Published online 17 May 2016

Nucleic Acids Research, 2016, Vol. 44, No. 13 ***e119***
doi: 10.1093/nar/gkw444

Differential distribution improves gene selection stability and has competitive classification performance for patient survival

Dario Strbenac[1], Graham J. Mann[2,3], Jean Y.H. Yang[1] and John T. Ormerod[1,4,*]

[1]School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia, [2]Melanoma Institute Australia, University of Sydney, NSW 2060, Australia, [3]Centre for Cancer Research, Westmead Millennium Institute, University of Sydney, Westmead NSW 2145, Australia and [4]ARC Centre of Excellence for Mathematical & Statistical Frontiers, University of Melbourne, Parkville VIC 3010, Australia

Received June 19, 2015; Revised March 18, 2016; Accepted May 09, 2016

ABSTRACT

A consistent difference in average expression level, often referred to as differential expression (DE), has long been used to identify genes useful for classification. However, recent cancer studies have shown that when transcription factors or epigenetic signals become deregulated, a change in expression variability (DV) of target genes is frequently observed. This suggests that assessing the importance of genes by either differential expression or variability alone potentially misses sets of important biomarkers that could lead to improved predictions and treatments. Here, we describe a new approach for assessing the importance of genes based on differential distribution (DD), which combines information from differential expression and differential variability into a unified metric. We show that feature ranking and selection stability based on DD can perform two to three times better than DE or DV alone, and that DD yields equivalent error rates to DE and DV. Finally, assessing genes via differential distribution produces a complementary set of selected genes to DE and DV, potentially opening up new categories of biomarkers.

INTRODUCTION

A central theme in disease diagnosis using genomic data is using the change in average expression as the main measure of differences between different classes. An example of one of the earliest studies built a classifier for different subtypes of leukaemia based on finding a set of genes that are uniformly high in one class and low in the other (1). Since then, a wide range of studies have been made to determine important biomarkers (features), with applications to predicting survival outcomes (2,3), disease subtypes (4,5), drug sensitivity (6,7) and even behavioural characteristics (8). Apart from RNA expression, many studies have utilised other kinds of biological data, such as protein (9) and metabolite (10) data. Reviews of classification methods based on differences in average expression levels can be found elsewhere (11–13).

Recently, variances of expression (differential variability) have been found to differ in numerous gene expression data sets (14,15). One biological interpretation for this is that increased variability of the RNA level of a particular gene, caused by the loss of precise regulation of its expression, may follow disruption of transcription factors or epigenetic signals by pathogenic processes, leading to greater variation of the expression level between samples within the affected class (16). The study by Ho *et al*. (14) found that highly variable genes are highly co-expressed with many fewer genes than are genes with lower variability in their expression. Genes with higher variability are typically associated with the disease state, although high variability *per se* may also be evolutionarily conserved property of some gene systems that serves a potentially beneficial purpose in some gene systems (17).

Motivated by these biological insights, two main statistical approaches have been developed to assess the association of genes with important disease phenotypes, such as natural history (prognosis), using differential variability. In the earliest proposal, mixture models were utilised, which can also measure differential expression (18). Another method, diffVar (19), which is based on testing absolute deviations from class means in a linear modelling framework, is limited to discovering changes in variability. These methods were able to find genes with differential variability between conditions, but neither study developed a corresponding metric for feature selection or investigated its potential as a biomarker in a prognostic setting.

The potential of differential variability to aid in classification was recently demonstrated for the first time in a comprehensive study of DNA methylation in a number of

*To whom correspondence should be addressed. Tel: +61 2 9351 5883; Email: john.ormerod@sydney.edu.au

cervical cancer data sets (20). A differential variability classifier based on adaptive index models (21) outperformed a differential methylation classifier for predicting early-stage cancer, although there was no difference in classification performance at later stages. This suggests that traditional differential expression (DE) classifiers disregard important differences which are present in real data sets.

Differential variability attempts to use the characteristics of dysregulated networks to provide a new approach to assessing the importance of genes. However, it omits useful information from changes in locations between classes. Here, we propose a novel metric based on identifying genes with differential distribution to simultaneously identify genes that are differentially expressed, differentially variable or both. As such, differential distribution (DD) aims to avoid the need for ad-hoc DE and DV classifier aggregation algorithms. Biologically, a change in distribution such as unimodality to multimodality suggests that a gene has an expression range which must be maintained for healthy cellular function. Increases in variability can be similarly interpreted. Furthermore, we extend such feature selection criteria into a classification setting. To date, no research literature has examined this kind of classification for biological problems. In other fields, such as engineering, DD classification by kernel density estimate voting has been shown to perform slightly better than methods like LDA (22) on a simulated data set and on low-dimensional data sets from physics and chemistry, motivating its exploration in high-dimensional biological data sets. This leads us to propose using DD metrics as a type of discrimination measure for identifying candidate genes of interest as well as using those metrics in a novel classification scheme for omics data sets. DV and DD are, for the first time, characterised in terms of model stability, something which is known to be lacking for DE feature selection (23). Additionally, we systematically examine the performance of all three classification schemes based on their prognostic error rate and biological relevance.

MATERIALS AND METHODS

Data sets

Three experimental data sets were used for comparison of selection and classification performance; two measuring RNA expression on microarrays and one utilising RNA-seq. All values from microarrays were transformed to the $\log_2$ scale. Cases in each data set were partitioned into a *good prognosis* class and a *poor prognosis* class. One independent validation melanoma data set was utilised for cross-study validation. An external database (MalaCards) was used for evaluation of feature selection in terms of previously disease-associated genes.

Melanoma. The raw microarray expression and clinical data are available from GEO as GSE54467. The samples were assayed on the Illumina Human WG-6 BeadChip microarray, version 3. Previously defined classes for this data set (3) of poor prognosis as death less than one year from metastasis (n = 22) and good prognosis as survival of more than four years with no signs of recurrence (n = 25) are considered. Raw data were NEQC normalised (24) and probes which had less than ten samples with a detection P-value of < 0.01 were removed from further analysis.

Serous ovarian cancer. Processed microarray data generated by the study GSE13876 were obtained from the Bioconductor package curatedOvarianData (25). Gene expression was measured with Operon Human (version 3) microarrays developed by the Netherlands Cancer Institute. Based on a density plot of survival times for all samples, we defined poor prognosis as death within two years (n = 22) and good prognosis as survival of five or more years (n = 25).

Lung Adenocarcinoma. The processed data were obtained from TCGA Data Portal on 16 May 2014. Poor prognosis cases were defined as those who died less than one year from diagnosis and good prognosis cases as those who lived for over four years, with no signs of recurrence. This resulted in a total of 18 poor prognosis and 18 good prognosis samples. Sequencing was performed on an Illumina HiSeq 2000 instrument. Normalised gene count values were used. Genes that had fewer than 10 counts in fewer than 10 samples were removed from further consideration.

Melanoma validation. The raw microarray expression and clinical data are available from GEO as GSE65904. The samples were assayed on the Illumina HumanHT-12 BeadChip microarray, version 4. Good and poor classes had the same definition used in the previously introduced melanoma data set. This resulted in a total of 22 poor prognosis and 25 good prognosis samples.

The follow-up time densities using all samples from each data set show that the TCGA lung cancer data set has relatively few cases with long follow-up times, while the other data sets have a greater variety of times (Supplementary Figure S1).

MalaCards. Cards for melanoma, ovarian cancer and lung cancer were downloaded from the MalaCards website (http://www.malacards.org) on 25 March 2014. These are gene lists for particular diseases with scores for each gene proportional to that gene's association to the disease in the published literature (26). In the ovarian cancer data set, some features were annotated with multiple RefSeq symbols. For those genes, the maximum MalaCard score of matching symbols was chosen.

