HHS Public Access

Author manuscript

IEEE/ACM Trans Comput Biol Bioinform. Author manuscript; available in PMC 2019 November 01.

Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2018 ; 15(6): 1991–1998. doi:10.1109/TCBB.2018.2858755.

"Super Gene Set" Causal Relationship Discovery from Functional Genomics Data

Zongliang Yue,
Informatics Institute, the University of Alabama at Birmingham, Birmingham, AL 35233, US. zongyue@uab.edu

Michael T. Neylon,
School of Informatics and Computing, Indiana University, Indianapolis, IN 46202, US. michael.t.neylon@gmail.com

Thanh Nguyen,
Informatics Institute, the University of Alabama at Birmingham, Birmingham, AL 35233, US. thamnguy@uab.edu

Timothy Ratliff, and
Purdue University Center for Cancer Research, West Lafayette, IN 47906, US. tlratliff@purdue.edu.

Jake Y. Chen
Informatics Institute, the University of Alabama at Birmingham, Birmingham, AL 35233, US. jakechen@uab.edu

Abstract

In this article, we present a computational framework to identify "causal relationships" among super gene sets. For "causal relationships", we refer to both stimulatory and inhibitory regulatory relationships, regardless of through direct or indirect mechanisms. For super gene sets, we refer to "pathways, annotated lists, and gene signatures", or PAGs. To identify causal relationships among PAGs, we extend the previous work on identifying PAG-to-PAG regulatory relationships by further requiring them to be significantly enriched with gene-to-gene co-expression pairs across the two PAGs involved. This is achieved by developing a quantitative metric based on PAG-to-PAG Co-expressions (PPC), which we use to infer the likelihood that PAG-to-PAG relationships under examination are causal—either stimulatory or inhibitory. Since true causal relationships are unknown, we approximate the overall performance of inferring causal relationships with the performance of recalling known r-type PAG-to-PAG relationships from causal PAG-to-PAG inference, using a functional genomics benchmark dataset from the GEO database. We report the area-under-curve (AUC) performance for both precision and recall to be 0.81. By applying our framework to a myeloid-derived suppressor cells (MDSC) dataset, we further demonstrate that this framework is effective in helping build multiscale biomolecular systems models with new insights on regulatory and causal links for downstream biological interpretations.

Keywords

super gene set; causal; PAG; systems biology

1 Introduction

Gene-set, network, and pathway analysis (GNPA) (1) has become the first choice for gaining insight into the underlying biology of differentially expressed genes and proteins (2). Briefly, GNPA helps the biologists explain the experimental results from existing biological domain-knowledge and potentially discover new biological mechanisms (2-4). GNPA analysis could be conducted in two directions. In the first direction, from the significant gene list retrieved from statistical analysis on the experimental results, the biologists search for which biological phenomenon (also called Gene Ontology (5)) enriching in the list (6-8), or which well-annotated pathways being 'similar' to the gene list (9-12). In the second direction, instead of highlighting the gene list from the experimental results, the biologists compare the gene expression patterns between the well-annotated gene set from the literature and the one showed in the experimental result (4,13-15). Both of these directions require robust statistical methods (2) and Gene-set-network-and pathway (which is also called PAG (16) or super-gene-set (17)) databases (16, 18-21). The development of GNPA methods and GNP databases enables shifting biological analysis from individual gene level to the PAG level. In the other words, from the data analytical point of view, the biological dataset could be represented by the PAG features instead of the gene features.

In the 'PAG paradigm', there are still many unexplored questions on the relationship among the PAGs, especially on the causal relationship. It is known that one gene may participate in multiple biological processes, and genes in different PAGs interact. The limitation among most of the well-established GNPA techniques is that these techniques often return individual PAGs from the biological input and treats the PAGs independently (2), and ignore the potential relationship among the PAGs. To answer the PAGs relationship question, in (16), we define two types of relationships: share-gene (m-type) and regulatory (r-type) relationships. Compared to the shared-gene relationship, the regulatory relationship has better explanatory power since it is derived from gene-gene regulation datasets. However, due to the noisy and incomplete of domain-knowledge on gene-gene regulation datasets, the PAG regulatory relationship is more difficult to annotate. Here, by annotation, we refer to determining the directionality of the PAG-to-PAG relationship, and whether the relationship is stimulatory or inhibitory. Integrating the gene expression profile into domain-knowledge interactome and topological data is a promising solution for this question. For example. Martini et al show that we can discover signal paths among the pathways from the expression data (22). Pepe et al, while discovering the perturbation of expression in a single pathway, shows that there exist connections among the perturbation in multiple pathways (23).

In this work, we propose a framework to enrich the annotation for r-type PAG-to-PAG relationship by integrating the gene expression profile and the existing relationships in PAGER (16). Therefore, we can form two types of network (or relationship) among the PAGs: the regulatory network and the coexpression network. Our experiment shows that the coexpression relationship could be used to re-discover the regulatory relationship with high precision and recall. Therefore, the framework name allows annotating the PAG-to-PAG relationship specific to the tissues, organs and biological condition specified in the expression profile. By this integration, we can infer the causal relationship among the PAGs,

which is the major innovation in this paper. We demonstrate the framework name in annotating cancer-specific gene expression profile and show how we can infer the causal relationship among different biological phenomenon enriched in the PAGs.

2 Methods

2.1 Recall PAGs and r-type PAG-PAG relationship from PAGER

There are three types of the PAGs in PAGER (16) A *P*-type PAG contains a connected set of molecules (genes/proteins/metabolites), among which some detail of curated mechanism of actions, e.g., protein interactions, reactions, or gene regulations, are available (24). *A*-type PAG contains a curated list of genes/proteins identified from a specific biological context, e.g. a shared GO category or a shared protein family without mechanism of actions. A G-type PAG contains a list of genes/proteins derived from any given high throughput Omics experiment, e.g. functional genomics, without annotation. (25) We give every PAG a cohesion coefficient score (CoCo score), which could be simply understood as 'PAG quality', to help assess the degree of biological relevance for each PAG. The CoCo score shows how single genes in a PAG connect: if genes inside a PAG strongly connect to the other, compared to the random connection, the PAG should have high CoCo score. The CoCo score shows how single genes in a PAG connect: if genes inside a PAG strongly connect to the other, compared to the random connection, the PAG should have high CoCo score.

