Journal of Visualized Experiments

Selecting multiple biomarker subsets with similarly effective binary classification performances

--Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE57738R3
Full Title:	Selecting multiple biomarker subsets with similarly effective binary classification performances
Keywords:	Biomarker detection; feature selection; OMIC; binary classification; filter; wrapper; extreme learning machine (ELM).
Corresponding Author:	Fengfeng Zhou Jilin University Changchun, Jilin CHINA
Corresponding Author's Institution:	Jilin University
Corresponding Author E-Mail:	fengfengzhou@gmail.com
First Author:	Xin Feng
Other Authors:	Xin Feng
	Shaofei Wang
	Quewang Liu
	Han Li
	Jiamei Liu
	Cheng Xu
	Weifeng Yang
	Yayun Shu
	Weiwei Zheng
	Bingxin Yu
	Mingran Qi
	Wenyang Zhou
Author Comments:	This article was invited by Dr. Ronald Myers.
Additional Information:	
Question	**Response**
If this article needs to be "in-press" by a certain date, please indicate the date below and explain in your cover letter.	

TITLE:

Selecting Multiple Biomarker Subsets with Similarly Effective Binary Classification Performances

AUTHORS AND AFFILIATIONS:

Xin Feng[1], Shaofei Wang[1], Quewang Liu[1], Han Li[2], Jiamei Liu[2], Cheng Xu[2], Weifeng Yang[2], Yayun Shu[2], Weiwei Zheng[1], Bingxin Yu[3], Mingran Qi[4], Wenyang Zhou[1], Fengfeng Zhou[1]

[1]College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, China

[2]College of Software, Jilin University, Changchun, Jilin, China

[3]Ultrasonography Department, China-Japan Union Hospital of Jilin University, Changchun, Jilin, China

[4]Department of Pathogenobiology, College of Basic Medical Science, JilinUniversity, Changchun, Jilin, China

CORRESPONDING AUTHOR:

Fengfeng Zhou (email: FengfengZhou@gmail.com)

KEYWORDS:

Biomarker detection, feature selection, OMIC, binary classification, filter, wrapper, extreme learning machine, ELM

SHORT ABSTRACT

Existing algorithms generate one solution for a biomarker detection dataset. This protocol demonstrates the existence of multiple similarly effective solutions and presents a user-friendly software to help biomedical researchers investigate their datasets for the proposed challenge. Computer scientists may also provide this feature in their biomarker detection algorithms.

LONG ABSTRACT

Biomarker detection is one of the more important biomedical questions for high-throughput 'omics' researchers, and almost all existing biomarker detection algorithms generate one biomarker subset with the optimized performance measurement for a given dataset. However, a recent study demonstrated the existence of multiple biomarker subsets with similarly effective or even identical classification performances. This protocol presents a simple and straightforward methodology for detecting biomarker subsets with binary classification performances, better than a user-defined cutoff. The protocol consists of data preparation and loading, baseline information summarization, parameter tuning, biomarker screening, result visualization and interpretation, biomarker gene annotations, and result and visualization exportation at publication quality. The proposed biomarker screening strategy is intuitive and demonstrates a general rule for developing biomarker detection algorithms. A user-friendly graphical user interface (GUI) was developed using the programming language Python, allowing biomedical researchers to have direct access to their results. The source code and manual of kSolutionVis can be downloaded from http://www.healthinformaticslab.org/supp/resources.php .

INTRODUCTION:

Binary classification, one of the most commonly investigated and challenging data mining problems in the biomedical area, is used to build a classification model trained on two groups of samples with the most accurate discrimination power[1-7]. However, the big data generated in the biomedical field has the inherent "large p small n" paradigm, with the number of features usually much larger than the number of samples[6,8,9]. Therefore, biomedical researchers have to reduce the feature dimension before utilizing the classification algorithms to avoid the overfitting problem[8,9]. Diagnosis biomarkers are defined as a subset of detected features separating patients of a given disease from healthy control samples[10,11]. Patients are usually defined as the positive samples, and the healthy controls are defined as the negative samples[12].

Recent studies have suggested that there exists more than one solution with identical or similarly effective classification performances for a biomedical dataset[5]. Almost all the feature selection algorithms are deterministic algorithms, producing only one solution for the same dataset. Genetic algorithms may simultaneously generate multiple solutions with similar performances, but they still try to select one solution with the best fitness function as the output for a given dataset[13,14].

Feature selection algorithms can be roughly grouped as either filters or wrappers[12]. A filter algorithm chooses the top-*k* features ranked by their significant individual association with the binary class labels based on the assumption that features are independent of each other[15-17]. Although this assumption does not hold true for almost all real-world datasets, the heuristic filter rule performs well in many cases, for instance, the mRMR (Minimum Redundancy and Maximum Relevance) algorithm, the Wilcoxon test based feature filtering (WRank) algorithm, and the ROC (Receiver operating characteristic) plot based filtering (ROCRank) algorithm. mRMR, is an efficient filter algorithm because it approximates the combinatorial estimation problem with a series of much smaller problems, comparing to the maximum-dependency feature selection algorithm, each of which only involves two variables, and therefore uses pairwise joint probabilities which are more robust[18,19]. However, mRMR may underestimate the usefulness of some features as it does not measure the interactions between features which can increase relevancy, and thus misses some feature combinations that are individually useless but are useful only when combined. The WRank algorithm calculates a non-parametric score of how discriminative a feature is between two classes of samples, and is known for its robustness for outliers[20,21]. Furthermore, the ROCRank algorithm evaluates how significant the Area Under the ROC Curve (AUC) of a particular feature is for the investigated binary classification performance[22,23].

On the other hand, a wrapper evaluates the pre-defined classifier's performance of a given feature subset, iteratively generated by a heuristic rule, and creates the feature subset with the best performance measurement[24]. A wrapper generally outperforms a filter in the classification performance but runs slower[25]. For example, the Regularized Random Forest (RRF)[26,27] algorithm uses a greedy rule, by evaluating the features on a subset of the training data at each random forest node, whose feature importance scores are evaluated by the Gini index. The choice of a new feature will be penalized if its information gain does not improve that of the chosen features. Additionally, the Prediction Analysis for Microarrays (PAM)[28,29] algorithm, also a wrapper algorithm, calculates a centroid for each of the class labels, and then selects features to shrink the gene centroids toward the overall class centroid. PAM is robust for outlying features.

Multiple solutions with the top classification performance may be necessary for any given dataset. Firstly, the optimization goal of a deterministic algorithm is defined by a mathematical formula, *e.g.*, minimum error rate[30], which is not necessarily ideal for biological samples. Secondly, a dataset may have multiple, significantly different, solutions with similar effective or even identical performances. Almost all existing feature selection algorithms will randomly select one of these solutions as the output[31].

This study will introduce an informatics analytic protocol for generating multiple feature selection solutions with similar performances for any given binary classification dataset. Considering that most biomedical researchers are not familiar with informatic techniques or computer coding, a user-friendly graphical user interface (GUI) was developed to facilitate the rapid analysis of biomedical binary classification datasets. The analytic protocol consists of data loading and summarizing, parameter tuning, pipeline execution, and result interpretations. With a simple click, the researcher is able to generate the biomarker subsets and publication-quality visualization plots. The protocol has been tested using the transcriptomes of two binary classification datasets of Acute Lymphoblastic Leukemia (ALL), *i.e.*, ALL1 and ALL2[12]. The datasets of ALL1 and ALL2 were downloaded from the Broad Institute Genome Data Analysis Center, available at http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi. ALL1 contains 128 samples with 12,625 features. Of these samples, 95 are B-cell ALL and 33 are T-cell ALL. ALL2 includes 100 samples with 12,625 features as well. Of these samples, there are 65 patients that suffered relapse and 35 patients that did not. ALL1 was an easy binary classification dataset, with a minimum accuracy of four filters and four wrappers being 96.7%, and 6 of the 8 feature selection algorithms achieving 100%[12]. While ALL2 was a more difficult dataset, with the above 8 feature selection algorithms achieving no better than 83.7% accuracy[12]. This best accuracy was achieved with 56 features detected by the wrapper algorithm, Correlation-based Feature Selection (CFS).

PROTOCOL:

Note: The following protocol describes the details of the informatics analytic procedure and pseudo-codes of the major modules. The automatic analysis system was developed using Python version 3.6.0 and the Python modules pandas, abc, numpy, scipy, sklearn, sys, PyQt5, sys, mRMR, math and matplotlib. The materials used in this study are listed in the **Table of Materials**.

1. Prepare the Data Matrix and Class Labels

1.1. Prepare the data matrix file as a TAB- or comma-delimited matrix file, as illustrated in **Figure 1A**.

Note: Each row has all the values of a feature, and the first item is the feature name. A feature is a probeset ID for the microarray-based transcriptome dataset or may be another value ID like a cysteine residue with its methylation value in a methylomic dataset. Each column gives the feature values of a given sample, with the first item being the sample name. A row is separated into columns by a TAB (**Figure 1B**) or a comma (**Figure 1C**). A TAB-delimited matrix file is recognized by the file extension .tsv, and a comma-delimited matrix file has the extension .csv. This file may be generated by saving a matrix as either the .tsv or .csv format from software such as Microsoft Excel. The data matrix may also be generated by computer coding.

1.2. Prepare the class label file as a TAB- or comma-delimited matrix file (**Figure 1D**), similar to the data matrix file.

Note: The first column gives the sample names, and the class label of each sample is given in the column titled **Class**. Maximal compatibility is considered in the coding process, so that additional columns may be added. The class label file may be formatted as a .tsv or .csv file. The names in the column **Class** may be any terms, and there may be more than two classes of samples. The user may choose any two of the classes for the following analysis.

2. Load the Data Matrix and Class Labels

2.1. Load the data matrix and class labels into the software. Click the button **Load data matrix** to choose the user-specified data matrix file. Click the button **Load class labels** to choose the corresponding class label file.

Note: After both files are loaded, kSolutionVis will conduct a routine screen of the compatibility between the two files.

2.2. Summarize the features and samples from the data matrix file. Estimate the size of the data matrix file.

2.3. Summarize the samples and classes from the class label file. Estimate the size of the class label file.

2.4. Test whether each sample from the data matrix has a class label. Summarize the numbers of the samples with the class labels.

3. Summarize and Display the Baseline Statistics of the Dataset

3.1. Click the button **Summarize**, without any specified keyword input, and the software will display 20 indexed features and the corresponding features names.

Note: Users need to specify the feature name they wish to find to see its baseline statistics and corresponding value distribution among all input samples.

3.2. Provide a keyword, *e.g.* “1000_at”, in the textbox F**eature** to find a specific feature to be summarized. Click the button **Summarize** to get the baseline statistics for this given feature.

Note: The keyword may appear anywhere in the target feature names, facilitating the search process for users.

3.3. Click the button **Summarize** to find more than one feature with the given keyword, and then specify the unique feature ID to proceed with the above step of summarizing one particular feature.

4. Determine the Class Labels and the Number of Top-ranked Features

4.1. Choose the names of Positive (“P (33)”) and Negative (“N (95)”) classes in the dropdown boxes **Class Positive** and **Class Negative**, as shown in **Figure 2** (middle).

Note: It is suggested to choose a balanced binary classification dataset, *i.e.*, the difference between the numbers of positive and negative samples is minimal. The number of samples is also given in parenthesis after the name of each class label in the two dropdown boxes.

4.2. Choose 10 as the number of top-ranked features (parameter *pTopX*) in the dropdown box **Top_X (?)** for a comprehensive screen of the feature-subset.

Note: The software automatically ranks all the features by the *P-value* calculated by a t-test of each feature comparing the positive and negative classes. A feature with a smaller *P-value* has a better discriminating power between the two classes of samples. The comprehensive screening module is computationally intensive. The parameter *pTopX* is 10 by default. Users can change this parameter in the range of 10 to 50, until they find satisfying feature subsets with good classification performances.

5. Tune System Parameters for Different Performances

5.1. Choose the performance measurement (*pMeasurement*) Accuracy (*Acc*) in the dropdown box **Acc/bAcc (?)** for the selected classifier Extreme Learning Machine (ELM). Another option of this parameter is the measurement Balanced Accuracy (*bAcc*).

Note: Let TP, FN, TN, and FP be the numbers of true positives, false negatives, true negatives and false positives, respectively. The measurement *Acc* is defined as (TP+TN)/(TP+FN+TN+FP), which works best on a balanced dataset[6]. But a classifier optimized for *Acc* tends to assign all the samples to the negative class if the number of negative samples is much larger than that of positive ones. The *bAcc* is defined as (Sn+Sp)/2, where Sn = TP/(TP+FN) and Sp = TN/(TN+FP) are the correctly predicted rates for positive and negative samples, respectively. Therefore, bAcc normalizes the prediction performances over the two classes, and may lead to a balanced prediction performance over two unbalanced classes. *Acc* is the default choice of *pMeasurement*. The software uses the classifier ELM by default to calculate the classification performances. The user may also choose a classifier from SVM (Support Vector Machine), KNN (k Nearest Neighbor), Decision Tree, or Naïve Bayes.

5.2. Choose the cutoff value 0.70 (parameter *pCutoff*) for the specified performance measurement in the input box **pCutoff:**.

Note: Both *Acc* and *bAcc* range between 0 and 1, and the user may specify a value pCutoff∈[0, 1] as the cutoff to display the matched solutions. The software carries out a comprehensive feature-subset screening, and an appropriate choice of *pCutoff* will make the 3D visualization more intuitive and explicit. The default value for *pCutoff is* 0.70.

6. Run the Pipeline and Produce the INTERACTIVE VISUALIZED RESULTS

6.1. Click the button **Analyze** to run the pipeline and generate the visualization plots, as shown in **Figure 2** (bottom).

Note: The left table gives all the feature subsets and their *pMeasurement* calculated by the 10-fold cross validation strategy of the classifier ELM, as described previously[5]. Two 3D scatter plots and two-line plots are generated for the feature-subset screening procedure with the current parameter settings.

6.2. Choose 0.70 as the default value of the *pMeasurement* cutoff (parameter *piCutoff*, input box **Value**), and 10 as the default of the number of best feature subsets (parameter *piFSNum*).

Note: The pipeline is executed using the parameters *pTopX*, *pMeasurement,* and *pCutoff*. The detected feature subsets may be further screened using the cutoff *piCutoff*, however *piCutoff* cannot be smaller than *pCutoff*. Therefore, *piCutoff* is initialized as *pCutoff* and only the feature subsets with the performance measurement ≥ *piCutoff* will be visualized. The default value of *piCutoff* is *pCutoff*. Sometimes kSolutionVis detects many solutions, and only the best *piFSNum* (default: 10) feature subsets will be visualized. If the number of feature subsets detected by the software is smaller than *piFSNum*, all the feature subsets will be visualized.

