The Model-Based Study of the Effectiveness of Reporting Lists of Small Feature Sets Using RNA-Seq Data

Cancer Informatics
Volume 16: 1–15
© The Author(s) 2017
DOI: 10.1177/1176935117710530
SAGE

Eunji Kim[1], Ivan Ivanov[2], Jianping Hua[3], Johanna W Lampe[4], Meredith AJ Hullar[4], Robert S Chapkin[5] and Edward R Dougherty[1,3]

[1]Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX, USA. [2]Department of Veterinary Physiology & Pharmacology, Texas A&M University, College Station, TX, USA. [3]Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University, College Station, TX, USA. [4]Public Health Sciences Division, Cancer Prevention, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [5]Program in Integrative Nutrition & Complex Diseases, Texas A&M University, College Station, TX, USA.

ABSTRACT: Ranking feature sets for phenotype classification based on gene expression is a challenging issue in cancer bioinformatics. When the number of samples is small, all feature selection algorithms are known to be unreliable, producing significant error, and error estimators suffer from different degrees of imprecision. The problem is compounded by the fact that the accuracy of classification depends on the manner in which the phenomena are transformed into data by the measurement technology. Because next-generation sequencing technologies amount to a nonlinear transformation of the actual gene or RNA concentrations, they can potentially produce less discriminative data relative to the actual gene expression levels. In this study, we compare the performance of ranking feature sets derived from a model of RNA-Seq data with that of a multivariate normal model of gene concentrations using 3 measures: (1) ranking power, (2) length of extensions, and (3) Bayes features. This is the model-based study to examine the effectiveness of reporting lists of small feature sets using RNA-Seq data and the effects of different model parameters and error estimators. The results demonstrate that the general trends of the parameter effects on the ranking power of the underlying gene concentrations are preserved in the RNA-Seq data, whereas the power of finding a good feature set becomes weaker when gene concentrations are transformed by the sequencing machine.

KEYWORDS: Classification, feature ranking, ranking power, RNA-Seq

RECEIVED: December 28, 2016. **ACCEPTED:** April 13, 2017.

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 1017 words, excluding any confidential comments to the academic editor.

TYPE: Methodology

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded in part by grants from the National Institutes of Health, U01 CA-162077, R35CA-197707, and P30ES-023512.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Eunji Kim, Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX 77843, USA.
Email: eunjikim@tamu.edu

Introduction

Ranking feature sets for phenotype classification based on gene expression can be viewed as gene selection and is a key issue for cancer informatics. Because ranking feature sets is often based on error estimates of the designed classifiers and error estimators based on training data from small samples tend to perform poorly, exhibiting optimistic bias or high variance, a feature set with a low error estimate cannot be automatically declared to be credible. Also, it is important to choose an error estimator which yields a reliable ranking for the feature sets.[1] Furthermore, when confronted with a small sample, feature selection algorithms often fail to find good feature sets. The problem is exacerbated for high-dimensional data, ie, data sets with feature sets of high cardinality. It is difficult to find a good feature set in the small sample setting even when one uses a mathematically favorable gene concentration/expression model.[2] These observations suggest that it is prudent to report a list of potential feature sets rather than attempting to find the best feature set. In addition to the unreliability of feature selection and error estimation, the accuracy of classification depends on the manner in which the phenomena are transformed into data by the measurement technology. High-throughput sequencing technologies such as next-generation sequencing (NGS) have recently emerged as popular tools to quantify gene transcripts. However, NGS technologies pose new computational and statistical challenges because their applications result in nonlinear transformations of the underlying gene concentration distributions. A recent study showed that an NGS pipeline could lead to transformation degradation in classification performance.[3] In this article, we address the effects of the nonlinear transformation induced by the sequencing machine and the choice of error estimators on feature-set ranking.

The development of NGS technologies enables simultaneous measurements of the abundance of messenger RNA (mRNA) transcripts, and such information can be used to detect differential gene expression and design gene expression–based classifiers for phenotypic discrimination and medical diagnosis or prognosis. RNA-Seq provides discrete counting measurements for the gene expression levels.[4] All RNA-Seq data generation follows a similar protocol, starting with shearing samples to generate millions of small RNA fragments. These fragments are then converted to complementary DNA (cDNA), and the adapter sequences are ligated to their ends.

This collection, referred to as a library, is then sequenced, which produces millions of short sequence reads that correspond to individual cDNA fragments. Finally, those reads are mapped to a reference genome. The number of reads mapped to a gene on the reference genome defines the count data, which is a discrete measure of the respective gene expression levels.

Much of the literature concerning the statistical representation of RNA-Seq data models it via a negative binomial[5,6] or Poisson[7] distribution. The Poisson model is parameterized by its mean and it is already known that RNA-Seq data may exhibit more variability than the single Poisson distribution parameter. The negative binomial distribution can mitigate this overdispersion problem, allowing the variance to exceed the mean; however, when dealing with a relatively small number of samples, it is difficult to accurately estimate the dispersion parameter of the negative binomial model. Therefore, in this article, we focus on a hierarchical multivariate Poisson model.[3] Specifically, gene concentration levels are extracted from a log-normal distribution, and their subsequent processing by the sequencing instrument is modeled via a Poisson process. The hierarchical model is not as restrictive as the simple Poisson model and can be considered as a compromise between the Poisson and negative binomial models in the small-sample setting.[8] The simulated NGS data follow a conditionally Poisson distribution, and the marginal distribution of the data is a mixture of Poisson and Gaussian distributions.

Although multivariate data offer the potential for finding features for phenotypic discrimination, large-scale and high dimensionality classification problems with small sample sizes can result in overfitting of the data. A variety of feature selection algorithms for classification have been proposed over the past decades.[9,10] Feature selection has inherent problems due to its combinatorial nature and sampling procedures. To select a subset of k features out of n potential features and be assured that it provides an optimal classifier with minimum error among all optimal classifiers for subsets of size k, all $\binom{n}{k}$ possible sets must be checked to guarantee that the best one is selected.[11] In other words, nothing but an exhaustive search can assure finding the best feature set. In practice, feature selection must proceed from sample data, which leads to the well-known peaking phenomenon, ie, the tendency of achieving improved classification performance with an increasing number of features only to a point, beyond which more features lead to degradation of the classification accuracy.[12-16] Therefore, employing too many features in a small-sample setting yields poorer classification accuracy, thereby leading to the need for feature selection. This raises a critical question: can one expect a feature selection algorithm to yield a feature set whose error is close to that of an optimal feature set?

A good feature selector is expected to report a list of feature sets without missing the true target. Thus, ranking of feature sets becomes a key issue for classification. Unfortunately, for small samples, error estimators deployed to perform the ranking of the feature sets suffer from different degrees of imprecision. Moreover, there is little correlation between the errors of the selected feature set and a close-to-optimal feature set.[17] When the number of samples is small, using resampling-based classifier error estimators such as cross-validation and bootstrap is risky owing to the substantial variance[18] and lack of regression with the true error,[18-21] which is exacerbated in the presence of feature selection.[22,23] Hence, it is important to choose a computationally feasible error estimator that yields rankings that better correspond to rankings produced by the true errors.

Often, when ordering a list of feature sets based on the estimated errors, the smaller estimates tend to be biased optimistically and the larger estimates tend to be biased pessimistically.[2] Thus, reporting a list of feature sets is preferred compared with providing a single good feature set, the idea being that some in the list of top-performing feature sets will be close to optimal.[2] This approach assures that there is at least one feature set on the list whose true classification error is within some given tolerance of the best feature set with high probability. Given the list, one can either focus on the feature sets in the list for further sampling or take a classical wet-lab approach to determine which ones are predictive of the phenotype of interest.[2]

In this study, we investigate the effects of the nonlinear transformation induced by NGS technology and the choice of error estimators on feature-set ranking. Quantification of changes in feature-set lists due to a measurement technology requires a baseline to compare, ie, underlying gene concentration as the biological ground truth. This can be accomplished via simulated data experiments. For this purpose, we used a model-based approach and provided a distribution from which the synthetic data arise. We also consider an application of the proposed methodology to real RNA-Seq data as an example of one possible way to derive power curves that estimate the goodness of the feature-set ranking under user-defined settings.

We focus on the linear discriminant analysis (LDA) classification rule, and our work is neither a comparison study of different pattern classifiers nor a model selection study. The rationale for focusing on LDA classifiers is based on our previous studies.[1,24] The performance of 7 different classification rules on real patient data was compared in terms of the expected classification error, for different sample sizes and dimensionality.[24] Classification rules considered were LDA, quadratic discriminant analysis, nearest mean classification, 1-nearest neighbor, 3-nearest neighbor, Classification And Regression Trees (CART) with a stopping rule that ends splitting when there are 6 or fewer sample points in a node, and a neural network with 4 nodes in the hidden layer. As a result, LDA has proved to be a very robust classification rule, which is effective for a wide range of sample sizes, and therefore, we focus on the LDA classification rule.

