Accurate highly variable gene selection using RECODE in scRNA-seq data analysis

Yusuke Imoto

Institute for the Advanced Study of Human Biology
Kyoto University
Kyoto, Japan
imoto.yusuke.4e@kyoto-u.ac.jp

Abstract

Accurate selection of highly variable genes (HVGs) is essential in single-cell RNA sequencing (scRNA-seq) data analysis, as it enables the identification of functionally important genes and the characterization of cell types and states. However, HVG selection is often confounded by technical noise inherent in the scRNA-seq measurement process, leading to the misidentification of biologically relevant genes. In this study, we propose a novel HVG selection method based on the RECODE-denoised variance. RECODE is a recently developed noise reduction algorithm that models technical noise in scRNA-seq data as a random sampling process and removes it through a theoretically grounded denoising procedure. Benchmarking on 10x Genomics PBMC data revealed that our method outperformed conventional approaches in capturing known marker genes. Moreover, it demonstrated robustness to changes in normalization parameters and downsampling of cell numbers. Importantly, the same denoised data used for HVG selection can be directly reused in downstream analysis, ensuring consistency throughout the scRNA-seq data analysis pipeline.

Keywords: Highly variable gene selection, Single-cell RNA-sequencing data, Technical noise, Dropout, RECODE.

Introduction

Recent advances in single-cell RNA sequencing (scRNA-seq) have enabled the analysis of transcriptomic profiles at the resolution of individual cells, greatly contributing to our understanding of developmental processes, cell type classification, and disease mechanisms [13]. In such analyses, the accurate selection of highly variable genes (HVGs), which reflect cell-to-cell heterogeneity, is essential to identify functional genes and decipher gene regulatory networks [19]. However, scRNA-seq data are known to contain a large amount of technical noise arising from processes such as library preparation and sequencing [1]. As a result, observed variability statistics often fail to accurately reflect true biological variability. When technical noise is not properly corrected, cell-type-specific genes and functional genes that drive cellular differentiation may be missed, and low-variability non-functional genes may be erroneously selected.

To address this issue, existing approaches have attempted to correct bias in variability statistics, such as observed variance or dispersion (variance-to-mean ratio), which tend to depend on mean expression levels.

For example, Satija et al. [11] proposed an HVG selection method using normalized dispersion, which is modified by a z-score transformation according to bins of average expression. As its newer version, Stuart et al. [12] introduced an HVG selection method using a normalized variance estimated by non-linear regression based on a negative binomial model [4]. While both methods show practical utility, they rely on empirical normalization of observed statistics and do not accurately account for the impact of noise. Consequently, their theoretical justifications remain limited. Moreover, the normalized variability statistics produced by conventional HVG selection methods are not directly applicable to downstream analyses such as clustering or pathway enrichment. As a result, these approaches fail to provide consistent and integrated analytical outcomes throughout the scRNA-seq data analysis pipeline.

To overcome this limitation, we previously developed RECODE (resolution of the curse of dimensionality), a method for reducing technical noise across entire single-cell transcriptome datasets based on high-dimensional statistics [7]. RECODE mathematically models technical noise as random sampling noise, derived from the measurement process of single-cell omics data, and applies eigenvalue modification techniques theoretically derived in high-dimensional statistics [18] to remove the noise component. More recently, the incorporation of eigenvector modification [17] has further improved the accuracy of noise reduction [5]. The application of RECODE to single-cell omics data has been shown to improve a wide range of downstream analyses, including clustering, trajectory inference, and differential expression analysis, leading to enhanced detection of functional genes and rare cell types [2, 6, 3, 8, 9, 10, 16].

In this study, we propose a simple and theoretically grounded method to select HVGs using the variance after applying RECODE. Using the widely studied PBMC dataset, we demonstrate through comparative experiments with conventional methods (Seurat, Seurat v3) that our approach achieves superior accuracy and robustness in detecting known marker genes. Our method is readily applicable to standard scRNA-seq analysis pipelines and offers HVG selection that is less affected by technical noise.

Results

Technical noise floor curve disturbs HVG selections

To investigate the influence of technical noise on variability statistics, we first visualized the relationship between mean expression per gene and various measures of variability across different normalizations (Figure 1a). In scRNA-seq data, we expect that a substantial number of genes, such as housekeeping or cell-cycle-related genes, are expressed consistently across most cell types and exhibit near-zero biological variance. Accordingly, variability measures should, in principle, contain many values close to zero and remain largely independent of mean expression.

However, as shown by the red dashed lines in Figure 1a, the observed distributions of dispersion and variance after total-count normalization exhibit a strong lower bound that increases monotonically with mean expression. Even after log-normalization, the lower envelope of variance remains non-negligible, peaking around mean expression levels of 1–2 on the log scale. These lower bounds are not of biological origin but rather arise from technical noise intrinsic to the scRNA-seq measurement process.

We refer to this increasing baseline as the *technical noise floor (TNF) curve*. Genes whose expression variability lies near the TNF curve are likely dominated by noise rather than true biological differences, and as a result, HVG selection in these regions is particularly error-prone. Therefore, reducing or equalizing the TNF curve across the range of expression levels is essential for comparing variances on an equal footing and improving the precision of HVG selection.