6.3. Collect and interpret the features detected by the software, as shown in **Figure 3**.

Note: The table in the left box shows the detected feature subsets and their performance measurements. The names of the first three columns are "F1", "F2", and "F3". The three features in each feature subset are given in their ranking order in one row (F1 < F2 < F3). The last column gives the performance measurement (*Acc* or *bAcc*) of each feature subset, and its column name (*Acc* or *bAcc*) is the value of *pMeasurement*.

7. Interpret the 3D Scatter Plots-Visualize and Interpret the Feature Subsets with Similarly Effective Binary Classification Performances Using 3D Scatter Plots

7.1. Click the button **Analyze** to generate the 3D scatter plot of the top 10 feature subsets with the best classification performances (*Acc* or *bAcc*) detected by the software, as shown in **Figure 3** (middle box). Sort the three features in a feature subset in ascending order of their ranks and use the ranks of the three features as the F1/F2/F3 axes, *i.e.*, F1 < F2 < F3.

Note: The color of a dot represents the binary classification performance of the corresponding feature subset. A dataset may have multiple feature subsets with similarly effective performance measurements. Therefore, an interactive and simplified scatter plot is necessary.

7.2. Change the value to 0.70 in the input box **pCutoff:** and click the button **Analyze** to generate the 3D scatter plot of the feature subsets with the performance measurement ≥ *piCutoff*, as seen in **Figure 3** (right box). Click the button **3D tuning** to open a new window to manually tune the viewing angles of the 3D scatter plot.

Note: Each feature subset is represented by a dot in the same way as above. The 3D scatter plot was generated in the default angle. To facilitate the 3D visualization and tuning, a separate window will be opened by clicking the button **3D tuning**.

7.3. Click the button **Reduce** to reduce the redundancy of the detected feature subsets.

Note: If users wish to further select the feature triplets and minimize the redundancy of the feature subsets, the software also provides this function using the mRMR feature selection algorithm. After clicking the **Reduce** button, kSolutionVis will remove those redundant features in the feature triplets and regenerate the table and the two scatter plots mentioned above. The removed features of the feature triplets will be replaced by the key word in the table. The values of **None** in the F1/F2/F3 axis will be denoted as the value of *piFSNum* (the range of the normal value of F1/F2/F3 is [1, top_x]). Therefore, the dots that include a **None** value may appear to be "outlier" dots in the 3D plots. The manually tunable 3D plots may be found in "Manual tuning of the 3D dot plots" in the supplementary material.

8. Find Gene Annotations and Their Associations with Human Diseases

Note: Steps 8 to 10 will illustrate how to annotate a gene from the sequence level of both DNA and protein. Firstly, the gene symbol of each biomarker ID from the above steps will be retrieved from the database DAVID[32], and then two representative web servers will be used to analyze this gene symbol from the levels of DNA and protein, respectively. The server GeneCard provides a comprehensive functional annotation of a given gene symbol, and the Online Mendelian Inheritance in Man database (OMIM) provides the most comprehensive curation of disease-gene associations. The server UniProtKB is one of the most comprehensive protein database, and the server Group-based Prediction System (GPS) predicts the signaling phosphorylation's for a very large list of kinases.

8.1. Copy and paste the web link of the database DAVID into a web browser and open the web page of this database. Click the link **Gene ID Conversion** seen in **Figure 4A** and input the feature

IDs 38319_at/38147_at/33238_at of the first biomarker subset of the dataset ALL1 (**Figure 4B**). Click the link **Gene List** and click **Submit List** as shown in **Figure 4B**. Retrieve the annotations of interest and click **Show Gene List** (**Figure 4C**). Get the list of gene symbols (**Figure 4D**).

Note: The gene symbols retrieved here will be used for further functional annotations in the next steps.

8.2. Copy and paste the web link of the database Gene Cards into a web browser and open the web page of this database. Search a gene's name CD3D in the database query input box and find the annotations of this gene from Gene Cards[33,34], as shown in **Table 1** and **Figure 5A**.

Note: Gene Cards is a comprehensive gene knowledgebase, providing nomenclature, genomics, proteomics, subcellular localization, and involved pathways and other functional modules. It also provides external links to various other biomedical databases like PDB/PDB_REDO[35], Entrez Gene[36], OMIM[37], and UniProtKB[38]. If the feature name is not a standard gene symbol, use the database ENSEMBL to convert it[39]. CD3D is the name of the gene T-Cell Receptor T3 Delta Chain.

8.3. Copy and paste the web link of the database OMIM into a web browser and open the web page of this database. Search a gene's name CD3D and find the annotations of this gene from the database OMIM[37], as shown in **Table 1** and **Figure 5B**.

Note: OMIM serves now as one of the most comprehensive and authoritative sources of human gene connections with inheritable diseases. OMIM was initiated by Dr. Victor A. McKusick to catalog the disease-associated genetic mutations[40]. OMIM now covers over 15,000 human genes and over 8,500 phenotypes, as of December 1st 2017.

9. Annotate the Encoded Proteins and the Post-Translational Modifications

9.1. Copy and paste the web link of the database UniProtKB into a web browser and open the web page of this database. Search a gene's name CD3D in the query input box of UniProtKB and find the annotations of this gene from the database[38], as shown in **Table 1** and **Figure 5C**.

Note: UniProtKB collects a rich source of annotations for proteins, including both nomenclature and functional information. This database also provides external links to other widely used databases, including PDB/PDB_REDO[35], OMIM[37], and Pfam[41].

9.2. Copy and paste the web link of the web server GPS into a web browser and open the web page of this web server. Retrieve the protein sequence encoded by the biomarker gene CD3D from the UniProtKB database[38] and predict the protein's post-translational modification (PTM) residues using the online tool GPS, as shown in **Table 1** and **Figure 5D**.

Note: A biological system is dynamic and complicated, and the existing databases collect only known information. Therefore, biomedical prediction online tools as well as offline programs may provide useful evidence to complement a hypothesized mechanism. GPS has been developed and improved for over 12 years[7,42] and may be used to predict a protein's PTM residues in a given peptide sequence[43,44]. Tools are also available for various research topics, including the prediction of a protein's subcellular location[45] and transcription factor binding motifs [46] among others .

10. Annotate Protein-Protein Interactions and Their Enriched Functional Modules

10.1. Copy and paste the web link of the web server String into a web browser and open the web page of this web server. Search the list for the genes CD3D and P53, and find their orchestrated properties using the database String[47]. The same procedure may be carried out using another web server, DAVID[32].

Note: Besides the aforementioned annotations for individual genes, there are many large-scale informatics tools available to investigate the properties of a group of genes. A recent study demonstrated that individually bad marker genes might constitute a much-improved gene set[5]. Therefore, it's worth the computational cost to screen for more complicated biomarkers. The database String may visualize the known or predicted interaction connections, and the David server may detect the functional modules with significant phenotype-associations in the queried genes[47],[32]. Various other large-scale informatics analysis tools are also available.

11. Export the Generated Biomarker Subsets and the Visualization Plots

11.1. Export the detected biomarker subsets as a .tsv or .csv text file for further analysis. Click the button **Export the Table** under the table of all the detected biomarker subsets and choose which text format to save as.

11.2. Export the visualization plots as an image file. Click the button **Save** under each plot and choose which image format to save as.

Note: The software supports the pixel format .png and the vector format .svg. The pixel images are good for displaying on the computer screen, while the vector images may be converted to any resolution required for journal publication purposes.

REPRESENTATIVE RESULTS

The goal of this workflow (**Figure 6)** is to detect multiple biomarker subsets with similar efficiencies for a binary classification dataset. The whole process is illustrated by two example datasets ALL1 and ALL2 extracted from a recently-published biomarker detection study[12,48]. A user may install kSolutionVis by following the instructions in the supplementary materials.

Dataset ALL1 profiled 12 625 transcriptomic features of 95 B-cell and 33 T-cell ALL patient blood samples. While dataset ALL2 detected the expression levels of 12 625 transcriptomic features for 65 ALL patients who relapsed after the treatment and 35 ALL patients who did not. For the user's convenience, both transcriptomic datasets and their class labels are provided in version

1.4 of the software. Both datasets are in the subdirectory “data” of the software’s source code directory.

The two datasets, ALL1 and ALL2, were formatted as .csv files and loaded into the software using the **Load data matrix** and **Load class labels** buttons, as shown in **Figure 7A**-**B**. **Figure 7A** shows that all 128 samples with 12 625 features were loaded, and all 128 samples also have class labels. The final data matrix has 95 negative samples (B-cell ALL) and 33 positive samples (T-cell ALL). Additionally, users may also determine which class label is the positive class label (**Figure 7A,** bottom). If the class label file defines more than two classes, users may want to choose which two class labels to investigate. Similar operations were also conducted for the difficult dataset ALL2, as shown in **Figure 7B**.

The value distributions of the features in the data matrix may be investigated by clicking the button **Summarize** while searching for a user-specific keyword in the feature names, as shown in **Figure 8**. **Figure 8A** illustrates the histogram of feature 1012_at in the dataset ALL1. Furthermore, as seen in **Figure 8B**, the same feature 1012_at has a similar distribution of expression in both datasets. If no keyword was specified by the user, some feature names would be listed to help the users to decide which features to summarize.

The easier dataset ALL1 screened the top 10 ranked features (*pTopX*) for biomarker subsets with the *pMeasurement Acc* ≥ 0.90 (*pCutoff*). After clicking the button **Run**, the algorithm was executed, and the results as seen in **Figure 9A**, were illustrated in the bottom part of the software after a few seconds. From this, 120 qualified biomarker subsets were detected and listed in the left table of **Figure 9A**. ALL1 was an easy-to-discriminate dataset, in that it has 57 triplet biomarker subsets with 100% in *Acc*. This protocol emphasizes the existence of multiple similarly effective solutions for a binary classification problem. Therefore, the first 3D scatter plot may illustrate more than 10 (parameter *piFSNum*) biomarker subsets, if they have the classification performance *Acc* (parameter *pMeasurement*) ≥ that of the top 10 ranked (parameter *piFSNum*) biomarker subset. The user may also choose to display fewer biomarker subsets by changing the parameter *piCutoff* in the parameter box above the table in **Figure 9A**. The manual tuning of the 3D plots may be found in the section **Manual tuning of the 3D dot plots** in the supplementary material.

Furthermore, all the results may be exported as external files for further analysis by clicking the button **Export the Table** under the table or scatter plots, as shown in **Figure 9**.

The first biomarker subset (38319_at, 38147_at, and 33238_at) for the dataset ALL1 was chosen for functional investigations, as shown in **Figure 9A**. The search module of ENSEMBL (http://useast.ensembl.org/Multi/Search/New?db=core) annotated these three features as a gene cluster of differentiation 3 delta (CD3D, 38319_at), Signaling Lymphocytic Activation Molecule-Associated gene (SH2D1A, 38147_at) and Lymphocyte Cell-Specific Protein-Tyrosine Kinase (LCK, 33238_at). Furthermore, the gene-disease association database OMIM[37,40] suggested that the gene CD3D encodes the delta subunit of the T-cell antigen receptor complex and is involved in the 11q23 translocations frequently observed in acute leukemia in humans[49,50]. OMIM also suggested that genomic mutations within the gene SH2D1A in the chromosome region of Xq25 may be associated with B-cell leukemia[51,52]. Additionally, OMIM also highlighted a possible T-cell ALL associated fusion event of the LCK and beta T-cell receptor (TCRB)[53]. Users may investigate other functional aspects of these biomarkers with their gene symbols, *e.g.*, gene function annotations in Entrez Gene[36], protein function annotations in UniProtKB[38] or Pfam[41], 3D protein structures in PDB/PDB_REDO[35], and PTM residues in GPS[7,42-44]. The interacting sub-network (database string[47]) and enriched functional modules (database David[32]) may also be screened for these biomarkers as an entirety. Various other databases or web servers may also facilitate the annotations and *in silico* predictions using the symbols or primary gene/protein sequences of these genes.

As seen in **Table 2**, the necessity of detecting more than one solution with identical or similarly effective performances is evident, with 57 groups of features with binary classification accuracies of 100% between B-cell and T-cell ALL samples. These particular biomarker subsets were called the perfect solutions. Quite a few biomarkers appeared in these perfect solutions repeatedly, suggesting that they may represent the key differences, on the molecular level, between B- and T-cell ALL. If the biomarker detection algorithm stops at detecting the first perfect solution of three genes CD3D/SH2D1A/LCK, another perfect solution CD74/HLA-DPB1/PRKCQ will be missed. For example, HLA-DPB1 is known to be significantly associated with the pediatric T-cell ALL but not B-cell ALL[54].

The three features of the first biomarker subset of ALL2 were chromatin assembly factor 1 subunit B (CHAF1B, 36912_at), exonuclease 1 (EXO1, 36041_at), and signal transducer and activator of transcription 6 (STAT6, 41222_at). CHAF1B was observed to be highly expressed in leukemia cell lines and the antibody against the CHAF1B encoded protein was significantly developed in acute myeloid leukemia (AML) patients[55]. EXO1 was lost in some cases of acute leukemia[56], and upregulated in the leukemia cell line HL-60[R]. It also has been found to negatively regulate the alternative lengthening of telomeres (ALT) pathway, which facilitated the formation of ALT-associated PML (promyelocytic leukemia) bodies (APBs)[57]. STAT6 was phosphorylated to activate the pro-survival and proliferative signaling pathway in the cases of relapsed AML[58]. Taken together, the three genes were associated with the development and relapse of leukemia, but no explicit evidence was published on their associations with the ALL relapse. This may represent an interesting topic for further investigation.

The same annotation procedure may be conducted on any biomarker subset for ALL1 and ALL2. The three biomarkers investigated in the above section were not identified as relapse biomarkers in the dataset ALL2, as shown in **Figure 9B**. This suggests that biomarkers are phenotype-specific, which is another major challenge for biomarker detection, alongside the existence of multiple similarly effective solutions.

Some technical modules were implemented and described here for interested users. The error handling module provides informative messages for the user when errors occur during the execution of the software. The main error messages are listed and explained in “Error messages” in the supplementary material. A parallel calculation of the biomarkers was implemented for computers with more than one CPU core. The detailed improvements to the running time may be found in "Parallel running time" in the supplementary material. The data suggests that the usage of more CPU cores may not improve the running time due to the cost of switching between different CPU cores.