Methods

Ranking power

The *ranking power* is a measure of the goodness of a ranked list of classification feature sets and is defined as follows[2]:

$$\Delta_{D,d}^{n,r}(m) = P(\varepsilon_1 - \varepsilon_0 < r)$$

where ε_1 is the lowest test error for the feature sets in a ranked list of length m sorted by their estimated errors, and ε_0 is the test error of the classifier computed for the Bayes features. Specifically, to compute the ranking power, consider all the possible feature sets of size d among the number D of total features. Then, rank them according to their estimated errors and obtain the top m feature sets, $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_m$. Define ε_1 as the lowest test error of the classifier among the m feature sets considered. The ith lowest estimated error $\hat{\varepsilon}_{(i)}$ corresponds to the feature set $\mathcal{F}_{(i)}$, but the ith lowest test error $\varepsilon_{(i)}$ will likely not correspond to the feature set $\mathcal{F}_{(i)}$.

The ranking power provides the probability that given a ranked list of m feature sets, there is at least one feature set in that list with an error that is close to that of the best feature set. The ranking power depends on the list length m, the total number D of features, the number d of selected features, and sample size n. The original ranking power definition takes into account the difference between the smallest test error of the classifier in a given list of feature sets and the respective test error of the Bayes feature set. However, it is often desirable to consider the magnitude of the Bayes error. Thus, we propose the following modification to the ranking power definition:

$$\Delta_{D,d}^{n,c}(m) = P(\varepsilon_1 - \varepsilon_0 < c \cdot \varepsilon_0)$$

This modification allows for an explicit comparison of the difference between ε_1 and ε_0 with the magnitude of the test error of the Bayes feature set as represented by the parameter c. For example, $c = 0.01$ indicates that we are only interested in ranked lists of features sets where the feature set with the smallest test error differs from the test error of the Bayes feature set by less than 1% of the test error of the Bayes feature set. For any given ε_0, there is a clear relationship between the value of r in the original definition of the ranking power and the parameter c in the modified version above. Thus, for the purpose of comparing our simulation results with those from the previous study by Zhao et al,[2] we report the values of the parameter r.

Ranking power of the gene expression concentration generated from the multivariate normal (MVN) distribution[25,26] is computed by the probability of the following inequality:

$$\varepsilon_{1,MVN} - \varepsilon_{0,MVN} < c \cdot \varepsilon_{0,MVN}$$

where $\varepsilon_{1,MVN}$ is the lowest test error for the feature sets in the MVN ranked list and $\varepsilon_{0,MVN}$ is the test error of the Bayes feature set in the MVN model. In the same way, ranking power of the NGS data is calculated by the probability of the following:

$$\varepsilon_{1,NGS} - \varepsilon_{0,NGS} < c \cdot \varepsilon_{0,NGS}$$

The same Bayes feature set in the MVN model is used as the Bayes feature set of the NGS model and $\varepsilon_{0,NGS}$ is the respective test error of the Bayes feature set in the NGS data. The smallest test error for the feature sets in the NGS ranked list is $\varepsilon_{1,NGS}$.

Length of extensions

Gene expression concentration is the biological ground truth and has often been modeled by the MVN distribution.[25,26] We use the MVN model to assess the effects of the NGS transformation on the ranking power and the composition of the ranked lists of feature sets. In general, when one desires to compare 2 ranked lists of feature sets, one is interested how a particular feature set is ranked in each one of the 2 lists. Although there are several possible ways to measure this difference in the ranking, we focus on the ranking of a top-performing feature set from 1 of the 2 lists in the other list. To achieve the desired comparison, we introduce the following notation: $\mathcal{F}_{MVN}$ denotes the feature set ranked at the top in the list of feature sets obtained using the MVN model of gene concentrations and the rank of $\mathcal{F}_{MVN}$ in the respective NGS list is denoted as τ_{NGS}. Similarly, τ_{MVN} is the rank of the top feature set $\mathcal{F}_{NGS}$ from the NGS list in the respective MVN ranked list of feature sets.

Bayes features

The Mahalanobis distance provides a way to calculate the Bayes error. If class densities are Gaussian, the Bayes error can be simply calculated using only sample mean vectors μ_i and sample covariance matrices Σ_i of class i. The Mahalanobis distance Δ is given as follows:

$$\Delta = \sqrt{(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)}$$

where Σ denotes the average covariance matrix given by $\Sigma = P(c_1) \cdot \Sigma_1 + P(c_2) \cdot \Sigma_2$ and $P(c_i)$ is a priori class probability of class $i = 1, 2$. Equal prior probabilities for the classes and equal covariance matrices are assumed in our model. Therefore, the Bayes error for any feature set $\mathcal{F}$ of size d is $\Phi(-\Delta/2)$, where Φ is the standard normal cumulative distribution function. $\mathcal{F}_{bayes}$ denotes the feature set having the largest Mahalanobis distance and, accordingly, the minimum Bayes error.

Bayes features of a hierarchical model cannot be easily found as in the Gaussian case. Simulated NGS data are the mixed form of Poisson and Gaussian distributions, so there is no analytical formula for the Bayes error. The Bayes error of the hierarchical model can be estimated using Monte Carlo sampling. In this study, Bayes features of the MVN are used as the Bayes features of the NGS data in the biological context. Although Bayes features of the MVN are not equal to those of the transformed data, MVN Bayes features reflect the biological ground-truth markers.

The Models for Gene Concentrations and NGS Data

Two different types of synthetic data are generated for simulation experiments: (1) actual gene expression concentration, called MVN and (2) Poisson-transformed MVN data, denoted as NGS, which emulate NGS reads.

Multivariate Gaussian model

Gene concentration levels can be modeled using a log-normal distribution,[27-29] and the hybrid multivariate Gaussian model proposed in Zhao et al[2] is adopted in this article. Genes/features are categorized into 2 groups: markers and nonmarkers. There is a total of $D = \upsilon + \eta$ features and υ and η represent the number of markers and nonmarkers in the model, respectively. Markers resemble genes associated with diseases and they have 2 class-conditional Gaussian distributions with equally likely classes and common covariance matrix Σ. The mean vectors for the markers are $\mu_0 = m_0 \times (0, 0, \ldots, 0)^T$ and $\mu_1 = m_1 \times (a_1, a_2, \ldots, a_\upsilon)^T$ for class 0 and class 1, respectively, where m_0 and m_1 are scalars and υ denotes the total number of marker features generated. To mimic real experimental situations, where every marker performs well but not exactly the same, all elements of vector μ_1 are not equal to one another. μ_1 is an equally spaced vector with $a_1 = 1$ and $a_\upsilon = 0.8$. The covariance matrix Σ is blocked and each block Σ_ρ has variance σ_μ^2 along the diagonal and correlation coefficient ρ off the diagonal:

$$\Sigma = \begin{bmatrix} \Sigma_\rho & 0 & \cdots & 0 & 0 \\ 0 & \Sigma_\rho & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & \Sigma_\rho & 0 \\ 0 & 0 & \cdots & 0 & \Sigma_\rho \end{bmatrix}$$

$$\text{where } \Sigma_\rho = \sigma_\mu{}^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

Different blocks correspond to different gene regulatory pathways[25,26] and model the assumption that groups of genes in the same pathway are biologically or functionally correlated and interacting with each other, whereas genes in different pathways are uncorrelated. Nonmarkers are uncorrelated and modeled as 1-dimensional zero-mean random Gaussian noise, with a total of η features.

The hierarchical multivariate Poisson model

The gene expression levels in NGS data are measured by the number of reads that are mapped to the corresponding gene in the reference genome. Thus, NGS-type data values are discrete with nonnegative integers. Several statistical models for NGS data based on the negative binomial model or Poisson distribution have been proposed.[5-7] In this article, the hierarchical multivariate Poisson model[3] is adopted. It assumes that the sequencing facility samples mRNA concentration through a Poisson process, and the expected number of reads is the mean of the Poisson distribution. Read count for a sample point *i* and the *j*th gene is $X_{i,j}$. It is obtained by the generalized linear model[30] for a given s_i:

$$p\left(X_{i,j} \mid s_i\right) \sim Poisson\left(s_i \exp\left(\lambda_{i,j} + \theta_{i,j}\right)\right)$$

where s_i denotes the sequencing depth for the *i*th sample point in the model and is randomly generated from a uniform distribution, $U(\alpha, \beta)$, where $\alpha > 0$ and $\beta > \alpha$. To generate count data for RNA-Seq reads, the hybrid Gaussian model is fed to the pipeline as $\lambda_{i,j}$, the *j*th gene expression level in a sample point *i*. The value is perturbed by $\theta_{i,j}$, which reflects technical effects associated with the experiment and is drawn from a Gaussian distribution:

$$\theta_{i,j} \sim N\left(0, \left|m_1 - m_0\right| COV\right)$$

where *COV* is the coefficient of variation. Once the NGS data are generated, the features are normalized in a way that each feature is zero mean and unit standard deviation across all the sample points.

Implementation

Figure 1 presents a general overview of the simulation employed herein. General implementation follows a similar simulation procedure proposed in Zhao et al[2]:

1. Set up a hybrid Gaussian model with υ marker features and η nonmarkers to yield $D = \upsilon + \eta$ features. Find the Bayes feature set $\mathcal{F}_{bayes}$ of size *d*.
2. Generate a large test set of independent data using the MVN model.
3. For every feature set of size d, design an LDA classifier and compute its estimated and test errors. Compute the test error $\varepsilon_{0,MVN}$ for $\mathcal{F}_{bayes}$.
4. Rank all feature sets by their estimated errors based on the training data and select the top m of them to form the MVN ranked list.
5. Let $\varepsilon_{1,MVN}$ be the lowest test error in the top m list. If $\varepsilon_{1,MVN} - \varepsilon_{0,MVN} < c \cdot \varepsilon_{0,MVN}$, set $count_{MVN} = count_{MVN} + 1$.

