2019

Multiplatform biomarker identification using a data-driven approach enables single-sample classification

Ling Zhang

Ishwor Thapa

Christian Haas

Dhundy Raj Bastola

Zhang *et al. BMC Bioinformatics* (2019) 20:601
https://doi.org/10.1186/s12859-019-3140-7

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Multiplatform biomarker identification using a data-driven approach enables single-sample classification

Ling Zhang, Ishwor Thapa, Christian Haas and Dhundy Bastola*

Abstract

Background: High-throughput gene expression profiles have allowed discovery of potential biomarkers enabling early diagnosis, prognosis and developing individualized treatment. However, it remains a challenge to identify a set of reliable and reproducible biomarkers across various gene expression platforms and laboratories for single sample diagnosis and prognosis. We address this need with our Data-Driven Reference (DDR) approach, which employs stably expressed housekeeping genes as references to eliminate platform-specific biases and non-biological variabilities.

Results: Our method identifies biomarkers with "built-in" features, and these features can be interpreted consistently regardless of profiling technology, which enable classification of single-sample independent of platforms. Validation with RNA-seq data of blood platelets shows that DDR achieves the superior performance in classification of six different tumor types as well as molecular target statuses (such as *MET* or *HER2*-positive, and mutant *KRAS*, *EGFR* or *PIK3CA*) with smaller sets of biomarkers. We demonstrate on the three microarray datasets that our method is capable of identifying robust biomarkers for subgrouping medulloblastoma samples with data perturbation due to different microarray platforms. In addition to identifying the majority of subgroup-specific biomarkers in CodeSet of nanoString, some potential new biomarkers for subgrouping medulloblastoma were detected by our method.

Conclusions: In this study, we present a simple, yet powerful data-driven method which contributes significantly to identification of robust cross-platform gene signature for disease classification of single-patient to facilitate precision medicine. In addition, our method provides a new strategy for transcriptome analysis.

Keywords: Biomarkers, Single-sample, Classification, Cancer, RNA-Seq, Microarray

Background

Identification of reliable and reproducible biomarkers can contribute to reveal patterns of disease heterogeneity. Recent advances in High-throughput sequencing (HTS) technology, such as microarray [1, 2] and RNA-Seq [3–5] have enabled us to profile entire gene expression at low costs. The massive amounts of gene expression profile data generated by HTS have provided a great opportunity to identify reliable biomarkers which facilitate diagnosis, prognosis or treatment of patients. The technological biases across gene expression platforms and non-biological variabilities make it challenging to identify robust gene signature for cross-platform and cross-laboratory classification. Several techniques have been developed to eliminate platform-specific biases [6–8]. However, these methods require multiple samples when processing transcriptome data, which is infeasible for analysis of biomarkers in samples obtained from single patient.

Since gene expression data are high-dimensional data, an important research aim in analysis of transcription profiles is the discovery of small subset of biomarkers containing the most discriminant information, also known as feature selection [9], for accurate assignment of molecular subtype of disease. During the past years, numerous gene selection methods have been developed based on gene expression data and applied in disease classification.

*Correspondence: dkbastola@unomaha.edu
School of Interdisciplinary Informatics, University of Nebraska at Omaha, 110 S 67th St, Omaha, NE 68182, USA

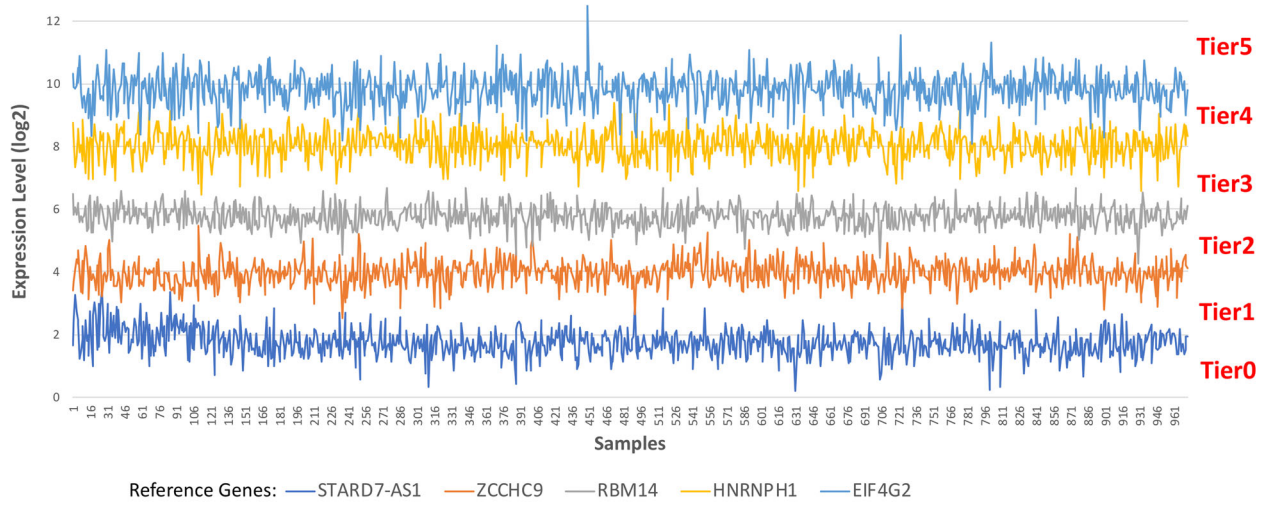
In general, the gene selection methods fall into four categories: filter methods [10–12], wrapper methods [13–15], hybrid methods [16, 17] and embedded methods [18, 19]. However, there is a lack of feature selection methods designed to select robust features (or genes) which enable cross-platform classification of single disease sample.

To address these challenges, we present a Data-Driven Reference (DDR) approach to identify robust cross-platform gene signature for classification of single-sample from various platforms. Our DDR algorithm consists of three main steps: 1) the stably expressed housekeeping genes are employed as references to create a contingency table for each gene using given gene expression dataset; 2) Fisher's exact tests are applied in contingency tables to identify differentially expressed genes (DEGs) as potential biomarkers between two conditions; 3) the categories which the expression levels of biomarkers fall into based on selected reference genes serve as input to the classifier. The reference genes are the housekeeping genes whose expression values remain relatively constant across all samples from different conditions. The categories generated by stably expressed reference genes represent the relative positions of biomarkers, which have a consistent interpretation across gene expression platforms and eliminate sample-specific biases. We illustrate DDR's utility through various evaluations and comparisons with gene signatures identified by existing methods. We demonstrate that DDR method contributes significantly to identification of robust cross-platform gene signature for disease classification of single-patient to facilitate precision medicine.

Results

Identification of potential biomarkers in various expression platforms

Differential expression analysis has been widely used to identify potential biomarkers for diagnosis and prognosis [20]. Using DDR to identify discriminant genes between two conditions involves first two steps: constructing the contingency table for each gene from expression data based on selected reference genes, and then, using the Fisher's exact test to determine if there is a significantly different expression for that gene between two groups (see "Methods" section for details). For example, five reference genes (*STARD7-AS1*, *ZCCHC9*, *RBM14*, *HNRNPH1*, and *EIF4G2*) from TCGA-BRCA RNA-Seq dataset were selected, so that log2-fold-changes between expressions of two consecutive reference genes were around 2 (Fig. 1 and Additional file 2). Gene expression heatmap was constructed to show the relative expression patterns of the top 20 (ranked based on FDR values) most significant DEGs in comparison between triple-negative breast



	Tier0	Tier1	Tier2	Tier3	Tier4	Tier5
TNBC	59	39	12	4	1	0
Others	21	33	29	189	379	207

