N = 10^4$	
Germline gene	p	aSPU	aSPU-LR	aSPU	aSPU-LR
BRCA1	140	429.56	17.90	3538.92	152.20
BRCA2	128	395.87	17.16	3340.97	161.52
HLA-A	702	2157.39	84.11	17813.93	816.15
NOTCH1	121	381.00	17.76	3382.10	154.99
SMARCB1	245	732.43	28.18	5741.07	275.15
STAT3	128	395.70	17.35	3289.93	159.69

improvement over aSPU. All the results in Table 6 show time savings of over 20 times. The time of randomized SVDs only accounts for a small portion of the total aSPU-LR time. For example, for germline gene BRAC1, the randomized SVD time is 0.06 seconds while the total time of aSPU-LR with $N = 10^4$ permutations is 152.20 seconds.

In the meantime, aSPU-LR also produces reliable association results. Fig. 6 shows the QQ plots of $-\log_{10}(p\text{-values})$ obtained from aSPU and aSPU-LR tests. The p-values from aSPU-LR align closely with those from aSPU.

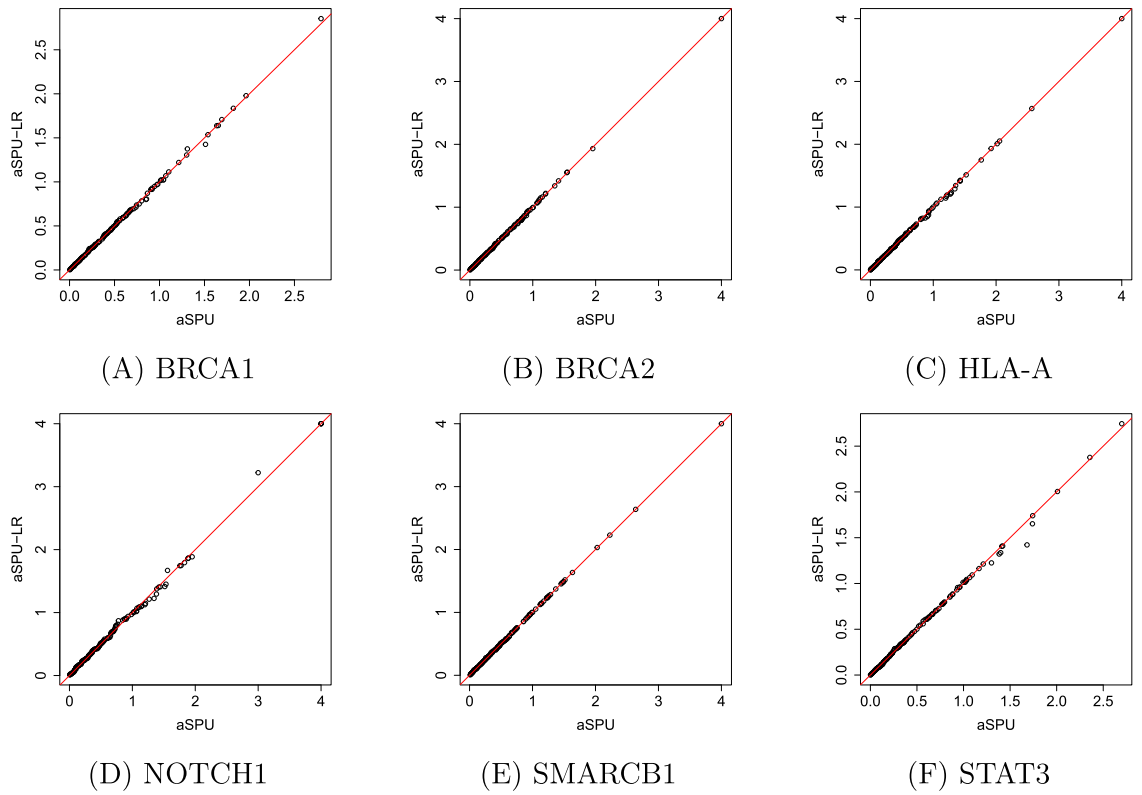


Fig. 6. QQ plots of $-\log_{10}(p\text{-values})$ obtained from aSPU and aSPU-LR tests of associations between 6 germline variation genes and 156 somatic mutation genes. To calculate the p-values, the number of permuations in both aSPU and aSPU-LR is $N = 10^4$.

Specifically, for some somatic mutation genes that demonstrate the most significant associations, as identified by aSPU, the actual p-values from the two tests are given in Table 7. The p-values from aSPU-LR are close to those from aSPU, confirming the effectiveness of aSPU-LR.

These results show that low-rank approximations can greatly accelerate association studies of aSPU while maintaining similar reliability.

6.2. *Fast selection of effective parameters*

We have also used low-rank approximations to select γ parameters that are likely the most effective for aSPU. We focused on the interactions between the above 6 germline variation genes and the somatic mutation of the gene CTNNB1 (since this gene showed

Table 7
Comparison of p-values from aSPU and aSPU-LR for the association tests between the 6 germline variation genes and selected driver somatic mutation genes. To calculate the p-values, the number of permuations in both aSPU and aSPU-LR is $N = 10^4$.