Overall, the PAGER collects and organizes 18,607 regular PAGs using a three-letter-code PAG classification system. It includes 3,153 *P*-type PAGs, 8,117 A-type PAGs and 7,337 G-type PAGs. We emphasize that these counts are from the 2015 version of PAGER. We only performed the analysis on this PAGER version. All PAGs are given a cohesion coefficient score (CoCo score) to help assess the degree of biological relevance for each PAG beyond random chance. Due to the limitation of time, we were not able to perform analysis on the updated PAGER version (26), published in 2017.

PAGER also contains 72,824 regulatory relationships (r-type relationships) among the PAGs. Two PAGs are considered having r-type relationship if the gene-gene regulations between two PAGs are significantly more than random (overrepresentative). In opposite, the under-representative relationship implies that gene-gene regulations between two PAGs are significantly less than random. We quantified the r-type relationships by applying the hypergeometric distribution statistics and recorded the probability mass function (pmf) as the score. In this paper, we further narrow down the number of r-type relationship to 24,686, after setting threshold pmf <0.01.

2.2 Gene expression dataset

To construct PAG-PAG co-expression relationship from functional genomic data, we applied the microarray dataset GSE32474 (27,28). The dataset contains 174 tissue-specific samples from 59 cell-lines covering 9 types of cancer tissues: Breast, Central Nervous System, Colon, Leukemia, Melanoma, Non-Small Cell Lung, Ovarian, Prostate, and Renal from the NCI-60 panel. The dataset covers the expression of 20,638 genes.

2.3 Apply NP-HART to increase the number of PAGs and r-type relationship

To increase the number of PAGs and regulatory PAG-to-PAG relationships, we applied NP-HART (New PAGs Heuristic Algorithm Based on Relationship Topology) (17) to generate the virtual PAG (denoted as m'PAG) which have direct relationships to the regular PAGs in PAGER. A virtual PAG collects single genes having upstream or downstream regulations with genes from an existing PAGER's PAG. The NP-HART has two steps. First, we heuristically group the single genes regulating or being regulated from an existing PAGER's PAG as long as the group significantly connects to the existing PAG, threshold by pmf < 0.01. Second, if the group cohesion is too low, we heuristically remove single genes from the group until the cohesion, marked by CoCo score, reaches 0.1 and forms the virtual m'PAG. The NPHART adds 87 virtual PAGs and 130 r-type PAG-PAG relationship, which increases these two overall statistics to 38,466 PAGs and 24,816 relationships.

2.4 Characterize causal PAG-to-PAG relationships detail by E-GGCC analysis

To annotate the r-type PAG-to-PAG relationships and further infer causal relationships among the PAGs, we developed enrichment of gene-gene co-expression correlation (E-GGCC) analysis. E-CGCC analysis results show how strongly genes in two PAGs co-express either positively or negatively. Compared to existing correlation methods, our analysis use discretization method, which does not require strong assumptions on the absolute scale of expression value (29). The detail of E-GGCC analysis is as follow.

2.4.1 Score expression correlation among genes with new GGC metric—To overcome the inappropriateness of the Euclidian distance in the conventional Pearson Correlation, when the absolute expression levels of functionally related genes are highly different (29), we constructed a new Gene-Gene expression Co-expression Correlation (GGC) metric using discretization method. We started E-GGCC analysis by computing the gene-gene co-expression (GCC) with discretization. In a given microarray dataset, for each gene, we discretized the absolute gene expression into three values: −1 for low expression, 0 for nonnal expression and +1 for high expression. In this study, we discretized the top 1/3 absolute gene expression values as +1, 1/3 bottom absolute expression values as −1 and the other values as 0. For two genes a and b, let C+(a, b) be the count of samples where a and b have the same nonzero discrete expressions, let C−(a, b) be the count of samples where a and b have the opposite nonzero discrete expressions, and C*(a, b) be the count of samples in other scenarios where one of a and b has zero discrete expression. We defined the positive GGC+ (a,b) and the negative GGC−(a,b) as follows:

$$\mathrm{GGC}^{+}(a,b)=\frac{C^{+}(a,b)}{C^{+}(a,b)+C^{-}(a,b)+C^{*}(a,b)} \quad (1)$$

$$\mathrm{GGC}^{-}(a,b)=\frac{C^{-}(a,b)}{C^{+}(a,b)+C^{-}(a,b)+C^{*}(a,b)} \quad (2)$$

We united GGC+ and GGC− to calculate the final GGC as

$$\mathrm{GGC}(\mathrm{A},\mathrm{B}) = \mathrm{sign}(\mathrm{GGC}^{+}(a,b) - \mathrm{GGC}^{-}(a,b)) \times \mathrm{Max}(\mathrm{GGC}^{+}(a,b), \mathrm{GGC}^{-}(a,b)) \quad (3)$$

Here, the sign function shows the type of expression correlation between gene A and gene B. If GGC+(a, b) > GGC−(a, b), there is more evidence showing that A and B have positive expression correlation; therefore, the sign function returns positive, and vice versa. The maximum function (Max) shows the co-expression strength. After calculating the GGC scores, we removed correlation with GGC between −0.5 and 0.5 since they were insignificant coexpression. More explanation on this threshold option could be found in the results section, where we also apply network power law analysis to justify our choice of GGC threshold.