Fig. 1 Expression levels of five data-driven reference genes from TCGA-BRCA RNA-Seq samples and tiers classified based on reference genes (top). The contingency table of *ESR1* expression based on expression levels of reference genes (bottom)

cancer (TNBC) and the other subtypes (Additional file 1: Figure S2A). Most of top 20 DEGs were down-regulated in TNBC samples compared with the other subtypes of breast cancer. TNBC has a poor prognosis compared with other types of breast cancer due to lack of therapeutic targets. In this study, we examined top 10 up-regulated genes (long non-coding RNA, LINC02188, was not included) (Additional file 1: Figure S2B), and at least six genes have been very recently (*BCL11A*, *FOXC1*, *CDCA7*, *PSAT1*, *UGT8*, and *GABRP*) experimentally validated for clinical or functional relevance in growth and metastasis of TNBC [21–26]. Four other genes (*B3GNT5*, *PPP1R14C*, *RGMA*, and *HAPLN3*) were also computationally selected as signature genes in TNBC[27–29]. The DDR was also performed in LUAD RNA-Seq dataset from TCGA and DEGs between LUAD samples and healthy samples were identified based on selected reference genes (Additional file 3 and Additional file 1: Figure S3). The method presented here can be applied as well to microarray expression data. Four reference genes (Additional file 1: Figure S4) were selected from expression microarray data (Accession: GSE62872), so that the differences between expressions of two consecutive reference genes were around 2. Then, the DEGs between prostate cancer and health were identified by DDR and ranked by adjusted *p*-value (Additional file 4).

To assess DDR's ability to detect DEGs, we compared it with the tools widely used in differential expression analysis in various platforms. The Fisher's exact test is a non-parametric test in the sense that it does not assume that the RNA-Seq read counts or microarray expression data across samples are based on the theoretical probability distribution. On the contrary, current popular tools, such as *DESeq* [30], *DESeq2* [31] and *edgeR* [32], use a negative binomial distribution to model RNA-Seq read counts for assessing differential expression. Linear models for microarray (*limma*) [33] uses linear models based on empirical Bayes method to identify DEGs. To compare DDR with existing tools for analysis of RNA-Seq data, a gene was declared as significantly differentially expressed if FDR (or adjusted *p*-value) was less than 0.01 in *EdgeR* and *DESeq2* methods, or FDR (adjusted p-value) was less than 0.1 in *DESeq* and DDR methods. We measured the precision and recall of the identified DEGs using the DEGs from the datasets of 230 TCGA-BRCA samples (115 TNBC samples and randomly selected 115 other subtypes) and 118 TCGA-LUAD samples (59 normal tissue samples and randomly selected 59 LUAD samples) as the gold standard. The precision and recall values in both datasets for different methods and different numbers of samples per group are illustrated in Fig. 2. Two *EdgeR* methods reported high values for precision in both datasets across different sample sizes (Fig. 2a and d). *DESeq2* achieved high performance in precision similar to *EdgeR* methods in BRCA dataset, but showed a slight decrease in LUAD dataset (Fig. 2a and d). For DDR, the precision values remained relatively high in LUAD datasets when sample size was reduced and was slightly reduced in BRCA dataset (Fig. 2a and d). On the contrary, *DESeq* showed lower values for precision with respect to all other tools. The recall values rapidly decreased for all the tools when the number of

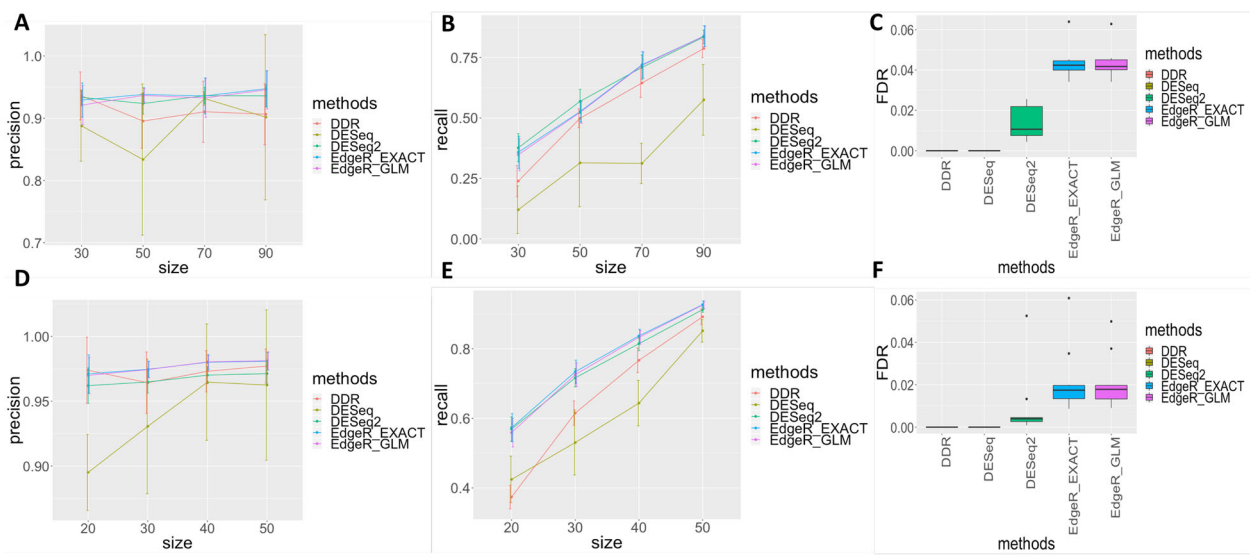


Fig. 2 Precision and recall of methods in two TCGA RNA-Seq datasets: precision (**a**) and recall (**b**) in BRCA data and precision (**d**) and recall (**e**) in LUAD data. The false discovery rate (FDR) on the basis of mock comparisons generated using two datasets: 116 TNBC samples from TCGA-BRCA dataset (**c**) and 59 normal samples from LUAD (**f**)

samples per group was decreased (Fig. 2b and e). *DESeq2*, *EdgeR_GLM* and *EdgeR_EXACT* outperformed the other methods and *DESeq* was the worst-performing method in analysis for both datasets. DDR resulted in intermediate values of recall with respect to all other tools (Fig. 2b and e).

To evaluate the false discovery rate (FDR) of the tools in analysis of RNA-Seq data, we generated mock comparisons from two datasets: the first consisted of 115 triple-negative breast cancer samples by randomly dividing the samples into two non-overlapping groups (57 samples for one group and 58 samples for the other group) and second consisted of 59 normal samples from TCGA-LUAD dataset by randomly dividing the samples into two non-overlapping groups (29 samples for one group and 30 samples for the other group). The median FDRs of both *EdgeR* methods were higher compared with the other methods in both datasets (Fig. 2c and f). *DESeq2* performed better than *EdgeR* methods and controlled the FDRs well (around 0.01 in BRCA dataset (Fig. 2c) and < 0.01 in LUAD dataset (Fig. 2f)). DDR and *DESeq* demonstrated extremely better control on false discovery rate compared with the other tools. It is essential to control false positives so that reliable and reproducible biomarkers can be identified.

Finally, we compared the overlaps of DEGs identified by the different methods through computing overlap coefficient (Szymkiewicz-Simpson coefficient) [34]. The overlaps between the methods are listed in Additional file 1: Table S4. In LUAD dataset, 80% of DEGs, identified using DDR (FDR $<$ 0.1), coincided with DEGs identified using *EdgeR* (FDR $<$ 0.01) or *DESeq2* (adjusted p-value $<$ 0.01). The use of DDR and *DESeq2* (or *EdgeR*) algorithms achieved higher overlap rate in DEG results from BRCA dataset. DESeq generated DEG list overlapped poorly with that from DDR ($<$ 52%) in both datasets. It is no surprise that the highest overlap percentages were observed between *DESeq* and *DESeq2* DEG lists or between *EdgeR_EXACT* and *EdgeR_GLM DEG* lists.

