A peer-reviewed version of this preprint was published in PeerJ on 19 January 2017.

View the peer-reviewed version (peerj.com/articles/2888), which is the preferred citable publication unless you specifically need to cite this preprint.

Zhu X, Ching T, Pan X, Weissman SM, Garmire L. 2017. Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. PeerJ 5:e2888 https://doi.org/10.7717/peerj.2888

PeerJ Preprints NOT PEER-REVIEWED

Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization

Xun Zhu[1, 2], Travers Ching[1, 2], Xinghua Pan[3], Sherman Weissman[3], Lana Garmire[2, *]

[1] Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa, Honolulu, Hawaii, United States of America

[2] Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, United States of America

[3] Department of Genetics, Yale University, New Haven, Connecticut, United States of America

[*] Corresponding author.

Email address: lgarmire@cc.hawaii.edu

PeerJ Preprints NOT PEER-REVIEWED

Abstract

Single-cell RNA-Sequencing (scRNA-Seq) is a cutting edge technology that enables the understanding of biological processes at an unprecedentedly high resolution. However, well suited bioinformatics tools to analyze the data generated from this new technology are still lacking. Here we have investigated the performance of non-negative matrix factorization (NMF) method to analyze a wide variety of scRNA-Seq data sets, ranging from mouse hematopoietic stem cells to human glioblastoma data. In comparison to other unsupervised clustering methods including K-means and hierarchical clustering, NMF has higher accuracy even when the clustering results of K-means and hierarchical clustering are enhanced by t-SNE. Moreover, NMF successfully detect the subpopulations, such as those in a single glioblastoma patient. Furthermore, in conjugation with the modularity detection method FEM, it reveals unique modules that are indicative of clinical subtypes. In summary, we propose that NMF is a desirable method to analyze heterogeneous single-cell RNA-Seq data, and the NMFEM pipeline is suitable for modularity detection among single-cell RNA-Seq data.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1839v2 | CC-BY 4.0 Open Access | rec: 9 Mar 2016, publ: 9 Mar 2016

PeerJ Preprints NOT PEER-REVIEWED

Introduction

The advancement of technologies has enabled researchers to separate individual cells from a bulk and sequence their transcriptomes at the single cell level, known as single-cell RNA-Sequencing (scRNA-Seq). This technology has reached an unprecedented fine resolution to reveal the program of gene expression within cells(Kumar et al., 2014). It was used to detect heterogeneity within the cell population, and it has greatly enhanced our understanding of the regulatory programs involved in systems such as glioblastoma(Patel et al., 2014), neuronal cells(Usoskin et al., 2014), or pluripotent stem cells (PSCs)(Kumar et al., 2014). It was also used to delineate cell types and subpopulations in differentiating embryonic cells(Treutlein et al., 2014). Other applications include uncovering multilineage priming processes involved in the initial organogenesis(Brunskill et al., 2014), and substantiating the hypothesis of inter-blastomere differences in 2- and 4-cell mouse embryos(Biase, Cao & Zhong, 2014). Indeed, ScRNA-Seq has already made profound impacts on our understanding of the diversity, complexity, and irregularity of biological activities in cells. It will continue to provide more transformative insights in the near future(Pan, 2014).

However, relative to the experimental technology, the bioinformatics tools to analyze scRNA-Seq data are still lagging behind. Given the large amount of noise in the scRNA-Seq data, it is unclear if the tools developed for population-level RNA-Seq differential expression analysis, such as DESeq2(Love, Huber & Anders, 2014) and EdgeR(Robinson, McCarthy & Smyth, 2010), are desirable to identify subpopulations in scRNA-Seq data. Recently, a couple of methods have been reported in the scRNA-Seq analysis domain (Brennecke et al., 2013; McDavid et al., 2013; Kharchenko, Silberstein

& Scadden, 2014). For example, a statistical variance model based on gamma distribution was developed to account for the high technical noise occurring in scRNA-seq experiments, such that genes with high squared correlation of variations (CV^2) relative to mean expression are identified as “significantly differentially expressed” between two conditions(Brennecke et al., 2013). Another Bayesian approach was proposed for scRNA-Seq differential expression analysis, by utilizing a probabilistic model of expression-magnitude distortions that commonly observed in noisy single-cell experiments(Kharchenko, Silberstein & Scadden, 2014). This method later was used for classification of sensory neurons using scRNA-Seq(Usoskin et al., 2014). On the other hand, an R package Monocle was developed recently for single-cell lineage construction(Trapnell et al., 2014). However, it is not clear if all these new methods are suitable for detecting subpopulations in single cells. Moreover, none of the packages mentioned above offers functionalities for modularity identification. For the purpose of network module detection, one has to either use the RNA-Seq transcriptome data as the input for packages such as Module Networks in Genomica(Segal et al., 2003), or use the discovered important genes as seeds to combine with other downstream module detection packages. The fast accumulation of scRNA-Seq data requires new tools to study single-cell transcriptome more efficiently.

Previously, NMF has been applied to other areas in computational biology, such as molecular pattern discovery(Brunet et al., 2004), class comparison and prediction(Gao & Church, 2005), cross-platform and cross-species analysis(Tamayo et al., 2007), and identify subpopulations of cancer patients with mutations in similar network regions. Moreover, NMF has been applied to gene expression profiling studies, in both array(Qi et al., 2009) and population-level RNA-Seq platforms(Brunet et al., 2004). Compared to other methods, it showed multiple advantages, such as less sensitivity to a priori selection of genes or initial conditions and the ability to detect context-dependent patterns of gene expression(Rajapakse, Tan & Rajapakse, 2004). Based on these properties, we hypothesize that NMF is less prone to the influence of noise in the scRNA-Seq data, and thus it can detect a group of genes that robustly differentiate single cells from different conditions. In this report, we demonstrate the capabilities of NMF in scRNA-Seq data analysis in these following aspects: (1) accurate clustering of single cells from different conditions in an unsupervised manner; (2) stratification of subpopulations within the same pool of single cells; (3) detection of meaningful genes, pathways and modules associated with differences among populations and subpopulations. We also combine NMF with the modified, seed based module detection tool Functional Epigenetic Modules (FEM)(Jiao, Widschwendter & Teschendorff, 2014), and provide the scientific community with a streamlined modularity detection R package called NMFEM.

Results

The workflow for a typical single-cell analysis using NMF is shown in Fig. 1. Briefly, the pipeline can take raw reads in FastQ files, align and count them to the RefSeq transcriptome, or use raw count data directly as the input matrix. The input data matrix is then subject to quality control and normalization steps. The normalized matrix is operated on by NMF, which clusters the samples into sub-populations and enlists the feature genes that separate the sub-populations. In order to display the insightful biological modules, the feature genes are then used as the seeds for a functional modularity detection algorithm FEM(Jiao, Widschwendter & Teschendorff, 2014), which identifies hotspots in

PeerJ Preprints NOT PEER-REVIEWED

the interactome with the scRNA-Seq profiling. We applied this workflow to four scRNA-Seq data sets, varying from mouse hematopoietic stem cells to human glioblastoma primary cancer cells.

NMF accurately clusters RNA-Seq data from hematopoietic stem cell differentiation

We first compared the accuracies of NMF in unsupervised clustering, compared to two other commonly used methods: K-means and hierarchical clustering (Hclust) algorithms. We tested these clustering methods on a data set composed of mouse hematopoietic stem cells (HSCs) and stage 1 multipotent progenitor cells (MPP1). These cells were classified using the combined CD62L and CD97 cell surface markers. In order to evaluate the performance of the clustering methods, we removed the cell surface marker based labels. As shown in the PCA plots in Fig. 2A, NMF is the most accurate method, while K-means and hierarchical clustering are much worse. These observations can be quantitatively supported by the results of pairwise Rand measure, a metric that describes the percentage of agreement on a pair of samples belonging to the same group (Fig. 2C). Even though the two cell types are closely related on cell lineage, NMF achieves an overall impressive Rand measure of 83.6% to classify RNA-Seq data by patient ID. In contrast, K-means and hierarchical clustering have much lower Rand measures of 50.6% and 49.7%, respectively (Fig. 2C). Additionally, we plotted the consensus heatmaps of two of the methods — NMF and K-means, which clearly shows the higher accuracy of NMF over K-means (S1 Fig.).

Next we investigated the effect of t-SNE modification on NMF, K-means and hierarchical clustering (Fig. 2B). t-SNE is a dimension reduction method that works by

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1839v2 | CC-BY 4.0 Open Access | rec: 9 Mar 2016, publ: 9 Mar 2016

minimizing the KL-divergence between the distribution of original distances and the distances in the lower-dimensional space. Methods such as K-means are usually conjugated with t-SNE(Van der Maaten & Hinton, 2008) to improve the accuracy of clustering and to be used as a method of visualization in 2-dimensional space(Van der Maaten & Hinton, 2008; Bushati et al., 2011; Junker et al., 2014). However, since NMF is not a distance-based method, applying t-SNE does not improve rather worsen the clustering results of NMF (Fig. 2B and 2C). With the two features extracted by t-SNE, NMF loses its ability to extract meta-genes and to conduct component decomposition, as demonstrated by the clustering accuracy (measured by Rand measure) before and after using t-SNE. On the contrary, K-means and hierarchical clustering have improved accuracies after the application of t-SNE (Fig. 2B and 2C). However, since the differences between HSC vs. MPP1 are very subtle, the ability of t-SNE to improve the clustering accuracy is limited (Fig. 2C).