FIGURE & TABLE LEGENDS:

Figure 1. The example dataset extracted from the transcriptome dataset ALL1 has the first six features of the first nine samples of ALL1. The data matrix was formatted in (a) the visualization form, (b) the TAB-delimited text format file, and (c) the comma-delimited text format file. (d) The class label data was formatted in the visualization form. Due to the TAB character is invisible, it is illustrated as **[TAB]** in (b). The column **Platform** gives the microarray platform **Affy** in (b), and is not a required data column.

Figure 2. Graphical user interface of the software. The baseline statistics are summarized in the top left box. Users may search for features of interest and investigate the value distributions in the two top right boxes. All the parameters for biomarker detection procedure may be tuned in the middle horizontal bar. All the biomarker subsets and their corresponding visualized distributions may be found in the bottom part.

Figure 3. Biomarker subsets and their visualizations generated. Users may further refine the table and two 3D scatter plots using the parameters *piCutoff* and *piFSNum*.

Figure 4. Gene annotations of the feature IDs detected in this study. Take the three feature IDs 38319_at/38147_at/33238_at of the first biomarker subset of the dataset ALL1. (a) Get the ID conversion module by clicking the link **Gene ID Conversion**. (b) Input the feature IDs in the red box 1, choose the feature type in the red box 2 (default "AFFYMETRIX_3PRIME_IVT_ID" is correct for this study), choose **Gene List** in the red box 3, and click **Submit List** in the red box 4. (c) Get all the functional annotations in this page and click **Show Gene List** to get the gene symbols of these queried features. (d) Get the gene symbols of the queried feature IDs.

Figure 5. Annotations and enrichment analysis of the detected feature subsets. (a) Gene annotations from Gene Card. (b) OMIM describes the disease associations of each feature/gene. (c) Annotate the protein encoded by the gene of interest in the database UniProtKB. (d) Predict the tyrosine phosphorylation residues in the given protein using the online tool GPS. A red box was added to show the user where to click to input the query data. The primary sequence of the example protein CD3D may be retrieved as the FASTA format from the red box in (c), and input in the query window by click the red box in (d).

Figure 6. Workflow of kSolutionVis. Each module of the software was described in the above protocol.

Figure 7. Baseline statistics of the two representative datasets. The numbers of samples, features and classes in (a) ALL1 and (b) ALL2 are calculated. The file sizes of the data matrix and class labels are also detected. And a new data matrix is extracted from the samples with class

labels.

Figure 8. Histogram visualization of the feature 1012_at in the two datasets. Both baseline statistics and histogram were generated for (a) ALL1 and (b) ALL2.

Figure 9. Biomarker subsets and the scatter plots of the two datasets. Users may change the parameters in the second row of parameter boxes to further refine the lists of biomarker subsets and 3D scatter plots for the datasets (a) ALL1 and (b) ALL2.

Table 1. Websites for annotating and analyzing the detected biomarkers. A list of useful online tools that help annotate the detected biomarkers.

Table 2. Annotations of all the features from the dataset ALL1. This is a binary classification dataset between B-cell and T-cell ALL samples. The gene symbols were collected for all the microarray features in the last three columns.

DISCUSSION

This study presents an easy-to-follow multi-solution biomarker detection and characterization protocol for a user-specified binary classification dataset. The software puts an emphasis on user-friendliness and flexible import/export interfaces for various file formats, allowing a biomedical researcher to investigate their dataset easily using the GUI of the software. This study also highlights the necessity of generating more than one solution with similarly effective modeling performances, previously ignored by many existing biomarker detection algorithms. In the future, newly developed biomarker detection algorithms may include this option by recording all the intermediate biomarker subsets with sufficient modeling performances.

In this protocol, steps 1 and 5 are of most importance, as the software is a fully automatic system that relies on correctly formatted input files. It was found that during our testing step, the mis-match of sample names from data matrix and class labels files may cause errors in the software, where the software will pop out a warning dialog about this error. Therefore, if the user finds no samples were loaded from the data matrix or class label files, the troubleshooting trick is to double-check whether the sample names in the two input files are inconsistent. If no dots were visualized in the 3D scatter plots, this may be due to the parameter *pCutoff* being higher than the best solution. In this instance, the troubleshooting trick is to lower the cutoff of the classification performance measurement (parameter *pCutoff*). However, the maximum performance measurement achieved by the biomarker subsets may be still blocked by the cutoff for a difficult dataset. A warning dialog will give this best performance measurement, and the user may choose a smaller cutoff to continue further analysis.

The main limitations of the software are its slow calculation speed and its ability to only focus on, at most, three features. Feature selection is an NP-hard problem, defined as a computational problem whose globally optimal solution cannot be resolved within polynomial time[59]. The comprehensive biomarker subset screening step consumes a high volume of computational power. The running time complexity of kSolutionVis is O(n^3) where n is the parameter *pTopX*. Additionally, this multiple-biomarker detection algorithm focuses on visualizing the screen of features, therefore confining the number of the features to three or fewer. This limitation may impede some users who may work on difficult problems and wish to find feature subsets consisting of more than three features. However, the software visualizes feature subsets in the 3D space and it's difficult to directly visualize feature subsets in more than three dimensions. In addition, based on the representative results presented above, the multiple feature triplets selected by kSolutionVis is a highly effective method in classification and shows significant results with important biomedical meaning.

The software represents useful complementary software to the existing feature selection algorithms. In the field of biomedicine, feature selection is termed biomarker, with the goal to find a subset of features achieving improved modeling performance[60-62]. The software is a comprehensive screening tool of all the triplet biomarker subsets based on the strategy proposed in a recent study[5]. The two representative datasets screened by the software's protocol, and their results demonstrate the existences of quite a few solutions with similarly effective or even identical modeling performances. However, heuristic rules[63-66] may be employed to find sub-optimal solutions, but such algorithms have a strong tendency to produce only one solution, ignoring many other solutions with similarly effective or even identical modeling performances. Therefore, the computer power and the lengthy running time of the software are worthwhile to ensure a more comprehensive detection of potential biomarkers in the future.

The representative results were calculated on two transcriptome datasets, however, the software handles input data in various standard file formats and may also be used to analyze other 'omic' datasets, including proteomics and metabolomics. Additionally, parallelization may speed up the calculation of the biomarker detection module in the software. There is some multi-core hardware including GPGPU (General-Purpose Graphical Processing Unite) and Intel Xeon Phi processors available for this purpose. However, these technologies require different coding strategies and will be considered in the next version of the software.

ACKNOWLEDGEMENTS

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040400) and the startup grant from Jilin University. Anonymous reviewers and biomedical testing users were appreciated for their constructive comments on improving the usability and functionality of kSolutionVis.

DISCLOSURES

We have no conflicts of interest related to this report.

REFERENCES

1 Heckerman, D. *et al.* Genetic variants associated with physical performance and anthropometry in old age: a genome-wide association study in the ilSIRENTE cohort. *Scientific Reports* **7**, 15879, doi:10.1038/s41598-017-13475-0 (2017).

2 Li, Z. *et al.* Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nature Genetics* **49**, 1576-1583, doi:10.1038/ng.3973 (2017).

3 Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols* **9**, 1192-1212, doi:10.1038/nprot.2014.071 (2014).

4 Harrison, R. N. S. *et al.* Development of multivariable models to predict change in Body Mass Index within a clinical trial population of psychotic individuals. *Scientific Reports* **7**, 14738, doi:10.1038/s41598-017-15137-7 (2017).

5 Liu, J. *et al.* Multiple similarly-well solutions exist for biomedical feature selection and classification problems. *Scientific Reports* **7**, 12830, doi:10.1038/s41598-017-13184-8 (2017).

6 Ye, Y., Zhang, R., Zheng, W., Liu, S. & Zhou, F. RIFS: a randomly restarted incremental feature selection algorithm. *Scientific Reports* **7**, 13013, doi:10.1038/s41598-017-13259-6 (2017).

7 Zhou, F. F., Xue, Y., Chen, G. L. & Yao, X. GPS: a novel group-based phosphorylation predicting and scoring method. *Biochemical and Biophysical Research Communications* **325**, 1443-1448, doi:10.1016/j.bbrc.2004.11.001 (2004).

8 Sanchez, B. N., Wu, M., Song, P. X. & Wang, W. Study design in high-dimensional classification analysis. *Biostatistics* **17**, 722-736, doi:10.1093/biostatistics/kxw018 (2016).

9 Shujie, M. A., Carroll, R. J., Liang, H. & Xu, S. Estimation and Inference in Generalized Additive Coefficient Models for Nonlinear Interactions with High-Dimensional Covariates. *Annals of Statistics* **43**, 2102-2131, doi:10.1214/15-AOS1344 (2015).

10 Li, J. H. *et al.* MiR-205 as a promising biomarker in the diagnosis and prognosis of lung cancer. *Oncotarget* **8**, 91938-91949, doi:10.18632/oncotarget.20262 (2017).

11 Lyskjaer, I., Rasmussen, M. H. & Andersen, C. L. Putting a brake on stress signaling: miR-625-3p as a biomarker for choice of therapy in colorectal cancer. *Epigenomics* **8**, 1449-1452, doi:10.2217/epi-2016-0128 (2016).

12 Ge, R. *et al.* McTwo: a two-step feature selection algorithm based on maximal information coefficient. *BMC Bioinformatics* **17**, 142, doi:10.1186/s12859-016-0990-0 (2016).

13 Tumuluru, J. S. & McCulloch, R. Application of Hybrid Genetic Algorithm Routine in Optimizing Food and Bioengineering Processes. *Foods* **5**, doi:10.3390/foods5040076 (2016).

14 Gen, M., Cheng, R. & Lin, L. *Network models and optimization: Multiobjective genetic algorithm approach*. (Springer Science & Business Media, 2008).

15 Radovic, M., Ghalwash, M., Filipovic, N. & Obradovic, Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics* **18**, 9, doi:10.1186/s12859-016-1423-9 (2017).

16 Ciuculete, D. M. *et al.* A methylome-wide mQTL analysis reveals associations of methylation sites with GAD1 and HDAC3 SNPs and a general psychiatric risk score. *Translational Psychiatry* **7**, e1002, doi:10.1038/tp.2016.275 (2017).

17 Lin, H. *et al.* Methylome-wide Association Study of Atrial Fibrillation in Framingham Heart Study. *Scientific Reports* **7**, 40377, doi:10.1038/srep40377 (2017).

18 Wang, S., Li, J., Yuan, F., Huang, T. & Cai, Y. D. Computational method for distinguishing lysine acetylation, sumoylation, and ubiquitination using the random forest algorithm with a feature selection procedure. *combinatorial chemistry & high throughput screening*, doi:10.2174/1386207321666171218114056 (2017).

19 Zhang, Q. *et al.* Predicting Citrullination Sites in Protein Sequences Using mRMR Method and Random Forest Algorithm. *combinatorial chemistry & high throughput screening* **20**, 164-173, doi:10.2174/1386207319666161227124350 (2017).

20 Cuena-Lombrana, A., Fois, M., Fenu, G., Cogoni, D. & Bacchetta, G. The impact of climatic variations on the reproductive success of Gentiana lutea L. in a Mediterranean mountain area. *International journal of biometeorology* doi:10.1007/s00484-018-1533-3 (2018).

21 Coghe, G. *et al.* Fatigue, as measured using the Modified Fatigue Impact Scale, is a predictor of processing speed improvement induced by exercise in patients with multiple sclerosis: data from a randomized controlled trial. *Journal of Neurology,* doi:10.1007/s00415-018-8836-5 (2018).

22 Hong, H. *et al.* Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. The case of Dayu County, China. *Science of the Total Environment* **630**, 1044-1056, doi:10.1016/j.scitotenv.2018.02.278 (2018).

23 Borges, D. L. *et al.* Photoanthropometric face iridial proportions for age estimation: An investigation using features selected via a joint mutual information criterion. *Forensic Science International* **284**, 9-14, doi:10.1016/j.forsciint.2017.12.011 (2018).

24 Kohavi, R. & John, G. H. Wrappers for feature subset selection. *Artificial intelligence* **97**, 273-324 (1997).

25 Yu, L. & Liu, H. Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research* **5**, 1205-1224 (2004).

26 Wexler, R. B., Martirez, J. M. P. & Rappe, A. M. Chemical Pressure-Driven Enhancement of the Hydrogen Evolving Activity of Ni2P from Nonmetal Surface Doping Interpreted via Machine Learning. *Journal of American Chemical Society,* doi:10.1021/jacs.8b00947 (2018).

27 Wijaya, S. H., Batubara, I., Nishioka, T., Altaf-Ul-Amin, M. & Kanaya, S. Metabolomic Studies of Indonesian Jamu Medicines: Prediction of Jamu Efficacy and Identification of Important Metabolites. *Molecular Informatics* **36**, doi:10.1002/minf.201700050 (2017).

28 Shangkuan, W. C. *et al.* Risk analysis of colorectal cancer incidence by gene expression analysis. *PeerJ* **5**, e3003, doi:10.7717/peerj.3003 (2017).

29 Chu, C. M. *et al.* Gene expression profiling of colorectal tumors and normal mucosa by microarrays meta-analysis using prediction analysis of microarray, artificial neural network, classification, and regression trees. *Disease Markers* **2014**, 634123, doi:10.1155/2014/634123 (2014).

30 Fleuret, F. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research* **5**, 1531-1555 (2004).

31 Pacheco, J., Alfaro, E., Casado, S., Gámez, M. & García, N. A GRASP method for building classification trees. *Expert Systems with Applications* **39**, 3241-3248 (2012).

32 Jiao, X. *et al.* DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* **28**, 1805-1806, doi:10.1093/bioinformatics/bts251 (2012).

33 Rappaport, N. *et al.* Rational confederation of genes and diseases: NGS interpretation via GeneCards, MalaCards and VarElect. *Biomedical Engineering OnLine* **16**, 72, doi:10.1186/s12938-017-0359-2 (2017).

34 Rebhan, M., Chalifa-Caspi, V., Prilusky, J. & Lancet, D. GeneCards: integrating information about genes, proteins and diseases. *Trends in Genet* **13**, 163 (1997).

35 Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **1**, 213-220, doi:10.1107/S2052252514009324 (2014).

36 Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* **39**, D52-57, doi:10.1093/nar/gkq1237 (2011).

37 Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* **43**, D789-798, doi:10.1093/nar/gku1205 (2015).

