Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Bisection Grover's Search Algorithm and Its Application in Analyzing CITE-seq Data

Ping Ma, Yongkai Chen, Haoran Lu & Wenxuan Zhong

To cite this article: Ping Ma, Yongkai Chen, Haoran Lu & Wenxuan Zhong (2025) Bisection Grover's Search Algorithm and Its Application in Analyzing CITE-seq Data, Journal of the American Statistical Association, 120:549, 52-63, DOI: 10.1080/01621459.2024.2404259

To link to this article: https://doi.org/10.1080/01621459.2024.2404259

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 120, NO. 549, 52–63: Applications and Case Studies
https://doi.org/10.1080/01621459.2024.2404259

Taylor & Francis
Taylor & Francis Group

Bisection Grover's Search Algorithm and Its Application in Analyzing CITE-seq Data

Ping Ma*, Yongkai Chen*, Haoran Lu, and Wenxuan Zhong

Department of Statistics, University of Georgia, Athens, GA

ABSTRACT

With the rapid development of quantum computers, researchers have shown quantum advantages in physics-oriented problems. Quantum algorithms tackling computational biology problems are still lacking. In this article, we demonstrate the quantum advantage in analyzing CITE-seq data. CITE-seq, a single-cell technology, enables researchers to simultaneously measure expressions of RNA and surface protein detected by antibody-derived tags (ADTs) in the same cells. CITE-seq data hold tremendous potential for identifying ADTs associated with targeted genes and identifying cell types effectively. However, both tasks are challenging since the best subset of ADTs needs to be identified from enormous candidate subsets. To surmount the challenge, we develop a quantum algorithm named bisection Grover's search (BGS) for the best subset selection of ADT markers in CITE-seq data. BGS takes advantage of quantum parallelism by integrating binary search and Grover's algorithm to enable fast computation. Theoretical results are provided to show the privilege of BGS in the estimation error and computational complexity. The empirical performance of the BGS algorithm is demonstrated on both the IBM quantum computer and simulator. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received November 2023
Accepted September 2024

KEYWORDS
Binary search; Curse of dimensionality; Grover's algorithm; Quantum computing; Single-cell; Variable selection

1. Introduction

Recent breakthroughs in quantum computers have shown quantum advantage (aka quantum supremacy), that is, quantum computers outperform the classical computers for solving specific problems (Shor 1999; Arute et al. 2019; Zhong et al. 2020; Wu et al. 2021). More importantly, there are already general-purpose programmable quantum computing devices available to the public, for example, IBM Quantum Experience, Microsoft Quantum, and Amazon Braket. Such quantum devices are commonly known as noisy intermediate-scale quantum (NISQ) devices (Preskill 2018). It is reported that NISQ with a thousand qubits has been released (Castelvecchi 2023). Despite the success of quantum computers, the investigated problems are highly physics-oriented and may not necessarily appeal to researchers in other fields, for example, statistics and data science (Wang 2022). The key challenge in leveraging quantum advantages in these fields is identifying practical applications where the integration of statistical analysis and quantum computing can effectively overcome computational bottlenecks.

In this article, we develop a highly versatile quantum algorithm named Bisection Grover's search (BGS) for best model selection, which exhibits excellent performance and computational efficiency in its applications, specifically in the analysis of CITE-seq data. BGS takes advantage of quantum parallelism by integrating binary search and Grover's algorithm to enable fast computation. We show that the proposed BGS is consistent as long as sufficient quantum bits are available. We also show that the BGS algorithm is nearly quadratic speed-up of classical algorithms in terms of time complexity. In addition, we provide the code that implements the proposed algorithm in the IBM quantum computer and the IBM quantum simulator with practical guidance.

Brief context and background. The advantages of quantum computing rely on the fact that quantum bits carry more information than classical bits. Different from classical computers built on classical bits having a *state* of either 0 or 1, quantum computers operate on quantum processing units, quantum bits (or qubits), which can be in a state 0, 1, or both simultaneously due to the superposition property (Nielsen and Chuang 2010). The p qubits create 2^p different states for the system that are superposed on each other. The superposition enables researchers to perform computations using all of those 2^p states simultaneously, which is also called quantum parallelism. Quantum parallelism circumvents the time/space tradeoff of classical parallel computing through its ability to hold exponentially many units of information in a linear amount of physical space. In addition, a quantum computer has some logic gates, which a classical computer does not have, enabling faster computation than a classical computer (Nielsen and Chuang 2010).

Despite the impressive achievements made possible by quantum computers, quantum algorithms for solving statistical or data science problems are still lacking. This deficiency primarily stems from the fact that many classical optimal algorithms are the culmination of intellectual efforts by generations of scientists. Developing a quantum algorithm that outperforms the classical optimal algorithms is intellectually challenging. In

CONTACT Ping Ma pingma@uga.edu; Wenxuan Zhong wenxuan@uga.edu Department of Statistics, University of Georgia, Athens, GA.

*These authors contributed equally to this work.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

particular, there are significant technical obstacles in developing such quantum algorithms. First, even though quantum computing enjoys quantum parallelism, accessing the result is not straightforward. Taking measurements of a quantum system induces a superposition of quantum states collapsing into one state in a certain probability and permanently changing the state of the system. Second, many quantum algorithms often depend on an oracle function to recognize if an outcome is a solution or not. For example, Grover's algorithm requires an oracle function that can recognize solutions. However, such an oracle function is usually not available in many statistics or data science problems. Third, extensive efforts have been devoted on analyzing the algorithmic aspect of quantum algorithms. In spite of these impressive algorithmic results, the theoretical analysis that addresses statistical aspects of quantum algorithms is still lacking. Without theoretical statistics insight, researchers may not effectively develop new quantum algorithms to tackle data science problems in a principal statistics framework.

To overcome these obstacles, we develop a BGS algorithm for the analysis of CITE-seq data. CITE-seq (Cellular Indexing of Transcriptomes and Epitopes by sequencing) is an innovative technology for studying single-cell biology (Stoeckius et al. 2017). Different from single-cell RNA-seq measuring RNA expression alone, CITE-seq enables researchers to simultaneously measure RNA and surface protein expression in the same cells. In particular, in CITE-seq, surface proteins detected by antibody-derived tags (ADTs) are transcriptomically profiled alongside RNA transcripts using single-cell RNA-seq. By using CITE-seq data of cells, we can study the following two biological problems: (a) identifying ADTs associated with the targeted genes and (b) designing panels of markers for cell type identifications. Successfully tackling these two problems hinges on the effective selection of the best subset of ADTs. The best ADT subset selection problem is generically hard with classical computing since it is computationally demanding and time-consuming due to the need to evaluate an extensive number of candidate subsets. Even for 30 ADTs, that is, $p = 30$, we have more than one billion candidate subsets, which poses significant challenges for classical algorithms to select the best subset.

Identification of ADTs associated with marker genes of interest. Despite the fact that RNAs and proteins are produced from the same genes, they provide some complementary information on cell states due to post-transcriptional and post-translational gene regulation. Since the CITE-seq data is often of high dimension and contains lots of redundant information across different RNAs and ADTs, selecting a parsimonious set of RNAs and ADTs is crucial to leverage the information of the CITE-seq data. We aim to develop a method to identify a set of ADTs that best explains the variations of the expressions of an RNA in a regression model.

Panel design for cell type identification. Efficiently identifying immune cell types is a key objective in immunological research and clinical diagnostics. A recent study finds that CITE-seq can effectively identify compact sets of immunophenotypic markers (ADTs) for characterizing different cell types (Hao et al. 2021). Such a set of markers is referred to as a panel. Including too many markers in a panel can result in a decreased signal-to-noise ratio and increased background noise in downstream experimental validation, for example, the spillover-spreading error inherent to the fluorescence in flow cytometry (Ferrer-Font et al. 2021). Here, we identify the best subset of ADTs for each cell type using logistic regressions by setting the cell type of interest as one and other cell types as zero. The resulting panel of ADTs can improve accuracy in identifying immune cell populations while optimizing resource allocation.

Statistically, our problem can be described as follows: CITE-seq yields a sample with RNA expressions and p ADT expressions $x_{i,1}, x_{i,2}, \ldots, x_{i,p}$ in the ith cell, where $i = 1, \ldots, n$ and n is the number of cells. In the problem of identifying the ADTs associated with the targeted gene, we model the expressions of RNA and ADTs through the linear regression model. Given a specific gene of interest, we represent its expression as y_i in the ith cell, serving as the response variable in the regression model. We assume that only a subset, denoted by $\mathcal{A}$, of ADTs co-express with the gene of interest and are used as the predictor variables. In the panel design for the cell identification problem, we consider the logistic regression model. The cell type of the ith cell is denoted by z_i, which is either one if it is a cell of interest, or zero otherwise. Once again, we assume that only a subset, denoted by $\mathcal{A}$, of ADTs is the marker for the cell type of interest. The primary research interest is identifying the subset $\mathcal{A} \subseteq \{1, \ldots, p\}$ effectively and efficiently. Here, we choose the subset $\mathcal{A}$ using the Bayesian information criterion (BIC) (Schwarz 1978), which is given by

$$\mathrm{BIC}(\mathcal{A}) = |\mathcal{A}| \log n - 2 \log L(\mathcal{A}), \tag{1}$$

where $|\mathcal{A}|$ is the subset size, n is the number of cells in the training dataset, $L(\mathcal{A})$ is the maximized value of the likelihood function of the fitted model for the corresponding subset.