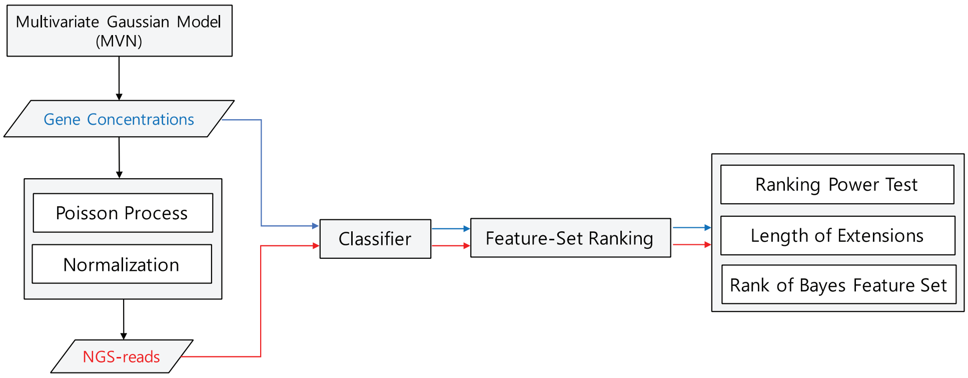


Figure 1. An overview of the simulation. Two different types of synthetic data are generated: (1) multivariate normal (MVN) and (2) next-generation sequencing (NGS). Data sets are generated from a multivariate Gaussian model and a hierarchical multivariate Poisson model. Subsequently, the data sets are fed to the same test modules: classification, error estimation, and feature-set ranking.

Table 1. Model parameters for generating synthetic data.

EXP NO.	D	υ	n	σ_μ^2	ρ	B	d
1	{50,100,150}	{5,10,20}	40	1	0.8	5	{2,3}
2	150	10	{40,80,120}	1	0.8	5	2
3	150	10	40	{0.5,1,2}	0.8	5	2
4	150	10	40	1	{0.1,0.5,0.8}	5	2
5	150	10	40	1	0.8	{2,5,10}	2
6	100	5	40	1	0.8	5	2
	200	10					
	300	15					

6. The MVN data generated from steps (1) and (2) are fed to the Poisson transformation pipeline to obtain the NGS data.
7. Repeat steps (3) through (5) for the NGS data. Use the same Bayes feature set $\mathcal{F}_{bayes}$ to compute the test error $\varepsilon_{0,NGS}$. If $\varepsilon_{1,NGS} - \varepsilon_{0,NGS} < c \cdot \varepsilon_{0,NGS}$, set $count_{NGS} = count_{NGS} + 1$.
8. Repeat steps (1) through (7) N times to get $\Delta^{n,c}_{D,d\,MVN}(m) = count_{MVN} / N$ and $\Delta^{n,c}_{D,d\,NGS}$ (m) = $count_{NGS} / N$.
9. Compare MVN and NGS lists and obtain τ_{MVN} and τ_{NGS}.
10. Find the ranks of Bayes feature sets in the MVN and NGS lists. Denote them as B_{MVN} and B_{NGS}, respectively.

Simulation Parameters

RNA-Seq technology can provide different numbers of reads per sample, depending on many factors, such as quality of the sample, the desired coverage, and sample multiplexing. To deal with this issue, a previous study[3] examined a variety of ranges of the sequencing depth and NGS-read counts for real RNA-Seq experiments, and the parameters α and β are chosen accordingly. Therefore, our selections for the model parameters reflect how real data behave because they take into account a range of NGS-read counts one can expect from real data. Our study is model based and we do not focus on the problems of inference or parameter estimation from data. Thus, we adopt the parameters' ranges/values from the work by Ghaffari et al.[3]

Parameters for the sequencing depth $s_i \sim U(\alpha, \beta)$ are set to $\alpha = 9$, $\beta = 11$, and $COV = 0.05$; $m_0 = 0$, $m_1 = 1$ are used for the distribution of technical effects, $\theta_{i,j}$. Simulation setups and the list of parameters used for the multivariate Gaussian model are provided in Table 1. Experiment numbers in Table 1 correspond to the parameter setting of each experiment in Table 2 and Supplementary Tables 1 and 2. Absolute bound r is used for comparisons between our results and those in Zhao et al.[2] Corresponding values for the relative significance of the difference c are provided in Table 2.

Because there is no closed form to calculate the true errors of designed classifiers, large independent test sets are generated. When using independent test data, the root mean square between the true and estimated error is bounded above by $1/2\sqrt{n_{test}}$.[3] Test sample of size $n_{test} = 10000$ is generated and samples are divided equally between the 2 classes.

Table 2. Mean of ε_0 in the multivariate normal and next-generation sequencing list and relative differences between ε_0 and ε_1 with respect to ε_0.

EXP NO.	PARAMETERS		$E[\varepsilon_{0,MVN}]$	c_{MVN} (r = 0.03)	$E[\varepsilon_{0,NGS}]$	c_{NGS} (r = 0.03)
1 (d = 2)	υ = 5	D = 50	0.2557	0.1173	0.2993	0.1002
		D = 100	0.2559	0.1172	0.2990	0.1003
		D = 150	0.2558	0.1173	0.2994	0.1002
	υ = 10	D = 50	0.2558	0.1173	0.2990	0.1003
		D = 100	0.2557	0.1173	0.2988	0.1004
		D = 150	0.2558	0.1173	0.2992	0.1003
	υ = 20	D = 50	0.2559	0.1172	0.2992	0.1003
		D = 100	0.2556	0.1174	0.2988	0.1004
		D = 150	0.2559	0.1173	0.2992	0.1003
1 (d = 3)	υ = 5	D = 50	0.2557	0.1173	0.2993	0.1002
		D = 100	0.2559	0.1172	0.2990	0.1003
		D = 150	0.2558	0.1173	0.2994	0.1002
	υ = 10	D = 50	0.2558	0.1173	0.2990	0.1003
		D = 100	0.2557	0.1173	0.2988	0.1004
		D = 150	0.2558	0.1173	0.2992	0.1003
	υ = 20	D = 50	0.2559	0.1172	0.2992	0.1003
		D = 100	0.2556	0.1174	0.2988	0.1004
		D = 150	0.2559	0.1173	0.2992	0.1003
2	n, bresub	40	0.2558	0.1173	0.2992	0.1003
		80	0.2498	0.1201	0.2956	0.1015
		120	0.2479	0.1210	0.2947	0.1018
	n, loo	40	0.2558	0.1173	0.2994	0.1002
		80	0.2497	0.1201	0.2958	0.1014
		120	0.2479	0.1210	0.2946	0.1018
3	σ_μ^2	0.5	0.1734	0.1731	0.2190	0.1370
		1	0.2558	0.1173	0.2992	0.1003
		2	0.3266	0.0919	0.3737	0.0803
4	ρ	0.1	0.2557	0.1173	0.2995	0.1002
		0.5	0.2559	0.1172	0.2991	0.1003
		0.8	0.2558	0.1173	0.2992	0.1003
5	B	2	0.2626	0.1142	0.3042	0.0986
		5	0.2558	0.1173	0.2992	0.1003
		10	0.2535	0.1184	0.2972	0.1009
6	D = 100, υ = 5		0.2559	0.1172	0.2990	0.1003
	D = 200, υ = 10		0.2557	0.1173	0.2991	0.1003
	D = 300, υ = 15		0.2558	0.1173	0.2996	0.1001

Abbreviations: BRESUB, bolstered resubstitution; LOO, leave-one-out.

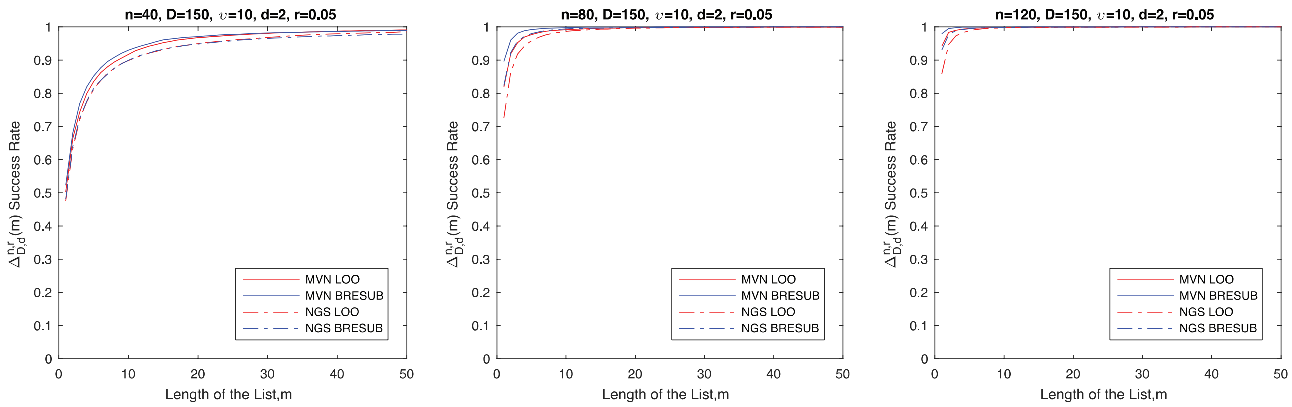


Figure 2. Power curves for different error estimators and sample size *n*. Solid: multivariate normal (MVN), dashed: next-generation sequencing (NGS), red: leave-one-out (LOO), and blue: bolstered resubstitution (BRESUB).