38 Boutet, E. *et al.* UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods in Molecular Biology* **1374**, 23-54, doi:10.1007/978-1-4939-3167-5_2 (2016).

39 Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res*, doi:10.1093/nar/gkx1098 (2017).

40 McKusick, V. A. & Amberger, J. S. The morbid anatomy of the human genome: chromosomal location of mutations causing disease. *Journal of Medical Genetics* **30**, 1-26 (1993).

41 Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44**, D279-285, doi:10.1093/nar/gkv1344 (2016).

42 Xue, Y. *et al.* GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Research* **33**, W184-187, doi:10.1093/nar/gki393 (2005).

43 Deng, W. *et al.* GPS-PAIL: prediction of lysine acetyltransferase-specific modification sites from protein sequences. *Scientific Reports* **6**, 39787, doi:10.1038/srep39787 (2016).

44 Zhao, Q. *et al.* GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic Acids Research* **42**, W325-330, doi:10.1093/nar/gku383 (2014).

45 Wan, S., Duan, Y. & Zou, Q. HPSLPred: An Ensemble Multi-Label Classifier for Human Protein Subcellular Location Prediction with Imbalanced Source. *Proteomics* **17**, doi:10.1002/pmic.201700262 (2017).

46 Zhang, H., Zhu, L. & Huang, D. S. WSMD: weakly-supervised motif discovery in transcription factor ChIP-seq data. *Scientific Reports* **7**, 3217, doi:10.1038/s41598-017-03554-7 (2017).

47 Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* **43**, D447-452, doi:10.1093/nar/gku1003 (2015).

48 Chiaretti, S. *et al.* Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood* **103**, 2771-2778, doi:10.1182/blood-2003-09-3243 (2004).

49 Rowley, J. D. *et al.* Mapping chromosome band 11q23 in human acute leukemia with biotinylated probes: identification of 11q23 translocation breakpoints with a yeast artificial chromosome. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 9358-9362 (1990).

50 Rabbitts, T. H. *et al.* The chromosomal location of T-cell receptor genes and a T cell rearranging gene: possible correlation with specific translocations in human T cell leukaemia. *Embo Journal* **4**, 1461-1465 (1985).

51 Yin, L. *et al.* SH2D1A mutation analysis for diagnosis of XLP in typical and atypical patients. *Human Genetics* **105**, 501-505 (1999).

52 Brandau, O. *et al.* Epstein-Barr virus-negative boys with non-Hodgkin lymphoma are mutated in the SH2D1A gene, as are patients with X-linked lymphoproliferative disease (XLP). *Human Molecular Genetics* **8**, 2407-2413 (1999).

53 Burnett, R. C., Thirman, M. J., Rowley, J. D. & Diaz, M. O. Molecular analysis of the T-cell acute lymphoblastic leukemia-associated t(1;7)(p34;q34) that fuses LCK and TCRB. *Blood* **84**, 1232-1236 (1994).

54 Taylor, G. M. *et al.* Genetic susceptibility to childhood common acute lymphoblastic leukaemia is associated with polymorphic peptide-binding pocket profiles in HLA-DPB1*0201. *Human Molecular Genetics* **11**, 1585-1597 (2002).

55 Wadia, P. P. *et al.* Antibodies specifically target AML antigen NuSAP1 after allogeneic bone marrow transplantation. *Blood* **115**, 2077-2087, doi:10.1182/blood-2009-03-211375 (2010).

56 Wilson, D. M., 3rd *et al.* Hex1: a new human Rad2 nuclease family member with homology to yeast exonuclease 1. *Nucleic Acids Research* **26**, 3762-3768 (1998).

57 O'Sullivan, R. J. *et al.* Rapid induction of alternative lengthening of telomeres by depletion of the histone chaperone ASF1. *Nature Structural & Molecular Biology* **21**, 167-174, doi:10.1038/nsmb.2754 (2014).

58 Lee-Sherick, A. B. *et al.* Aberrant Mer receptor tyrosine kinase expression contributes to leukemogenesis in acute myeloid leukemia. *Oncogene* **32**, 5359-5368, doi:10.1038/onc.2013.40 (2013).

59 Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *Journal of machine learning research* **3**, 1157-1182 (2003).

60 John, G. H., Kohavi, R. & Pfleger, K. in *Machine learning: proceedings of the eleventh international conference.* 121-129.

61 Jain, A. & Zongker, D. Feature selection: Evaluation, application, and small sample performance. *IEEE transactions on pattern analysis and machine intelligence* **19**, 153-158 (1997).

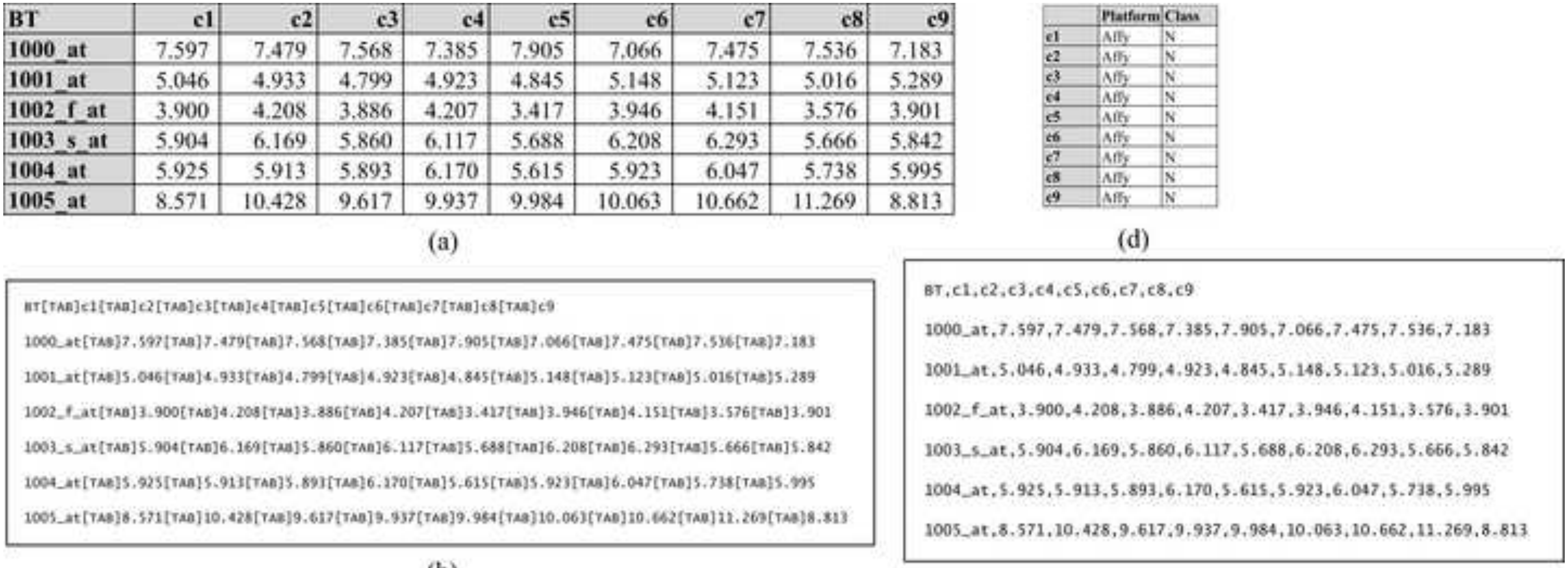
62 Taylor, S. L. & Kim, K. A jackknife and voting classifier approach to feature selection and classification. *Cancer Informatics* **10**, 133-147, doi:10.4137/CIN.S7111 (2011).

63 Andresen, K. *et al.* Novel target genes and a valid biomarker panel identified for cholangiocarcinoma. *Epigenetics* **7**, 1249-1257, doi:10.4161/epi.22191 (2012).

64 Guo, P. *et al.* Gene expression profile based classification models of psoriasis. *Genomics* **103**, 48-55, doi:10.1016/j.ygeno.2013.11.001 (2014).

65 Xie, J. & Wang, C. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. *Expert Systems with Applications* **38**, 5809-5815 (2011).

66 Zou, Q., Zeng, J., Cao, L. & Ji, R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* **173**, 346-354 (2016).



BT	c1	c2	c3	c4	c5	c6	c7	c8	c9
1000_at	7.597	7.479	7.568	7.385	7.905	7.066	7.475	7.536	7.183
1001_at	5.046	4.933	4.799	4.923	4.845	5.148	5.123	5.016	5.289
1002_f_at	3.900	4.208	3.886	4.207	3.417	3.946	4.151	3.576	3.901
1003_s_at	5.904	6.169	5.860	6.117	5.688	6.208	6.293	5.666	5.842
1004_at	5.925	5.913	5.893	6.170	5.615	5.923	6.047	5.738	5.995
1005_at	8.571	10.428	9.617	9.937	9.984	10.063	10.662	11.269	8.813

(a)

	Platform	Class
c1	Affy	N
c2	Affy	N
c3	Affy	N
c4	Affy	N
c5	Affy	N
c6	Affy	N
c7	Affy	N
c8	Affy	N
c9	Affy	N

(d)

```
BT[TAB]c1[TAB]c2[TAB]c3[TAB]c4[TAB]c5[TAB]c6[TAB]c7[TAB]c8[TAB]c9

1000_at[TAB]7.597[TAB]7.479[TAB]7.568[TAB]7.385[TAB]7.905[TAB]7.066[TAB]7.475[TAB]7.536[TAB]7.183

1001_at[TAB]5.046[TAB]4.933[TAB]4.799[TAB]4.923[TAB]4.845[TAB]5.148[TAB]5.123[TAB]5.016[TAB]5.289

1002_f_at[TAB]3.900[TAB]4.208[TAB]3.886[TAB]4.207[TAB]3.417[TAB]3.946[TAB]4.151[TAB]3.576[TAB]3.901

1003_s_at[TAB]5.904[TAB]6.169[TAB]5.860[TAB]6.117[TAB]5.688[TAB]6.208[TAB]6.293[TAB]5.666[TAB]5.842

1004_at[TAB]5.925[TAB]5.913[TAB]5.893[TAB]6.170[TAB]5.615[TAB]5.923[TAB]6.047[TAB]5.738[TAB]5.995

1005_at[TAB]8.571[TAB]10.428[TAB]9.617[TAB]9.937[TAB]9.984[TAB]10.063[TAB]10.662[TAB]11.269[TAB]8.813
```

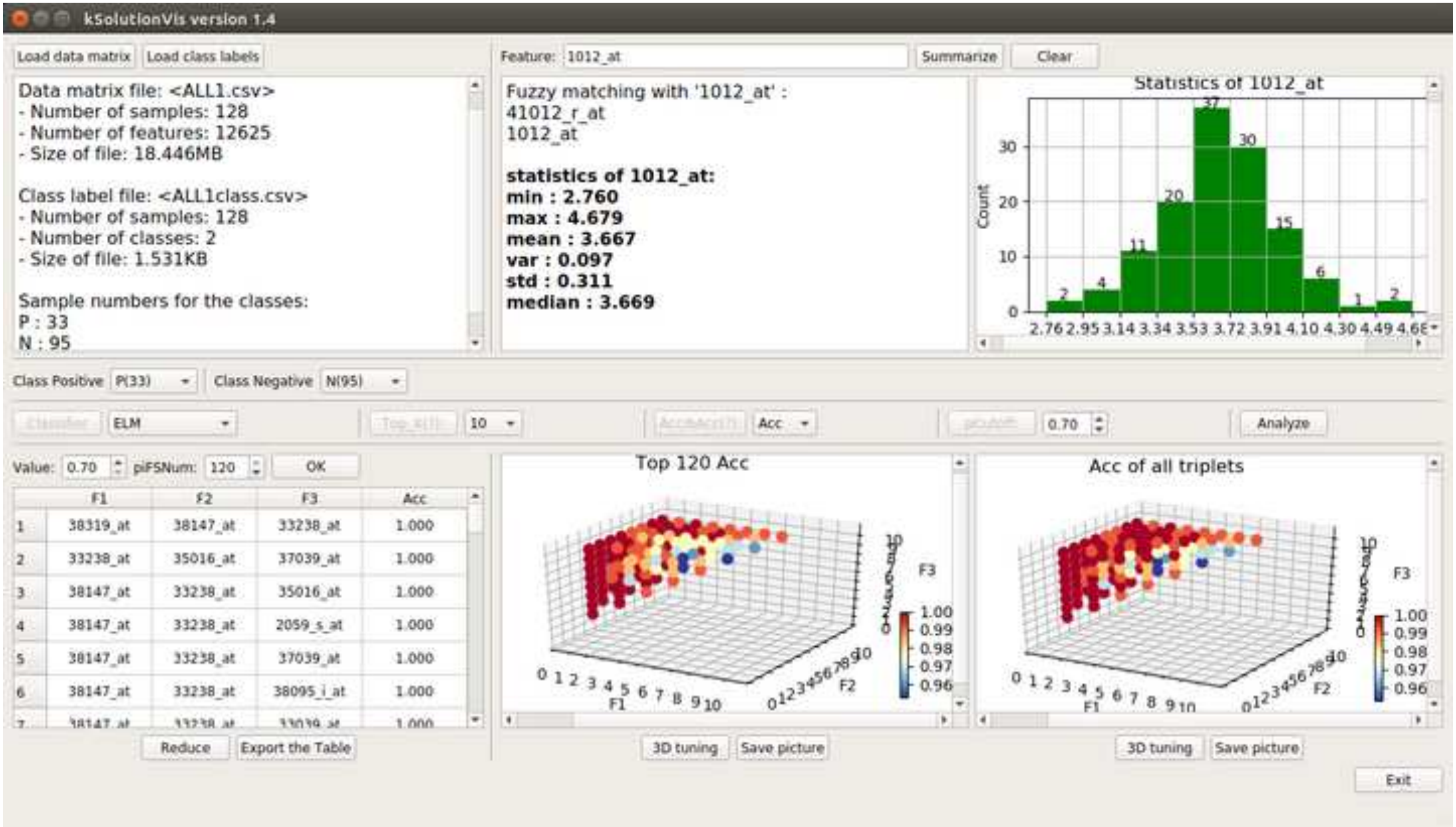
(b)

```
BT,c1,c2,c3,c4,c5,c6,c7,c8,c9

1000_at,7.597,7.479,7.568,7.385,7.905,7.066,7.475,7.536,7.183

1001_at,5.046,4.933,4.799,4.923,4.845,5.148,5.123,5.016,5.289

1002_f_at,3.900,4.208,3.886,4.207,3.417,3.946,4.151,3.576,3.901

1003_s_at,5.904,6.169,5.860,6.117,5.688,6.208,6.293,5.666,5.842

1004_at,5.925,5.913,5.893,6.170,5.615,5.923,6.047,5.738,5.995

1005_at,8.571,10.428,9.617,9.937,9.984,10.063,10.662,11.269,8.813
```