In the literature, many quantum-based feature and model selection methods have been explored. He et al. (2018) leverages the searching abilities of Grover's algorithm and proposes the quantum versions of the forward selection and backward elimination algorithm, achieving quadratic speedup for each step of addition or deletion. Chakraborty et al. (2020) and Li et al. (2022) explore the quantum benefits of graph theory to solve the graph-theoretic feature selection problems. Recently, the promise of high efficiency for solving the quadratic unconstrained binary optimization (QUBO) with quantum optimizers has sparked widespread interest in its study, leading to a surge in research investigating quantum-based feature selection within the QUBO model (Von Dollen et al. 2021; Turati, Dacrema, and Cremonesi 2022; Mücke et al. 2023). Nonetheless, the quantum algorithm specifically designed for the aforementioned global optimization searching problems is still lacking. Furthermore, most existing researches focus on conceptual and theoretical exploration. The practical implications of leveraging quantum computing to enhance our understanding of single-cell multi-omics studies remain largely unexplored.

To overcome the aforementioned limitations, we propose BGS to select the best subset. In this method, we randomly choose a subset as our benchmark subset. This benchmark subset bisects all subsets into two partitions, an oracle set consisting of subsets having smaller BICs than that of the benchmark subset, and a non-oracle set consisting of rest subsets. BGS starts with an initial superposition where all candidate subsets are encoded with equal weights. BGS then iteratively updates the superposition toward the oracle set and yields a new superposition. The new superposition is then measured and collapses to a new subset. The BIC of the new subset is compared with that of the benchmark set. If the new subset is better than the benchmark set, the new one replaces the benchmark one. We then repeat the above procedures until no more replacements can be made. Different from Grover's algorithm, our BGS algorithm does not require any oracle function to recognize the best subset.

Our methodological contributions are as follows. First, we take advantage of quantum parallelism by integrating binary search and Grover's algorithm to design a new iteration algorithm and only access the outcome at the end of each iteration. In this way, we can greatly harness the power of quantum computing. Second, our proposed algorithm does not require an oracle function to recognize the best subset. This relaxation expands the scope of applications and overcomes the obstacle of the impractical assumption that the best subset needs to be known a priori in Grover's algorithm. Third, we integrate Grover's algorithm with the quantum counting algorithm to get a high probability of identifying the best subset.

Our theoretical contribution is that we derive the error bound for the proposed BGS algorithm and show that the error can be controlled to be arbitrarily small as long as we have a sufficient number of qubits. It is the consequence of powerful quantum parallelism and accurate quantum counting. Moreover, we derive the time and space complexity of the proposed algorithm.

Our empirical contribution is that we conduct the empirical analysis in an IBM quantum computer and IBM quantum simulator. Our empirical results are consistent with our theoretical analysis and demonstrate the quantum advantage of the proposed algorithm over classical algorithms. Our exploration facilitates future quantum algorithm developments targeting other biological problems.

2. Methods

2.1. Overview

To select the best subset $\mathcal{A}$ that minimizes (1), one possible quantum computing method we could use is Grover's algorithm (Grover 1996). Grover's algorithm is a quantum search algorithm. The intuition behind Grover's algorithm lies in its ability to leverage quantum superposition to enhance the search process. Grover's algorithm initializes the qubits in a superposition of all possible states, effectively exploring multiple search paths simultaneously. Grover's algorithm then employs a technique known as amplitude amplification (Brassard et al. 2002) to enhance the probability of finding the desired solution. This technique involves iteratively applying a sequence of quantum operations to amplify the amplitude of the target state while suppressing the amplitudes of other states. This approach leads to a quadratic speedup compared to classical search algorithms, making it highly efficient for large-scale search problems. However, there are three key difficulties in applying Grover's algorithm to our problem. First, in our problem, we do not know the oracle state priori. Consequently, it is not clear how to design Grover's operation. Second, even if we know the oracle state in Grover's algorithm, the number of Grover's rotations needs to be determined accurately. Otherwise, Grover's algorithm may not be as effective as what algorithmic analysis prescribes. Third, theoretical analysis, including error quantification and algorithmic complexity, for taking into account the above two difficulties and uncertainties is still lacking. The details of Grover's algorithm are relegated to Section A.2 in the supplementary material.

We shall now develop a novel quantum search algorithm named bisection Grover's search (BGS) to address these difficulties. Recall that all subsets of $\{1, \ldots, p\}$ naturally correspond to a state in the orthonormal basis $\mathcal{B} = \{|b_0\rangle, \ldots, |b_{D-1}\rangle\}$ of a Hilbert space $\mathcal{H}$. Notice that only p qubits are sufficient to represent these states. We define a state loss function $g(\cdot) : \mathcal{H} \to \mathbb{R}$,

$$g\left(|b_j\rangle\right) \equiv \mathrm{BIC}(\mathcal{A}_j), \tag{2}$$

where state $|b_j\rangle \in \mathcal{B}$ is a vector in $\mathcal{H}$ that corresponds to the subset $\mathcal{A}_j$. Hence, the best model corresponds to the state that minimizes the state loss function. We assume there exists a sole oracle state in $\mathcal{B}$ that minimizes $g(\cdot)$. Suppose the oracle state denoted by $|b^\star\rangle$ is unique and satisfies

$$g(|b^\star\rangle) < g(|b_j\rangle) \quad \text{for } |b_j\rangle \in \mathcal{B} \text{ and } |b_j\rangle \neq |b^\star\rangle \,. \tag{3}$$

The idea of the proposed BGS algorithm can be sketched as follows. Since the oracle state is unknown, the BGS algorithm randomly selects a state and regards all the states with smaller BICs than that of the selected state as oracle states. In particular, BGS randomly selects a state in $\mathcal{B}$ and calculates its BIC. We name this state the *benchmark state*. This benchmark state bisects the set of all states into two subsets: a subset of *oracle states* and a subset of *non-oracle states*. Oracle states are those states having smaller BICs than the benchmark state, and non-oracle states are all other states. BGS then rotates the uniform superposition of all states via Grover's operations toward the superposition of the oracle states. Once taking the measurement, the rotated superposition collapses to one of the oracle states. If the collapsed state's BIC is smaller than that of the benchmark state, the benchmark state is replaced by it. BGS then iterates the above steps. Otherwise, we output this collapsed state.

The quantum circuits for evaluation of the state loss function $g(\cdot)$ in all D states are illustrated in Figure 1(a) and (b). The operation $\mathbf{S}_\mathbf{X}$ encodes the ADTs $\mathcal{A}_j$ when the input state is $|b_j\rangle$. The operation $\mathbf{S}_{\boldsymbol{y}}$ encodes the response $\boldsymbol{y}$. The state $|\mathbf{X}^{\mathcal{A}_j}\rangle$ encodes the matrix $\mathbf{X}^{\mathcal{A}_j}$ as suggested by Schuld, Sinayskiy, and Petruccione (2016). The operation $\mathbf{U}_g$ computes the BIC value $g(|b_j\rangle)$ with the input $\mathbf{X}^{\mathcal{A}_j}$ and the encoding of the response $|\boldsymbol{y}\rangle$. The detailed description of these quantum circuits is relegated to Section B in supplementary material.

Remark. $\mathbf{U}_g$ can be implemented with quantum speedup. Notice that quantum computing methods for regression and classification offer exponential or polynomial speed-up compared to their classical counterparts (Biamonte et al. 2017), for example, $O(\log(n))$ complexity for fitting linear regression models compared to $O(n)$ complexity in classical computing methods (Schuld, Sinayskiy, and Petruccione 2016) for linear regression.

Remark. The inherent advantages of quantum parallelism enable the efficient evaluation of the state loss function $g(\cdot)$ in

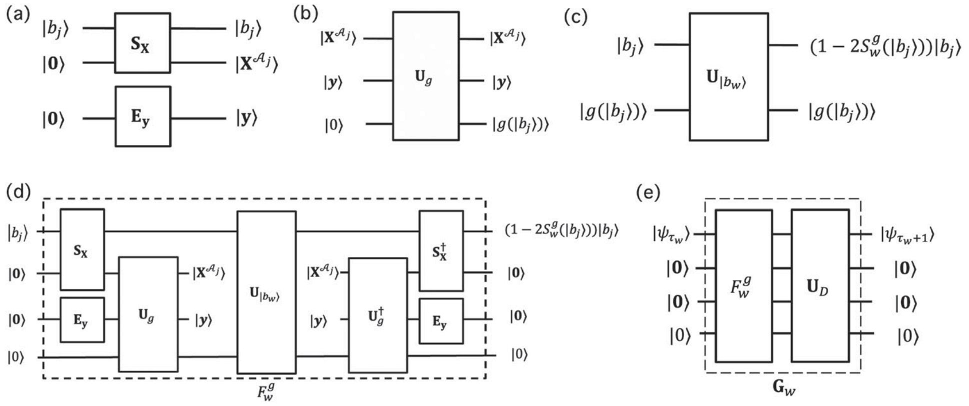

Figure 1. The quantum circuits for illustration. (a) The circuit encodes the ADT subset. (b) The circuit to calculate the state loss function $g(\cdot)$. (c) The circuit to flip the sign of oracle states. (d) The assembled circuit for the operation F^g_w. (e) The circuit for the operation $\mathbf{G}_w$ in the BGS algorithm.

all D states. That is, all D values can be computed simultaneously in a single operation by leveraging the power of quantum computing.