BRCA1	Somatic	CTNNB1	EGFR	APC	ARHGAP35	PTCH1	PCBP1
	aSPU	0.0016	0.0109	0.0151	0.0203	0.0222	0.0231
	aSPU-LR	0.0014	0.0105	0.0146	0.0196	0.0229	0.0230
BRCA2	Somatic	CTNNB1	PIK3CA	PPP2R1A	PA2G4	KRAS	AXIN1
	aSPU	< 0.0001	0.0111	0.0283	0.0287	0.0389	0.0447
	aSPU-LR	< 0.0001	0.0117	0.0277	0.0282	0.0381	0.0458
HLA-A	Somatic	CTNNB1	RRAGC	EZH2	TNFRSF14	SRSF7	ACVR2A
	aSPU	< 0.0001	0.0027	0.0088	0.0096	0.0120	0.0172
	aSPU-LR	< 0.0001	0.0027	0.0089	0.0098	0.0117	0.0178
NOTCH1	Somatic	KRAS	CTNNB1	PIK3CA	SMAD4	TBR1	EGFR
	aSPU	< 0.0001	< 0.0001	0.0010	0.0112	0.0129	0.0130
	aSPU-LR	< 0.0001	< 0.0001	0.0006	0.0214	0.0138	0.0130
SMARCB1	Somatic	CTNNB1	NFE2L2	SMO	AKT1	STAT3	CCND3
	aSPU	< 0.0001	0.0023	0.0059	0.0094	0.0232	0.0305
	aSPU-LR	< 0.0001	0.0023	0.0059	0.0093	0.0232	0.0304
STAT3	Somatic	CTNNB1	SPOP	PCBP1	ACVR2A	EEF1A1	PTPN11
	aSPU	0.0020	0.0044	0.0098	0.0182	0.0208	0.0380
	aSPU-LR	0.0018	0.0042	0.0099	0.0182	0.0223	0.0379

significant associations with most of the 150 germline variation genes in the previous study (Chen et al., 2023)). Following the procedure in Section 3.2, we took the dataset given by each germline variation gene and the somatic mutation of CTNNB1, and performed random sampling to select 80% of the samples as a data subset. Such random sampling was repeated to obtain $M = 1000$ subsets. The aSPU test was applied to these subsets to compute ρ_γ in (17). On the other hand, with our aSPU-LR method, we can quickly obtain approximations to ρ_γ as mentioned in Section 3.2. That is, the SNP matrix is compressed into a low-rank form (18)