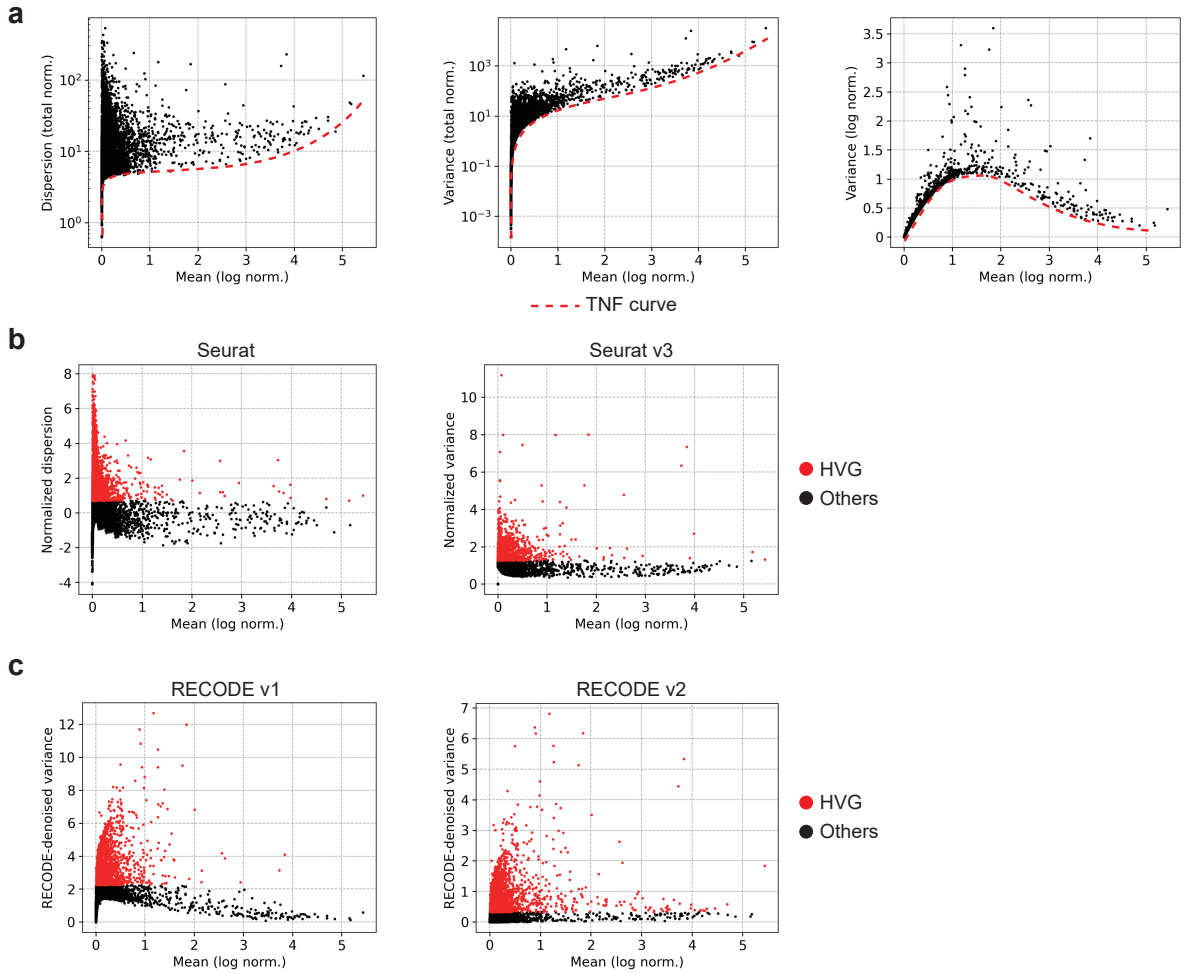

Figure 1: Overview of technical-noise structure and HVG selections. **a**, Scatter plots of mean expression per gene (x-axis) versus variability measures (y-axis) under different normalization schemes. Left: dispersion after total-count normalization; center: variance after total-count normalization; right: variance after log-normalization. Each black dot represents a gene. The red dashed line denotes the technical noise floor (TNF) curve, representing the lower bound of variability statistics elevated by technical noise. **b**, Comparison of highly variable gene (HVG) selection by two conventional methods. The y-axis shows the variability measure used by each method: normalized dispersion (Seurat) and normalized variance (Seurat v3). **c**, HVG selection by RECODE v1 and RECODE v2. The y-axis shows the RECODE-denoised variance after noise correction and normalization. In **b**–**c**, genes highlighted in red are those selected as HVGs with a cutoff of `n_top_genes` $= 2{,}000$, and genes in black are others.

Existing and RECODE-based methods for HVG selection

We summarize two widely used methods for highly variable gene (HVG) selection and introduce our proposed approach based on RECODE-denoised variance. The Seurat method [11] classifies genes according to their mean expression and then z-score normalizes the dispersion (variance-to-mean ratio) within each bin. The resulting normalized dispersion values are used to select HVGs. Seurat has been widely adopted due to its scalability and ease of use, and its practical utility has been validated in benchmark studies such as Yip et al. [19]. Seurat version 3 (Seurat v3) [12] improves upon the earlier method by using normalized

variances obtained through variance-stabilizing transformation, in which the mean-variance relationship is modeled based on a negative binomial distribution [4]. These methods serve as the default HVG selection procedures in the popular Python package `scanpy` [15] and the R package `Seurat` [12], and are thus adopted as baselines in our subsequent comparisons.

Our proposed method is based on RECODE version 2 (RECODE v2) [5], which denoises scRNA-seq data by reducing technical noise arising from the random sampling process. We then apply total-count normalization and log normalization, two standard preprocessing steps in scRNA-seq analysis, to the RECODE-denoised data. Total-count normalization, also known as library size normalization or total count scaling, scales gene expression so that the total count in each cell is fixed to a constant value, known as the *target sum* c^{TS}. Log normalization applies a $\log(x+1)$ transformation to each entry. While $c^{\mathrm{TS}} = 10^4$ is commonly used in scRNA-seq data analysis, the proposed method adopts $c^{\mathrm{TS}} = 10^5$. The rationale for this choice is empirically justified in later sections, and the full implementation details are described in the Methods.

All three methods aim to mitigate the influence of technical noise on variability statistics, effectively correcting the lower bound imposed by noise (Figure 1b). Notably, both Seurat v3 and RECODE v2 achieve a flattened TNF curve, meaning that the lower bound of the variability statistic becomes nearly constant with respect to the mean expression level. This may suggest improved comparability of gene variances and greater reliability in HVG selection, as the influence of technical noise no longer distorts comparisons of variability across genes. In the next section, we quantitatively evaluate the performance of these methods using numerical benchmarks.

All methods aim to mitigate the influence of technical noise on variability statistics, effectively correcting the lower bound (TNF curve) imposed by noise (Figure 1b, c). Seurat uses z-score normalization within gene expression bins, which leads to an irregular and fragmented lower bound of the normalized dispersion across the range of mean expression values. Seurat v3 achieves a nearly horizontal lower bound by performing nonlinear regression based on a negative binomial model. However, this correction tends to elevate the lower bound in low-expression regions, indicating incomplete noise removal in that range. As a reference, RECODE v1 attenuates the TNF curve but does not fully flatten it. In contrast, RECODE v2 achieves a nearly flat TNF curve with a lower bound approaching zero across all expression levels, suggesting that technical noise has been effectively removed and that variability statistics now primarily reflect biological differences. This flattening improves comparability of gene variances and reliability in HVG selection, as technical noise no longer distorts variability comparisons across genes. In the next section, we quantitatively evaluate the performance of these methods using numerical benchmarks.