2.4.2 Score the co-expression between two PAGs by PPC metric.—Similar to the CoCo score in PAGER (16), we constructed the PAG-to-PAG Co-expression Correlation (PPC) to address the biological significance of r-type PAG-to-PAG relationship using hypergeometric distribution. Figure 1 provides one example of E-GGCC calculation. For two PAG i and j, let m(i) be the count of genes in i, m(j) be the count of genes in j and m(i, j) be the count of overlapping genes between i and j. To calculate the positive correlation (PPC +), we used the cumulative distribution function (CDF) of four parameters N, K, n and k, defined as

$$\mathrm{PPC(i,\,j)}^{+} = \mathrm{sign}\left(\frac{\mathrm{n}}{\mathrm{k}} - \frac{\mathrm{N}}{\mathrm{K}}\right) \times -\log_{10}\left(\sum_{\mathrm{t}=\mathrm{k}}^{\min(\mathrm{n},\mathrm{K})} \frac{\binom{\mathrm{K}}{\mathrm{t}}\binom{\mathrm{N}-\mathrm{K}}{\mathrm{n}-\mathrm{t}}}{\binom{\mathrm{N}}{\mathrm{n}}}\right) \quad (4)$$

where N = (m(i)+m(j)−2m(i,j)) × (m(i)+m(j)−2m(i,j)−1) / 2 is the count of theoretical gene-gene expression correlations in both i and j, K is the count of actual positive gene-gene expression correlations in both i and j, n = (m(i)−m(i,j))×(m(j)−m(i,j)) is the count of theoretical expression correlation between genes in i and genes in j, and k is the count of actual positive gene-gene expression correlations between genes in i and genes in j. To calculate the negative correlation (PPC−) between i and j, we applied (4) with the same definition for N and n, but K is the count of actual negative gene-gene expression correlations in both i and j, and k is the count of actual negative gene-gene expression correlations between genes in i and genes in j. In the right side of (4), the sum operator represents the definition of CDF in hypergeometric distribution as in (30). In the right side of (4), the sum operator represents the definition of CDF in hypergeometric distribution as in (30). Then, we estimated the positive or negative correlation using the final PPC(i, j) defined as:

$$\mathrm{PPC}(i,j)=\begin{cases}\mathrm{PPC}(i,j)^{+} & \text{if } \mathrm{PPC}(i,j)^{+} > 2*\mathrm{PPC}(i,j)^{-}\\ \mathrm{PPC}(i,j)^{-} & \text{if } \mathrm{PPC}(i,j)^{-} > 2*\mathrm{PPC}(i,j)^{+}\\ 0 & \text{otherwise}\end{cases} \quad (5)$$

2.4.3 Validate the PPC metric by the capability of recalling r-type PAG-PAG relationship.—

To validate the PPC metric, we drew the receiver-operating-characteristic curve (ROC) when using it to separate true r-type PAG-PAG relationship from the random PAG-PAG relationship. Here, the positive set contains significant overrepresentative (16) r-type PAG-to-PAG relationships in PAGER with pmf < 0.01 (refer to section 2.1). The negative set contains 10,000 under-representative r-type PAG-to-PAG relationships, and then randomly seed the false set's sample size equal to the true set's sample size. The positive set is from the causal PAG-to-PAG relationship's randomly chosen in PAGER (10). Table 1 shows how to setup confusion matrix when the PPC score is used to classify between the over-representative and the under-representative relationships. We expect that the area-under-curve (AUC) using the real expression data (GSE32474) should be significantly higher than the AUC using the random expression data, which should receive values close to 0.5. In this work, we generated random expression data by randomly set the discrete expression for each gene in each sample by −1, 0 or 1.

2.5 Case study: infer PAG-PAG causal relationship by applying E-GGCC to the myeloid-derived suppressor cells expression data set

We demonstrate the capability of the E-GGCC analysis in inferring PAG-PAG causal relationship in myeloid-derived suppressor cells case study, as showed in Figure 2. The (MDSC) related PAGs were identified at Purdue University Center for Cancer Research. Starting with the expression data for MDSC (31), we identified 1,105 differentially expressed genes, which includes 576 over-expressed genes (N+) and 529 under-expressed genes (N−), according to the methods in (17). From these differentially expressed genes, we queried PAGER with NP-HART to find the N+/N− associated and the r-type relationship among these PAGs (17). Finally, we applied the E-GGCC algorithm to annotate these r-type as 'stimulatory' (strong PPC+ score) or 'inhibitory' (strong PPC− score) and inferred the causal relationships in the r-type PAG-to-PAG's network.

3 Results

3.1 The gene-gene expression correlation identified from functional genomic data

In Figure 3, we show the distribution of the absolute value of GGC score for all possible gene-gene co-expressions, which justifies our choice of choosing −0.5 and 0.5 as the threshold for selecting GGC score. The distribution of the absolute value of GGC is close to the normal form, which implies that the distribution of all GGC is in bi-normal form. Here we choose GGC cutoff threshold to be 0.5, because we believe that it's a conveniently applicable threshold that ensures 57,576 significant expression correlations with p-value ≤ 5.86×10^{-5}. There are 57,576 correlations satisfying this threshold condition.

The gene degree of significant positive gene-gene correlation pairs and of negative gene-gene correlation pairs follow the power law shown in Figure 4. The linear regression R^2 of significant positive gene-gene expression correlation pairs is 0.93, and R^2 of significant negative gene-gene expression correlation pairs is 0.88.

3.2 The PPC score could re-discover the r-type PAG-to-PAG relationship

In Figure 5, we show that the PPC could be applied to classify between existing r-type PAG-PAG relationships and the under-representative relationships. With the GSE32474 dataset, the PPC score achieves AUC = 0.81. Meanwhile, with the random expression dataset, the PPC only achieves AUC = 0.50. This result implies that the PPC score is consistent with the real biological expression correlation patterns. We found that the optimal PPC threshold for the confusion matrix (table 1) is 1.0 to reach the precision of 0.60 and recall of 0.01. From this threshold, we were able to characterize 12,212 r-type PAG-PAG relationships as either 'stimulatory effect' (PPC+ > PPC−) or 'inhibitory effect' (PPC+ < PPC−). The statistics of PAGs, PAG-PAG relationships and characterization could be found in Table 2.

We have present top 4 significantly causal PAG-to-PAG relationships (top 8% in the pair of two PAGs' size ≤ 1000, and at least one is P-type PAG) in PAGER. We found several DNA replicate related pathways, e.g. WIG000672, WIG001985, WIG000802 have the stimulatory effect on the DNA binding or DNA repair-related PAGs, e.g. TAX001140, MAX001341 in Table 3.

3.3 PAG-PAG causal relationship in MDSC gene expression case study

In our MDSC network in Figure 6, we identify 239 PAGs and 191 PAG-PAG relationships in the MDSC as shown in Table 4. Similar to (17) we chose the central PAG as FEX001153 (16). Among these PAGs these, two m'PAGs, U'_1 (11 genes), U'_2 (7 genes) are in the upstream of FEX001153 and one is a downstream m'PAG, D'_1 (11 genes) is in the downstream of FEX001153. We found 3 causal PAG-to-PAG have stimulatory effects and 3 causal PAG-to-PAG have inhibitory effects using E-GGCC (Table 5,6). The PAG MAX003319 has a stimulatory effect on the PAG FEX001153.

