The Author(s) *BMC Genomics* 2016, 17(Suppl 13):1025
DOI 10.1186/s12864-016-3317-7

BMC Genomics

RESEARCH Open Access

A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data

Yongli Hu[1,4*], Takeshi Hase[2], Hui Peng Li[3], Shyam Prabhakar[3], Hiroaki Kitano[2], See Kiong Ng[1], Samik Ghosh[2] and Lawrence Jin Kiat Wee[1,4]

From 15th International Conference On Bioinformatics (INCOB 2016)
Queenstown, Singapore. 21-23 September 2016

Abstract

Background: The ability to sequence the transcriptomes of single cells using single-cell RNA-seq sequencing technologies presents a shift in the scientific paradigm where scientists, now, are able to concurrently investigate the complex biology of a heterogeneous population of cells, one at a time. However, till date, there has not been a suitable computational methodology for the analysis of such intricate deluge of data, in particular techniques which will aid the identification of the unique transcriptomic profiles difference between the different cellular subtypes. In this paper, we describe the novel methodology for the analysis of single-cell RNA-seq data, obtained from neocortical cells and neural progenitor cells, using machine learning algorithms (Support Vector machine (SVM) and Random Forest (RF)).

Results: Thirty-eight key transcripts were identified, using the SVM-based recursive feature elimination (SVM-RFE) method of feature selection, to best differentiate developing neocortical cells from neural progenitor cells in the SVM and RF classifiers built. Also, these genes possessed a higher discriminative power (enhanced prediction accuracy) as compared commonly used statistical techniques or geneset-based approaches. Further downstream network reconstruction analysis was carried out to unravel hidden general regulatory networks where novel interactions could be further validated in web-lab experimentation and be useful candidates to be targeted for the treatment of neuronal developmental diseases.

Conclusion: This novel approach reported for is able to identify transcripts, with reported neuronal involvement, which optimally differentiate neocortical cells and neural progenitor cells. It is believed to be extensible and applicable to other single-cell RNA-seq expression profiles like that of the study of the cancer progression and treatment within a highly heterogeneous tumour.

Keywords: Single-cell RNA-seq, Machine learning, Network reconstruction, Systems biology

* Correspondence: huy@i2r.a-star.edu.sg; yongli@bic.nus.edu.sg
[1]Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore, Singapore
[4]The Systems Biology Institute, Singapore Node hosted at the Institute for Infocomm Research, A*STAR, Singapore, Singapore
Full list of author information is available at the end of the article

Background

The advent of sequencing technology has brought about the unprecedented ability to sequence individual single cells. Now, the distinct gene expression profiles of seemingly similar yet genetically heterogeneous subpopulations of cells within different tissue types can be elucidated with the use of single-cell sequencing technology. The study of such subpopulations within tumours is especially important in the study of differential reactivity of patients to drug treatments and that of acquired drug resistance within cancer patients [1, 2]. The complex underlying transcriptomic dynamics elucidated will enhance our understanding of the distinct gene expression signatures of different carcinomas or subpopulations within disparate tumour tissues which will ultimately aid in the optimization of cancer treatments.

A major challenge, however remains, is that of a suitable computational analytic pipeline for the analysis of single-cell RNA-Seq transcriptomic data. To address this problem, this paper proposes the identification of the unique gene expression profile within each subpopulation through traditional statistical methodology, geneset enrichment analysis (GSEA), machine learning algorithms where genes identified are subsequently used to build predictive classifiers for cell type prediction. Computational analysis of RNA-Seq transcriptomic data using machine learning algorithms, particularly that of supervised learning algorithms, like rule-based machine learning techniques [3], Support Vector Machine (SVM)-based [4, 5] and network-based approaches [6], is not new. However, this paper is the first to utilize a combination of two different machine learning algorithms (SVM and Random Forest (RF)) on single-cell RNA-seq transcriptomic data to identify the key signatures of different cell types for cell type prediction. Using single-cell RNA-seq expression data from neocortical cells and those of neural progenitor cells as inputs, we have identified a set of 38 key genes which optimally differentiates developing neocortical cells and those of neural progenitor cells.

Further, relevance of the differentially expressed genes in neuronal cell differentiation were also investigated using network-based approaches where the gene regulatory networks (GRNs) inferred elucidated the potential underlying interactions/functions of the key hub genes (eg, genes that regulate many genes in neuronal cells but do not regulate genes in neuronal progenitor cells) which could be further validated in wet-lab experimentation [7, 8]. In summary, this paper described a novel computational pipeline for the study of single-cell RNA-Seq transcriptomic data where key genes identified were used, with high accuracy, to predict distinct neuronal cell subtypes where such a system could be used to uncover the different subpopulations within a newly sequenced brain tissue. In addition, downstream network studies lend a systems-level relevance where potential underlying relationships are unravelled and potentially be used for targeted for treatment in neuronal developmental diseases.

Results

Prefiltering of genes

The summary of the methodology employed in this paper is summarized in Fig. 1.

Raw data, downloaded from the NCBI databased, was filtered (the criteria used for data filtering is reported in the materials and methods section) to obtain a total 8681 (~15%) genes which represents the dataset denoted by "all genes". A total of 65 samples (15 NPC and 50 developing neuronal cells) were used as training data to SVM and RF algorithms. Significance of the cell types assignment is further validated using Pvclust [9], which employs a multiple bootstrap (50,000) resampling algorithm to calculate the approximately unbiased (AU) probability values for cluster distinctions which is shown in red in Fig. 2.

SVM and RF-based classification of neuronal cell types on all expressed genes

Each set of gene expression data, extracted from different feature selection methods, was used to train two different machine learning (ML) models, namely a SVM and a RF classifier. The classification was reduced to a two-class problem where the predictor was designed to identify potential neocortical neuronal cells from NPCs.

Due to the limited number of samples, 65 (15 NPC and 50 neuronal cells) used in this study, the data was not separated into training and testing set for the construction of SVM-based classifiers. Instead, leave-one-out (LOO) cross validation was carried out. However, this was deemed unnecessary for RF as classifiers were built by aggregating a large number of different decision trees, predictors built with the random forests algorithm is expected to have low variance and low bias.

