ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

G. Fiscon

NETWORK-BASED MODEL FOR STUDYING CHRONIC OBSTRUCTIVE PULMONARY DISEASE

R. 5, 2018

Giulia Fiscon - Institute for System Analysis and Computer Science "Antonio Ruberti" (IASI), CNR, Via dei Taurini 19, 00185 Rome, Italy. Email: giulia.fiscon@iasi.cnr.it.

ISSN: 1128-3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

via dei Taurini 19, 00185 ROMA, Italy

tel. ++39-06-49937101/02
fax ++39-06-49937106
email: iasi@iasi.cnr.it
URL: http://www.iasi.cnr.it

Abstract

Recently, we developed SWIM, a wizard-like software that integrates gene expression data with network topological properties for identifying a small pool of genes (i.e., switch genes) critically associated with drastic changes in cell phenotype. SWIM was amenable to detect switch genes in different organisms and cell conditions, with the potential to uncover important players in biologically relevant scenarios, including but not limited to human diseases. In this work, we present an application of SWIM to the microarray gene expression profiling of a large sample of resected lung tissues from subjects with severe chronic obstructive pulmonary disease (COPD). COPD is a progressive and obstructive lung disease characterized by an airflow obstruction that leads to a chronic inflammation and for which no cure is known. We aimed to find switch genes in the comparison between 111 severe COPD cases and 40 control smokers, freely available from GEO dataset GSE76925. We found 397 differentially expressed genes (DEGs) at a 10% FDR; and four of them – DLG2, ELMO1, NNT, SPAG16 - were at significant GWAS loci. From DEGs, SWIM built the COPD correlation network, in which two nodes are connected if the absolute value of the Pearson correlation coefficient for their expression profiles is greater than 0.5. This network encompasses 355 DEGs. Partitioning the COPD correlation network in communities, we found 3 modules, ranging in size from 408 genes in module 1, 387 genes in module 2, and 126 genes in module 3. In particular, module 2 was found enriched for B cell pathways, and included SERPINE2, CD79A, BCL2, POU2AF1, BCL11A that were previously considered as putative interactors of genes at COPD GWAS loci [3]. Then, SWIM identified 33 switch genes in COPD correlation network; all switch genes except one (E2F6) resulted up-regulated in COPD case with respect to control smokers; 29 switch genes are protein coding, including 2 transcription factors, ZNF143 and E2F6. The top differentially expressed switch gene was ZNF143 which negatively interacts in the network (i.e. highly negatively correlated) with NNT, a known COPD GWAS gene (default p-value < 10^{-5}) from the NHGRI-EBI Catalog (www.ebi.ac.uk/gwas/), and positively interacts with BCL2 (i.e. highly positively correlated).

Key words: SWIM tool, network medicine, COPD, batch effect

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a type of obstructive lung disease characterized by long-term breathing problem and progressive airflow obstruction, followed by chronic inflammation. It is the third leading cause of mortality worldwide, for which no cure is known [1].

Here, we apply SWitchMiner (SWIM) [2] - a software that we recently developed to unveil a small pool of genes (called "switch genes") critically associated with drastic changes in cell phenotype – to the microarray expression data on a large sample of severe COPD cases and control smokers available from [3], and freely downloadable from Gene Expression Omnibus (GEO) repository. Our aim was to identify switch genes associated to the transition between the control smokers and COPD cases.

When dealing with gene expression data, technical heterogeneity or batch effects (i.e., different experiment times, laboratory conditions, handlers, reagent lots, etc.) have proven challenging and constitute one lively discussed complication with such studies [4]. In fact, technical sources of variation added to the samples could be confounded with an outcome of interest and lead to incorrect conclusions. To shed light on this controversy, we were interested in testing the performance of SWIM both with and without considering the batch effects on the GEO dataset under study and analyzing the obtained results in both cases.

2. Material and Methods

2.1 Dataset

We downloaded the GEO dataset GSE76925 available from the original study of *Morrow et al. Scientific reports 2017* [3]. The authors of [3] performed microarray gene expression profiling in lung or airway tissues from subjects with chronic obstructive pulmonary disease (COPD) by using HumanHT-12 BeadChips (Illumina, San Diego, CA). They collected a total of 111 COPD cases and 40 control smokers with normal lung function. RNA and DNA were simultaneously extracted from the homogenized lung tissue using the AllPrep kit (Qiagen, Valenica, CA). RNA quality was assessed on a BioAnalyzer (Agilent, Santa Clara, CA).

In order to map probes into genes, we used the platform GPL10558 (Illumina HumanHT-12 V4.0 expression beadchip) available from GEO.

2.2 Software

In this study, we exploited the software SWitchMiner (SWIM) [2], which implements an integrated network analysis able to extract from genome-wide expression data key players (i.e. switch genes) marking the shift from one condition to another in a complex biological network. SWIM algorithm is composed of well-defined steps shown in Fig.1.

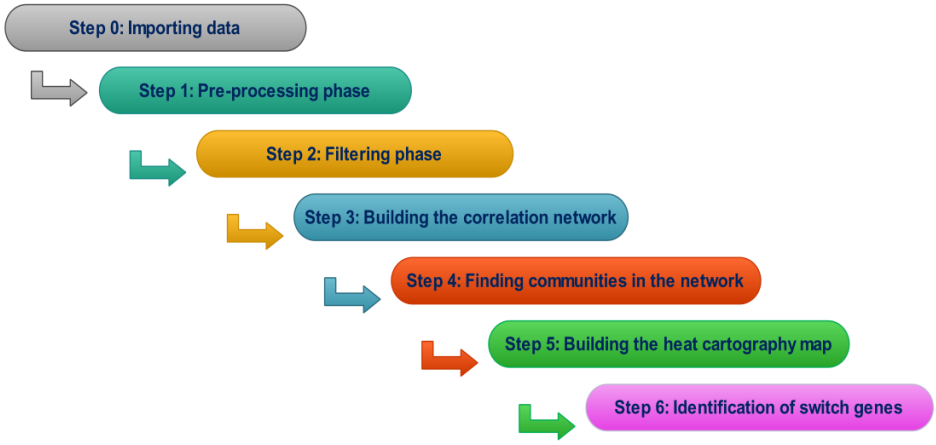

Figure 1 SWIM algorithm steps.

2.3 Models

We were interested to test the performance of SWIM on the two following linear models (Table 1):

- **model #1** without considering the batch effects at a 5% FDR (in the following referred as *without-sva*)
- **model #2** considering batch effects at a 10% FDR (in the following referred as *with-sva*). The higher false discovery rate threshold was chosen in order to have a sizeable number of differentially expressed genes.

In order to make SWIM results comparable with the results obtained in the original study [3], we used for both linear models the same pipeline of authors to compute the differentially expressed genes (DEGs), which consists in using R statistical software (v 3.4.4) and the package limma, instead of using SWIM DEGs calculation steps (steps 1-2 of Fig. 1). Thus, the linear models were fitted to the expression data for each probe by using least squares regression and considering age, sex and race as clinical phenotypes, and pack years as further covariate variable. Then, an empirical Bayes shrinkage method was used to obtain a moderated t-test statistic and its p-value in limma. Adjustment for multiple testing controlled for false discovery rate (FDR).

Finally, in the linear regression model #2, microarray batch effects were identified by using surrogate variables, obtained via the R/Bioconductor package SVA, as further covariate variables.

Model #	Variable of interest	Model
1	COPD	EXP ~ COPD + age + sex + race + pack-years
2	COPD	EXP ~ COPD + age + sex + race + pack-years + surrogate_variables

Abbreviations: EXP = expression values, COPD = Chronic Obstructive Pulmonary disease

Table 1. Linear regression models for association with outcome of interest

3. Results and discussion

In this section we present and discuss the results obtained by SWIM for model #1 without considering the batch effects (without-sva) and model #2 considering batch effects (with-sva).

3.1 Model #1 *without-sva*

Starting from 32831 probes, we obtained 1011 significantly differentially expressed (DE) probes (FDR < 0.05) mapped to 963 genes (Fig. 2).

We found 586 DE probes (58%) down-regulated in COPD cases and the remaining 425 DE ones (42%) up-regulated (Fig. 2a).

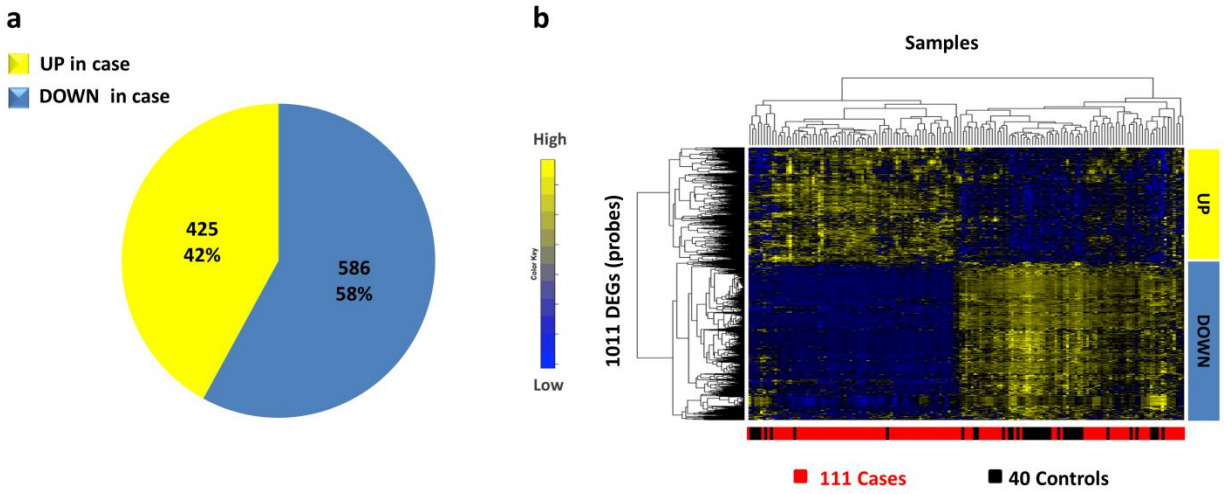

Figure 2. Differentially expressed probes. (a) Pie chart represents the percentages of differentially expressed probes that are up-/down-regulated in COPD cases in comparison to control subjects. (b) Differentially expressed profiles are clustered according to probes (rows) and samples (columns) by using Pearson correlation distance as metrics. Heat map colors represent different expression levels increasing from blue to yellow.