(c)

Figure 2

Click here to download Figure Figure2.jpg



Data matrix file: <ALL1.csv>
- Number of samples: 128
- Number of features: 12625
- Size of file: 18.446MB

Class label file: <ALL1class.csv>
- Number of samples: 128
- Number of classes: 2
- Size of file: 1.531KB

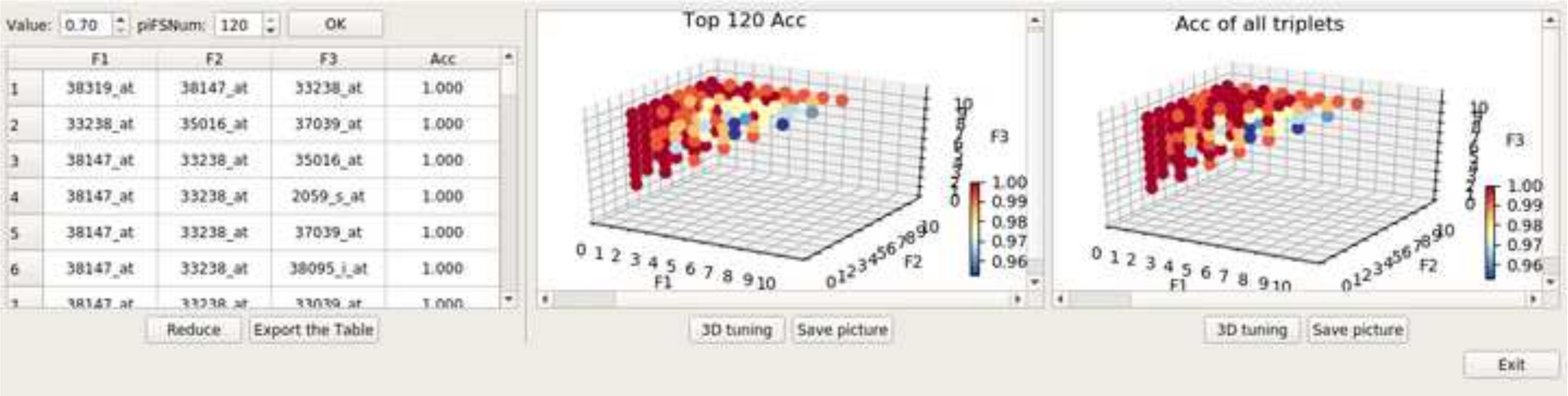
Sample numbers for the classes:
P : 33
N : 95

Fuzzy matching with '1012_at' :
41012_r_at
1012_at

statistics of 1012_at:
min : 2.760
max : 4.679
mean : 3.667
var : 0.097
std : 0.311
median : 3.669

	F1	F2	F3	Acc
1	38319_at	38147_at	33238_at	1.000
2	33238_at	35016_at	37039_at	1.000
3	38147_at	33238_at	35016_at	1.000
4	38147_at	33238_at	2059_s_at	1.000
5	38147_at	33238_at	37039_at	1.000
6	38147_at	33238_at	38095_i_at	1.000
7	38147_at	33238_at	33039_at	1.000

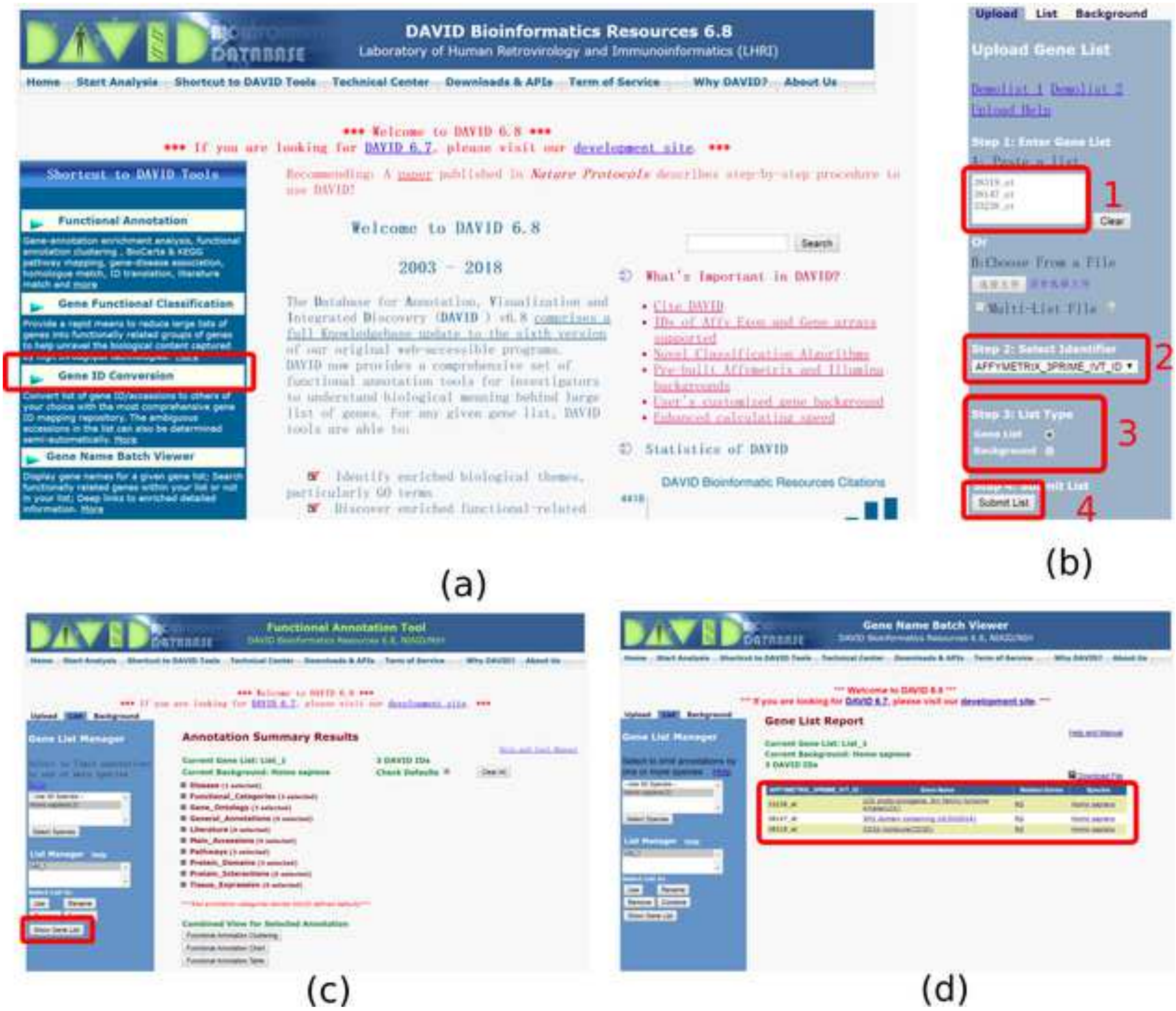
Figure 3



Value: 0.70 piFSNum: 120

	F1	F2	F3	Acc
1	38319_at	38147_at	33238_at	1.000
2	33238_at	35016_at	37039_at	1.000
3	38147_at	33238_at	35016_at	1.000
4	38147_at	33238_at	2059_s_at	1.000
5	38147_at	33238_at	37039_at	1.000
6	38147_at	33238_at	38095_i_at	1.000
7	38147_at	33238_at	33039_at	1.000

Figure 4



DAVID Bioinformatics Resources 6.8
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

Home | Start Analysis | Shortcut to DAVID Tools | Technical Center | Downloads & APIs | Term of Service | Why DAVID? | About Us

*** Welcome to DAVID 6.8 ***
*** If you are looking for DAVID 6.7, please visit our development site. ***

Shortcut to DAVID Tools

Functional Annotation
Gene-annotation enrichment analysis, functional annotation clustering, BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification
Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion
Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer
Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

Recommending: A paper published in *Nature Protocols* describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.8

2003 - 2018

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 comprises a full Knowledgebase update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related

Search

What's Important in DAVID?

- Cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

Statistics of DAVID

DAVID Bioinformatic Resources Citations

Upload | List | Background

Upload Gene List

Demolist 1 Demolist 2
Upload Help

Step 1: Enter Gene List
A: Paste a list

```
38319_at
38147_at
33238_at
```

Clear

Or
B:Choose From a File

□ Multi-List File

Step 2: Select Identifier
AFFYMETRIX_3PRIME_IVT_ID

Step 3: List Type
Gene List ●
Background ○

Step 4: Submit List
Submit List

Functional Annotation Tool

Annotation Summary Results

Current Gene List: List_1 | 3 DAVID IDs
Current Background: Homo sapiens | Check Defaults

- Disease
- Functional_Categories
- Gene_Ontology
- General_Annotations
- Literature
- Main_Accessions
- Pathways
- Protein_Domains
- Protein_Interactions
- Tissue_Expression

Combined View for Selected Annotation
Functional Annotation Clustering
Functional Annotation Chart
Functional Annotation Table

Gene List Manager
List Manager
Show Gene List

Gene Name Batch Viewer

*** Welcome to DAVID 6.8 ***
*** If you are looking for DAVID 6.7, please visit our development site. ***

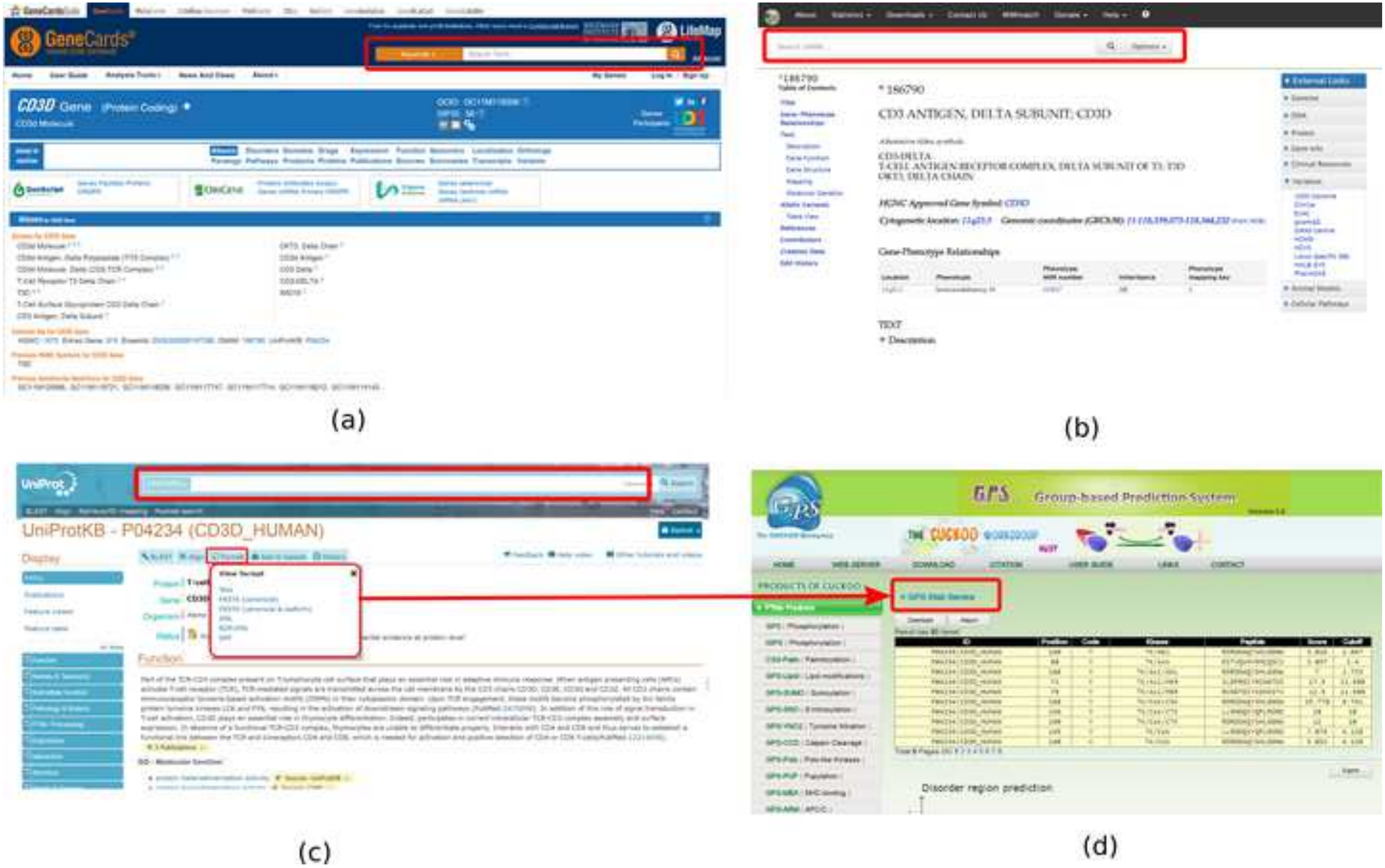
Gene List Report

Current Gene List: List_1
Current Background: Homo sapiens
3 DAVID IDs

AFFYMETRIX_3PRIME_IVT_ID	Gene Name	Related Genes	Species
33238_at	[illegible]	RG	Homo sapiens
38147_at	[illegible]	RG	Homo sapiens
38319_at	[illegible]	RG	Homo sapiens

Figure 5

Click here to download Figure Figure5.jpg


GeneCards
LifeMap
CD3D Gene (Protein Coding)
* 186790
CD3 ANTIGEN, DELTA SUBUNIT; CD3D
CD3-DELTA
Gene-Phenotype Relationships
TEXT
Description
UniProt
UniProtKB - P04234 (CD3D_HUMAN)
Display
Function
GPS
Group-based Prediction System
THE CUCKOO WORKGROUP
HOME
WEB SERVER
DOWNLOAD
CITATION
USER GUIDE
LINKS
CONTACT
PRODUCTS OF CUCKOO
Disorder region prediction

(a)

(b)

(c)

(d)