We shall now present the proposed method in detail.

2.2. Bisection and Grover's Operations

BGS randomly chooses an initial benchmark state $|b_w\rangle$ from the set $\mathcal{B}$. Let us define a local evaluation function $S^g_w(|b_j\rangle)$ as

$$S^g_w(|b_j\rangle) = \begin{cases} 1, & \text{if } g(|b_j\rangle) < g(|b_w\rangle), \\ 0, & \text{otherwise.} \end{cases} \qquad (4)$$

Using this local evaluation function, $|b_w\rangle$ bisects basis set $\mathcal{B}$ into two subsets: the subset of oracle states: $\mathcal{B}_w = \{|b_j\rangle : S^g_w(|b_j\rangle) = 1\}$ and the subset of non-oracle states: $\mathcal{B}^c_w = \{|b_j\rangle : S^g_w(|b_j\rangle) = 0\}$. If $\mathcal{B}_w$ is a null set, that is, $|b_w\rangle$ happens to be the unique oracle state $|b^\star\rangle$, BGS can output $|b_w\rangle$ as the final result. If $\mathcal{B}_w$ is not empty, BGS proceeds as follows. Let D_w be the cardinality of $\mathcal{B}_w$, that is, the number of states in $\mathcal{B}_w$. BGS initializes a uniform superposition as

$$|\psi_0\rangle = \frac{1}{\sqrt{D}} \sum_{j=0}^{D-1} |b_j\rangle \equiv \sqrt{\frac{D_w}{D}}\,|\phi_w\rangle + \sqrt{\frac{D-D_w}{D}}\,|\zeta_w\rangle, \qquad (5)$$

where $|\phi_w\rangle = \frac{1}{\sqrt{D_w}} \sum_{j\in\mathcal{B}_w} |b_j\rangle$ and $|\zeta_w\rangle = \frac{1}{\sqrt{D-D_w}} \sum_{j\in\mathcal{B}^c_w} |b_j\rangle$. Notice that $|\phi_w\rangle$ is the linear combination of the oracle states, whereas $|\zeta_w\rangle$ is the linear combination of the non-oracle states. Define as θ_w the angle between $|\psi_0\rangle$ and $|\zeta_w\rangle$. By the definition, we have

$$\sin(\theta_w) = \sqrt{D_w/D}. \qquad (6)$$

Since $|\phi_w\rangle$ and $|\zeta_w\rangle$ are orthonormal, these two superpositions form a two-dimensional Hilbert space. This two-dimensional Hilbert space can be represented as the column space of a $2^p \times 2$ matrix $\mathbf{\Gamma}_w$, where the two columns are the corresponding vector representations of $|\phi_w\rangle$ and $|\zeta_w\rangle$. The initial state vector then can be written as $|\psi_0\rangle = \sin\theta_w\,|\phi_w\rangle + \cos\theta_w\,|\zeta_w\rangle$. We now rotate this state vector toward $|\phi_w\rangle$. Analogous to Grover's algorithm, we define a flip operation F^g_w, where $F^g_w\,|b_j\rangle = \left(1 - 2S^g_w(|b_j\rangle)\right)|b_j\rangle$. The implementation details of the flip operation are illustrated in Figure 1(c) and (d). The operation $|\mathbf{U}_{|b_w\rangle}\rangle$ flips the sign of the input state $|b_j\rangle$ if $g(|b_j\rangle) < g(|b_w\rangle)$, that is, $S^g_w(|b_j\rangle) = 1$. Consequently, the flip operation can be constructed by assembling the operations for evaluation of the state loss function $g(\cdot)$ and $|\mathbf{U}_{|b_w\rangle}\rangle$. Notice that to avoid the unnecessary entanglement created by qubits encoding $|\mathbf{X}^{\mathcal{A}_j}\rangle$s and $|g(|b_j\rangle)\rangle$, we use a standard approach by uncomputing the evaluation of the state loss function with the conjugate transpose of the operations $\mathbf{S_X}$ and $\mathbf{U}_g$, that is, $\mathbf{S}^\dagger_{\mathbf{X}}$ and $\mathbf{U}^\dagger_g$, respectively.

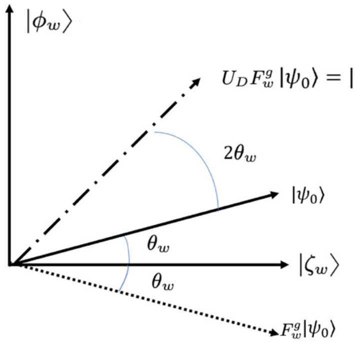

Figure 2. Geometrical illustration of the two operations in the first rotation step of the BGS algorithm. After taking the diffusion operation $\mathbf{U}_D$ and the flip operation F^g_w, the initial state vector $|\psi_0\rangle$ moves to $|\psi_1\rangle$, which is closer to the oracle states $|\phi_w\rangle$.

The diffusion operation $\mathbf{U}_D$ remains the same as the one defined in (A.8). We thus have Grover's operation (with respect to $|\phi_w\rangle$) $\mathbf{G}_w = \mathbf{U}_D F^g_w$. Figure 2 provides a visualization of the two operations in the first rotation of BGS. It is easy to derive that $\mathbf{G}_w$ can be decomposed as follows,

$$\mathbf{G}_w = \mathbf{\Gamma}_w \begin{pmatrix} \cos 2\theta_w & \sin 2\theta_w \\ -\sin 2\theta_w & \cos 2\theta_w \end{pmatrix} \mathbf{\Gamma}^\top_w + \tilde{\mathbf{\Gamma}}_w \tilde{\mathbf{\Gamma}}^\top_w, \qquad (7)$$

where $\tilde{\mathbf{\Gamma}}_w$ is a $2^p \times (2^p - 2)$ matrix of which columns are orthonormal and orthogonal with $\mathbf{\Gamma}$. The detailed derivation can be found in sec. 6 of Nielsen and Chuang (2010).

After we apply Grover's operation τ_w times to it, $|\psi_0\rangle$ becomes

$$|\psi_{\tau_w}\rangle = \sin\left((2\tau_w+1)\theta_w\right)|\phi_w\rangle + \cos\left((2\tau_w+1)\theta_w\right)|\zeta_w\rangle. \qquad (8)$$

As long as $(2\tau_w + 1)\theta_w \le \pi/2$, BGS gradually amplifies the amplitudes of the oracle states and suppresses the amplitudes of non-oracle states. Notice that all the states' amplitudes are updated simultaneously in this procedure, resulting in highly efficient searching.

Once a measurement is taken, $|\psi_{\tau_w}\rangle$ collapses to a state $|b_{w^{new}}\rangle \in \mathcal{B}_w$. The probability that $|b_{w^{new}}\rangle = |b_j\rangle$ is

$$\mathbb{P}(|b_{w^{new}}\rangle = |b_j\rangle) = \begin{cases} \frac{\sin^2((2\tau_w+1)\theta_w)}{D_w}, & \text{if } S_w^g(|b_j\rangle) = 1, \\ \frac{\cos^2((2\tau_w+1)\theta_w)}{D-D_w}, & \text{otherwise.} \end{cases} \qquad (9)$$

Now the output is $|b_{w^{new}}\rangle = |b_j\rangle$. If $S_w^g(|b_j\rangle) = 0$, BGS does not update benchmark state $|b_w\rangle$. Otherwise, that is, $S_w^g(|b_j\rangle) = 1$, BGS replaces benchmark state $|b_w\rangle$ by $|b_{w^{new}}\rangle$, that is, updates $|b_w\rangle$ to an oracle state.

The quantum circuit for implementing Grover's operation $\mathbf{G}_w$ is visualized in Figure 1(e).

2.3. Estimating the Number of Grover's Operations

BGS only updates the benchmark state if the superposition collapses to an oracle state. By (9), the probability that the superposition collapses to an oracle state is the highest if $(2\tau_w + 1)\theta_w = \pi/2$. Thus, the efficiency of BGS highly depends on the choice of τ_w, the number of Grover's operations. Since θ_w is also unknown, a proper choice of τ_w relies on an accurate estimate of θ_w. Note that we cannot estimate θ_w from (6) since $\mathcal{B}_w$ is also unknown.

We propose to estimate θ_w using the quantum counting algorithm (Brassard, Høyer, and Tapp 1998). Since the presentation of the quantum counting algorithm is lengthy, we relegate the details to Section A.3 in supplementary material. The key idea of the quantum counting algorithm is sketched as follows. Since $0 \leq \theta_w < \pi$, the T-digit binary approximation of θ_w is $\theta_w/\pi \approx 0.e_1e_2\cdots e_T = \sum_{k=1}^{T} e_k 2^{-k}$, where e_k is either 0 or 1. The quantum counting algorithm employs a divide-and-conquer strategy by dividing the problem of estimating θ_w into T sub-problems of estimating each digit e_k separately for $k = 1, \ldots, T$ with T additional (auxiliary) qubits. The estimates of e_k for $k = 1, \ldots, T$ are then summed together yielding the estimate of θ_w. We refer to the T auxiliary qubits as the *rotation register*. Technically, the information of θ_w is first passed on to each auxiliary qubit in the rotation register. The amplitude of the uniform superposition in each auxiliary qubit is then varied with respect to θ_w so that the probability of measuring the corresponding e_k is amplified.