Feature selection and classification

All analyses were carried out using the Bioconductor package ClassifyR (27), developed by the authors of this study. ClassifyR is a framework that allows users to carry out classification using several resampling methods (including cross-validation), and calculate several different performance measures, such as classification error rates and variable inclusion frequencies. Parallel processing is implemented and a flexible framework for including user-defined feature selection and classification functions is available. This allows new kinds of classifiers to be systematically tested, once they become available.

Each classification type was performed in a similar way. We used a 5-fold cross-validation scheme where the samples

have been repeatedly resampled with replacement 100 times to create 100 versions of each data set. For each iteration of cross-validation:

(i) The training data were first processed by a feature selection function. In each case, the feature selection function chose the set of features by testing each gene individually, ranking them by a score and calculating resubstitution error rates for the top x ranked features. Values of x considered ranged from 10 to 150 in increments of 10. The value of x which obtained the lowest balanced error rate determined the size of the set of top features selected.
(ii) For our proposed DV method, we performed a transformation of expression values. For each feature, all samples' expression levels were subtracted from the median expression level of the training set, and absolute values taken. This transforms the data into a form that allows a classifier with linear decision boundaries to be applied.
(iii) The samples assigned to the training set in the current iteration were used for model building.
(iv) The samples assigned to the test set in the current iteration had their classes predicted.

Note that the choice of using the resubstitution error for feature selection is a pragmatic one. Ideally, a nested cross-validation could be used. However, we found this approach was not feasible in practice and used resubstitution error instead. A summary of the various feature characteristics and classifiers for them is presented in Figure 1.

The particular feature selection methods and classifier utilised are distinct for each type of change.

DE and classification. (i) For microarray data, genes were ranked on their moderated t-statistics using the implementation in limma (28). Training and prediction for the microarray data sets was performed using diagonal linear discriminant analysis (DLDA). (ii) For RNA-seq data, genes were ranked based on a likelihood ratio test statistic of negative binomial generalised linear models using the implementation in edgeR (29). Poisson linear discriminant analysis (PLDA) was used to determine a decision boundary and make predictions, as it been demonstrated that DLDA finds suboptimal boundaries for count data, whereas PLDA finds the correct boundary (30). A power transformation was applied to eliminate overdispersion, making PLDA applicable to RNA-seq count data.

DV and classification. For microarray data, the normalised values were directly used. For RNA-seq data, the mean-variance trend was removed by using the regularised logarithm transformation (31) of DESeq2, to avoid detecting DV features simply caused by DE. Features were then ranked based on either their Bartlett statistic or their Levene statistic and selection was applied. The Bartlett test tends to choose features with a small number of outliers, whereas the Levene statistic is robust to outliers. Before training and prediction, feature values were calculated as the absolute value of the difference of each measurement with the median of all samples in the training set. Thus, if the values originally came from a normal distribution the transformed values would follow a half-normal distribution. Fisher's linear discriminant analysis (FLDA) was used for classification.

DD and classification. We considered four approaches for assessing the differences between two different distributions (class 1 and class 2). These are the differences of medians and deviations, the Kolmogorov–Smirnov distance, the log-likelihood ratio and simply combining the results of individual DE and DV selections. Motivated by the success of finding DE genes by considering the absolute differences in medians for the melanoma data set (3), the differences of medians and deviations (DMD) is defined as

$$\mathrm{DMD} = |median_1 - median_2| + \left| Q_{n_1} - Q_{n_2} \right|$$

where $median_1$ and $median_2$ represent the median expression values of class 1 and 2, respectively. The values Q_{n_1} and Q_{n_2} represent the robust scale estimator (32) for class 1 and 2, respectively. The Kolmogorov–Smirnov (KS) distance is simply defined as the greatest distance between the empirical cumulative distribution functions of the two classes. Thirdly, we used a log-likelihood ratio statistic with robust estimates of the location and scale:

$$LLR = -2\Big(\sum_{i=1}^{s_1+s_2} \log_e f(x_i|\mu = median, \sigma^2 = Q_n^2) - \sum_{i=1}^{s_1} \log_e f(x_i|\mu = median_1, \sigma^2 = Q_{n_1}^2) - \log_e f(x_i|\mu = median_2, \sigma^2 = Q_{n_2}^2)\Big)$$

where subscript i denotes membership of class i, s_i denotes the number of samples in class i and f is the probability density function of the normal distribution. Note that LLR is the log of the likelihood ratio statistic for testing whether the two classes come from the same normal distribution or two different normal distributions where robust estimators of the mean and standard deviation are used in place of the maximum likelihood estimators. Terms without subscripts use values from samples in both classes. Finally, we consider ensemble feature selection by combining the selections which use the limma moderated t-test and Bartlett test by a simple union of sets. This is a naïve way to jointly capture features that are changing means and also those that are changing variances.

For the chosen features, a kernel density estimate was built for each of the two classes using a Gaussian smoothing kernel and bandwidth calculated by Silverman's rule, which are the default settings of the density function in R. For the RNA-seq data set, counts were transformed by the regularised log method, to prevent feature selection being biased towards differentially expressed genes, because of overdispersion of count data. To predict a sample from the test set, a naïve Bayes classifier was used for each feature. Two variants of the classifier were considered. Firstly, each feature votes for the class that has the maximum *a posteriori* estimate. This is referred to as *unweighted voting*. Also, the differences between class densities, the distances from an observation to the nearest non-zero crossover point of the two densities, and the sums of those two weights were calculated and summed over all features, with the sign of the sum determining the class prediction. This is termed *weighted voting*. Intuitively, the crossover distance weighting captures how far away a measurement is from the nearest substantial observation of the class with lower density at the measurement point.

The novel DD classification approach is summarised by a flowchart (Figure 2).

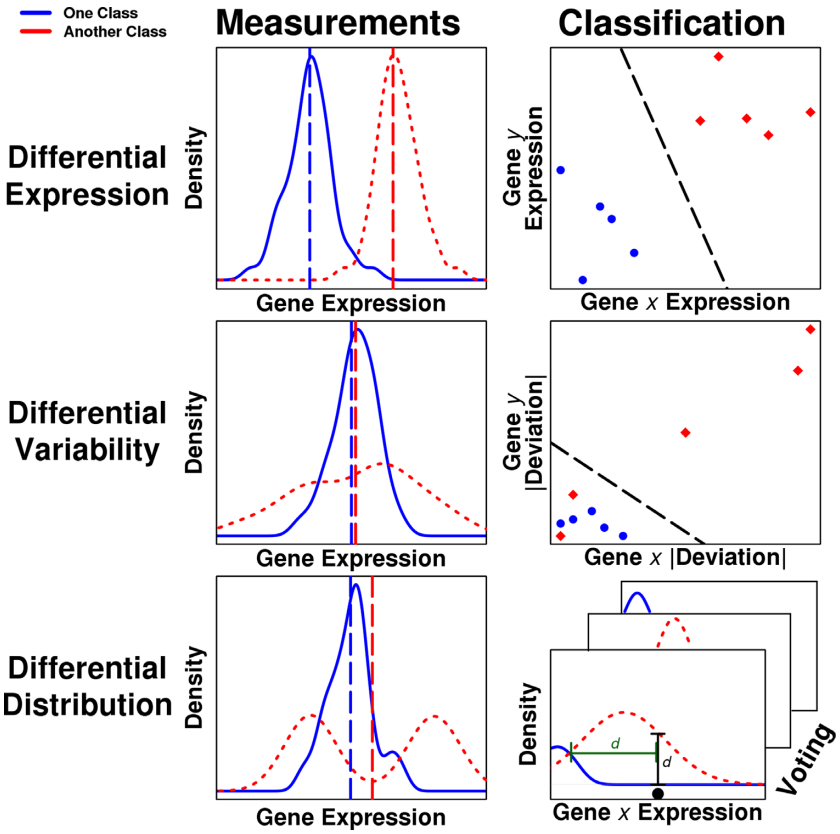

Figure 1. Summary of feature types and classifiers. For each of differential expression, differential variability and differential distribution, a representative gene profile is given, and an illustration of the classification process given. In the left column, the dashed vertical lines represent the means of the class distributions. In the right column, the variables x and y denote two different genes in a data set. The bottom right panel illustrates that each gene from the selected gene set votes independently in differential distribution classification.