Quantitative comparison of HVG selection methods

To evaluate the accuracy of HVG selection, we benchmarked each method on the publicly available 10x Genomics 3k Peripheral Blood Mononuclear Cells (PBMCs) scRNA-seq dataset. For validation, we considered 17 canonical marker genes associated with distinct immune cell types. These include, for example, *IL7R* (CD4 T cells), *CD8A* and *CD8B* (CD8 T cells), and *CD14* (CD14$^+$ monocytes).

Table 1 shows the rank of each marker gene under different HVG selection methods. Figure 2a visualizes the distribution of these marker genes with respect to each method's variability statistic. In addition to the comparing methods Seurat, Seurat v3, and RECODE v2, we sometimes include RECODE v1 for reference.

The dataset contains a total of 32,738 genes. Although Seurat ranks many marker genes higher, a large number of them still fall outside the commonly used cutoff of 2,000 genes for HVG selection, indicating insufficient sensitivity. However, Seurat v3 improves substantially over Seurat, assigning higher ranks to nearly all marker genes. Remaerkebly, RECODE shows even greater improvements in both versions v1 and v2. As shown in Figure 2b, RECODE v2 achieves the highest accuracy in marker gene ranking. For example, under a typical HVG selection cutoff of `n_top_genes` $= 2{,}000$, REOCDE v2 can detect all the

Table 1: Gene-wise rankings of highly variable genes across different methods.

		HVG rank			
Gene	**Mean Exp.**	**Seurat**	**Seurat v3**	**RECODE v1**	**RECODE v2**
LYZ	1.8449	274	2	2	3
CST3	1.2575	1051	18	4	5
NKG7	0.9085	562	16	5	4
S100A8	0.8881	485	12	3	2
IL7R	0.8681	2545	2056	54	68
LGALS3	0.5127	1304	1082	34	34
GNLY	0.4991	218	5	9	6
FCGR3A	0.4408	1941	85	29	28
CD79A	0.4221	786	55	20	22
MS4A1	0.3533	2199	580	79	77
CD14	0.3107	4424	2041	82	72
CD8B	0.2747	3157	2218	185	224
CD8A	0.2371	8106	9970	374	547
MS4A7	0.2274	3431	359	202	158
KLRB1	0.1066	1328	582	968	1195
PPBP	0.0732	66	1	69	30
FCER1A	0.0453	381	10	791	514

marker genes and only RECODE v2 is able to capture key marker genes such as *IL7R*, *CD8A*, *CD8B*, and *CD14*, while Seurat and Seurat v3 fail to detect them (Figure 2c).

Importantly, these improvements are particularly notable for genes with intermediate expression levels. As seen in Table 1, RECODE v2 shows large gains in ranking over Seurat v3 for genes with mean expression between 0.20 and 0.87. Empirically, this expression range is enriched for functional marker genes [1], suggesting that RECODE v2's improvement is both quantitatively and biologically meaningful.

Effect of total-count parameter on HVG selection

We next examined how the total-count normalization parameter, denoted c^{TS} (target_sum), affects HVG selection. This parameter sets the total count to which each cell's expression values are scaled during total-count normalization. Since this scaling directly influences the input to the subsequent log-transformation, it alters the effective dynamic range of expression values. For example, as shown in Figure 3a and Table 2, when $c^{\mathrm{TS}} = 10^4$, highly expressed genes such as *LYZ* tend to rank higher. However, as c^{TS} increases, genes with moderate expression levels such as *GNLY* and *NKG7* become more prominent in the ranking. This indicates that increasing c^{TS} relatively amplifies variability in the low-to-moderate expression range, thereby shifting the focus of variability to different regions.

Typical cells contain many low-expression genes that are highly susceptible to technical noise (dropout) in scRNA-seq data. Therefore, smaller values such as $c^{\mathrm{TS}} = 10^4$ are commonly used to suppress the amplification of noise. In contrast, bulk RNA-seq pipelines often adopt count per million (CPM) normalization, which corresponds to $c^{\mathrm{TS}} = 10^6$, as technical noise has a much smaller impact in such datasets. In our benchmark, we observed that increasing c^{TS} degraded the HVG selection accuracy of Seurat, likely due to insufficient noise correction (Figure 3b).

In contrast, the median ranks of marker genes under both RECODE v1 and v2 remained stable or improved slightly with increasing c^{TS} (Figure 3b and Table 2). Notably, the rankings of low-expression genes (e.g.,