We benchmarked DDR approach against *limma* by using prostate cancer microarray data. Similarly as in analysis of RNA-Seq data, we used 240 samples (randomly selected 120 samples from each group) as the gold standard and measured the precision (Additional file 1: Figure S5A) and recall (Additional file 1: Figure S5B) for both methods in different sample size per group. *limma* performed better in term of precision. The recall values were systematically lower than precision values for DDR and *limma*. Additional file 1: Figure S5C shows that both DDR and *limma* appeared extremely conservative in controlling FDR in this analysis. DDR and *limma* DEG lists achieved 83% overlap with each other.

Feature selection and cross-platform single-sample classification

In this section, we provided an example of using DDR to select signature genes between TNBC and other types of BRCA using TCGA RNA-Seq dataset, and use features of selected genes to classify BRCA samples from a different expression profiling platform (Microarray. GEO accession: GSE27447). DDR was applied to TCGA-BRCA RNA-Seq dataset to identify a list of ranked DEGs (ranked by adjusted *p*-value) (Additional file 2). The small subset of 4 genes (*ESR1*, *AGR2*, *AGR3*, *FOXA1*) was selected as biomarkers for classification based on adjusted *p*-value ($< 1 \times 10^{-60}$) and Expression Distance (ED $>$2.5). Combination of FDR and ED for selection of signature genes enables not only identifying genes containing most discriminant information but also leading to more reproducible biomarkers. Fisher's exact tests were employed to identify top DEGs as potential biomarkers, which come with the most differentially relative positions in comparison to reference genes, named as "built-in" features. Since the positions are relative, they can be robust features and be used for single sample classification. These "built-in" features (Additional file 5) served as input to train classifiers. Here, we compared the performance of different classifiers from *Scikit-learn* [35] for TNBC classification using categorized expression of four signature genes as feature. From Additional file 1: Figure S6, it can be seen that SVM achieved slightly better performance (Accuracy: 94%) though the other classifiers performed as well on classification task. Most of non-TNBC breast cancer samples were correctly predicted (Accuracy: 97%), whereas the proportion of mis-assigned TNBC samples was higher (Additional file 1: Table 1A). To evaluate the capacity of four selected signature genes and SVM classifier trained on TCGA-BRCA dataset in cross-platform classification of single-samples, the microarray dataset containing 5 TNBC samples and 14 non-TNBC samples was collected from GEO (Accession: GSE27447) [36]. GSE27447 data (.CEL files) were normalized by *affy* package in R. When using tiered classifications of four genes selected above (Additional file 5) based on reference genes from GSE27447 dataset as input to SVM classifier trained on TCGA-BRCA dataset, 5/5 (100%) and 11/14 (79%) were classified correctly to TNBC and non-TNBC, respectively (Table 1b). These examples demonstrate DDR's ability to identify robust biomarkers for cross-platform classification of single patient.

Identification and analysis of different cancer subtypes using RNA-Seq of tumor-educated platelets

Molecular information in non-invasive liquid biopsy offers the promise of detection and classification of cancer subtypes [37, 38]. In this study, we employed RNA-Seq data of blood platelets to evaluate DDR in its ability

Table 1 Classification Performance on (**A**) TCGA-BRCA RNA-Seq Dataset and (**B**) GSE27447 BRCA Microarray Test Dataset using 4 Signature Genes

A

Actual Class \ Predicted Class	TNBC	Other
TNBC	75%	25%
Other	3%	97%

Accuracy: 94%
Recall: 86%

B

Actual Class \ Predicted Class	TNBC	Other
TNBC	5/5(100%)	
Other	3/14(21%)	11/14(79%)

Accuracy: 84%
Recall: 89%

to identify biomarkers for the classification of cancer subtypes and status of therapy-targeting genes. The platelets as liquid biopsy are capable of carrying RNA molecules from tumor tissues (educating), and serve as potential non-invasive biomarker source for detecting and monitoring cancers [38, 39]. These platelets are known as Tumor-Educated Platelets (TEP), of which the RNA profiles could be used to subgroup the cancers [38].

Here, we applied DDR to identify the subsets of discriminant genes from RNA-Seq data of TEP from GEO (Accession: GSE68086) [38] and employed SVM classifier on classification tasks. Furthermore, we compared the classification performance of our method with that of Best et al. (see "Discussion" section for details) [38]. To classify pan-cancer samples representing six tumor types (breast cancer (BRCA, n = 39), colorectal cancer (CRC, n = 42), glioblastoma (GBM, n = 40), non-small cell lung cancer (NSCLC, n = 60), hepatobiliary cancer (HBC, n = 14), and pancreatic cancer (PAAD, n = 35)) and healthy donors (HD, n = 55), the DDR was applied to each pairwise comparison among groups to identify DEGs, and then a small subsets of biomarkers were selected from DEG lists of pairwise comparisons based on adjusted *p*-values and ED (see Additional file 6 for details). These subsets of biomarkers were merged and duplicate genes were removed to generate a list comprising 596 genes (Additional file 7) for pan-cancer classification. The tiered categorizations of 596 genes were used as input to multi-class One-versus-One (OvO) SVM classifier to yield overall accuracy of 72% (Table 2a). Similarly, for discriminating three different types of adenocarcinomas (CRC, PAAD and HBC) from gastro-intestinal tract, we selected 144 genes (Additional file 7) and performed OvO SVM classifier, yielding an overall accuracy of 80% (Table 2b).

Table 2 Classification Performance on (**A**) Multi-class Cancers and (**B**) Gastro-intestinal Tract Cancers

A

Actual Class \ Predicted Class	Healthy	BrCa	CRC	GBM	HBC	NSCLC	PAAD	Genes: 596
Healthy	95%	2%		2%		1%		
BrCa	1%	73%	8%			12%	6%	
CRC		8%	65%			16%	11%	
GBM	19%		1%	73%		5%	2%	
HBC	5%		14%	3%	55%	15%	8%	
NSCLC	5%	13%	15%	2%		61%	4%	
PAAD		6%	13%		1%	8%	72%	

Accuracy: 72%, Recall: 70%

B

Actual Class \ Predicted Class	CRC	HBC	PAAD	Genes: 144
CRC	87%	1%	12%	
HBC	10%	78%	12%	
PAAD	25%	3%	72%	

Accuracy: 80%, Recall: 79%

Best et al.[38] also reported that RNA profiles from TEP could be used to discriminate tumor patients with different status of therapy-targeting oncogenes, such as *KRAS*-mut vs *KRAS*-wt in CRC, HBC, NSCLC, and PAAD patients, *EGFR*-mut vs *EGFR*-wt in NSCLC patients, *MET*+ vs *MET*- in NSCLC patients, *PIK3CA*-mut vs *PIK3CA*-wt in BRCA, *HER2*+ vs *HER2*- in BRCA, as well as triple-negative breast cancer. The number of biomarkers (Additional files 6 and 7) and classification accuracies were presented in Tables 3a–i.

Identification and analysis of biomarkers for subgrouping medulloblastoma microarray data

Medulloblastoma (MB) is the most common malignant brain tumor in children and represents approximately 20% of childhood brain tumors [40]. Transcriptional profiling of MB identified four distinct molecular subgroups: WNT (Wnt signaling pathway), SHH (sonic hedgehog signaling pathway), Group 3 (G3) and Group 4 (G4) [41]. The nearest shrunken centroid and *t* test were combined to select 22 medulloblastoma subgroup-specific signature genes (CodeSet), and then medulloblastoma samples were subgrouped by measuring the expression level of 22 subgroup-specific genes [42, 43]. Here, the DDR method was applied for identifying signature genes for each MB subgroup using microarray dataset (Accession: GSE37418 [44]), and these genes were used to subgroup medulloblastoma samples from different microarray platforms. Four reference genes (*C1orf127*, *ZNF347*, *WDR70* and *HNRNPK*) (Additional file 1: Figure S7) were selected and DEGs (adjusted p-value $< 1 \times 10^{-5}$) for each subgroup were identified by performing DDR for each subgroup against the other subgroups (Additional file 8). Then, top 5 genes (non-coding RNAs were not included) for each subgroup were selected based on ED values. A total of 20 genes included: WNT (*WIF1*, *GAD1*, *DKK2*, *TRDV3*, *SHOX2*), SHH (*PDLIM3*, *EYA1*, *HHIP*, *CRB1*, *SFRP1*), G3 (*TRIM58*, *GABRA5*, *PALMD*, *NPR3*, *HLX*), G4 (*EOMES*, *NWD2*, *PTPN5*, *RBM24*, *UNC5D*), and their expression heatmap is presented in Fig. 3a. Among these 20 genes, 12 overlap with medulloblastoma subgroup-specific signature genes (CodeSet) from NanoString Technologies, Inc. [42]. The tiered categorizations of 20 signature genes (Additional file 9) were used as input to OneVsRestClassifier from *Scikit-learn* [35] using SVM over 1000 Monte Carlo cross-validation (MCCV) iterations to yield overall accuracy of 99% and recall of 99% (Table 4a).