Data simulation

(i) Background expression: To estimate a realistic set of expression values of unchanged features, we resampled from the Melanoma data set after we excluded all features that had any potential changes. To define 'changed' features, we applied six methods to rank features in terms of DE, DV or DD, from largest to smallest. These were based on moderated t-statistic, a Bartlett statistic, a Levene statistic, a DMD distance, a KS test statistic and a likelihood ratio statistic. Features that appeared in the top 20% of any of the six lists were excluded from the unchanged feature set. This gave 9453 unchanged features, and 300 of these features were randomly chosen to be changed to create seven simulated data sets.

(ii) Features of interest: Seven simulated data sets were generated, with varying proportions of features of interest. These included DE, DV, differentially skewed (DS) and differentially modal (DM) features. For each data set, the changing features and their magnitudes were chosen by randomly choosing a class and a direction of change, with both the classes and directions being equally likely. To add noise for DE features, 10% to 30% of randomly chosen samples in the unchanged class were also changed by the sampled amount of change. The random sampling of change magnitude was repeated for each feature. The samples of the unchanged class that were changed were constant for all features. Additionally,

(a) The DE features amount of change was sampled from a log-normal distribution with mean 1 and standard deviation 1. The change was applied by adding or subtracting the change value from the measurements.

(b) Two varieties of DV features were simulated; *consistent* and *outlier*. Each variety was equally as likely to be applied to a feature.
Consistent: For a particular feature, the standard deviation of the chosen samples were increased or decreased by a number sampled from a log-normal distribution with mean 1 and standard deviation 1. This enlarges or shrinks the spread of data symmetrically. Lastly, the values were shifted to keep the original mean.
Outlier: Between 10% and 30% of samples in a randomly chosen class had their expression values increased or decreased by an amount sampled from a $Uniform(2, 5)$ distribution. This simulates another frequently observed pattern of DV in biological data sets.

(c) For DS features, the median expression value of the change class was calculated and either the values lower or higher than this value were chosen to be shifted by a value calculated from multiplying their distance from the median value by a skewing factor. The skewing factor

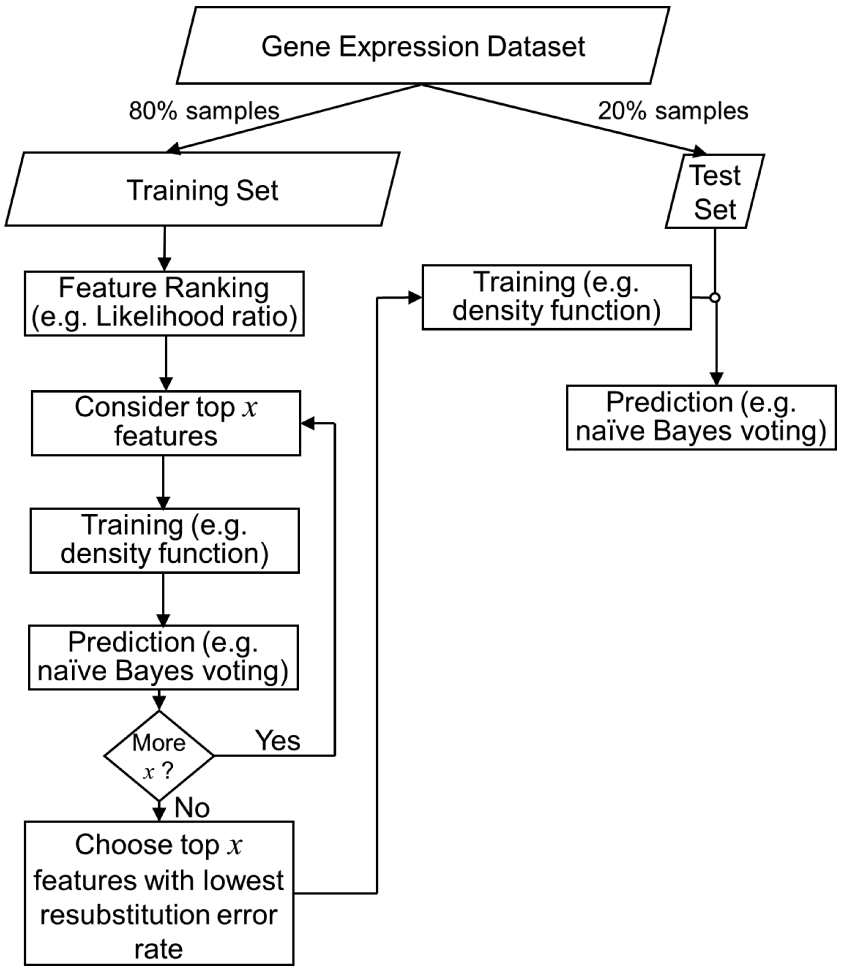

Figure 2. Flowchart of differential distribution classification. A single fold of 5-fold cross-validation is shown. Features are ranked and chosen based on the resubstitution error rate. For each sample in the test set, its class is predicted based on votes made by each of the selected features, based on the class densities fitted during the training process.

used came from independent samples from a log-normal distribution with mean 1 and standard deviation 1.

(d) DM features were created by calculating the mean of the change class and sampling two mean changes from a log-normal distribution with mean 1 and standard deviation 1. One change is the distance to the simulated lower mode's mean and the other is the distance to the simulated higher mode's mean. Additionally, two standard deviation values were sampled from a log-normal distribution with mean 1 and standard deviation $\frac{1}{2}$. Once the changed means and standard deviations were found, random samples (of the same number originally above and below the median value) were drawn from the normal distribution with the mean and standard deviation of each mode. This creates a bimodal distribution.

Performance evaluation

Feature ranking stability was assessed by considering the feature rankings of every cross-validation iteration of a particular classification. The top t ranked features were considered for each iteration, and every possible pair-wise intersection was done. The average size of the intersections was converted to a percentage by dividing by t. A range of t values between 10 and 100 were assessed. To assess feature selection stability, overlap percentages between all pairs of selected feature sets were calculated as

$$overlap_{i,j} = \frac{\left|features_i \cap features_j\right|}{\left|features_i \cup features_j\right|} \times 100.$$

The indices i and j are for different iterations of the cross-validation loop of a particular classification. Feature ranking commonality and feature selection commonality are similarly defined, except that the feature set comparison was between two kinds of classification, rather than within a classification.

Balanced error rate is one performance measure of prediction we considered, and can be thought of as the average error rate for the two classes, denoted positive and negative. This is defined as

$$BER = \frac{\left(\frac{FP}{P} + \frac{FN}{N}\right)}{2}$$

where FP denotes the number of false positives and P denotes the number of samples in the positive class. Similarly,

FN and N represent the number of false negative and N the number of samples in the negative class.Source code to reproduce all of the following results is provided as three supplementary PDF files.