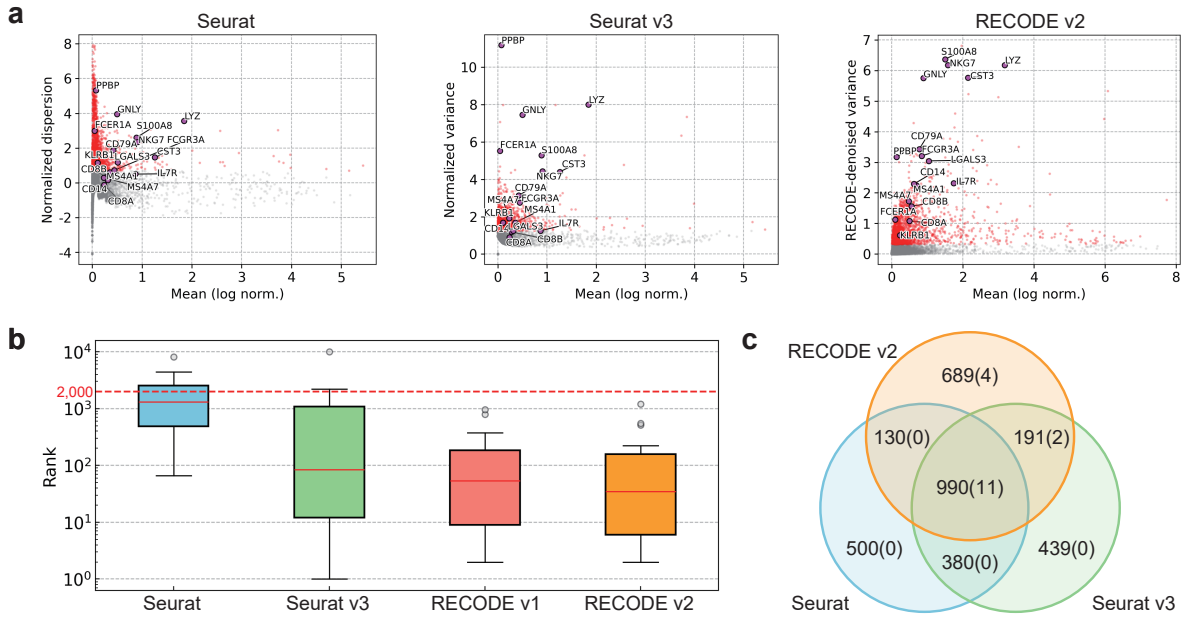

Figure 2: The performance of HVG selection methods evaluated with marker genes. **a**, Scatter plots corresponding to Figure 1b, with known marker genes highlighted in red. **b**, Box-plot comparison of the ranks assigned to marker genes by four methods (Seurat, Seurat v3, RECODE v1, RECODE v2); lower values indicate better detection. **c**, Venn diagram showing the overlap among the 2,000 HVGs selected by Seurat, Seurat v3, and RECODE v2. Numbers in parentheses indicate the count of marker genes contained in each region.

PPBP, *FCER1A*, *KLRB1*; mean expression $\lesssim 0.1$) improved at higher c^{TS}, suggesting that RECODE successfully denoised low-count features and enabled their reliable detection.

These results demonstrate that our RECODE-based method allows users to tailor HVG selection to target specific expression ranges by tuning c^{TS}. In this study, we adopted $c^{\mathrm{TS}} = 10^5$ as the default setting, as it offered a favorable tradeoff between high average marker gene rank and low rank variability (Figure 3b).

Robustness of HVG selection under cell downsampling

Finally, we investigated the robustness of HVG selection against reduced cell numbers. Using the same PBMC dataset, we randomly downsampled the number of cells to $n = 2{,}000$, $1{,}000$, 500, and 100 and performed HVG selection for each case. For each n, this process was repeated 10 times, and the average rank of each marker gene was computed across repetitions (Table 3, Figure 4).

Because the raw scRNA-seq data are sparse, downsampling results in a reduction in the number of expressed genes per cell. As a consequence, we observed a general decrease in gene-wise variance following downsampling (Figure 4a). Despite this, both Seurat and RECODE v2 remained computationally stable across all downsampled sizes, though a mild degradation in marker-gene ranking was observed as the cell number decreased (Figure 4b). In particular, RECODE v2 maintained reliable performance for genes with moderate to high expression levels (mean expression ≥ 0.25), with average ranks remaining below the commonly used cutoff `n_top_genes` $= 2{,}000$ even for $n = 100$. By contrast, Seurat v3 failed to produce results at $n = 500$ and $n = 100$ due to numerical instability, indicating a lack of robustness under strong downsampling conditions.

Table 2: Marker-gene ranks of proposed method (RECODE v2) at different `target_sum` (TS) values for total-count normalization.

		HVG rank		
Gene	**Mean Exp.**	$c^{\mathrm{TS}} = 10^4$	$c^{\mathrm{TS}} = 10^5$	$c^{\mathrm{TS}} = 10^6$
LYZ	1.8449	2	3	6
CST3	1.2575	8	5	5
NKG7	0.9085	7	4	3
S100A8	0.8881	4	2	2
IL7R	0.8681	70	68	135
LGALS3	0.5127	54	34	28
GNLY	0.4991	10	6	4
FCGR3A	0.4408	45	28	23
CD79A	0.4221	30	22	21
MS4A1	0.3533	77	77	96
CD14	0.3107	87	72	86
CD8B	0.2747	377	224	301
CD8A	0.2371	764	547	714
MS4A7	0.2274	234	158	176
KLRB1	0.1066	1531	1195	1157
PPBP	0.0732	55	30	17
FCER1A	0.0453	805	514	392

Table 3: Average marker-gene ranks of proposed method (RECODE v2) under random downsampling of cell number ($n = 2{,}000$, 1000, 500, 100). Each rank is averaged over 10 repetitions.

		Average HVG rank			
Gene	**Mean Exp.**	$n = 2{,}000$	$n = 1{,}000$	$n = 500$	$n = 100$
LYZ	1.8449	6.1	2.2	3.8	2.5
CST3	1.2575	5.1	6.3	9.5	18.1
NKG7	0.9085	2.8	4.5	5.2	43.8
S100A8	0.8881	2.3	2.8	2.2	3.6
IL7R	0.8681	63.1	121.2	156.0	167.6
LGALS3	0.5127	58.5	118.6	160.0	118.1
GNLY	0.4991	4.4	4.7	4.4	32.4
FCGR3A	0.4408	23.7	25.5	34.0	100.8
CD79A	0.4221	27.9	27.0	31.7	301.3
MS4A1	0.3533	86.5	98.0	109.3	475.1
CD14	0.3107	83.3	50.7	53.6	47.2
CD8B	0.2747	274.5	301.4	346.8	698.3
CD8A	0.2371	520.0	518.5	527.8	1623.6
MS4A7	0.2274	172.1	219.1	213.5	169.4
KLRB1	0.1066	965.7	779.6	1449.9	2643.7
PPBP	0.0732	28.7	25.4	128.9	2761.1
FCER1A	0.0453	447.0	461.7	1368.1	4367.3