The 1011 DE probes include 197 DE probes out of 214 DE probes of the original work [3], about 92%. This made us quite confident that we were not losing basic information avoiding to consider batch effect but with the advantage of gaining a greater number of differentially expressed genes (DEGs).

The top differentially expressed probe was ILMN_3187508 mapped to KRT18P13 gene (referred as FLJ40504 gene in Table 1 of the original work [3], where it is the top 2 of DEGs). The top 1 DEG of the original work [3], HMGB1, falls in the top 4 of our DEGs.

Seven previously identified genome-wide significant COPD GWAS genes (default p-value < 10^{-5}) from the NHGRI-EBI Catalog (www.ebi.ac.uk/gwas/) were significantly differentially expressed at a 5% FDR (Table 2).

Gene name	DEG FDR	GWAS p-value
DLG2	0.006	8.00E-08
EFCAB2	0.044	2.00E-06
GOLGA8B	0.025	8.00E-08
KCNK1	0.033	2.00E-07
NNT	0.006	2.00E-06
NUDT12	0.014	2.00E-06
SPAG16	0.006	5.00E-06

Table 2. Differentially expressed COPD GWAS genes

The DEG matrix of 1011 rows (DE probes) and 151 columns (111 COPD cases + 40 control samples) was used as input of SWIM that run from step 3 (Fig. 1) to build the correlation network based on the Pearson correlation coefficient between the expression profiles of any pair of the differentially expressed probes. In this network, two nodes are connected if the absolute value of the Pearson correlation coefficient for their expression profiles is greater than a given threshold (for this study the selected threshold is 0.53 corresponding to the 75th percentile of the entire correlations' distribution). The correlation network encompasses 960 nodes corresponding to 960 probes mapped to 945 genes (Fig. 3a).

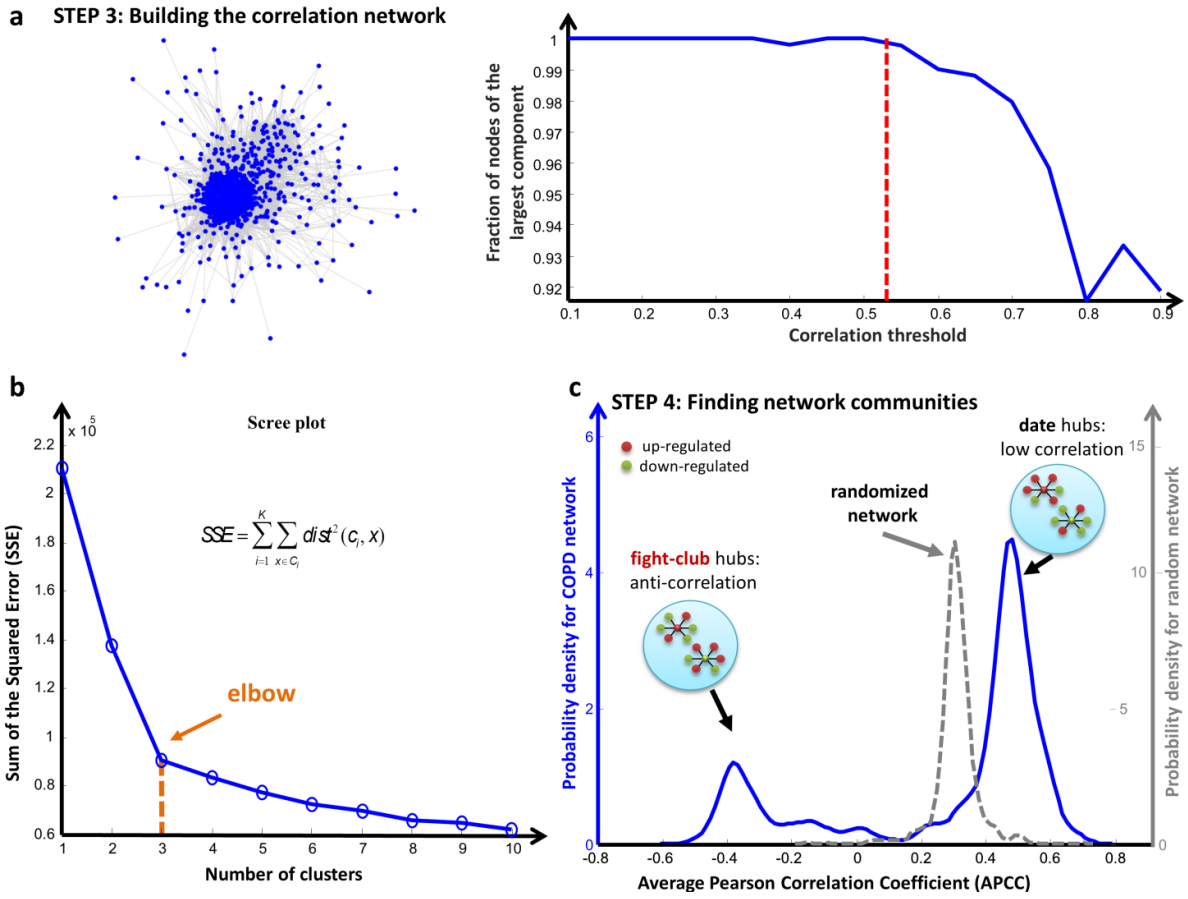

Figure 3. COPD correlation network and fight-club hubs. (a) [left] Representation of COPD correlation network; [right] Connectivity of the COPD correlation network. The x-axis represents the Pearson correlation threshold (ρ = 0.53 or 75th percentile) varying in the chosen range, while the y-axis represents the fraction of nodes populating the largest component. The dashed red lines correspond to the selected threshold. Note that y=1 means that all nodes fall in the largest component and thus the network is fully connected; otherwise more components exist. (b) Scree plot. The position of the elbow suggests a reasonable choice of the number of clusters for the k-means. (c) Probability distribution of APCC for hubs (i.e., node with degree greater than 5) identified in the correlation network built from the COPD

expression dataset (blue solid line) and in its randomized counterpart obtained by shuffling the edges but preserving the degree of each node (grey dashed line). In this case, the peak corresponding to party hubs was less evident and thus not highlighted because the number of party hubs was much lower than the number of date hubs.

In order to detect the community structure of the network, SWIM used the k-means clustering algorithm, which partitions n objects (here network nodes) into a predefined number N of clusters. The quality of clustering was evaluated by minimizing the Sum of the Squared Error (SSE), depending on the distance of each object to its closest centroid. As distance measure, SWIM used $dist(x,y) = 1 - \rho(x,y)$, where $\rho(x,y)$ is the Pearson correlation between expression profiles of nodes x and y. A reasonable choice of the number of clusters is suggested by the position of an elbow in the SSE plot computed as function of N (Fig. 3b). The final COPD correlation network consisted of N=3 clusters, ranging in size from 408 genes in module 1, 387 genes in module 2, and 126 genes in module 3.

Communities in the network were constructed to group together co-expressed genes prior to pathway analysis. Pathways enrichment tests were performed, using KEGG database, for genes in all the three clusters and we found that pathways enriched in the cluster 2 are more specific demonstrating enrichment for B cell related processes (p-value = 2.62e-03, FDR = 0.45). In particular, cluster 2 includes SERPINE2, CD79A, BCL2, POU2AF1, BCL11A.

SWIM next searched for specific topological properties of the correlation network using the date/party/fight-club hub classification system [2], based on the Average Pearson Correlation Coefficients (APCCs) between the expression profiles of each hub (i.e., node with degree greater than 5) and its nearest neighbors (Fig. 3c): date hubs display low co-expression with their partners (i.e. low positive APCC values, APCC < 0.5); party hubs have a high co-expression with their partners (i.e. high positive APCC values, APCC ≥ 0.5); fight-club hubs show an inversely correlated profile with their partners (i.e. negative APCC values). In COPD network, SWIM found 216 fight-club hubs (mapped to 210 genes), 459 date hubs (mapped to 440 genes), and 240 party hubs (mapped to 234 genes).

Finally, SWIM assigned a role to each node based on its inter-cluster and intra-cluster interactions. This is reached by defining two statistics: the clusterphobic coefficient K_π and the within-module degree z_g. The clusterphobic coefficient measures the “fear” of being confined in a cluster, in analogy with the claustrophobic disorder. The global within-module degree z_g measures how “well-connected” each node is to other nodes in its own community. According to K_π and z_g values, the plane is divided into seven regions (R1-R7), each defining a specific node role. High z_g values correspond to nodes that are hubs within their module (local hubs), while high values of K_π identify nodes that interact mainly outside their community, i.e. having much more external than internal links. SWIM colored each node in the plane identified by z_g and K_π according to its APCC value, thus defining a heat cartography map (Fig. 4).