RESULTS

Differential variability classification has inefficient feature selection but good error rate under simulation

Both DV feature selection methods selected an undesirably large proportion of features from the background feature set; typically about 50% of selections (Figure 3A). As expected, features simulated as DE were rarely selected. DV features comprised almost all of the features that were selected. DM features were chosen in similar proportions to their presence in each simulated data set.

The Bartlett statistic had a consistently better median BER than feature selection by the Levene statistic (Figure 3B). The BERs were higher for data sets in which the simulated DV proportion is smaller. The interquartile range of BERs for these data sets was also moderately larger. Therefore, feature selection based on the Bartlett statistic is used for biological data set comparisons.

Differential distribution classification performs well under simulation

Varieties of DD were examined that incorporate different choices of how distance between classes is measured to compare weighted and unweighted voting schemes. DD classification was found to perform well across a range of possible simulation settings.

Firstly, we investigated whether the four different selection methods selected differing proportions of simulated changed features. The features chosen by DMD mirrored the pattern of simulated changes most closely (Figure 4A). The KS statistic selected no DV features for most simulations. Regardless of how common they were, it favoured the selection of features that were simulated to be DE. The likelihood ratio selection chose features in proportions similar to which they were simulated in, but always chose more unchanged features than the DMD method did. About half of the features chosen were those that had been simulated to be unchanged. In terms of BER (Figure 4B), the DMD selection statistic with crossover distance weighted classification had the lowest error rate for most of the simulated data sets. Likelihood ratio and naïve ensemble selection had quite variable balanced error rates between data sets, particularly for the height and sum of differences weightings. Because it selected features in the desired proportions and has good error rates across all seven data sets, the DMD statistic combined with crossover distance weighted voting is used for biological data set comparisons.

Differential distribution classification has similar accuracy to existing methods in cancer data sets

DE, DV and DD classification all had a similar error profile. Figure 5A shows the balanced error rates for all three data sets classified by all three types of classification. For melanoma, the median balanced error rate for the resampling and folding validation was 23%, 26% and 24% for DE, DV and DD, respectively. A similar pattern was observable for the other two cancers. Ovarian cancer was always the most difficult to classify whereas lung cancer had the lowest error rate of each classification type. The spread of the BER values was large. For example, each of the BER distributions had values as low as 5% or as high as 50%, depending on the iteration of the cross-validation. By chance, it would be expected to obtain BERs of 50%, so all three classifications almost always performed much better than classification at random. DD classification had the largest increase of its BER in an independent data set, however (Supplementary Figure S2).

The error rates of individual patients were also similar between methods. That is, some samples were systematically classified poorly for all three gene assessment types, while others were typically classified correctly. The percentage of misclassifications of every sample was calculated for every method and plotted as an error map (Figure 5B). Each heatmap compares the error rate for each sample when classified with each of the three classification types. Darker shading indicates correct predictions are made in more cross-validations. For example, melanoma patient TB52 was classified poorly by all three methods, whereas TB36 was always classified well. There are more misclassified samples in the smaller class for ovarian cancer. The lung cancer data set has balanced class sizes and no tendency in misclassification was observed.

Differential distribution selection identifies different sets of genes to existing assessment types, many which are known disease-related genes

Considering all the samples for each data set, the top 50 genes ranked by differential distribution had only minor overlaps with both differentially expressed and differentially variable genes (Figure 6A). Ovarian cancer had the smallest overlaps between methods, with at most two genes in common between each possible pair of selections. For the ovarian cancer and lung cancer data sets, one gene was common to all three selection types. Regardless of the data set considered, all pairwise overlaps between selection types were 12% or less of the size of the set union.

The selected features also had little overlap (Figure 6B). For melanoma and lung cancer, only one feature was chosen by all three assessment types. For all three cancers, the DE and DV selections had just one gene in common. There were no common genes between DE and DD selection for ovarian cancer. The size of the gene list chosen also varied widely from data set to data set. DE selection gave the most compact sets, ranging from 10 to 30 features. DD selection for lung cancer gave the biggest set of features, choosing all 150 that were considered. It also chose 110 features for the melanoma data set, while the other two methods chose 10 or 20 features.

Although these sets of genes had little overlap, the DD-selected genes were enriched for those of biological significance (Figure 6C). Considering the top-ranked genes, DD selection provided the most disease-associated genes for the melanoma and ovarian cancer data sets over different sub-

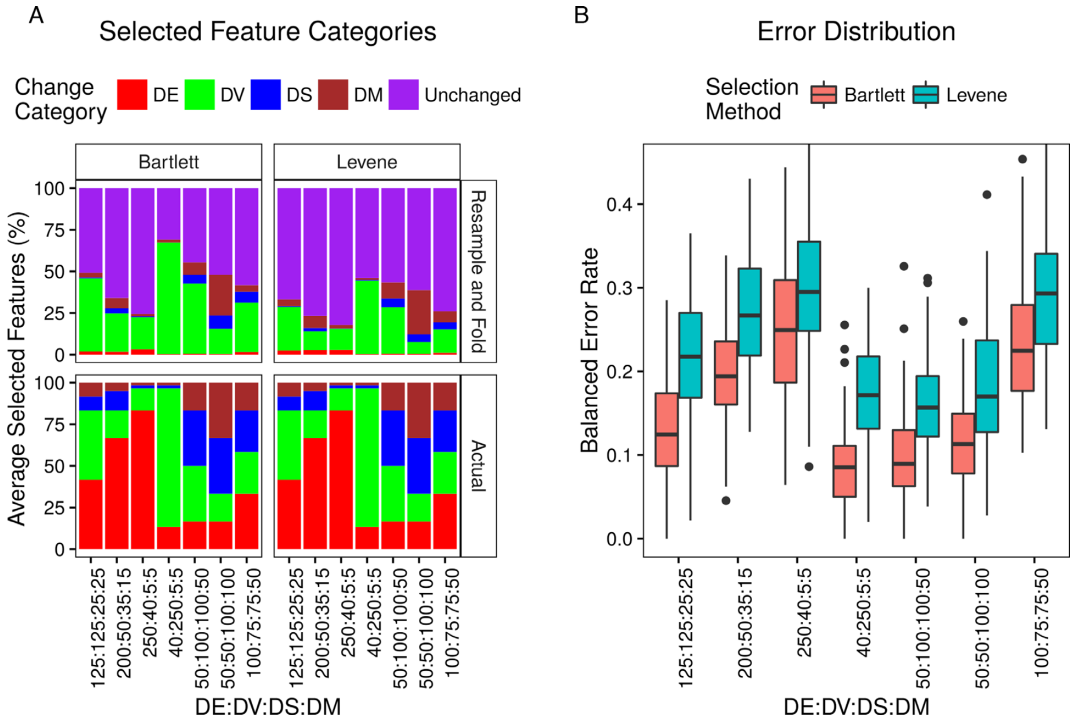

Figure 3. Feature selection proportions and balanced error rates of DV classification for seven simulated data sets. (**A**) Proportions of selected genes. The average percentage of selected genes that are in the specified simulated change categories over all cross-validations is shown. The bottom row shows the proportions of simulated changes. (**B**) Balanced error rates of class predictions. The distributions of error rates across all cross-validation iterations are shown as boxplots.

sets of top-ranked genes. For lung cancer, it selected almost as many high-scoring genes as DE. DV selection provided the lowest cumulative score of previously disease-associated genes for every data set. The low recall of DV was clearest for the lung cancer data set, where the cumulative sum plateaus at top rankings no higher than 50. The top-ranking DV features of ovarian cancer were almost entirely unassociated with the disease.