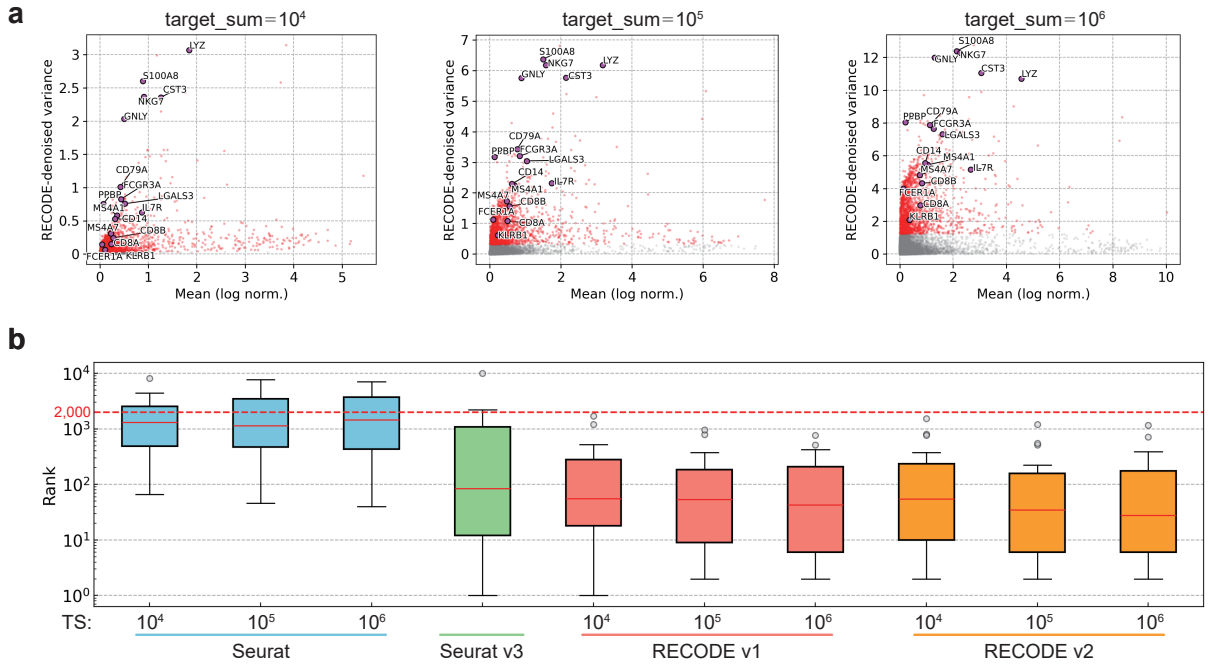

Figure 3: Influence of the total-count normalisation parameter c^{TS} on variability profiles and marker-gene ranking. **a**, Scatter plots of gene-wise mean log-normalised expression (x-axis) versus RECODE-denoised variance (y-axis) for three values of $c^{\mathrm{TS}} = 10^4, 10^5, 10^6$. Grey points indicate all genes; red points denote those selected as HVGs at `n_top_genes` $= 2{,}000$. Labeled points are established marker genes. **b**, Box plots of marker-gene ranks produced by each method as a function of parameter c^{TS}. Within Seurat and RECODE (v1, v2), three boxes correspond to $c^{\mathrm{TS}} = 10^4, 10^5, 10^6$ (labels beneath the x-axis). Seurat v3 is independent of parameter c^{TS} and therefore appears as a single box.

Low-expression genes were more frequently excluded from the proposed HVG selection as the number of cells decreased due to downsampling. This is likely due to increased sparsity in the count matrix, fewer cells express these genes after downsampling, thereby degrading the accuracy of noise correction. These findings suggest that sufficient cell numbers are necessary to reliably detect low-expression HVGs. On the other hand, for identifying highly variable genes with moderate to high expression levels, our proposed method (RECODE v2) remains accurate even under extreme downsampling conditions.

Discussion

In this study, we proposed a novel method for selecting highly variable genes (HVGs) based on the variance after denoising by RECODE (RECODE-denoised variance) and evaluated its effectiveness through comparison with existing methods. RECODE mathematically models the technical noise inherent in scRNA-seq data as random sampling noise and removes it based on high-dimensional statistics theory [7]. Our method applies total-count normalization and log normalization to RECODE-denoised data, and uses the resulting variance as the HVG selection criterion.

We first examined how technical noise affects statistical measures of variability in scRNA-seq data. We observed that dispersion and variance showed a lower bound dependent on mean expression, which we termed the technical noise floor (TNF) curve (Figure 1). This TNF structure distorts HVG selection. In contrast,