The classification capacity of identified signature genes and OneVsRest classifier was evaluated on two

Table 3 Classification Performance for Molecular Pathway Diagnostics

A

Actual Class \ Predicted Class	KRAS mut	KRAS wt	Genes: 9
CRC			
KRAS mut	92%	8%	
KRAS wt	4%	96%	

Accuracy: 95% Recall: 94%

B

Actual Class \ Predicted Class	KRAS mut	KRAS wt	Genes: 20
PAAD			
KRAS mut	100%	0	
KRAS wt	16%	84%	

Accuracy: 94% Recall: 92%

C

Actual Class \ Predicted Class	KRAS mut	KRAS wt	Genes: 16
NSCLC			
KRAS mut	81%	19%	
KRAS wt	8%	92%	

Accuracy: 87% Recall: 86%

D

Actual Class \ Predicted Class	KRAS mut	KRAS wt	Genes: 30
CRC PAAD NSCLC HBC			
KRAS mut	69%	31%	
KRAS wt	12%	88%	

Accuracy: 80% Recall: 79%

E

Actual Class \ Predicted Class	EGFR mut	EGFR wt	Genes: 18
NSCLC			
EGFR mut	80%	20%	
EGFR wt	6%	94%	

Accuracy: 88% Recall: 87%

F

Actual Class \ Predicted Class	MET+	MET-	Genes: 10
NSCLC			
MET+	90%	10%	
MET- wt	0	100%	

Accuracy: 96% Recall: 95%

G

Actual Class \ Predicted Class	PIK3CA mut	PIK3CA wt	Genes: 3
BrCa			
PIK3CA mut	51%	49%	
PIK3CA wt	0	100%	

Accuracy: 90% Recall: 76%

H

Actual Class \ Predicted Class	HER2+	HER2-	Genes: 8
BrCa			
HER2+	83%	17%	
HER2- wt	0	100%	

Accuracy: 93% Recall: 91%

I

Actual Class \ Predicted Class	TNBC	OTHER	Genes: 57
BrCa			
TNBC	99%	1%	
OTHER	0	100%	

Accuracy: 100% Recall: 100%

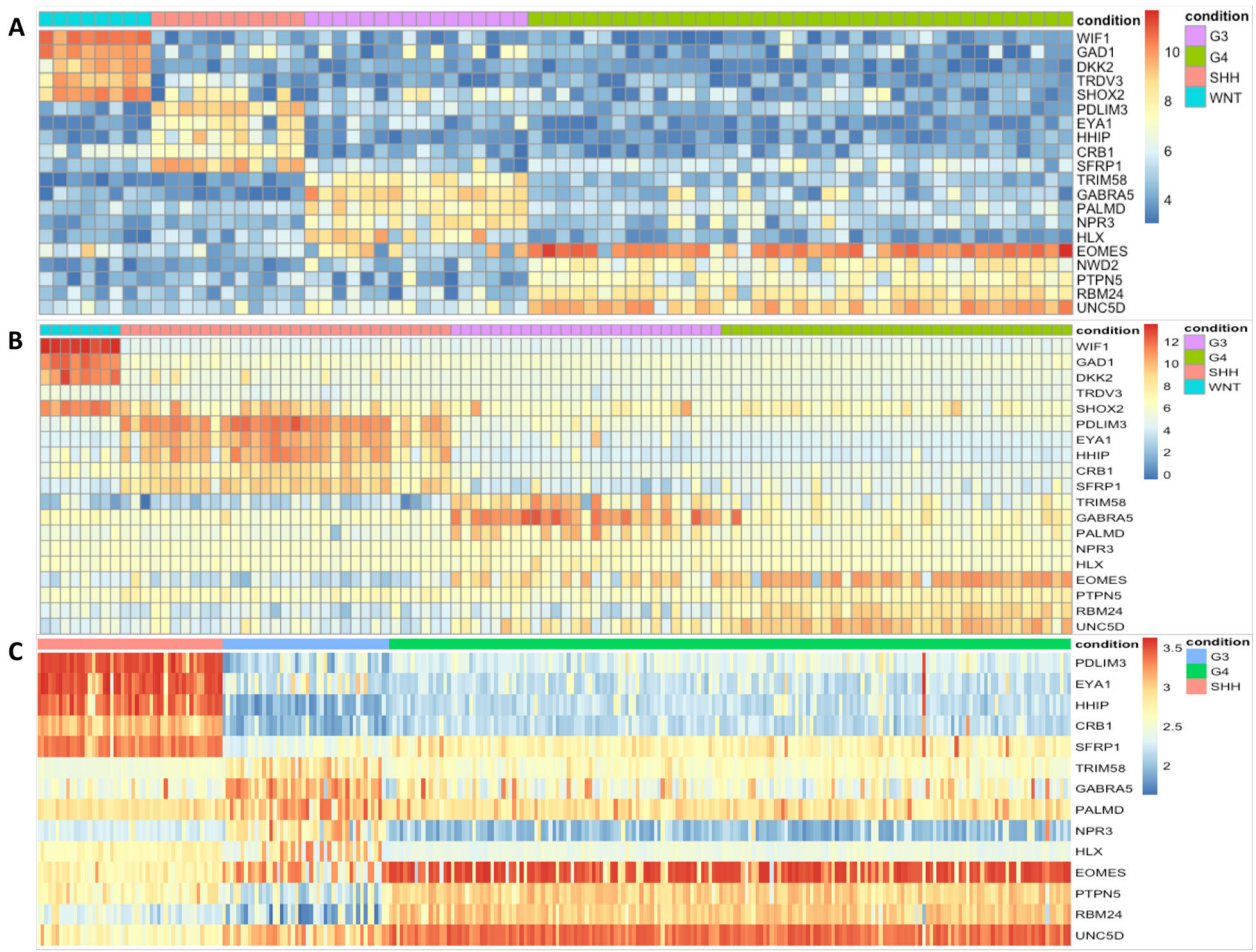


Fig. 3 The expression heatmaps for signature genes in GSE37418[44] (**a**), GSE21140[41] (**b**), and GSE37382[45] (**c**)

independent datasets ((Accession: GSE21140 [41] and GSE37382 [45]) from different microarray platforms. Since signature gene *NWD2* from training dataset (GSE37418) was not available in GSE21140 datasets, the other 19 signature genes (Additional file 9) were used for classification (Fig. 3b). Using OneVsRest SVM classifier trained on GSE37418 dataset, 94/103 (∼ 91%) samples from GSE21140 were assigned to appropriate subgroups (Table 4b). In GSE21140 dataset, all WNT and SHH samples were correctly classified, seven G3 samples were misclassified to G4 subgroup (7/27), and two G4 cases were misclassified to G3 group (Table 4b).

