BMC Microbiology

RESEARCH ARTICLE Open Access

The identification of co-expressed gene modules in *Streptococcus pneumonia* from colonization to infection to predict novel potential virulence genes

Sadegh Azimzadeh Jamalkandi[1], Morteza Kouhsar[2], Jafar Salimian[1] and Ali Ahmadi[3*]

Abstract

Background: *Streptococcus pneumonia* (pneumococcus) is a human bacterial pathogen causing a range of mild to severe infections. The complicated transcriptome patterns of pneumococci during the colonization to infection process in the human body are usually determined by measuring the expression of essential virulence genes and the comparison of pathogenic with non-pathogenic bacteria through microarray analyses. As systems biology studies have demonstrated, critical co-expressing modules and genes may serve as key players in biological processes. Generally, Sample Progression Discovery (SPD) is a computational approach traditionally used to decipher biological progression trends and their corresponding gene modules (clusters) in different clinical samples underlying a microarray dataset. The present study aimed to investigate the bacterial gene expression pattern from colonization to severe infection periods (specimens isolated from the nasopharynx, lung, blood, and brain) to find new genes/gene modules associated with the infection progression. This strategy may lead to finding novel gene candidates for vaccines or drug design.

Results: The results included essential genes whose expression patterns varied in different bacterial conditions and have not been investigated in similar studies.

Conclusions: In conclusion, the SPD algorithm, along with differentially expressed genes detection, can offer new ways of discovering new therapeutic or vaccine targeted gene products.

Keywords: *Streptococcus pneumonia*, Transcriptome analysis, SDP algorithm, Systems biology

* Correspondence: aliahmadigorgani@gmail.com
[3]Molecular Biology Research Center, Systems Biology and Poisonings Institute, Baqiyatallah University of Medical Sciences, Tehran, Iran
Full list of author information is available at the end of the article

Background

Streptococcus pneumonia (pneumococci) is a common bacterial pathogen in children, immunocompromised individuals, and the elderly. It infects the upper respiratory tract (especially nasopharynx) of 27–65% of children and 10% of adults. Pneumococci can cause severe infections in susceptible hosts through a highly flexible gene expression capacity, allowing it to move from the nasopharynx and adapt to highly sterile body sites, including lung, blood, and brain. It causes a wide range of disorders, from otitis media and sinusitis to severe infections, such as bacteremia, pneumonia, and meningitis [1]. Hence, despite available pneumococcal treatments and effective vaccines, pneumococci is one of the 12 highly invasive pathogens causing more deaths than any other infectious diseases in the world [2]. A small pneumococcal genome size (3000–5000 genes) confirms that transcriptional events play a critical role in an adaptive and smart behavior [3]. Accordingly, many studies have investigated the pathogenicity behavior of the pathogen by measuring, through microarray experiments, the expression of essential virulence genes and comparing it with that of non-pathogenic bacteria in different niches during colonization and invasion [4]. As systems biology studies have demonstrated, critical co-expressing modules and genes may serve as the key player in biological processes. Generally, Sample Progression Discovery (SPD) is a computational approach traditionally used to decipher biological progression trends and their corresponding gene module clusters in different clinical samples underlying a microarray dataset. This approach is used in progression-based diseases, including cancer, chronic pulmonary obstructive disease (COPD), and basic cellular processes, including cell differentiation [5]. The SPD framework tries to cluster genes into modules of co-expressed genes, construct modules' minimum spanning tree (MST), select modules corresponding to common MSTs, and, according to all genes of all selected modules, reconstruct a general MST [6]. Because some essential virulence genes may not necessarily be differentially expressed genes (DEGs), we aimed to use SPD to define if some critical niche dependent-, co-expressed modules, and genes are necessary for bacterial translocation through different host niches during the pathogenicity adventure of the pathogen. In the present study, we wonder if some genes are ignored during the DEGs detection approach while they may potentiate to serve as key candidate regulators in the pathology trend of pneumococcal diseases. We aimed to analyze pneumococcal's gene expression behavior in both colonization and invasion states by focusing on two microarray datasets of pneumococcus derived from the nasopharynx, lung, blood, and brain of mice. The data were analyzed through a machine-learning algorithm to detect those genes related to the infection progression from the nasopharynx to the lung, blood, and brain. We finally found some key expression modules and genes that could distinguish precisely between different clinical samples.

Results

Co-expressed modules detected by SPD

According to the aim of the present study, the two extracted datasets were pooled to analyze with limma. Each spot file contained 3297 unique probes, of which 943 included control probes that were filtered out in the pre-processing step. Then, the gene expression matrix was reconstructed as the input of SPD. As a result of the SPD algorithm, many co-expressed modules were obtained based on time series data in each group. After investigating all modules and their MST, modules that significantly separated the samples based on their gene expression patterns and source tissues (including nasopharynx, lung, blood, and brain) were selected to further analysis (Table 1, S1 File). Although the enrichment analysis and literature mining could not find any meaningful information about many of the modules due to limited genetic annotations and enrichment tools on *Streptococcus pneumonia*, some critical modules were identified in each group, including genes involved in essential cellular processes. Among these, top modules are analyzed in the following section (S1 Table).

Nasopharynx-lung progression

Regarding the nasopharynx-lung expression data, a total of 182 modules were detected by SPD, two of which (modules 14 and 71) were selected as the best results based on their MST, representing the invasion of bacteria from the nasopharynx to the lung (Fig. 1). As shown in Fig. 1, based on MST and hierarchical clustering, results showed that these modules' gene expression pattern is significantly different between lung and nasopharynx. The genes involved in module 14 are mostly enriched to the "purine metabolism" pathway and "*'de novo'* inosine monophosphate (IMP) biosynthetic" process (S2 Table). These pathways and processes are related to biofilm formation [7, 8]. In module 71, three genes (SP_2173, SP_2175, and SP_2176) are involved in host immune system defensive mechanisms against infections, including the "Cationic antimicrobial peptide (CAMP) resistance" pathway that contains six genes [9, 10]. Another pathway in module 71 is the "Two-component regulatory system," a pathway that regulates the expression of pneumococcal surface antigen A protein and consequently controls the virulence of bacteria and its resistance oxidative stress [11, 12]. The two-component system is also related to Cellobiose Metabolism and the interaction of the bacteria with its environment [13, 14].

Table 1 The number of modules extracted from the data by the SPD algorithm

Sample source	No. of all modules	No. of selected modules	No. of top modules (No. of each module's genes)
Nasopharynx and lung	182	30	2 (13, 10)
Lung and blood	160	10	2 (7, 9)
Blood and brain	138	13	2 (3, 14)
Nasopharynx, lung, and blood	179	15	2 (5, 8)
Lung, blood, and brain	160	12	1 (3)
Nasopharynx, lung, blood, and brain	169	18	2 (4, 3)

Literature mining demonstrated that the module 14 genes are very important in pneumococcal infection. For instance, SP_0050 (purH), SP_0202 (nrdD), SP_0054 (purK) and SP_0205 (nrdG) are early pneumococcal response genes in human lung epithelial cells [15]. Iron starvation condition up-regulates the expression of nrdD, SP_0204, and SP_0205 genes, whereas it down-regulates purK, SP_1405, and SP_1460 genes [16]. In addition, SP_1780, SP_1405 (spxA), and SP_0274 (polC) are reported to be pneumococcal essential genes for pulmonary infection [17].

Lung-blood progression

Applying the SPD algorithm on lung-blood expression data detected 160 co-expressed modules, two of which (modules 22 and 101) were selected for further analysis. As shown in Fig. 2, these modules' expression pattern was significantly different between lung and blood samples. All proteins in module 22 interact physically with each other based on bacterial interactome (Fig. 2a). The genes in this module significantly enriched the "Ribosome" pathway and some significant processes, such as "nitrogen compound metabolic process," "primary metabolic process," and "response to stimulus" (S2 Table). The nitrogen compound metabolic process and primary metabolic process are dysregulated in copper resistance *Streptococcus pneumonia* [18]. No enrichment result was found for module 101, though some critical previously reported genes were present in this module. For example, Giefing et al. [19] introduced the SP_1923 and SP_1891 genes as vaccine candidates.