Passing the information of θ_w to an auxiliary qubit in the rotation register can be achieved via the *controlled Grover operation*. The controlled Grover operation acts on the joint system of one auxiliary qubit in the rotation register and the *BGS register* that consists of the qubits for running BGS. In particular, the auxiliary qubit in the rotation register is initialized at the uniform superposition, that is, $\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. Applying the controlled Grover operation is equivalent to applying the Grover operation to the BGS register only if the auxiliary qubit in the rotation register is at the state $|1\rangle$. In this way, a single controlled Grover operation rotates the auxiliary qubit by angle $2\theta_w$ (see Section A.1 in supplementary material for the technical details). Repeatedly applying different numbers of the controlled Grover operations is needed to solve different sub-problems. For example, applying 2^{k-1} controlled Grover operations is needed to estimate e_k. The inverse quantum Fourier transform (QFT) is then applied to each auxiliary qubit, the measurement of which is the solution of the corresponding sub-problem with a high probability. Finally, the measurement of the joint system of the rotation register and the BGS register is used to estimate θ_w.

Using the quantum counting algorithm, we get the estimate of θ_w, denoted by $\hat{\theta}_w$. By (6), we estimate the number of oracle states via $\hat{D}_w = [D\sin^2(\hat{\theta}_w)]$, where $[\cdot]$ denotes rounding to the nearest integer. Hence, the estimate of the number of iterations is

$$\hat{\tau}_w = \left[\frac{\pi}{4\arcsin\left(\sqrt{\hat{D}_w/D}\right)} - \frac{1}{2} \right]. \qquad (10)$$

The estimation error of the quantum counting algorithm is established in the following lemma. The proof of this lemma is relegated in Section A.3 in supplementary material.

Lemma 2.1. Let $\hat{\theta}_w$ be the estimate of the θ_w using the quantum counting algorithm with T qubits in the rotation register. Assuming that $\theta_w \in \left(0, \frac{\pi}{4}\right)$. For any $\varepsilon \in \left(0, \frac{1}{4}\right)$, we have

$$\mathbb{P}\left(\frac{1}{\pi}\left|\hat{\theta}_w - \theta_w\right| > \varepsilon + \frac{1}{2^T}\right) < \frac{1}{2^{T+1}\varepsilon}. \qquad (11)$$

The computational complexity of this quantum counting algorithm is $O\left(p2^T\right)$.

Remark. The condition of $\theta_w \in \left(0, \frac{\pi}{4}\right)$ implies that there are less than half of the states have smaller BICs than the benchmark state. To satisfy this condition, we can take a simple approach to choose the benchmark state. Instead of randomly selecting one state as the benchmark state, we randomly select several states and choose the one with the smallest BIC as the benchmark state. For example, if m states are randomly selected, then the condition is satisfied with the probability $1 - \frac{1}{2^m}$, which is close to 1 for a moderate m.

Notice that the computational complexity of the quantum counting algorithm is of linear order in p and exponential order in T. Moreover, we will show that the computational cost of the quantum counting algorithm is of the lower order compared to the overall computational cost of the BGS algorithm.

2.4. BGS Algorithm

The details of BGS are summarized in Algorithm 1.

Note that at the beginning of each iteration, that is, (b1) in Algorithm 1, the quantum state is always initialized at the uniform superposition of all states specified in (5). This is due to the no-cloning theorem (Wootters and Zurek 1982), which states the impossibility of creating an independent and identical copy of an arbitrary unknown quantum state. This feature distinguishes the iterative algorithms in quantum computers from those on classical computers.

Furthermore, iterations in the BGS algorithm are terminated if $\sin(\hat{\theta}_w)$ is less than an error tolerance δ. By (6), the value of $\sin(\theta_w)$ dictates the number of oracle states D_w. If $\sin(\theta_w) = 0$, there is no oracle state, that is, $D_w = 0$. Hence,

Algorithm 1 Bisection Grover's search

```
Input: An orthonormal basis set $\mathcal{B}$ of size $D = 2^p$, a state loss function $g(\cdot)$ that maps a state in $\mathcal{B}$ defined in (2), initial benchmark state $|b_w\rangle$, and error tolerance $\delta$.
Initialization: Define a local evaluation function $S_w^g(\cdot)$ such that $S_w^g(|b_j\rangle) = 1$ if $g(|b_j\rangle) < g(|b_w\rangle)$ and $S_w^g(|b_j\rangle) = 0$ if $g(|b_j\rangle) \geq g(|b_w\rangle)$.
(a). Use the quantum counting algorithm to get the initial estimates $\hat{\theta}_w$ and $\hat{\tau}_w$.
repeat
    repeat
        (b1). Repeatedly apply Grover's operations defined in (7) for $\hat{\tau}_w$ times.
        (b2). Measure the quantum register and denote the measurement result by $|b_{w^{new}}\rangle$.
    until $S_w^g(|b_{w^{new}}\rangle) = 1$.
    (c). Set $|b_w\rangle = |b_{w^{new}}\rangle$ and update $S_w^g(\cdot)$ accordingly.
    (d). Use the quantum counting algorithm to get the updated estimates $\hat{\theta}_w$ and $\hat{\tau}_w$.
until $\sin(\hat{\theta}_w) \leq \delta$.
Output: The final benchmark state $|b_w\rangle$.
```

the benchmark state $|b_w\rangle$ must be the solution state $|b^\star\rangle$ in the BGS algorithm. If $\sin(\theta_w) = \sqrt{1/D}$, there is one oracle state, that is, $D_w = 1$. In that case, the BGS algorithm may yield a state with the second-best BIC value. The larger value of $\sin(\theta_w)$ is, the higher probability that the BGS algorithm yields an inaccurate estimate. However, smaller $\sin(\theta_w)$ implies more iteration steps and longer computation time. Thus, one may set appropriate error tolerance δ in the iteration stopping rule to strike a balance between estimation accuracy and computational cost.

3. Theoretical Analysis

In this section, we present the theoretical analysis for the proposed BGS method. Theoretical results are established for estimation error and computational cost. The theorems also provide guidance for selecting the parameters, that is, error tolerance δ and the rotation register's number of qubits T in Algorithm 1. All proofs for this section are relegated to Section C in supplementary material.

The following theorem is on the estimation error of BGS.

Theorem 3.1 (Estimation error of BGS). Assume all the conditions in Lemma 2.1 are satisfied. If the error tolerance $\delta < 1/\sqrt{D}$, and the rotation register's number of qubits $T = \Omega\left(\log \frac{\log D}{(1/\sqrt{D}-\delta)}\right)$, the final output state $|b_w\rangle$ has the following error bound,

$$\mathbb{P}\left(|b_w\rangle \neq |b^\star\rangle\right) \lesssim \frac{\log D}{2^T\left(1/\sqrt{D}-\delta\right)}. \tag{12}$$

Remark. If the error tolerance violates $\delta \geq 1/\sqrt{D}$, (12) no longer holds, and the error probability $\mathbb{P}\left(|b_w\rangle \neq |b^\star\rangle\right)$ is lower bounded by a constant close to $\frac{1}{2}$ (see Proposition C.5 in supplementary material).

Under a slightly stronger condition on δ, we can establish a more explicit relationship between the error bound and D as well as T.

Corollary 3.2. Suppose all the conditions in Theorem 3.1 are satisfied. If we further require $\sqrt{D}\delta$ to be upper bounded away from 1, that is, $\delta \leq (1-c)/\sqrt{D}$ for some positive constant $c < 1$, we have

$$\mathbb{P}\left(|b_w\rangle \neq |b^\star\rangle\right) \lesssim \frac{\sqrt{D}\log D}{c2^T}. \tag{13}$$

Remark. Since $D = 2^p$, we can rewrite (13) as $\mathbb{P}\left(|b_w\rangle \neq |b^\star\rangle\right) \lesssim p2^{\frac{p}{2}-T}/c$.

Note that the conditions $T = \Omega\left(\log \frac{\log D}{(1/\sqrt{D}-\delta)}\right)$ and $\delta \leqslant (1-c)/\sqrt{D}$ together imply that $T = \Omega\left(\log\left(\sqrt{D}\log D\right)\right)$. Thus, (13) indicates that we can control the error probability $\mathbb{P}\left(|b_w\rangle \neq |b^\star\rangle\right)$ to be arbitrarily small with a proper T of the order $\Theta\left(\log\left(\sqrt{D}\log D\right)\right)$.

According to Theorem 3.1 and Corollary 3.2, we provide guidance for selecting the error tolerance δ and the number of qubits T in the rotation register. Note that the estimation error (13) diverges if T is too small, while the computation is too expensive if T is too large. Thus, we suggest setting $T = \left\lceil \log(\sqrt{D}\log D) + 5\right\rceil$ in practice. As for δ, intuitively, small δ needs more iterations for the algorithm to converge. Hence, we set $\delta = \frac{1}{2\sqrt{D}}$, which is bounded away from $\frac{1}{\sqrt{D}}$ but not too close to zero.

We further establish the following theorem regarding the computational cost of BGS.