Table 4 Classification Performances on medulloblastoma samples (**A**) cross-validation analysis for GSE37418 dataset, (**B**) GSE21140 dataset, and (**C**) GSE37382 dataset

A

Predicted Class

Actual Class	WNT	SHH	Group 3	Group 4
WNT	100%			
SHH		100%		
Group 3			100%	
Group 4			3%	97%

Accuracy: 99%, Recall: 99%

B

Predicted Class

Actual Class	WNT	SHH	Group 3	Group 4
WNT	8/8(100%)			
SHH		33/33(100%)		
Group 3			20/27(74%)	7/27(26%)
Group 4			2/35(6%)	33/35(94%)

Accuracy: 91%, Recall: 92%

C

Predicted Class

Actual Class	SHH	Group 3	Group 4
SHH	51/51(100%)		
Group 3		33/46(72%)	13/46(28%)
Group 4	1/188(0.5%)	2/188(1.0%)	185/188(98.4%)

Accuracy: 94%, Recall: 90%

There were only three subgroups of samples (SHH, G3 and G4) in GSE37382 dataset and 14 signature genes (Additional file 9 and *NWD2* was not included) were used for subgrouping. Using OneVsRest SVM classifier trained on GSE37418 dataset, 51/51 (100%) SHH samples, 33/46 (72%) G3 samples and 185/188 (98%) G4 samples were correctly classified to appropriate subgroups, respectively, which resulted in accurate classification of 94% in GSE37382 dataset (Table 4c). To better characterize non-SHH/non-WNT (G3 and G4) MB, we applied DDR using the same reference genes as above to identify DEGs between G3 (adjusted p value $< 1 \times 10^{-4}$) and G4 (adjusted p value $< 1 \times 10^{-5}$) (Additional file 8), then three up-regulated genes (non-coding RNAs were not included) with maximum ED values from DEGs were selected for G3 and G4, respectively (Additional file 1: Figure S8A). Using SVM classifier, we correctly classified G3 and G4 subgroups with average 96% accuracy using MCCV (Table 5a). Subsequent validation using six signature genes and SVM classifier trained on GSE37418 dataset, yielded accuracies of 90% and 94% in GSE21140 and GSE37382 datasets, respectively, when subgrouping G3 and G4 (Additional file 1: Figures S8B and S8C). More G3 cases were correctly assigned in both validation datasets compared with predictions above (Table 5b and c). It is worthwhile to note that the expression level of *EN2*, which improved G3/G4 classification performance, has been reported to alter glioma cell morphology [46].

Discussion

In the past decades, a wide variety of methods have been developed to identify biomarkers (feature selections) for classification of diseases using gene expression profiling. However, these approaches posed serious reproducibility challenge when classifying cross-platform samples individually due to technological and platform biases. To overcome this limitation, we present a simple, yet powerful data-driven method that does not require distribution-based modeling for gene expression analysis and it identifies potential biomarker genes with "built-in" features (categorized tiers based on reference genes) for the classification of single-sample from distinct platforms.

The huge amount of gene expression profiling data has been accumulated over the past decades and deposited in public databases such as GEO [47] and TCGA [48]. These data can be great resources to detect significantly differentially expressed genes (DEGs). These DEGs may be considered as potential biomarkers for disease classification or therapeutic targets [20]. The expression values are grouped in discrete intervals based on the expression levels of reference genes before applying Fisher's exact test. In this study, we evaluated the ability of DDR to identify DEGs in multiple platforms through comparing with several well-established methods, *edgeR* [32], *DESeq* [30] and *DESeq2* [31] on two RNA-Seq datasets: TCGA-BRCA and TCGA-LUAD, and *limma* [33] on a microarray prostate dataset. Although not a best performer, DDR still had relatively high precision and recall values for detecting differentially expressed genes across gene expression profiling platforms. The overlaps between DDR and *DESeq2* (and *edgeR*) were higher (>80%). These results demonstrated that DDR retains information of expression values well on analysis of DEGs. More importantly, DDR controlled the number of false positives better, which guaranteed identifying reliable biomarkers. Unlike *edgeR*, *DESeq*, *DESeq2* and *limma*, DDR is a data-driven non-parametric method which requires fewer assumption about data and is robust to outliers. As a result, it deals with cross-platform profiling gene expression and technical bias well. The utilization of reference genes at different expression levels effectively combines p-value and fold-change to identify reliable DEGs (biomarkers). In addition, employing logarithmic expression levels when selecting reference genes provides wiggle room to deal with overdispersion in RNA-Seq data. Our method provides a better methodological advantage to identify reliable and reproducible potential biomarkers from various expression profiling platforms. DDR can also be employed to detect DEGs among multiple conditions by designing appropriate contingency tables.

Table 5 Classification Performances between Group 3 and Group 4 on (A) cross-validation analysis for GSE37418 dataset, (B) GSE21140 dataset, and (C) GSE37382 dataset

A

Actual Class \ Predicted Class	Group 3	Group 4
Group 3	94%	6%
Group 4	2%	98%

Accuracy: 96%, Recall: 96%

B

Actual Class \ Predicted Class	Group 3	Group 4
Group 3	23/27(85%)	4/27(15%)
Group 4	2/35(6%)	33/35(94%)

Accuracy: 90%, Recall: 90%

C

Actual Class \ Predicted Class	Group 3	Group 4
Group 3	36/46(78%)	10/46(22%)
Group 4	5/188(3%)	183/188(97%)

Accuracy: 94%, Recall: 88%

TCGA is a comprehensive molecular profiling project that compiles clinical and genomic data from samples of different human tumor types [48], and provide a great opportunity to identify genome-wide biomarkers for the classification of cancer conditions. In this study, we employed TCGA-BRCA datasets as training datasets and used DDR to identify small subsets of biomarkers for discriminating TNBC from the other subtypes of BRCA. The "built-in" features used as input to the classifiers effectively eliminated platform-based biases and avoided perpetuating biases from one sample into another. DDR's ability to identify cross-platform features and classify single sample can leverage information from gene expression data that have been accumulated over the past decade and integrate them with samples now being profiled with next generation technologies. Additionally, the simplicity of these features makes them robust across various classifiers in spite of our using of SVM classifier for the classification in this study. The ability of DDR in classifying individuals into appropriate disease groups makes it an ideal choice in personalizing tool and a therapeutic strategy based on specific subgroups of cancer.

Tumor-educated platelets (TEP) is able to serve as potential noninvasive source of tumor-related RNA biomarkers [38, 39, 49]. Best et al. employed *edgeR* using RNA-Seq profiling of TEP to yield DEGs for pinpointing the location of primary tumor with 71% accuracy across six types of tumors and distinguish cancer patients with different molecular subtypes as well. In this study, we applied the DDR method to identify potential biomarkers using RNA-Seq data of TEP and identify multiclass cancer and molecular subclass. We were successful in achieving comparable classification performance with fewer biomarkers compared with results reported by Best et al. [38]. Much smaller sets of biomarkers associated with cancer molecular subtypes (e.g. *MET* and *HER2*-positive, or *EGFR*, *PIK3CA* and *KRAS* mutations) were identified by DDR as compared to biomarker sets from Best et al., which makes it more practical in blood-based cancer diagnostics and therapeutic target identification. Selection of smaller subset of biomarkers can reduce over-fitting and computational complexity by removing redundant features.

DDR employs Fisher's exact test, a non-parametric method, to analyze gene expression profiling, so it could be equally applied in the analysis of microarray data. To evaluate the performance of DDR in identification of signature genes and classification of microarray data, we applied DDR to a published microarray dataset of medulloblastoma (GSE37418) and derived 20 signature genes for medulloblastoma subgroups. Among 20 signature genes, 12 genes overlay NanoString codeset [42] from a commercial instrument system, which suggests our method is reliable in discovering biomarkers. The application of the classifier trained on GSE37418 dataset yielded high accuracy rate for subgrouping WNT, SHH and G4 in two independently validated dataset, confirming classification reproducibility of biomarkers identified by DDR. G3 and G4 subgroups display more similarity to each other in transcriptional profiling compared with WNT and SHH subgroups [41], so it is a challenge to discriminate G3 from G4. In this study, we applied DDR to identify signature genes which were differently expressed between G3 and G4, and achieved improvement in G3 assignment. A recent study suggests that non-SHH/non-WNT medulloblastoma may comprise of three subgroups rather than just G3 and G4 [50], which may explain the observed low accuracy for G3 and G4 subtyping.