Blood-brain progression

In blood-brain data, 138 modules were identified, among which, based on MST results, modules 130 and 87 are

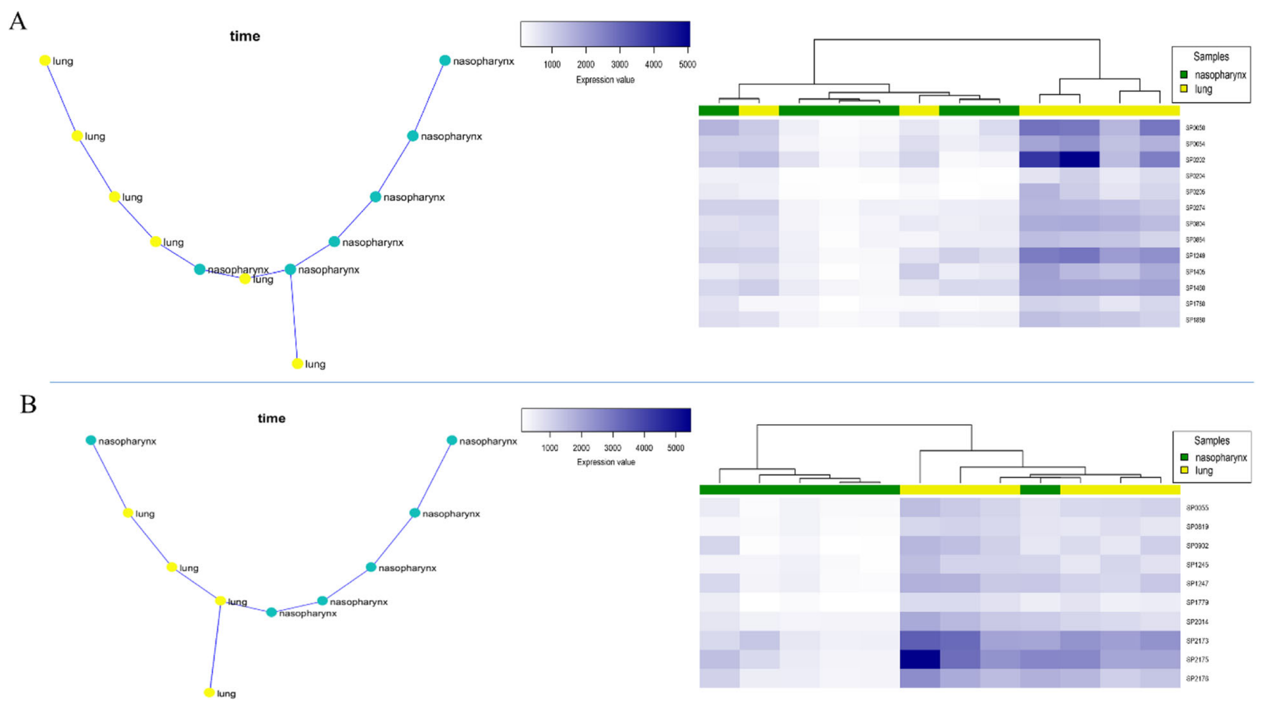

Fig. 1 a The co-expressed module 14 represents the infection progression in the nasopharynx to lung (Left: MST, Right: hierarchical clustering based on gene expression data). **b** The co-expressed module 71 represents the infection progression in the nasopharynx to lung (Left: MST, Right: hierarchical clustering based on gene expression data)

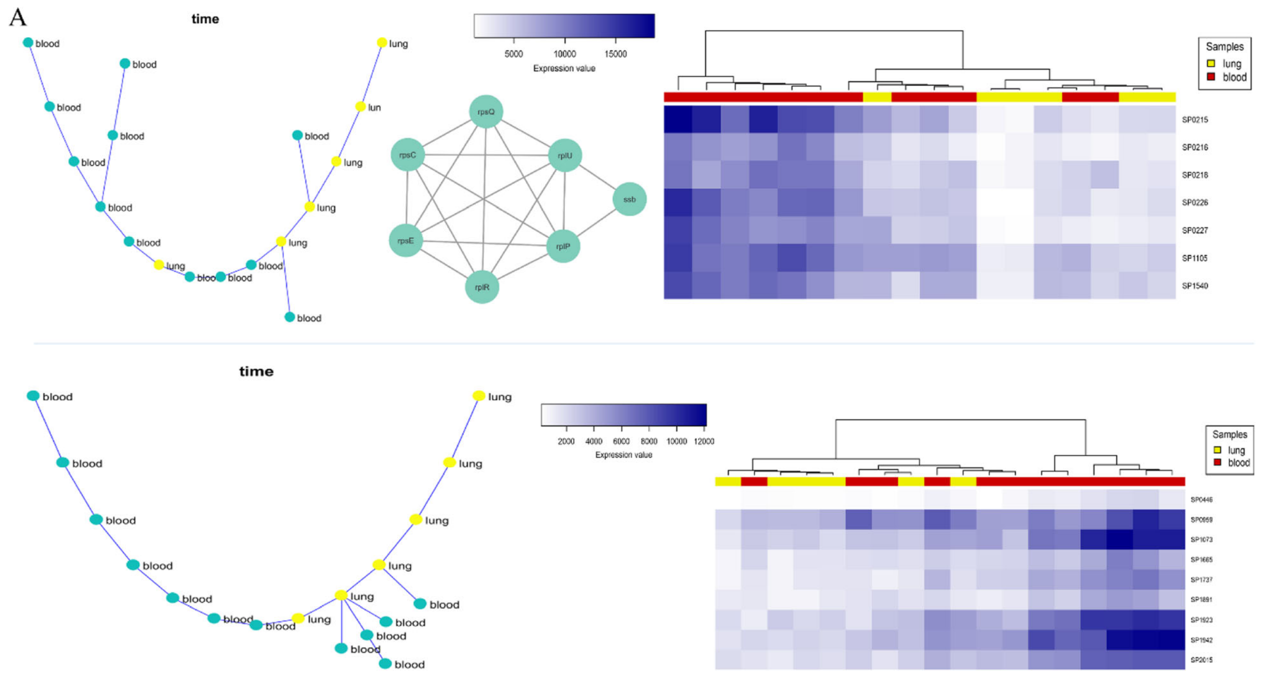

Fig. 2 a The co-expressed module 22 represents the infection progression in lung to blood (Left to right: MST, protein interaction network, and hierarchical clustering based on gene expression data). **b** The co-expressed module 101 represents the infection progression in lung to blood (Left: MST, Right: hierarchical clustering based on gene expression data)

present mostly co-expressed genes in the infection progression from blood to lung (Fig. 3). Interestingly, although module 130 contained only three genes, the MST and hierarchical clustering results showed that the expression pattern was significantly different between the blood and brain. We could not find any pathway or process for these genes through enrichment analysis, but in the STRING database, the gene SP_2146 interacts with four other genes, including SP_2144, SP_2145, SP_1654, and SP_0648 (*bgaA*) (Fig. 3a). In addition, SP_2146, SP_1654 and SP_0648 are involved in “other glycan degradation” pathway (*p*-value = 3.00e-05). Robb et al. [20] demonstrate that this pathway is required for full virulence in *Streptococcus pneumonia*.

For module 87, no result was obtained from enrichment analysis; however, the gene SP_0176 (ribAB) interacts with three other genes (SP_0177 or ribE, SP_0175 or ribH, and SP_0178 or ribD) in the bacterial interactome. These four genes were significantly enriched to the “Riboflavin metabolism” pathway (*p*-value = 6.82e-06), a critical pathway in pneumococcal infections that its regulatory factors have been previously introduced as novel drug targets [21, 22].