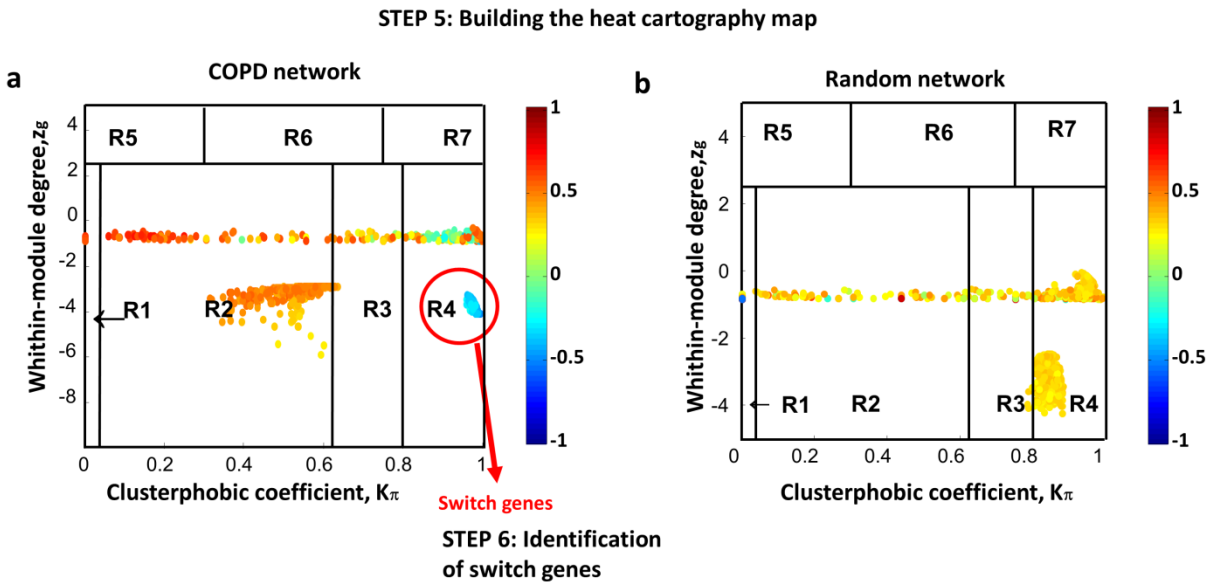

Figure 4. COPD switch genes. (a-b) Heat cartography maps of COPD and randomized network. Dots correspond to nodes in the networks. Each node is colored according to the value of the APCC between its expression profile and that of its nearest neighbors in the network.

SWIM identifies "switch genes" as a special subclass of fight-club hubs falling in R4 region. In particular, switch genes have to satisfy the following topological and expression features:

1. being not a hub in their own cluster (z_g<2.5)
2. having many links outside their own cluster (K_π > 0.8)
3. having a negative average weight of their incident links (APCC < 0)

We found 209 switch genes in COPD correlation network, i.e. 215 mapped probes. All switch genes except one (CREB5) resulted up-regulated in COPD (Fig. 5a).

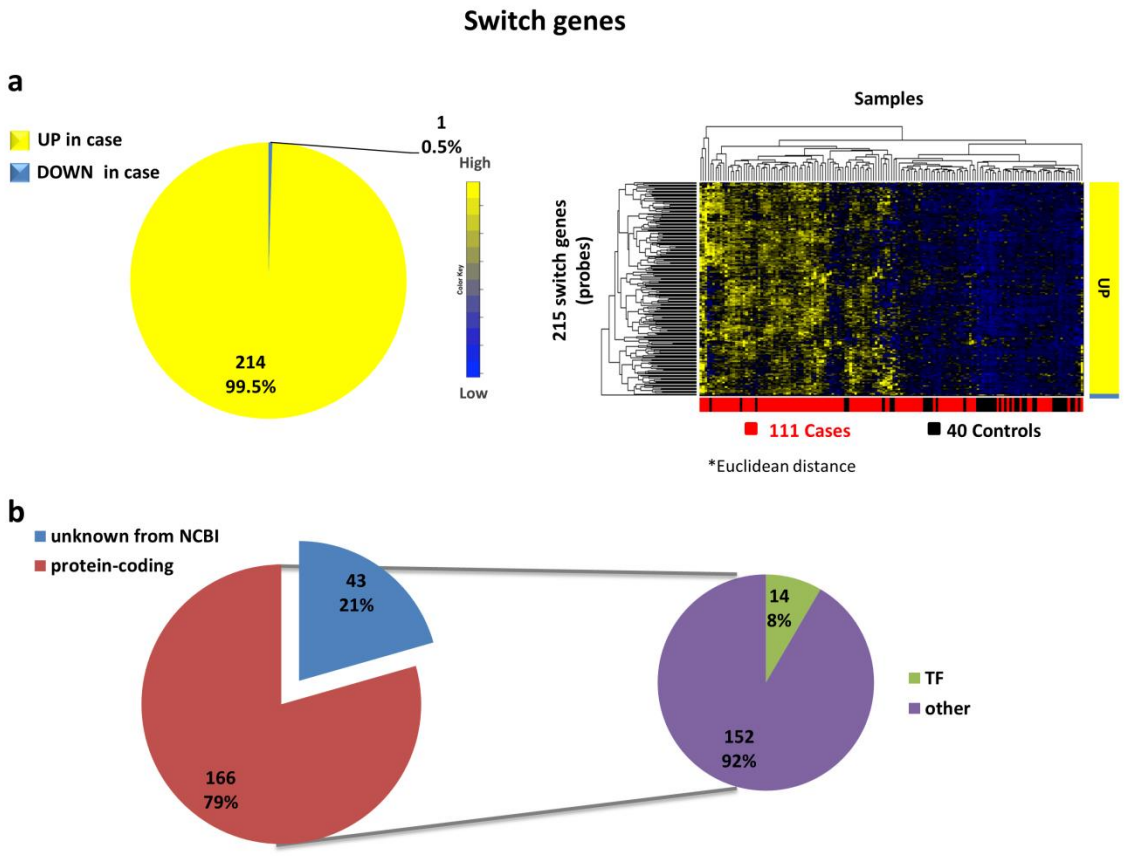

Figure 5. Characterization of COPD switch genes. (a) [left] Pie chart represents the percentages of switch genes (in term of probes) that are up-/down-regulated in COPD cases in comparison to control subjects. [right] Switch genes are clustered according to probes (rows) and samples (columns) by using Euclidean distance as metrics. Heat map colors represent different expression levels increasing from blue to yellow. (b) The larger pie chart [left] represents the classification of switch genes according with their molecular type. The smaller pie charts [right] highlight the number of transcription factors among the protein coding switch genes.

The list of switch genes included 166 protein coding, among which 14 transcription factors (Fig. 5b) that are listed in Table 3.

ID	symbol	pval	FDR	logFC	direction	module
ILMN_1674399	ZNF143	1.01E-07	0.000427413	0.469363476	UP	3
ILMN_1758705	IRX6	1.97E-05	0.006221702	0.826415201	UP	1
ILMN_1802457	MAX	0.000257482	0.020618042	0.691155442	UP	1
ILMN_1786015	CTCF	0.000570573	0.031966669	0.362498054	UP	2
ILMN_2182647	PINX1	0.00089779	0.039831543	0.496870123	UP	1
ILMN_1702125	HOXB7	0.000943855	0.040827344	0.554879367	UP	2
ILMN_1736954	ZBTB7B	0.000956897	0.040900751	0.439518737	UP	2
ILMN_1652486	THAP7	0.00096412	0.040948302	0.458378619	UP	3
ILMN_1728677	CREB5	0.000981018	0.04128695	-0.993689388	DOWN	1
ILMN_1786658	BOLA3	0.00113289	0.043552576	0.296349025	UP	1
ILMN_1694603	SMARCC1	0.001143712	0.043792857	0.160122239	UP	2
ILMN_1745365	PKNOX1	0.001226047	0.045227358	0.269893723	UP	2
ILMN_1798808	FOXH1	0.001300487	0.046441584	0.312860548	UP	2
ILMN_1806713	ZNF18	0.001508128	0.049781968	0.244776105	UP	1

Table 3. Switch genes that show a transcription factor activity

The results of module classification in the COPD correlation network, let us to test whether one or more switch genes are highly correlated or anti-correlated with one or more of the genes included in module 2, such as SERPINE 2, CD79A, BCL2, POUF2AF1, BCL11A. We found some switch genes that are strongly positively correlated with them (Table 4).

Positive nearest neighbors in cluster 2	Switch genes
SERPINE 2	PPAT ALDH18A1
CD79A	EEF1B2 SP140 SSR4 LOC100188949
BCL2	TOP2B RAPGEF6 ZNF143 SLC46A3 LOC730246 TMC8 LIX1L EPRS
POU2AF1	ARMET EEF1B2 SSR4 SP140 LOC100188949
BCL11A	TSSC1

Table 4 Positive nearest neighbors in cluster 2

None of the previously identified genome-wide significant COPD GWAS genes (default p-value < 10^{-5}) from the NHGRI-EBI Catalog (www.ebi.ac.uk/gwas/) were identified as switch genes. However, we found some switch genes that were highly negatively correlated with six of them (i.e. EFCAB2, GOLGA8B, KCNK1, NNT, NUDT12, SPAG16).

The top 26 significant switch genes include 9 differentially expressed genes of the original work [3], with ZNF143 as the top one (Fig. 6a-b).