We repeated the same analytical comparisons with another set of dendritic cell differentiation data(Schlitzer et al., 2015), and obtained similar conclusion. That is, NMF has better accuracy than distance-based methods such as K-means and hierarchical clustering, even when the other two methods are boosted by t-SNE (S2 Fig.).

NMF discovers uniquely important genes in mouse embryonic lung distal epithelium development

Unlike other conventional differential expression test methods that explicitly model the relationships between the variance and mean in the RNA-Seq data, NMF selects the important genes by Kullback–Leibler divergence (KL-divergence)(Yang et al., 2011). Note, these “important genes” are by no means “differentially expressed (DE) genes”, as defined by the differential gene expression (DGE) statistical tests. For comparison, we include the recently developed methods for single-cell transcriptome analysis, including Monocle(Trapnell et al., 2014), MAST(McDavid et al., 2013) as well as SCDE(Kharchenko, Silberstein & Scadden, 2014), as well as DESeq2 and EdgeR, two commonly used differential gene selection methods for the bulky RNA-Seq data. We chose another set of mouse embryonic lung distal epithelial cells reported by Treutlein et al.(Treutlein et al., 2014), and focus on the single cells from E14.5 and E16.5 stages, where the RNA-Seq data are so similar that even PCA analysis cannot separate clearly (S3 Fig.). Given that rich experiential knowledge has been accumulated on their developmental process, this dataset allows us to empirically evaluate the results obtained from different RNA-Seq analysis tools.

We present the characteristics of “important genes” detected by each method in the MA-plots (Fig. 3). The uniquely identified genes from these methods vary greatly (Fig. 3 and S4 Fig. A). In contrast with all other compared methods, NMF selects genes that are sufficiently expressed in many samples, with a strong preference to select genes around a specific expression level (FPKM 2.740) and but not genes expressed too lowly or too highly (S4 Fig. A). On the other hand, a fair amount of genes selected by MAST, SCDE, and Monocle have very little numerical differences between E14.5 and E16.5 stages. A considerable amount of genes selected by DESeq2 and EdgeR have average low expressions but large variance (Fig. 3). Many of them have zero count in all samples of E16.5 stage. Since lowly expressed genes usually have much higher levels of noise, this suggests that DESeq2 and EdgeR may have detected the expression patterns that are less reliable(Brennecke et al., 2013).

PeerJ Preprints NOT PEER-REVIEWED

Such a group of intermediately expressed genes identified by NMF are robust and unlikely a random sample from all expressed genes, since the density distribution of the top 500 genes in NMF per drop-one-out resampling is clearly distinctive from that of random background gene expression (S4 Fig B). The reason that NMF tends to avoid the extremely lowly expressed genes is that KL-divergence intrinsically penalizes lowly expressed genes as A_{ij} can be seen as the weight of (log $\left(\frac{A_{ij}}{(WH)_{ij}}\right)$) in the formula (see Methods). The lower the original expression level, the weaker that gene can affect the clustering, and thus less likely to be selected as a feature gene by NMF. On the other hand, the highly expressed genes typically have extreme spikes among a few samples, and are also less likely to be selected as feature genes, as the signal linearity of NMF prefers to select genes with consistent expression levels in each cluster.

Important genes selected by NMF yield biologically meaningful modules

We next asked if the important genes detected by NMF convey unique and meaningful biological functions. Towards this, we examined the modularity potentials and used the same number of 500 top genes selected by the eight methods above as the initial seeds for the module detection software FEM(Jiao, Widschwendter & Teschendorff, 2014). FEM is a versatile method that can be adapted to identify hotspots in the interactome with the differential expression profiling, using the seed inputs from external programs including NMF, DESeq2, EdgeR, MAST, SCDE, or Monocle. We present the results of the top 5 most significant modules for each of the eight methods. Within each top module, we conducted Gene Ontology (GO) enrichment analysis and list the top two GO terms (Table 1).

In comparison, the methods that are established on similar assumptions have higher degrees of agreements on the detected top modules (Table 1) as well as genes in common (S5 Fig.), as expected. For examples, SCDE, MAST and Monocle have more similar results than others; whereas DESeq2 and EdgeR tend to agree to each other better since they were designed for bulky cell RNA-Seq. Interestingly, all methods except EdgeR, detected that the transcription-related processes play important role from E14.5 to E16.5. NMF finds two unique modules for “mRNA destabilization” (seed gene Pnn) and “rRNA processing” (seed gene exosc4) (Table 1 and Fig. 4). These results are very interesting as mRNA-destabilizing inflammatory RNA-binding proteins were previous reported to be important in the regulation of miR-155 biogenesis in lung epithelial cells with cystic fibrosis condition(Bhattacharyya et al., 2013). Exosc4 is part of the exosome complex, which has the function of degrading various types of RNA molecules. Since E14.5 cells are prior to sacculation and E16.5 cells are in the early stage of sacculation, the exosc4-centered module may indicate the fast turnover of RNA material associated with the cell growth/apoptosis activities in the process of embryonic lung morphological changes.

Additionally, NMF identifies a module related to “G-protein coupled receptor signaling pathway” (seed gene Gna13), which is also shared by DESeq2 and EdgeR methods (Table 1 and Fig. 4). This may indicate active intracellular signal changes during the early phase of embryonic lung epithelial cells. This observation is coherent with another unique module found by NMF, which is related to bone morphogenetic protein (BMP) pathway (seed gene Smad4). BMP pathway previously was verified to have important roles in signal transduction, transcription and adhesion in epithelial bud development, including lung epithelial cells(Jamora et al., 2003). Moreover, BMPs play important roles in different stem cell systems, including embryonic stem cells(Zhang & Li, 2005).

In summary, due to the mechanism of identifying correlated genes rather than genes with numerical differences, NMF is able to extract very unique biological information from different classes of single cells.

NMF identifies tumor sub-populations among a single glioblastoma patient

Detecting the subpopulations of single cells within the same bulk is an even subtler problem, in comparison to the issue of accurate clustering of mixed populations. To examine the potential of NMF in this aspect, we next tested the scRNA-Seq data from the five individual glioblastoma patients as reported by Patel, AP et al.(Patel et al., 2014) Interestingly, the consensus clustering results generated from NMF show that among the five patients, only patient MGH28 (Fig. 5A-B) and MGH31 (S6 Fig. A-B) have two distinct subpopulations.

To investigate further the characteristics of the two subpopulations in MGH28, we retrieved the top 500 ranked genes that differentiate these two subpopulations and conducted KEGG pathway enrichment analysis on them. A pathway named “pathogenic Escherichia coli infection” stands out as the most significantly altered pathway between the two subpopulations (FDR < 1E-03) (Fig. 5C). Further examination of this pathway reveals that multiple genes involved in cell mobility are enriched, including ACTG1, ACTB, CTTN, YWHAZ, CDC42, TUBB, RHOA, ROCK, ARPC5, TUBA1A, NCL, TUBA1B, and TUBA1C (Fig. 5D). Glioblastoma is among the most heterogeneous tumors in human, and mainly have pro-neuron and mesenchymal phenotypes. The latter is associated with more invasive and infiltrating phenotype. Our results indicate that some cells in patient MGH28 have mesenchymal phenotype. Coincidently, Patel, AP et al also concluded MGH28 as mesenchymal glioblastoma, by comparing the scRNA-Seq signatures to those from TCGA glioblastoma RNA-Seq data(Patel et al., 2014).

Interestingly, we also found that patient MGH31 has the same enriched KEGG pathway term of "pathogenic Escherichia coli infection" (S6 Fig. C). Almost all of the important genes in this pathway from patient MGH31 (S6 Fig. D) overlap those from patient MGH28 mentioned above (Fig. 5D). The only exceptions are NCL unique to MGH28, and CDC42 and ROCK2 unique to MGH31. The almost identical genes found in the same pathway that differentiates the subpopulations of both MGH28 and MGH31 suggest that MGH31 may also be classified as mesenchymal glioblastoma, similar to MGH28.

Discussion and conclusions

Due to the high noise levels within scRNA-Seq data(Brennecke et al., 2013), the conventional approaches, which aim to detect numerical differences of gene expression in cell bulks under different conditions, may be limited. Previous applications of NMF to fields such as face reorganization(Rajapakse, Tan & Rajapakse, 2004), image compression(Yuan & Oja, 2005; Monga & Mıhçak, 2007) and sound decomposition(Smaragdis, 2004), have proven successful. Here we propose to utilize NMF as a desirable method for scRNA-Seq analysis. We believe that the pattern based feature extraction ability of NMF can meet the demands to identify genes that signify the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1839v2 | CC-BY 4.0 Open Access | rec: 9 Mar 2016, publ: 9 Mar 2016

differences within the noisy scRNA-Seq data. The in-depth analyses on multiple public and private data sets in this study have provided supports from several aspects.