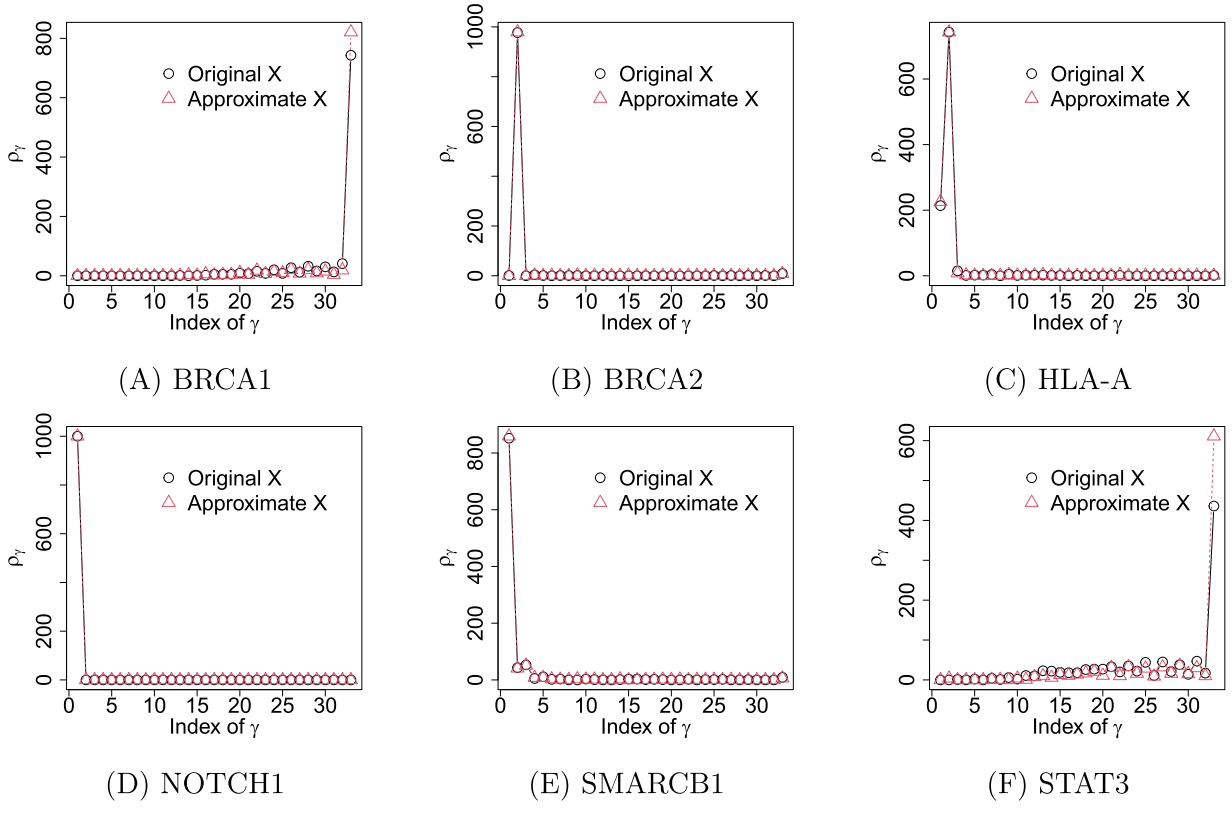


Fig. 7. Low-rank parameter selection with rank $\tilde{r} = 10$ for the association tests between 6 germline variation genes and the somatic mutation of CTNNB1, where the horizontal axis is for the indices of γ in the set S in (26). The number of permutations for both aSPU and aSPU-LR is $N = 10^3$.

with a very small rank $\tilde{r} = 10$. Then the randomized sampling is just performed on the factor $\tilde{Q}_{\tilde{r}}$ in (18). For such small $\tilde{r}$, the speed improvement of aSPU-LR over aSPU is even more dramatic than that in Table 6.

Fig. 7 shows the ρ_γ values produced by aSPU and aSPU-LR. Clearly, the patterns match very well. Thus, aSPU-LR can quickly identify effective γ parameters. Those identified effective γ's reflect some association patterns. For instance, in Fig. 7, for the germline genes NOTCH1 and SMARCB1, $\gamma = 1$ corresponds to the highest ρ_γ value and it is overwhelmingly larger than those with other γ's. Thus, $\gamma = 1$ is viewed as the most effective parameter, which further indicates that the association directions between most SNPs within the germline gene and the somatic mutation of CTNNB1 are likely the same. Similarly, $\gamma = 2$ is the most effective parameter for genes BRCA2 and HLA-A, indicating likely different association directions. Also, $\gamma = \infty$ is the most effective parameter for genes BRCA1 and STAT3, indicating that there is likely a dominant significant SNP in the associations.

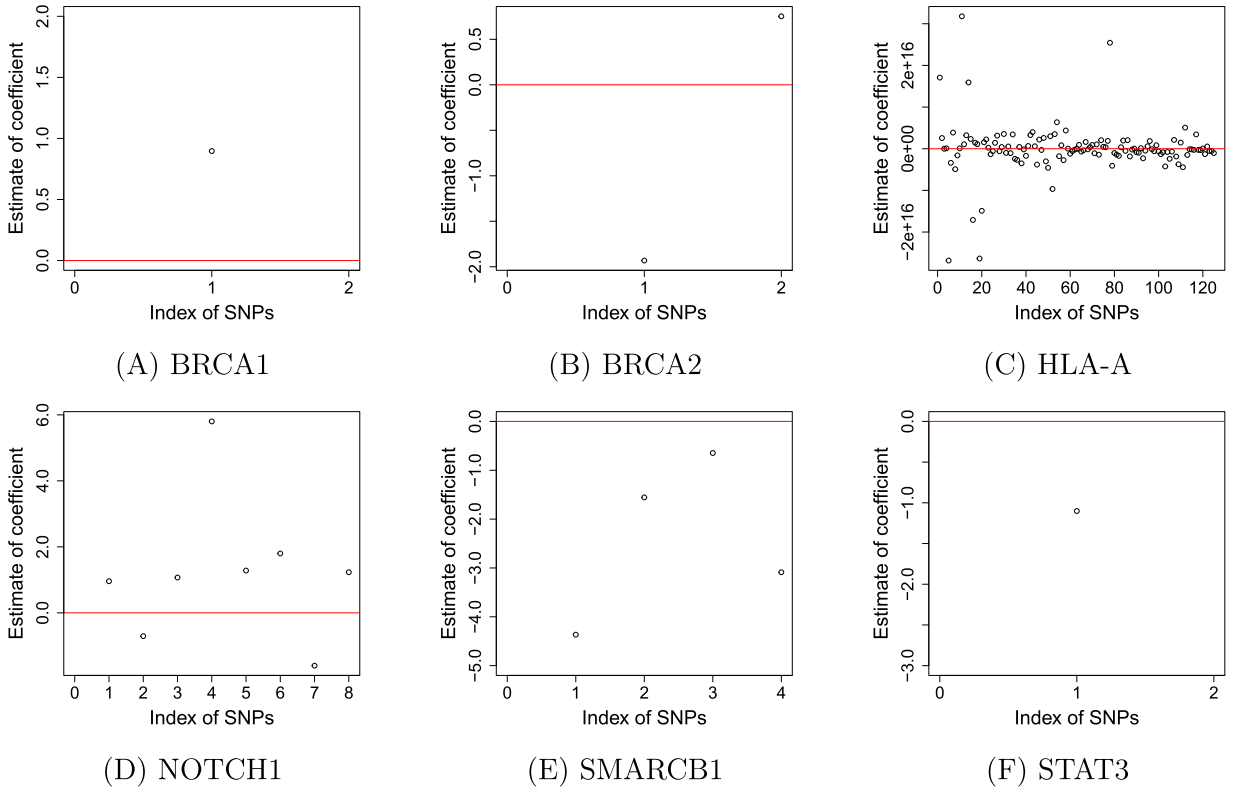
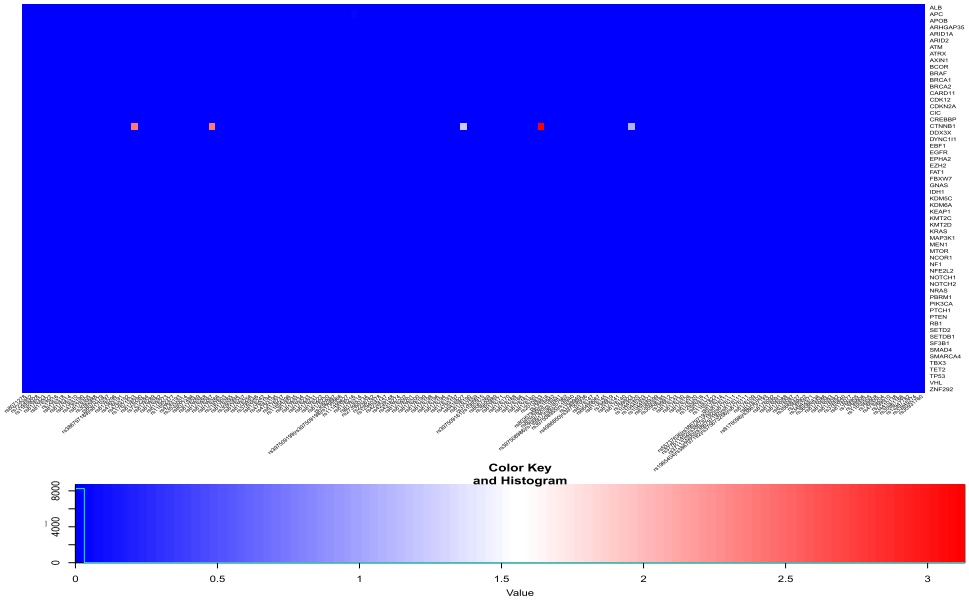