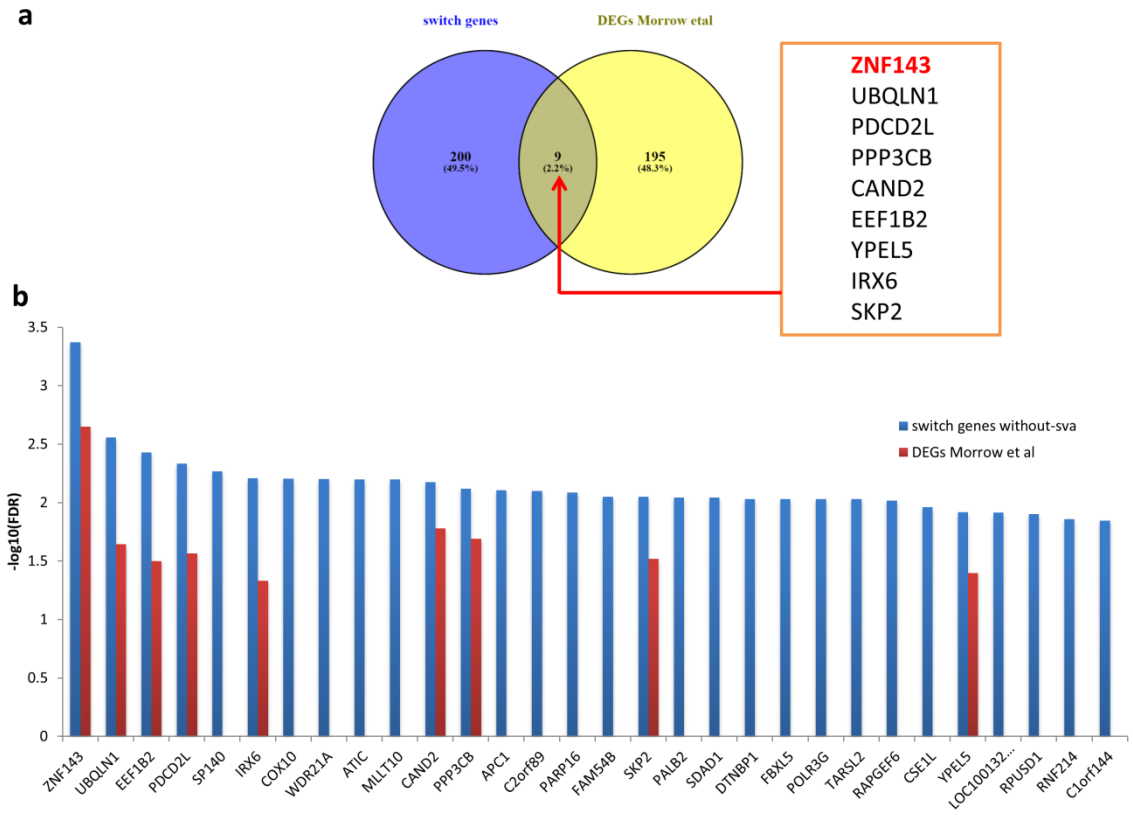

Figure 6. Switch genes and DEGs of the original work [3]. (a) Venn diagram of switch genes and DEGs of the original work [3]. (b) Histogram of FDR (-log10 transformation) for the first 30 statistically significant switch genes (blue bars) that included all the 9 DEGs (red bars) shared with the original work.

ZNF143 is up-regulated in COPD cases and shows a log 2 fold-change (FC) of 0.47 (Fig. 7).

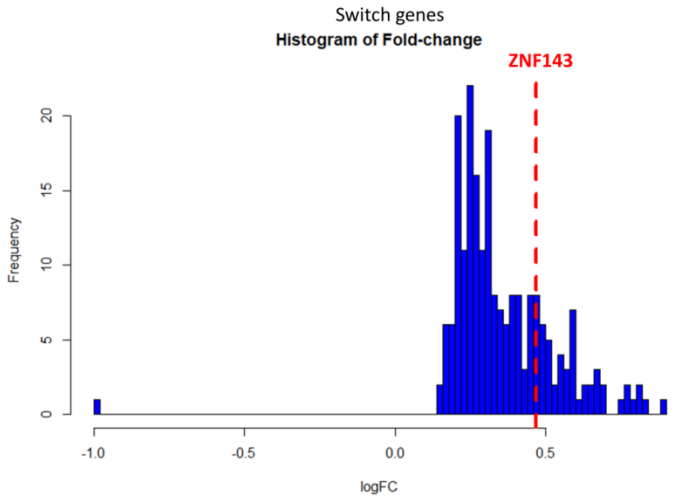

Figure 7. Fold-change histogram for COPD switch genes. The position of ZNF143 switch gene in the log 2 FC distribution is highlighted.

Among the negative nearest neighbors of ZNF143 (i.e. the set of adjacent nodes of ZNF143 in the correlation network that are highly anti-correlated with ZNF143, $\rho < -0.53$), we found 7 of the top 20 DEGs of the

original work [3], including their top one gene HMGB1 (Fig. 8 top right) identified as an important interacting partner of AGER, a gene implicated by GWAS for emphysema susceptibility.

Among the positive nearest neighbors of ZNF143 (i.e. the set of its adjacent nodes of ZNF143 in the correlation network that are highly positively correlated with ZNF143, $\rho > 0.53$), we found 7 DEGs of the original work [3], including the BCL2 implicated in COPD via regulation of apoptosis (Fig. 8 bottom right).

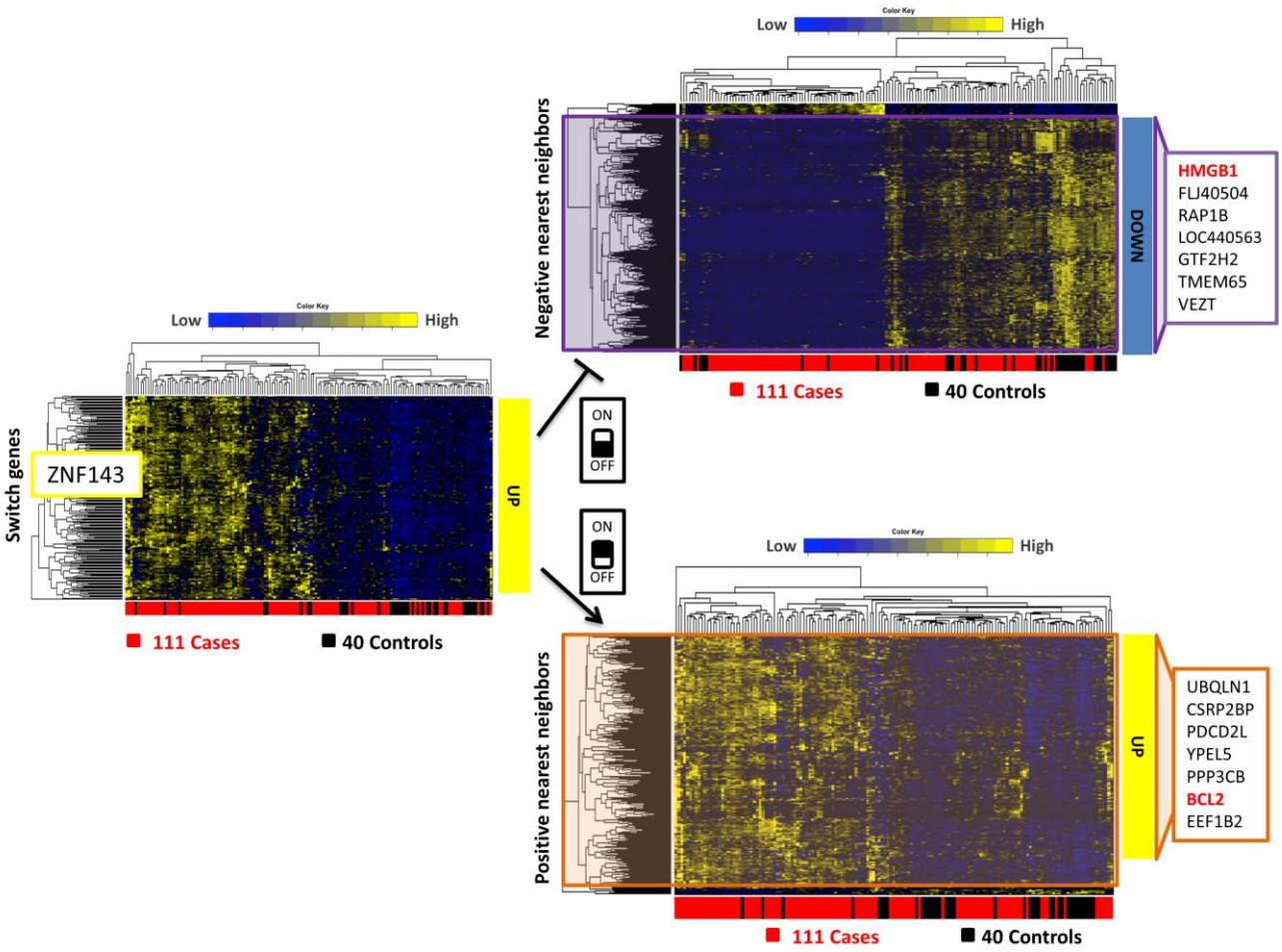

Figure 8. Heatmap of switch genes, their negative/positive nearest neighbors. Genes are clustered according to probes (rows) and samples (columns) by using Euclidean distance as metrics. Heat map colors represent different expression levels increasing from blue to yellow. The top statistically significant switch gene ZNF143 is highlighted (left) together with its negative (top right) and positive (bottom right) nearest neighbors.

Finally, we studied the effect of targeted removal of date/party/fight-club hubs and switch genes on the COPD correlation network topology (Fig. 9). Strikingly, the removal of 15% fight-club produces a drastic increase of the average shortest path (i.e. where the shortest path between two nodes is the minimum number of edges connecting them and the average shortest path is the mean of the shortest paths for all possible pairs of nodes in the network), indicating a very rapid disintegration of the network into multiple components (Fig. 9a). This behavior is very similar to the effect caused by the deletion of date hubs that are known to be higher-level connectors between groups. On the contrary, the random removal of nodes does not affect the integrity of the network letting almost unchanged the average shortest path. Evaluating the contribute of switch genes to the robustness of the network, we found a drastic effect upon removal of switch genes (Fig. 9b) that mirrors the behavior observed for fight-club hubs, as expected because 99% of them are switch genes.