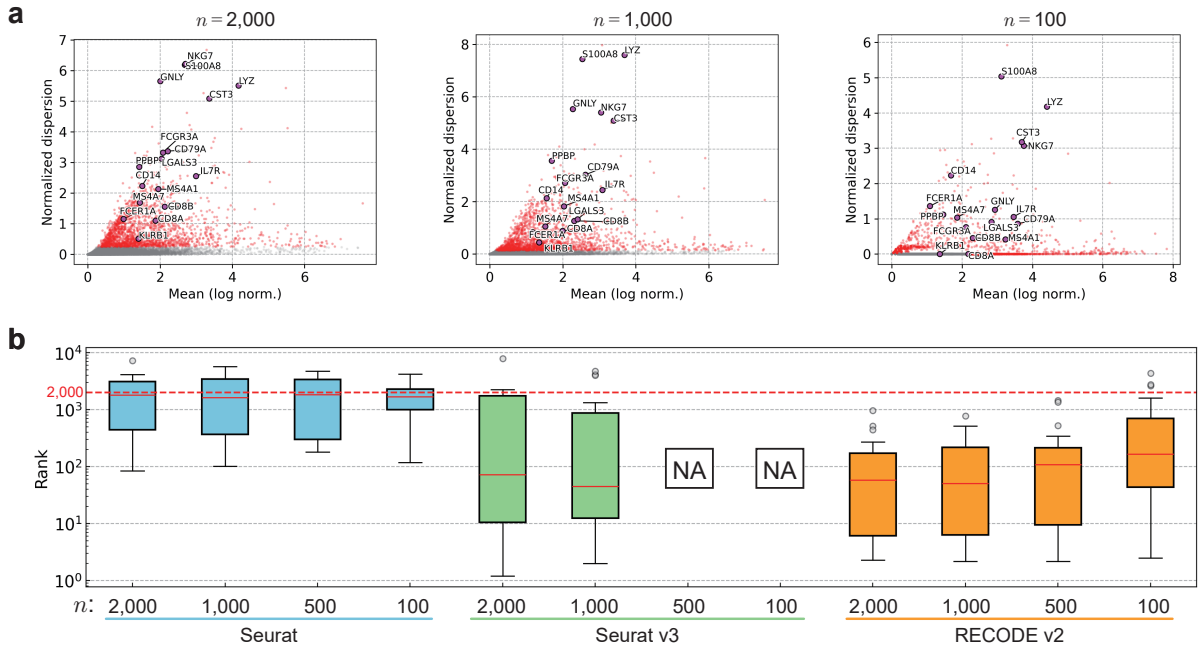

Figure 4: Robustness of HVG selection under cell downsampling. **a**, Scatter plots showing gene-wise mean log-normalized expression (x-axis) versus normalized dispersion (y-axis) after downsampling the number of cells to $n = 2{,}000$, $1{,}000$, and 100. Red points indicate genes selected as HVGs (`n_top_genes` $= 2{,}000$); labeled black points represent known marker genes. **b**, Box plots of marker-gene ranks produced by each method (Seurat, Seurat v3, RECODE v2) under different levels of downsampling ($n = 2{,}000$, $1{,}000$, 500, 100). Each box represents the distribution of average HVG ranks for 17 marker genes, calculated over 10 random subsampling replicates. Effect of cell downsampling on variability profiles and HVG selection. Seurat v3 failed at 500 and 100 cells owing to numerical errors; these cases are indicated as "NA".

RECODE normalization flattens the TNF curve, demonstrating its ability to eliminate the expression-level dependency in variability metrics.

We assessed HVG selection performance using the 3k PBMC scRNA-seq dataset provided by 10x Genomics and evaluated the ranking of 17 known marker genes. Our results demonstrated that the proposed method ranked more marker genes in the top ranks compared with Seurat and Seurat v3 (Figure 2, Table 1). In particular, several marker genes for major cell types, such as CD4 T cells, CD8 T cells, and CD14+ monocytes, were selected within the top 2,000 genes only by RECODE v2. This indicates that RECODE successfully removes technical noise, enabling biologically important low- to mid-expression genes to be properly detected as HVGs.

We then investigated the sensitivity to the parameter c^{TS} used in total-count normalization. While the selection performance of conventional methods varied considerably depending on c^{TS}, RECODE consistently showed stable accuracy across different settings (Figure 3, Table 2). In particular, we adopted $c^{\mathrm{TS}} = 10^5$ in this study, as it provided high accuracy and low variability in marker gene detection. Moreover, increasing c^{TS} led to an improvement in the ranking of low-expression genes, suggesting that RECODE allows flexible HVG selection according to the expression level of interest.

We also evaluated the robustness of each method under cell downsampling. RECODE v2 remained stable even with as few as 100 cells (Figure 4, Table 3). In contrast, Seurat v3 frequently failed when the number of cells was reduced to 500 or fewer, indicating limited robustness under low-sample conditions. For RECODE

v2, mid- to high-expression genes (mean expression ≥ 0.25) retained rankings within the top 2,000 even under severe downsampling, demonstrating the method's resilience. Although the detection of low-expression HVGs requires a sufficient number of cells, RECODE remains accurate for moderate-to-high expression HVGs even when the number of cells is extremely small.

These findings highlight the versatility of RECODE-based HVG selection. For example, adjusting the target sum c^{TS} enables the user to focus on different expression ranges of interest. In addition, the robustness to cell number is advantageous not only in low-quality datasets but also when isolating specific rare cell types for targeted analysis.

Most importantly, because RECODE reduces the noise in the scRNA-seq data itself, the same denoised data used for computing RECODE-denoised variance can be directly reused for downstream analysis. This ensures methodological consistency between HVG selection and downstream analysis, providing a key advantage over conventional approaches that often apply distinct models to each step.

Although we did not evaluate multi-batch datasets in this study, an extended version of RECODE, called iRECODE [5], has been developed to simultaneously reduce technical and batch noises. This suggests potential for accurate HVG selection in multi-batch integration scenarios as well.

In summary, we have shown that RECODE-based HVG selection is theoretically robust to technical noise and achieves higher accuracy and stability than standard methods. This approach provides a practical and flexible tool for cell-type identification and functional gene discovery in scRNA-seq analysis. Future work will focus on evaluating its generalizability to a wider variety of datasets and integration with other preprocessing pipelines.

Methods

Overview of RECODE

We provide an overview of the resolution of the curse of dimensionality (RECODE) algorithm. For details, refer to the Supplementary File of Imoto–Nakamura et al. [7] and the Methods section of Imoto [5].

Let $X = (x_{ij}) \in \mathbb{Z}_{\geq 0}^{d \times n}$ denote the raw count matrix of scRNA-seq data, where x_{ij} represents the observed RNA (UMI) count for gene i in cell j. Here, d and n indicate the number of genes and cells, respectively. Cells with zero total count ($\sum_i x_{ij} = 0$) and genes with zero total count ($\sum_j x_{ij} = 0$) are excluded from the analysis. RECODE begins by applying a noise variance–stabilizing normalization (NVSN), defined as:

$$F(x_{ij}) = \frac{x_{ij}}{t_j \sqrt{n^{-1} \sum_k x_{ik} t_k^{-2}}},$$

where $t_j = \sum_i x_{ij}$ is the total count for cell j. This normalization equalizes the variance of random sampling noise, which originates from the scRNA-seq measurement process and is mathematically modeled as technical noise. This step is also essential to ensure compatibility with the subsequent eigenvalue and eigenvector correction procedures developed in the field of high-dimensional statistics.