Differential distribution is the most stable method of ranking and selecting features

In cross-validation using the cancer data sets, differential distribution selection yielded more features in common over all pairs of cross-validations. Considering the highest ranking genes, the DD-selected features were typically two to three to times more stably ranked than DE features, depending on the data set (Figure 7A). For lung cancer, the stability was as much as six times higher. DD feature ranking was also more stable than DV feature ranking for the ovarian and lung cancer data sets. The lung cancer data set benefitted the most from DD selection, with stabilities ranging between 30% and 35%. The lung cancer data set had the highest selection stability overall, whereas the ovarian cancer data set achieved the lowest.

Feature selection based on differential distribution was also the most stable (Figure 7B). The DD median selection score was about twice as large as the second highest median score, except for melanoma, where it was nearly identical to DV's score. For the ovarian and lung cancers, the second highest median score was from DV selection. The median score of DE selection was the lowest in every data set. The interquartile range for DD scores was the largest, except for melanoma, where DV had a slightly larger spread. The expression distribution of the most frequently selected feature in each data set and of each feature type is illustrated (Supplementary Figure S3). The most stable genes for ovarian cancer have much less noticeable differences between survival classes than the most stable genes for the other two cancers.

DISCUSSION

Stable and accurate prediction of sample classes from gene expression signatures or other omics data sets remains a challenging problem in cancer prognosis and omics research. Previous research (19), as well as our own, has found that DE and DV methods select rather different sets of genes. Feature selection using all samples in a data set had a minimal overlap, ranging from 1% to 3% for the cancers considered. These large differences in selected sets motivated the development of measures that combine the characteristics of both DE and DV methods. Here, we have developed a kernel density-based DD measure with a corresponding prognostic algorithm and showed that it performs well in terms of classification and stability on both simulated and three sets of real high-dimensional transcriptome data.

Feature selection stability is an important problem, because if a feature is selected infrequently in a resampling procedure, that feature may have been selected by chance and not related to the outcome of interest, which limits its

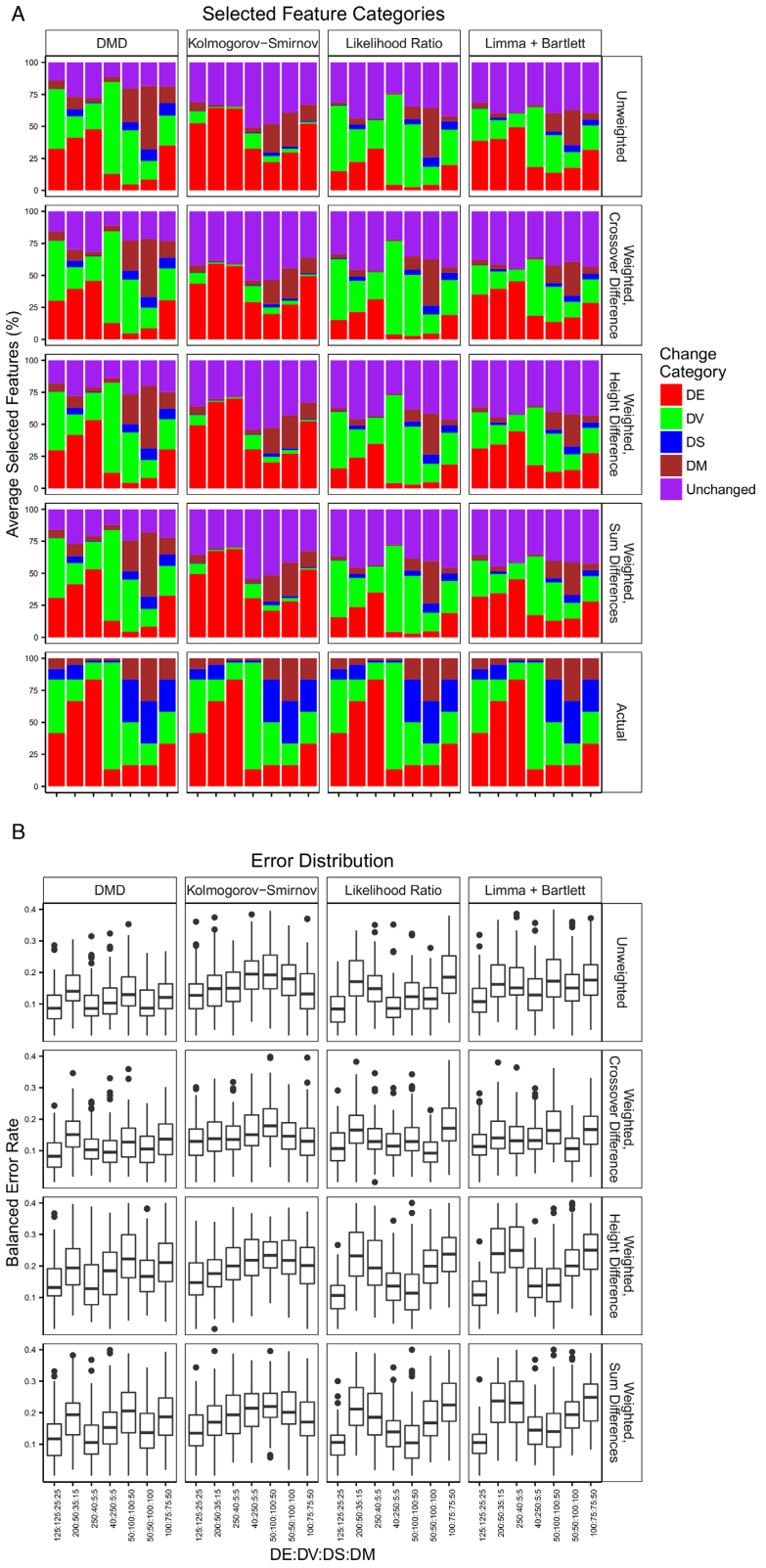

Figure 4. Feature selection proportions and balanced error rates of DD classification for seven simulated data sets. (**A**) Proportions of selected genes. The first four rows of panels contain results for unweighted and three weighted gene voting modes of the naïve Bayes classifier. The fifth row shows the proportion of simulated changes. The columns contain results for DMD, Kolmogorov–Smirnov, Likelihood ratio and ensemble of moderated t-statistic and Bartlett statistic feature selection. The average percentage of selected genes that are in the specified simulated change categories over all cross-validations is shown. (**B**) Balanced error rates of class predictions. The distributions of balanced error rates across all cross-validation iterations are shown as boxplots.