Fig. 8. Estimated coefficients β_j's for the SNPs with p-values less than 0.05 in a logistic regression model. The SNPs shown in each panel are listed below in ascending orders of SNP indices, together with the corresponding p-values:
(A) BRCA1: rs2070833 ($p = 3.86e-07$);
(B) BRCA2: rs76370881 ($p = 9.19e-05$), rs2320236 ($p = 1.95e-02$);
(C) HLA-A: 126 SNPs with p less than 0.05;
(D) NOTCH1: rs7856092 ($p = 3.57e-02$), rs6563 ($p = 2.01e-02$), rs75842134 ($p = 2.96e-02$), rs3124597 ($p = 9.83e-03$), rs3124600 ($p = 2.00e-02$), rs4077029 ($p = 4.99e-02$), rs9411208 ($p = 4.02e-02$), rs3124607 ($p = 3.17e-02$);
(E) SMARCB1: rs5760033 ($p = 1.18e-02$), rs738798($p = 1.74e-02$), rs738800 ($p = 2.92e-02$), rs2330624 ($p = 3.31e-02$);
(F) STAT3: rs62075782 ($p = 1.44e-05$).

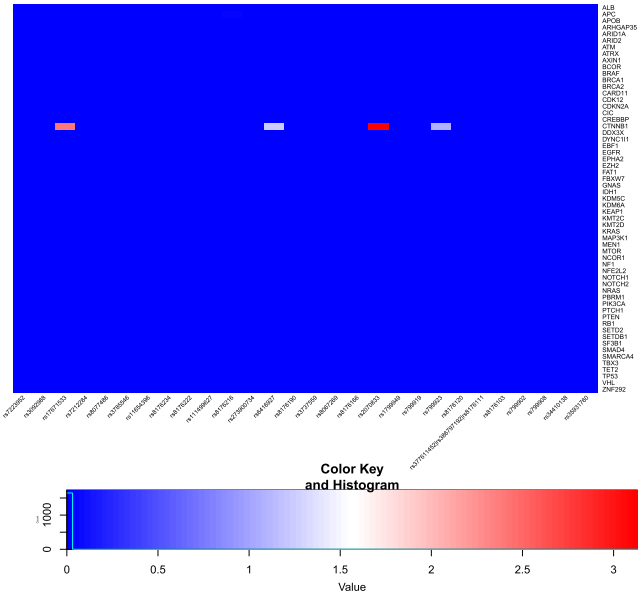
To verify the association patterns suggested by the effective γ parameters, we did the following. We used Fisher's exact tests to select the most significant SNPs (within a gene) that are associated with the somatic mutation of CTNNB1. After the false discovery rate (FDR) multiple testing correction, we kept the SNPs with FDR less than 0.05. These significant SNPs were then put into a logistic regression model as in (2). The estimated coefficients β_j's for the SNPs with p-values less than 0.05 in the model were plotted in Fig. 8. The results of association directions in the plots are consistent with what we predict using the effective γ values. For example, for genes NOTCH1 and SMARCB1, the association directions between most SNPs and the somatic mutation are the same. The corresponding effective γ is 1. For genes BRCA2 and HLA-A, the associations between most SNPs within each gene and the somatic mutation are in different directions. The corresponding effective γ is 2. For genes BRCA1 and STAT3, there is only one dominant SNP (rs2070833 for BRCA1 and rs62075782 for STAT3) that is associated with the somatic mutation and the corresponding effective γ is ∞.