SVM and RF classifiers built with the filtered high dimensional single-cell RNA-seq expression dataset, consisting of more than 8000 transcripts (Table 1), yields an accuracy of 95.3 and 76.9% respectively (Table 2). It seems like classifiers built with all transcripts, sans those of low expression, are able produce classifiers of a reasonably high accuracy, however, the quality of such classifiers needs to be co-ordinately investigated. To this end, the Matthews correlation coefficient (MCC) was used to validate the quality of the classifiers constructed. A coefficient of 1 represents a prefect prediction while that of 0 indicates a classifier producing predictions similar to random prediction. The SVM classifiers were far superior to the cognate RF classifiers having a MCC of 0.91 and 0, respectively. Thus, there is a need for

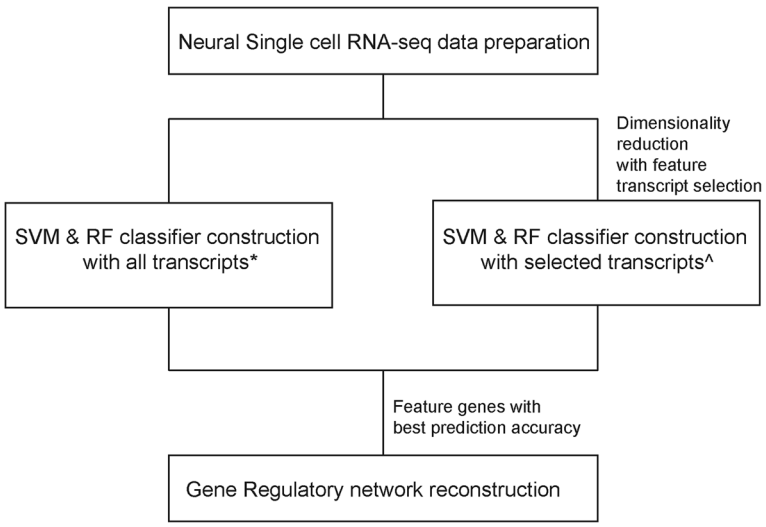

Fig. 1 Workflow for data analysis carried out in this paper. *All genes refers to the set of genes filtered by expression values and ^ selected genes refers to the optimal set of genes identified by geneset enrichment analysis (GSEA), statistical and machine-learning approaches (See Methods for more information)

additional feature transcript selection process to enhance both the accuracy and the quality of the constructed classifiers. Additionally, the construction of classifiers based on all transcripts are computational inefficient and the inclusion of large number of "noisy genes" will obscure important underlying signatures of each phenotypic class due to data overfitting and this will greatly limit the accuracy and quality of the classifiers [10]. On a more biological note, such a method fails to identify a subset of key genes which

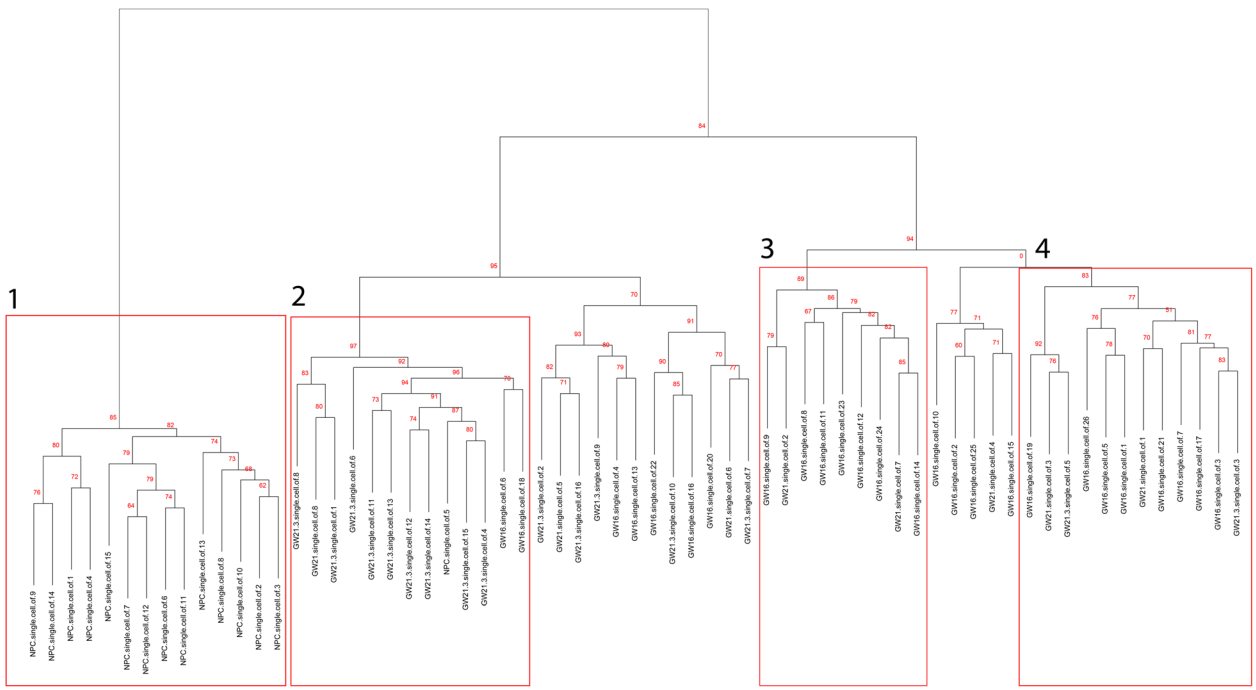

Fig. 2 Clustering of 65 neuronal cells. The approximately unbiased (AU) probability value at each node is shown in red font. There are four distinct clusters (red boxes labelled 1–4) with an AU higher than 80. Box1 comprises of mainly NPCs while boxes 2–4 primarily consists of neuronal cells only

Table 1 Genes/features selected by disparate feature selection techniques

Feature selection techniques[a]	Features/Genes (No.)
Filtered by low expression	8281
GSVA feature enrichment	1161
sRAP	837
SVM-RFE	38
RF-based Positive MDA	3339
T-test	60

[a]Feature selection is based on five different methodologies based on machine learning algorithms (SVM and RF) and also that of traditional differentially expressed genes (*sRAP*), *t*-test based analysis (*limma*) and genes in deregulated pathways (*GSVA*)

might have important biological applications in novel biomarker discovery.

SVM and RF-based classification of neuronal cell types with enhanced feature selection

In this study, a total of five different feature selection techniques were employed in for dimensionality reduction and they are pathway-based selection by GSVA, statistical-based selection by sRAP and *T*-test approaches and ML-based selection by SVM-RFE and RF-based positive MDA approaches. The number of transcripts selected by each feature elimination method can be found in Table 1. Additionally, the corresponding accuracies and MCC of each classifier are listed in Table 2.