Nasopharynx-lung-blood progression

Given the data derived from the nasopharynx-lung-blood progression, 179 modules were detected by the SPD algorithm, among which modules 95 and 103 had the best clustering results (Fig. 4). As shown in Fig. 4, the gene expression patterns of these modules can cluster the samples. The heatmaps in Fig. 4 shows that as the infection progresses from the nasopharynx to lung and then to blood, the gene expression values are simultaneously decreased in module 95 and increased in module 103. The genes in module 95 and 103 are significantly enriched to “Ascorbate and Aldarate” and “Cysteine and methionine” metabolic pathways, respectively (S2 Table).

Lung-blood-brain progression

Regarding the lung-blood-brain data, 160 modules were detected, in which just one module (module 30) had appropriate results. As shown in Fig. 5, this module contained three genes, including SP_0171, SP_0391 or *CbpF*, and SP_1762.

Nasopharynx-lung-blood-brain progression (full progression)

In this section, we considered all samples’ data to find co-expressed modules that their expression patterns were significantly changed throughout all samples. We could identify 169 modules, among which two modules, including modules 34 and 144, had the optimal results (Fig. 6). Module 34 contained four genes, including SP_0171, SP_0256, SP_0391, and SP_1762, three of which are common with module 30 genes detected in the lung-

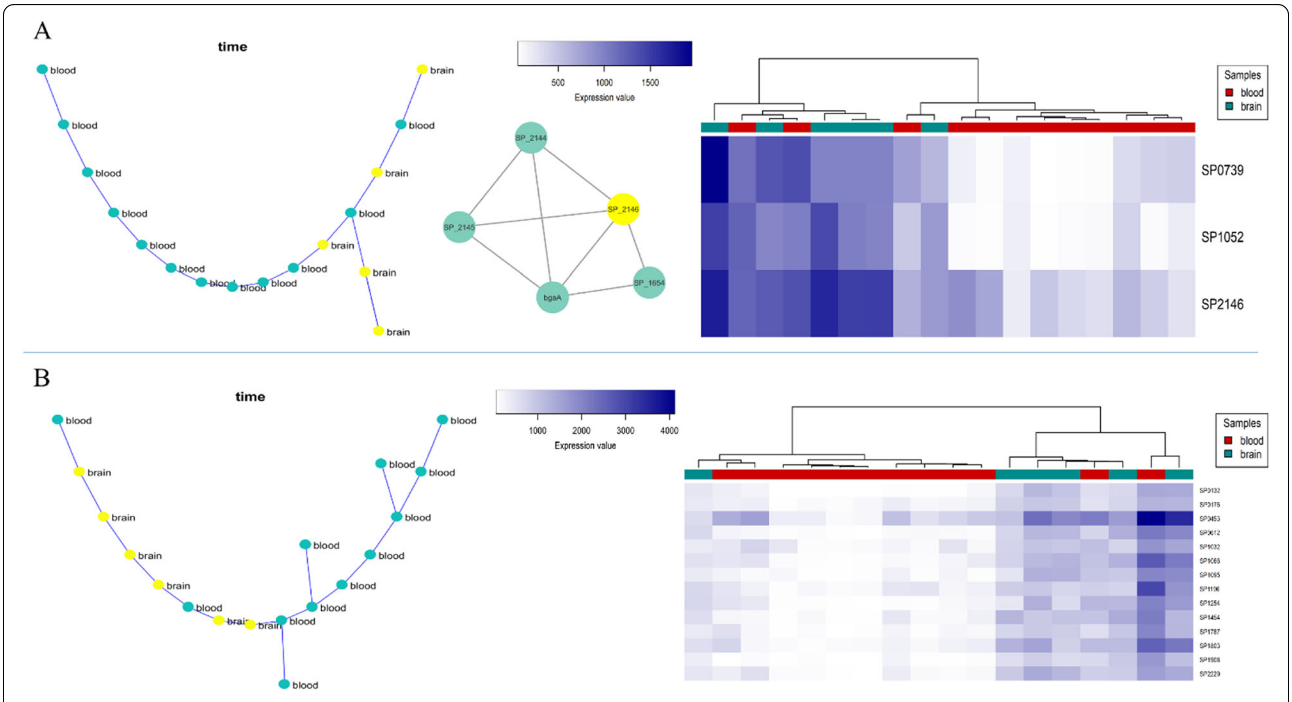

Fig. 3 a The co-expressed module 130 represents the infection progression in the blood to the brain (Left to right: MST, the genes interact with SP_2146 based on STRING database, hierarchical clustering based on gene expression data). **b** The co-expressed module 87 represents the infection progression in the blood to the brain (Left: MST, Right: hierarchical clustering based on gene expression data)

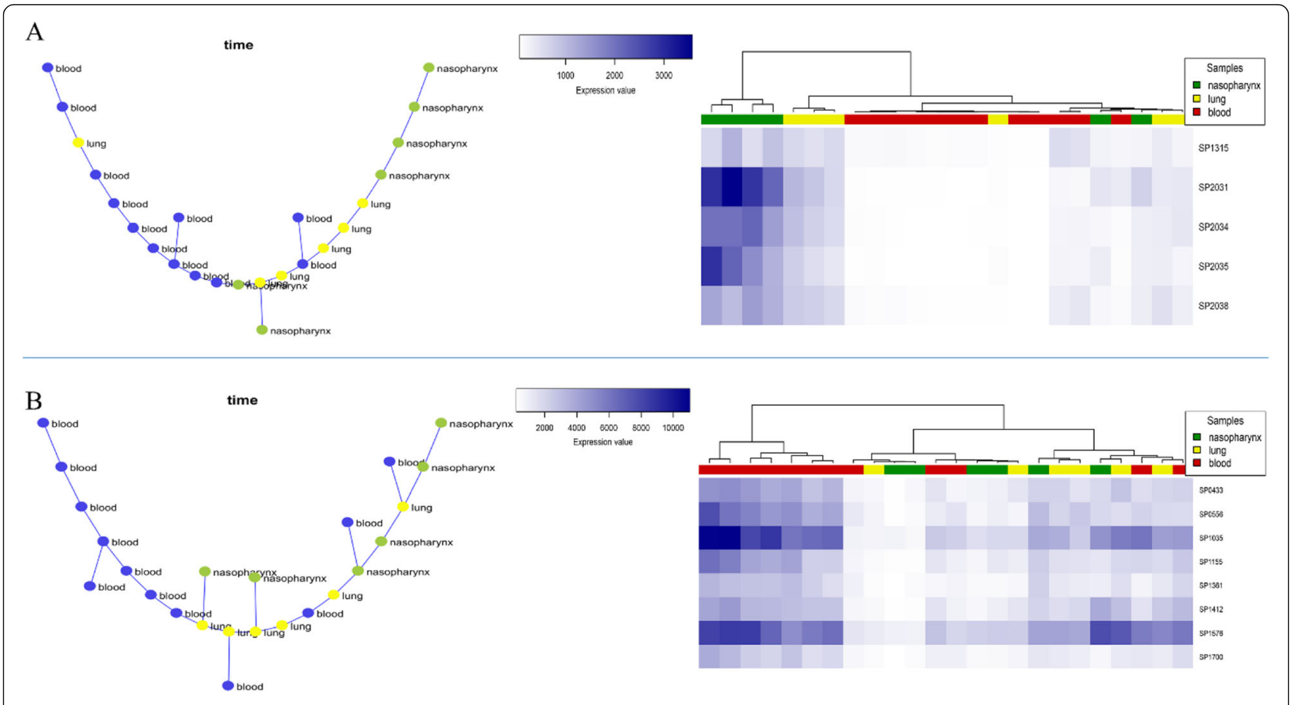

Fig. 4 a The co-expressed module 95 represents the infection progression from the nasopharynx to lung and blood (Left: MST, Right: hierarchical clustering based on gene expression data). **b** The co-expressed module 103 represents the infection progression from the nasopharynx to lung and blood (Left: MST, Right: hierarchical clustering based on gene expression data)