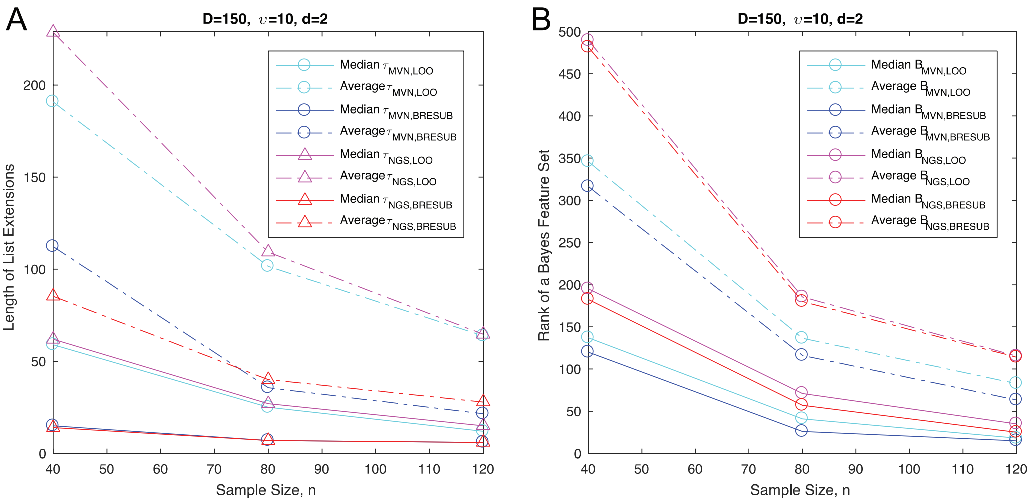


Figure 3. Effects of different error estimators and sample size n on (A) length of list extensions and (B) rank of a Bayes feature set. Solid: median; dashed: average; cyan: MVN, LOO; blue: MVN, BRESUB; pink: NGS, LOO; red: NGS, BRESUB. BRESUB indicates bolstered resubstitution; LOO, leave-one-out, MVN, multivariate normal; NGS, next-generation sequencing.

Results

Synthetic data

Effects of $D, \upsilon, n, \sigma^2_\mu, \rho, B, d$ and proportion of υ to D are studied.

Effects of error estimators. Previous literature shows that cross-validation methods, especially leave-one-out (LOO) estimators, display large variance.[31,32] This variance results in a widely dispersed deviation between the true and estimated errors of a classifier, thereby making cross-validation unreliable for ranking feature sets in the small-sample setting. It has been shown that bolstering and resubstitution-based feature ranking outperform LOO cross-validation–based feature ranking for discovering top-performing feature sets for classification when using small samples.[1] Previous studies[1,33] are based on a Gaussian mixture model and microarray-based patient data. In this article, we examine the effects of error estimators on the ranking of feature sets of RNA-Seq data. Two different error estimators, bolstered resubstitution (BRESUB) and LOO, are used to sort the lists. Figure 2 indicates that the hit rate of finding a good feature set in the list sorted by BRESUB error estimators is higher than the success rate of the LOO-based list. Figure 3A shows that both $\tau_{MVN,LOO}$ and $\tau_{NGS,LOO}$ are larger than $\tau_{MVN,BRESUB}$ and $\tau_{NGS,BRESUB}$, respectively, which implies that LOO mixes up the orders more harshly than BRESUB. Moreover, Figure 3B shows that the ranks of Bayes feature sets in the LOO-based list are larger than that of the bolstered resubstitution-based list. All of these results suggest that LOO estimators perform poorly with RNA-Seq data, producing less accurate ranking orders compared with the list sorted by BRESUB error estimators.

Effects of the sample size, n. A larger sample size generally leads to better performance of classification and ranking feature sets. The results of NGS shown in Figures 2 and 3 are in accordance with this expectation. As sample size increases, ranking power curves for NGS are also elevated. For both types of data, monotonic decrease in extension length and Bayes rank in median is observed as sample size gets larger.

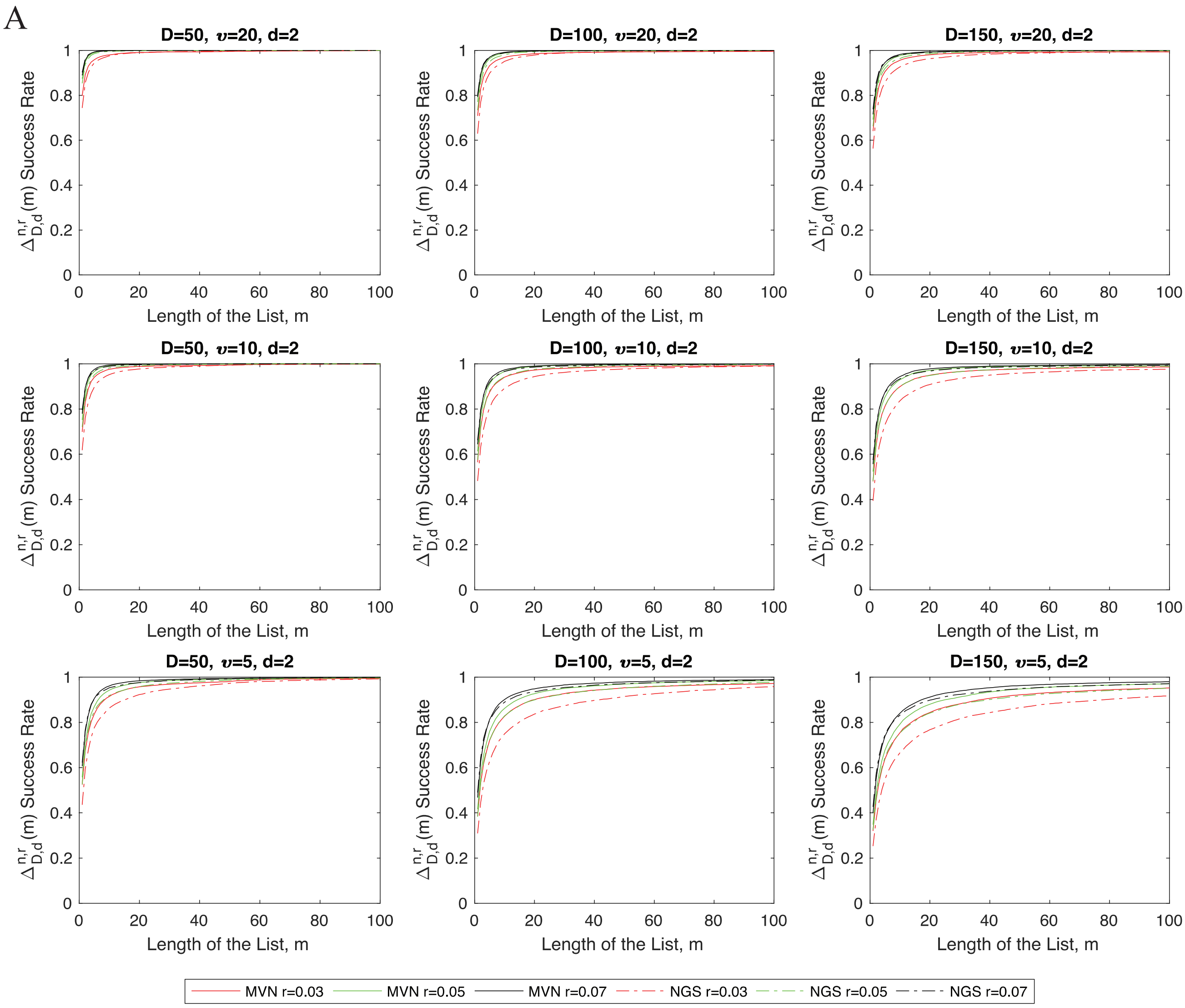
Effects of the total number of features D and the number of marker features, υ. Figure 4A represents the effects of the total number D of features and the number υ of marker features on the ranking power curves when the final number d of selected features is 2. The ranking power curves for $d = 3$ are provided in Figure 4B. Zhao et al[2] have shown that the power curves are lowered in the MVN model as the total number of features increases. Figure 4A and B shows analogous results in the RNA-Seq model. The plots also indicate that for a fixed value of D, the power increases as υ increases. This is not surprising because the prior information provided by the biologist becomes richer, containing more markers.

Figure 5 illustrates the effects of increasing D and υ on τ_{MVN} and τ_{NGS}. As D gets larger, a monotonic increase in median and average extension length is observed in both models. In Figure 5B and D, no obvious trend can be discerned in terms of the mean, nor is there any consistency. However, the median extension length exhibits a slight increasing trend.