The feature selection process decreased the number of transcripts analysed by ~60% to more than 95%. The best classifier constructed was that with features selected by SVM-RFE. This selection gave the best prediction accuracies and MCC, 100% and 1 respectively, for both the SVM and RF classifiers. GSEA has enabled scientists to identify sets or group of deregulated genes where preliminary insights to alternation of cellular mechanisms under different biological conditions can be studied [10].

Table 2 Accuracy of RF and SVM classifiers on the neuronal dataset

Genes selected	Accuracy (%)[a]		MCC^	
	SVM	RF	SVM	RF
All genes[b]	95.3	76.9	0.91	0.00
GSVA feature enrichment	98.5	76.9	0.87	0.00
sRAP	100	76.9	1.00	0.00
SVM-RFE	100	100	1.00	1.00
RF-based Positive MDA	100	76.9	1.00	0.00
T-test	100	97.0	1.00	0.91

The accuracy of the SVM predictors were obtained from LOO cross validation. SVM and RF classifiers were constructed with each set of data listed in Table 2
[a]All percentages are rounded off to three significant figures
[b]Transcripts with a total expression of zero and/or having more than six samples with expression levels less than one were excluded
^Matthews correlation coefficient (MCC) rounded to 2 decimal places

Given the usefulness of such a methodology, in this paper, we explore the impact on GSEA gene selection and prediction accuracy. Classifiers built with on GSVA-enriched genes did not considerably increase the prediction accuracy of the classifiers as a mere 3.2% increase in accuracy of the SVM predictor was obtained.

Nevertheless, it is interesting to note that RF classifiers generally have a lowered level of accuracy and a poor MCC value as compared to the cognate SVM models. For example, the RF classifier constructed using RF-based Positive MDA gene selection approach have an accuracy of 76.9, ~23% lower than that of the SVM classifier built with the same data. Also, the SVM classifier produces a perfect predictor (MCC = 1) while that of the RF classifier performs no better than random prediction (MCC = 0). This observation could be an indication of the shortcoming of tree-based ML methods to build high quality classifiers with single-cell RNA-seq expression data.

Network-level differences between NPCs and neuronal cells and their biological relevance in neuronal development

Gene regulatory networks (GRNs) among these transcripts inferred from RNA-seq expression profiles of SVM-RFE genes are useful in the investigation of system-level differences between the two cell types. Hub-genes (genes/transcripts that have a large number of regulatory interactions with other genes/transcripts) identified in GRNs might play potentially key roles in the maintenance of a particular cellular state. Thus, "differential hub genes (DHGs)" that are hub-genes/transcripts in neuronal cells (or NPCs) but not hub-genes in NPCs (or neuronal cells) could have important roles to differentiate the two cell types.

In order to investigate network-level difference (eg, DHGs) between the two cell types, GRNs were inferred (see Methods for more details) in neuronal cells (see Fig. 3a), NPCs (see Fig. 3b) from RNA-seq expression profiles and the structure of the two GRNs were subsequently compared (Fig. 3c). As observed, a large number of regulatory interactions are activated in one cell type but are not activated in the other cell type (blue links represents regulatory interactions activated in NPCs but not in neuronal cells, while red links represents those activated in neuronal cells but not in NPCs). For example, several interactions of the *Homeobox protein orthopedia* (OTP) gene (red-colored node in Fig. 3c) are activated in neuronal cells but not in NPCs and this is indicative that OTP gene is a potentially important gene which is possibly regulated in neuronal cells but not in NPCs.

In order to identify DHGs between the two cell types, we used a representative network metric, "degree", which is defined as the number of links to the transcript.

