PLOS ONE

RESEARCH ARTICLE

Inferring gene and protein interactions using PubMed citations and consensus Bayesian networks

Anthony Deeter[1,2]*, **Mark Dalman**[3,4], **Joseph Haddad**[2], **Zhong-Hui Duan**[1,2]

1 Integrated Bioscience, University of Akron, Akron, Ohio, United States of America, **2** Department of Computer Science, University of Akron, Akron, Ohio, United States of America, **3** College of Public Health, Department of Biostatistics, Environmental Health Sciences and Epidemiology, Kent State University, Kent, Ohio, United States of America, **4** College of Podiatric Medicine, Department of Preclinical Sciences, Kent State University, Kent, Ohio, United States of America

* aed27@zips.uakron.edu

Citation: Deeter A, Dalman M, Haddad J, Duan Z-H (2017) Inferring gene and protein interactions using PubMed citations and consensus Bayesian networks. PLoS ONE 12(10): e0186004. https://doi.org/10.1371/journal.pone.0186004

Editor: Frank Alexander Feltus, Clemson University, UNITED STATES

Received: April 19, 2017

Accepted: September 22, 2017

Published: October 19, 2017

Data Availability Statement: The source code of the software presented is freely available in the web repository GitHUB https://github.com/Timer/bayesian-learning.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The PubMed database offers an extensive set of publication data that can be useful, yet inherently complex to use without automated computational techniques. Data repositories such as the Genomic Data Commons (GDC) and the Gene Expression Omnibus (GEO) offer experimental data storage and retrieval as well as curated gene expression profiles. Genetic interaction databases, including Reactome and Ingenuity Pathway Analysis, offer pathway and experiment data analysis using data curated from these publications and data repositories. We have created a method to generate and analyze consensus networks, inferring potential gene interactions, using large numbers of Bayesian networks generated by data mining publications in the PubMed database. Through the concept of network resolution, these consensus networks can be tailored to represent possible genetic interactions. We designed a set of experiments to confirm that our method is stable across variation in both sample and topological input sizes. Using gene product interactions from the KEGG pathway database and data mining PubMed publication abstracts, we verify that regardless of the network resolution or the inferred consensus network, our method is capable of inferring meaningful gene interactions through consensus Bayesian network generation with multiple, randomized topological orderings. Our method can not only confirm the existence of currently accepted interactions, but has the potential to hypothesize new ones as well. We show our method confirms the existence of known gene interactions such as *JAK-STAT-PI3K-AKT-mTOR*, infers novel gene interactions such as *RAS- Bcl-2* and *RAS-AKT*, and found significant pathway-pathway interactions between the JAK-STAT signaling and Cardiac Muscle Contraction KEGG pathways.

Introduction

With over tens of millions citations, PubMed, which offers free access to the National Library of Medicine's MEDLINE database, contains a great wealth of biomedical literature records [1].

An investigator can use PubMed's search capabilities to locate and read publications using simple data mining techniques such as keyword, author, year, and publication-name searches. The issue arises that with such a vast storage of information, to manually search and aggregate this information in order to discover potential relationships among the data requires an immense amount of time. In order to maximize the retrieval and use of the information available within PubMed, many post-processing computational tools have been developed. Plikus [2] provides a comprehensive group of third-party search interfaces [3–5] as well as systems that associate PubMed literature with ontology databases [6–8].

As PubMed provides a platform to access extensive amounts of information from biomedical literature, other data repositories such as GEO [9] and GDC [10] allow for the storage and retrieval of extensive amounts of genomic data from micro-array analysis [11, 12] and next-generation sequencing [13–15]. In part, the desire to determine genetic interactions using this data has led to the adaptation of commonly used bioinformatics methods including Neural Networks [16], Support Vector Machines [17], and Bayesian Networks [18, 19]. Additionally, past studies have inferred genetic interactions through the discovery of the co-occurrence of gene names within the abstracts [20, 21] and the linguistic analysis of relevant records acquired from PubMed [22]. We have created a system with which to combine the statistical power of Bayesian networks with the genetic information contained within citations accessible from the PubMed database in order to infer interactions among genes. Current genetic interaction software systems include Reactome [23, 24], Ingenuity Pathway Analysis (IPA) [25], the KEGG pathway database [26, 27], BioGrid [28], and IntAct [29]. Reactome is a curated, peer-reviewed database of genetic pathways composed of an organisms complete set of genetic reactions [30]. IPA has several features including pathway analysis, predictive causal analysis, and next-generation sequencing data analysis, all of which utilize their Ingenuity Knowledge Base; a repository of expert-curated biological interactions and functional annotations. While similar to these, our system does not rely on the input of expert biologists curating experiment data or publication information.