We found several literatures supporting the PAG-PAG causal relationship showed in Figure 6. For the PAG MAX003319 has a stimulatory effect on the PAG FEX001153, the report (32) presented the same conclusion as B lymphocyte stimulator enhancement of humoral immune response.

There are two PAGs, MAX002771, WIG001016ve inhibitory effect on the PAG FEX001153. Yoneyama et al (33) reported that the cellular protein retinoic acid-inducible gene I (RIG-I) senses intracellular viral infection and triggers a signal for innate antiviral responses, which is strong evidence to support RIG-I/MDA5 signaling having a positive effect on immune system. Similarly, Sakaguchi et al (34) reported that FOXP3+ TReg cells can also suppress antitumor immune responses and favor tumor progression.

FEX001153 has a stimulatory effect on two PAGs, MAX001758 and MAX002449. Wherry et al (35) indicate that lineage relationship and protective immunity of memory CD8 T cell subsets, which support the point of immune system stimulatory effect on CD8 T cell.

FEX001153 has inhibitory effect on a PAG MAX001454. Müller-Schmah et al (36) revealed the mechanism of immune response acting as long-lasting disease control in spontaneous remission of MLL/AF9-positive acute myeloid leukemia response acting as long-lasting disease control in spontaneous remission of MLL/AF9-positive acute myeloid leukemia.

4 Conclusions

In this work, we presented a new way to infer causal PAG-to-PAG regulatory relationships. Our approach could be adaptive to different gene expression data sources, which are varied by subjects, treatments and phenotypes. Compared to previous work in discovering relationships among the pathways (22,23), our work has two advantages. First, we are able to integrate domain-knowledge gene-gene relationship to reduce potential noise discovery from co-expression analysis (37). Second, our discretization approach in handling the expression data requires fewer assumptions on the scale and distribution of the data. These assumptions (29,38) are required in common statistical analysis but may not be satisfied in expression data. Our framework could reveal several causal relationships among the PAGs in MDSC case study.

However, the advantages our framework associate with several limitations. First, by relying on the existing PAG− PAG relationship, our framework has limited capability in discovering novel PAG-PAG relationships. Second, by relying on simple discretization approach for co-expression analysis, our approach is less systematic and may not able to evaluate the statistical significance of the discovered gene-gene co-expression. In addition, the quality and coverage of causal relationship found in this framework depend on the comprehensiveness of the expression data. As shown in our case study, the MDSC case is independent from the GSE32474 dataset. In addition, the GSE32474 does not contain any cell lines for MDSC. This fact explains why we could only rediscover 6 causal relationships among 191 existing r-type PAG-PAG relationships. Therefore, the users should choose the expression dataset carefully and in accordant to the expected PAG-PAG relationships.

Acknowledgments

We thank Itika Arora and Nafisa Bulsara for the revision of the manuscript.

Biography

Zongliang Yue received the BS degree in life science at Capital Normal University, China. He received the MS degree in Bioinformatics in school of informatics and computing at Indiana University, US. He is a PhD student in Genetic, Genomics and Bioinformatics from

University of Alabama at Birmingham, US. His main areas of research include bioinformatics and system biology.

Michael Neylon earned a BA in Neuroscience from Vanderbilt University and an MS in Bioinformatics from Indiana University. Michael occasionally collaborates with his former graduate lab but primarily works in bioinformatics research in the pharmaceutical industry.

Thanh Nguyen received his B.S. (2012) in Computer Science at Indiana University Purdue University Indianapolis (IUPUI). He is completing his PhD program at Department of Computer science at IUPUI. He is also the visiting scientist at the Informatics Institute - the University of Alabama at Birmingham. His primary research interests include Machine Learning, Artificial Intelligence, and Data Mining. He is also interested in applying Machine Learning and AI techniques to solve challenges in Bioinformatics, System Biology and Health Informatics.

Dr. Timothy Ratliff is a Professor of Comparative pathology at department of Comparative pathology at Purdue University. He received the BS at University of Texas. He received the MS at Texas A & M. He received PhD at University of Arkansas. He focuses on Myeloid-derived Suppressor Cell-Medicated Control of T Cell Immunity and pathway analysis.

Dr. Jake Y. Chen is a Professor of Genetics and Computer Science, Chief Bioinformatics Officer of the newly established Informatics Institute, and Head of the Informatics Section of the Genetics Department at the University of Alabama at Birmingham. He holds a BS degree in Biochemistry and Molecular Biology from Peking University, and both MS and PhD degrees in Computer Science and Engineering from the University of Minnesota. He has

more than 20 years of bioinformatics R&D experience, including biological data mining, computational systems biology, and translational bioinformatics, with more than 150 peer-reviewed publications. His research focuses on building quantitative biomolecular systems models from genomic and clinical big data, thus helping understand, simulate, and predict complex disease biology outcomes. Prior to join UAB, he holds tenured faculty positions at Indiana University School of Informatics and Computing and at Purdue University Computer and Information Science Department. He is also an entrepreneur who created several startup companies to make emerging biomedical data easy to interpret and use by growing Medicine 2.0 stakeholders.