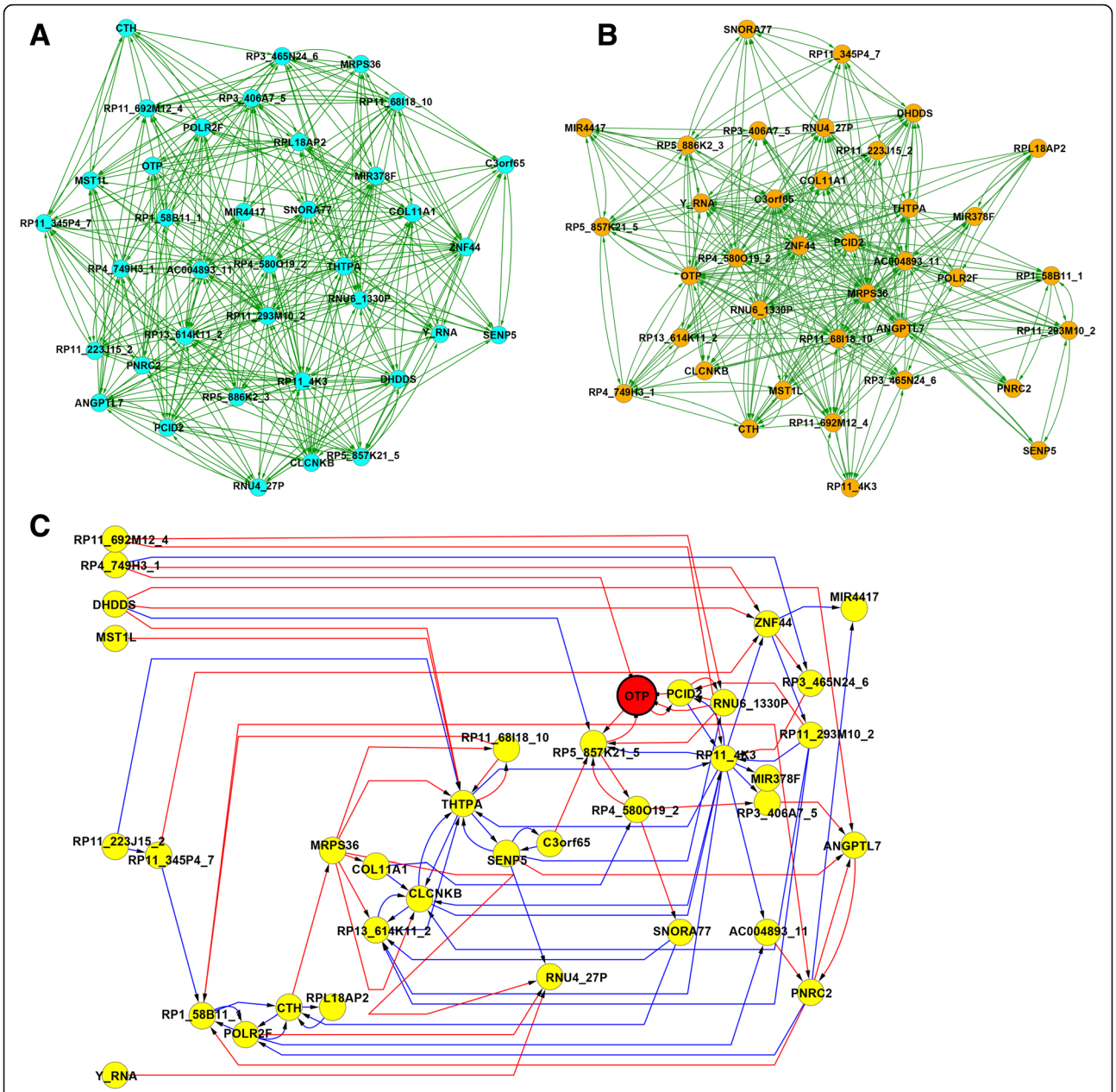

Fig. 3 GRN in NPCs (**a**) neuronal cells (**b**) and differential GRN between the two cell types. **a** GRN in NPCs. Nodes represent transcripts, while links between two nodes represent regulatory interactions between two transcripts in NPCs. Gene regulatory interactions with high confidence score (confidence score > 0.75) in NPCs are shown in the diagram. **b** GRN in neuronal cells. Nodes represent transcripts, while links between two nodes represent regulatory interactions between two transcripts in neuronal cells. Gene regulatory interactions with high confidence score (confidence score > 0.75) in neuronal cells are shown in the diagram. **c** Differential GRN between two cell lines. Nodes are transcripts. Red links represent gene regulatory interactions that are activated in neuronal cells but not activated in NPCs, while blue links represents those activated in NPCs but not in neuronal cells. In this diagram, we assumed that a regulatory interaction is activated in neuronal cells (or NPCs) but is not activated in NPCs (or neuronal cells), if difference in confidence score between the two cell types is greater than 0.75, eg, an interaction whose confidence scores are 0.99 and 0.20 in neuronal cells and NPCs, respectively. Note that OTP is a representative DHG (see Table 3) and thus the gene is highlighted in red

For weighted network, d_i, degree of a gene i is defined as, $d_j = \sum_{j=1}^{N} W_{i,j}$, where N is number of transcripts in a GRN and $w_{i,j}$ is weight (in this study, we used confidence score for a link as weight) for a regulatory interaction between two genes i and j. If a transcript have high-degree in a cell type, such transcript is defined as hub genes/transcripts in the cell type. In order to identify DHGs, for each of genes, we calculated difference in degrees between two cell types. For example, degree of MRPS36 in neuronal cells and that in NPCs are 57.08

and 33.99 respectively, and the degree difference of MPS is 23.09 (= |57.08 − 33.99|). Then, we ranked the transcripts according to their degree difference (Table 3). In Table 3, highly ranked transcripts (transcripts with high degree difference) are DHGs and may play an important role in differentiating neuronal cells from NPCs.

Table 3 Degree of difference between neuronal cells and NPCs in different genes

Genes	Degree in neuronal cells[a]	Degree in NPCs[a]	Degree difference between neuronal cells and NPCs[a]
MRPS36	57.1	34.0	23.1
RP11_4K3	23.0	44.0	21.0
SENP5	19.9	35.7	15.8
CLCNKB	23.9	39.6	15.7
POLR2F	23.7	39.2	15.4
OTP	48.3	33.1	15.2
RP1_58B11_1	24.6	39.3	14.7
RP11_293M10_2	32.4	45.9	13.5
RP3_465N24_6	25.6	39.1	13.5
SNORA77	30.0	43.2	13.2
C3orf65	50.8	38.0	12.8
ANGPTL7	49.7	37.9	11.8
RP4_580O19_2	47.2	36.9	10.2
RNU4_27P	45.1	35.0	10.0
COL11A1	25.8	35.3	9.46
RP11_68I18_10	47.9	39.3	8.67
Y_RNA	44.1	35.8	8.32
THTPA	29.9	37.5	7.61
ZNF44	45.7	38.6	7.11
CTH	28.9	35.9	7.02
RP11_692M12_4	42.8	35.8	7.02
RP11_345P4_7	32.0	38.8	6.82
RP13_614K11_2	36.5	43.3	6.81
MIR4417	30.7	37.2	6.55
RP11_223J15_2	36.7	42.5	5.77
RNU6_1330P	47.1	41.4	5.67
AC004893_11	45.1	39.9	5.23
RP3_406A7_5	33.4	37.5	4.06
MIR378F	39.1	35.5	3.63
PNRC2	33.7	30.6	3.15
RP5_886K2_3	43.5	40.6	2.89
RP5_857K21_5	34.9	33.2	1.61
PCID2	35.8	37.3	1.47
MST1L	38.5	37.4	1.00
RP4_749H3_1	39.5	38.9	0.624
RPL18AP2	36.0	36.6	0.573
DHDDS	41.1	40.7	0.414

[a]Degree of difference is corrected to three significant figures

Among the top ranked DHGs, we identified two potential key transcripts (the mitochondrial ribosomal protein S36 (MRPS36) and OTP) which could be responsible for the differentiation of NPCs from neuronal cells. MRPS36 is reported to be important in the maintenance of an undifferentiated state as overexpression of MRPS36 retards cell proliferation and delays cell cycle, helping cells maintain their undifferentiated state [11]. Similarly, the protein OTP is reportedly expressed in the hypothalamus during mammalian embryonic brain development and is a key determinant underlying FezI and Pac1-mediated of hypothalamic neural differentiation [12, 13]. These results indicate that the putative transcripts present within the top ranked DHGs could act as candidate targets for further experimental validation of their role in neuronal development. It is pertinent to note that as shown in Table 3, several of the SVM-RFE genes did not have significant difference in network property (degree) between the two cell types. This can signify that at the network level, all the SVM-RFE genes (DHGs) may not play biologically relevant role in differentiating the two cell types.