6.3. Selection of important SNPs

The low-rank approximations to the SNP matrices X were also used to identify representative SNPs following the strategy in Section 4.2. The rank of the low-rank approximations is still r in (23). Through fast pivoting, we selected r representative SNPs from

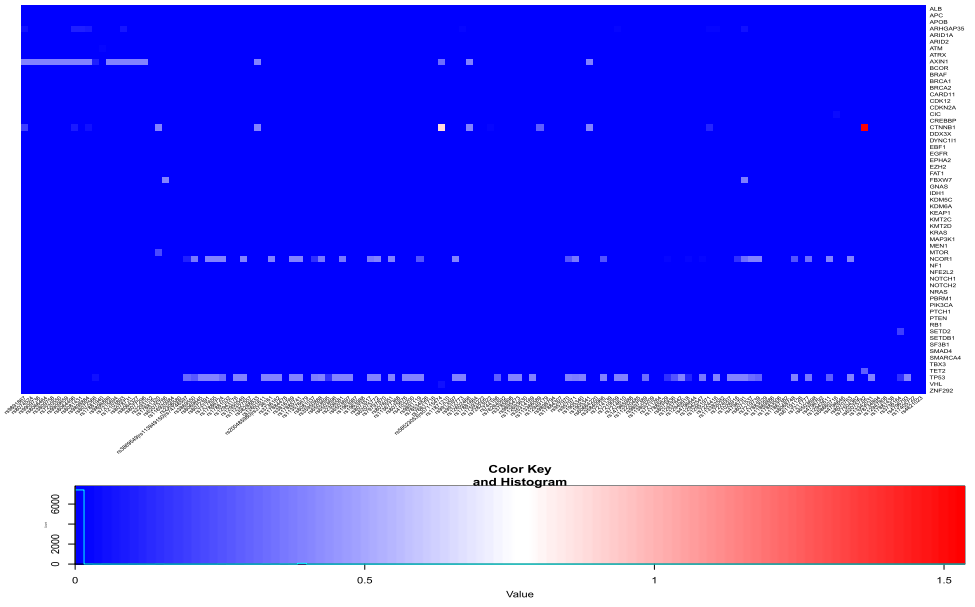


(A) BRCA1

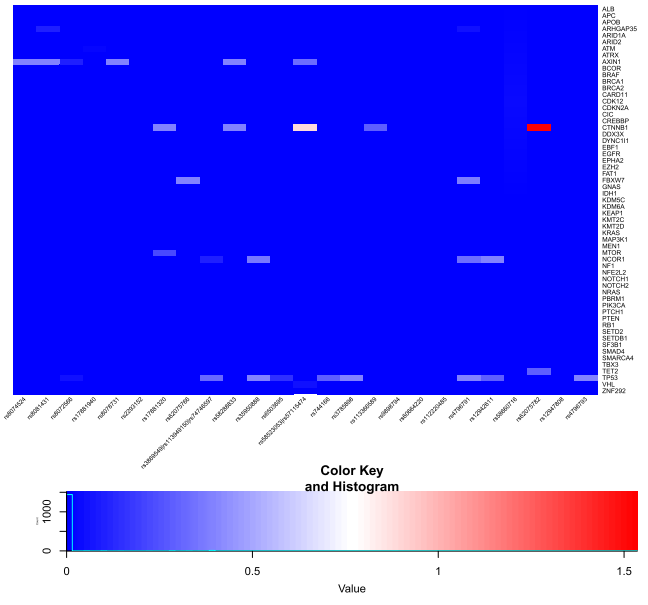


(B) Representative SNPs in BRCA1

Fig. 9. Heatmap of $-\log_{10}$(FDR) from Fisher's exact tests between germline variations of BRCA1 and somatic mutations of driver genes as in (A), and (B) is for a subset of (A) corresponding to the representative SNPs identified by fast pivoting, where the color range of the heatmaps is from blue to red with darker red indicating more significant associations, and the threshold corresponding to FDR = 0.05 is $-\log_{10}(0.05) \approx 1.3$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(A) STAT3



(B) Representative SNPs in STAT3

Fig. 10. Heatmap of $-\log_{10}$(FDR) from Fisher's exact tests between germline variations of STAT3 and somatic mutations of driver genes as in (A), and (B) is for a subset of (A) corresponding to the representative SNPs identified by fast pivoting, where the color range of the heatmaps is from blue to red with darker red indicating more significant associations, and the threshold corresponding to FDR = 0.05 is $-\log_{10}(0.05) \approx 1.3$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

a germline variation gene. At the same time, we performed Fisher's exact tests to study the associations between each SNP within the germline gene and the somatic mutations of driver genes (with mutation frequency $> 1\%$) and calculated the corresponding p-values. The p-values were adjusted by FDR multiple testing correction and the heatmaps of the FDR values were plotted in Figs. 9 and 10 for two genes as examples.