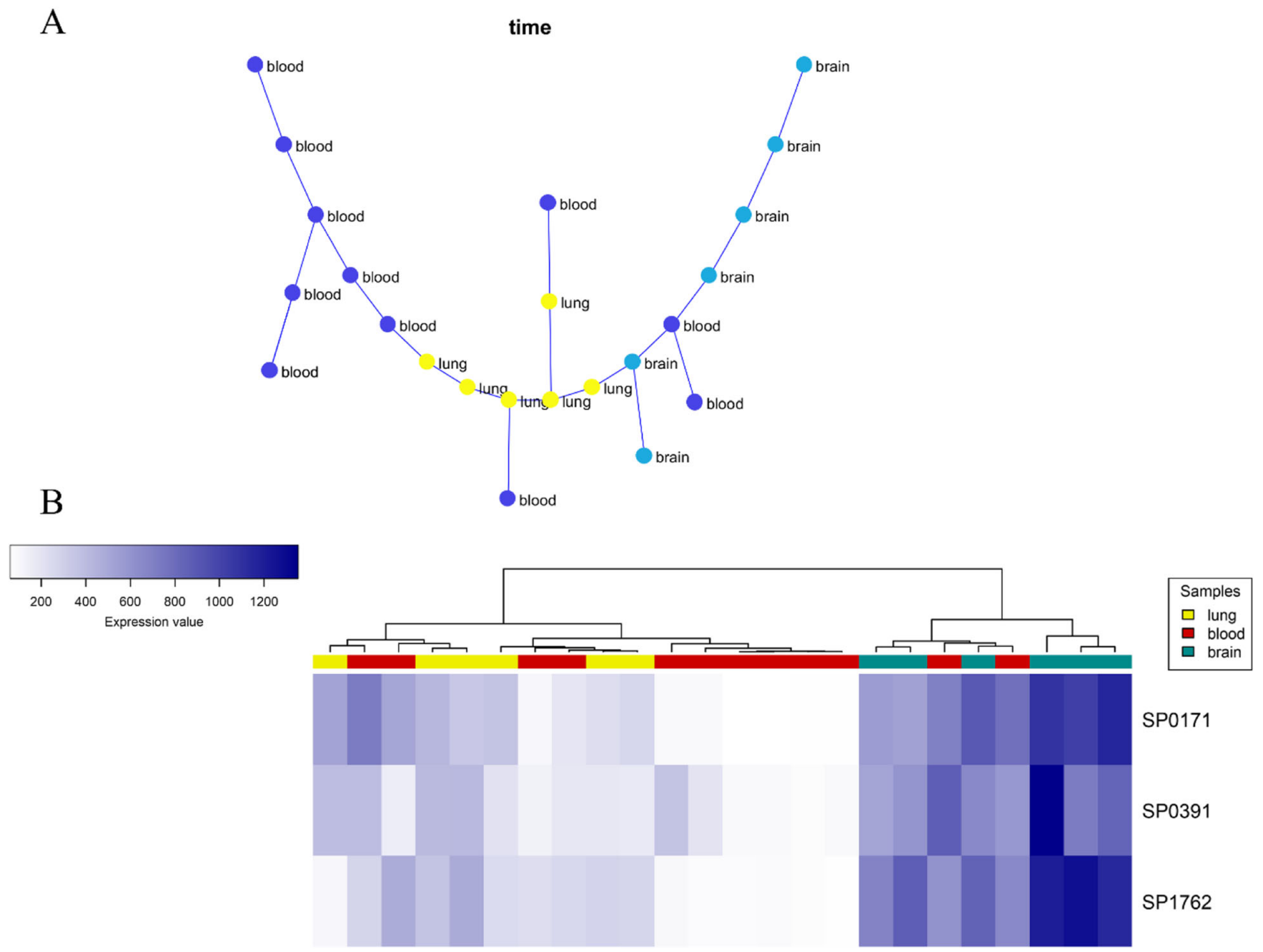

Fig. 5 The co-expressed module 30 represents the pneumococcal infection in the lung-blood-brain progression (**a**: MST, **b**: hierarchical clustering based on gene expression data)

blood-brain progression data. No pathways or processes were found via enrichment analysis for this module. Regarding module 144, three genes, including SP_2031, SP_2034, and SP_2035, were found. As shown in Fig. 6, the module's MST and hierarchical clustering results indicated a significantly different gene expression pattern compared to all samples. Besides, this module's genes were significantly enriched to the "Ascorbate and aldarate metabolism" pathway (S2 Table).

Differentially expressed genes vs. SPD results

Considering the infection progression from the nasopharynx to the brain, we compared the differentially expressed genes (DEGs, S2 File) with the genes obtained from SPD analysis when the pneumococci transfer from one sample niche to another one. The number of co-expressed SPD detected genes and DEGs is shown in Table 2. Also, the Venn diagram of the SPD identified genes and DEGs is shown in Fig. 7. As shown in Table 2, the number of DEGs (with both logFC thresholds) when the bacteria move from blood to the brain is higher than in other conditions. This evidence may indicate the high bacterial transcriptome alteration through infection progression from blood to the brain. In contrast, the transcriptome alteration is the least when the infection progressed from the lung to blood (comparing with two other conditions shown in the table). Approximately 10, 15, and 25% of the DEGs are co-expressed and detected by SPD in nasopharynx vs. lung, lung vs. blood, and blood vs. brain conditions.

Regarding the genes in the SPD modules (S1 file), particularly the modules mentioned earlier, we can find a group of genes that are not classified as DEGs; however, they are critical for the infection process based on previous reports. For instance, as mentioned above, SP_0054 is an early response gene in human lung epithelial cells, while SP_0274 is a crucial gene in pulmonary infection [8, 17]. Table 3 shows some critical infection-related genes detected by the SPD algorithm, while they are not assigned to DEGs.