Histograms of length of extensions and the rank of the Bayes feature set are illustrated in Figure 6. It is a skewed heavy-tailed distribution with the mean farther out in the long tail than the median. Because the mean is highly vulnerable to outliers, it should be interpreted with caution when extreme values are present. Focusing on the median values, which are less affected by outliers, an increasing trend of median extension length is exhibited as υ gets larger. This is because it becomes more competitive to rank at the top as more markers enter into the data and the one which occupies the top becomes more variable, thereby resulting in the increase in extension length to match 2 lists.

The monotonic increase in median rank of Bayes feature pair is presented in Figure 7B and D, as υ increases. As more marker features are included, there are more feature pairs which perform as well as a Bayes feature pair. Therefore, the Bayes feature set is no longer a unique and distinguishing feature pair, and the multitude of marker features obscures the Bayes feature pairs.

Effects of the variance σ_μ^2 in the marker model. Figure 8A shows the effect of the variance in the marker model. Higher variance



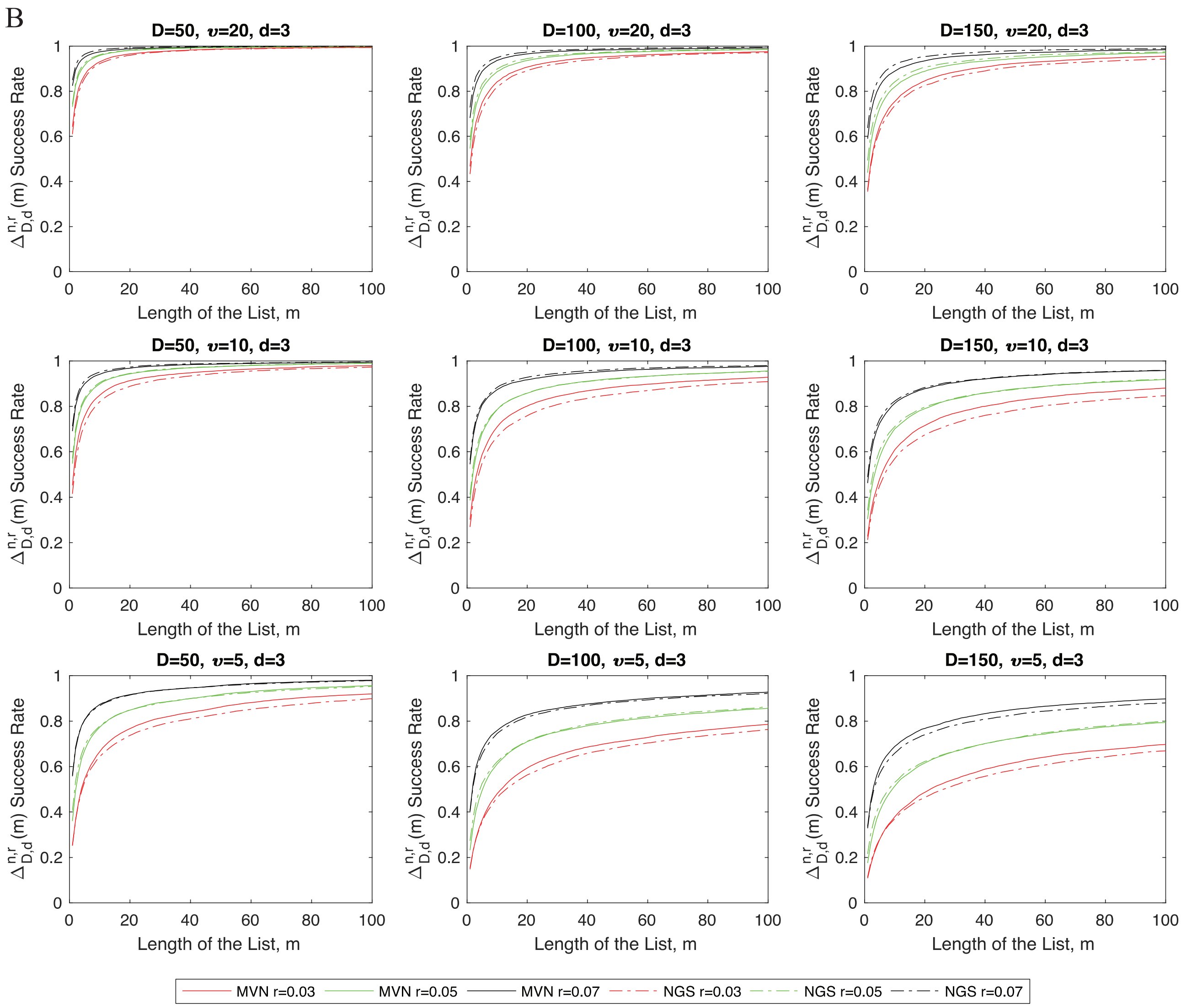


Figure 4. Power curves for different D and υ when (A) $d = 2$ and (B) $d = 3$. MVN indicates multivariate normal; NGS, next-generation sequencing.

results in larger overlaps of the 2 distributions, which leads to degradation of classification performance and increasing difficulty of finding top-performing feature sets. Therefore, the success rates of both models decrease as variance increases. When $\sigma_\mu^2 = 2.0$, the power curve of the NGS model is higher than that of MVN. This does not necessarily mean that it is better to use the RNA-Seq model to detect a good feature set when the problem is difficult. A better interpretation is that mixing is so extensive that even the underlying gene concentrations are useless for finding a good feature set. Figure 8A also shows that both extension length and rank of Bayes feature sets increase as variance increases.

Effects of the correlation ρ in the covariance matrix. Zhao et al[2] have shown that a higher correlation makes it slightly harder to find good features in the MVN model. Figure 8B indicates that the same applies to the RNA-Seq model. As ρ increases, ranking power of both MVN and RNA-Seq models decreases. Curves for median extension length and the rank of Bayes feature sets are almost flat with respect to the correlation.

Effects of the number of blocks B in the covariance matrix. Different blocks represent different metabolic/biologic pathways, and as the number of blocks increases, genes may become spread among more pathways and may increase the power to find good features. Zhao et al[2] showed that it is easier to find good features with more blocks. Figure 8C demonstrates that the ranking power becomes higher as B increases in the RNA-Seq model. When there are only 2 blocks, RNA-Seq exhibits a higher success rate compared with the MVN model, but it is very unlikely to have only 2 pathways in real data. Figure 8C shows decreasing extension length with larger B, which is consistent with the power curve. No specific trend is observed in the rank of the Bayes feature set with respect to B.