The representative SNPs identified by our fast pivoting strategy are roughly consistent with the significant SNPs given by Fisher's exact tests. For example, for the germline gene BRCA1 (Fig. 9), the significant SNPs that are strongly associated with the CTNNB1

somatic mutation picked by Fisher's exact tests are:

$$\text{rs2070833(the most significant one : } p\text{-value} = 3.3860e-8, \text{FDR-value} = 7.3949e-4),$$
$$\text{rs17671533, rs8176267, rs6416927, and rs799923.}$$

The representative SNPs identified by our fast pivoting approach include all the above SNPs except rs8176267. Interestingly, we find that the set of representative SNPs identified by our fast pivoting method typically includes the most significant SNP from Fisher's exact test. For example, for the germline gene STAT3 (Fig. 10), the most significant SNP from the Fisher's exact tests is rs62075782 (p-value= $1.4578e-6$, FDR-value= $2.9109e-2$) which is in the set of representative SNPs from our method.

To further demonstrate that we have extracted major genetic information, we performed the aSPU test with the representative SNPs of a gene and compared to the results of aSPU with all the SNPs in the gene. S in (25) was used in the tests. As shown in Fig. 11, the QQ plots indicate that the p-values from aSPU with the representative SNPs and with all the SNPs mostly align.

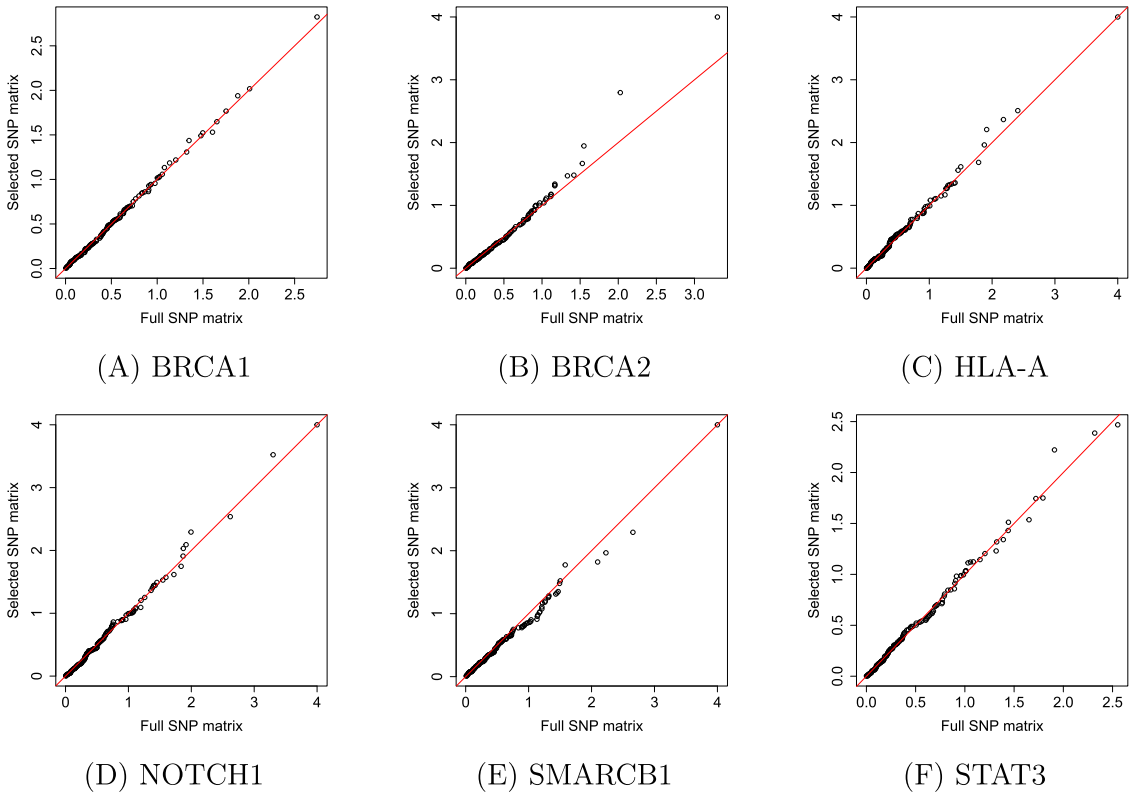


Fig. 11. QQ plots of $-\log_{10}(p\text{-values})$ from aSPU with all the SNPs and aSPU with the representative SNPs. To calculate the p-values, the number of permutations for both aSPU and aSPU-LR is $N = 10^4$.