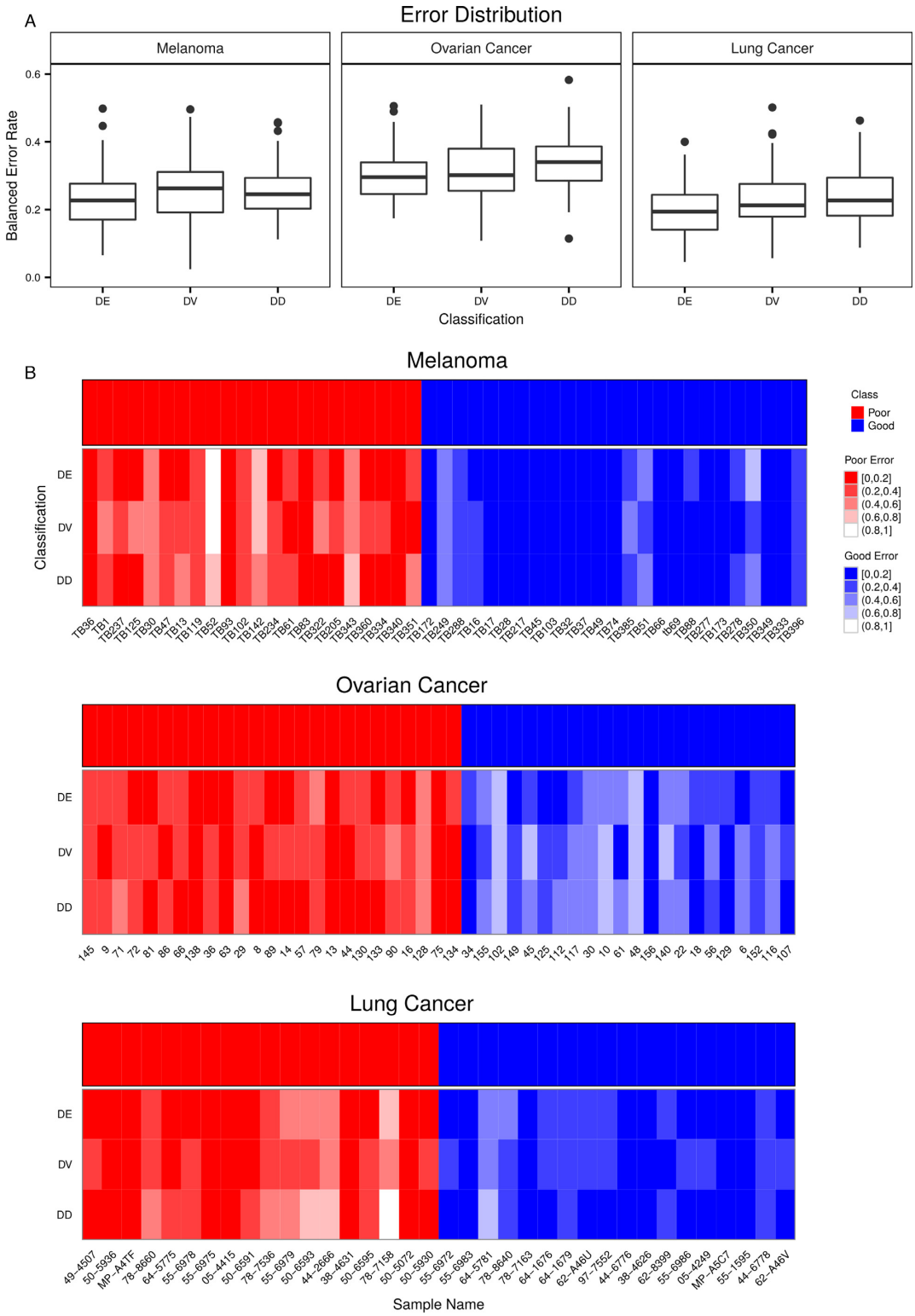

Figure 5. Cross-validated balanced error rates and sample-wise error rates. (**A**) Distribution of balanced error rates over all iterations of cross-validation. (**B**) Sample-wise error rates. Each patient is one column of a heatmap. Each classification type is one row of a heatmap. Details of the selection and classifier algorithms are provided in the Materials and Methods section. The error rates are binned into five equally sized bins. Colour scales are shaded by class colour, with a darker colour indicating less frequent misclassification than a lighter colour.

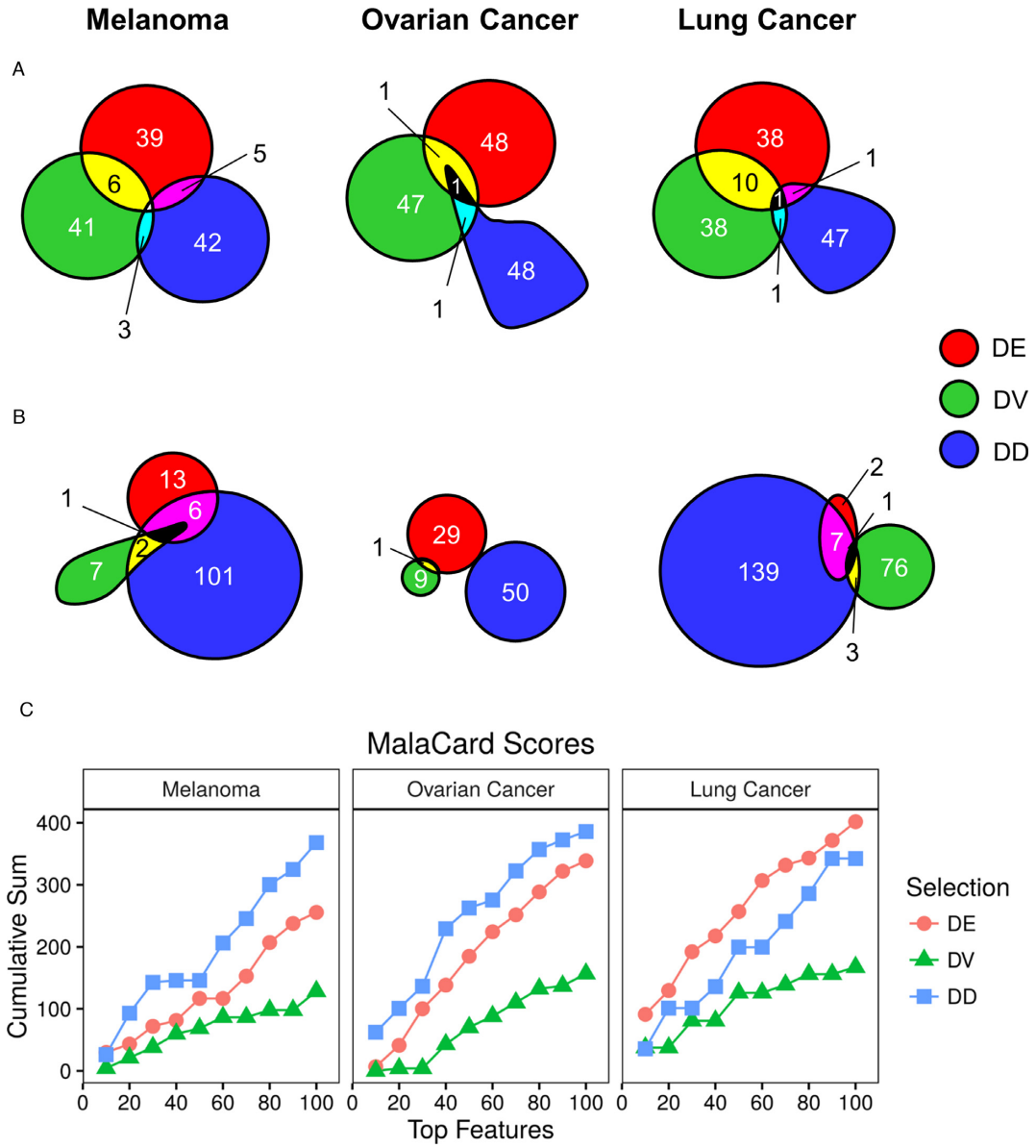

Figure 6. Feature ranking and selection overlaps. (**A**) Overlaps between three feature selection types for three cancer data sets. The 50 highest-ranked genes of each method are used. All samples are used. (**B**) Overlaps between three feature selection approaches for three cancer data sets. The genes selected by best resubstitution error rate of each method are used. All samples are used. (**C**) Cumulative MalaCards scores for the most frequently selected features in cross-validation.

translational potential. The sources of selection instability have been previously characterised and ensemble feature selection proposed as a solution (33). This, however, is computationally costly, as it requires the training of many similar models. Also, the combining of features from different models is subjective, and depends on a user-specified parameter; the minimum number of models the feature was selected in, in order to be included in the final model. In comparison, differential distribution selection by the DMD method has been shown to always be more stable than the popular moderated t-statistic and equally or more stable than the Bartlett statistic, without requiring the generation of multiple models and subjectively aggregating them. As shown by Figure 6C, for each data set, the DMD selection type chose more genes in common with a meta-analysis than DE or DV in two out of three data sets, suggesting that DD has more power than either of the alternatives.