We have demonstrated that NMF performs well relative to other popular clustering methods including K-means and hierarchical clustering, even when these methods in comparisons are boosted with t-SNE. Moreover, NMF is capable of identifying subpopulations within the same tumor sample, exemplified by the glioblastoma data here. Through NMF clustering, we found in that patients MGH28 and MGH31 both have a group of genes that can distinguish the single cells into two subpopulations. These genes include actins, tubulins and signaling molecules that can affect cell mobility. Thus we speculate that both MGH28 and MGH31 have mesenchymal phenotypes. The suspected mesenchymal phenotype of MGH28 from scRNA-Seq data alone is directly supported by Patel, AP et al.(Patel et al., 2014), where they used TCGA glioblastoma data and classified MGH28 as the mesenchymal type. On the other hand, the authors could not clearly classified MGH31 as the mesenchymal type, although they suspected two genetic clones from this patient. Here with NMF based subpopulation identification and comparisons of characteristic genes, our analysis confirms the existence of two subpopulations and further, the clinical subtype of MGH31.

In summary, we have demonstrated that NMF is a desirable method capable of accomplishing various tasks in scRNA-Seq data analysis, from reclassifying populations of single cells, identifying subpopulations, to revealing meaningful genes, gene sets and modules of biological significance. We expect the new workflow named NMFEM to have wide applications in the field of scRNA-Seq bioinformatics analysis.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1839v2 | CC-BY 4.0 Open Access | rec: 9 Mar 2016, publ: 9 Mar 2016

PeerJ Preprints NOT PEER-REVIEWED

Methods

Data sets

Glioblastoma

ScRNA-Seq data were retrieved from the original 875 samples of glioblastoma tumor cells in 5 patients, along with population and cell line controls (GSE57872)(Patel et al., 2014). For NMF, very minimal filtering was employed (filtering steps of other methods are detailed in a later section). First, genes with zero expression across all samples were removed so that 22704 out of 23710 genes (95.8%) remained. Next the smallest number of samples was removed so that at least one gene was expressed across all samples considered, as a quality requirement of DESeq2. As a result, 124 samples with the lowest amount of non-zero expression across all genes are removed, leaving 751 of 875 samples (85.8%).

Mouse lung epithelial cells

ScRNA-Seq data were retrieved from the original 201 samples of lung distal epithelial cells of embryonic mouse (GSE52583)(Treutlein et al., 2014). We filtered genes and samples following the sample procedure as in Glioblastoma data set, leaving 16168 out of 23420 genes (69.0%) and 199 out of 201 samples (99.0%).

Mouse HSCs and MPP1s

ScRNA-Seq data were extracted from mouse hematopoietic stem cells (HSCs) and early multipotent progenitors (MPP1s). The data were pre-processed into the format of a FPKM expression profile, which include 59 HSCs and 53 MPP1 single cells. We filtered genes and samples following the sample procedure as in Glioblastoma data set, leaving 12719 out of 21664 genes (58.7%) and 112 out of 112 samples (100.0%).

Mouse dendritic cells

ScRNA-Seq data were extracted from mouse macrophage DC progenitors (MDPs), common DC progenitors (CDPs), and Pre-DCs (GSE60781)(Schlitzer et al., 2015). We used the RPKM table provided by the authors. We filtered genes and samples following the same procedure as in Glioblastoma data set, leaving 15722 out of 29779 genes (52.8%) and 251 out of 251 samples (100.0%).

Single-cell RNA-Seq analysis

Read alignment

We downloaded the public datasets from NCBI The Gene Expression Omnibus (GEO) database(Edgar, Domrachev & Lash, 2002; Barrett et al., 2013), and retrieved the SRA files from The Sequence Read Archive (SRA)(Leinonen et al., 2011). We used the fastq-dump tool from SRA Toolkit to convert the SRA files into two pair-end FastQ files. We applied FastQC for quality control and Tophat2(Kim et al., 2013) for alignment to the reference genomes. The ready-to-use genome sequences and annotation files were downloaded from Illumina iGenomes page (http://support.illumina.com/sequencing/sequencing_software/igenome.html). For human build hg19 was used, and for mouse genome build mm10 was used(Karolchik et al., 2014).

PeerJ Preprints NOT PEER-REVIEWED

Read Counting

We used featureCounts(Liao, Smyth & Shi, 2014) to map and count the aligned BAM files to the RefSeq transcriptomes from the pre-built packages on Illumina iGenome website above. We used the options to count fragments instead of reads; paired-end distance was checked by featureCounts when assigning fragments to meta-features or features. We only took into account of fragments that have both ends aligned successfully and discarded chimeric fragments. Fragments mapped to multiple locations were counted. The command is "featureCounts -pPBCM --primary -T 6 -a <gtf_file> -o <output_file> <bam_file>".

Normalization of Counts

We used reads per kilo base per million (RPKM) to represent the gene expression level, where the length of each gene was calculated by UCSC RefSeq annotation table, by concatenating all the exons. We normalized the data using DESeq2.

Non-negative Matrix Factorization (NMF)

We used the R-package implementation of NMF(Gaujoux & Seoighe, 2010) to perform NMF analysis. NMF is mathematically approximated by: $A \approx WH$, where A (n by m) is the matrix representing the scRNA-Seq profile in this report, W is a slim weight matrix (n by k, where $n \gg k$), H is a wide matrix (k by m, where $m \gg k$), and all three of them are non-negative(Brunet et al., 2004).The column vectors in W are called meta-genes, which are higher-level abstraction of the original gene expression pattern. We used the method "*brunet*" to solve the approximation of A, which employs the multiplicative iterative algorithm described by the following rules:

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1839v2 | CC-BY 4.0 Open Access | rec: 9 Mar 2016, publ: 9 Mar 2016

PeerJ Preprints NOT PEER-REVIEWED

$$H_{au} \leftarrow H_{au} \frac{\sum_i \frac{W_{ia}V_{iu}}{(WH)_{iu}}}{\sum_k W_{ka}}$$

$$W_{ia} \leftarrow W_{ia} \frac{\sum_u \frac{H_{au}A_{iu}}{(WH)_{iu}}}{\sum_v H_{av}}.$$

The initialization of H_{au} and W_{ia} was generated as random seed matrices drawn from a uniform distribution within the same range as the entries in the matrix A. Since the starting matrices were randomized, we conducted an average of 30 simulations for each NMF run to obtain the consensus clustering results. We used Kullback–Leibler divergence (KL-divergence) as the distance function, as it has significantly better performance theorized in Yang et al.(Yang et al., 2011). The rank (k) is chosen by listing the clustering results of all possible k's (usually ranging from 2 to 5, as higher k values requires exponentially more time to run). k is chosen when the best cophenetic correlation coefficient is achieved, as proposed in Brunet et al. 2004(Brunet et al., 2004).

NMF package uses the *feature score* to measure the genes for different expression between sample groups, based on a method described in Kim et al.(Kim et al., 2013)

$$\text{FeatureScore}(i) = 1 + \frac{1}{\log_2 k} \sum_{q=1}^{k} p(i,q) \log_2 \, p(i,q),$$

where

$$p(i,\Omega) = \frac{W(i,\Omega)}{\sum_{q=1}^{k} W(i,q)}.$$

The feature score lies between 0 and 1, and is positively related to its factor-specificity. That is, a higher feature score indicates that the gene has more different expression patterns between sample groups (phenotypes)(Kim & Park, 2007). We select the top 500 genes of NMF based on this feature score.

Other packages used for detecting significant or important genes

We compared a series of computational methods to call “significant genes” with NMF. These methods are divided into three categories.

DE methods for bulky-level RNA-Seq: we used two most popular bulky-level RNA-Seq methods: DESeq2 and EdgeR, to compare on the results of DE genes.

DE methods for scRNA-Seq: three methods were investigated, with default settings of the packages. (1) Monocle: this is a versatile method (V. 1.0.0) that performs differential expression analysis between cell types or states, moreover places cells in order according to their progression through processes such as cell differentiation(Trapnell et al., 2014). (2) SCDE: this package (V 1.2.1) implemented in R is based on Bayesian method, where the individual genes were modeled explicitly as a mixture of the dropout and amplification events by the Poisson model and negative binomial model(Kharchenko, Silberstein & Scadden, 2014). (3) MAST: this method (V 1.0.1) implemented in R was originally used to detected DE genes in qPCR results of single cells. We selected the 500 genes with the lowest likelihood ratio test p-value using Hurdle Model provided by the package, as recommended by the authors(McDavid et al., 2013).

Data filtering for other scRNA-Seq methods: SCDE model deals with high level noise automatically and requires no filtering as stated by authors. For Monocle and MAST, we first removed the genes of high technical variations using the method as described in Brennecke et al. 2013(Brennecke et al., 2013), then performed filtering steps as instructed

PeerJ Preprints NOT PEER-REVIEWED

in each paper. Monocle filters out libraries that contained fewer than 1 million reads in its original report, in the case that reads in some data set do not meet this threshold (such as mouse embryonic lung epithelial cell data), we resorted to no sample filtering to be safe.