Figure 6

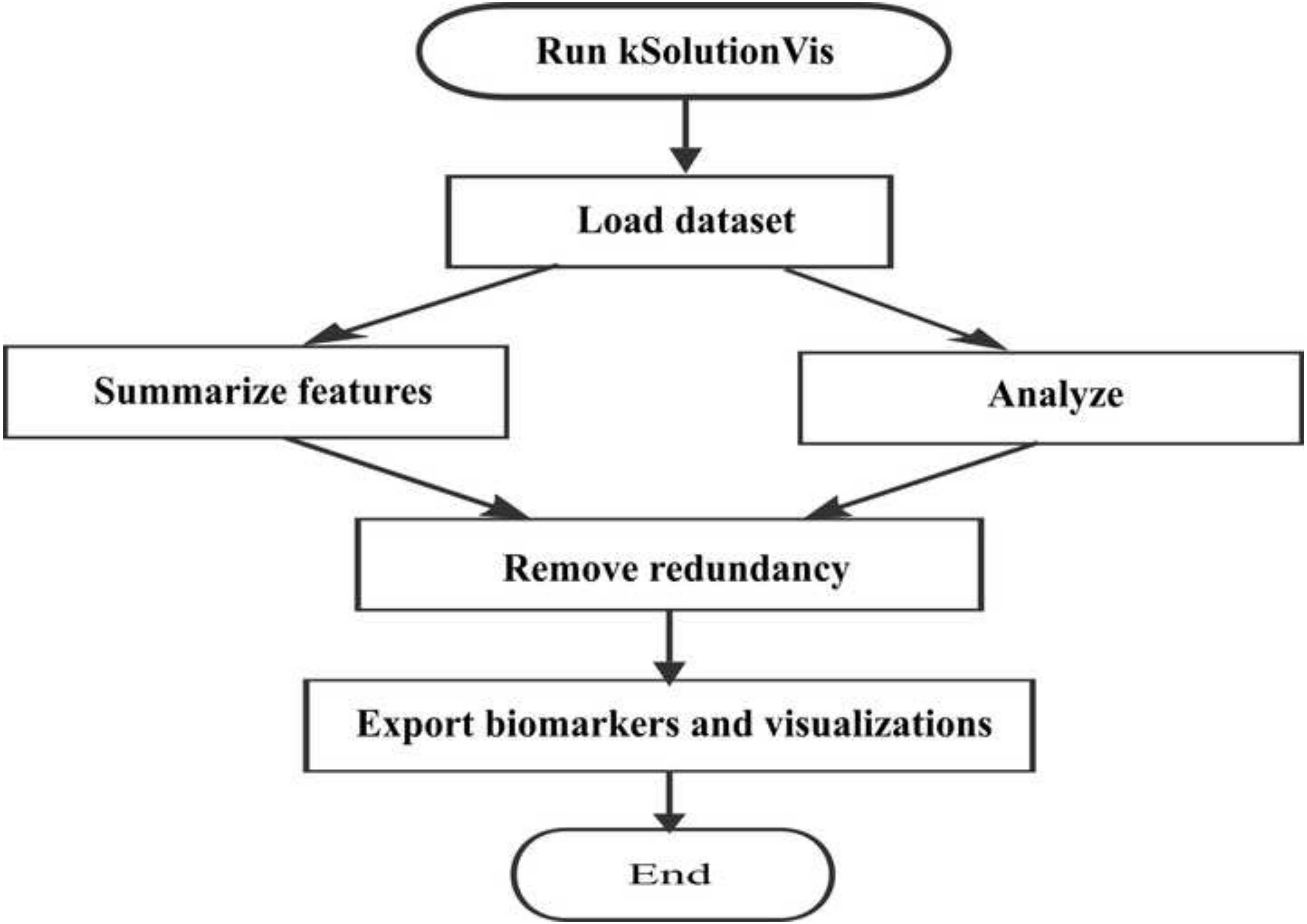
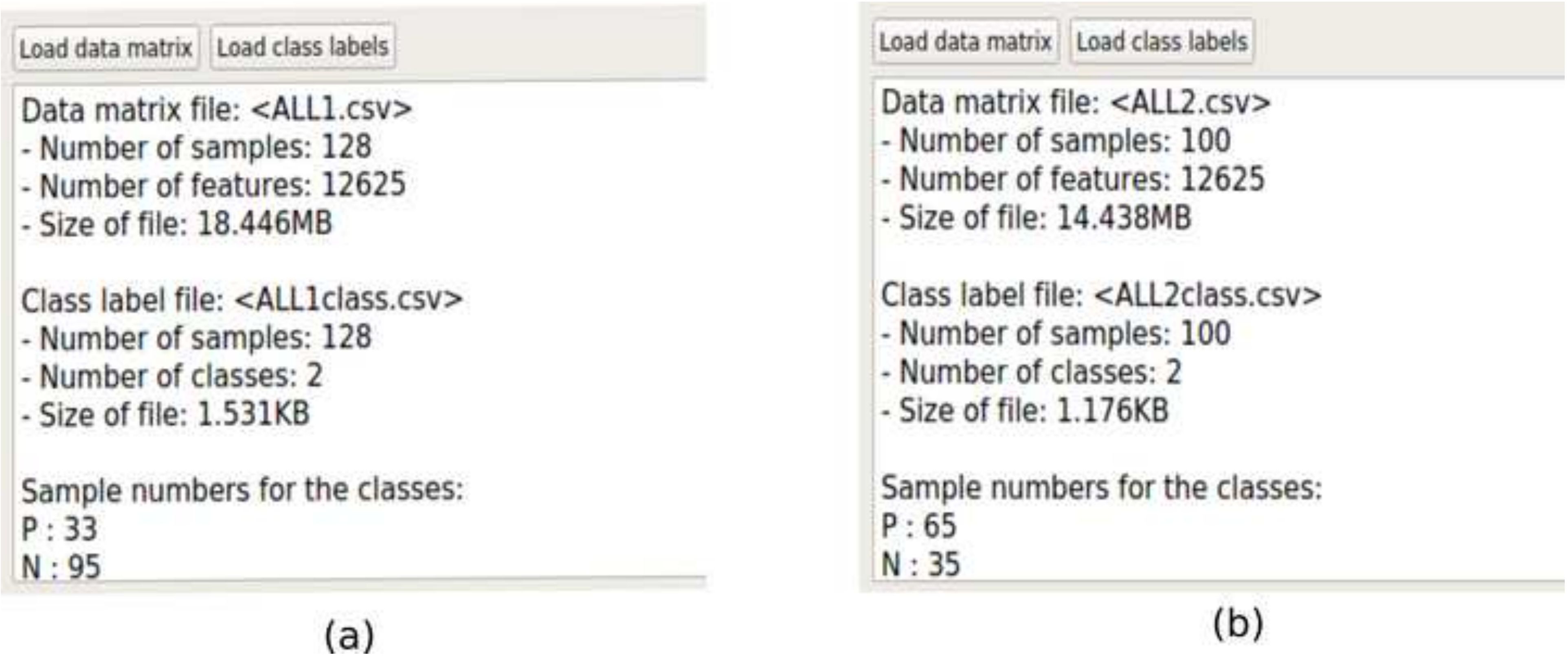

Run kSolutionVis
Load dataset
Summarize features
Analyze
Remove redundancy
Export biomarkers and visualizations
End

Figure 7

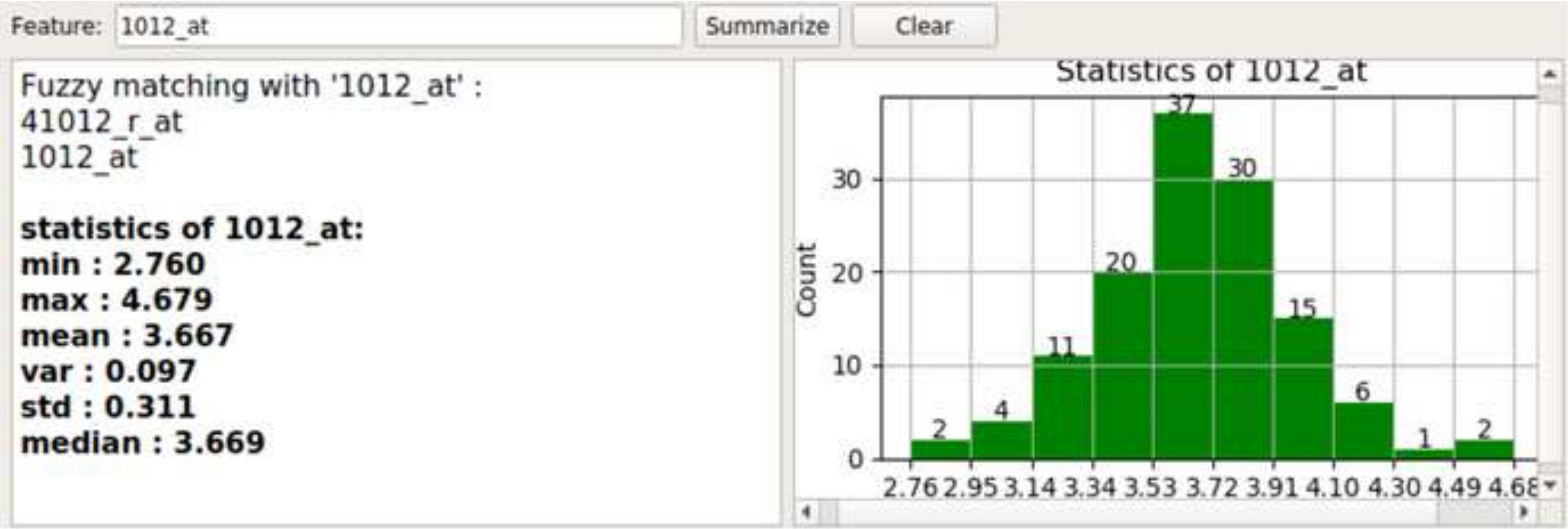


```
Data matrix file: <ALL1.csv>
- Number of samples: 128
- Number of features: 12625
- Size of file: 18.446MB

Class label file: <ALL1class.csv>
- Number of samples: 128
- Number of classes: 2
- Size of file: 1.531KB

Sample numbers for the classes:
P : 33
N : 95
```

(a)

```
Data matrix file: <ALL2.csv>
- Number of samples: 100
- Number of features: 12625
- Size of file: 14.438MB

Class label file: <ALL2class.csv>
- Number of samples: 100
- Number of classes: 2
- Size of file: 1.176KB

Sample numbers for the classes:
P : 65
N : 35
```

(b)

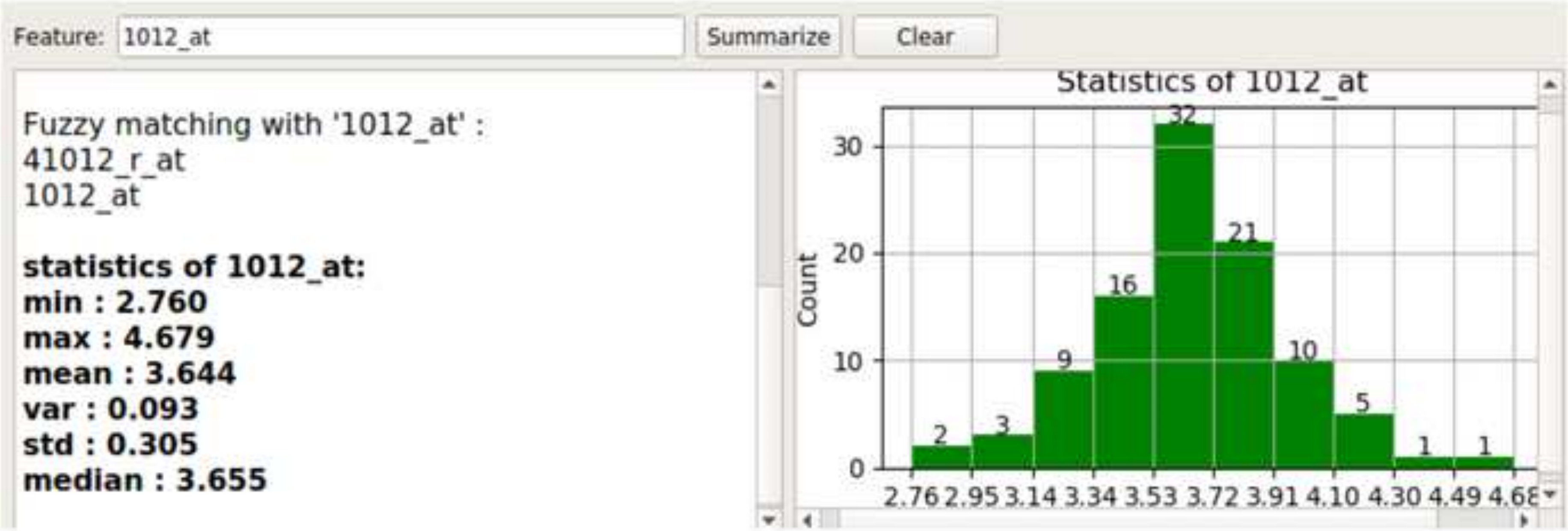


Feature: 1012_at

Fuzzy matching with '1012_at' :
41012_r_at
1012_at

statistics of 1012_at:
min : 2.760
max : 4.679
mean : 3.667
var : 0.097
std : 0.311
median : 3.669

(a)



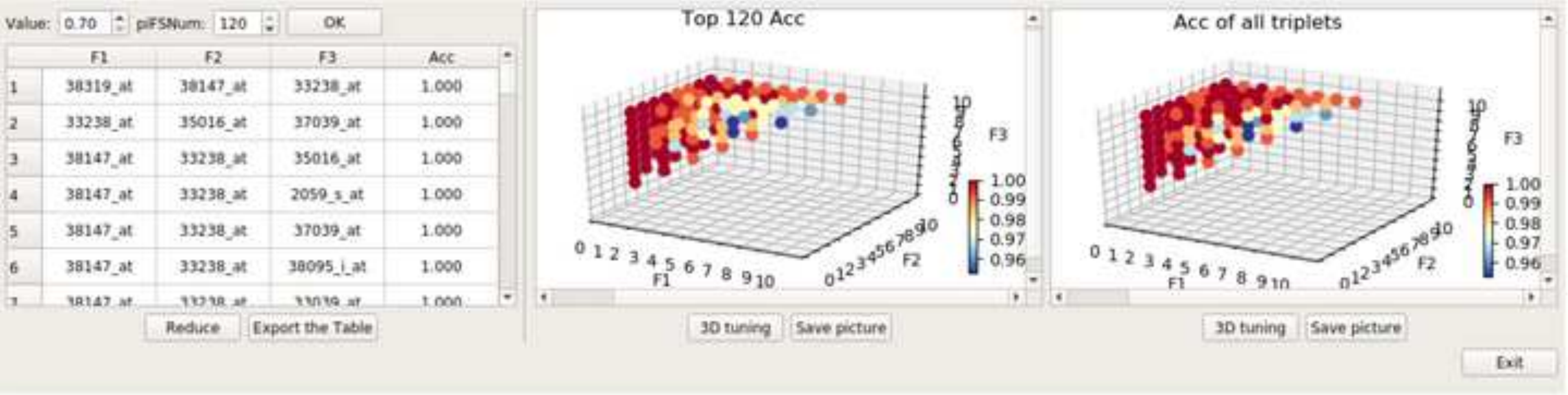
Feature: 1012_at

Fuzzy matching with '1012_at' :
41012_r_at
1012_at

statistics of 1012_at:
min : 2.760
max : 4.679
mean : 3.644
var : 0.093
std : 0.305
median : 3.655