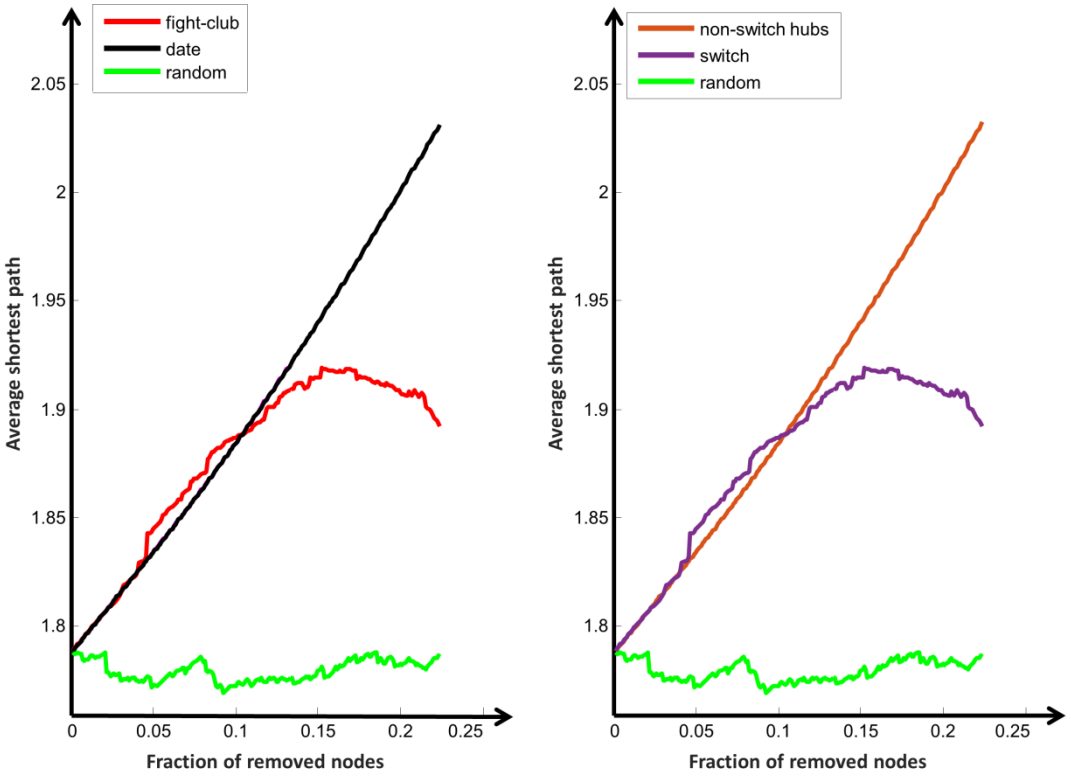

Figure 9. Robustness analysis for COPD correlation network. (a)-(b) For each class of hubs, nodes are sorted by decreasing degree and the first 215 (i.e. the number of switch) sorted nodes are selected to be removed. Then, the cumulative node deletion is computed by class (i.e. date hubs, fight-club hubs, switch genes, non-switch hubs, and randomly chosen nodes). The x-axis represents the cumulative fraction of removed nodes with respect to the total number of network nodes that is 960, while the y-axis represents the average shortest path. Each curve corresponds to the variation of the average shortest path of the COPD correlation network as function of the removal of nodes specified by the color of each curve. Note that party hubs curve is not shown because its contribution to the average shortest path resembles the contribution of random removal.

A summary Table of the SWIM run is provided in Table 5.

SWIM run for COPD cases vs control without-sva	
FDR threshold	0.05
Number of DE probes	1011 (mapped to 963 genes)
Pearson Correlation threshold	0.53 (75^{th} prc)
Number of network nodes	960 (mapped to 915 genes)
Number of fight club hubs	216 (mapped to 210 genes)
Number of date hubs	459 (mapped to 440 genes)
Numer of party hubs	240 (mapped to 234 genes)
Number of clusters	3
Number of switch genes	215 (mapped to 209 genes)

Table 5. Summary of SWIM running parameters

3.2 Model #2 *with-sva*

Starting from 32831 probes, we obtained 416 significantly differentially expressed (DE) probes (FDR < 0.1) mapped to 397 genes (Fig. 10), which are included for 88% (351 out of 397) in DEGs obtained by considering model #1 *without-sva*.

We found 263 DE probes (63%) down-regulated in COPD cases and the remaining DE 153 ones (37 %) up-regulated (Fig. 10a).

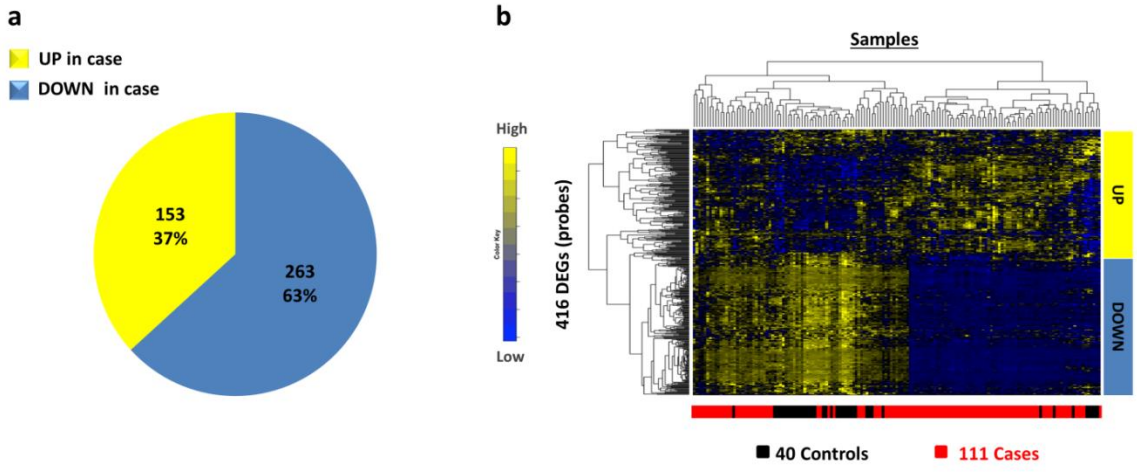

Figure 10. Differentially expressed probes. (a) Pie chart represents the percentages of differentially expressed probes that are up-/down-regulated in COPD cases in comparison to control subjects. (b) Differentially expressed profiles are clustered according to probes (rows) and samples (columns) by using Pearson correlation distance as metrics. Heat map colors represent different expression levels increasing from blue to yellow.

Four previously identified genome-wide significant COPD GWAS genes (default p-value < 10^{-5}) from the NHGRI-EBI Catalog (www.ebi.ac.uk/gwas/) were significantly differentially expressed at 5% FDR (Table 6).

Gene name	DEG FDR	GWAS p-value
DLG2	0.038	8.00E-08
ELMO1	0.029	4.00E-06
NNT	0.013	2.00E-06
SPAG16	0.054	5.00E-06

Table 6. Differentially expressed COPD GWAS genes

The DEG matrix of 416 rows (DE probes) and 151 columns (111 COPD cases + 40 control samples) was used as input of SWIM that run from step 3 (Fig. 1) to build the correlation network based on the Pearson correlation coefficient between the expression profiles of any pair of the differentially expressed probes. For this study the selected threshold is 0.5 corresponding to the 75th percentile of the entire correlations' distribution (Fig. 11a). The correlation network encompasses 369 nodes corresponding to 369 probes mapped to 355 genes, which are included for 92% (325 out of 355) in the network nodes of *without-sva* study.