Additionally, we experimented if introducing t-SNE, a dimension reduction method that was recently successfully applied to scRNA-Seq, would improve the results of NMF. We used the C++ accelerated R-package Rtsne (V 0.10), based on the original C++ implementation by van der Maaten et al.(van der Maaten, 2013)

Module detection package

We use Functional Epigenetic Modules (FEM) R package(Jiao, Widschwendter & Teschendorff, 2014) for module detection. FEM utilizes an expansion algorithm based on the z-score of the expression level, by using a list of seed genes as the starting points. It selects the top modules based on p-values calculated by a Monte Carlo method.

We modified the source code of the FEM package and changed the process of the seed gene selection. Instead of selecting the seed genes based on the z-score of the expression level, we directly plugged in a list of genes as the seed genes, which were generated from each of the compared method for important gene detection.

Measuring the performance of unsupervised clustering methods

Label assignments for PCA/t-SNE plots

Since multiple assignments of labeling to clusters are possible, for each clustering algorithm we iterated through all possible permutations of labeling and calculated the accuracy for each. The one with the best accuracy rate is picked as the *most favorable labeling* for the clustering algorithm and is used in plotting its PCA/t-SNE scatter-plots.

Confusion matrix

Confusion matrix C was calculated by the following formula:

$$C_{i,j} = |A_i \cap B_j|,$$

Where A_i is the set of samples that are labeled as class i according to the correct labelling, and B_j is the set of samples that are labeled as class j in the tested method(Stehman, 1997).

Chi-square test score

Chi-square test score S_{χ^2} was calculated from the chi-square test p-value p_{χ^2},

$$S_{\chi^2} = \log_{0.05} p_{\chi^2},$$

which in turn was calculated by the *chisq.test* function in R(Aguirre & Nikulin, 1994).The base of 0.05 was chosen so that a score larger than one indicates that the resulting p-value is significant.

Pair-wise Rand measure

Pair-wise Rand measure for clustering between the test and the reference is defined by

$$R = \frac{TP + TN}{TP + FP + FN + TN},$$

in which the four quantities TP, FP, FN, and TN are cardinals of the four sets of pairs. T/F means true/false based on the reference, and P/N means positive/negative results from the test. Specifically, a positive result (P) refers to a pair of samples clustered in the same group by the tested method; a true positive (TP) or true negative (TN) result represents the case where the agreements between the test and the reference clustering is reached(Rand, 1971).

Modularity detection and pathway Analysis

We used Functional Epigenetic Modules (FEM) package(Jiao, Widschwendter & Teschendorff, 2014) implemented in R for module detection. FEM utilizes SpinGlass algorithm(Reichardt & Bornholdt, 2006) based on the z-score of the expression level, by using a list of seed genes as the starting points. It selects the top modules based on p-values calculated from a Monte Carlo method. We modified the source code of the package to allow seed genes generated from other methods (NMF, DESeq2, EdgeR, SCDE, MAST and Monocle) that detect significant or important genes. In each case, we used top 500 most important genes as the seeds for FEM. We next compared biological meanings of the resulting modules by Gene Ontology (GO) or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis, implemented as DAVID Web Service in R(Huang, Sherman & Lempicki, 2008, 2009).

Data and code availability

The Glioblastoma, mouse lung distal epithelial and mouse dendritic cell data are downloaded from GSE57872, GSE52583, and GSE60781. The code used for this package can be found at https://github.com/lanagarmire/NMFEM, and https://github.com/lanagarmire/NMFEM_extra.

Author contributions

LXG envisioned the project. XZ conducted the data analysis, with assistance from TC. XP and SW communicated on bioinformatics analysis and provided the mouse HSC and MPP1 scRNA-Seq data. XZ and LXG wrote the draft. All authors have read, reviewed and agreed on the manuscript.

Acknowledgement

This research was supported by grants K01ES025434 awarded by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), P20 COBRE GM103457 awarded by NIH/NIGMS, and Medical Research Grant 14ADVC-64566 from Hawaii Community Foundation to L.X. Garmire.

Competing interests

The authors declare that they have no competing interests.

References

Aguirre N., Nikulin M. 1994. Chi-squared goodness-of-fit test for the family of logistic distributions. *Kybernetika* 30:214–222.

Barrett T., Wilhite SE., Ledoux P., Evangelista C., Kim IF., Tomashevsky M., Marshall KA., Phillippy KH., Sherman PM., Holko M., Yefanov A., Lee H., Zhang N., Robertson CL., Serova N., Davis S., Soboleva A. 2013. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* 41 :D991–D995. DOI: 10.1093/nar/gks1193.

PeerJ Preprints NOT PEER-REVIEWED

Bhattacharyya S., Kumar P., Tsuchiya M., Bhattacharyya A., Biswas R. 2013. Regulation of miR-155 biogenesis in cystic fibrosis lung epithelial cells: Antagonistic role of two mRNA-destabilizing proteins, KSRP and TTP. *Biochemical and Biophysical Research Communications* 433:484–488. DOI: 10.1016/j.bbrc.2013.03.025.

Biase F., Cao X., Zhong S. 2014. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome research*:gr–177725.

Brennecke P., Anders S., Kim JK., Kołodziejczyk AA., Zhang X., Proserpio V., Baying B., Benes V., Teichmann SA., Marioni JC. 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods*.

Brunet J-P., Tamayo P., Golub TR., Mesirov JP. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences* 101 :4164–4169.

Brunskill EW., Park J-S., Chung E., Chen F., Magella B., Potter SS. 2014. Single cell dissection of early kidney development: multilineage priming. *Development* 141:3093–3101.

Bushati N., Smith J., Briscoe J., Watkins C. 2011. An intuitive graphical visualization technique for the interrogation of transcriptome data. *Nucleic acids research* 39:7380–9. DOI: 10.1093/nar/gkr462.

Edgar R., Domrachev M., Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* 30 :207–210. DOI: 10.1093/nar/30.1.207.

Gao Y., Church G. 2005. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics* 21:3970–3975.

Gaujoux R., Seoighe C. 2010. A flexible R package for nonnegative matrix factorization. *BMC bioinformatics* 11:367.

Huang DW., Sherman BT., Lempicki RA. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* 4:44–57.

Huang DW., Sherman BT., Lempicki RA. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* 37:1–13.

Jamora C., DasGupta R., Kocieniewski P., Fuchs E. 2003. Links between signal transduction, transcription and adhesion in epithelial bud development. *Nature* 422:317–322. DOI: 10.1038/nature01458.

Jiao Y., Widschwendter M., Teschendorff AE. 2014. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics*:btu316.

Junker JP., Noël ES., Guryev V., Peterson KA., Shah G., Huisken J., McMahon AP., Berezikov E., Bakkers J., van Oudenaarden A. 2014. Genome-wide RNA Tomography in the Zebrafish Embryo. *Cell* 159:662–675. DOI: 10.1016/j.cell.2014.09.038.

Karolchik D., Barber GP., Casper J., Clawson H., Cline MS., Diekhans M., Dreszer TR., Fujita PA., Guruvadoo L., Haeussler M. 2014. The UCSC genome browser database: 2014 update. *Nucleic acids research* 42:D764–D770.

Kharchenko P V., Silberstein L., Scadden DT. 2014. Bayesian approach to single-cell differential expression analysis. *Nature methods* 11:740–742.

Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R., Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14:R36.

Kim H., Park H. 2007. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* 23:1495–1502.

Kumar RM., Cahan P., Shalek AK., Satija R., DaleyKeyser AJ., Li H., Zhang J., Pardee K., Gennert D., Trombetta JJ., Ferrante TC., Regev A., Daley GQ., Collins JJ. 2014. Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* 516:56–61. DOI: 10.1038/nature13920.

Leinonen R., Sugawara H., Shumway M., Collaboration on behalf of the INSD. 2011. The Sequence Read Archive. *Nucleic Acids Research* 39:D19–D21. DOI: 10.1093/nar/gkq1019.

Li Y., Ngom A. 2013. The non-negative matrix factorization toolbox for biological data mining. *Source code for biology and medicine* 8:1–15.

Liao Y., Smyth GK., Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923–930.

Love MI., Huber W., Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*.

van der Maaten L. 2013. Barnes-hut-sne. *arXiv preprint arXiv:1301.3342*.

Van der Maaten L., Hinton G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9:85.

McDavid A., Finak G., Chattopadyay PK., Dominguez M., Lamoreaux L., Ma SS., Roederer M., Gottardo R. 2013. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics (Oxford, England)* 29:461–7. DOI: 10.1093/bioinformatics/bts714.

Monga V., Mıhçak MK. 2007. Robust and secure image hashing via non-negative matrix factorizations. *Information Forensics and Security, IEEE Transactions on* 2:376–390.

Pan X. 2014. Single Cell Analysis: From Technology to Biology and Medicine. *Single cell biology* 3:106. DOI: 10.4172/2168-9431.1000106.