Next, the covariance matrix of the normalized matrix $F(X) = (F(x_{ij})) \in \mathbb{R}^{d \times n}$ is computed as $S = F(X)F(X)^{\mathrm{T}}/(n-1)$, where T denotes matrix transpose. The eigenvalue equation $Su_i = \lambda_i u_i$ ($\|u_i\| = 1$) is then solved for $i = 1, \ldots, d$. Yata and Aoshima proposed a modification of eigenvalues [18] defined by:

$$\widetilde{\lambda}_i := \begin{cases} \lambda_i - \dfrac{1}{\min\{n-1, d\} - i + 1} \displaystyle\sum_{k=i+1}^{\min\{n-1,d\}} \lambda_k, & \text{for } i = 1, \ldots, \ell, \\ 0, & \text{otherwise} \end{cases}$$

where $\ell \leq \min\{n-1, d\}$ is a parameter representing the essential dimensionality. This modification improves the convergence of empirical eigenvalues to their population counterparts, indicating mitigation of the curse of dimensionality caused by noise accumulation. In the RECODE algorithm, the optimal value of the essential dimensionality ℓ is automatically determined as

$$\ell^* \coloneqq \min\left\{k\ ;\ \sum_{i=k+1}^{d} \lambda_i \leq d-k\right\}.$$

RECODE version 1 (RECODE v1) modifies the raw count matrix X such that the eigenvalues λ_i match the modified eigenvalues $\widetilde{\lambda}_i$:

$$\mathrm{RECODE}^{\mathrm{v1}}(X) \coloneqq F^{-1}\left[U_{\ell^*}\widetilde{\Lambda}_{\ell^*}^{1/2}\Lambda_{\ell^*}^{-1/2}U_{\ell^*}^{\mathrm{T}}\left(F(X)-\overline{F(X)}\right)+\overline{F(X)}\right],$$

where $U_{\ell^*} = (u_1, \ldots, u_{\ell^*}) \in \mathbb{R}^{d\times\ell^*}$, $\widetilde{\Lambda}_{\ell^*}^{1/2} = \mathrm{diag}(\widetilde{\lambda}_1^{1/2}, \ldots, \widetilde{\lambda}_{\ell^*}^{1/2})$, and $\Lambda_{\ell^*}^{-1/2} = \mathrm{diag}(\lambda_1^{-1/2}, \ldots, \lambda_{\ell^*}^{-1/2})$. Recently, Yata and Aoshima also proposed a correction method for eigenvectors [17]:

$$\widetilde{u}_{ij} \coloneqq \begin{cases} \dfrac{u_{ij}}{\sum_{j=1}^{k_i} u_{i\sigma(j)}^2}, & \sigma^{-1}(j) \leq k_i, \\ 0, & \sigma^{-1}(j) > k_i, \end{cases}$$

where $\sigma : \{1, \ldots, d\} \to \{1, \ldots, d\}$ is a permutation such that $|u_{i\sigma(1)}| \geq \cdots \geq |u_{i\sigma(d)}|$, and k_i is the integer uniquely determined by the following condition:

$$\sum_{j=1}^{k_i} u_{i\sigma(j)}^2 < \frac{\widetilde{\lambda}_i}{\lambda_i}, \quad \sum_{j=1}^{k_i+1} u_{i\sigma(j)}^2 \geq \frac{\widetilde{\lambda}_i}{\lambda_i}$$

This eigenvector correction also alleviates the curse of dimensionality [17]. RECODE version 2 (RECODE v2) incorporates both the modified eigenvalues and eigenvectors:

$$\mathrm{RECODE}^{\mathrm{v2}}(X) \coloneqq F^{-1}\left[\widetilde{U}_{\ell^*}\widetilde{\Lambda}_{\ell^*}^{1/2}\Lambda_{\ell^*}^{-1/2}\widetilde{U}_{\ell^*}^{\mathrm{T}}\left(F(X)-\overline{F(X)}\right)+\overline{F(X)}\right],$$

where $\widetilde{U}_{\ell^*} = (\widetilde{u}_1, \ldots, \widetilde{u}_{\ell^*}) \in \mathbb{R}^{d\times\ell^*}$.

RECODE-based HVG selection

Let $\widetilde{X} = (\widetilde{x}_{ij}) \in \{\mathrm{RECODE}^{\mathrm{v1}}(X), \mathrm{RECODE}^{\mathrm{v2}}(X)\}$ denote the scRNA-seq data denoised by RECODE. We apply two standard preprocessing steps, total-count normalization and log normalization, to the RECODE-denoised data $\widetilde{X}$. Given a non-negative matrix $A \in \mathbb{R}_{\geq 0}^{d\times n}$, the total-count normalization f^{TN} is defined as follows:

$$[f^{\mathrm{TN}}(A)]_{ij} \coloneqq c^{\mathrm{TS}}\frac{a_{ij}}{\sum_i a_{ij}}$$

where c^{TS} is a scaling parameter known as the target sum or size factor. The log normalization f^{LN} is defined as:

$$[f^{\mathrm{LN}}(A)]_{ij} \coloneqq \log(a_{ij}+1)$$

In RECODE-based HVG selection, we first apply total-count normalization followed by log normalization, resulting in the matrix $f^{\mathrm{LN}} \circ f^{\mathrm{TN}}(\widetilde{X})$. We then compute the variance for each gene across cells and select the top `n_top_genes` genes with the highest variances as highly variable genes.

Implementation

RECODE-based HVG selection is implemented in the `screcode` Python package (version 2.1 or later). Given an `AnnData` object [14] representing scRNA-seq data, users can perform HVG selection using the following procedure:

```
import screcode
recode = screcode.RECODE()
adata = recode.fit_transform(adata)
recode.highly_variable_genes(adata)
```

By default, the top 2,000 genes with the highest variance are selected as HVGs. These genes are stored in `adata.var.RECODE_highly_variable`, and the RECODE-normalized variances used for ranking are stored in `adata.var.RECODE_normalized_variance`.

In addition, the RECODE-denoised data after applying total-count and log normalizations is stored in `adata.layers["RECODE_log"]`. Users can use the following commands to set this as the default data matrix and subset it to HVGs:

```
adata.X = adata.layers["RECODE_log"]
adata = adata[:, adata.var.RECODE_highly_variable]
```

This procedure allows seamless reuse of RECODE-denoised data in downstream analyses using Scanpy or other libraries.

For R users, RECODE can be used by calling Python from within R using `reticulate`. For detailed instructions and tutorials, please refer to the GitHub repository at https://github.com/yusuke-imoto-lab/RECODE and the tutorial page at https://yusuke-imoto-lab.github.io/RECODE/Tutorials/index.html.

HVG selection analysis

The peripheral blood mononuclear cell (PBMC) dataset and marker genes used in this study are consistent with those introduced in the Scanpy tutorial "Preprocessing and clustering 3k PBMCs." Specifically, the following 17 marker genes were used: *IL7R* (CD4 T cells), *CD8A* and *CD8B* (CD8 T cells), *CD14*, *LYZ*, *LGALS3*, and *S100A8* (CD14$^+$ monocytes), *MS4A1* and *CD79A* (B cells), *GNLY*, *NKG7*, and *KLRB1* (natural killer cells), *FCGR3A* and *MS4A7* (FCGR3A$^+$ monocytes), *FCER1A* and *CST3* (dendritic cells), and *PPBP* (megakaryocytes).