Conclusions

The main technical novelty of this work is the combination of data-driven reference genes with non-parametric Fisher's exact test for discovering potential biomarkers. This not only allowed us to identify differentially expressed genes but also help to extract corresponding "built-in" features based on reference genes. One of the exciting outcomes of these "built-in" features is their reliability and reproducibility in classifying disease samples involved with technical bias and cross-platform, which allow us to analyze single sample of disease. This study has shown that DDR can be a promising tool for the identification of biomarkers for precision medicine. And some expression assay (e.g. quantitative PCR) based on these biomarkers and reference genes can be easily developed for diagnosis, prognosis and developing individualized treatment in the future. Finally, although this study has focused on cancer classification, it could be equally useful in classification of other diseases such as Parkinson or Alzheimer's. In conclusion, we have developed a novel, reliable and reproducible data-driven method for identification of potential biomarkers for single-sample classification.

Methods

Data

RNA-Seq from The Cancer Genome Atlas (TCGA). All RNA-Seq read count data from TCGA lung adenocarcinoma (LUAD) project (n=594) were retrieved using the GDC Data Transfer Tool [51]. Both LUAD and normal data were collected, resulting in 535 cancerous and 59 normal tissue samples. All RNA-Seq read count data for breast invasive carcinoma (BRCA) project (n=1222) were downloaded from TCGA. Both cancer and normal samples were collected, resulting in 1109 BRCA and 113 normal samples. Clinical files were downloaded from TCGA data portal for all BRCA samples using GDC client tool. We identified 115 triple-negative breast cancer (TNBC)

samples and 858 samples of the other subtypes based on the annotation provided in the clinical files.

RNA-Seq from Gene Expression Omnibus (GEO). RPKM normalized RNA-Seq data of LUAD were downloaded from the Gene Expression Omnibus (GEO, accession: GSE40419) [52]. The dataset contains 87 lung adenocarcinomas samples and 77 corresponding normal samples.

Microarray data from GEO. We retrieved mRNA expression microarray data set from GEO under the accession number GSE62872 [53] which was generated using platform GPL19370. These 424 samples consisting of 264 samples of prostate tumor and 160 samples of normal tissue were used. Additionally, the microarray data of 5 TNBC samples and 14 non-TNBC samples were downloaded from GEO profile data of GSE27447 [36].

RNA-Seq of Tumor-Educated Platelets. The gene expression profiles of 285 blood platelet samples were downloaded from GEO under the accession number GSE68086 [38]. The samples consisted of breast cancer (BRCA), colorectal cancer (CRC), glioblastoma (GBM), hepatobiliary cancer (HBC), non-small cell lung cancer (NSCLC), pancreatic cancer (PAAD) and healthy donors (HD) (Additional file 1: Table S1).

Microarray data of Medulloblastoma. All microarray data are downloaded from the GEO database under assession number GSE37418 [44], GSE21140 [41], and GSE37382 [45], respectively. The detailed number of samples for each subtype of medulloblastom from three datasets is listed in Additional file 1: Table S2.

Data preprocess

When analyzing RNA-Seq data, DDR expects count data as input. The count data can be obtained through TopHat-HTSeq pipeline [54, 55]. For microarray data, the data normalized by using Bioconductor package *affy* were used as input to DDR. Before running DDR, the input data were re-arranged to group the samples under the different conditions.

Selection of reference genes

Reference genes from RNA-Seq data were identified using a data-driven approach similar to that developed by Hoang et al [56]. The normalization of RNA-Seq read counts was performed using Trimmed Mean of M-values (TMM) in edgeR [32, 57], and then the normalized values were transformed into Counts Per Million (CPM). The CPM values for all genes in RNA-Seq datasets were used to generate two metrics across the samples, namely, the coefficient of variation (COV) and the maximum fold change (MFC). COV was calculated for each gene i by dividing the standard deviation (σ_i) of its CPM values by the mean (μ_i): $COV_i = \frac{\sigma_i}{\mu_i}$. MFC is the ratio of the maximum value over the minimum CPM expression value, also for each gene. The product score (PS), our final metric for each gene, was calculated by multiplying the COV by the MFC:

$$PS_i = COV_i \cdot MFC_i \tag{1}$$

The genes with the lowest product scores and those that were also included in the list of human housekeeping genes [58] were selected as candidate reference genes. The list of human housekeeping genes was obtained by analyzing data from the Human BodyMap 2.0 project across 16 human tissue types [58]. The top gene from candidate reference genes at a given range of expression was selected as the final reference gene for corresponding expression level. Let $r_1^{(k)}, r_2^{(k)}, ..., r_n^{(k)}$ denote the expression levels of reference genes selected above in sample k, so that $r_1^{(k)} < r_2^{(k)} < ... < r_n^{(k)}$. The gene expression level between $r_j^{(k)}$ and $r_{j+1}^{(k)}$ ($0 \leq j \leq n$) was assigned to Tier j in sample k, here $r_0^{(k)} = 0$ and $r_{n+1}^{(k)} = \infty$ (Fig. 1).

Identification of discriminant genes

When n reference genes were selected from previous step, $n+1$ tiers were generated based on the expression levels of these reference genes. The relative expression position (tier) of each gene in comparison to the reference genes from the same sample was obtained, and then frequency counts of the tiers were obtained across the samples from the same condition. These frequency counts were filled in the cells of a contingency table (each column represents the relative position (tier) of each gene in comparison to the reference genes and each row represents the different conditions. So, a $C \times T$ contingency table was created for each gene, here C is the number of conditions and T is the number of tiers, to display the numbers of samples from a particular condition in which that gene was assigned to a particular tier (Fig. 1). Then, Fisher's exact tests (FETs) were performed for the contingency tables to assess whether the expression level of the gene is independent or correlated with conditions or phenotypes. The resulting p-values were adjusted to account for multiple tests using the p.adjust function in R (method = 'fdr'). In addition to adjusted p-values, we defined expression distance (ED) for each gene to describe a quantity change of the gene expression between conditions. The ED can be used to select up- and down-regulated genes.

$$ED^{(i)} = \frac{\sum_{j=1}^{T} n_{A_j}^{(i)} \cdot j}{n_A} - \frac{\sum_{j=1}^{T} n_{B_j}^{(i)} \cdot j}{n_B} \tag{2}$$

$n_{A_j}^{(i)}$ = the number of samples from group A with gene i in tier j

n_A = the number of samples from group A

$n_{B_j}^{(i)}$ = the number of samples from group B with gene i in tier j

n_B = the number of samples from group B

Benchmarks comparison in identification of discriminant genes

To assess how well DDR performs for identification of DEGs in comparison to the current methods (*DESeq* [30], *DESeq2* [31] and *EdgeR* [32] for RNA-Seq data and *limma* [33] for microarray data), we used some real and simulated data. To compute the true positives, we randomly selected the subsets with different sizes (equal number for each condition) from full datasets, and the random selection was repeated 10 times to avoid sample selection bias. Larger sample size generally leads to increased precision, so the overlapped DEGs identified from full dataset and subset approximate the set of true positives. The tested methods were compared under precision and recall [59] defined as

$$precision(DE_{full}, DE_{subset}) = \frac{\#(DE_{full} \cap DE_{subset})}{\#DE_{subset}} \quad (3)$$

$$recall(DE_{full}, DE_{subset}) = \frac{\#(DE_{full} \cap DE_{subset})}{\#DE_{full}} \quad (4)$$

DE_{full} = the set of identified DEGs in the full data set

DE_{subset} = the set of identified DEGs in a subset of the data

To evaluate the false discovery rate of different methods, we randomly assigned the equal sizes of samples from the same condition without replacement into two groups and the procedure was repeated 10 times. All samples were from the same condition, which means that there should not be any real DEGs, so DEGs identified from simulated datasets arise by chance alone. The false discovery rate was defined as ratio of the number of identified DEGs from simulated dataset to the number of identified DEGs by comparing two conditions from complete dataset. We compared the overlaps of the identified DEGs between the methods by Szymkiewicz-Simpson overlap coefficient.

$$overlap(X, Y) = \frac{|X \cap Y|}{min(|X|, |Y|)} \quad (5)$$

Here, X and Y are the lists of DEGs identified by two methods, respectively.