Theorem 3.3 (Time and space complexity). Under all the conditions of Lemma 2.1, if $T = \Theta\left(\log\left(\sqrt{D}\log D\right)\right)$ and $\delta = \frac{1}{2\sqrt{D}}$, we have (1) the expected time complexity of Algorithm 1 is $O\left(\sqrt{D}\left(\log D\right)^3\right)$; (2) the space complexity of Algorithm 1 is $O\left(\log(\sqrt{D}\log D)\right)$.

Remark. Given the stochastic nature of quantum measurement outcomes and the heuristic-based iterative approach of BGS, the time complexity is random. Therefore, we are reporting

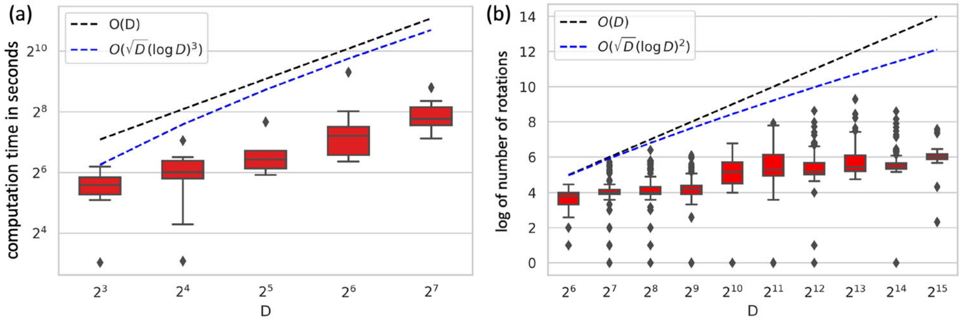

Figure 3. (a) The computation time of the BGS algorithm for the experiments conducted in the quantum computing environment. The boxplots of the actual computation time of the replicated experiments are plotted for varying D. The dashed black line denotes the order of the computation time of the exhaustive search $O(D)$. The dashed blue line represents the order of the theoretical computation time upper-bound $O(\sqrt{D}(\log D)^3)$. (b) The results of the experiments conducted in a quantum simulator. The boxplots of the log-transformed number of rotations of the replicated experiments are plotted for varying D. The dashed black line denotes the order $O(D)$. The dashed blue line represents the order of the theoretical upper bound of the number of rotations $O(\sqrt{D}(\log D)^2)$.

the upper bound of the expected time complexity. On a quantum computer, the time complexity of one rotation operation is $O\left(\log D\right)$ (Koike and Okudaira 2009). Our proposed BGS algorithm needs $O\left(\sqrt{D}(\log D)^2\right)$ rotations in expectation.

Note that search algorithms in classical computers usually have a time complexity of $O(D)$. Thus, our BGS method provides a nearly quadratic speed-up. When D is large, this improvement over classical algorithms is very significant. Moreover, Grover's algorithm has a time complexity of $O\left(\sqrt{D}\log D\right)$. The time complexity of the BGS has a moderate increase over that of Grover's algorithm while without requiring the oracle state as input.

4. Empirical Performance

We assess the empirical performance of the BGS algorithm through simulation studies using IBM Quantum Experience. Since IBM Quantum Experience only offers seven qubits for public access, we evaluate the performance of BGS for varying the number of predictors up to the maximum in the real quantum computing environment. In addition, IBM provides a quantum simulator to mimic the real quantum computing environment. We thus evaluate the performance of BGS for a relatively large number of predictors using the simulator. We implemented the proposed BGS algorithm using a Qiskit Python development kit[1] provided by IBM Quantum Experience. The performance of the BGS algorithm is assessed in both linear regression models and weighted logistic regression models. We present the results of linear regression models here. The results of weighted logistic regression models are relegated to Section D in supplementary material.

For the linear regression model, replicated samples are generated according to

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n, \tag{14}$$

where $y_i \in \mathbb{R}$, $\boldsymbol{x}_i \in \mathbb{R}^p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$, and $\epsilon_i \in \mathbb{R}$. We set $n = 1000$, $\{\boldsymbol{x}_i\}_{i=1}^n \overset{\text{iid}}{\sim} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, the (i,j)th entry of $\boldsymbol{\Sigma}$ equals $0.7^{|i-j|}$. The first $\lfloor p/2 \rfloor$ entries in $\boldsymbol{\beta}$ are set to 1, while the remaining entries are set to 0. We set $\sigma^2 = \frac{1}{3}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}$, and $\{\epsilon_i\}_{i=1}^n \overset{\text{iid}}{\sim} N(0, \sigma^2)$. We generate 10 replicated samples for each $p \in \{3, 4, 5, 6, 7\}$ for real quantum computing, and 100 replicated samples for each $p = \{6, 7, \ldots, 15\}$ for quantum simulating .

We apply our BGS algorithm to these samples and compared its performance with the best subset selection (BSS) via an exhaustive search.

4.1. Computational Complexity

Recall that Theorem 3.3 states the time complexity of our proposed BGS algorithm is $O(\sqrt{D}(\log D)^3)$. In particular, the time complexity of Grover's rotation is $O(\log D)$, and BGS needs $O(\sqrt{D}(\log D)^2)$ times of rotations.

In Figure 3(a), we plot the actual computation time of the BGS algorithm for a limited range of D conducted in the quantum computer. Next, we examine the empirical computational complexity of the proposed BGS for a larger range of D in the quantum simulator. Due to the fact that the quantum simulator is realized by a classical computer, some quantum steps are implemented through classical computational methods. For example, Grover's rotation is implemented via multiplying a $D \times D$ rotation matrix with a D-dimensional vector, which gives rise to the fact that the actual computational complexity of one Grover's rotation in the simulator is $O(D^2)$. Therefore, the real computation time of the BGS algorithm in a simulator does not faithfully represent its computation time in a quantum computer. For a fair comparison, we plot the number of rotations of the BGS algorithm as a surrogate of its computation time in Figure 3(b).

Both the results in Figure 3(a) and (b) show that the computational cost of our proposed BGS gradually increases as D gets larger. However, we note that both the growth rate of the computation time and the growth rate of the numbers of rotations are indeed upper bounded as shown in our theoretical analysis. In Figure 3(a), we notice that the computation time of running BGS is large even when the number of predictor sets D is small. This computational cost is primarily spent on the quantum machine warm-up.

We also notice that the computational cost has a large variability, which is attributed to the fact that BGS is a stochastic

[1] *https://qiskit.org/*

Table 1. The percentage of the true subset being selected among 100 replicates by BSS and proposed BGS.

Method	$p = 6$	$p = 7$	$p = 8$	$p = 9$	$p = 10$	$p = 11$	$p = 12$	$p = 13$	$p = 14$	$p = 15$
BSS	96%	100%	99%	98%	99%	99%	99%	99%	98%	99%
BGS	98%	99%	98%	98%	98%	98%	97%	99%	97%	98%

algorithm. Nevertheless, the advantage of the proposed BGS over the classical algorithm becomes more significant as D gets larger.

4.2. Performance of Selection

To evaluate the subset selection accuracy of the BGS algorithm, we compare it with the best subset selection (BSS) via emulation in a classical computer. For the BSS, the subset with the smallest BIC is selected. One hundred replicated samples of size 1000 are generated according to model (14) for each $p = \{6, 7, \dots, 15\}$. We report the frequency of the true subset being selected among 100 replicates in Table 1.

We observe that the frequencies in Table 1 are close or equal to 100% for both BGS and BSS under all settings. The frequencies of the true subset being selected by the BGS are almost as good as those by the BSS.

5. Application

We evaluate the performance of the BGS algorithm in two CITE-seq studies.

In each study, we randomly divide the whole dataset into a training dataset, consisting of CITE-seq expressions of 75% of the cells, and a testing dataset consisting of expressions of 25% of the cells. We apply BGS to the training dataset to fit models and select the best subsets, and then conduct predictions in the testing dataset. We repeat this experiment 100 times. In each experiment, we also apply the differential expression methods in the Seurat package V5 (Hao et al. 2023) with six options: Wilcoxon Rank Sum test (Seurat-wilcox), Student's t-test (Seurat-t), likelihood-ratio test, (Seurat-bimod), ROC analysis (Seurat-roc), a logistic regression-based method (Seurat-LR) and a hurdle model-based method (Seurat-MAST), and the differential expression methods in the SCANPY package V1.9.2 (Wolf, Angerer, and Theis 2018) with three options: Wilcoxon Rank Sum test (SCANPY-wilcox), Student's t-test (SCANPY-t), and a logistic regression-based method (SCANPY-LR). Further details can be found in Sections E and F in supplementary material.

5.1. Identification of ADTs Associated with the Marker Gene of Interest

In a study of the human cord blood mononuclear cells (CBMC), Stoeckius et al. (2017) profiled a total of 8617 cells, including 1727 CD14+ Monocytes cells, using CITE-seq. For each cell, 13 cell-surface protein markers are quantified via sequencing their corresponding antibody-derived tags (ADTs), and RNA expression levels are measured. Among the sequenced RNA, CD14 RNA is an important immune response gene and has a high expression level in some monocytes (Rawat et al. 2021). We are interested in the association between the expression level of the CD14 RNA and these 13 ADTs in the CD14+ Monocytes cells. We preprocess the expression data using the Seurat package V5 (details are provided in Section E in supplementary material).

We identify the ADTs associated with CD14 RNA through a regression approach (Zhong et al. 2005; Zamdborg and Ma 2009; Liu et al. 2024). In particular, we regress the expression level of CD14 RNA on the expression levels of 13 ADTs. We aim to identify the best subset of ADTs for explaining the variation of CD14 RNA. For each candidate subset of the 13 ADTs, we fit the corresponding linear regression model to the training dataset and calculate the BIC. We then apply the BGS algorithm to select the best subsets.