For comparative evaluation, we applied the widely used methods Seurat and Seurat v3. Both were implemented using the function `scanpy.pp.highly_variable_genes` in Python. For Seurat, total-count normalization followed by log normalization was applied before HVG selection. Except in the section "Effect of total-count parameter on HVG selection," the total-count normalization parameter c^{TS} was set to 10^4. Although Seurat provides an option to filter genes based on mean expression, we disabled this filter to enable direct comparison of the variability statistics between methods.

Seurat v3 was applied directly to the raw count matrix, as required by its negative binomial-based modeling framework. The average expression values shown in Tables 1–3 represent gene-wise means computed after applying total-count normalization with $c^{\mathrm{TS}} = 10^4$ and log normalization to the raw count matrix.

Data Availability

The scRNA-seq dataset used in this study is the 3k Peripheral Blood Mononuclear Cells (PBMCs) dataset provided by 10x Genomics. This publicly available dataset can be accessed from the 10x Genomics website: https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k

Code Availability

The Python and R implementations of the RECODE algorithm, including functionality for highly variable gene selection, are publicly available at the following GitHub repository: https://github.com/yusuke-imoto-lab/RECODE

References

[1] P. Brennecke, S. Anders, J. K. Kim, et al. "Accounting for technical noise in single-cell RNA-seq experiments". *Nature methods* 10.11 (2013), pp. 1093–1095.

[2] S. Fujioka, M. Fujioka, Y. Imoto, et al. "Single-cell multiomic analysis revealed the differentiation, localization, and heterogeneity of IL10+ Foxp3–follicular T cells in humans". *International Immunology* (2025).

[3] S. Gyobu-Motani, Y. Yabuta, K. Mizuta, et al. "Induction of fetal meiotic oocytes from embryonic stem cells in cynomolgus monkeys". *The EMBO Journal* 42.9 (2023). DOI: 10.15252/embj.2022112962.

[4] C. Hafemeister and R. Satija. "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". *Genome Biology* 20.1 (2019). DOI: 10.1186/s13059-019-1874-1.

[5] Y. Imoto. "Comprehensive Noise Reduction in Single-Cell Data with the RECODE Platform". *bioRxiv* (2024). DOI: 10.1101/2024.04.18.590054.

[6] Y. Imoto and Y. Hiraoka. "V-Mapper: topological data analysis for high-dimensional data with velocity". *Nonlinear Theory and Its Applications, IEICE* 14.2 (2023), pp. 92–105. DOI: 10.1587/nolta.14.92.

[7] Y. Imoto, T. Nakamura, E. G. Escolar, et al. "Resolution of the curse of dimensionality in single-cell RNA sequencing data analysis". *Life Sci Alliance* 5.12 (2022). DOI: 10.26508/lsa.202201591.

[8] Y. Murase, R. Yokogawa, Y. Yabuta, et al. "In vitro reconstitution of epigenetic reprogramming in the human germ line". *Nature* 631.8019 (2024), pp. 170–178. DOI: 10.1038/s41586-024-07526-6.

[9] M. Nakajima-Koyama, M. Kabata, J. Lee, et al. "The balance between IFN- and ERK/MAPK signaling activities ensures lifelong maintenance of intestinal stem cells". *Cell Reports* 44.3 (2025), p. 115286. DOI: 10.1016/j.celrep.2025.115286.

[10] F. Nakamura, Y. Nakano, and S. Yamada. "Fine construction of gene coexpression network analysis using GTOM and RECODE detected a critical module of neuroblastoma stages 4 and 4S". *Hereditas* 161.1 (2024). DOI: 10.1186/s41065-024-00342-y.

[11] R. Satija, J. A. Farrell, D. Gennert, et al. "Spatial reconstruction of single-cell gene expression data". *Nature Biotechnology* 33.5 (2015), pp. 495–502. DOI: 10.1038/nbt.3192.

[12] T. Stuart, A. Butler, P. Hoffman, et al. "Comprehensive Integration of Single-Cell Data". *Cell* 177.7 (2019), 1888–1902.e21. DOI: 10.1016/j.cell.2019.05.031.

[13] F. Tang, C. Barbacioru, Y. Wang, et al. "mRNA-Seq whole-transcriptome analysis of a single cell". *Nature Methods* 6.5 (2009), pp. 377–382. DOI: 10.1038/nmeth.1315.

[14] I. Virshup, S. Rybakov, F. J. Theis, et al. "anndata: Access and store annotated data matrices". *Journal of Open Source Software* 9.101 (2024), p. 4371. DOI: 10.21105/joss.04371.

[15] F. A. Wolf, P. Angerer, and F. J. Theis. "SCANPY: large-scale single-cell gene expression data analysis". *Genome biology* 19 (2018), pp. 1–5. DOI: doi.org/10.1186/s13059-017-1382-0.

[16] T. Yachimura, H. Wang, Y. Imoto, et al. "scEGOT: single-cell trajectory inference framework based on entropic Gaussian mixture optimal transport". *BMC Bioinformatics* 25.1 (2024). DOI: 10.1186/s12859-024-05988-z.

[17] K. Yata and M. Aoshima. "Automatic sparse PCA for high-dimensional data". *Statistica Sinica* 35 (2025), pp. 1069–1090. DOI: 10.5705/ss.202022.0319.

[18] K. Yata and M. Aoshima. "Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations". *Journal of Multivariate Analysis* 105.1 (2012), pp. 193–215. DOI: 10.1016/j.jmva.2011.09.002.

[19] S. H. Yip, P. C. Sham, and J. Wang. "Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data". *Briefings in Bioinformatics* 20.4 (2019), pp. 1583–1589. DOI: 10.1093/bib/bby011.

Acknowledgment.

This research was partially supported by the JST CREST (Grant Number JPMJCR24Q1).

Author Contributions

Yusuke Imoto is the sole author of this study. He conceived the research idea, designed the methodology, conducted all analyses, and wrote the manuscript.

Competing Interests

Y. Imoto holds a patent relating to the RECODE method granted by Kyoto University.