Discussion

The advancement of high throughput sequencing technologies has brought about an unprecedented ability for scientists to analyze the highly complex eukaryotic transcriptome by RNA-Seq. As compared to its predecessors, RNA-Seq has a very high signal-to-noise ratio and very large dynamic range. Reproducibility of RNA-Seq sequencing is also high and is able to provide high correlation across biological and technical replicates [14–16]. Further, single-cell RNA-seq techniques were developed, allowing finer insights to be elucidated with respect to the dynamics of disparate cellular differentiation, responses to stimulation and the stochastic nature of transcription within individual cells within a tissue or a tumour. Though it is still expensive to carry out RNA-seq sequencing in the current paradigm, it is expected that the sequencing cost will significantly decrease within the next few years [17].

In view of the impending information overflow, there is a concurrent need to develop more efficient techniques for the analysis such big data, especially for the construction of predictive models which can aid the identification and classification of different cell types as described in this paper. This work is one of the first to analyse single-cell RNA-seq profiles for the construction of predictive classifiers for neuronal cells and NPC. Also, classification accuracy of different models, built with features selected by different methods (ML based, GSEA or traditional DE genes based methods), have also been critically assessed. Further, we integrated the classification results with a

network inference pipeline to infer potential regulatory network amongst genes/transcripts (GRN) and identify network signatures for specific cell types.

Four key insights were obtained from this piece of work. First, transcripts selected by ML algorithm build better classifiers with enhanced accuracy, up to 100%, as compared to DE genes selected by traditional methods (sRAP and GSVA). Second, models built with differentially expressed transcripts selected from biological pathway-based methods (GSVA) proved to be inferior to that of models built with highly deregulated genes identified by traditional statistical means (sRAP and *T*-test). While pathway-based techniques are able to lend biological relevance to selected genes, genes selected by such methodology might not be able to capture gene expression signature of the disparate cell types studied. Thirdly, accuracy differs between classifiers built by different ML algorithms where RF, as compared to SVM, is unable to produce high accuracy for the high dimensional data analysed in this paper. Finally, the system-level analysis of the set optimal transcripts, using GRN inference analysis, is useful in the identification of hub transcripts which defines the different cell types investigated. Candidate transcripts identified by GRNs have potential biological correlates which are important in providing biological insights to cellular development. It is believed that such a workflow is extensible and applicable to other single-cell RNA-seq expression profiles like that of the study of the cancer progression within the highly heterogeneous cancer cells within a tumour [18].

Conclusion

The advancement in sequencing technologies have always brought along immense computational challenges to accurately and rapidly analyse the large amount of data generated from such experiments. Single-cell sequencing will inevitably become the gold standard for the study of genetic/transcriptomic aberrations, thus, concurrent efforts need to be placed in parallel to devise computational pipelines which can effectively analyse such big data. In addition to the use of legacy algorithms passed down for the era of microarray data analysis, there is a need to inject novelty and creativity in the analysis of single-cell RNA-seq data. This can be achieved using the combination of machine learning algorithms like SVM and RF and network reconstruction algorithms as reported in this paper. We have demonstrated that predictors built from transcripts selected using machine-learning based feature selection techniques which outperforms the commonly used statistical techniques or geneset-based approaches. Also, the novel incorporation of network reconstruction techniques have led to the identification of existing interactions and also potentially new interaction networks are identified which can be further validated in a smaller number of wet lab experiments as candidate biomarker genes. We believe that such a pipeline is extensible to other single-cell RNA-seq datasets, including those of tumor samples where the intricate transcriptomic complexity of the highly heterogeneous tumor can be unravelled for the design of personalized treatment for individual patient.

Methods

Data preparation for Single-cell RNA-Seq

Single-cell RNA gene expression profiles of neural cells from Pollen et al. [19] were used for this study as training data for the SVM/RF classifiers and will be called the "Fluidigm neural dataset" in this study. The data contained expression profiles of four neuronal cell populations, 65 samples in total, including (i) neural non-progenitor cells (NPCs), (ii) cells from the germinal zone of human cortex at gestational week (GW) 16 (GW16), (iii) 21 (GW21) and (iv) a subset of cells at GW21 which were furthered cultured for 3 weeks (GW21 + 3). Raw reads, obtained from the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) [20] under accession number SRP041736 [3], were mapped to the reference genome using Tophat2 (v2.1.0) [21] and were subjected to Fragments per Kilobase of Exon per Million Fragments Mapped (FPKM) normalization, using cuffdiff (v2.2.1) [22], prior to downstream analysis. Expression levels were logged prior to filtering. Genes having log FPKM of greater than one in more than six cells were included in the analyses. Samples were assigned labels of NPCs and those of the neuronal cells (inclusive of expression data obtained from cells of GW16, GW21 and GW21 + 3).

Construction of support vector machine (SVM) and random forest (RF) predictive models for the identification of fetal neocortical neuronal cells from NPCs

Classification accuracies of each disparate dataset were explored using two different machine learning algorithms - Support Vector Machine (SVM) and Random Forest (RF).

The SVM predictors were built with the LIBSVM package [23] and the RF predictors were built with the randomForest package in the R programming environment [24].

The detailed methodology for the construction of a SVM classifier can be obtained from the article by Burges [25] and a brief description of the SVM algorithm from Wee et al. [26]. Briefly, the SVM algorithm is based on the structural risk minimization principle from statistical learning theory [27]. A set of positive (single-cell RNA-Seq neocortical transcription data) and negative (single-cell RNA-Seq NPC transcription data) examples were represented by the feature vectors x_i (i = 1, 2,....N) with

corresponding labels $y_i \in \{+1,-1\}$. To classify the data as NPCs or neuronal cells, the SVM trains a classifier by mapping the input samples, using a kernel function (radial basis function (RBF) in this study), onto a high- dimensional space, and then seeking a separating hyperplane that differentiates the two classes with maximal margin and minimal error. Parameter optimization was carried out for g, which determines the capacity of the RBF kernel, and the regularization parameter C using leave-one-out (LOO) cross-validation. The optimal g and C values obtained from the optimization processes were used subsequently for training the entire training set to create the final SVM classifier.

RF is a tree-based classifier where classification is carried out by aggregating the votes for all trees built from different subsamples, randomly selected, with replacement, within the training set, from the training dataset. As the classifier is built by aggregating a large number of different decision trees, predictors built with the random forests algorithm is expected to have low variance and low bias. The number of trees (T) was set to 20,000 and the number of features to consider at each split in the decision tree (m) obtained from the optimization processes were used subsequently for training the entire training set to create the final RF classifier [28, 29].