References

1. Huang da W, Sherman BT and Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research, 37, 1–13. [PubMed: 19033363]
2. Khatri P, Sirota M and Butte AJ (2012) Ten years of pathway analysis: current approaches and outstanding challenges. PLoS computational biology, 8, e1002375. [PubMed: 22383865]
3. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulski KS, Halloran P and Yasui Y (2009) Gene-set analysis and reduction. Briefings in bioinformatics, 10, 24–34. [PubMed: 18836208]
4. Nam D and Kim SY (2008) Gene-set approach for expression pattern analysis. Briefings in bioinformatics, 9, 189–197. [PubMed: 18202032]
5. Gene Ontology C, Blake JA, Dolan M, Drabkin H, Hill DP, Li N, Sitnikov D, Bridges S, Burgess S, Buza T et al. (2013) Gene Ontology annotations and resources. Nucleic acids research, 41, D530–535. [PubMed: 23161678]
6. Huang da W, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC et al. (2007) DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic acids research, 35, W169–175. [PubMed: 17576678]
7. Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA and Tainsky MA (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. Nucleic acids research, 31, 3775–3781. [PubMed: 12824416]
8. Berriz GF, King OD, Bryant B, Sander C and Roth FP (2003) Characterizing gene sets with FuncAssociate. Bioinformatics, 19, 2502–2504. [PubMed: 14668247]
9. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102, 15545–15550. [PubMed: 16199517]
10. Glaab E, Baudot A, Krasnogor N, Schneider R and Valencia A (2012) EnrichNet: network-based gene set enrichment analysis. Bioinformatics, 28, i451–i457. [PubMed: 22962466]
11. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A et al. (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research, 44, W90–97. [PubMed: 27141961]
12. Khatri P, Draghici S, Ostermeier GC and Krawetz SA (2002) Profiling gene expression using onto-express. Genomics, 79, 266–270. [PubMed: 11829497]
13. Ben-Shaul Y, Bergman H and Soreq H (2005) Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. Bioinformatics, 21, 1129–1137. [PubMed: 15550480]
14. Virtaneva K, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri MA, Bloomfield CD, de La Chapelle A and Krahe R (2001) Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. Proceedings of the

National Academy of Sciences of the United States of America, 98, 1124–1129. [PubMed: 11158605]

15. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS and Park PJ (2005) Discovering statistically significant pathways in expression profiling studies. Proceedings of the National Academy of Sciences of the United States of America, 102, 13544–13549. [PubMed: 16174746]
16. Yue Z, Kshirsagar MM, Nguyen T, Suphavilai C, Neylon MT, Zhu L, Ratliff T and Chen JY (2015) PAGER: constructing PAGs and new PAG-PAG relationships for network biology. Bioinformatics, 31, i250–257. [PubMed: 26072489]
17. Chen JY, Yue Z, Neylon MT, Nguyen T, Bulsara N, Arora I and Ratliff T (2016) Towards Constructing “Super Gene Sets” Regulatory Networks. 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), **DOI:** 10.1109/BIBM.2016.7822534.**DOI:**
18. Culhane AC, Schroder MS, Sultana R, Picard SC, Martinelli EN, Kelly C, Haibe-Kains B, Kapushesky M, St Pierre AA, Flahive W et al. (2012) GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. Nucleic acids research, 40, D1060–1066. [PubMed: 22110038]
19. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P and Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. Bioinformatics, 27, 1739–1740. [PubMed: 21546393]
20. Kanehisa M, Furumichi M, Tanabe M, Sato Y and Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic acids research, 45, D353–D361. [PubMed: 27899662]
21. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B et al. (2017) The Reactome Pathway Knowledgebase. Nucleic acids research.
22. Martini P, Sales G, Massa MS, Chiogna M and Romualdi C (2013) Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic acids research, 41, e19. [PubMed: 23002139]
23. Pepe D and Grassi M (2014) Investigating perturbed pathway modules from gene expression data via structural equation models. BMC bioinformatics, 15, 132. [PubMed: 24885496]
24. Institute, N.H.G.R. (2012). National Human Genome Research Institute.
25. Hieter P and Boguski M (1997) Functional genomics: it’s all how you read it. Science, 278, 601–602. [PubMed: 9381168]
26. Yue Z, Zheng Q, Neylon MT, Yoo M, Shin J, Zhao Z, Tan AC and Chen JY (2017) PAGER 2.0: an update to the pathway, annotated-list and gene-signature electronic repository for Human Network Biology. Nucleic acids research.
27. Pfister TD, Reinhold WC, Agama K, Gupta S, Khin SA, Kinders RJ, Parchment RE, Tomaszewski JE, Doroshow JH and Pommier Y (2009) Topoisomerase I levels in the NCI-60 cancer cell line panel determined by validated ELISA and microarray analysis and correlation with indenoisoquinoline sensitivity. Molecular cancer therapeutics, 8, 1878–1884. [PubMed: 19584232]
28. Kohn KW, Zeeberg BM, Reinhold WC and Pommier Y (2014) Gene expression correlations in human cancer cell lines define molecular interaction networks for epithelial phenotype. PloS one, 9, e99269. [PubMed: 24940735]
29. Weirauch MT (2011) Gene coexpression networks for the analysis of DNA microarray data. Applied statistics for network biology: methods in systems biology, 215–250.
30. Rice J (2006) Mathematical statistics and data analysis. Nelson Education.
31. Cimen Bozkus C, Elzey BD, Crist SA, Ellies LG and Ratliff TL (2015) Expression of Cationic Amino Acid Transporter 2 Is Required for Myeloid-Derived Suppressor Cell-Mediated Control of T Cell Immunity. J Immunol, 195, 5237–5250. [PubMed: 26491198]
32. Do RK, Hatada E, Lee H, Tourigny MR, Hilbert D and Chen-Kiang S (2000) Attenuation of apoptosis underlies B lymphocyte stimulator enhancement of humoral immune response. J Exp Med, 192, 953–964. [PubMed: 11015437]
33. Yoneyama M, Kikuchi M, Matsumoto K, Imaizumi T, Miyagishi M, Taira K, Foy E, Loo YM, Gale M, Jr., Akira S et al. (2005) Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity. J Immunol, 175, 2851–2858. [PubMed: 16116171]

34. Sakaguchi S, Miyara M, Costantino CM and Hafler DA (2010) FOXP3+ regulatory T cells in the human immune system. Nat Rev Immunol, 10, 490–500. [PubMed: 20559327]

35. Wherry EJ, Teichgraber V, Becker TC, Masopust D, Kaech SM, Antia R, von Andrian UH and Ahmed R (2003) Lineage relationship and protective immunity of memory CD8 T cell subsets. Nat Immunol, 4, 225–234. [PubMed: 12563257]

36. Müller-Schmah C, Solari L, Weis R, Pfeifer D, Scheibenbogen C, Trepel M, May AM, Engelhardt R and Lübbert M (2012) Immune response as a possible mechanism of long-lasting disease control in spontaneous remission of MLL/AF9-positive acute myeloid leukemia. Annals of Hematology, 91, 27–32. [PubMed: 21959947]