Patel AP., Tirosh I., Trombetta JJ., Shalek AK., Gillespie SM., Wakimoto H., Cahill DP., Nahed B V., Curry WT., Martuza RL., Louis DN., Rozenblatt-Rosen O., Suvà ML., Regev A., Bernstein BE. 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344:1396–1401. DOI: 10.1126/science.1254257.

Qi Q., Zhao Y., Li M., Simon R. 2009. Non-negative matrix factorization of gene expression profiles: a plug-in for BRB-ArrayTools. *Bioinformatics* 25:545–547.

Rajapakse M., Tan J., Rajapakse J. 2004. Color channel encoding with NMF for face recognition. In: *Image Processing, 2004. ICIP'04. 2004 International Conference on*. IEEE, 2007–2010.

Rand WM. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 66:846–850.

Reichardt J., Bornholdt S. 2006. Statistical mechanics of community detection. *Physical Review E* 74:16110.

Robinson MD., McCarthy DJ., Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140.

Schlitzer A., Sivakamasundari V., Chen J., Sumatoh HR Bin., Schreuder J., Lum J., Malleret B., Zhang S., Larbi A., Zolezzi F. 2015. Identification of cDC1-and cDC2-committed DC progenitors reveals early lineage priming at the common DC progenitor stage in the bone marrow. *Nature immunology* 16:718–728.

Segal E., Shapira M., Regev A., Pe'er D., Botstein D., Koller D., Friedman N. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics* 34:166–176.

Smaragdis P. 2004. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In: *Independent Component Analysis and Blind Signal Separation*. Springer, 494–499.

Stehman S V. 1997. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment* 62:77–89.

Tamayo P., Scanfeld D., Ebert BL., Gillette MA., Roberts CWM., Mesirov JP. 2007. Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proceedings of the National Academy of Sciences* 104:5959–5964.

Trapnell C., Cacchiarelli D., Grimsby J., Pokharel P., Li S., Morse M., Lennon NJ., Livak KJ., Mikkelsen TS., Rinn JL. 2014. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature biotechnology* 32:381.

Treutlein B., Brownfield DG., Wu AR., Neff NF., Mantalas GL., Espinoza FH., Desai TJ., Krasnow MA., Quake SR. 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509:371–375.

Usoskin D., Furlan A., Islam S., Abdo H., Lönnerberg P., Lou D., Hjerling-Leffler J., Haeggström J., Kharchenko O., Kharchenko P V. 2014. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature neuroscience* 18:145–153.

Yang Z., Zhang H., Yuan Z., Oja E. 2011. Kullback-Leibler divergence for nonnegative matrix factorization. In: *Artificial Neural Networks and Machine Learning–ICANN 2011*. Springer, 250–257.

Yuan Z., Oja E. 2005. Projective nonnegative matrix factorization for image compression and feature extraction. In: *Image Analysis*. Springer, 333–342.

Zhang J., Li L. 2005. BMP signaling and stem cell regulation. *Developmental Biology*

284:1–11. DOI: 10.1016/j.ydbio.2005.05.009.

Tables

Table 1. Comparison of the top 5 modules selected by FEM with seed genes generated by NMF and other differential expression detection methods. The other compared methods include MAST, SCDE, Monocle, DESeq2 and EdgeR. GO analysis was performed on each module, and the top 2 most enriched GO terms are listed along with their p-values. Connectivity is computed by taking the average of the degree number of all the nodes in the graph. The p-value for each module was calculated by FEM's internal Monte Carlo procedure.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1839v2 | CC-BY 4.0 Open Access | rec: 9 Mar 2016, publ: 9 Mar 2016

PeerJ Preprints NOT PEER-REVIEWED

Figure legends

Fig. 1: The workflow of NMFEM. The input can be either FastQ files or a raw counts table. If FastQ files are used, they are aligned using TopHat and counted using FeatureCounts (steps shown in brackets). The input or calculated raw counts table are filtered by samples and genes, converted into RPKMs using gene lengths, and normalized by samples. We then run NMF method on them to detect subpopulations, and find the feature genes separating the detected subpopulations. Finally, we feed the feature genes as seed genes in FEM, and generate PPI gene modules that contain highly differentially expressed genes.

Fig. 2: Comparisons among clustering methods on the HSC vs. MPP1 scRNA-Seq data.

(A) The PCA scatter-plots of the samples, based on their log normalized expression level. Colors indicate the most favorable labeling that can be assigned to the clustering result generated by each method. The correctly and incorrectly labeled samples are marked by dot (•) and cross (x), respectively. Confusion matrices of the methods in comparison are inserted on the top-right corner of each sub-panel. The closer the matrix is to a diagonal matrix, the more accurate the method is. (B) The scatter-plots of the samples for K-means and hierarchical clustering, after t-SNE based dimension reduction. (C) Rand measures of the methods in comparison, before and after t-SNE. Rand measure ranges from 0 to 1, where a higher value indicates a greater clustering accuracy.

Fig. 3: MA-plots of significant or important genes defined by different methods.

Shown are scRNA-Seq data in the mouse lung distal epithelial cell E14.5 vs. E16.5

PeerJ Preprints NOT PEER-REVIEWED

samples. The blue color highlights the genes selected as “the most significant” by the corresponding methods. X-axis (A-value) is the mean of the gene expression, and y-axis (M-value) is the difference of the gene expression between E16.5 and E14.5 stages.

Fig. 4: Network of top 5 modules using the seed genes generated by NMF.

Shown are module detection results in the FEM package, using the top 500 most important genes detected by NMF in Fig. 3. ScRNA-Seq data in the mouse lung distal epithelial cell E14.5 vs. E16.5 samples are compared, where the red and green colors indicate up- and down-regulation of genes in E16.5 relative to E14.5, respectively. The top 5 modules are selected by the p-values calculated from the internal Monte-Carlo method in the FEM package (Table 1).

Fig. 5: Using NMF to identify subpopulations in a single glioblastoma tumor from patient MGH28.

(A) The consensus heat map generated from NMF. The two subpopulation clusters are the evident 2 red squares, marked out by number 1 and 2. The brightness indicates the confidence level of two subpopulations. (B) The PCA plot of scRNA-Seq samples from patient MGH28, the discovered subpopulations are coded in red and blue colors. (C) The results of KEGG/BioCarta Pathway enrichment analysis. The line of significance (to the right of which meaning the FDR less than 0.05) is shown. (D) The protein interaction diagram of the KEGG pathway “Pathogenic E. Coli infection”. The proteins coded by the genes detected by NMF are highlighted yellow, with the gene names marked below.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1839v2 | CC-BY 4.0 Open Access | rec: 9 Mar 2016, publ: 9 Mar 2016

Supporting Information

S1 Fig. The consensus map of NMF and K-means methods run on the HSC vs. MPP1 dataset. The columns and rows are samples. The brightness indicates the confidence of the method to assign the samples in the same group.

S2 Fig. (A) comparison of t-SNE two-dimensional scatter-plots of the mouse dendritic cell scRNA-Seq data. Colors indicate the most favorable labeling that can be assigned to the clustering result generated by each method. The correctly and incorrectly labeled samples are marked by dot (•) and cross (x), respectively. (B) Rand measures of the methods in comparison, before and after t-SNE. Rand measure ranges from 0 to 1, where a higher value indicates a greater clustering accuracy.

S3 Fig. PCA plot of the mouse epithelial cell data set. The groups that are most difficult to separate (E14.5 vs. E16.5) are circled out.

S4 Fig. (A) The kernel density estimation (KDE) plot showing the frequency of log expression values of "important genes" that separate E14.5 vs. E16.5, as detected by the various methods in comparison. (B) KDE plot of frequency of genes appear in the 71 Jackknife runs. For a certain x-value (frequency), a higher y-value (density) means that a higher percentage of genes appear around this frequency among the 71 runs. The blue block is the top 500 genes selected by NMF and the red block is all the genes in the filtered data used by NMF.

S5 Fig. The heatmap of the characteristic genes (E14.5 vs. E16.5) found in common pair-wise by the various methods. The dendrogram at the bottom shows the hierarchical clustering results using the distance measured by the inverse of the number of overlapping genes.

S6 Fig. Using NMF to identify subpopulations in a single glioblastoma tumor from Patient MGH31

(A) The consensus heat map generated from NMF. The two subpopulation clusters are the evident 2 red squares, marked out by number 1 and 2. The brightness indicates the confidence level of two subpopulations. (B) The PCA plot of scRNA-Seq samples from patient MGH31, the discovered subpopulations are coded in red and blue colors. (C) The results of KEGG/BioCarta Pathway enrichment analysis. The line of significance (to the right of which meaning the FDR less than 0.05) is shown. (D) The protein interaction diagram of the KEGG pathway “Pathogenic E. Coli infection”. The proteins coded by the genes detected by NMF are highlighted yellow, with the gene names marked below.