Feature extraction and dimensionality reduction

Additionally, dimensionality reduction was carried out to obtain optimal subsets of gene/features for classifier construction and they are as listed below.

(i) Selection of genes from deregulated pathways using geneset enrichment analysis (GSEA). A non-parametric, unsupervised G was carried out with the Gene Set Variation Analysis (*GSVA*) package [30] in the R programming environment [24]. The original ensemble gene (ENSG) identifiers were mapped to their cognate HUGO Gene Nomenclature Committee (HGNC)/Uniprot identifiers using the biomaRt package in R [31]. This was carried out in order to permit the mapping of genes to that of the curated C2 geneset (September 2014), obtained from the Broad Institute's Molecular Signatures database version 4.0 (MSigDB) [10], for gene set analysis. Manual curation of the HGNC/Uniprot identifiers was subsequently carried out to obtain a curated list of identifiers. Genes with ENSG codes that were not matched to any symbols were removed. Also, gene fragments sharing the same symbol were excluded from analysis. The identities of the up and down-regulated pathways (p-value < 0.005), together with the corresponding genes within these genesets, were identified and reported.

(ii) Selection of differentially expressed (DE) genes using the R package Simplified RNA-Seq Analysis Pipeline (*sRAP*) [32, 33].

(iii) Selection of a subset of genes with the highest ranking criterion based on SVM-based classification [34] (SVM-RFE) using the R package *pathClass* [35]. SVM-RFE is an iterative gene selection process where features, expression values of different genes obtained from single-cell RNAseq experiments, with the smallest ranking criterion are recursively removed when the ranking criterion for all features are computed from the SVM-classifiers.

(iv) Selection of genes with positive mean decrease in accuracy (MDA) from RF analyses where selected feature genes are deemed to reduce classification error.

(v) Selection of DE genes using two-tailed T-test based analysis using R package *limma* [36] (p-value < 0.05).

Evaluation of model performance

A set of statistical variables were established to evaluate the performance, Accuracy and Matthews correlation coefficient (MCC) [37], of the SVM and RF classifiers. Only LOO cross validation was carried out for the SVM classifiers.

Inference of gene regulatory networks in neuronal cells and NPCs

A plethora of network-inference algorithms are now available and have been used to infer GRNs from gene expression datasets. As mentioned in Marbach et al. [38] and Hase et al. [39], different network-inference algorithms have different strength and weakness and complement with each other. Thus, by integrating heterogeneous network-inference algorithms, we can take advantage of their strengths to recover high-quality gene regulatory networks [38, 39].

Therefore, in this study, we selected 14 representative network-inference algorithms that are based on heterogeneous statistical techniques and integrated results from the selected algorithms. The selected algorithms includes six mutual information based methods (ARACNE [40], CLR [41], MRNET [42], RELNET [43], C3NET [44], and BC3NET [45]), two correlation based method (Spearman's correlation and Pearson's correlation [43]), one Bayesian network based method (SiGN-BN [46]), two random forest based method (GENIE3 [47] with two different parameter settings, see Table 4 for the details), two regression based method (TIGRESS [48] with two different parameter settings, see Table 4 for the details), and one method with both of ordinary differential equation based recursive optimization and mutual information (NARROMI [49]). The set of 14 algorithms includes several high-performance

Table 4 Parameter used to optimize each network-inference algorithms

Network-inference algorithms	Parameter optimization setting[c]
GENIE3-A[a]	K = "all", nb.trees = 10,000
GENIE3-B[a]	K = "sqrt", nb.trees = 10,000
TIGRESS-A[b]	scoring = "area"
TIGRESS-B[b]	scoring = "max"
ARACNE	eps = 0.1
BC3NET	boot = 10, alpha1 = 0.99, alpha2 = 0.99
SiGN-BN	Number of iteration of bootstrap method = 1,000

[a]GENIE3-A and -B represent two different parameter settings for GENIE3 algorithm used in this study
[b]TIGRESS-A and -B represent two different parameter settings for TIGRESS algorithm used in this study
[c]We used default settings for parameters that are not shown in this table

algorithms, ie, GENIE3 is the winner of both of DREAM4 (DREAM, Dialogue on Reverse Engineering Assessment and Methods) and DREAM5 network inference challenges [38, 47], while, in DREAM5, TIGRESS and CLR (and MRNET) are the best performer among regression techniques and that among mutual information techniques, respectively [38]. Zhang et al. demonstrated that NARROMI outperforms GENIE3 [49].

Integration of results from multiple network-inference algorithms

In this study, we have used a computational framework, "Top1net", to integrate results from the selected 14 network-inference algorithms [39].

Each of individual network-inference algorithms calculates confidence score for each gene pair and, an interaction between a gene pair with higher confidence score is more likely to be true positive interaction [39, 40, 50]. Top1net applies bagging method introduced by Breiman [51], to integrate confidence scores for each of gene pairs from multiple individual network-inference algorithms [39]. Top1net assumes that, if at least one network-inference algorithm assigns high confidence score to a gene pair, one gene in the pair has a regulatory interaction with another gene [39]. As network-inference algorithms tend to assign high confidence scores to true positive interactions, Top1net would recover a large number of true positive interactions in a GRN. The procedure of Top1Net is composed of three steps.

Step 1

From an expression dataset, an individual network-inference algorithm assigns confidence score for each of gene pairs and the gene pairs are ranked according to their confidence scores, ie, a gene pair with highest confidence score has the rank value of 1.

Step 2

We normalized ranked scores from each algorithm by scaling from 0 to 1 and used the normalized ranked scores (NRSs) as confidence scores by the algorithm. If a pair of genes i and j has rank value of $g_{i,j}$ by an algorithm, the $NRS_{i,j}$ of the gene pair by the algorithm is defined as, $NRS_{i,j} = \frac{N(N-1)+1-g_{i,j}}{N(N-1)}$, where N represents the number of genes in the gene expression dataset.

Step 3

We integrate NRSs from the algorithms by Top1net. For example, if we used the 14 network-inference algorithms to calculate 14 NRSs for each gene pairs. For each gene pairs, Top1net used the highest NRS among 14 NRSs as the confidence score of the gene pairs. For example, if the algorithms assign 14 NRSs, 0.98, 0.85, 0.8, 0.69, 0.65, 0.63, 0.62, 0.61, 0.58, 0.55, 0.53, 0.51, 0.50 and 0.35 for the gene pair, Top1net used 0.98 as the confidence score for the interaction between the gene pair.