Lastly, we can gain some biological insights from the representative SNPs. Even without the outcome information (somatic mutation), we were still able to identify some important SNPs. Such SNPs are likely driver SNPs that are significantly associated with most somatic mutations and may potentially contribute to cancers. For example, the BRAC1 SNP rs2070833 was reported to be the risk factor of endometrial carcinoma (Zheng et al., 2015) and triple-negative breast cancer (Zhao et al., 2021). The STAT3 SNP rs62075782 was reported to be a key SNP associated with cancer risks and susceptibility. This is useful in personalized cancer immunotherapy (Yu et al., 2007).

Remark 6.1. ICGC real-world analysis like in Chen et al. (2023) investigates the overall trend of the interactions between germline variations and somatic mutations across different cancer types. Such pan-cancer data analysis enjoys a couple of advantages such as the effective discovery of shared cancer mechanisms, high statistical powers, and cross-cancer therapeutic insights. On the other hand, these studies have heavy computational burden due to large matrix computations. Our low-rank approximation strategies open new opportunities to resolve this issue, making this pan-cancer data analysis both efficient and reliable.

7. Discussions

This work has shown a set of novel techniques for GWASs based on low-rank approximations of data matrices. In the aSPU-LR testing scheme, SNP matrices X were compressed via randomized SVDs to significantly reduce the cost of computing score vectors.

This overcomes a major efficiency bottleneck of aSPU tests when a large number of permutations are used for accurate p-value calculations. In the meantime, it keeps high statistical powers and similar type I error rates. Next, a parameter selection scheme was designed by using low-rank approximations to X to quickly identify some effective parameters in aSPU tests. Such effective parameters suggest important association patterns and are potentially useful for further accelerating association studies. Third, the low-rank approximations were also used to quickly extract important genetic information by identifying representative SNPs. This needs only little extra costs.

We showed the efficiency and the reliability of the proposed ideas through comprehensive simulation studies and ICGC real data analysis. The test method aSPU-LR has significant efficiency advantages over aSPU, even with modest data sizes. It also produces association results without compromising the reliability. By further using a small rank in the parameter selection process, we were able to quickly find the most effective parameters that suggest association patterns between some germline variations and somatic mutations. The procedure of identifying representative SNPs additionally enabled the extraction of biomarker information that would be unavailable to methods like PCA.

The research in this article provides several novel ideas of applying low-rank approximations to GWASs. We have utilized aSPU tests as an example, but the work further sheds light onto more general association studies. That is, low-rank techniques can serve as a valuable tool to accelerate association studies whenever there are extensive matrix operations, and the techniques further make it feasible to quickly extract important information that is otherwise too expensive to get.

The low-rank techniques also open opportunities to new research directions and further improvements of the ideas. For example, we may improve the SNP selection quality by replacing the Gram-Schmidt process with the method of Gu and Eisenstat (1996). We may further improve the efficiency by replacing the randomized SVD with a more efficient low-rank approximation method from Xia (2024). We would also like to incorporate the outcome/response vector information in the SNP selection process like in least angle regression (Efron et al., 2004), but with low-rank approximations to gain high efficiency. Additionally, the efficient association tests and information extraction make it possible to perform comprehensive studies of large datasets like ICGC so as to uncover many more biological findings. In particular, these low-rank approximation strategies are especially attractive for pathway-level association studies because of the large numbers of SNPs. This will appear in future work.

Code availability

Relevant codes are available from the following link: https://github.com/chenstatistics/lowrank.

Data availability

The ICGC data are provided at https://www.nature.com/articles/s41467-020-16785-6#data-availability.

Acknowledgement

We sincerely appreciate the valuable comments and suggestions from the editor and two anonymous referees which have greatly helped to improve the manuscript. The first author would also like to thank Professor Tao Wang from the Medical College of Wisconsin for some helpful discussions.

References

Barfield, R., Qu, C., Steinfelder, R.S., Zeng, C., Harrison, T.A., Brezina, S., Buchanan, D.D., Campbell, P.T., Casey, G., Gallinger, S., Giannakis, M., Gruber, S.B., Gsur, A., Hsu, L., Huyghe, J.R., Moreno, V., Newcomb, P.A., Ogino, S., Phipps, A.I. et al., 2022. Association between germline variants and somatic mutations in colorectal cancer. Sci. Rep. 12, 10207.

Carter, H., Marty, R., Hofree, M., Gross, A.M., Jensen, J., Fisch, K.M., Wu, X., Deboever, C., Nostrand, E. L.V., Song, Y., Wheeler, E., Kreisberg, J.F., Lippman, S.M., Yeo, G.W., Gutkind, J.S., Ideker, T., 2017. Interaction landscape of inherited polymorphisms with somatic events in cancer. Cancer Discov. 7 (4), 410–423.

Chen, H., Huffman, J.E., Brody, J.A., Wang, C. et al., 2019. Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. Am. J. Human Genetic. 104 (2), 260–274.

Chen, Z., Liang, H., Wei, P., 2023. Data-adaptive and pathway-based tests for association studies between somatic mutations and germline variations in human cancers. Genet. Epidemiol. 47 (8), 617–636.