PeerJ Preprints NOT PEER-REVIEWED

seed	size	connectivity	p_values	first_term	first_fisher	second_term	second_fisher
NMF							
Gna13	32	4.6875	0.004	G-protein coupled receptor signaling pathway	1.80E-13	semaphorin-plexin signaling pathway	2.50E-13
Med31	73	8.136986301	0.009	stem cell maintenance	1.40E-13	RNA metabolic process	1.90E-13
Smad4	52	4.230769231	0.017	BMP signaling pathway	0.00012	regulation of BMP signaling pathway	0.00031
Exosc4	42	7.857142857	0.022	rRNA catabolic process	1.10E-16	rRNA processing	4.70E-16
Pnn	14	3.857142857	0.023	mRNA destabilization	0.000028	RNA destabilization	0.000059
MAST							
Hdac2	92	5.869565217	0	chromatin organization	6.10E-29	negative regulation of nucleic acid-templated transcription	1.50E-27
Dld	73	8.02739726	0.001	carboxylic acid metabolic process	1.80E-29	oxoacid metabolic process	9.00E-29
Sdhb	33	7.696969697	0.006	aerobic respiration	3.80E-17	tricarboxylic acid cycle	8.10E-17
Ndufv2	24	7.666666667	0.008	oxidation-reduction process	0.000000065	response to protozoan	0.00024
Twistnb	46	13.13043478	0.012	transcription from RNA polymerase III promoter	3.70E-14	nucleobase-containing compound biosynthetic process	6.10E-13
SCDE							
Polr2l	75	12.88	0.002	nucleobase-containing compound biosynthetic process	2.50E-14	aromatic compound biosynthetic process	5.90E-14
Ndufv2	24	7.666666667	0.007	oxidation-reduction process	0.000000065	response to protozoan	0.00024
Sdhb	33	7.696969697	0.008	aerobic respiration	3.80E-17	tricarboxylic acid cycle	8.10E-17
Ldha	33	7.696969697	0.01	aerobic respiration	3.80E-17	tricarboxylic acid cycle	8.10E-17
Polr2b	79	10.75949367	0.014	nucleobase-containing compound biosynthetic process	2.60E-18	transcription, DNA-templated	4.50E-18
Monocle							
Hdac2	92	5.869565217	0	chromatin organization	6.10E-29	negative regulation of nucleic acid-templated transcription	1.50E-27
Rabgap1	10	8	0.005	single-organism catabolic process	0.0014	cellular catabolic process	0.0017
Sdhb	33	7.696969697	0.006	aerobic respiration	3.80E-17	tricarboxylic acid cycle	8.10E-17
Twistnb	46	13.13043478	0.006	transcription from RNA polymerase III promoter	3.70E-14	nucleobase-containing compound biosynthetic process	6.10E-13
Ndufv2	24	7.666666667	0.013	oxidation-reduction process	0.000000065	response to protozoan	0.00024
DESeq2							
Aldh6a1	36	8.111111111	0.003	aerobic respiration	1.00E-16	tricarboxylic acid cycle	2.00E-16
Gfm2	10	8	0.005	single-organism catabolic process	0.0014	cellular catabolic process	0.0017
Polr2l	75	12.88	0.006	nucleobase-containing compound biosynthetic process	2.50E-14	aromatic compound biosynthetic process	5.90E-14
Twistnb	46	13.13043478	0.006	transcription from RNA polymerase III promoter	3.70E-14	nucleobase-containing compound biosynthetic process	6.10E-13
Gna13	32	4.6875	0.008	G-protein coupled receptor signaling pathway	1.80E-13	semaphorin-plexin signaling pathway	2.50E-13
EdgeR							
Aldh6a1	36	8.111111111	0.004	aerobic respiration	1.00E-16	tricarboxylic acid cycle	2.00E-16
Gna13	32	4.6875	0.012	G-protein coupled receptor signaling pathway	1.80E-13	semaphorin-plexin signaling pathway	2.50E-13
Tpr	58	12.24137931	0.016	proteolysis involved in cellular protein catabolic process	5.00E-18	cellular protein catabolic process	1.30E-17
Thbs1	16	3.875	0.017	cell adhesion	0.00001	biological adhesion	0.00001
Por	12	7.333333333	0.018	single-organism catabolic process	0.000018	cellular catabolic process	0.000058

Table 1.Comparison of the top 5 modules selectedby FEM with seed genes generated by NMF and other differential expression detection methods. The other compared methods include MAST, SCDE, Monocle, DESeq2 and EdgeR. GO analysis was performed on each module, and the top 2 most enriched GO terms are listed along with their p-values. Connectivity is computed by taking the average of the degree number of all the nodes in the graph. The p-value for each module was calculated by FEM's internal Monte Carlo procedure.

PeerJ Preprints
NOT PEER-REVIEWED

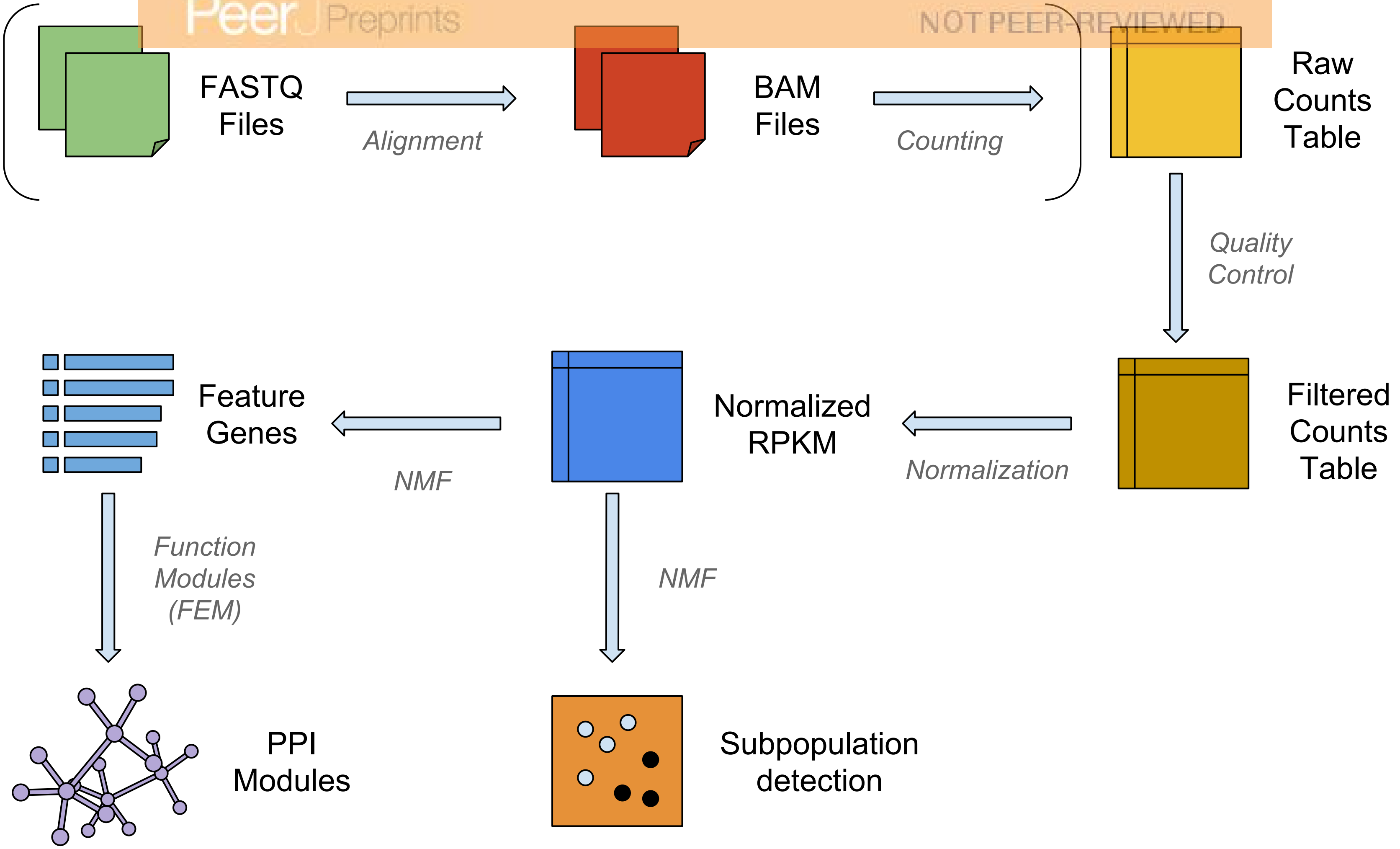
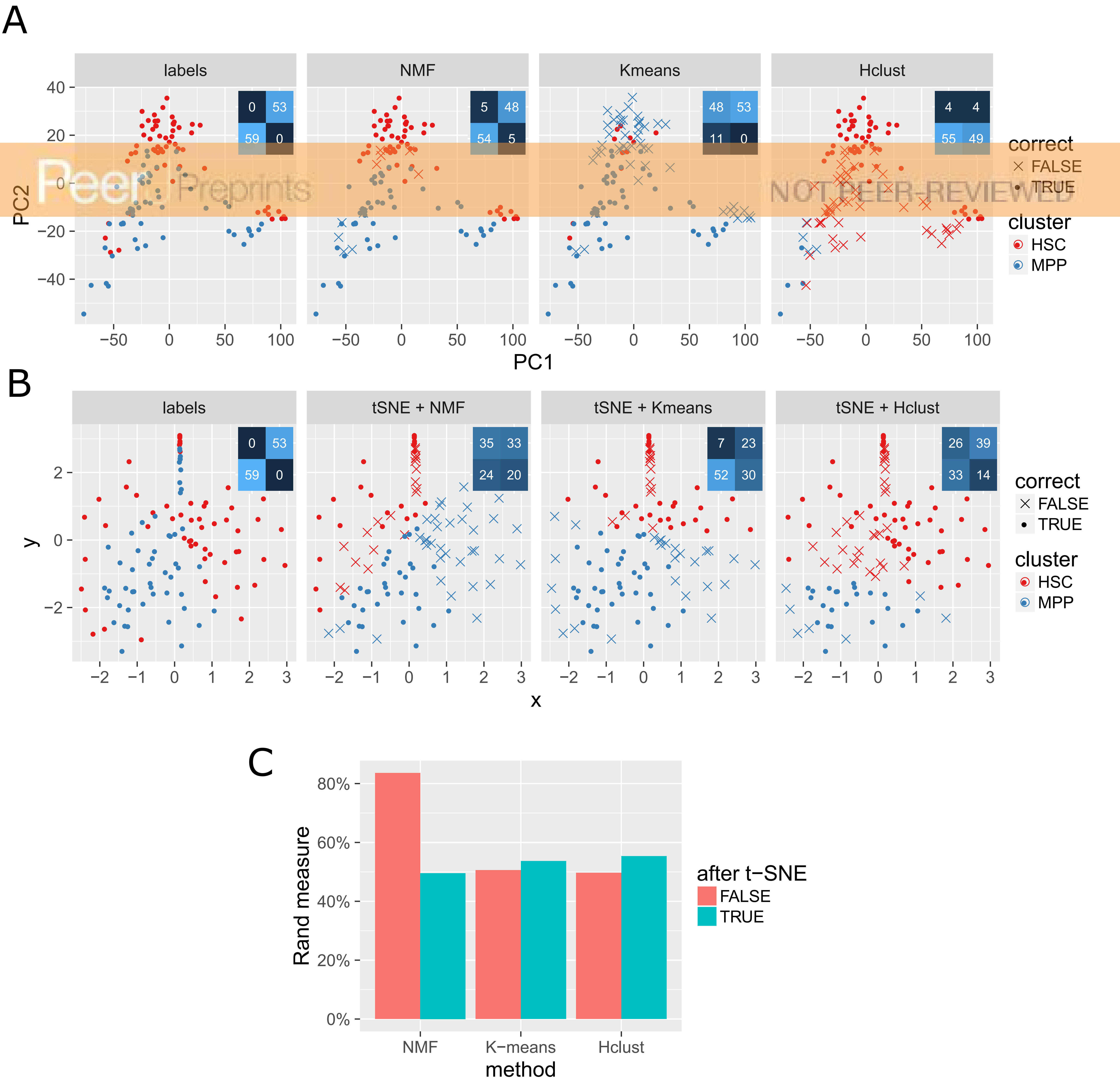