RNA-seq expression profiles for GRN inference

Only 37 genes identified by SVM-RFE method were used for the inference of gene regulatory interactions within neuronal cells and NPCs as a single gene, RP4_803A2_1, was excluded for having expression values of 0 across all NPC samples.

Packages and parameters for individual network inference algorithms

To infer GRNs by individual algorithms, we used MINET package [52] for ARACNE, CLR, MRNET and RELNET, c3net packages for C3NET, bc3net packages for BC3NET, source code obtained from http://www.montefiore.ulg.ac.be/~huynh-thu/software.html for GENIE3, source code obtained from GP-DREAM network inference website (http://dream.broadinstitute.org/) for TIGRESS, and source code from http://comp-sysbio.org/narromi.htm for NARROMI. For SiGN-BN [28], we used software on the super-computing resource that was provided by Human Genome Center, the Institute of Medical Science, and the University of Tokyo. For PCC and SCC, we used R function ("cor" function) to calculate Pearson's and Spearman's correlation coefficient.

Nine, out of the 14, algorithms required optimization and the parameter settings for the nine network-inference algorithms are shown in Table 1. More information on algorithm customization can be found in references [40, 45–49] and also within the manuals written for the individual algorithms.

Acknowledgements

The authors would like to express their sincere gratitude to all individuals who have helped in this paper.

Declaration
This article has been published as part of *BMC Genomics* Volume 17 Supplement 3, 2016: 15th International Conference On Bioinformatics (INCOB 2016). The full contents of the supplement are available online at https://bmcgenet.biomedcentral.com/articles/supplements/volume-17-supplement-3.

Funding
Publication of this article was funded by the Agency for Science, Technology and Research (A*STAR) Joint Council (JCO) Project grant, EC-2013-063. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Availability of data and materials
Single-cell RNA gene expression profiles of neural cells from Pollen et al. [19] were used for this study and raw reads can be obtained from the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) [20] under accession number SRP041736 [3].

Authors' contributions
Supervised the research: SP, HK, SKN, SG and LJKW. Conceived the experiments: SG, LJKW. Designed the experiments and analyses: YH, TH. Performed the experiments: YH, TH, HPL. Analyzed the data: YH, TH. Wrote the paper: YH, TH. All authors read and approved the final manuscript.

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

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore, Singapore. [2]The Systems Biology Institute, Falcon Building 5 F, 5-6-9 Shirokanedai, Minato, Tokyo 108-0071, Japan. [3]Computational and Systems Biology, Genome Institute of Singapore, A*STAR, 60 Biopolis Street, Genome, #02-01, Singapore 138672, Singapore. [4]The Systems Biology Institute, Singapore Node hosted at the Institute for Infocomm Research, A*STAR, Singapore, Singapore.

Published: 22 December 2016

References

1. Suzuki A, Matsushima K, Makinoshima H, Sugano S, Kohno T, Tsuchihara K, et al. Single-cell analysis of lung adenocarcinoma cell lines reveals diverse expression patterns of individual cells invoked by a molecular target drug treatment. Genome Biol. 2015;16:66. doi:10.1186/s13059-015-0636-y. PubMed PMID: 25887790, PubMed Central PMCID: PMC4450998.
2. Kim KT, Lee HW, Lee HO, Kim SC, Seo YJ, Chung W, et al. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome Biol. 2015;16:127. doi:10.1186/s13059-015-0692-3. PubMed PMID: 26084335, PubMed Central PMCID: PMC4506401.
3. Cestarelli V, Fiscon G, Felici G, Bertolazzi P, Weitschek E. CAMUR: Knowledge extraction from RNA-seq cancer data through equivalent classification rules. Bioinformatics. 2015;32(5):697–704. doi:10.1093/bioinformatics/btv635.
4. Yao F, Zhang C, Du W, Liu C, Xu Y. Identification of gene-expression signatures and protein markers for breast cancer grading and staging. PLoS One. 2015;10(9):e0138213. doi:10.1371/journal.pone.0138213. PubMed PMID: 26375396, PubMed Central PMCID: PMC4573873.
5. Chen L, Xuan J, Riggins RB, Clarke R, Wang Y. Identifying cancer biomarkers by network-constrained support vector machines. BMC Syst Biol. 2011;5:161. doi:10.1186/1752-0509-5-161. PubMed PMID: 21992556, PubMed Central PMCID: PMC3214162.
6. Sundaramurthy G, Eghbalnia HR. A probabilistic approach for automated discovery of perturbed genes using expression data from microarray or RNA-Seq. Comput Biol Med. 2015;67:29–40. doi:10.1016/j.compbiomed.2015.07.029.
7. Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. Cell. 2011;144(6):986–98. doi:10.1016/j.cell.2011.02.016. PubMed PMID: 21414488, PubMed Central PMCID: PMC3102045.
8. Ahmad FK, Deris S, Othman NH. The inference of breast cancer metastasis through gene regulatory networks. J Biomed Inform. 2012;45(2):350–62. doi:10.1016/j.jbi.2011.11.015.
9. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006;22(12):1540–2. doi:10.1093/bioinformatics/btl117.
10. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50. doi:10.1073/pnas.0506580102. PubMed PMID: 16199517, PubMed Central PMCID: PMC1239896.
11. Chen YC, Chang MY, Shiau AL, Yo YT, Wu CL. Mitochondrial ribosomal protein S36 delays cell cycle progression in association with p53 modification and p21(WAF1/CIP1) expression. J Cell Biochem. 2007;100(4):981–90. doi:10.1002/jcb.21079.
12. Kaji T, Nonogaki K. Role of homeobox genes in the hypothalamic development and energy balance. Front Biosci (Landmark Ed). 2013;18:740–7.
13. Blechman J, Borodovsky N, Eisenberg M, Nabel-Rosen H, Grimm J, Levkowitz G. Specification of hypothalamic neurons by dual regulation of the homeodomain protein Orthopedia. Development. 2007;134(24):4417–26. doi:10.1242/dev.011262.
14. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods. 2008;5(7):613–9. doi:10.1038/nmeth.1223.
15. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320(5881):1344–9. doi:10.1126/science.1158441. PubMed PMID: 18451266, PubMed Central PMCID: PMC2951732.
16. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621–8. doi:10.1038/nmeth.1226.
17. Nagalakshmi U, Waern K, Snyder M. RNA-Seq: a method for comprehensive transcriptome analysis. Curr Protoc Mol Biology/edited by Frederick M Ausubel [et al.]. 2010;Chapter 4:Unit 4 11. 1–3. doi:10.1002/0471142727.mb0411s89.
18. Hou Y, Fan W, Yan L, Li R, Lian Y, Huang J, et al. Genome analyses of single human oocytes. Cell. 2013;155(7):1492–506. doi:10.1016/j.cell.2013.11.040.
19. Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol. 2014;32(10):1053–8. doi:10.1038/nbt.2967. PubMed PMID: 25086649, PubMed Central PMCID: PMC4191988.
20. Kodama Y, Shumway M, Leinonen R. International nucleotide sequence database C. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res. 2012;40(Database issue):D54–6. doi:10.1093/nar/gkr854. PubMed PMID: 22009675, PubMed Central PMCID: PMC3245110.
21. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36. doi:10.1186/gb-2013-14-4-r36. PubMed PMID: 23618408, PubMed Central PMCID: PMC4053844.
22. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511–5. doi:10.1038/nbt.1621. PubMed PMID: 20436464, PubMed Central PMCID: PMC3146043.
23. Chang C-C, Lin C-J. LIBSVM : a library for support vector machines. ACM Trans Intell Syst Technol. 2011;2(3):27.
24. R_Core_Team. A Language and Environment for Statistical Computing Vienna. Austria: R Foundation for Statistical Computing; 2015 [cited 2015]. Available from: http://www.r-project.org/.
25. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov. 1998;2(2):121–67. doi:10.1023/a:1009715923555.
26. Wee LJ, Simarmata D, Kam YW, Ng LF, Tong JC. SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction. BMC Genomics. 2010;11 Suppl 4:S21. doi:10.1186/1471-2164-11-S4-S21. PubMed PMID: 21143805; PubMed Central PMCID: PMC3005920.
27. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–397.

28. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
29. Treeratpituk P, Giles CL. Disambiguating Authors in academic publications using random forests. In: JCDL '09 Proceedings of the 9th ACM/IEEE-CS joint conference. 2009. p. 39–48. doi:10.1145/1555400.1555408.
30. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7. doi:10.1186/1471-2105-14-7. PubMed PMID: 23323831, PubMed Central PMCID: PMC3618321.
31. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439–40. doi:10.1093/bioinformatics/bti525.
32. Warden CD, Kanaya N, Chen S, Yuan YC. BD-Func: a streamlined algorithm for predicting activation and inhibition of pathways. PeerJ. 2013;1:e159. doi:10.7717/peerj.159. PubMed PMID: 24058887, PubMed Central PMCID: PMC3775632.
33. Warden CD, Yuan Y-C, Wu X. Optimal calculation of RNA-Seq fold-change values. Int J Comput Bioinformatics In Silico Model. 2013;2(6):285–92.
34. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389–422.
35. Johannes M, Frohlich H, Sultmann H, Beissbarth T. pathClass: an R-package for integration of pathway knowledge into support vector machines for biomarker discovery. Bioinformatics. 2011;27(10):1442–3. doi:10.1093/bioinformatics/btr157.
36. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. doi:10.1093/nar/gkv007. PubMed PMID: 25605792; PubMed Central PMCID: PMC4402510.
37. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975;405(2):442–51.
38. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9(8):796–804. doi:10.1038/nmeth.2016. PubMed PMID: 22796662, PubMed Central PMCID: PMC3512113.
39. Hase T, Ghosh S, Yamanaka R, Kitano H. Harnessing diversity towards the reconstructing of large scale gene regulatory networks. PLoS Comput Biol. 2013;9(11):e1003361. doi:10.1371/journal.pcbi.1003361. PubMed PMID: 24278007, PubMed Central PMCID: PMC3836705.
40. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006;7 Suppl 1:S7. doi:10.1186/1471-2105-7-S1-S7. PubMed PMID: 16723010; PubMed Central PMCID: PMC1810318.
41. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5(1):e8. doi:10.1371/journal.pbio.0050008. PubMed PMID: 17214507, PubMed Central PMCID: PMC1764438.
42. Meyer PE, Kontos K, Lafitte F, Bontempi G. Information-theoretic inference of large transcriptional regulatory networks. EURASIP J Bioinform Syst Biol. 2007:79879. doi:10.1155/2007/79879. PubMed PMID: 18354736; PubMed Central PMCID: PMC3171353.
43. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput. 2000:418–29.
44. Altay G, Emmert-Streib F. Inferring the conservative causal core of gene regulatory networks. BMC Syst Biol. 2010;4:132. doi:10.1186/1752-0509-4-132. PubMed PMID: 20920161, PubMed Central PMCID: PMC2955605.
45. de Matos SR, Emmert-Streib F. Bagging statistical network inference from large-scale gene expression data. PLoS One. 2012;7(3):e33624. doi:10.1371/journal.pone.0033624. PubMed PMID: 22479422, PubMed Central PMCID: PMC3316596.
46. Tamada Y, Shimamura T, Yamaguchi R, Imoto S, Nagasaki M, Miyano S. Sign: large-scale gene network estimation environment for high performance computing. Genome Inform. 2011;25(1):40–52.
47. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010;5(9). doi:10.1371/journal.pone.0012776. PubMed PMID: 20927193; PubMed Central PMCID: PMC2946910.
48. Haury AC, Mordelet F, Vera-Licona P, Vert JP. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. BMC Syst Biol. 2012;6:145. doi:10.1186/1752-0509-6-145. PubMed PMID: 23173819, PubMed Central PMCID: PMC3598250.
49. Zhang X, Liu K, Liu ZP, Duval B, Richer JM, Zhao XM, et al. NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. Bioinformatics. 2013;29(1):106–13. doi:10.1093/bioinformatics/bts619.
50. Altay G, Emmert-Streib F. Revealing differences in gene network inference algorithms on the network level by ensemble methods. Bioinformatics. 2010;26(14):1738–44. doi:10.1093/bioinformatics/btq259.
51. Breiman L. Bagging predictors. Mach Learn. 1996;24:123–40.
52. Meyer PE, Lafitte F, Bontempi G. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics. 2008;9:461. doi:10.1186/1471-2105-9-461. PubMed PMID: 18959772; PubMed Central PMCID: PMC2630331.