We also select the subset of ADTs using the differential expression methods of the Seurat and SCANPY packages. Using the subset of ADTs, we fit a linear regression for CD14 RNA on the training set, and we evaluate the methods by the mean square error between the predicted and observed values of CD14 RNA expression level on the testing dataset. For a fair comparison, we ensure that all methods have the same number of selected ADTs, ranging from 1 to 12. The natural logarithms of the ratios of the mean squared errors (MSEs) and BIC of Seurat and SCANPY relative to BGS are computed. The resulting log MSE ratios and BIC ratios are presented in Figure 4(a) and (b) through boxplots. Notably, for all numbers of selected ADTs, most boxplots are situated above zero. This observation indicates that BGS consistently outperforms the other two methods across various scenarios.

We now examine the identified ADTs associated with CD14 RNA by the BGS algorithm. The best subset of ADTs is selected among 2^{13} possible candidate subsets. We identify the two most significant ADTs associated with CD14 RNA. By fitting the linear regression model using CD4 and CD14 as the predictors on all CD14+ Monocytes cells, the coefficient of CD4 is -0.1388, and the coefficient of CD14 is 0.1333. Notice that CD14 ADT is positively associated with CD14 RNA, which is consistent with the fact that CD14 RNA is the coding gene for the CD14 protein. Additionally, we observe negative relationships between CD4 and CD14 RNA. This outcome is consistent with the existing literature. A study on the clinical outcomes of dedifferentiated liposarcoma patients (Schroeder et al. 2021) has shown a negative relationship between CD4 and CD14.

5.2. Panel Design for Immune Cell Type Identification

In a CITE-seq experiment, Hao et al. (2021) reported the measurement of expression levels for 228 ADTs and 33,538 RNAs in 161,764 peripheral blood mononuclear cells (PBMCs). These cells encompassed 57 distinct cell types, which were identified and annotated based on differentially expressed RNAs. Additionally, 10 ADT markers were identified for each cell type.

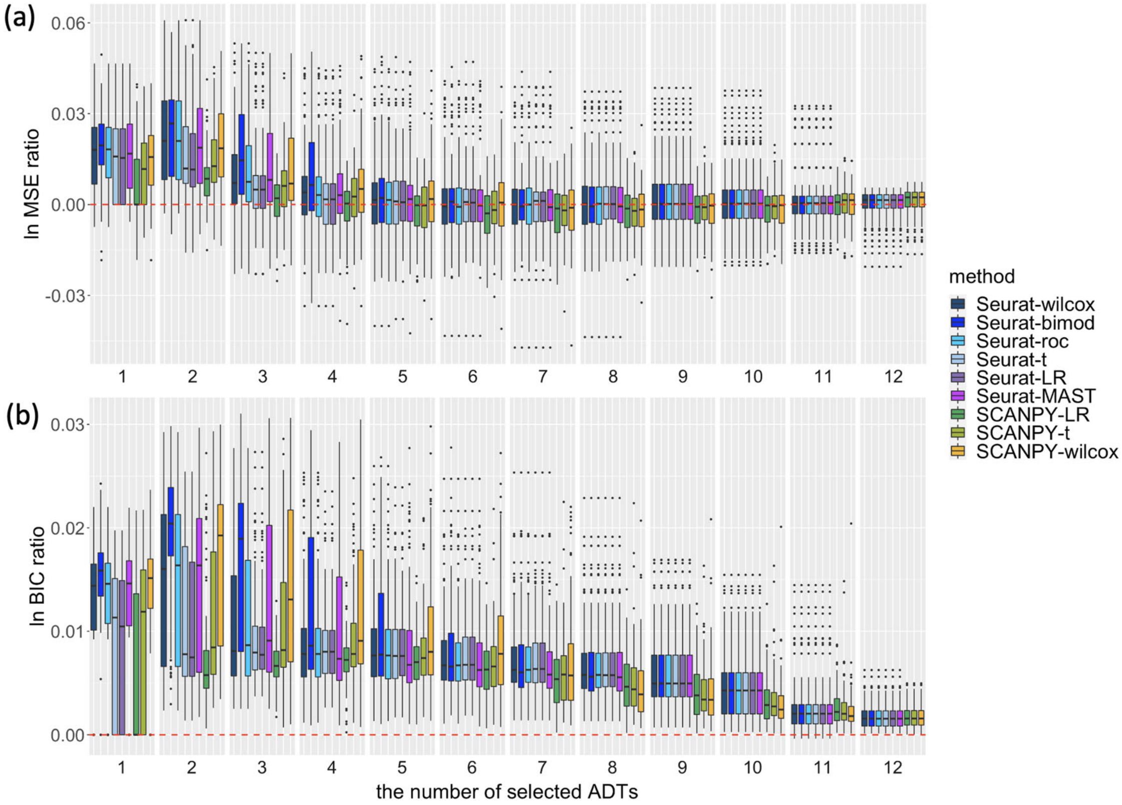

Figure 4. Comparison of the performance of BGS, Seurat, and SCANPY on the CD14+ Monocytes cells from CBMC data. The evaluation is based on the mean squared error (MSE) and BIC between the predicted and observed values of CD14 RNA expression level across 100 testing datasets. (a) The x-axis represents the number of selected ADTs, while the y-axis represents the log ratio of the MSE for each method compared to the MSE of the proposed BGS algorithm. The red dashed line indicates a ratio of zero. (b) The x-axis represents the number of selected ADTs, while the y-axis represents the log ratio of the BIC for each method compared to the BIC of the proposed BGS algorithm.

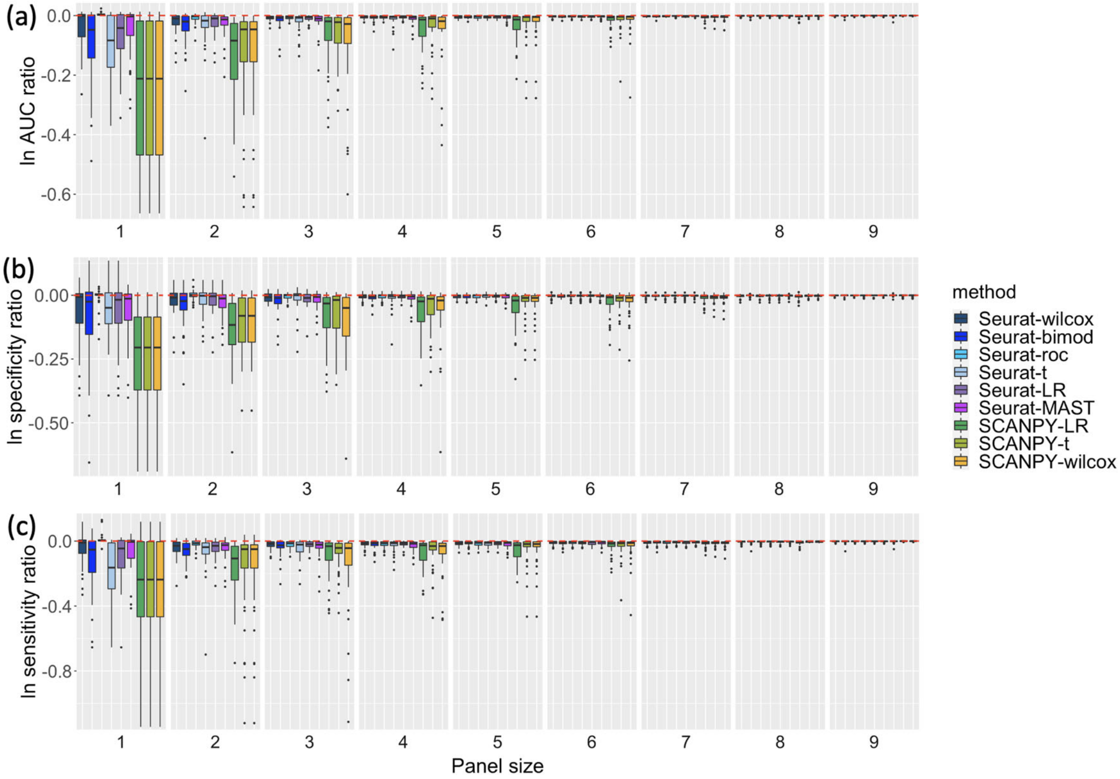

Figure 5. The prediction performance of the 41 cell types is compared for the Seurat and SCANPY methods with respect to the BGS algorithm. The panel size is shown on the X-axis, while the Y-axis represents the log ratio of the evaluation metric of each method compared to that of the proposed BGS algorithm. The evaluation metrics reported are AUC in (a), specificity in (b), and sensitivity in (c). The red dashed line in each plot indicates a ratio of zero, serving as a reference point for comparison.