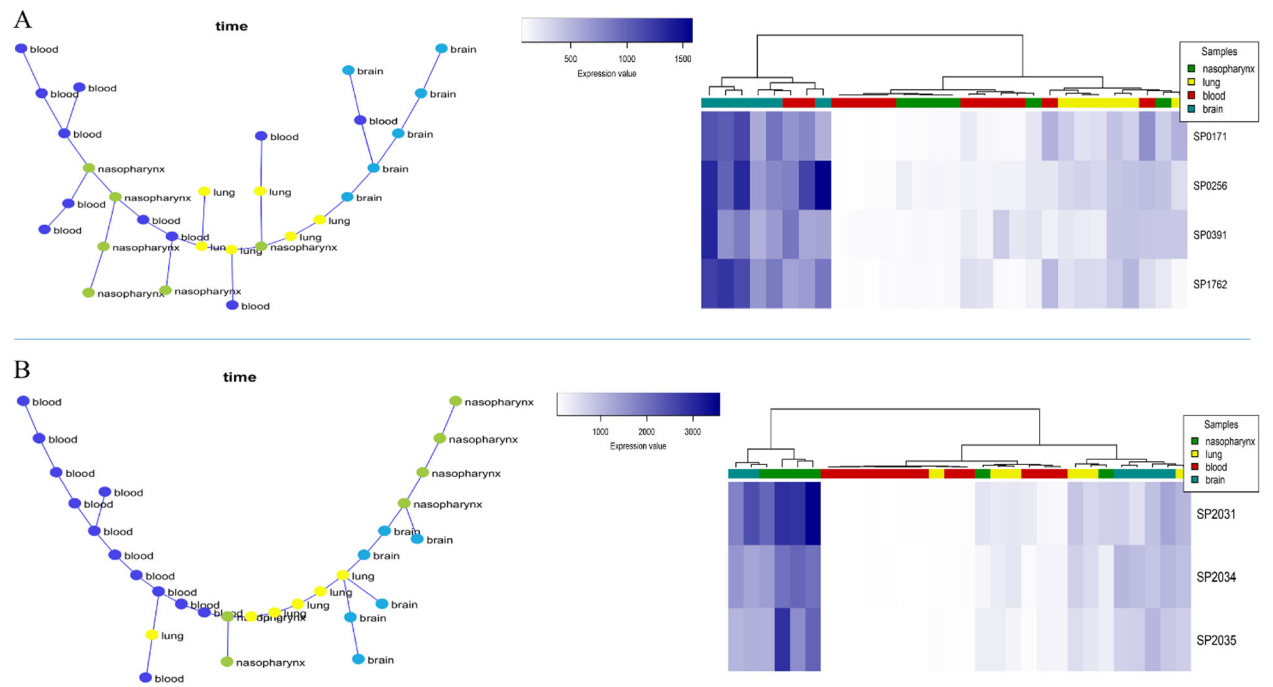

Fig. 6 a The co-expressed module 34 represents the infection progression in the nasopharynx-lung-blood-brain model (Left: MST, Right: hierarchical clustering based on gene expression data). **b** The co-expressed module 144 represents the infection progression in the nasopharynx to lung, blood, and brain (Left: MST, Right: hierarchical clustering based on gene expression data)

Discussion

Extensive transcriptomic changes occur when pneumococci migrate from the nasopharynx into the lung, blood, or brain. Available pneumococcal gene expression studies rely on only DEG genes during bacterial transmission from one body niche to another. According to the systems biology approach, sometimes a gene may not have a significant expression level; however, it could play an important role in the complex system of gene regulation and disease progression. Our study aimed to predict these genes signatures and related alterations during infection progression from the nasopharynx to other niches. Acccordingly, we tried to apply a method on transcriptome data to extract a subset of genes undergoing a spectrum of expression changes between niches. Because co-expressed genes often share common pathways and are involved in common cellular processes [25], appropriate methods should try to extract co-expressed modules instead of single or unrelated altered genes. Therefore, we selected the SPD algorithm as an appropriate method to identify co-expressed modules representing sample progression in transcriptome data. Although other approaches such as differentially co-expressed module identification [26] and Atomic Regulons can be used [27], the SPD algorithm can compare multiple niches simultaneously. Though the feature selection algorithms [28] can detect gene alterations in multiple conditions, they ignore gene-gene interactions and thus were not suitable for our study.

Table 2 The number of DEGs and co-expressed SPD detected genes

Condition	#SPD	#DEG	#Common	logFC
Nasopharynx vs. lung	632	729	196	0.7
		446	105	1
Lung vs. blood	220	644	59	0.7
		327	33	1
Blood vs. brain	485	976	161	0.7
		821	126	1

*P-value< 0.05

SPD is performed to obtain different sets of genes specific to every niche. For example, SP_0446 and SP_0959 genes were detected in lung-blood migration data. These genes were previously reported as dysregulated genes in the early response in *THP-1 human macrophages* [29]. *In blood-brain* migration *data,* SP_2144 was detected in module 26. This gene, along with two others (SP_2145 and SP_1654), was previously reported as virulence-related genes in *S. pneumonia* [17, 30, 31]. These three genes physically interact with module 130 (detected in blood-brain migration) based on *S. pneumonia interactome.* SPD detects SP_0171 and SP_0391 (cbpF) in module 30 in lung-blood-brain migration data. SP_0171 is a ROK family protein expressed differentially in the early response to THP-1 human macrophage [29]. SP_0391

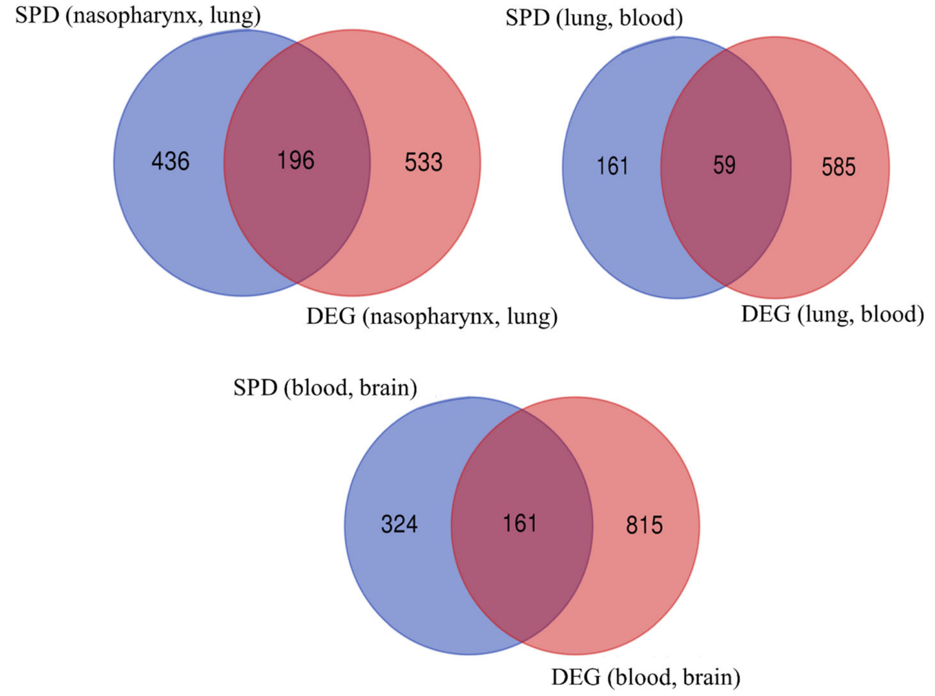

Fig. 7 The Venn diagram of the SPD results and DEGs (|logFC| > 0.7 and *p*-value< 0.5)

(*cbpF*) is an important choline-binding protein that was reported previously as a virulence factor in *S. pneumonia* [32–34]. In full progression data, SP_0391 and SP_0256 are two important co-expressed genes that were detected by SPD in module 34. SP_0391 is an important virulence factor of pneumococci [32–34], and SP_0256 is up-regulated in response to penicillin [23]. Furthermore, enrichment analysis revealed some important pathways and processes which may play an important role in pneumococcal infection. For instance, the genes in module 22, detected in lung-blood progression data, significantly enriched to "nitrogen compound metabolic process." Lui et al. demonstrated that the genes deferentially expressed in children with acute otitis media due to *Streptococcus pneumonia* are significantly enriched in this process [35]. As mentioned in the result section, in nasopharynx-lung-blood migration data, module 95 and 103 are significantly enriched to "Ascorbate and aldarate

Table 3 The key infection-related genes were not categorized as DEGs but detected by the SPD algorithm

SPD Modules	Gene name	Role in the infection process	Reference (s)
14 (Nasopharynx, lung)	SP_0054	Early response gene in human lung epithelial cells	[8]
	SP_0274	Essential gene in lung infection	[17]
	SP_1460	Involved in iron starvation condition	[16]
	SP_1780	Essential gene in pulmonary infection	[17]
17 (Nasopharynx, lung)	SP_2176	Enriched in Two-component system which controls the virulence and bacterial resistance to oxidative stress	[11, 12]
101(lung, blood)	SP_1923	Vaccine candidate gene	[19]
22 (lung, blood)	SP_0215	Enriched in nitrogen compound metabolism and primary metabolic process which is dysregulated in copper resistance in Streptococcus pneumonia	[18]
	SP_1540		
	SP_1105	Enriched in the metabolic process of nitrogen compounds	[18]
130 (blood, brain)	SP_0739	Up-regulated in response to exposure to penicillin	[23]
	SP_1052	Contributes to virulence in mice	[24]