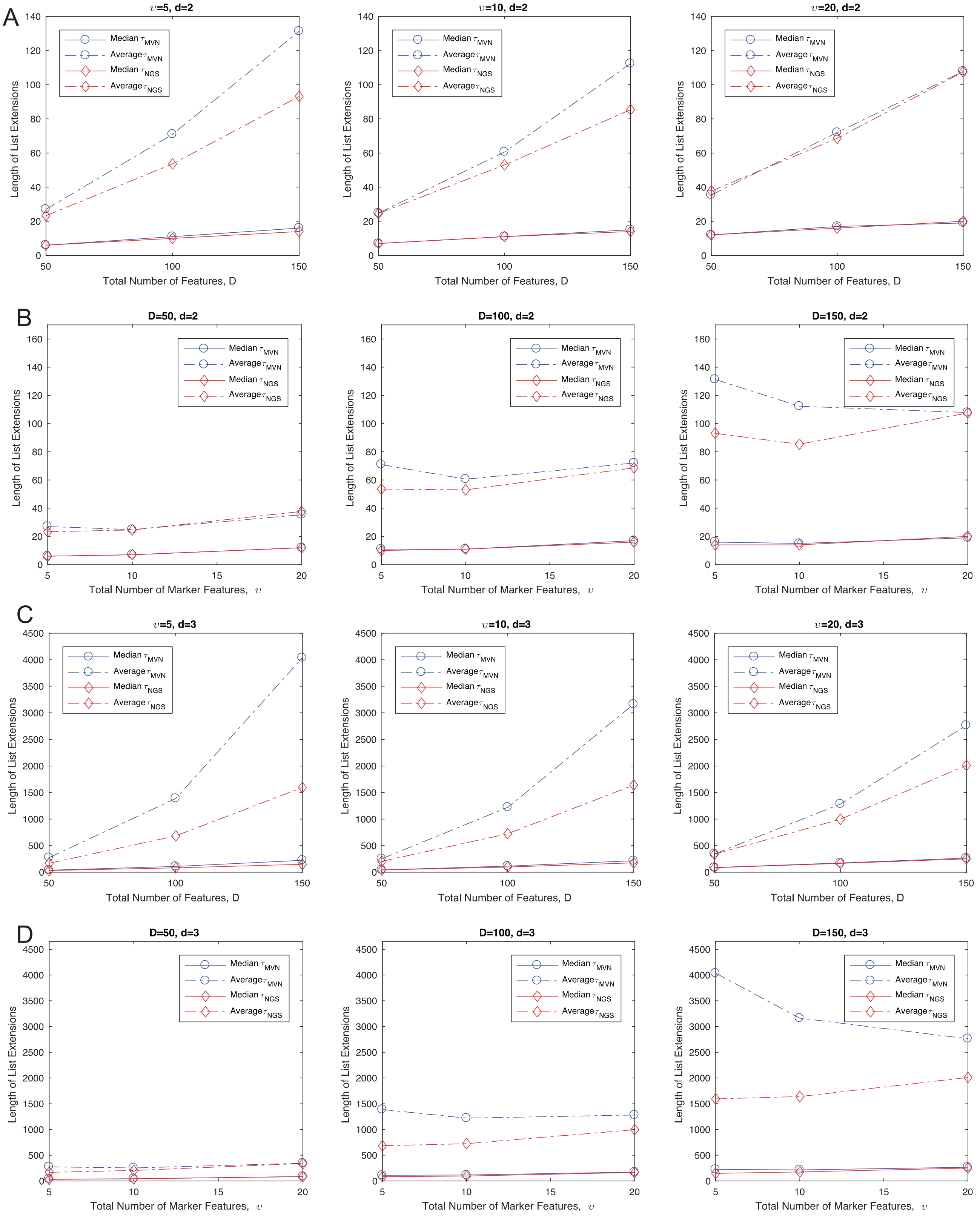


Figure 5. Effects of different *D* and υ on length of list extensions for $d = 2$ are presented in (A) and (B). Graphs for $d = 3$ are presented in (C) and (D). Solid: median, dashed: average, blue: MVN, red: NGS. MVN indicates multivariate normal; NGS, next-generation sequencing.

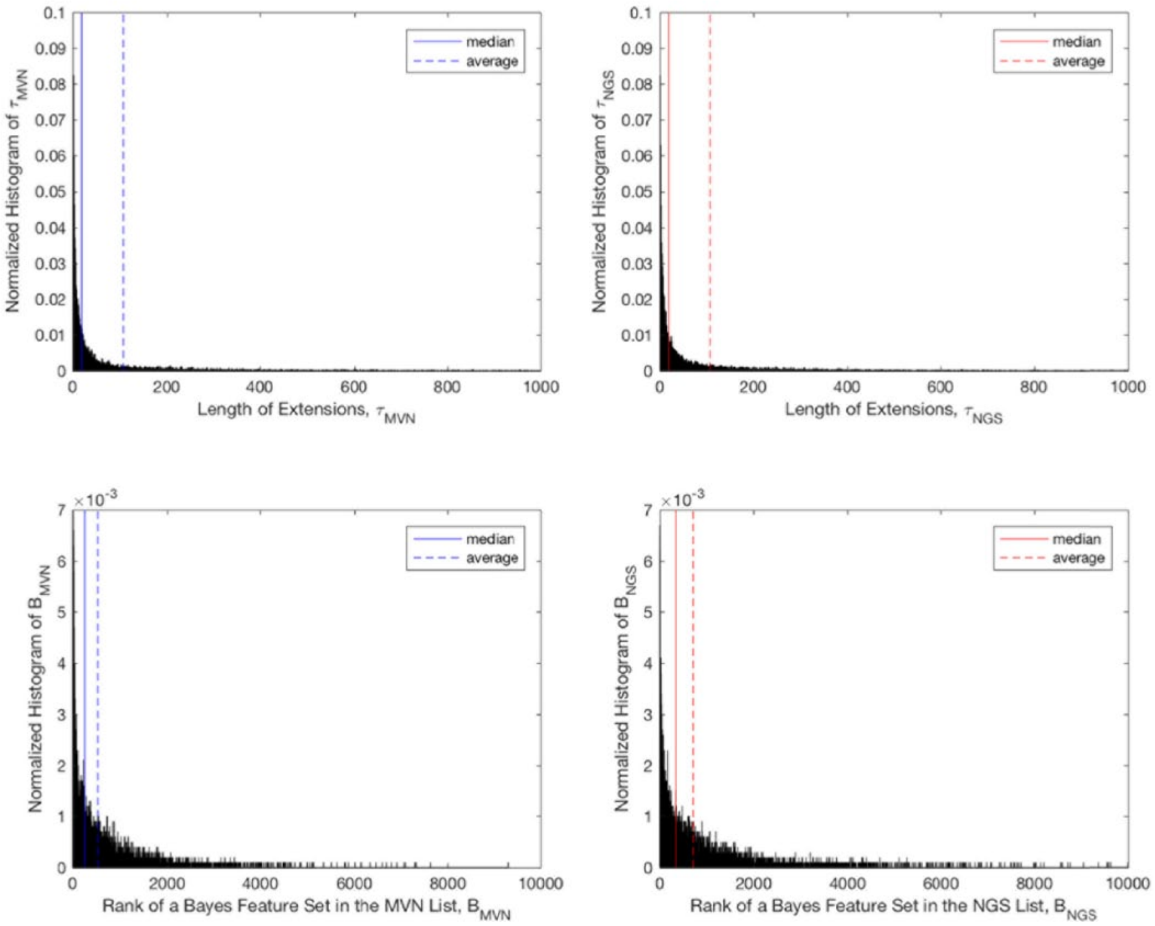


Figure 6. Histogram of length of list extensions and rank of a Bayes feature set. Solid: median, dashed: average, blue: MVN, red: NGS. MVN indicates multivariate normal; NGS, next-generation sequencing.

Effects of increasing D at the same rate υ increases. To examine the effects of increasing D at the same rate υ increases, the proportion of marker features in the data were fixed at 0.05 with the total number of features ranging from 100 to 300. Figure 9A shows that when the proportion υ / D is kept constant, the power curves are relatively unchanged as the number of total features D increases. However, Figure 9B shows that the extension length and the rank of the Bayes feature sets increase under the same conditions, pointing to the increased difficulty of the problem as the number of total features increases.

An example of feature-set ranking for a real data set

We consider a real RNA-Seq data set from a randomized, double-blind crossover intervention of flaxseed lignan extract and placebo.[34] Colonic mucosal biopsies from healthy participants are used to characterize the site-specific global gene expression signatures associated with stromal versus epithelial tissue. The data provide insight into the gene expression landscape of the normal epithelium and stroma prior to the onset of intestinal tumorigenesis. This is noteworthy because the development of cancer is intimately linked to cross talk between cancer cells and the surrounding stromal cells.[34] The data set consists of 29 epithelium and 30 stroma biopsies from the sigmoid colon. Epithelium samples belong to class 0, and stroma samples are labeled as class 1. In total, 960 intestinal genes were selected using prior biological knowledge.[35] Out of 960 genes, 259 stromal genes and 9 epithelial genes were included which were shown to be highly expressed in stroma and epithelium, respectively.[36,37] Repeated random subsampling holdout[38] method was employed on the data set. Twenty samples were randomly selected and used for training, and the remaining data samples were assigned to the test set. This process was repeated 10 000 times with different subsamples to improve the reliability of the holdout estimate.[38] The proportion of samples from each class was kept the same in both the training and test sets. For every feature set of size 2, we designed an LDA classifier and computed its estimated and test errors. Bolstered resubstitution error estimators were used to sort the feature sets. As there is no analytical way to obtain a set of Bayes features for real data,