Churchill, G.A., Doerge, R.W., 1994. Empirical threshold values for quantitative trait mapping. Genetics 130, 963–971.

Deng, Y., He, Y., Xu, G., Pan, W., 2022. Speeding up monte carlo simulations for the adaptive sum of powered score test with importance sampling. Biometrics 78, 261–273.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32 (2), 407–499.

Faye, L.L., Machiela, M.J., Kraft, P., Bull, S.B., Sun, L., 2013. Re-ranking sequencing variants in the post-gwas era for accurate causal variant identification. PLoS Genet. 9 (8), 1003609.

Golub, G., Van Loan, C., 2013. Matrix Computations, 4th Ed., Johns Hopkins University Press.

Gu, M., Eisenstat, S.C., 1996. Efficient algorithms for computing a strong rank-revealing qr factorization. SIAM J. Sci. Comput. 17, 848–869.

Halko, N., Martinsson, P.G., Tropp, J., 2011. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. 53, 217–288.

Kwak, I.-Y., et al., 2021. aSPU: Adaptive sum of powered score test. https://cran.r-project.org/web/packages/aSPU/index.html.

Liberty, E., Woolfe, E., Martinsson, P.G., Rokhlin, V., Tygert, M., 2007. Randomized algorithms for the low-rank approximation of matrices. Proc. Natl. Acad. Sci. USA 104, 20167–20172.

Ling, A.S., Hay, E.H., Aggrey, S.E., Rekaya, R., 2021. Dissection of the impact of prioritized QTL-linked and -unlinked SNP markers on the accuracy of genomic selection. BMC Genomic Data.

Loh, P., Tucker, G., Bulik-Sullivan, B.K., Finucane, B.J., Salem, H.K., Chasman, R.M., Ridker, D.I., Neale, P.M., Berger, B.M., Patterson, B., Price, N., A, L., 2015. Efficient bayesian mixed model analysis increases association power in large cohorts. Nat. Genet. 47 (3), 284–290. Vilhj1lmsson.

Mamidi, T. K.K., Wu, J., Hicks, C., 2019. Integrating germline and somatic variation information using genomic data for the discovery of biomarkers in prostate cancer. BMC Cancer 19, 229.
Pan, W., 2009. Asymptotic tests of association with multiple snps in linkage disequilibrium. Genet. Epidemiol. 33, 497–507.
Pan, W., Kim, J., Zhang, Y., Shen, X., Wei, P., 2014. A powerful and adaptive association test for rare variants 197, 1081–1095.
Pan, W., Kwak, I., Wei, P., 2015. A powerful pathway-based adaptive test for genetic association with common or rare variants. Am. J. Hum. Genet. 97, 86–98.
Pan, W., Shen, X., 2011. Adaptive tests for association analysis of rare variants. Genet. Epidemiol. 35, 381–388.
Ramroop, J.R., Gerber, M.M., Toland, A.E., 2019. Germline variants impact on somatic events during tumorigenesis. Trend Genetic. 35 (7), 515–526.
The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020. Pan-cancer analysis of whole genomes. Nature 578, 82–93.
Tippett, L. H.C., 1931. The Methods of Statistics. London, Williams and Norgate, Ltd.
Udell, U., Townsend, A., 2019. Why are big data matrices approximately low rank? SIAM J. Math. Data Sci. 1, 144–160.
Vali-Pour, M., Lehner, B., Supek, F., 2022. The impact of rare germline variants on human somatic mutation processes. Nat. Commun. 13, 3724.
Vosoughi, A., Zhang, T., Shohdy, K.S., Vlachostergios, P.J., Wilkes, D.C., Bhinder, B., Tagawa, S.T., Nanus, D.M., Molina, A.M., Beltran, H., Sternberg, C.N., Motanagh, S., Robinson, B.D., Xiang, J., Fan, X., Chung, W.K., Rubin, M.A., Elemento, O., Sboner, A., Faltas, B.M., 2020. Common germline-somatic variant interactions in advanced urothelial cancer. Nat. Commun. 11, 6195.
Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X., 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89, 82–93.
Xia, J., 2024. Making the nystr6m method highly accurate for low-rank approximations. SIAM J. Sci. Comput. 46, 1076–A1101.
Xia, J., Xi, Y., Gu, M., 2012. A superfast structured solver for toeplitz linear systems via randomized sampling. SIAM J. Matrix Anal. Appl. 33, 837–858.
Yu, H., Marcin, K., Drew, P., 2007. Crosstalk between cancer and immune cells: role of stat3 in the tumour microenvironment. Nature Rev. Immunol. 7 (1), 41–51.
Zhao, H., Feng, Y., Yang, J., 2021. The clinical feature of triple-negative breast cancer in Beijing. China. https://doi.org/10.1101/2021.08.03.21261573
Zheng, L., Song, A., Chen, L., Liu, D., Li, X., Guo, H., Tian, X., Fang, W., 2015. Association of genetic polymorphisms in aurka, brca1, ccne1 and cdk2 with the risk of endometrial carcinoma and clinicopathological parameters among chinese han women. Eur. J. Obstetric. Gynecol. Reproduct. Biol. 184, 65–72.