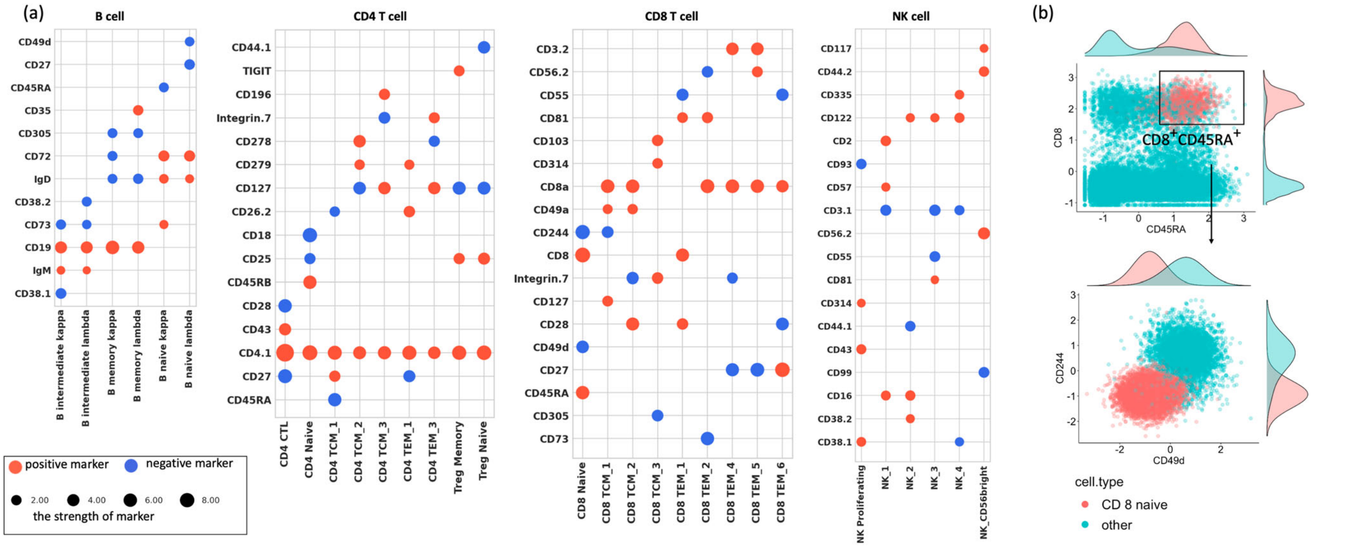

Figure 6. The best four-marker panels selected by the BGS algorithm. (a) Dot plots of the best panels are presented. Each cell type is represented by a column, with the cell type name displayed in the bottom margin. Within each column, four dots are positioned to represent the best four markers, and their corresponding names are indicated in the left margin. The dots are color-coded as red or blue, representing the positive or negative coefficients of the markers in the corresponding logistic regression model. The size of each dot reflects the strength of the corresponding marker, determined by the absolute value of the coefficient. (b) The best four-marker panel for CD8 naive cells (CD8, CD45RA, CD49d, CD244) is shown. The marginal distributions of these markers' expressions are displayed on the top and right margins. Each dot represents a cell, and the color of the dot indicates the cell type. The upper subfigure illustrates the scatterplot of CD8 and CD45RA for all PBMC cells, while the lower subfigure shows the scatterplot of CD49d and CD244 specifically for the screened CD8$^+$CD45RA$^+$ cells.

Our study aims to evaluate the performance of our proposed BGS algorithm in identifying panels of ADT markers for classifying various cell types. To ensure a fair comparison with the results reported in Hao et al. (2021), we restrict the panel size to be smaller than 10. We focus on classifying cells for the major cell types, specifically those with more than 500 sequenced cells, resulting in 41 cell types of interest. For each cell type of interest, we assign that particular cell type a label of one, while labeling all other cell types as zero. Consequently, we encounter highly unbalanced cell samples, with significantly fewer cells labeled as one compared to those labeled as zero. To mitigate potential biases arising from unbalanced samples, we employ weighted logistic regression models (King and Zeng 2001) on the training dataset for each candidate panel, accounting for the sample weights. We calculate the BIC and employ our BGS algorithm to identify the best panel with the optimal BIC. For each of the 41 cell types, we also use the Seurat and SCANPY methods to identify the differentially expressed ADTs specific to that cell type, using the training dataset. These selected ADTs are then used to fit weighted logistic regression models. Finally, we assess the performance of the fitted models on the testing dataset by calculating metrics such as the area under the ROC curve (AUC), sensitivity, and specificity.

For each of the 41 cell types, we calculate the log ratios of the evaluation metrics for the Seurat and SCANPY methods compared to those of the proposed BGS algorithm. A log ratio less than zero indicates that BGS outperforms the corresponding method. The results are visualized through boxplots in Figure 5(a)–(c), for AUC, sensitivity, and specificity, respectively. To maintain a reasonable scale of the figure, the outliers below the lower bound of the Y-axis are omitted. Across all three metrics and considering the four different panel sizes, the majority of the boxplots are positioned below zero. This observation indicates that BGS consistently outperforms the other methods across a wide range of scenarios.

Figure 6 (a) showcases dot plots depicting the best four-marker panels selected by the BGS algorithm for 30 different cell types, which can be categorized into four coarse cell types: B cells, CD4 T cells, CD8 T cells, and NK cells. Upon examining the panels within each coarse cell type, we observe that several best panels share common markers. For instance, in the case of CD4 T cells, all best panels include CD4.1 as a marker with a large strength. However, when comparing panels across different coarse cell types, we find that the best panels exhibit distinct marker compositions. In addition, the panels identified by our BGS algorithm are consistent with the literature. For instance, CD19$^+$IgD$^-$ are well-known markers for B memory cells; IgD$^+$ is a marker for B naive cells (Kaminski et al. 2012); CD4$^+$CD25$^+$CD127$^-$ are markers for Treg cells (Liu et al. 2006), and CD3$^-$CD122$^+$ are markers for NK cells (Farag and Caligiuri 2006).

Moreover, we make an intriguing discovery regarding the combined use of CD49d and CD244 as highly informative markers, in addition to the typical markers CD8$^+$CD45RA$^+$ (Nguyen et al. 2016), for identifying CD8 naive T cells. This novel finding has not been previously reported. Although previous studies consistently demonstrate that CD8 naive cells typically exhibit low expression levels of both CD49d and CD244, which may increase upon activation, the specific roles and implications of CD49d and CD244 in immune responses have not been actively studied until very recently (White, Cross, and Kedl 2017; Berard and Tough 2002; Agresta, Hoebe, and Janssen 2018). The up-regulation of CD49d facilitates efficient access to inflamed peripheral tissues and enhances responsiveness to inflammatory signals, while up-regulated CD244 signaling activates CD8 naive T cells. However, using CD244 or CD49d alone may not be sufficient as informative markers, as illustrated by the marginal distributions of CD49d and CD244 for the $CD8^+CD45RA^+$ cell population in Figure 6(b). In contrast, our selected panel ($CD8^+CD45RA^+CD244^-CD49d^-$) demonstrates high accuracy in screening CD8 naive cells, as evidenced by an average AUC of 0.998 in the test datasets across 100 replications. Some additional results for this study are reported in Section H in supplementary material.

Note that the design of the experimental setup is constrained by the hardware currently available. Consequently, the experiments are conducted on relatively low-dimensional problems. Nonetheless, we expect that our algorithm will scale effectively and yield more significant insights as quantum computing technology advances.

6. Conclusion

Given that many statistical problems involve computationally intensive tasks, statisticians are particularly intrigued by the potential of quantum computers (Wang and Song 2020; Wang and Liu 2022). Consequently, a natural question is whether these computers will benefit the statisticians in solving some statistics or data science problems. If the answer is affirmative, what kind of statistics problems should statisticians resort to quantum computers? Unfortunately, the general answer to this question remains elusive.

In this article, we answer this question by showing the benefit of quantum computing in single-cell biology problems that statisticians have been working on extensively (Agarwal, Wang, and Zhang 2020). In particular, we developed the bisection Grover's search algorithm for selecting the best subset and demonstrated its advantages in identifying the ADTs associated with targeted genes and designing panels for cell type identifications. We established the theoretical properties of our proposed algorithm. We also demonstrated the empirical performance of the proposed algorithm in a NISQ device and a simulator.

It is worth noting that the BGS algorithm is highly versatile and flexible. It has the potential to be seamlessly integrated into other machine learning methods, such as deep neural networks, which are known for their remarkable fitting or expressive power and predictive capabilities (Beer et al. 2020; Abbas et al. 2021). This integration enables researchers to harness the expressive power offered by machine learning models while capitalizing on the computational efficiencies provided by quantum computing to tackle more complex problems.

One key feature that distinguishes quantum algorithms from classical algorithms is quantum parallelism, which enables us to develop a unique approach to addressing multi-modal biological problems. The potential application of quantum algorithms in the realm of biological problems extends far beyond the scope presented here. For instance, quantum algorithms hold promise for more effectively analyzing spatial transcriptomics data, unlocking new insights and capabilities in the field.

Supplementary Materials

The supplementary materials contain technical preliminaries and details, additional simulation and experiment results, theoretical results and the corresponding proofs.

Acknowledgments

The authors would like to thank the editor, associate editor, and the three committed reviewers for their careful and constructive review.

Disclosure Statement

The authors report there are no competing interests to declare.

Data Availability Statement

The human cord blood mononuclear cells dataset (Stoeckius et al. 2017) is available at the NCBI Gene Expression Omnibus (GEO; *https://www.ncbi.nlm.nih.gov/geo/*) with accession no.GSE100866. The peripheral blood mononuclear cells dataset (Hao et al. 2021) is available at New York Genome Center (*https://atlas.fredhutch.org/nygc/multimodal-pbmc/*). Source code for the BGS algorithm is freely available on GitHub (*https://github.com/StatCYK/BGS-for-CITEseq/*).