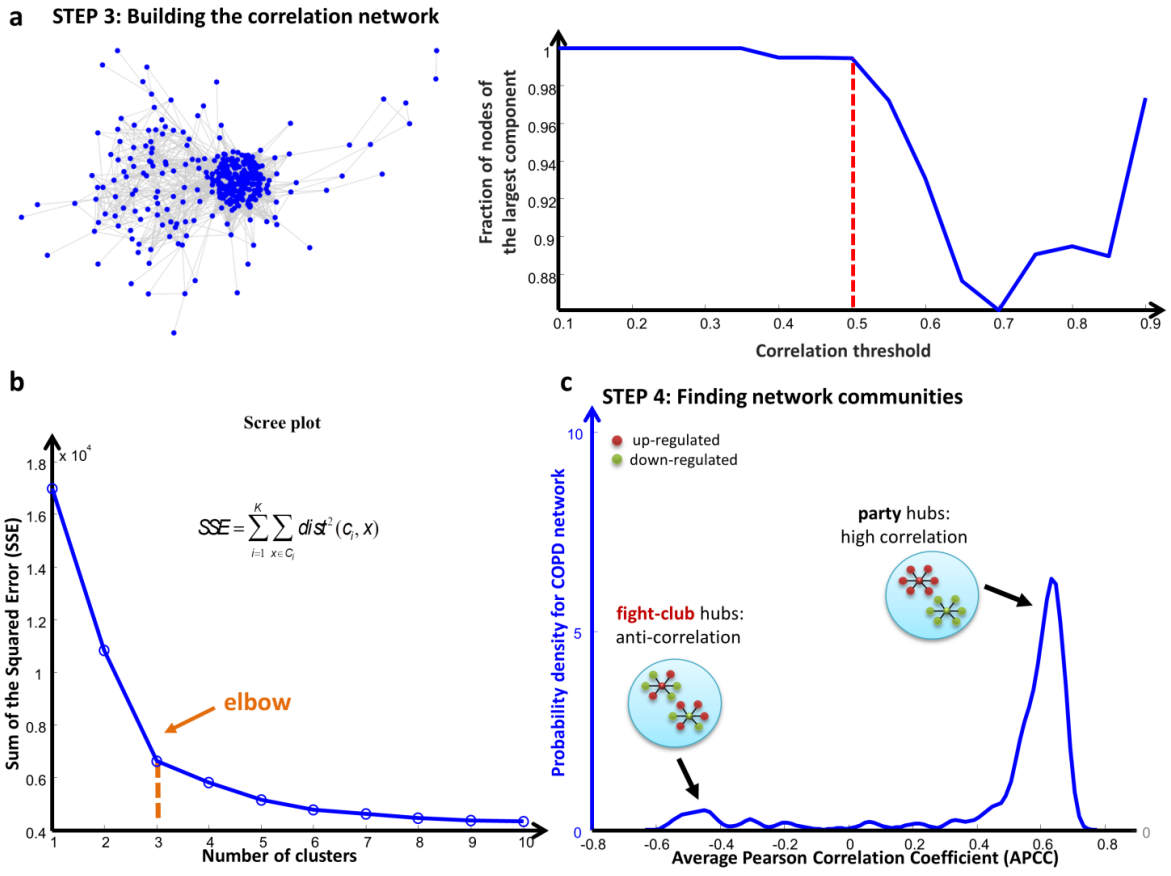

Figure 11. COPD correlation network and fight-club hubs. (a) [left] Representation of COPD correlation network; [right] Connectivity of the COPD correlation network. The x-axis represents the Pearson correlation threshold (ρ = 0.53 or 75th percentile) varying in the chosen range, while the y-axis represents the fraction of nodes populating the largest component. The dashed red lines correspond to the selected threshold. Note that y=1 means that all nodes fall in the largest component and thus the network is fully connected; otherwise more components exist. (b) Scree plot. The position of the elbow suggests a reasonable choice of the number of clusters for the k-means. (c) Probability distribution of APCC for hubs (i.e., node with degree greater than 5) identified in the correlation network built from the COPD expression dataset. In this case, the peak corresponding to date hubs was less evident and thus not highlighted because the number of date hubs was much lower than the number of party hubs.

The SWIM step of partitioning the network in communities led to a final COPD correlation network consisting of N=3 clusters (Fig. 11b), ranging in size from 26 genes in module 1, 155 genes in module 2, and 174 genes in module 3. Once again, we found an enrichment in KEGG pathways for B cell related processes (p-value = 3e-03, FDR = 0.46) for genes in cluster 2. As before, cluster 2 includes SERPINE2, CD79A, BCL2, POU2AF1, BCL11A.

The SWIM step of classifying nodes in date/party/fight-club hubs (Fig. 11c) led to 37 fight club hubs (mapped to 37 genes), 37 date hubs (mapped to 37 genes), and 258 party hubs (mapped to 248 genes).

Finally, the SWIM step of building the heat cartography map (Fig. 12) led to the identification of 33 switch genes in COPD correlation network, i.e. 33 mapped probes.

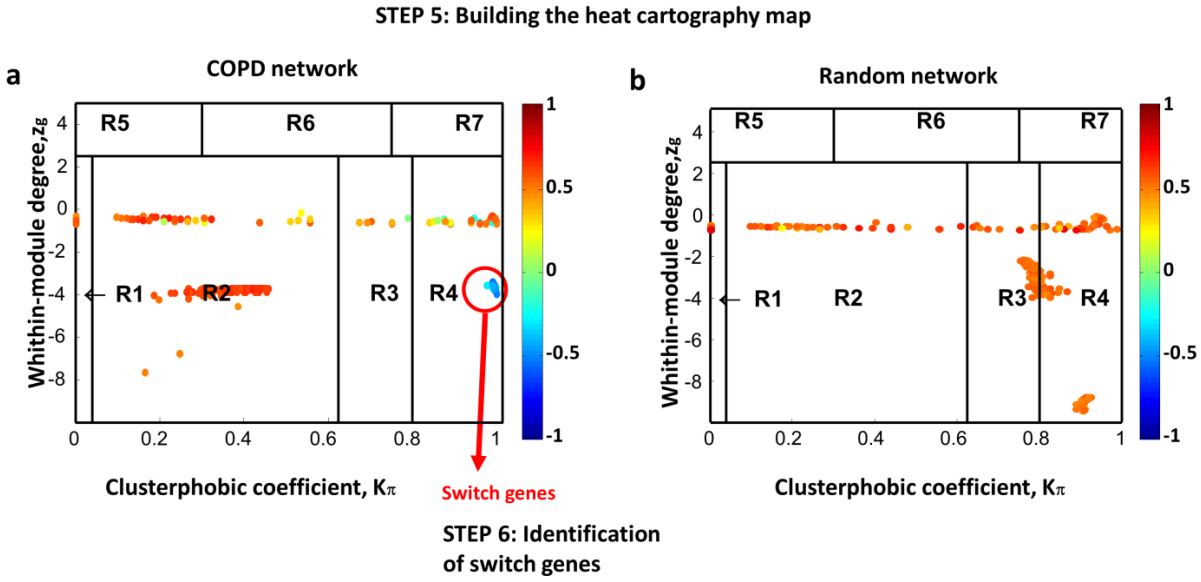

Figure 12. COPD switch genes. (a-b) Heat cartography maps of COPD and randomized network. Dots correspond to nodes in the networks. Each node is colored according to the value of the APCC between its expression profile and that of its nearest neighbors in the network.

Note that the 88% (29 out of 33) of switch genes here identified are included in first top 55 significantly switch genes obtained in the model #1 *without-sva* (Fig. 13).

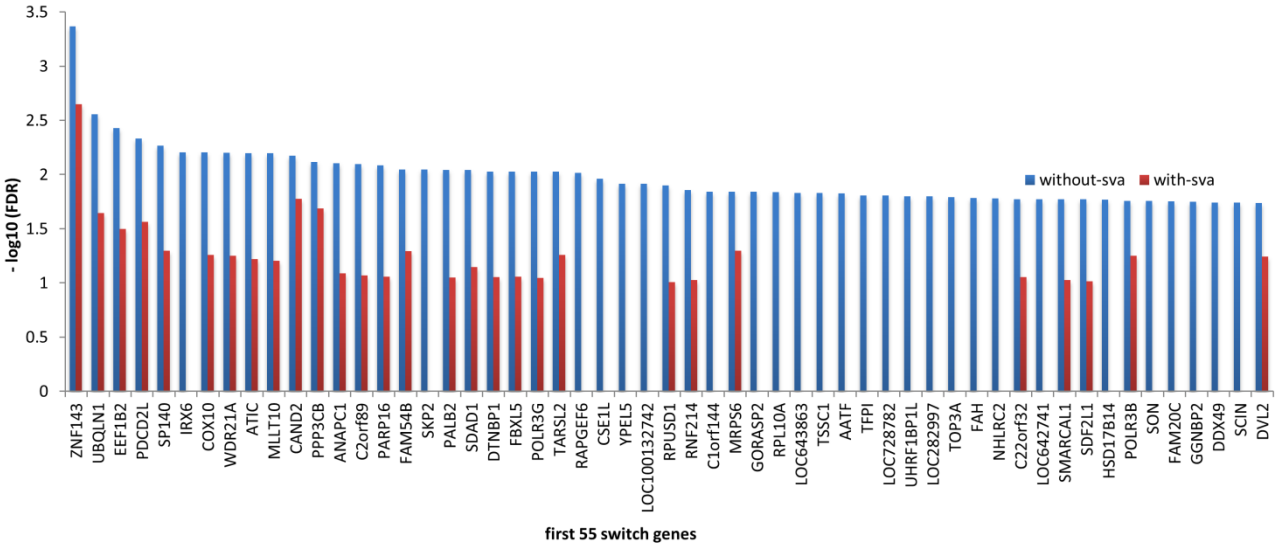

Figure 13. Switch genes in *without-sva* and *with-sva* models. Histogram of FDR (-log10 transformation) for switch genes obtained from model #1 *without-sva* (blue bars) and model #2 *with-sva* (red bars). The x axis shows the first 55 most statistically significant switch genes of the *without-sva* analysis that include all the 29 common switch genes obtained from *with-sva* analysis.

All switch genes except one (E2F6) resulted up-regulated in COPD (Fig. 14a).