metabolism" and "Cysteine and methionine metabolism" pathways, respectively. Previous studies demonstrated that ascorbic acid metabolism affects the expression of some critical genes in the pathogenicity of *S. pneumonia* [36]. Also, methionine synthesis has a critical role in bacterial growth and virulence [37]. Identified genes may be applied as antibacterial therapeutic targets and vaccine candidates after more investigations, including determining the cellular location. There are some methodological limitations and problems in this study: 1) Data acquisition; at first, we tried to search for microarray and RNA-Seq data from public datasets, such as GEO and ArrayExpress, using the keyword "Streptococcus pneumonia." There was a great limitation since the systems biology approach requires a high number of clinical samples. Unfortunately, we found that high-throughput data are scarce for most bacteria, including *S. pneumonia*, and in contrast to many other human-related fields, including cancer studies, the field of bacterial pathogenesis in systems biology is relatively in its infancy. Accordingly, we had to perform our study on currently available data. We finally found only two appropriate pneumococcal transcriptome datasets that could be integrated to increase the data volume. However, regarding the advent of systems biology approaches in medical bacteriology, this field will be definitely developed in the future. In spite of limited studies with a small sample size, the sample size is still critical to achieving precise statistical significance in systems biology. Pooling data from similar studies, if logically permissible, could overcome to some extent the problems. Therefore, it is imperative to conduct larger volume studies and use a high number of samples to generate applicable high-throughput data. Due to the emerging of high-throughput technology, such as RNA-Seq, the data limit will be diminished, and in future studies, machine-learning approaches, such as SPD, could be applied to new appropriate datasets to extract significant results. 2) Another critical issue is that available databases are not exclusively devoted to bacteria, and their search tools are publicly designed, making it difficult to explore bacterial data. We could only search and use available datasets in spite of their limitations. For this reason, we were not capable of interpreting some of the obtained results. Accordingly, developing a comprehensive bacteria-specific database storing transcriptomic (or other bacterial omics) data, along with specially designed bacterial searching tools, is a valuable and essential step in developing and advancing systems biology studies to more in understanding the pathogenesis of infectious diseases. 3) The lack of an appropriate enrichment tool is another challenging issue in our research. STRING was the only relatively suitable enrichment tool; however, it is not specific for bacteria and may cover very poorly pneumococcal genes. Providing powerful and user-friendly enrichment tools for bacterial pathogens is highly needed. 4) The next challenge was to interpret the results to obtain functional information about these genes and their associated biological processes and pathways through databases. Although there are various databases and many identified biological processes for human and mouse genes, it does not cover most bacteria, including pneumococci. Likewise, no annotation on the function or biological processes was available in databases for numerous discovered pneumococcal genes in this study. This greatly affects the enrichment process because we were not able to provide any interpretation for many modules or gene clusters, although the results showed significant issues.

As a note, although given the small genotype differences between various serotypes, it would be better to use data from one serotype alone for the study; however, for some reason, we utilized only two datasets. First, our study is based on identifying bacterial invasion patterns and based on this, and we can approximately categorize invasive serotypes. Second, we needed to investigate the pathogenesis pattern in several ecological niches, from the normal flora in the nasopharynx to the complete pathogenicity (including meningitis and sepsis). Hence, we used only two studies. Third, each of these studies alone had a small sample size, and by pooling them, the sample volume was increased. Consequently, we believe that the lack of sufficient omics data, bacteria-specific databases, and appropriate tools are the main drawbacks of systems biology and computational research to analyze bacterial pathogens, such as Streptococcus pneumonia. In conclusion, this is the first study using the SPD algorithm to assess the transcriptome pattern of pneumococci in different niches, regardless of the expression fold change of genes. Although some of the genes obtained here are entirely unknown, our results show that the expression patterns of these genes are different in different niches, and some of them interact with important well-known genes at the protein level, emphasizing their importance for more closely recognition. It seems that this approach could identify new essential genes involving in various pneumococcal pathogenesis that have been disregarded in the conventional method (fold change expression analysis). This approach could identify important novel genes. This approach is not specific to pneumococci and is applicable if there is adequate and appropriate data, database, and enrichment tools for any other pathogen.

Methods

Data preparation

Figure 8 shows our method workflow. The first step of our study is data preparation. The datasets used in this

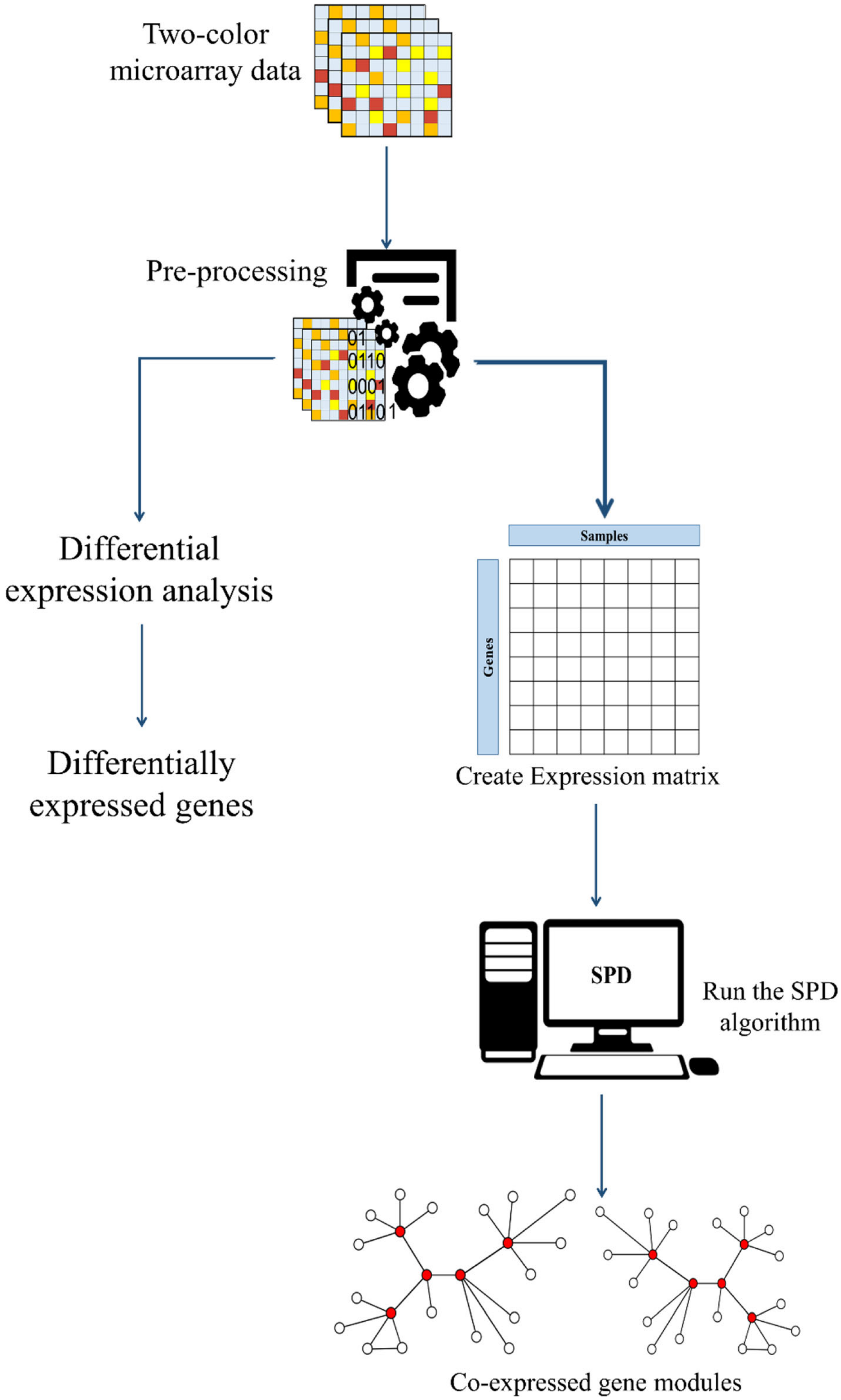
Two-color microarray data
Pre-processing
01
0110
0001
01101
Differential expression analysis
Differentially expressed genes
Samples
Genes
Create Expression matrix
SPD
Run the SPD algorithm
Co-expressed gene modules

Fig. 8 Method workflow

study were downloaded from ArrayExpress database (https://www.ebi.ac.uk/arrayexpress/) with the accession numbers of E-BUGS-130 and E-BUGS-133. The two-color microarray technology had been used to extract the total RNA expression of the bacteria in 48, 72, and 96 h' time points in a mice intranasal challenge experiment [38, 39]. E-BUGS-130 contains the blood and brain samples, and E-BUGS-133 contains the nasal, blood, and lung samples. The pneumococcal strains included the strain WCH43 (serotype 4) and WCH16 (serotype 6A) isolated from blood. The array's design is available at the ArrayExpress database with the accession number of A-BUGS-14, and the gene annotations are based on TIGR4 and R6 strains.