Funding

This work was partially supported by the U.S. National Science Foundation DMS-1925066, DMS-1903226, DMS-2124493, DMS-2311297, DMS-2319279, DMS-2318809 and the National Institutes of Health NIH R01GM152814.

ORCID

Ping Ma http://orcid.org/0000-0002-5728-3596

References

Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A., and Woerner, S. (2021), "The Power of Quantum Neural Networks," *Nature Computational Science*, 1, 403–409. [62]

Agarwal, D., Wang, J., and Zhang, N. R. (2020), "Data Denoising and Post-Denoising Corrections in Single Cell RNA Sequencing," *Statistical Science*, 35, 112–128. [62]

Agresta, L., K. H. Hoebe and E. M. Janssen (2018), "The Emerging Role of CD244 Signaling in Immune Cells of the Tumor Microenvironment," *Frontiers in Immunology*, 9, 2809. [61]

Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J. C., Barends, R., et al. (2019), "Quantum Supremacy Using a Programmable Superconducting Processor," *Nature*, 574, 505–510. [52]

Beer, K., Bondarenko, D., Farrelly, T., Osborne, T. J., Salzmann, R., Scheiermann, D., et al. (2020), "Training Deep Quantum Neural Networks," *Nature Communications*, 11, 808. [62]

Berard, M., and Tough, D. F. (2002), "Qualitative Differences Between Naive and Memory T Cells," *Immunology*, 106, 127. [61]

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., and Lloyd, S. (2017), "Quantum Machine Learning," *Nature*, 549, 195–202. [54]

Brassard, G., Hoyer, P., Mosca, M., and Tapp, A. (2002), "Quantum Amplitude Amplification and Estimation," *Contemporary Mathematics*, 305, 53–74. [54]

Brassard, G., Høyer, P., and Tapp, A. (1998), "Quantum Counting," in *International Colloquium on Automata, Languages, and Programming*, pp. 820–831, Springer. [56]

Castelvecchi, D. (2023), "IBM Releases First-Ever 1,000-Qubit Quantum Chip," *Nature*, 624, 238–238. [52]

Chakraborty, S., Shaikh, S. H., Chakrabarti, A., and Ghosh, R. (2020), "A Hybrid Quantum Feature Selection Algorithm Using a Quantum Inspired Graph Theoretic Approach," *Applied Intelligence*, 50, 1775–1793. [53]

Farag, S. S., and Caligiuri, M. A. (2006), "Human Natural Killer Cell Development and Biology," *Blood Reviews*, 20, 123–137. [61]

Ferrer-Font, L., Small, S. J., Lewer, B., Pilkington, K. R., Johnston, L. K., Park, L. M., et al. (2021), "Panel Optimization for High-Dimensional Immunophenotyping Assays Using Full-Spectrum Flow Cytometry," *Current Protocols*, 1, e222. [53]

Grover, L. K. (1996), "A Fast Quantum Mechanical Algorithm for Database Search," in *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '96, New York, NY, USA, pp. 212–219, Association for Computing Machinery. [54]

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., et al. (2021), "Integrated Analysis of Multimodal Single-Cell Data," *Cell*, 184, 3573–3587. [53,59,61,62]

Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., et al. (2023), "Dictionary Learning for Integrative, Multimodal and Scalable Single-Cell Analysis," *Nature Biotechnology*, 42, 293–304. [59]

He, Z., Li, L., Huang, Z., and Situ, H. (2018), "Quantum-Enhanced Feature Selection with Forward Selection and Backward Elimination," *Quantum Information Processing*, 17, 1–11. [53]

Kaminski, D. A., Wei, C., Qian, Y., Rosenberg, A. F., and Sanz, I. (2012), "Advances in Human B Cell Phenotypic Profiling," *Frontiers in Immunology*, 3, 302. [61]

King, G., and Zeng, L. (2001), "Logistic Regression in Rare Events Data," *Political Analysis*, 9, 137–163. [61]

Koike, T., and Okudaira, Y. (2009), "Time Complexity and Gate Complexity," *Physical Review A*, 82, 042305. [58]

Li, Y., Zhou, R.-G., Xu, R., Luo, J., Hu, W., and Fan, P. (2022), "Implementing Graph-Theoretic Feature Selection by Quantum Approximate Optimization Algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, 35, 2364–2377. [53]

Liu, W., Putnam, A. L., Xu-Yu, Z., Szot, G. L., Lee, M. R., Zhu, S., et al. (2006), "Cd127 Expression Inversely Correlates with foxp3 and Suppressive Function of Human CD4+ T Reg Cells," *The Journal of Experimental Medicine*, 203, 1701–1711. [61]

Liu, Y., Chen, Y., Lu, H., Zhong, W., Yuan, G.-C., and Ma, P. (2024), "Orthogonal Multimodality Integration and Clustering in Single-Cell Data," *BMC Bioinformatics*, 25, 164. [59]

Mücke, S., Heese, R., Müller, S., Wolter, M., and Piatkowski, N. (2023), "Feature Selection on Quantum Computers," *Quantum Machine Intelligence*, 5, 11. [53]

Nguyen, H. H., Kim, T., Song, S. Y., Park, S., Cho, H. H., Jung, S.-H., et al. (2016), "Naïve CD8+ T Cell Derived Tumor-Specific Cytotoxic Effectors as a Potential Remedy for Overcoming TGF-β Immunosuppression in the Tumor Microenvironment," *Scientific Reports*, 6, 1–10. [61]

Nielsen, M. A., and Chuang, I. L. (2010), *Quantum Computation and Quantum Information*, Cambridge: Cambridge University Press. [52,55]

Preskill, J. (2018), "Quantum Computing in the NISQ Era and Beyond," *Quantum*, 2, 79. [52]

Rawat, K., Pal, A., Banerjee, S., Pal, A., Mandal, S. C., and Batabyal, S. (2021), "Ovine CD14-an Immune Response Gene has a Role Against Gastrointestinal Nematode Haemonchus Contortus—A Novel Report," *Frontiers in Immunology*, 12, 664877. [59]

Schroeder, B. A., LaFranzo, N. A., LaFleur, B. J., Gittelman, R. M., Vignali, M., Zhang, S., et al. (2021), "CD4+ T cell and M2 Macrophage Infiltration Predict Dedifferentiated Liposarcoma Patient Outcomes," *Journal for Immunotherapy of Cancer*, 9, e002812. [59]

Schuld, M., Sinayskiy, I., and Petruccione, F. (2016), "Prediction by Linear Regression on a Quantum Computer," *Physical Review A*, 94, 022342. [54]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [53]

Shor, P. W. (1999), "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer," *SIAM Review*, 41, 303–332. [52]

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., et al. (2017), "Large-Scale Simultaneous Measurement of Epitopes and Transcriptomes in Single Cells," *Nature Methods*, 14, 865. [53,59,62]

Turati, G., Dacrema, M. F., and Cremonesi, P. (2022), "Feature Selection for Classification with qaoa," in *2022 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pp. 782–785, IEEE. [53]

Von Dollen, D., Neukart, F., Weimer, D., and Bäck, T. (2021), "Quantum-Assisted Feature Selection for Vehicle Price Prediction Modeling," arXiv preprint arXiv:2104.04049. [53]

Wang, Y. (2022), "When Quantum Computation Meets Data Science: Making Data Science Quantum," *Harvard Data Science Review*, 4. DOI:10.1162/99608f92.ef5d8928. [52]

Wang, Y., and Liu, H. (2022), "Quantum Computing in a Statistical Context," *Annual Review of Statistics and Its Application*, 9, 479–504. [62]

Wang, Y., and Song, X. (2020), "Quantum Science and Quantum Technology," *Statistical Science*, 35, 51–74. [62]

White, J. T., Cross, E. W., and Kedl, R. M. (2017), "Antigen-Inexperienced Memory cd8+ t Cells: Where They Come From and Why We Need Them," *Nature Reviews Immunology*, 17, 391–400. [61]

Wolf, F. A., Angerer, P., and Theis, F. J. (2018), "Scanpy: Large-Scale Single-Cell Gene Expression Data Analysis," *Genome Biology*, 19, 1–5. [59]

Wootters, W. K., and Zurek, W. H. (1982), "A Single Quantum Cannot be Cloned," *Nature*, 299, 802–803. [56]

Wu, Y., Bao, W.-S., Cao, S., Chen, F., Chen, M.-C., Chen, X., et al. (2021), "Strong Quantum Computational Advantage Using a Superconducting Quantum Processor," *Physical Review Letters*, 127, 180501. [52]

Zamdborg, L., and Ma, P. (2009), "Discovery of Protein–DNA Interactions by Penalized Multivariate Regression," *Nucleic Acids Research*, 37, 5246–5254. [59]

Zhong, H.-S., Wang, H., Deng, Y.-H., Chen, M.-C., Peng, L.-C., Luo, Y.-H., et al. (2020), "Quantum Computational Advantage Using Photons," *Science*, 370, 1460–1463. [52]

Zhong, W., Zeng, P., Ma, P., Liu, J. S., and Zhu, Y. (2005), "RSIR: Regularized Sliced Inverse Regression for Motif Discovery," *Bioinformatics*, 21, 4169–4175. [59]