Assessment of genes via DV remains a potentially desirable type of classification when one seeks to classify samples using mostly experimentally unexplored genes. The top ranked DV genesets for each data set had the least overlap with currently well-known disease-associated genes, as defined by MalaCards (Figure 6C). Although this observation may seem concerning at first, it is actually expected and can explained by publication bias. Almost all prior research work on biomarkers has focussed on obtaining markers that have a systematic change in expression between conditions. Only one study has attempted classification with differential

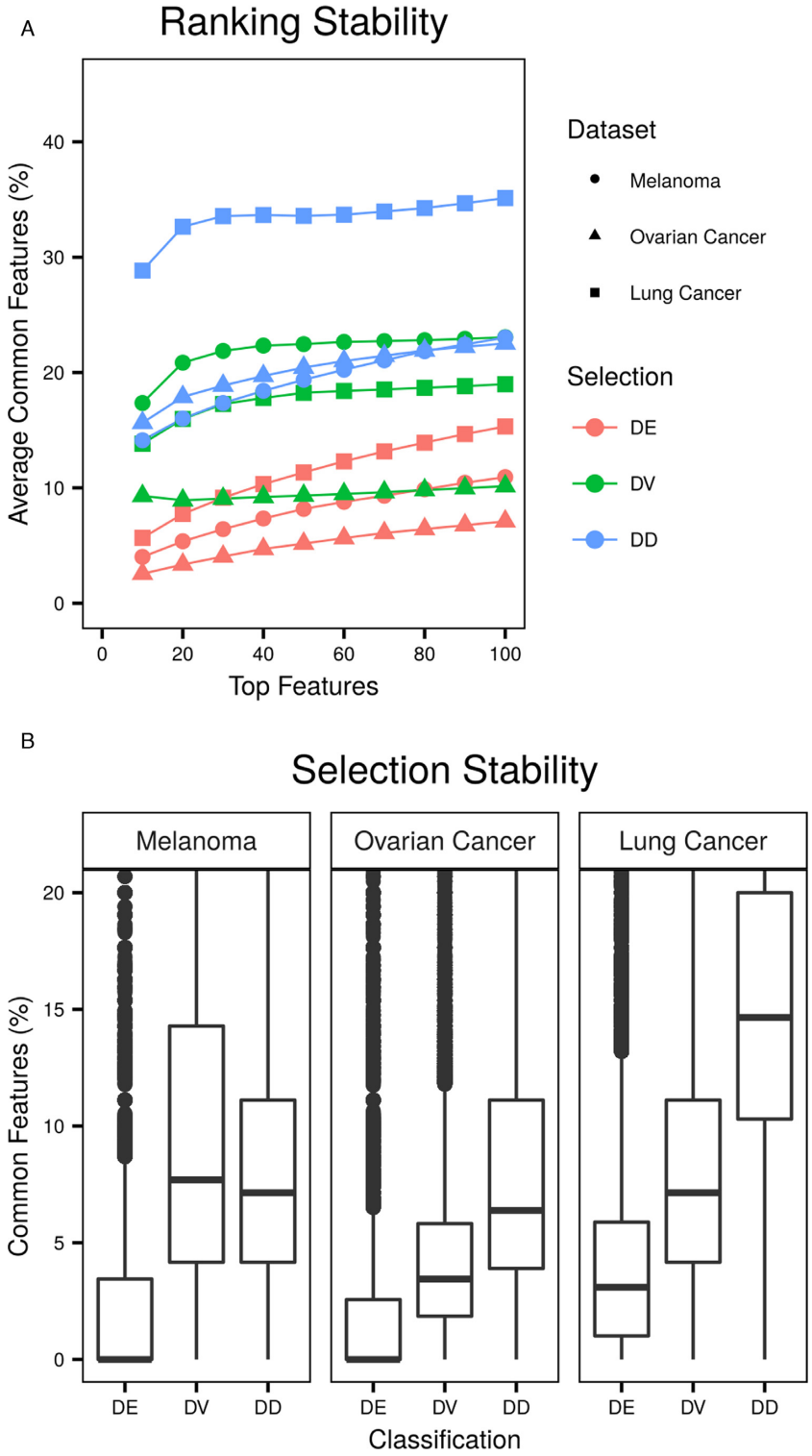

Figure 7. Cross-validation feature ranking and selection stability. For each pair of comparisons, the number of genes in common is divided by the number of genes in the union and converted to a percentage. (**A**)The average pairwise overlap of the top ranked genes is calculated for all iterations of cross-validation. Shapes represent data sets and colours represent different types of classification. (**B**) The distribution of the pairwise overlaps of the selected genes is calculated for all iterations of cross-validation. From left to right, the number of data points which are greater than 20% and are not shown as points is: 475, 16 879, 5301, 305, 593, 5884, 963, 2489 and 29 384.

variability (20) and the lack of variability-associated disease genes in public databases is likely to end, once more studies begin to consider DV or DD classification. Also, in an independent melanoma data set, DV classification had the best BER, while the median BERs worsened by three times as much for DV and DD. A limitation of this comparison is that there was only one relevant independent data set available for evaluation. It would be important to determine if this error rate robustness holds for more independent data sets.

Figure 5 shows that ovarian cancer had a higher BER than the other two data sets for all three classification types, as well as more patients with high patient-specific error rates. The difficulty of ovarian cancer survival prediction has been demonstrated recently (35). The lowest FDR value obtained from fitting a Cox regression model to each gene was 0.85. Selecting the four best genes simply based on odds ratio magnitudes and testing them on an independent data set found that none of them could be validated.

Figure 5B demonstrates that a minority of patients were difficult to classify using DE, DV or DD. Although most patients in each class were classified correctly at least 80% of the time, a small number of patients were classified incorrectly in the majority of cross-validations. The frequent incorrect classification happened regardless of the type of classification done. We suggest that this could be as a result of differences in medical treatment or other unspecified confounding factors. For example, two patients could each have a gene signature that is associated with poor prognosis, but one patient may have received better surgical treatment than the other and, therefore, survive a long time. This issue has recently been explored in invasive breast carcinoma (36), where prognosis prediction was shown to be confounded by ER status, causing some samples to be systematically misclassified. Grouping patients by key features of their clinical data before creating separate omics data classifiers for those groups is a promising new research direction. It provides extra motivation for researchers not only to increase their sample sizes, but to obtain thorough clinical data, in order to make these analyses possible.

In summary, compared to DE and DV, assessing DD identifies association of different genes to the phenotype, selects these genes in a more stable manner, and provides competitive balanced error rates. DE classification only detects changes in means, and misses signatures of transcriptional deregulation. Our results show that the DD approach selects a different set of features with greater biological relevance than does DE, while maintaining good prognostic accuracy. DD classification is, therefore, a superior approach for assessing gene expression, providing good classification accuracy and a complementary set of biological features to DE or DV selection for biologists to pursue experimentally.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors thank Ellis Patrick for his thoughtful suggestions.