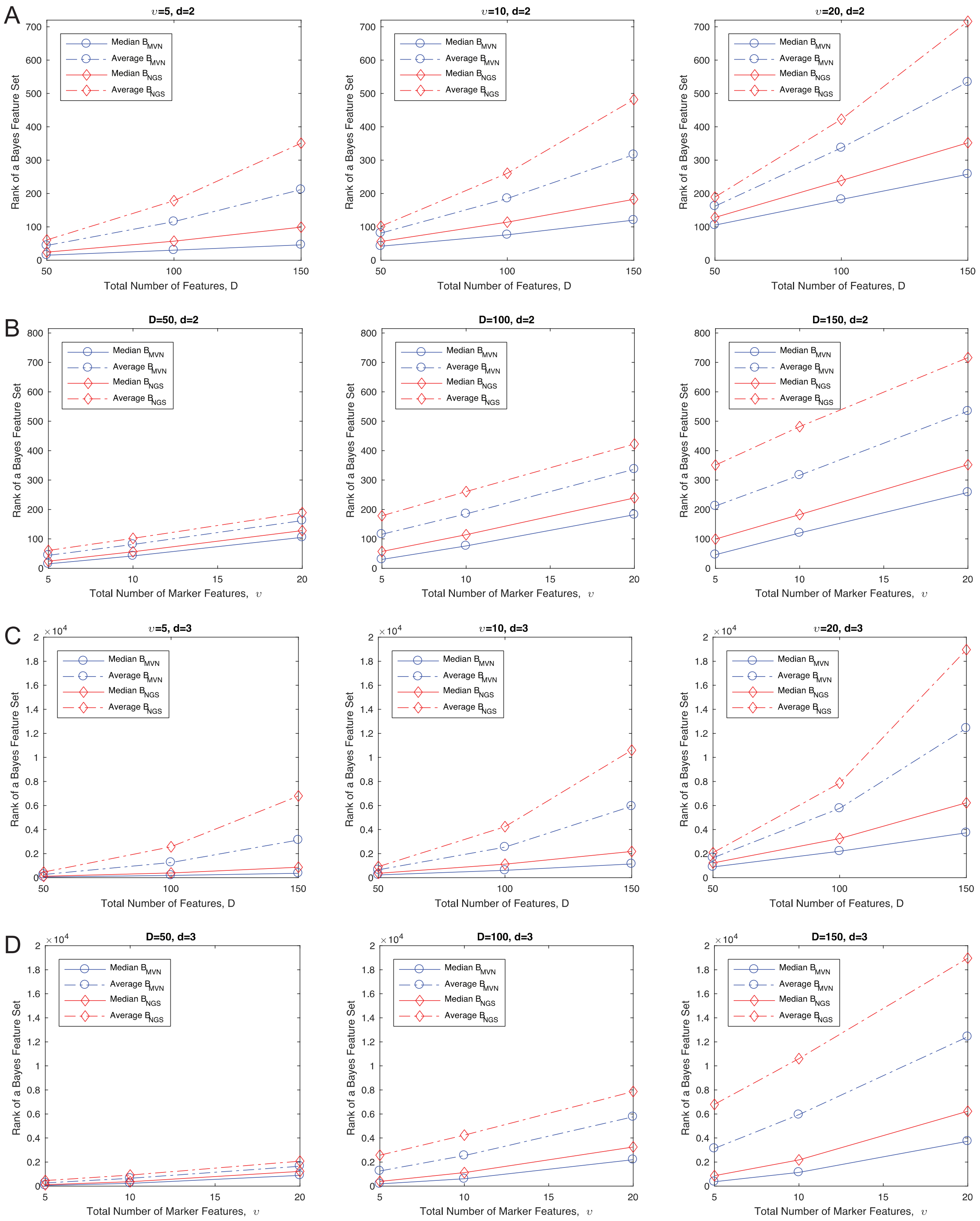


Figure 7. Effects of different *D* and υ on rank of a Bayes feature set for $d = 2$ are shown in (A) and (B). Graphs for $d = 3$ are presented in (C) and (D). Solid: median, dashed: average, blue: MVN, red: NGS. MVN indicates multivariate normal; NGS, next-generation sequencing.

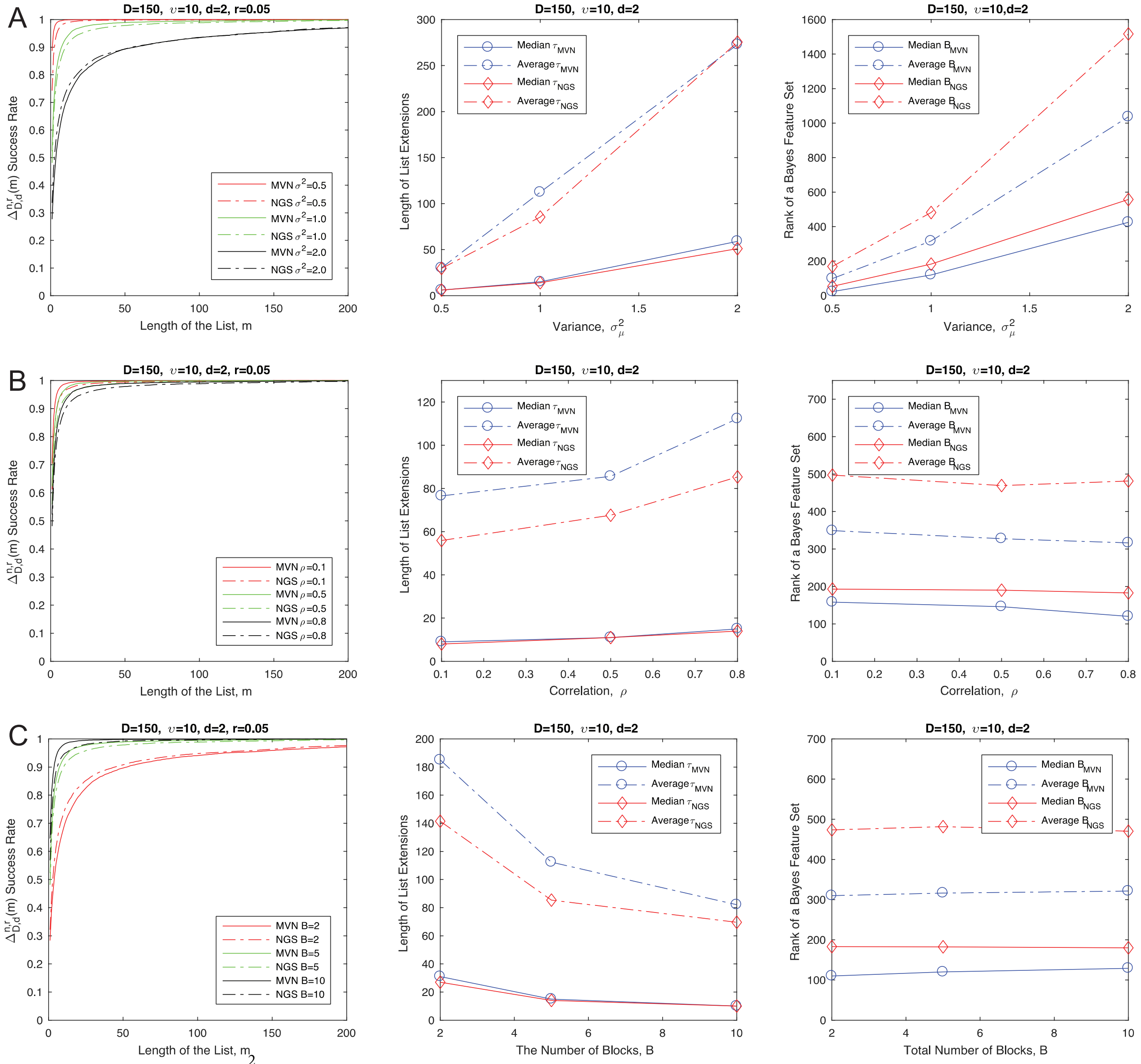


Figure 8. Effects of (A) variance, σ^2_μ; (B) correlation, ρ; and (C) the number of blocks, B on the ranking power, length of list extensions, and rank of a Bayes feature set. MVN indicates multivariate normal; NGS, next-generation sequencing.

we determined ε_0 empirically. Random subsampling was repeated 10 000 times, and the mean of the lowest test errors was taken as $\varepsilon_0(\varepsilon_0 = 0.189192)$.

Figure 10 shows the ranking power for this data set. The parameter $r = 0.03$ indicates that we are interested in ranked lists of feature sets where the feature set with the smallest test error differs from ε_0 less than 15.9% of the ε_0. Typically, when a smaller r is employed, a short list may miss interesting gene sets worthy of consideration. Therefore, the list should be further extended to increase the probability of the existence of candidate genes that provide a good approximation of the Bayes features. It is also important to note that a longer list does not always increase the number of candidate genes. As shown in Zhao et al,[2] some genes repeatedly appear in the list combined with other genes. Ranking power of the real data for large m is provided in Supplementary Figure 1.

Conclusions

This study examines the ranking performance of feature sets derived from a model of RNA-Seq data and compares it with that of an MVN model of gene concentrations. The results demonstrate that the general trends of the parameter effects on the ranking power of underlying gene concentrations are preserved in the RNA-Seq data; however, the power of finding a good feature set becomes weaker and the data become less discriminative when gene concentrations are transformed by the sequencing machine. Moreover, the consistency between the ranked lists of feature sets based on the MVN and the NGS