(b)

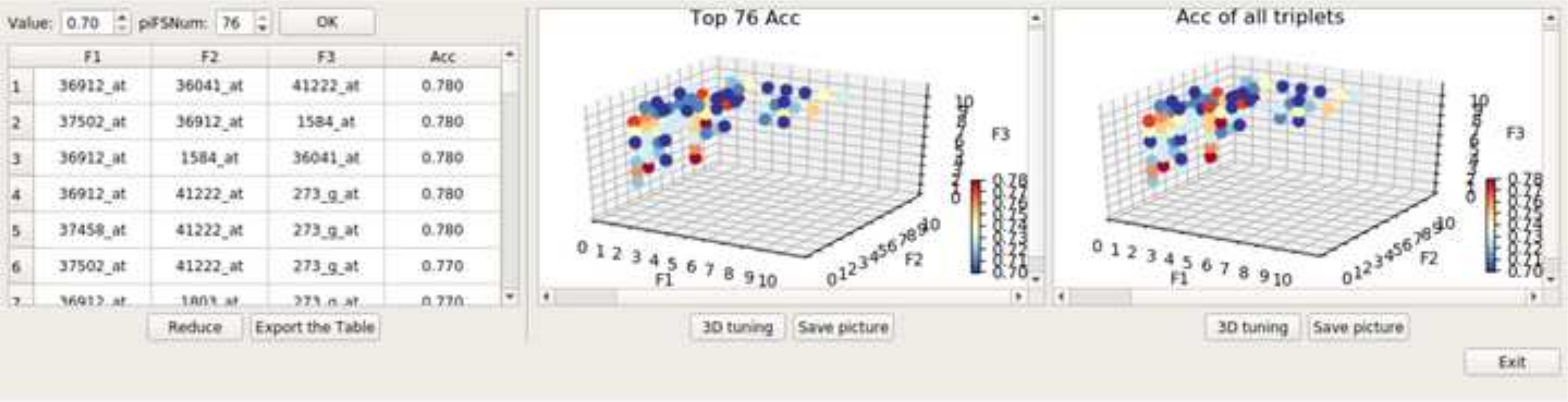
Figure 9



Value: 0.70 piFSNum: 120

	F1	F2	F3	Acc
1	38319_at	38147_at	33238_at	1.000
2	33238_at	35016_at	37039_at	1.000
3	38147_at	33238_at	35016_at	1.000
4	38147_at	33238_at	2059_s_at	1.000
5	38147_at	33238_at	37039_at	1.000
6	38147_at	33238_at	38095_i_at	1.000
7	38147_at	33238_at	33039_at	1.000

(a)



Value: 0.70 piFSNum: 76

	F1	F2	F3	Acc
1	36912_at	36041_at	41222_at	0.780
2	37502_at	36912_at	1584_at	0.780
3	36912_at	1584_at	36041_at	0.780
4	36912_at	41222_at	273_g_at	0.780
5	37458_at	41222_at	273_g_at	0.780
6	37502_at	41222_at	273_g_at	0.770
7	36912_at	1803_at	273_g_at	0.770

(b)

Web site	Link
GeneCards	http://www.genecards.org/cgi-bin/carddisp.pl?gene=CD3D
OMIM	https://omim.org/entry/186790?search=CD3D&highlight=cd3d
UniProtKB	http://www.uniprot.org/uniprot/P04234
GPS	http://gps.biocuckoo.org/
String	https://string-db.org/
David	https://david.ncifcrf.gov/

Functionality
Gene annotation
Gene-disease association
Protein annotation
Protein’s PTM prediction
Protein-protein interaction
Gene Set Enrichment Analysis

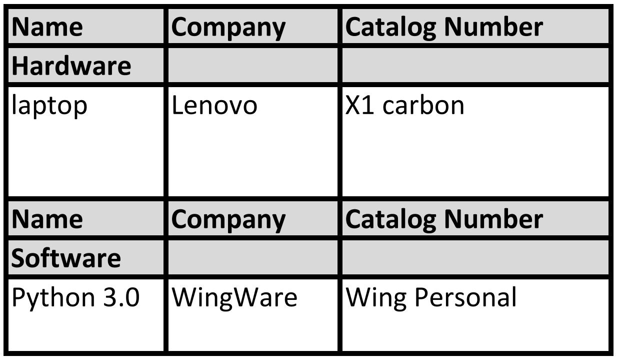
f1	f2	f3	Acc	Symbol1	Symbol2	Symbol3
38319_at	38147_at	33238_at	1.0000	CD3D	SH2D1A	LCK
33238_at	35016_at	37039_at	1.0000	LCK	CD74	HLA-DRA
38147_at	33238_at	35016_at	1.0000	SH2D1A	LCK	CD74
38147_at	33238_at	2059_s_a	1.0000	SH2D1A	LCK	LCK
38147_at	33238_at	37039_at	1.0000	SH2D1A	LCK	HLA-DRA
38147_at	33238_at	38095_i_a	1.0000	SH2D1A	LCK	HLA-DPB1
38147_at	33238_at	33039_at	1.0000	SH2D1A	LCK	TRAT1
38147_at	35016_at	2059_s_a	1.0000	SH2D1A	CD74	LCK
38147_at	35016_at	33039_at	1.0000	SH2D1A	CD74	TRAT1
38147_at	35016_at	38949_at	1.0000	SH2D1A	CD74	PRKCQ
38147_at	2059_s_at	37039_at	1.0000	SH2D1A	LCK	HLA-DRA
38147_at	2059_s_at	38095_i_a	1.0000	SH2D1A	LCK	HLA-DPB1
38147_at	37039_at	33039_at	1.0000	SH2D1A	HLA-DRA	TRAT1
38147_at	37039_at	38949_at	1.0000	SH2D1A	HLA-DRA	PRKCQ
38319_at	38147_at	35016_at	1.0000	CD3D	SH2D1A	CD74
38147_at	38833_at	38949_at	1.0000	SH2D1A	HLA-DPA1	PRKCQ
33238_at	35016_at	33039_at	1.0000	LCK	CD74	TRAT1
38319_at	38833_at	38949_at	1.0000	CD3D	HLA-DPA1	PRKCQ
33238_at	35016_at	38949_at	1.0000	LCK	CD74	PRKCQ
33238_at	2059_s_at	37039_at	1.0000	LCK	LCK	HLA-DRA
33238_at	37039_at	38095_i_a	1.0000	LCK	HLA-DRA	HLA-DPB1
33238_at	37039_at	33039_at	1.0000	LCK	HLA-DRA	TRAT1
33238_at	37039_at	38949_at	1.0000	LCK	HLA-DRA	PRKCQ
33238_at	38095_i_at	38949_at	1.0000	LCK	HLA-DPB1	PRKCQ
33238_at	38833_at	38949_at	1.0000	LCK	HLA-DPA1	PRKCQ
33238_at	33039_at	38949_at	1.0000	LCK	TRAT1	PRKCQ
35016_at	2059_s_at	33039_at	1.0000	CD74	LCK	TRAT1
35016_at	2059_s_at	38949_at	1.0000	CD74	LCK	PRKCQ
35016_at	38095_i_at	38949_at	1.0000	CD74	HLA-DPB1	PRKCQ
2059_s_at	37039_at	33039_at	1.0000	LCK	HLA-DRA	TRAT1
2059_s_at	38095_i_at	38949_at	1.0000	LCK	HLA-DPB1	PRKCQ
2059_s_at	38833_at	38949_at	1.0000	LCK	HLA-DPA1	PRKCQ
38319_at	33039_at	38949_at	1.0000	CD3D	TRAT1	PRKCQ
38147_at	38095_i_at	38949_at	1.0000	SH2D1A	HLA-DPB1	PRKCQ
38319_at	33238_at	38833_at	1.0000	CD3D	LCK	HLA-DPA1
38319_at	2059_s_at	38833_at	1.0000	CD3D	LCK	HLA-DPA1
38319_at	33238_at	33039_at	1.0000	CD3D	LCK	TRAT1
38319_at	33238_at	38095_i_a	1.0000	CD3D	LCK	HLA-DPB1
38319_at	33238_at	37039_at	1.0000	CD3D	LCK	HLA-DRA
38319_at	35016_at	38833_at	1.0000	CD3D	CD74	HLA-DPA1
38319_at	33238_at	2059_s_a	1.0000	CD3D	LCK	LCK
38319_at	35016_at	33039_at	1.0000	CD3D	CD74	TRAT1
38319_at	33238_at	35016_at	1.0000	CD3D	LCK	CD74
38319_at	35016_at	38949_at	1.0000	CD3D	CD74	PRKCQ
38319_at	2059_s_at	37039_at	1.0000	CD3D	LCK	HLA-DRA
38319_at	38147_at	38949_at	1.0000	CD3D	SH2D1A	PRKCQ

38319_at	38147_at	33039_at	1.0000	CD3D	SH2D1A	TRAT1
38319_at	33238_at	38949_at	1.0000	CD3D	LCK	PRKCQ
38319_at	2059_s_at	38095_i_a	1.0000	CD3D	LCK	HLA-DPB1
38319_at	38147_at	38833_at	1.0000	CD3D	SH2D1A	HLA-DPA1
38319_at	2059_s_at	33039_at	1.0000	CD3D	LCK	TRAT1
38319_at	38147_at	38095_i_a	1.0000	CD3D	SH2D1A	HLA-DPB1
38319_at	37039_at	33039_at	1.0000	CD3D	HLA-DRA	TRAT1
38319_at	38147_at	37039_at	1.0000	CD3D	SH2D1A	HLA-DRA
38319_at	38147_at	2059_s_a	1.0000	CD3D	SH2D1A	LCK
38319_at	2059_s_at	38949_at	1.0000	CD3D	LCK	PRKCQ
38319_at	35016_at	2059_s_a	1.0000	CD3D	CD74	LCK
2059_s_at	37039_at	38095_i_a	0.9922	LCK	HLA-DRA	HLA-DPB1
35016_at	33039_at	38949_at	0.9922	CD74	TRAT1	PRKCQ
2059_s_at	37039_at	38949_at	0.9922	LCK	HLA-DRA	PRKCQ
35016_at	2059_s_at	37039_at	0.9922	CD74	LCK	HLA-DRA
35016_at	37039_at	38949_at	0.9922	CD74	HLA-DRA	PRKCQ
35016_at	38833_at	38949_at	0.9922	CD74	HLA-DPA1	PRKCQ
2059_s_at	33039_at	38949_at	0.9922	LCK	TRAT1	PRKCQ
37039_at	38833_at	38949_at	0.9922	HLA-DRA	HLA-DPA1	PRKCQ
37039_at	33039_at	38949_at	0.9922	HLA-DRA	TRAT1	PRKCQ
38319_at	38095_i_at	38949_at	0.9922	CD3D	HLA-DPB1	PRKCQ
33238_at	37039_at	38833_at	0.9922	LCK	HLA-DRA	HLA-DPA1
38095_i_at	33039_at	38949_at	0.9922	HLA-DPB1	TRAT1	PRKCQ
33238_at	2059_s_at	38949_at	0.9922	LCK	LCK	PRKCQ
38319_at	38833_at	33039_at	0.9922	CD3D	HLA-DPA1	TRAT1
38833_at	33039_at	38949_at	0.9922	HLA-DPA1	TRAT1	PRKCQ
38147_at	33039_at	38949_at	0.9922	SH2D1A	TRAT1	PRKCQ
38319_at	37039_at	38833_at	0.9922	CD3D	HLA-DRA	HLA-DPA1
38147_at	2059_s_at	38949_at	0.9922	SH2D1A	LCK	PRKCQ
38147_at	38095_i_at	38833_at	0.9922	SH2D1A	HLA-DPB1	HLA-DPA1
38147_at	33238_at	38949_at	0.9922	SH2D1A	LCK	PRKCQ
38147_at	2059_s_at	33039_at	0.9922	SH2D1A	LCK	TRAT1
38319_at	37039_at	38949_at	0.9922	CD3D	HLA-DRA	PRKCQ
38319_at	38095_i_at	38833_at	0.9922	CD3D	HLA-DPB1	HLA-DPA1
38147_at	2059_s_at	38833_at	0.9922	SH2D1A	LCK	HLA-DPA1
33238_at	35016_at	2059_s_a	0.9922	LCK	CD74	LCK
38319_at	35016_at	38095_i_a	0.9922	CD3D	CD74	HLA-DPB1
33238_at	35016_at	38095_i_a	0.9922	LCK	CD74	HLA-DPB1
38319_at	35016_at	37039_at	0.9922	CD3D	CD74	HLA-DRA
38147_at	33238_at	38833_at	0.9922	SH2D1A	LCK	HLA-DPA1
38147_at	37039_at	38095_i_a	0.9844	SH2D1A	HLA-DRA	HLA-DPB1
38147_at	35016_at	38833_at	0.9844	SH2D1A	CD74	HLA-DPA1
38147_at	35016_at	38095_i_a	0.9844	SH2D1A	CD74	HLA-DPB1
35016_at	2059_s_at	38095_i_a	0.9844	CD74	LCK	HLA-DPB1
38147_at	37039_at	38833_at	0.9844	SH2D1A	HLA-DRA	HLA-DPA1
35016_at	2059_s_at	38833_at	0.9844	CD74	LCK	HLA-DPA1
38319_at	37039_at	38095_i_a	0.9844	CD3D	HLA-DRA	HLA-DPB1