37. Freytag S, Gagnon-Bartsch J, Speed TP and Bahlo M (2015) Systematic noise degrades gene co-expression signals but can be corrected. BMC bioinformatics, 16, 309. [PubMed: 26403471]

38. Stuart JM, Segal E, Koller D and Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science, 302, 249–255. [PubMed: 12934013]

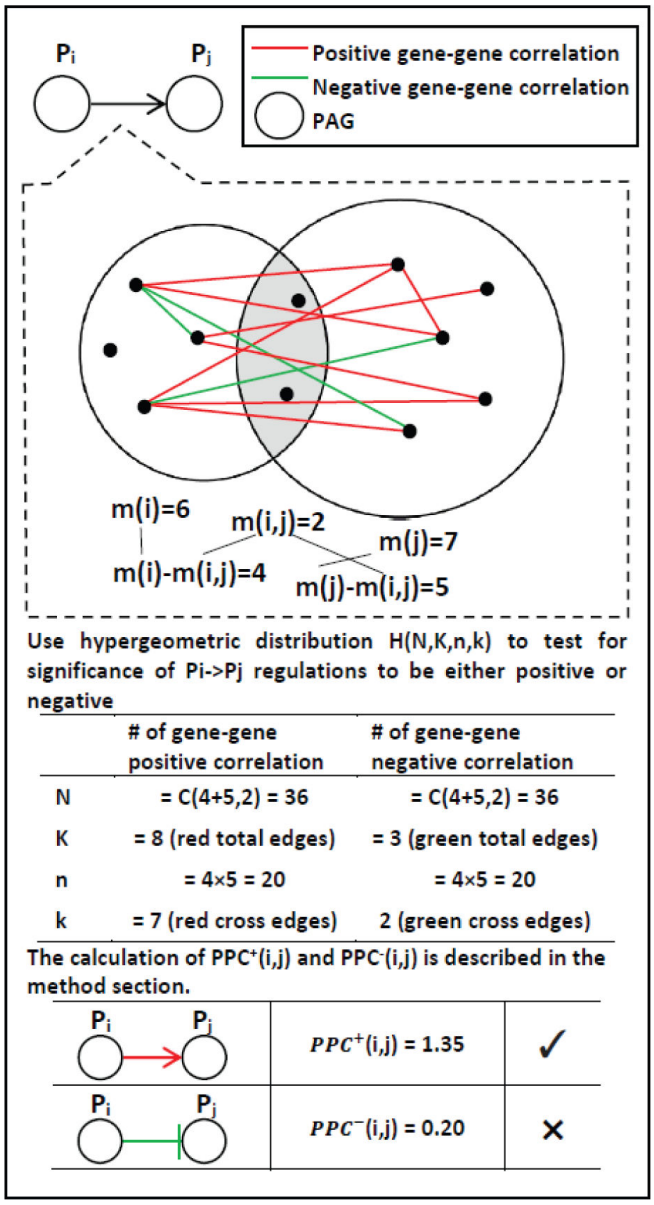

Figure 1.

An illustration showing how to determine Pi to Pj regulation details as either “positive” (stimulation) or “negative” (inhibition) based on PAG-to-PAG regulation relationships and gene-gene correlation data.

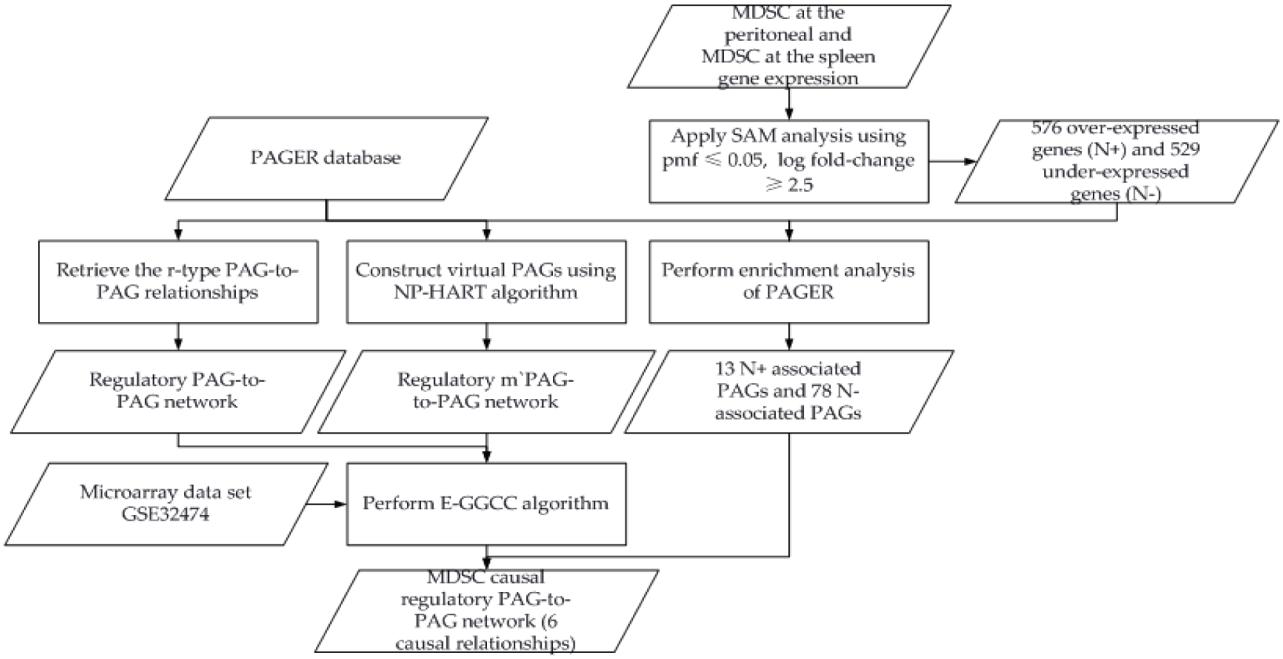

Figure 2.

An overview of the pipeline for constructing the MDSC causal regulatory PAG-to-PAG network.