Feature selection and classification

Small subset of biomarkers for classification was obtained from DEGs by filtering on the basis of low adjusted p-values (FDRs) and high EDs. The feature tables consisting of tiers to which selected biomarkers belong for samples served as input to the classifiers. An example of feature table for classifying TNBC and non-TNBC is shown in Additional file 1: Table S3.

Several machine learning classifiers from *Scikit-learn* [35] Python library were applied to classification. We selected Support Vector Machine (SVM) using RBF kernel with C=1 as final classification model for binary classification and OneVsOneClassifier (or OnevsRestclassifier) using SVM to deal with multi-class classification problems. The samples were randomly separated into training/testing sets with 90% of samples as training and 10% as testing. And then, we followed 1000 iterations in stratified cross-validation analysis which deals with imbalanced classes and used accuracy, precision, recall and F1 scores from *Scikit-learn* to assess the performance of our classification model.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-019-3140-7.

Additional file 1: The file contains eight figures (Figures S1–S8) and four tables (Tables S1-S4) as supplementary results.

Additional file 2: The table contains reference genes and DEGs identified by DDR in TCGA-BRCA RNA-Seq data.

Additional file 3: The table contains reference genes and DEGs identified by DDR in TCGA-LUAD RNA-Seq data.

Additional file 4: The table contains reference genes and DEGs identified by DDR in prostate cancer microarray data from GEO (Accession: GSE62872).

Additional file 5: The table contains "built-in" features of four selected biomarkers from TCGA-BRCA RNA-Seq data and microarray data from GEO (Accession: GSE27447) as input to classifiers.

Additional file 6: The table contains DEGs identified by DDR in RNA-Seq data of Tumor-Educated Platelets from GEO (Accession: GSE68086).

Additional file 7: The table contains "built-in" features from Tumor-Educated Platelets (Accession: GSE68086) as input to classifiers.

Additional file 8: The table contains DEGs identified by DDR in microarray datasets of medulloblastoma from GEO (Accession: GSE37418, GSE21140 and GSE37382).

Additional file 9: The table contains "built-in" features from medulloblastoma (Accession: GSE37418, GSE21140 and GSE37382) as input to classifiers.

Abbreviations

b TEP: Tumor-educated platelets; BRCA: Breast cancer; CRC: Colorectal cancer; DDR: Data-driven reference; DEGs: Differentially expressed genes; DT: Decision tree; ED: Expression distance; FDR: False discovery rate; G3: Group 3; G4: Group 4; GBM: Glioblastoma; GEO: Gene expression omnibus; HBC: Hepatobiliary cancer; HD: Healthy donors; HTS: High-throughput sequencing; KNN: K-nearest neighbors; LD: Linear discriminant; LUAD: Lung adenocarcinoma; MB: Medulloblastoma; NSCLC: Non-small cell lung cancer; PAAD: Pancreatic cancer; RF: Random forest; SHH: Sonic hedgehog signaling pathway; SVM: Support vector machine; TCGA: The cancer genome atlas program; TNBC: Triple-negative breast cancer; WNT: Wnt signaling pathway

Acknowledgments

We would like to thank Dr. Dario Ghersi for extensive conversation and Kaitlin Goettsch for her comments on earlier versions of this manuscript.

Availability and Requirements

- **Project name:** DDR
- **Project home page:** https://github.com/idellyzhang/DDR
- **Operating system(s):** Linux, macOS
- **Programming languages:** Python and R
- **License:** GPLv3
- **Any restrictions to use by non-academics:** License needed

Authors' contributions

LZ and DB conceived the study and designed research. LZ designed the algorithm. LZ and IT implemented the algorithm in R and Python. CH assisted

with method development. LZ drafted the manuscript. All authors read and approved the final manuscript.

Funding
No funding was received for the study.

Availability of data and materials
The source code and datasets are available at: https://github.com/idellyzhang/DDR.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 20 March 2019 Accepted: 9 October 2019
Published online: 21 November 2019