Data processing

After the pre-processing of samples, the background correction and the quantile normalization were applied using the limma R package [40]. The green and red color data were separated in the microarray results to reconstruct an expression matrix containing the expression data of the genes in each experimental condition. Next, the average expression value was replaced for duplicate probes or conditions in each dataset. Finally, the datasets' expression matrices were combined to reconstruct the final expression matrix (S3 File) for further analysis. In this matrix, the rows represented probes or genes, and the columns represented the experimental conditions (time points and samples).

Differential gene expression analysis

To compare the SPD results with Differentially Expressed Genes (DEGs), differential gene expression analysis was applied using the limma package [40] in the R programming environment. The genes that are differentially expressed from one niche to another when the infection progressed from nasopharynx to brain were extracted through t-test with p-value < 0.05 and $|\text{logFC}| > 0.7$ ($|\text{logFC}| > 1$ was also extracted).

Module detection and enrichment analysis

To detect co-expressed gene modules representing biological progressions behind time-series microarray data, the SPD algorithm was applied to the expression matrix [6]. Based on each detected module's expression, a Minimum Spanning Tree (MST, an acyclic graph with minimum total edge weights) was created as columns of the expression matrix for each sample. The weight of each edge denotes the Euclidean distance between two samples, and each tree represents a biological progression in all samples [6]. The pneumococcal infection progression begins from the nasopharynx and can extend to the lung, blood, and brain [41]. Considering the expression data in these niches in multiple time points, we categorized the data into six groups, including 1) [Nasopharynx and lung], 2) [Lung and blood], 3) [Blood and brain], 4) [Nasopharynx, lung, and blood], 5) [Lung, blood, and brain], and 6) finally [Nasopharynx, lung, blood, and brain]. Subsequently, the SPD algorithm was applied to each group with the correlation threshold of 0.95 and the minimum gene cut-off of one to predict the significant modules. After detecting the modules, we compared the MSTs in each group and selected those modules able to separate the body niches based on their expression patterns. These modules were chosen as the best results of the SPD algorithm for further analyses.

Gene set enrichment analysis and interaction assessment

The gene set enrichment analysis was performed for detected modules using the Comparative GO web tool [42, 43], the STRING database [44] based on KEGG pathways [45], and Gene Ontology Biological Processes [46, 47]. Besides, the STRING database was used for interaction analysis.

Hierarchical clustering and visualization

Hierarchical clustering and visualization were performed in the R programming environment with the Euclidean distance and Ward.D2 method [48]. The Venn diagrams were drawn with an online tool available at http://bioinformatics.psb.ugent.be/webtools/Venn/. Also, network visualization was performed via Cytoscape software [49].

Conclusions

In conclusion, this is the first study using the SPD algorithm to assess the transcriptome pattern of pneumococci in different niches, regardless of the expression fold change of genes. Although some of the genes obtained here are entirely unknown, our results show that the expression patterns of these genes are different in different niches, and some of them interact with well-known essential genes at the protein level, emphasizing their importance for more closely recognition. It seems that this approach could identify new essential genes involving in various pneumococcal pathogenesis that have been disregarded in the conventional method (fold change expression analysis). This approach can identify significant novel genes not only in pneumococci but also in other pathogens in the case of the availability of adequate and appropriate data, databases, and enrichment tools.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12866-020-02059-0.

Additional file 1:. SPD genes. All genes were detected by the SPD algorithm.

Additional file 2:. DEGs. Differentially expressed genes.

Additional file 3:. Expression matrix. The expression matrix was extracted from microarray samples.

Additional file 4: Table S1. Genes in the top modules. Co-expressed genes in the best modules extracted from the data by the SPD algorithm.

Additional file 5: Table S2. Enrichment Analysis. Biological process and pathway enrichment analysis results for the best modules obtained using the SPD algorithm.

Abbreviations

SPD: Sample Progression Discovery; COPD: Chronic pulmonary obstructive disease; MST: Minimum spanning tree; DEGs: Differentially expressed genes; IMP: Inosine monophosphate; CAMP: Cationic antimicrobial peptide

Acknowledgments

Not applicable.

Authors' contributions

SAJ: Conceptualization, result analysis, writing-review & editing. MK: implementation, formal analysis, investigation, writing-review & editing. JS: Conceptualization, writing-review & editing. AA: Conceptualization, Supervision, project administration, writing-review & editing. The author(s) read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets are available in the ArrayExpress database with the accession number of E-BUGS-130 and E-BUGS-133. The SPD algorithm source code and user guide are available at http://pengqiu.gatech.edu/software/SPD/index.html. Other source codes and materials are available upon request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Chemical Injuries Research Center, Systems Biology and Poisonings Institute, Baqiyatallah University of Medical Sciences, Tehran, Iran. [2]Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran. [3]Molecular Biology Research Center, Systems Biology and Poisonings Institute, Baqiyatallah University of Medical Sciences, Tehran, Iran.

Received: 25 June 2020 Accepted: 2 December 2020
Published online: 17 December 2020