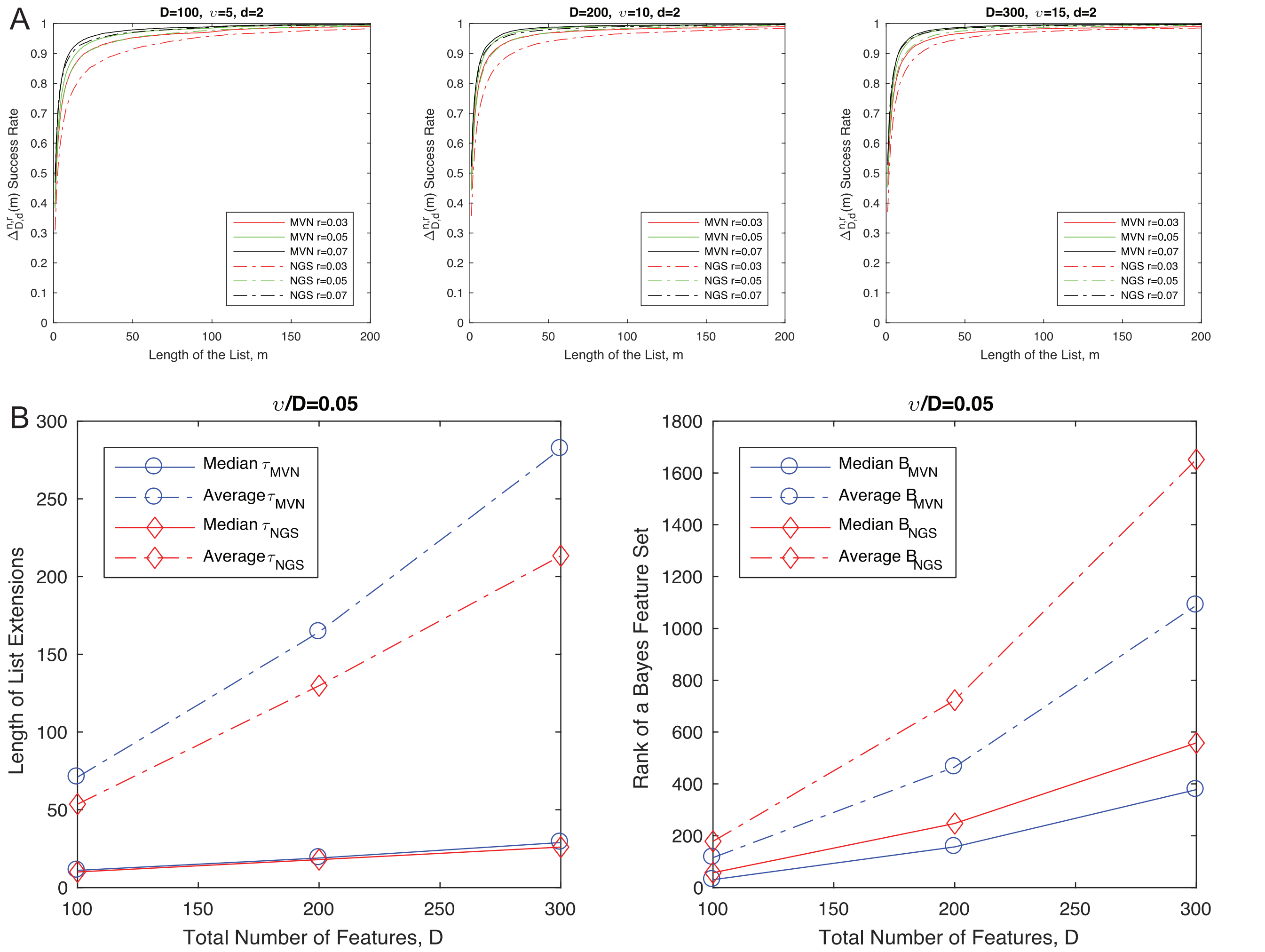


Figure 9. Effects of proportion of υ to D on (A) the ranking power, (B) length of list extensions, and (C) rank of a Bayes feature set. Ratio of υ to D remains the same as 0.05. MVN indicates multivariate normal; NGS, next-generation sequencing.

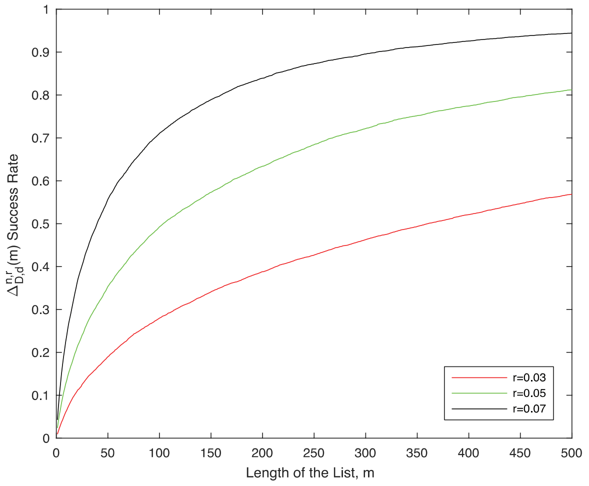


Figure 10. Power curves for a real data set where $n = 59$, $D = 960$, $d = 2$. Red: $r = 0.03$, green: $r = 0.05$, black: $r = 0.07$.

data is poor, which indicates unreliable classification performance in the case of RNA-Seq data.

Author Contributions

EK, II, JH, and ERD conceived and designed the experiments. EK analyzed the data. EK and II wrote the first draft of the manuscript. EK, II, JH, RSC, and ERD contributed to the writing of the manuscript. JWL, MAJH, and RSC provided a real data set used as an example. All authors reviewed and approved the final manuscript.

REFERENCES

1. Sima C, Braga-Neto U, Dougherty ER. Superior feature-set ranking for small samples using bolstered error estimation. *Bioinformatics*. 2005;21: 1046–1054.
2. Zhao C, Bittner ML, Chapkin RS, Dougherty ER. Characterization of the effectiveness of reporting lists of small feature sets relative to the accuracy of the prior biological knowledge. *Cancer Inform*. 2010;9:49–60.
3. Ghaffari N, Yousefi MR, Johnson CD, Ivanov I, Dougherty ER. Modeling the next generation sequencing sample processing pipeline for the purposes of classification. *BMC Bioinformatics*. 2013;14:307.
4. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008;5:621–628.
5. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11:R106.

6. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26:139–140.
7. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18:1509–1517.
8. Knight JM, Ivanov I, Dougherty ER. MCMC implementation of the optimal Bayesian classifier for non-Gaussian models: model-based RNA-Seq classification. *BMC Bioinformatics*. 2014;15:401.
9. Kudo M, Sklansky J. Comparison of algorithms that select features for pattern classifiers. *Pattern Recogn*. 2000;33:25–41.
10. Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance. *IEEE Trans Pattern Anal Mach Intell*. 1997;19:153–158.
11. Cover TM, Campenhout JMV. On the possible orderings in the measurement selection problem. *IEEE Trans Syst Man Cybern*. 1977;7:657–661.
12. Jain AK, Waller WG. On the optimal number of features in the classification of multivariate Gaussian data. *Pattern Recogn*. 1978;10:365–374.
13. Hughes G. On the mean accuracy of statistical pattern recognizers. *IEEE T Inform Theory*. 1968;14:55–63.
14. Sima C, Dougherty ER. The peaking phenomenon in the presence of feature-selection. *Pattern Recogn Lett*. 2008;29:1667–1674.
15. Hua J, Xiong Z, Dougherty ER. Determination of the optimal number of features for quadratic discriminant analysis via the normal approximation to the discriminant distribution. *Pattern Recogn*. 2005;38:403–421.
16. Hua J, Xiong Z, Lowey J, Suh E, Dougherty ER. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics*. 2005;21:1509–1515.
17. Sima C, Dougherty ER. What should be expected from feature selection in small-sample settings. *Bioinformatics*. 2006;22:2430–2436.
18. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*. 2004;20:374–380.
19. Hanczar B, Hua J, Dougherty ER. Decorrelation of the true and estimated classifier errors in high-dimensional settings. *EURASIP J Bioinform Syst Biol*. 2007;2007:38473.
20. Blaise H, Edward RD. On the comparison of classifiers for microarray data. *Curr Bioinform*. 2010;5:29–39.
21. Hanczar B, Dougherty ER. The reliability of estimated confidence intervals for classification error rates when only a single sample is available. *Pattern Recogn*. 2013;46:1067–1077.
22. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*. 2005;21:3301–3307.
23. Xiao Y, Hua J, Dougherty ER. Quantification of the impact of feature selection on the variance of cross-validation error estimation. *EURASIP J Bioinform Syst Biol*. 2007;2007:16354.
24. Braga-Neto U, Dougherty ER. *Classification*. Cairo, Egypt: Hindawi Publishing Corporation; 2005.
25. Shmulevich I, Dougherty ER. *Genomic Signal Processing*. Princeton, NJ: Princeton University Press; 2007.
26. Doughtery ER, Jianping H, Bittner ML. Validation of computational methods in genomics. *Curr Genomics*. 2007;8:1–19.
27. Hua J, Tembe WD, Dougherty ER. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recogn*. 2009;42:409–424.
28. Attoor S, Dougherty ER, Chen Y, Bittner ML, Trent JM. Which is better for cDNA-microarray-based classification: ratios or direct intensities. *Bioinformatics*. 2004;20:2513–2520.
29. Dalton LA, Dougherty ER. Application of the Bayesian MMSE estimator for classification error to gene expression microarray data. *Bioinformatics*. 2011;27:1822–1831.
30. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*. 2010;11:94.
31. Devroye L, Gyèorfi L, Lugosi G. *A Probabilistic Theory of Pattern Recognition*. New York: Springer; 1996.
32. Neto UMB, Dougherty ER. *Error Estimation for Pattern Recognition* (IEEE Press Series on Biomedical Engineering). Piscataway, NJ: IEEE Press; 2015. http://dx.doi.org/10.1002/9781119079507/
33. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*. 2002;347:1999–2009.
34. Knight JM, Kim E, Ivanov I, et al. Comprehensive site-specific whole genome profiling of stromal and epithelial colonic gene signatures in human sigmoid colon and rectal tissue. *Physiol Genomics*. 2016;48:651–659.
35. Zhao C, Ivanov I, Dougherty ER, et al. Noninvasive detection of candidate molecular biomarkers in subjects with a history of insulin resistance and colorectal adenomas. *Cancer Prev Res (Phila)*. 2009;2:590–597.
36. Mo A, Jackson S, Varma K, et al. Distinct transcriptional changes and epithelial-stromal interactions are altered in early-stage colon cancer development. *Mol Cancer Res*. 2016;14:795.
37. Calon A, Lonardo E, Berenguer-Llergo A, et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat Genet*. 2015;47:320–329.
38. Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. *IEEE Trans Pattern Anal Mach Intell*. 2000;22:4–37.