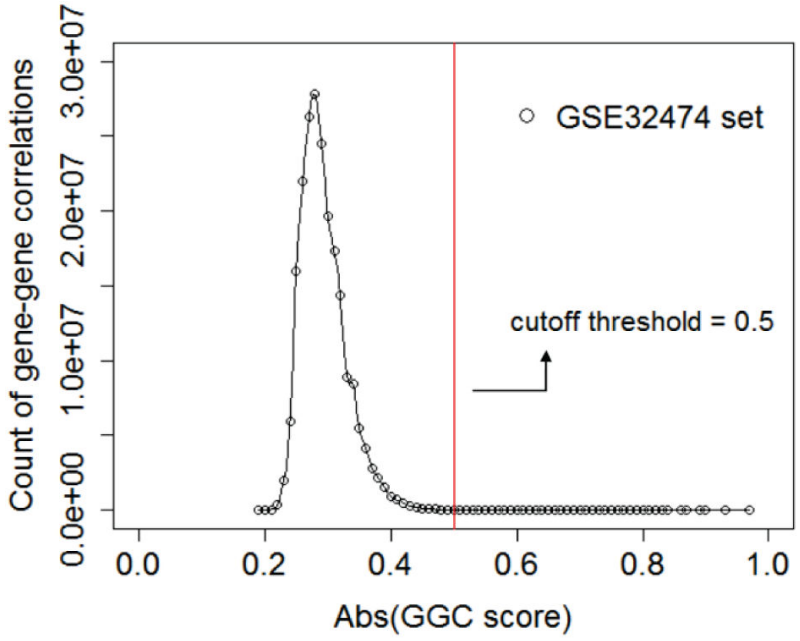

Figure 3. Distribution of GGC scores for gene expression correlation among all genes in PAGER.
The circle points stand for expression correlation calculated from the actual GSE32474 expression dataset. The plus points stand for expression correlation calculated from the random expression dataset.

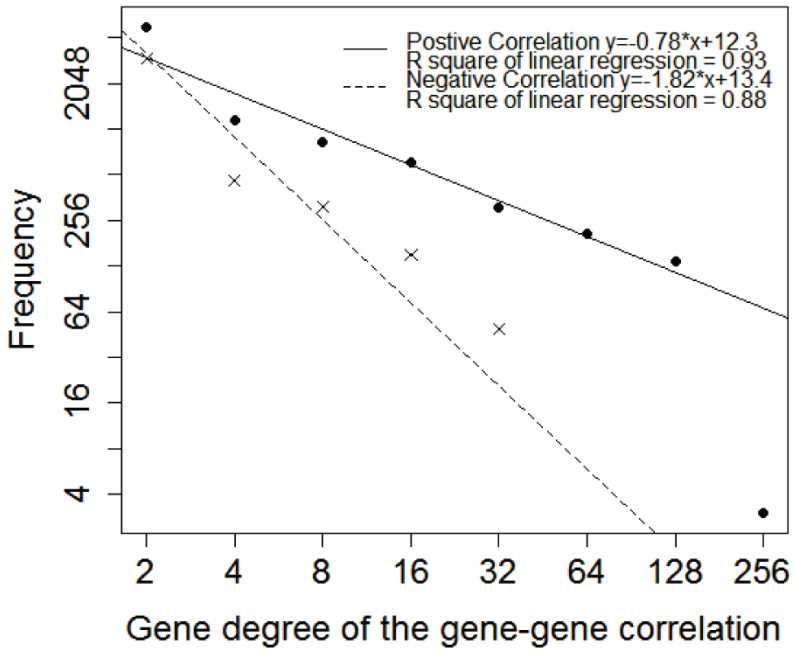

Figure 4. Gene degree of significant positive gene-gene correlation pairs and of negative gene-gene correlation pairs.

The linear regression R^2 of significant positive gene-gene correlation pairs is 0.93, and R^2 of significant negative gene-gene correlation pairs is 0.88.

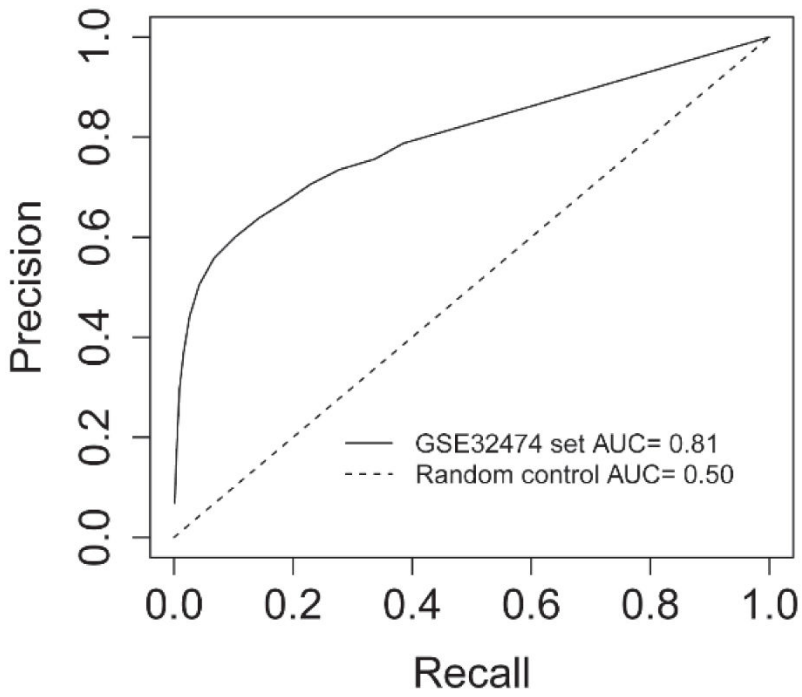

Figure 5. The receiver-operator characteristic (ROC) curve to demonstrate the prediction performance of identified causal PAG-to-PAG regulatory relationship details (stimulatory/ inhibitory or not).

The two curves show a comparison using the same significance tests with the use of either the GSE32474 real data set or a randomly reshuffled data set. The Area-under-curve (AUC) performances are also indicated.