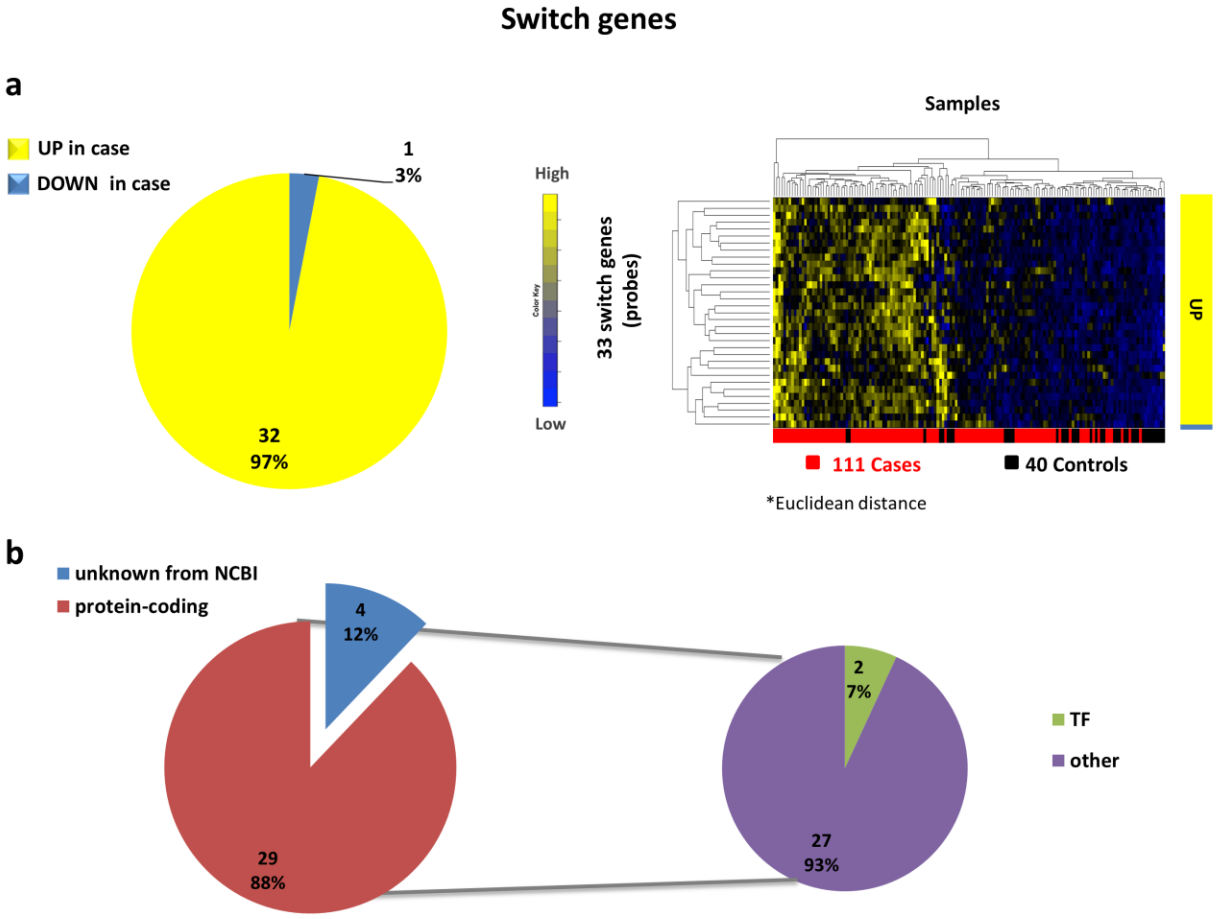

Figure 14. (a) [left] Pie chart represents the percentages of switch genes (in term of probes) that are up-/down-regulated in COPD cases in comparison to control subjects. [right] Switch genes are clustered according to probes (rows) and samples (columns) by using Euclidean distance as metrics. Heat map colors represent different expression levels increasing from blue to yellow. (b) The larger pie chart [left] represents the classification of switch genes according with their molecular type. The smaller pie charts [right] highlight the number of transcription factors among the protein coding switch genes.

The list of switch genes included 29 protein coding, among which 2 transcription factors (Fig. 14b) that are listed in Table 7.

ID	symbol	pval	FDR	logFC	direction	module
ILMN_1674399	ZNF143	1.40E-06	0.002241	0.346666	UP	1
ILMN_1656196	E2F6	0.000106	0.027667	-0.28345	DOWN	1

Table 7. Switch genes that show a transcription factor activity

By considering the results of module classification in the COPD correlation network, once again we found some switch genes that are strongly positively correlated with the genes included in module 2, such as CD79A, BCL2, POUF2AF1, BCL11A (Table 8).

Positive nearest neighbors in cluster 2	Switch genes
CD79A	EEF1B2 SP140 SDF2L1
BCL2	EEF1B2 ZNF143
POU2AF1	EEF1B2 SP140
BCL11A	SP140

Table 8 Positive nearest neighbors in cluster 2

None of the previously identified genome-wide significant COPD GWAS genes (default p-value < 10^{-5}) from the NHGRI-EBI Catalog (www.ebi.ac.uk/gwas/) were identified as switch genes. However, we found some switch genes that were highly negatively correlated with two of them (i.e. NNT, SPAG16).

The top 7 switch genes include 7 differentially expressed genes of the original work [3], with ZNF143 that is confirmed as the top one (Fig. 15).

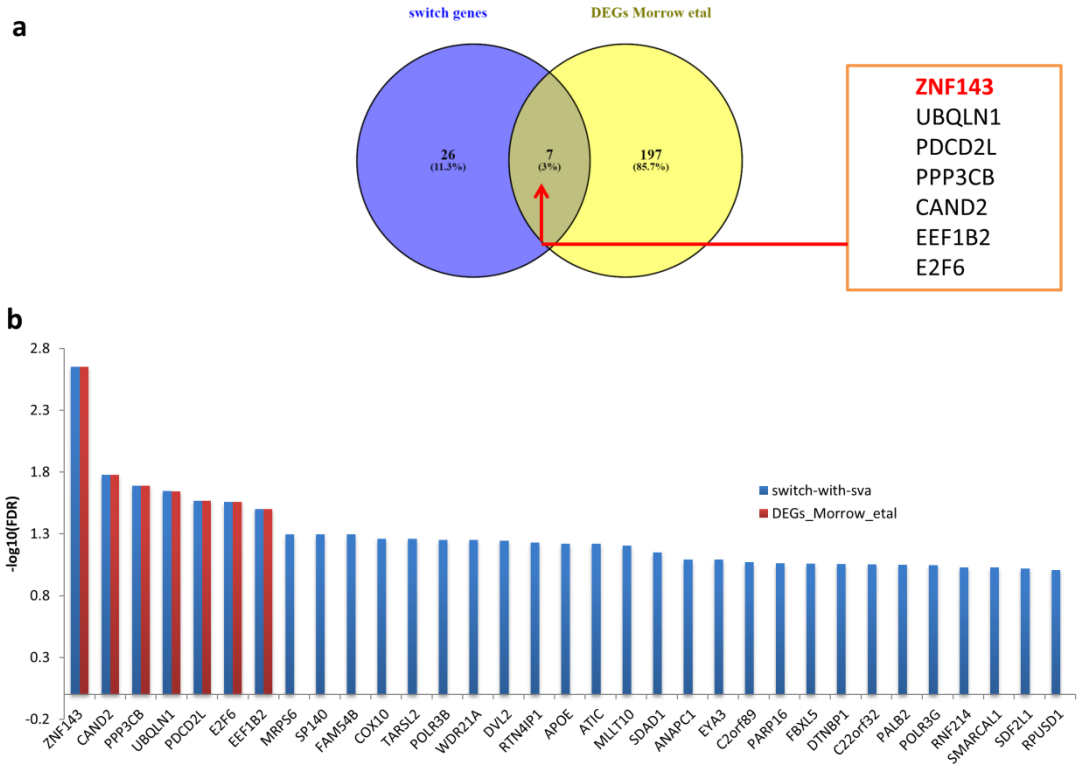

Figure 15. Switch genes and DEGs of the original work [3]. (a) Venn diagram of switch genes and DEGs of the original work [3]. (b) Histogram of FDR (-log10 transformation) for all switch genes (blue bars) that included the 7 DEGs (red bars) shared with the original work.

ZNF143 is up-regulated gene in COPD cases and shows a log 2 FC of 0.35 (Fig. 16).

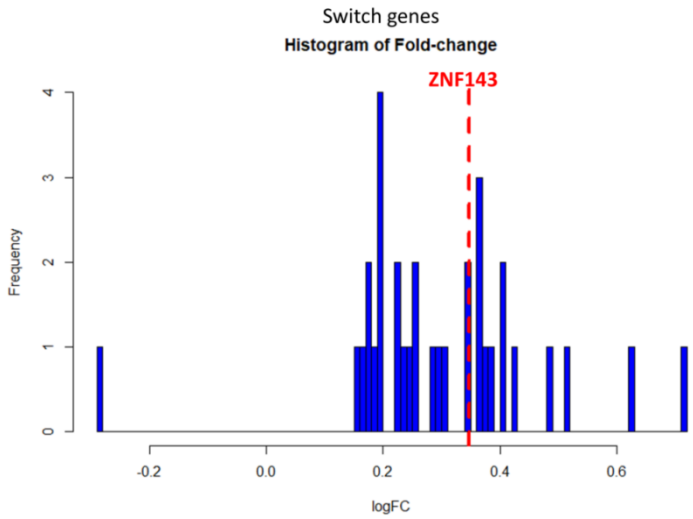

Figure 16. Fold-change histogram for COPD switch genes. The position of ZNF143 switch gene in the log 2 FC distribution is highlighted.

Among the negative nearest neighbors of ZNF143 (i.e. nodes highly anti-correlated with ZNF143, $\rho < -0.5$), we found 5 of the top 20 DEGs of the original work [3] (Fig. 17 top right). These 5 genes are included in the list of 7 genes obtained from model #1 *without-sva* as negative nearest neighbors of ZNF143 (only HMGB1 and FLJ40504 are missing).

Among the positive nearest neighbors of ZNF143 (i.e. nodes highly positively correlated with ZNF143, $\rho > 0.5$), we found the same 7 DEGs of the original work [3] obtained from model #1 *without-sva*.