37039_at	38095_i_at	38949_at	0.9844	HLA-DRA	HLA-DPB1	PRKCQ
38147_at	38833_at	33039_at	0.9844	SH2D1A	HLA-DPA1	TRAT1
38095_i_at	38833_at	38949_at	0.9844	HLA-DPB1	HLA-DPA1	PRKCQ
33238_at	35016_at	38833_at	0.9844	LCK	CD74	HLA-DPA1
38319_at	38095_i_at	33039_at	0.9844	CD3D	HLA-DPB1	TRAT1
2059_s_at	37039_at	38833_at	0.9844	LCK	HLA-DRA	HLA-DPA1
2059_s_at	38833_at	33039_at	0.9766	LCK	HLA-DPA1	TRAT1
2059_s_at	38095_i_at	33039_at	0.9766	LCK	HLA-DPB1	TRAT1
2059_s_at	38095_i_at	38833_at	0.9766	LCK	HLA-DPB1	HLA-DPA1
33238_at	2059_s_at	38095_i_a	0.9766	LCK	LCK	HLA-DPB1
35016_at	38095_i_at	33039_at	0.9766	CD74	HLA-DPB1	TRAT1
38147_at	38095_i_at	33039_at	0.9766	SH2D1A	HLA-DPB1	TRAT1
33238_at	2059_s_at	33039_at	0.9766	LCK	LCK	TRAT1
35016_at	37039_at	33039_at	0.9766	CD74	HLA-DRA	TRAT1
33238_at	38095_i_at	33039_at	0.9766	LCK	HLA-DPB1	TRAT1
33238_at	38833_at	33039_at	0.9766	LCK	HLA-DPA1	TRAT1
35016_at	38833_at	33039_at	0.9766	CD74	HLA-DPA1	TRAT1
33238_at	38095_i_at	38833_at	0.9688	LCK	HLA-DPB1	HLA-DPA1
37039_at	38833_at	33039_at	0.9688	HLA-DRA	HLA-DPA1	TRAT1
38147_at	35016_at	37039_at	0.9688	SH2D1A	CD74	HLA-DRA
33238_at	2059_s_at	38833_at	0.9688	LCK	LCK	HLA-DPA1
37039_at	38095_i_at	33039_at	0.9688	HLA-DRA	HLA-DPB1	TRAT1
38095_i_at	38833_at	33039_at	0.9609	HLA-DPB1	HLA-DPA1	TRAT1
35016_at	38095_i_at	38833_at	0.9609	CD74	HLA-DPB1	HLA-DPA1
37039_at	38095_i_at	38833_at	0.9531	HLA-DRA	HLA-DPB1	HLA-DPA1
35016_at	37039_at	38095_i_a	0.9531	CD74	HLA-DRA	HLA-DPB1
35016_at	37039_at	38833_at	0.9531	CD74	HLA-DRA	HLA-DPA1

Materials Table

Click here to download Materials Table Material Table.xlsx



Name	Company	Catalog Number
Hardware		
laptop	Lenovo	X1 carbon
Name	**Company**	**Catalog Number**
Software		
Python 3.0	WingWare	Wing Personal

Comments
Any computer works. Recommended minimum configuration: 1GB extra hard disk space, 1 GB memory, 2.0MHz CPU

Comments
Any python programming and running environments support Python version 3.0 or above

jove JOURNAL OF VISUALIZED EXPERIMENTS

1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

ARTICLE AND VIDEO LICENSE AGREEMENT

Title of Article: Selecting multiple biomarker subsets with similarly effective binary classification performances

Author(s): Xin Feng, Shaofei Wang, Quewang Liu, Jiamei Liu, Cheng Xu, Weifeng Yang, Yayun Shu, Weiwei Zheng, Fengfeng Zhou

Item 1 (check one box): The Author elects to have the Materials be made available (as described at http://www.jove.com/author) via: □ Standard Access ☒ Open Access

Item 2 (check one box):

☒ The Author is NOT a United States government employee.

□ The Author is a United States government employee and the Materials were prepared in the course of his or her duties as a United States government employee.

□ The Author is a United States government employee but the Materials were NOT prepared in the course of his or her duties as a United States government employee.

ARTICLE AND VIDEO LICENSE AGREEMENT

1. Defined Terms. As used in this Article and Video License Agreement, the following terms shall have the following meanings: "**Agreement**" means this Article and Video License Agreement; "**Article**" means the article specified on the last page of this Agreement, including any associated materials such as texts, figures, tables, artwork, abstracts, or summaries contained therein; "**Author**" means the author who is a signatory to this Agreement; "**Collective Work**" means a work, such as a periodical issue, anthology or encyclopedia, in which the Materials in their entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole; "**CRC License**" means the Creative Commons Attribution-Non Commercial-No Derivs 3.0 Unported Agreement, the terms and conditions of which can be found at: http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode; "**Derivative Work**" means a work based upon the Materials or upon the Materials and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Materials may be recast, transformed, or adapted; "**Institution**" means the institution, listed on the last page of this Agreement, by which the Author was employed at the time of the creation of the Materials; "**JoVE**" means MyJove Corporation, a Massachusetts corporation and the publisher of *The Journal of Visualized Experiments;* "**Materials**" means the Article and / or the Video; "**Parties**" means the Author and JoVE; "**Video**" means any video(s) made by the Author, alone or in conjunction with any other parties, or by JoVE or its affiliates or agents, individually or in collaboration with the Author or any other parties, incorporating all or any portion of the Article, and in which the Author may or may not appear.

2. Background. The Author, who is the author of the Article, in order to ensure the dissemination and protection of the Article, desires to have the JoVE publish the Article and create and transmit videos based on the Article. In furtherance of such goals, the Parties desire to memorialize in this Agreement the respective rights of each Party in and to the Article and the Video.

3. Grant of Rights in Article. In consideration of JoVE agreeing to publish the Article, the Author hereby grants to JoVE, subject to **Sections 4** and **7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Article in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Article into other languages, create adaptations, summaries or extracts of the Article or other Derivative Works (including, without limitation, the Video) or Collective Works based on all or any portion of the Article and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. If the "Open Access" box has been checked in **Item 1** above, JoVE and the Author hereby grant to the public all such rights in the Article as provided in, but subject to all limitations and requirements set forth in, the CRC License.

4. Retention of Rights in Article. Notwithstanding the exclusive license granted to JoVE in **Section 3** above, the Author shall, with respect to the Article, retain the non-exclusive right to use all or part of the Article for the non-commercial purpose of giving lectures, presentations or teaching classes, and to post a copy of the Article on the Institution's website or the Author's personal website, in each case provided that a link to the Article on the JoVE website is provided and notice of JoVE's copyright in the Article is included. All non-copyright intellectual property rights in and to the Article, such as patent rights, shall remain with the Author.

5. Grant of Rights in Video – Standard Access. This **Section 5** applies if the "Standard Access" box has been checked in **Item 1** above or if no box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby acknowledges and agrees that, Subject to **Section 7** below, JoVE is and shall be the sole and exclusive owner of all rights of any nature, including, without limitation, all copyrights, in and to the Video. To the extent that, by law, the Author is deemed, now or at any time in the future, to have any rights of any nature in or to the Video, the Author hereby disclaims all such rights and transfers all such rights to JoVE.

6. Grant of Rights in Video – Open Access. This **Section 6** applies only if the "Open Access" box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby grants to JoVE, subject to **Section 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Video in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Video into other languages, create adaptations, summaries or extracts of the Video or other Derivative Works or Collective Works based on all or any portion of the Video and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. For any Video to which this Section 6 is applicable, JoVE and the Author hereby grant to the public all such rights in the Video as provided in, but subject to all limitations and requirements set forth in, the CRC License.

7. Government Employees. If the Author is a United States government employee and the Article was prepared in the course of his or her duties as a United States government employee, as indicated in **Item 2** above, and any of the licenses or grants granted by the Author hereunder exceed the scope of the 17 U.S.C. 403, then the rights granted hereunder shall be limited to the maximum rights permitted under such statute. In such case, all provisions contained herein that are not in conflict with such statute shall remain in full force and effect, and all provisions contained herein that do so conflict shall be deemed to be amended so as to provide to JoVE the maximum rights permissible within such statute.

8. Likeness, Privacy, Personality. The Author hereby grants JoVE the right to use the Author's name, voice, likeness, picture, photograph, image, biography and performance in any way, commercial or otherwise, in connection with the Materials and the sale, promotion and distribution thereof. The Author hereby waives any and all rights he or she may have, relating to his or her appearance in the Video or otherwise relating to the Materials, under all applicable privacy, likeness, personality or similar laws.

9. Author Warranties. The Author represents and warrants that the Article is original, that it has not been published, that the copyright interest is owned by the Author (or, if more than one author is listed at the beginning of this Agreement, by such authors collectively) and has not been assigned, licensed, or otherwise transferred to any other party. The Author represents and warrants that the author(s) listed at the top of this Agreement are the only authors of the Materials. If more than one author is listed at the top of this Agreement and if any such author has not entered into a separate Article and Video License Agreement with JoVE relating to the Materials, the Author represents and warrants that the Author has been authorized by each of the other such authors to execute this Agreement on his or her behalf and to bind him or her with respect to the terms of this Agreement as if each of them had been a party hereto as an Author. The Author warrants that the use, reproduction, distribution, public or private performance or display, and/or modification of all or any portion of the Materials does not and will not violate, infringe and/or misappropriate the patent, trademark, intellectual property or other rights of any third party. The Author represents and warrants that it has and will continue to comply with all government, institutional and other regulations, including, without limitation all institutional, laboratory, hospital, ethical, human and animal treatment, privacy, and all other rules, regulations, laws, procedures or guidelines, applicable to the Materials, and that all research involving human and animal subjects has been approved by the Author's relevant institutional review board.

10. JoVE Discretion. If the Author requests the assistance of JoVE in producing the Video in the Author's facility, the Author shall ensure that the presence of JoVE employees, agents or independent contractors is in accordance with the relevant regulations of the Author's institution. If more than one author is listed at the beginning of this Agreement, JoVE may, in its sole discretion, elect not take any action with respect to the Article until such time as it has received complete, executed Article and Video License Agreements from each such author. JoVE reserves the right, in its absolute and sole discretion and without giving any reason therefore, to accept or decline any work submitted to JoVE. JoVE and its employees, agents and independent contractors shall have

full, unfettered access to the facilities of the Author or of the Author's institution as necessary to make the Video, whether actually published or not. JoVE has sole discretion as to the method of making and publishing the Materials, including, without limitation, to all decisions regarding editing, lighting, filming, timing of publication, if any, length, quality, content and the like.

11. Indemnification. The Author agrees to indemnify JoVE and/or its successors and assigns from and against any and all claims, costs, and expenses, including attorney's fees, arising out of any breach of any warranty or other representations contained herein. The Author further agrees to indemnify and hold harmless JoVE from and against any and all claims, costs, and expenses, including attorney's fees, resulting from the breach by the Author of any representation or warranty contained herein or from allegations or instances of violation of intellectual property rights, damage to the Author's or the Author's institution's facilities, fraud, libel, defamation, research, equipment, experiments, property damage, personal injury, violations of institutional, laboratory, hospital, ethical, human and animal treatment, privacy or other rules, regulations, laws, procedures or guidelines, liabilities and other losses or damages related in any way to the submission of work to JoVE, making of videos by JoVE, or publication in *JoVE* or elsewhere by JoVE. The Author shall be responsible for, and shall hold JoVE harmless from, damages caused by lack of sterilization, lack of cleanliness or by contamination due to the making of a video by JoVE its employees, agents or independent contractors. All sterilization, cleanliness or decontamination procedures shall be solely the responsibility of the Author and shall be undertaken at the Author's expense. All indemnifications provided herein shall include JoVE's attorney's fees and costs related to said losses or damages. Such indemnification and holding harmless shall include such losses or damages incurred by, or in connection with, acts or omissions of JoVE, its employees, agents or independent contractors.

12. Fees. To cover the cost incurred for publication, JoVE must receive payment before production and publication the Materials. Payment is due in 21 days of invoice. Should the Materials not be published due to an editorial or production decision, these funds will be returned to the Author. Withdrawal by the Author of any submitted Materials after final peer review approval will result in a US$1,200 fee to cover pre-production expenses incurred by JoVE. If payment is not received by the completion of filming, production and publication of the Materials will be suspended until payment is received.

13. Transfer, Governing Law. This Agreement may be assigned by JoVE and shall inure to the benefits of any of JoVE's successors and assignees. This Agreement shall be governed and construed by the internal laws of the Commonwealth of Massachusetts without giving effect to any conflict of law provision thereunder. This Agreement may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to me one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

A signed copy of this document must be sent with all new submissions. Only one Agreement required per submission.

CORRESPONDING AUTHOR:

Name:	Fengfeng Zhou		
Department:	College of Computer Science and Technology		
Institution:	Jilin University		
Article Title:	Selecting multiple biomarker subsets with similarly effective binary classification performances		
Signature:	Fengfeng Zhou	Date:	Dec 15, 2017

Please submit a signed and dated copy of this license by one of the following three methods:

1) Upload a scanned copy of the document as a pfd on the JoVE submission site;
2) Fax the document to +1.866.381.2236;
3) Mail the document to JoVE / Attn: JoVE Editorial / 1 Alewife Center #200 / Cambridge, MA 02139

For questions, please email submissions@jove.com or call +1.617.945.9051

May 9, 2018
Editorial Office of JoVE,
Dear Editor of JoVE,

Thank you for your decision on our submission. We appreciate the editorial comments and all the editorial comments are responded point-by-point. All the revisions in the manuscript were highlighted in red according to the comments.

Hope you and the anonymous reviewers will be satisfied with the current revision! Any further suggestions are also welcome!

Sincerely,
Fengfeng Zhou,
College of Computer Science and Technology,
Jilin University

Editorial comments:

The manuscript has been modified and the updated manuscript, **57738_R2**.docx, is attached and located in your Editorial Manager account. Please use the updated version to make your revisions.

RESPONSE:

We appreciate the editorial comment on improving our manuscript and have downloaded the formatted manuscript for further revisions.

Editorial comments:

1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues.

RESPONSE:

The full-text spelling problem was checked with the MS WORD Grammar Checking Function by pressing the hotkey F7, and corrections were made accordingly.

Editorial comments:

2. Please do not highlight a step without highlighting any of the sub-steps for filming.

RESPONSE:

We un-highlighted all the steps without a highlighted sub-step.

Editorial comments:

3. Step 7: There is no sub-step for step 7. Please add more sub-steps to provide the details.

RESPONSE:

The original step 7 and its “Note” explained this step well, so we combined step 7 into the previous step (step 6), and change the numbers of the other steps.

Editorial comments:

4. Please do not abbreviate journal titles for all references.

RESPONSE:

We changed the journal name abbreviations the full names in all the references and marked the corrections in red.