Fig. 1: The workflow of NMFEM. The input can be either FastQ files or a raw counts table. If FastQ files are used, they are aligned using TopHat and counted using FeatureCounts (steps shown in brackets). The input or calculated rawcounts table are filtered by samples and genes, converted into RPKMs using gene lengths, and normalized by samples. We then run NMF method on them to detect subpopulations, and find the feature genes separating the detected subpopulations. Finally,we feed the feature genes as seed genes in FEM, and generate PPI gene modules that contain highly differentially expressed genes.



PeerJ Preprints
NOT PEER-REVIEWED

Fig. 2: Comparisons among clustering methods on the HSC vs. MPP1 scRNA-Seq data.

(A) The PCA scatter-plots of the samples, based on their log normalized expression level. Colors indicate the most favorable labeling that can be assigned to the clustering result generated by each method. The correctly and incorrectly labeled samples are marked by dot (•) and cross (x), respectively.Confusion matrices of the methods in comparison are inserted on thetop-right corner of each sub-panel. The closer the matrix is to a diagonal matrix, the more accurate the method is. (B) The scatter-plots of the samples for K-means and hierarchical clustering, after t-SNE based dimension reduction. (C) Rand measures of the methods in comparison, before and after t-SNE. Rand measure ranges from 0 to 1, where a higher value indicates a greater clustering accuracy.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1839v2 | CC-BY 4.0 Open Access | rec: 9 Mar 2016, publ: 9 Mar 2016

PeerJ Preprints NOT PEER-REVIEWED

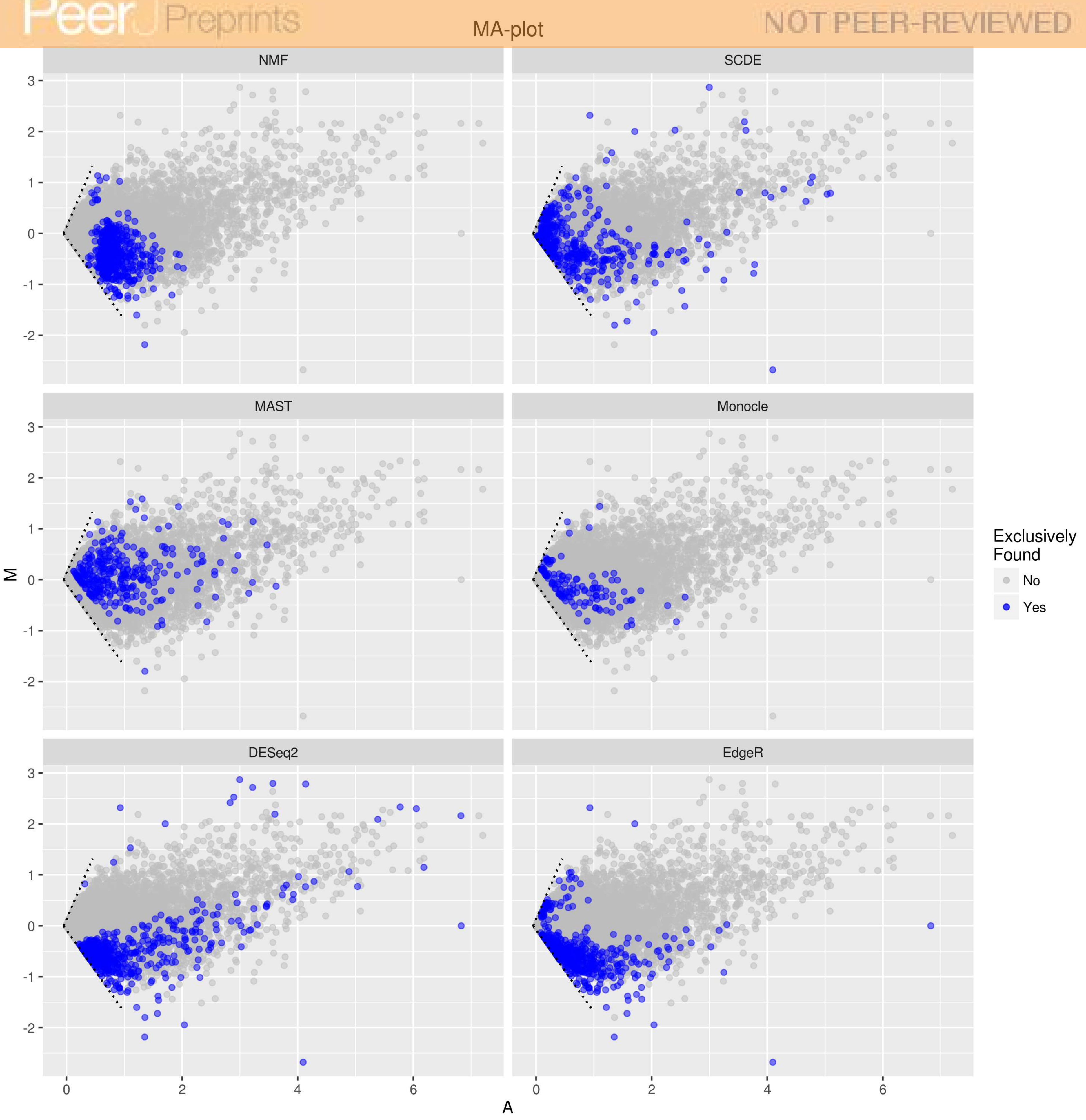


Fig. 3: MA-plots of significant or important genes defined by different methods.

Shown are scRNA-Seq data in the mouse lung distal epithelial cell E14.5 vs. E16.5 samples. The blue color highlights the genes selected as "the most significant" by the corresponding methods. X-axis (A-value) is the mean of the gene expression, and y-axis (M-value) is the difference of the gene expression between E16.5 and E14.5 stages.