FUNDING

Australian Government Department of Education and Training Australian Postgraduate Award (to D.S.); Australian Research Council DECRA Fellowship [DE130101670 to J.T.O.]; National Health and Medical Research Council of Australia Program [633004 to G.M.]; Australian Research Council Discovery [DP130100488 to J.Y.H.Y. and G.M.]; Cancer Institute New South Wales Translational Program [10TPG/1/02 to G.M.]. Funding for open access charge: Australian Research Council Discovery [DP130100488].
Conflict of interest statement. None declared.

REFERENCES

1. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
2. Eschrich,S., Yang,I., Bloom,G., Kwong,K.Y., Boulware,D., Cantor,A., Coppola,D., Kruhøffer,M., Aaltonen,L., Orntoft,T.F. *et al.* (2005) Molecular staging for survival prediction of colorectal cancer patients. *J. Clin. Oncol.*, **23**, 3526–3535.
3. Jayawardana,K., Schramm,S.-J., Haydu,L., Thompson,J.F., Scolyer,R.A., Mann,G.J., Müller,S. and Yang,J.Y.H. (2015) Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation, mRNA, microRNA, and protein information. *Int. J. Cancer*, **136**, 863–874.
4. Mills,K.I., Kohlmann,A., Williams,P.M., Wieczorek,L., Liu,W., Li,R., Wei,W., Bowen,D.T., Loeffler,H., Hernandez,J.M. *et al.* (2009) Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome. *Blood*, **114**, 1063–1072.
5. Marisa,L., de Reyniès,A., Duval,A., Selves,J., Gaub,M.P., Vescovo,L., Etienne-Grimaldi,M.-C., Schiappa,R., Guenot,D., Ayadi,M. *et al.* (2013) Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med.*, **10**, e1001453.
6. Gunther,E.C., Stone,D.J., Gerwien,R.W., Bento,P. and Heyes,M.P. (2003) Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 9608–9613.
7. Li,L., Fridley,B.L., Kalari,K., Niu,N., Jenkins,G., Batzler,A., Abo,R.P., Schaid,D. and Wang,L. (2014) Discovery of genetic biomarkers contributing to variation in drug response of cytidine analogues using human lymphoblastoid cell lines. *BMC Genomics*, **15**, 93.
8. Takahashi,M., Hayashi,H., Watanabe,Y., Sawamura,K., Fukui,N., Watanabe,J., Kitajima,T., Yamanouchi,Y., Iwata,N., Mizukami,K. *et al.* (2010) Diagnostic classification of schizophrenia by neural network analysis of blood-based gene expression signatures. *Schizophr. Res.*, **119**, 210–218.
9. Li,X., Hayward,C., Fong,P.-Y., Dominguez,M., Hunsucker,S.W., Lee,L.W., McLean,M., Law,S., Butler,H., Schirm,M. *et al.* (2013) A Blood-Based Proteomic Classifier for the Molecular Characterization of Pulmonary Nodules. *Sci. Transl. Med.*, **5**, 207ra142.
10. Kim,Y., Koo,I., Jung,B.H., Chung,B.C. and Lee,D. (2010) Multivariate classification of urine metabolome profiles for breast cancer diagnosis. *BMC Bioinformatics*, **11**, S4.
11. Lin,W.-J. and Chen,J.J. (2013) Class-imbalanced classifiers for high-dimensional data. *Brief. Bioinform.*, **14**, 13–26.
12. Ferté,C., Trister,A.D., Huang,E., Bot,B.M., Guinney,J., Commo,F., Sieberts,S., André,F., Besse,B., Soria,J.-C. *et al.* (2013) Impact of Bioinformatic procedures in the development and translation of high-throughput molecular classifiers in oncology. *Clin. Cancer Res.*, **19**, 4315–4325.
13. Domany,E. (2014) Using high-throughput transcriptomic data for prognosis: a critical overview and perspectives. *Cancer Res.*, **74**, 4612–4621.

14. Ho,J.W.K., Stefani,M., dos Remedios,C.G. and Charleston,M.A. (2008) Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*, **24**, i390–i398.
15. Hulse,A.M. and Cai,J.J. (2013) Genetic variants contribute to gene expression variability in humans. *Genetics*, **193**, 95–108.
16. Haraksingh,R.R. and Snyder,M.P. (2013) Impacts of variation in the human genome on gene regulation. *J. Mol. Biol.*, **425**, 3970–3977.
17. Mostafavi,S., Ortiz-Lopez,A., Bogue,M.A., Hattori,K., Pop,C., Koller,D., Mathis,D., Benoist,C., Blair,D.A., Dustin,M.L. *et al.* (2014) Variation and genetic control of gene expression in primary immunocytes across inbred mouse strains. *J. Immunol.*, **193**, 4485–4496.
18. Bar,H.Y., Booth,J.G. and Wells,M.T. (2014) A bivariate model for simultaneous testing in bioinformatics data. *J. Am. Stat. Assoc.*, **109**, 537–547.
19. Phipson,B. and Oshlack,A. (2014) DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biol.*, **15**, 465.
20. Teschendorff,A.E. and Widschwendter,M. (2012) Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics*, **28**, 1487–1494.
21. Tian,L. and Tibshirani,R. (2011) Adaptive index models for marker-based risk stratification. *Biostatistics*, **12**, 68–86.
22. Ghosh,A.K., Chaudhuri,P. and Sengupta,D. (2006) Classification using kernel density estimates: multiscale analysis and visualization. *Technometrics*, **48**, 120–132.
23. Cun,Y. and Fröhlich,H.F. (2012) Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC Bioinformatics*, **13**, 69.
24. Shi,W., Oshlack,A. and Smyth,G.K. (2010) Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Res.*, **38**, e204–e204.
25. Ganzfried,B.F., Riester,M., Haibe-Kains,B., Risch,T., Tyekucheva,S., Jazic,I., Wang,X.V., Ahmadifar,M., Birrer,M.J., Parmigiani,G. *et al.* (2013) curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *Database*, bat013.
26. Rappaport,N., Nativ,N., Stelzer,G., Twik,M., Guan-Golan,Y., Iny Stein,T., Bahir,I., Belinky,F., Morrey,C.P., Safran,M. *et al.* (2013) MalaCards: an integrated compendium for diseases and their annotation. *Database*, bat018.
27. Strbenac,D., Mann,G.J., Ormerod,J.T. and Yang,J.Y.H. (2015) ClassifyR: an R package for performance assessment of classification with applications to transcriptomics. *Bioinformatics*, **31**, 1851–1853.
28. Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, 3.
29. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
30. Witten,D.M. (2011) Classification and clustering of sequencing data using a Poisson model. *Ann. Appl. Stat.*, **5**, 2493–2518.
31. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
32. Rousseeuw,P.J. and Croux,C. (1993) Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.*, **88**, 1273–1283.
33. Yang,P., Zhou,B.B., Yang,J.Y.-H. and Zomaya,A.Y. (2013) Stability of feature selection algorithms and ensemble feature selection methods in bioinformatics. In: Elloumi,M and Zomaya,AY (eds). *Biological Knowledge Discovery Handbook*. John Wiley & Sons, Inc., Hoboken, pp. 333–352.
34. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
35. Lisowska,K.M., Olbryt,M., Dudaladava,V., Pamula-Pilat,J., Kujawa,K., Grzybowska,E., Jarzab,M., Student,S., Rzepecka,I.K., Jarzab,B. *et al.* (2014) Gene expression analysis in ovarian cancer - faults and hints from DNA microarray study. *Front. Oncol.*, **4**, 6.
36. Tofigh,A., Suderman,M., Paquet,E.R., Livingstone,J., Bertos,N., Saleh,S.M., Zhao,H., Souleimanova,M., Cory,S., Lesurf,R. *et al.* (2014) The prognostic ease and difficulty of invasive breast carcinoma. *Cell Rep.*, **9**, 129–142.