References

1. Weiser JN, Ferreira DM, Paton JC. Streptococcus pneumoniae: transmission, colonization and invasion. Nat Rev Microbiol. 2018;16(6):355–67.
2. Henriques-Normark B, Tuomanen EI. The pneumococcus: epidemiology, microbiology, and pathogenesis. Cold Spring Harb Perspect Med. 2013;3(7):a010215.
3. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, Pessia A, Aanensen DM, Mather AE, Page AJ. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014;46(3):305.
4. Mahdi LK, Van der Hoek MB, Ebrahimie E, Paton JC, Ogunniyi AD. Characterization of pneumococcal genes involved in bloodstream invasion in a mouse model. PLoS One. 2015;10(11):e0141816.
5. Mostafaei S, Kazemnejad A, Jamalkandi SA, Amirhashchi S, Donnelly SC, Armstrong ME, Doroudian M. Identification of novel genes in human airway epithelial cells associated with chronic obstructive pulmonary disease (COPD) using machine-based learning algorithms. Sci Rep. 2018;8(1):1–20.
6. Qiu P, Gentles AJ, Plevritis SK. Discovering biological progression underlying microarray samples. PLoS Comput Biol. 2011;7(4):e1001123.
7. Yadav MK, Kwon SK, Cho CG, Park SW, Chae SW, Song JJ. Gene expression profile of early in vitro biofilms of Streptococcus pneumoniae. Microbiol Immunol. 2012;56(9):621–9.
8. Chao Y, Marks LR, Pettigrew MM, Hakansson AP. Streptococcus pneumoniae biofilm formation and dispersion during colonization and disease. Front Cell Infect Microbiol. 2014;4:194.
9. LaRock CN, Nizet V. Cationic antimicrobial peptide resistance mechanisms of streptococcal pathogens. Biochim Biophys Acta. 2015;1848(11 Pt B):3047–54.
10. Joo HS, Fu CI, Otto M. Bacterial strategies of resistance to antimicrobial peptides. Philos Trans R Soc Lond Ser B Biol Sci. 2016;371(1695):20150292.
11. McCluskey J, Hinds J, Husain S, Witney A, Mitchell TJ. A two-component system that controls the expression of pneumococcal surface antigen a (PsaA) and regulates virulence and resistance to oxidative stress in Streptococcus pneumoniae. Mol Microbiol. 2004;51(6):1661–75.
12. Paterson GK, Blue CE, Mitchell TJ. Role of two-component systems in the virulence of Streptococcus pneumoniae. J Med Microbiol. 2006;55(Pt 4):355–63.
13. McKessar SJ, Hakenbeck R. The two-component regulatory system TCS08 is involved in cellobiose metabolism of Streptococcus pneumoniae R6. J Bacteriol. 2007;189(4):1342–50.
14. Gomez-Mejia A, Gamez G, Hammerschmidt S. Streptococcus pneumoniae two-component regulatory systems: the interplay of the pneumococcus with its environment. Int J Med Microbiol. 2018;308(6):722–37.
15. Song XM, Connor W, Hokamp K, Babiuk LA, Potter AA. Streptococcus pneumoniae early response genes to human lung epithelial cells. BMC Res Notes. 2008;1:64.
16. Jimenez-Munguia I, Calderon-Santiago M, Rodriguez-Franco A, Priego-Capote F, Rodriguez-Ortega MJ. Multi-omic profiling to assess the effect of iron starvation in Streptococcus pneumoniae TIGR4. PeerJ. 2018;6:e4966.
17. Hava DL, Camilli A. Large-scale identification of serotype 4 Streptococcus pneumoniae virulence factors. Mol Microbiol. 2002;45(5):1389–406.
18. Guo Z, Han J, Yang XY, Cao K, He K, Du G, Zeng G, Zhang L, Yu G, Sun Z, et al. Proteomic analysis of the copper resistance of Streptococcus pneumoniae. Metallomics. 2015;7(3):448–54.
19. Giefing C, Meinke AL, Hanner M, Henics T, Bui MD, Gelbmann D, Lundberg U, Senn BM, Schunn M, Habel A, et al. Discovery of a novel class of highly conserved vaccine antigens using genomic scale antigenic fingerprinting of pneumococcus with human antibodies. J Exp Med. 2008;205(1):117–31.
20. Robb M, Hobbs JK, Woodiga SA, Shapiro-Ward S, Suits MD, McGregor N, Brumer H, Yesilkaya H, King SJ, Boraston AB. Molecular characterization of N-glycan degradation and transport in Streptococcus pneumoniae and its contribution to virulence. PLoS Pathog. 2017;13(1):e1006090.
21. Long Q, Ji L, Wang H, Xie J. Riboflavin biosynthetic and regulatory factors as potential novel anti-infective drug targets. Chem Biol Drug Des. 2010;75(4):339–47.
22. Hartmann N, McMurtrey C, Sorensen ML, Huber ME, Kurapova R, Coleman FT, Mizgerd JP, Hildebrand W, Kronenberg M, Lewinsohn DM, et al. Riboflavin metabolism variation among clinical isolates of Streptococcus pneumoniae results in differential activation of mucosal-associated invariant T cells. Am J Respir Cell Mol Biol. 2018;58(6):767–76.
23. Rogers PD, Liu TT, Barker KS, Hilliard GM, English BK, Thornton J, Swiatlo E, McDaniel LS. Gene expression profiling of the response of Streptococcus pneumoniae to penicillin. J Antimicrob Chemother. 2007;59(4):616–26.
24. Brown JS, Gilliland SM, Spratt BG, Holden DW. A locus contained within a variable region of pneumococcal pathogenicity island 1 contributes to virulence in mice. Infect Immun. 2004;72(3):1587–93.
25. Saelens W, Cannoodt R, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. Nat Commun. 2018;9(1):1–12.
26. Ray S, Lall S, Bandyopadhyay S. CODC: a copula-based model to identify differential coexpression. NPJ Syst Biol Appl. 2020;6(1):1–13.
27. Faria JP, Davis JJ, Edirisinghe JN, Taylor RC, Weisenhorn P, Olson RD, Stevens RL, Rocha M, Rocha I, Best AA. Computing and applying atomic regulons to understand gene expression and regulation. Front Microbiol. 2016;7:1819.
28. Park S, Shin B, Shim WS, Choi Y, Kang K, Kang K. Wx: a neural network-based feature selection algorithm for transcriptomic data. Sci Rep. 2019;9(1):1–9.
29. Song XM, Connor W, Hokamp K, Babiuk LA, Potter AA. Transcriptome studies on Streptococcus pneumoniae, illustration of early response genes to THP-1 human macrophages. Genomics. 2009;93(1):72–82.
30. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, Heidelberg J, DeBoy RT, Haft DH, Dodson RJ, et al. Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. Science. 2001;293(5529):498–506.

31. Polissi A, Pontiggia A, Feger G, Altieri M, Mottl H, Ferrari L, Simon D. Large-scale identification of virulence genes from Streptococcus pneumoniae. Infect Immun. 1998;66(12):5620–9.
32. Gamez G, Castro A, Gomez-Mejia A, Gallego M, Bedoya A, Camargo M, Hammerschmidt S. The variome of pneumococcal virulence factors and regulators. BMC Genomics. 2018;19(1):10.
33. Gosink KK, Mann ER, Guglielmo C, Tuomanen EI, Masure HR. Role of novel choline binding proteins in virulence of Streptococcus pneumoniae. Infect Immun. 2000;68(10):5690–5.
34. Yamaguchi M, Goto K, Hirose Y, Yamaguchi Y, Sumitomo T, Nakata M, Nakano K, Kawabata S. Identification of evolutionarily conserved virulence factor by selective pressure analysis of Streptococcus pneumoniae. Commun Biol. 2019;2:96.
35. Liu K, Chen L, Kaur R, Pichichero M. Transcriptome signature in young children with acute otitis media due to Streptococcus pneumoniae. Microbes Infect. 2012;14(7–8):600–9.
36. Afzal M, Shafeeq S, Kuipers OP. Ascorbic acid-dependent gene expression in Streptococcus pneumoniae and the activator function of the transcriptional regulator UlaR2. Front Microbiol. 2015;6:72.
37. Basavanna S, Chimalapati S, Maqbool A, Rubbo B, Yuste J, Wilson RJ, Hosie A, Ogunniyi AD, Paton JC, Thomas G, et al. The effects of methionine acquisition and synthesis on Streptococcus pneumoniae growth and virulence. PLoS One. 2013;8(1):e49638.
38. Mahdi LK, Wang H, Van der Hoek MB, Paton JC, Ogunniyi AD. Identification of a novel pneumococcal vaccine antigen preferentially expressed during meningitis in mice. J Clin Invest. 2012;122(6):2208–20.
39. Ogunniyi AD, Mahdi LK, Trappetti C, Verhoeven N, Mermans D, Van der Hoek MB, Plumptre CD, Paton JC. Identification of genes that contribute to the pathogenesis of invasive pneumococcal disease by in vivo transcriptomic analysis. Infect Immun. 2012;80(9):3268–78.
40. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W. Smyth GK: limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
41. Mahdi LK, Deihimi T, Zamansani F, Fruzangohar M, Adelson DL, Paton JC, Ogunniyi AD, Ebrahimie E. A functional genomics catalogue of activated transcription factors during pathogenesis of pneumococcal disease. BMC Genomics. 2014;15:769.
42. Fruzangohar M, Ebrahimie E, Ogunniyi AD, Mahdi LK, Paton JC, Adelson DL. Comparative GO: a web application for comparative gene ontology and gene ontology-based gene selection in bacteria. PLoS One. 2013;8(3):e58759.
43. Fruzangohar M, Ebrahimie E, Adelson DL. A novel hypothesis-unbiased method for gene ontology enrichment based on transcriptome data. PLoS One. 2017;12(2):e0170486.
44. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607–13.
45. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353–61.
46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
47. The Gene Ontology C. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47(D1):D330–8.
48. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J Classif. 2014;31(3):274–95.
49. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.