References

1. Kuwabara PE. DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling. Briefings in functional genomics and proteomics. 2003;2(1):80–81. Oxford University Press.
2. Speed T. Statistical Analysis of Gene Expression Microarray Data. Boca Raton: CRC Press; 2003.
3. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by rna-seq. Nat Methods. 2008;5(7):621.
4. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. Transcriptome genetics using second generation sequencing in a caucasian population. Nature. 2010;464(7289):773.
5. Nagalakshmi U, Waern K, Snyder M. Rna-seq: a method for comprehensive transcriptome analysis. Curr Protoc Mol Biol. 2010;89(1):4–11.
6. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics. 2007;8(1):118–27.
7. Thompson JA, Tan J, Greene CS. Cross-platform normalization of microarray and rna-seq data for machine learning applications. PeerJ. 2016;4:1621.
8. Franks JM, Cai G, Whitfield ML. Feature specific quantile normalization enables cross-platform classification of molecular subtypes using gene expression data. Bioinformatics. 2018;34(11):1868–74.
9. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, de Schaetzen V, Duque R, Bersini H, Nowe A. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinforma (TCBB). 2012;9(4):1106–19.
10. Zhu S, Wang D, Yu K, Li T, Gong Y. Feature selection for gene expression using model-based entropy. IEEE/ACM Trans Comput Biol Bioinforma (TCBB). 2010;7(1):25–36.
11. Mandal M, Mukhopadhyay A. An improved minimum redundancy maximum relevance approach for feature selection in gene expression data. Procedia Technol. 2013;10:20–7.
12. Maulik U, Chakraborty D. Fuzzy preference based feature selection and semisupervised svm for cancer classification. IEEE Trans Nanobiosci. 2014;13(2):152–60.
13. Luo L-K, Huang D-F, Ye L-J, Zhou Q-F, Shao G-F, Peng H. Improving the computational efficiency of recursive cluster elimination for gene selection. IEEE/ACM Trans Comput Biol Bioinforma. 2011;8(1):122–9.
14. Sharma A, Imoto S, Miyano S. A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Trans Comput Biol Bioinforma (TCBB). 2012;9(3):754–64.
15. Li Y, Kang K, Krahn JM, Croutwater N, Lee K, Umbach DM, Li L. A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. BMC Genomics. 2017;18(1):508.
16. Shreem SS, Abdullah S, Nazri MZA, Alzaqebah M. Hybridizing relieff, mrmr filters and ga wrapper approaches for gene selection. J Theor Appl Inf Technol. 2012;46(2):1034–9.
17. El Akadi A, Amine A, El Ouardighi A, Aboutajdine D. A two-stage gene selection scheme utilizing mrmr filter and ga wrapper. Knowl Inform Syst. 2011;26(3):487–500.
18. Pang H, George SL, Hui K, Tong T. Gene selection using iterative feature elimination random forests for survival outcomes. IEEE/ACM Trans Comput Biol Bioinforma (TCBB). 2012;9(5):1422–31.
19. Liang Y, Liu C, Luan X-Z, Leung K-S, Chan T-M, Xu Z-B, Zhang H. Sparse logistic regression with a l 1/2 penalty for gene selection in cancer classification. BMC Bioinformatics. 2013;14(1):198.
20. Zhao X-M, Qin G. Identifying biomarkers with differential analysis. In: Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases. New York: Springer; 2013. p. 17–31.
21. Khaled WT, Lee SC, Stingl J, Chen X, Ali HR, Rueda OM, Hadi F, Wang J, Yu Y, Chin S-F, et al. Bcl11a is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. Nat Commun. 2015;6:6987.
22. Pan H, Peng Z, Lin J, Ren X, Zhang G, Cui Y. Forkhead box c1 boosts triple-negative breast cancer metastasis through activating the transcription of chemokine receptor-4. Cancer Sci. 2018;109(12):3794.
23. Ye L, Li F, Song Y, Yu D, Xiong Z, Li Y, Shi T, Yuan Z, Lin C, Wu X, et al. Overexpression of cdca7 predicts poor prognosis and induces ezh2-mediated progression of triple-negative breast cancer. Int J Cancer. 2018;143(10):2602–213.
24. Clem B, Metcalf S, Kruer T, Klinge C. Investigation of phosphoserine aminotransferase 1 and its role in breast cancer progression. In: FASEB JOURNAL, vol. 32. BETHESDA: FEDERATION AMER SOC EXP BIOL 9650 ROCKVILLE PIKE; 2018. p. 20814–3998.
25. Cao Q, Chen X, Wu X, Liao R, Huang P, Tan Y, Wang L, Ren G, Huang J, Dong C. Inhibition of ugt8 suppresses basal-like breast cancer progression by attenuating sulfatide–αvβ5 axis. J Exp Med. 2018;215(6):1679–2.
26. Sizemore GM, Sizemore ST, Seachrist DD, Keri RA. Gaba (a) receptor pi (gabrp) stimulates basal-like breast cancer cell migration through activation of extracellular-regulated kinase 1/2 (erk1/2). J Biol Chem. 2014;289(35):24102–13.
27. Segaert P, Lopes MB, Casimiro S, Vinga S, Rousseeuw PJ. Robust identification of target genes and outliers in triple-negative breast cancer data. Stat Methods Med Res. 20180962280218794722. https://doi.org/10.1177/0962280218794722.
28. Xiao B, Chen L, Ke Y, Hang J, Cao L, Zhang R, Zhang W, Liao Y, Gao Y, Chen J, et al. Identification of methylation sites and signature genes with prognostic value for luminal breast cancer. BMC Cancer. 2018;18(1):405.
29. Santuario-Facio SK, Cardona-Huerta S, Perez-Paramo YX, Trevino V, Hernandez-Cabrera F, Rojas-Martinez A, Uscanga-Perales G, Martinez-Rodriguez JL, Martinez-Jacobo L, Padilla-Rivas G, et al. a new gene expression signature for triple-negative breast cancer using frozen fresh tissue before neoadjuvant chemotherapy. Mol Med. 2017;23:101.
30. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):106.
31. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biol. 2014;15(12):550.
32. Robinson MD, McCarthy DJ, Smyth GK. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
33. Smyth GK. Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York: Springer; 2005. p. 397–420.
34. Vijaymeena M, Kavitha K. A survey on similarity measures in text mining. Mach Learn Appl Int J. 2016;3:19–28.
35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in python. J Mach Learn Res. 2011;12(Oct):2825–30.
36. Yang L, Wu X, Wang Y, Zhang K, Wu J, Yuan Y, Deng X, Chen L, Kim C, Lau S, et al. Fzd7 has a critical role in cell proliferation in triple negative breast cancer. Oncogene. 2011;30(43):4437.
37. Heitzer E, Perakis S, Geigl JB, Speicher MR. The potential of liquid biopsies for the early detection of cancer. NPJ Precis Oncol. 2017;1(1):36.
38. Best MG, Sol N, Kooi I, Tannous J, Westerman BA, Rustenburg F, Schellen P, Verschueren H, Post E, Koster J, et al. Rna-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. Cancer Cell. 2015;28(5):666–76.
39. Best MG, Wesseling P, Wurdinger T. Tumor-educated platelets as a noninvasive biomarker source for cancer detection and progression monitoring. Cancer Res. 2018;78(13):3407–12.

40. Kijima N, KaNemura Y. Molecular classification of medulloblastoma. Neurol Med Chir. 2016;56(11):687–97.
41. Northcott PA, Korshunov A, Witt H, Hielscher T, Eberhart CG, Mack S, Bouffet E, Clifford SC, Hawkins CE, French P, et al. Medulloblastoma comprises four distinct molecular variants. J Clin Oncol. 2011;29(11):1408.
42. Northcott PA, Shih DJ, Remke M, Cho Y-J, Kool M, Hawkins C, Eberhart CG, Dubuc A, Guettouche T, Cardentey Y, et al. Rapid, reliable, and reproducible molecular sub-grouping of clinical medulloblastoma samples. Acta Neuropathol. 2012;123(4):615–26.
43. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natal Acad Sci. 2002;99(10):6567–6572.
44. Robinson G, Parker M, Kranenburg TA, Lu C, Chen X, Ding L, Phoenix TN, Hedlund E, Wei L, Zhu X, et al. Novel mutations target distinct subgroups of medulloblastoma. Nature. 2012;488(7409):43.
45. Northcott PA, Shih DJ, Peacock J, Garzia L, Morrissy AS, Zichner T, Stütz AM, Korshunov A, Reimand J, Schumacher SE, et al. Subgroup-specific structural variation across 1000 medulloblastoma genomes. Nature. 2012;488(7409):49.
46. Wang L, Yang M, Liao S, Liu W, Dai G, Wu G, Chen L. Hsa-mir-27b is up-regulated in cytomegalovirus-infected human glioma cells, targets engrailed-2 and inhibits its expression. Exp Biol Med. 2017;242(12): 1227–33.
47. Clough E, Barrett T. The gene expression omnibus database. In: Statistical Genomics. New York: Springer; 2016. p. 93–110.
48. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM, Network CGAR, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113.
49. Joosse SA, Pantel K. Tumor-educated platelets as liquid biopsy in cancer patients. Cancer Cell. 2015;28(5):552–4.
50. Łastowska M, Trubicka J, Niemira M, Paczkowska-Abdulsalam M, et al. Medulloblastoma with transitional features between group 3 and group 4 is associated with good prognosis. J Neuro-Oncol. 2018;138(2):231–40.
51. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375(12):1109–12.
52. Seo J-S, Ju YS, Lee W-C, Shin J-Y, Lee JK, Bleazard T, Lee J, Jung YJ, Kim J-O, Shin J-Y, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res. 2012;22(11):2109–19.
53. Penney KL, Sinnott JA, Tyekucheva S, Gerke T, Shui IM, Kraft P, Sesso HD, Freedman ML, Loda M, Mucci LA, et al. Association of prostate cancer risk variants with gene expression in normal and tumor tissue. Cancer Epidemiol Prev Biomark. 2015;24(1):255–60.
54. Trapnell C, Pachter L, Salzberg SL. Tophat: discovering splice junctions with rna-seq. Bioinformatics. 2009;25(9):1105–11.
55. Anders S, Pyl PT, Huber W. Htseq—a python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–69.
56. Hoang VL, Tom LN, Quek X-C, Tan J-M, Payne EJ, Lin LL, Sinnya S, Raphael AP, Lambie D, Frazer IH, et al. Rna-seq reveals more consistent reference genes for gene expression studies in human non-melanoma skin cancers. PeerJ. 2017;5:3631.
57. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of rna-seq data. Genome Biol. 2010;11(3):25.
58. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. 2013;29(10):569–74.
59. Jaakkola MK, Seyednasrollah F, Mehmood A, Elo LL. Comparison of methods to detect differentially expressed genes between single-cell populations. Brief Bioinformat. 2016;18(5):735–43.

Publisher's Note