PeerJ Preprints
NOT PEER-REVIEWED

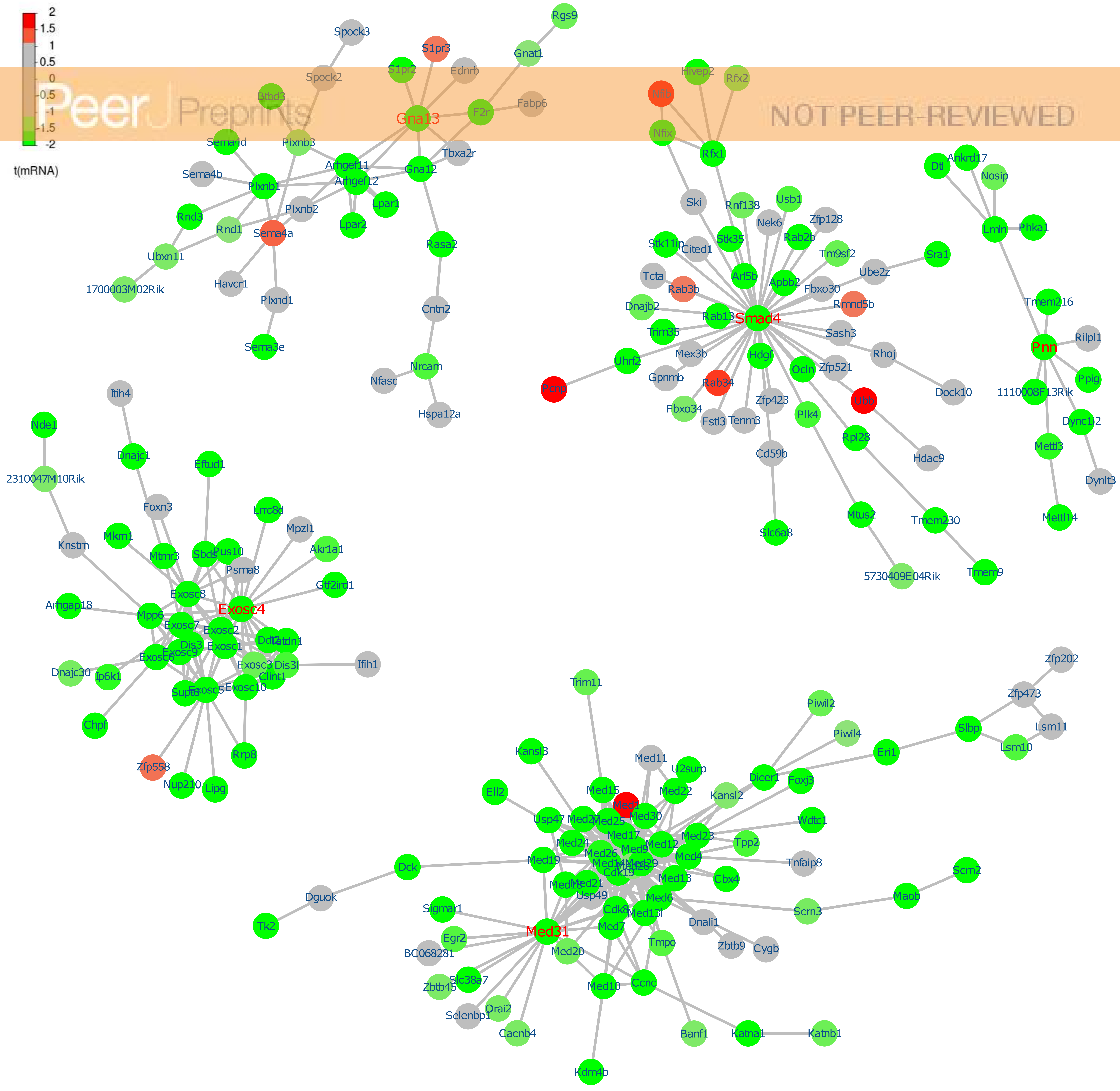
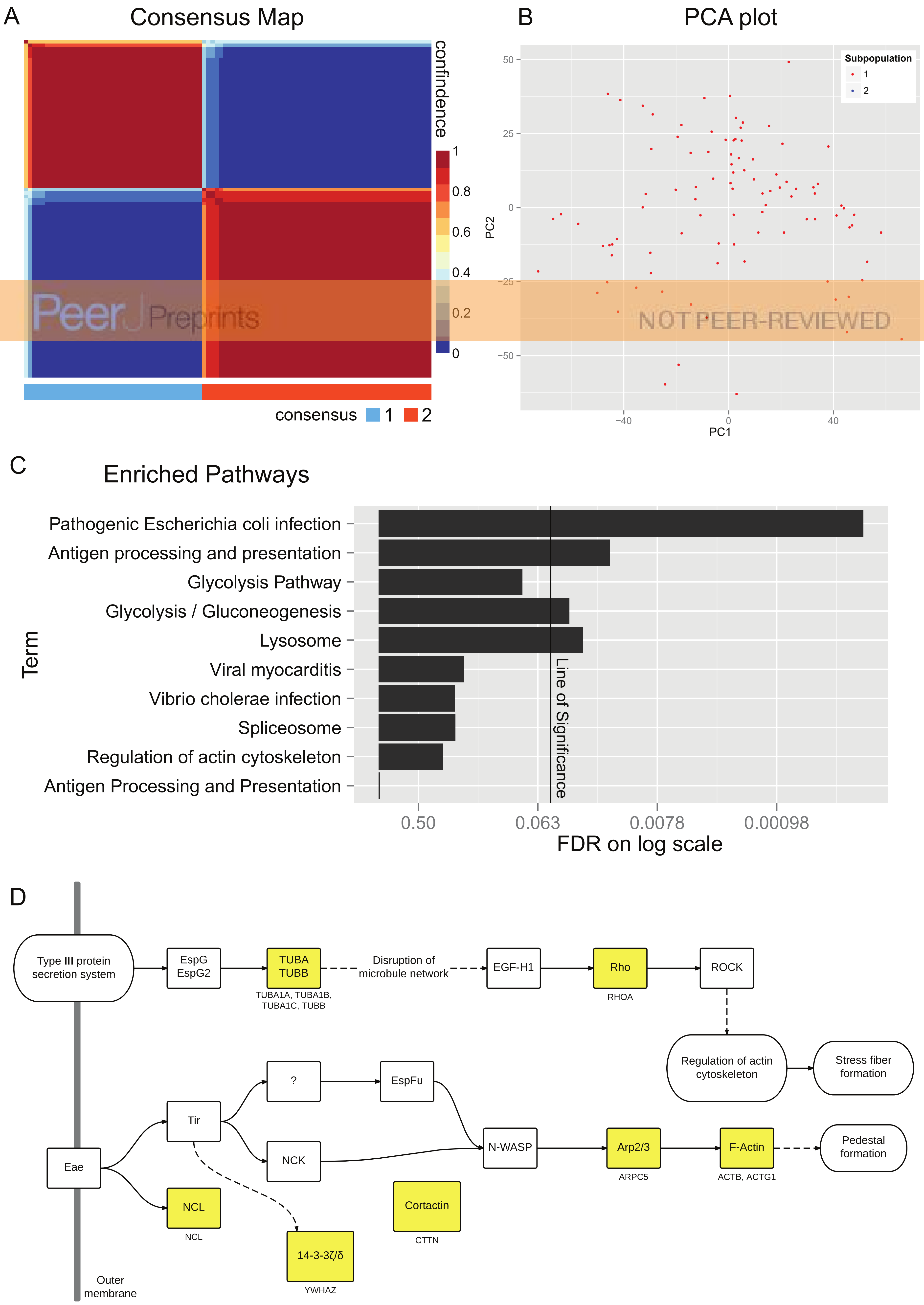


Fig. 4: Network of top 5 modules using the seed genes generated by NMF.

Shown are module detection results in the FEM package, using the top 500 most important genes detected by NMF in Fig. 3. ScRNA-Seq data in the mouse lung distal epithelial cell E14.5 vs. E16.5 samples are compared, where the red and green colors indicate up-and down-regulation of genes in E16.5 relative to E14.5, respectively. The top 5 modules are selected by the p-values calculated from the internal Monte-Carlo method in the FEM package (Table 1).



PeerJ Preprints

NOT PEER-REVIEWED

Fig. 5: Using NMF to identify subpopulations in a single glioblastoma tumor from patient MGH28.

(A) The consensusheat map generated from NMF. The two subpopulation clusters are the evident 2 red squares, marked out by number 1 and 2. The brightness indicates the confidence level of two subpopulations. (B) The PCA plot of scRNA-Seq samples from patient MGH28, the discovered subpopulations are coded in red and blue colors. (C) The results of KEGG/BioCarta Pathway enrichment analysis. The line of significance (to the right of which meaning the FDR less than 0.05) is shown. (D) The protein interaction diagram of the KEGGpathway “Pathogenic E. Coli infection”. The proteins coded by the genes detected by NMF are highlighted yellow, with the gene names marked below.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1839v2 | CC-BY 4.0 Open Access | rec: 9 Mar 2016, publ: 9 Mar 2016