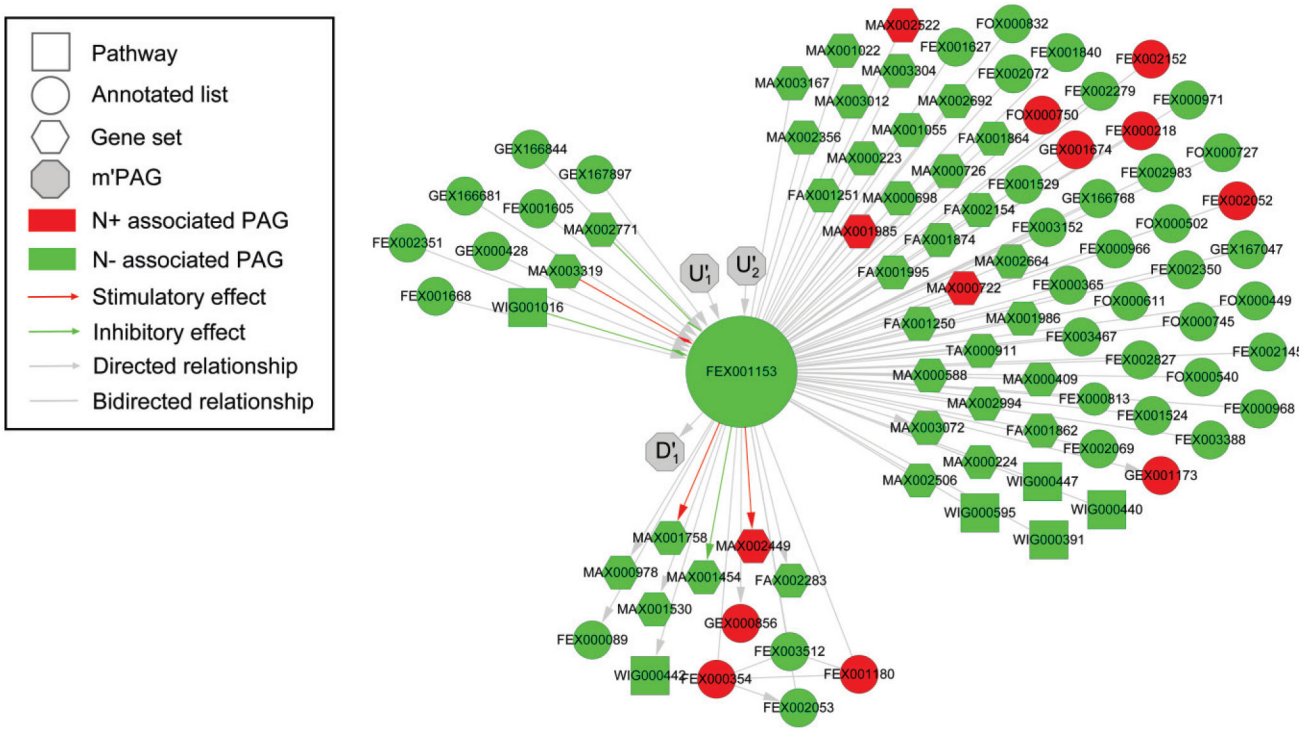

Figure 6. MDSC-specific network with causal mPAG-to-mPAG regulatory effect details.

The m'PAGs are colored in grey. Colored edges indicate either stimulatory effects (in red) or inhibitory effects (in green) among mPAG-to-mPAG regulatory relationships.

Table 1.

The confusion matrix for generating the ROC curve.

		PPC score as test	
		≥ PPC threshold	< PPC threshold
Pmf as T/F criteria	pmf ≤0.01 over representation	TP (12.212)	FN (8.267)
	pmf ≤0.01 under representation	FP (1.995)	TN (18.484)

Table 2.

An overview of PAG and PAG-to-PAG identified data statistics.

	Counts
Genes in PAGs	44,313
Gene-Gene Relationships	115,840
PPI	93,713
Gene Regulation	22,127
PAGs	38,446
PAGER original PAGs	38,379
NP-HART virtual PAGs	87
r-type PAG-PAG pairs	24,816
Stimulatory effect (PPC>1)	7,618
Inhibitory effect (PPC>1)	4,594

Table 3.

An example showing PAG-to-PAG relationship in the database.

PAG A	PAG A size	PAG A type	PAG A name	PAG B	PAG B size	PAG B type	PAG_B name	Causal type	PPC score
WIG000672	31	P	Genes involved in Activation of the pre-replicative complex	TAX001140	602	A	microtubule/chromatin interaction, structure-specific DNA binding, structure-specific DNA binding, plasmid binding	pos	144
WIG001985	41	P	DNA Replication	TAX001140	602	A	microtubule/chromatin interaction, structure specific DNA binding, structure-specific DNA binding, plasmid binding.	pos	122
WIG000802	30	P	Genes in-volved in DNA strand elongation	MAX001341	230	A	Genes involved in DNA repair, compiled manually by the authors.	pos	96
WIG000504	36	P	DNA replication	MAX001341	230	A	Genes involved in DNA repair, compiled manually by the authors.	pos	46

Table 4.

The statistics of PAG-PAG causal relationship in MDSC case study.

Categories	Count
PAGs	239
Up-regulated (N+)	26
Down-regulated (N−)	156
Other	57
PAG-PAG Regulations	191
Stimulatory effect	3
Inhibitory effect	3
Cannot characterize	185

Table 5.

PAGs' name in the MDSC network.

PAG ID	PAG name
FEX001153	Comprehensive List of Immune-Related Genes
MAX003319	B lymphocyte late differentiation genes (LDG): top genes down-regulated in plasma cells from tonsils (TPC) compared to those from bone marrow (BPC)
MAX002771	Genes up-regulated specifically in human thymus
WIG001016	Genes involved in Negative regulators of RIGI/MDA5 signaling
MAX001758	Genes down-regulated in the influenza-specific CD8+ T lymphocytes from bronchoalveolar lavage (BAL) compared to those from spleen
MAX002449	Genes down-regulated in T lymph
MAX001454	Myeloid leukemia model in mice with germ-line MLL-AF9 fusion knock-in [GeneID=4297;4300]: genes changed in comparison among the leukemic, preleukemic and wild-type animals

Table 6.

Causal PAGs identified in the MDSC-specific network relative to the immunity PAG

Direction	Effect	PAG ID	PPC
Upstream	stimulatory	MAX003319	3.4
	inhibitory	MAX002771	21.2
		WIG001016	2.5
Downstream	stimulatory	MAX001758	26.5
		MAX002449	15.8
	inhibitory	MAX001454	2.6