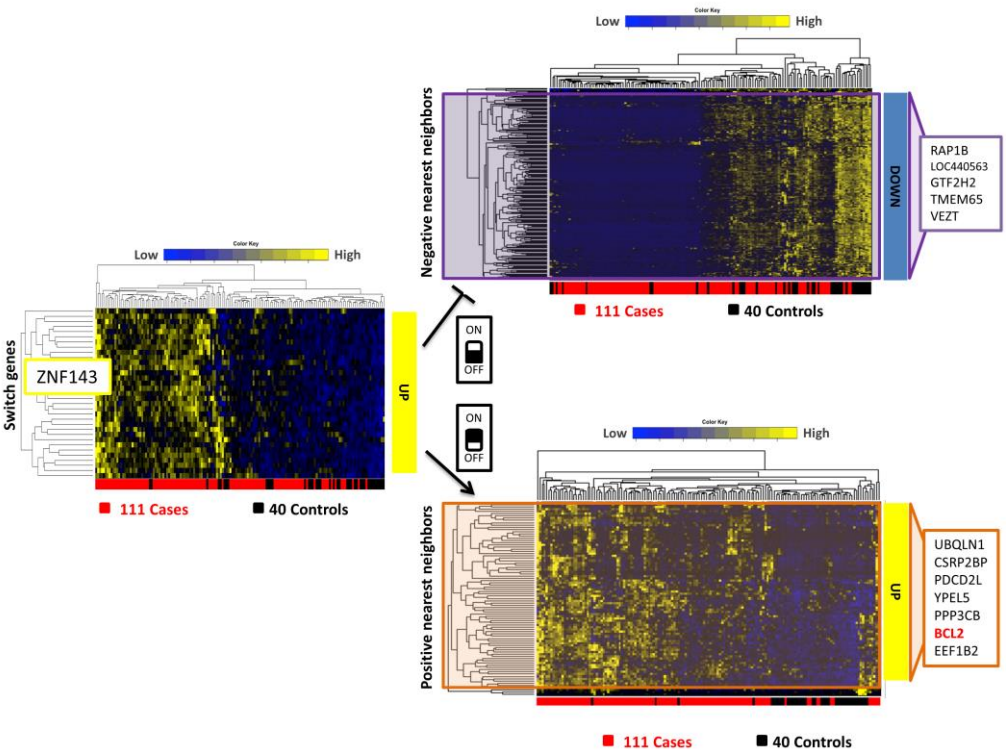

Figure 17. Heatmap of switch genes, their negative/positive nearest neighbors. Genes are clustered according to probes (rows) and samples (columns) by using Euclidean distance as metrics. Heat map colors represent different expression levels increasing from blue to yellow. The top statistically significant switch gene ZNF143 is highlighted (left) together with its negative (top right) and positive (bottom right) nearest neighbors.

Finally, the SWIM analysis for evaluating the effect of targeted removal of date/party/fight-club hubs and switch genes on the COPD correlation network topology showed, once again, a critical contribution of switch genes in preserving the integrity of the network, mirroring the effect caused by the deletion of fight-club hubs, as expected because 89% of them are switch genes (Fig. 17).

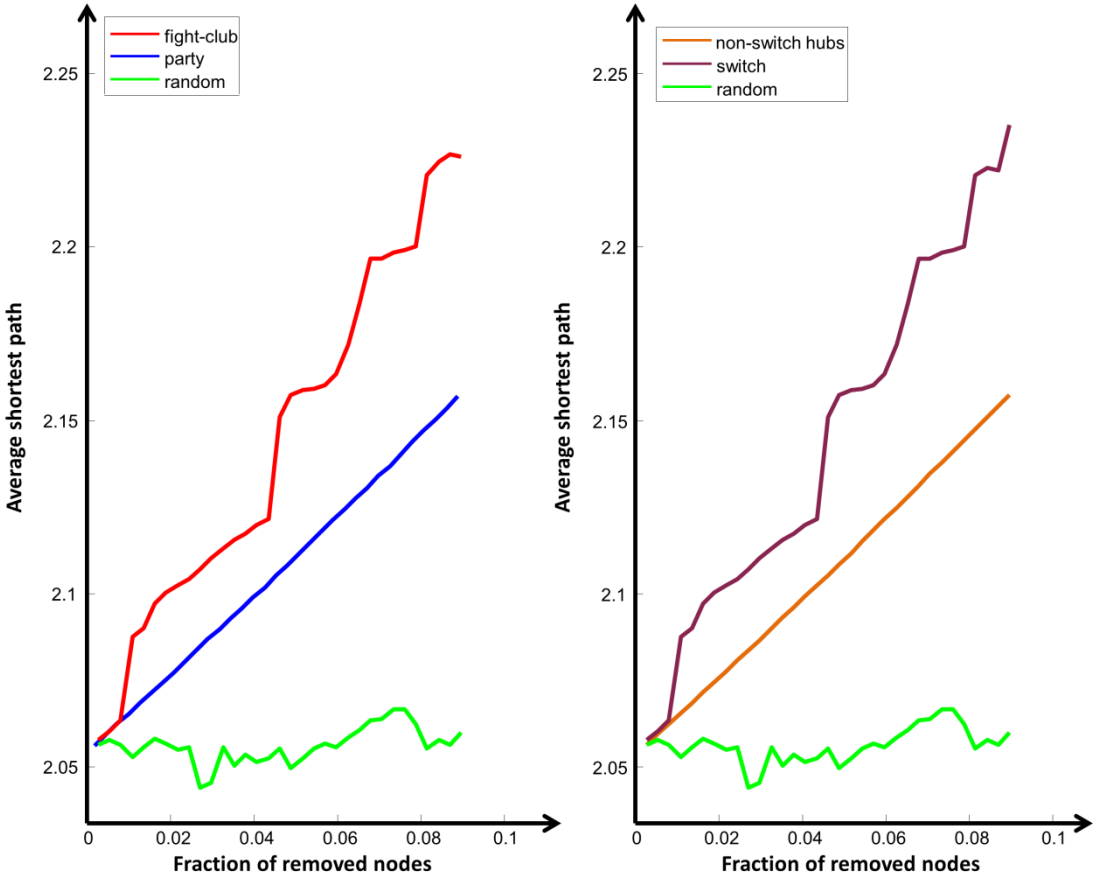

Figure 17. Robustness analysis for COPD correlation network. (a)-(b) For each class of hubs, nodes are sorted by decreasing degree and the first 33 (i.e. the number of switch) sorted nodes are selected to be removed. Then, the cumulative node deletion is computed by class (i.e. party hubs, fight-club hubs, switch genes, non-switch hubs, and randomly chosen nodes). The x-axis represents the cumulative fraction of removed nodes with respect to the total number of network nodes that is 369, while the y-axis represents the average shortest path. Each curve corresponds to the variation of the average shortest path of the COPD correlation network as function of the removal of nodes specified by the color of each curve. Note that date hubs curve is not shown because its contribution to the average shortest path resembles the contribution of random removal.

A summary Table of the SWIM run is provided in Table 9.

SWIM run for COPD cases vs control with-sva	
FDR threshold	0.1
Number of DE probes	416 (mapped to 397 genes)
Pearson Correlation threshold	0.5 (75^{th} prc)
Number of network nodes	369 (355 genes)
Number of fight club hubs	37 (mapped to 37 genes)
Number of date hubs	37 (mapped to 37 genes)
Numer of party hubs	258 (mapped to 248 genes)
Number of clusters	3
Number of switch genes	33 (mapped to 33 genes)

Table 9. Summary of SWIM running parameters

To conclude, from our analysis it turned out that without considering batch effects the noteworthy features of SWIM analysis remains unchanged (Table 10).

	without-sva	with-sva	% with-sva in without -sva
DEGs	963	397	88%
Correlation network nodes	915	355	92%
Switch genes	209	33	88%

Table 10 Comparison between *without-sva* and *with-sva* study

4. Conclusions

We run SWIM on the microarray gene expression profiling of a large sample of resected lung tissues from subjects with severe COPD [3]. Our aim was to find switch genes in the comparison between 111 COPD cases and 40 control smokers. We found 397 differentially expressed genes (DEGs) at a 10% FDR; 4 of them – DLG2, ELMO1, NNT, SPAG16 - were at significant GWAS loci. From DEGs, SWIM built the COPD correlation network, in which two nodes are connected if the absolute value of the Pearson correlation coefficient for their expression profiles is greater than 0.5. This network encompasses 355 DEGs. Partitioning the COPD correlation network in communities, we found 3 modules, ranging in size from 408 genes in module 1, 387 genes in module 2, and 126 genes in module 3. In particular, module 2 was found enriched for B cell pathways, and included SERPINE2, CD79A, BCL2, POU2AF1, BCL11A that were previously considered as putative interactors of genes at COPD GWAS loci [3]. Then, SWIM identifies switch genes in the COPD correlation network satisfying the following topological features:

1. being not a hub in their own cluster
2. having many links outside their own cluster
3. having a negative average weight of their incident links

We found 33 switch genes in COPD correlation network; all switch genes except one (E2F6) resulted up-regulated in COPD case with respect to control smokers; 29 switch genes are protein coding, including 2 transcription factors, ZNF143 and E2F6.

The top differentially expressed switch gene was ZNF143 which negatively interacts in the network (i.e. highly negatively correlated) with NNT, a known COPD GWAS gene (default p-value < 10^{-5}) from the NHGRI-EBI Catalog (www.ebi.ac.uk/gwas/), and positively interacts with BCL2 (i.e. highly positively correlated)

In Table 11 we reported all differentially expressed switch genes at a 5% FDR that negatively interacts with known COPD GWAS genes.

	GWAS genes	
	NNT	SPAG16
Switch genes	PPP3CB	PPP3CB
	ZNF143	
	UBQLN1	
	PDCDL2	
	E2F6	
	EEF1B2	

Table 11 Differentially expressed switch genes at a 5% FDR that highly negatively correlated with known COPD GWAS genes

Note that these conclusion were drawn up for the model #2 *with*-sva but remain true also in model #1 *without-sva*.

References

[1] Vestbo,J. et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. Am. J. Respir. Crit. Care Med. 187, 347–65 (2013).

[2] Paci et al. “SWIM: a computational tool to unveiling crucial nodes in complex biological networks”. Scientific Reports (2017), 7: 44797.

[3] Morrow et al. “Functional interactors of three genome-wide association study genes are differentially expressed in severe chronic obstructive pulmonary disease lung tissue.” Scientific reports (2017), 7: 44232.

[4] Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics. 2010;11:733–739.