The structure of a Bayesian network can be representative of the interactions among genes within a biological pathway. Interactions among the genes are represented using directed edges among nodes in the network [31]. Constructing Bayesian networks can be a complex and computationally intensive process. The amount of data available within PubMed is large and the search space involved in Bayesian network creation can grow exponentially without imposing restrictions on the network.

Methods involved in network construction generally fall into two categories: constraint-based and score-based. Constraint-based construction attempts to reduce the search space by placing restrictions on the structure of the network. The restrictions imposed on the network depend on the conditional influence nodes have on one another. This reduction of the search space allows for faster computation of potential network structures, but the potential for compounding error increases with each reduction [32]. Score-based methods, selecting the network structure based on a maximum score, have a higher degree of precision. The issue with higher precision scoring methods are their higher computational complexity, limiting score-based construction to low-dimensional data.

In order to reduce the computational runtime of Bayesian network creation, Sriram suggests a combination of both constraint- and score-based construction utilizing the K2 algorithm [33], KEGG pathway information and an initial topological ordering of the genes that reduces the initial search space to a smaller, low-dimensional set of data for use with a scoring method. One possible caveat of using the K2 algorithm is the potential bias introduced from the initial topological input. The utilization of prior knowledge to construct these topological inputs as well as the generation of consensus networks from multiple Bayesian networks can

help mitigate this bias [32]. Consensus networks have been utilized in the past to combine data from several networks into one; a standard use case being phylogenetic trees [34, 35]. Consensus networks have been utilized with Bayesian networks as well [36], but differ from ours in that they retain the directed edges and structure of a Bayesian network, whereas our method utilizes every edge and removes directed aspect.

We applied the idea of consensus network generation utilizing the K2 algorithm with data mined from the PubMed database. When using K2, a topological input is given, restricting the possible parents each node in the network can have. This topological ordering is normally constructed using expert knowledge of the data being examined. This can be accomplished by curating publications and experiment data. Additionally, the use of natural language processing like that integrated into Chilibot [22] can be a potential solution to generating knowledge of the data. Instead, in order to remove the need for expert curators and natural language processing, we expand the number of topological orderings supplied from several to many, and increase their potential orders to include not a specialized group of inputs, but an all-encompassing group of possibilities. This allows us to create extensive consensus networks comprised of a large number of individual Bayesian networks in order to remove this potential bias. In addition, the expanded input allows us to introduce the concept of network resolution. What we are calling network resolution is the ability to adjust the sensitivity of the findings from a broad, inclusive set of connections to a narrow, focused group with a higher potential to be true interactions. Using our method, investigators can not only confirm the existence of currently accepted genetic interactions through use of high-resolution networks, but can also hypothesize novel interactions by lowering the resolution to reveal additional, potential interactions.

Materials and methods

Datasets

In order to create our Bayesian networks, we constructed a set of prior knowledge about the genes groups we were interested in. KEGG pathways present a set of molecules involved in a biological system and an overview of their interactions in a sequence of coordinated events. We created two separate datasets in order to examine both the stability of our method across multiple numbers of topologies and differing sample sizes as well as the functionality of the method and its output. In this study, we used the gene groups in the JAK-STAT signaling (Fig 1) and Cardiac Muscle Contraction (Fig 2) pathways as constraints to reduce the search space and explore the relationships among those presented in PubMed. The JAK-STAT signaling pathway contains 31 functional gene groups putatively involved with signaling for development and homeostasis in mammals. We focused on the key molecules involved in the signaling cascade. The molecules in the extracellular space such as hormones and cytokines, as well as receptors, were not included in any network construction. The Cardiac Muscle Contraction pathway contains 13 functional gene groups putatively involved with the contraction of the heart. Three negative control genes were also included: *SCF*, *NOS*, and *AC5*.

For both data sets, composing the prior knowledge about the selected gene groups involved data mining PubMed for publications that contained a member from at least two separate gene groups. The selected members for each gene group were treated as a single entity when searching, and a successful search for any of the members within a group was treated as a successful search for the entire group. Each publication that met this requirement became a row in the initial Bayesian network prior knowledge publication list. After list compilation, regular expressions were used to examine the abstracts within the list to create the prior knowledge matrix, searching again for the presence of members from any of the gene groups. This matrix

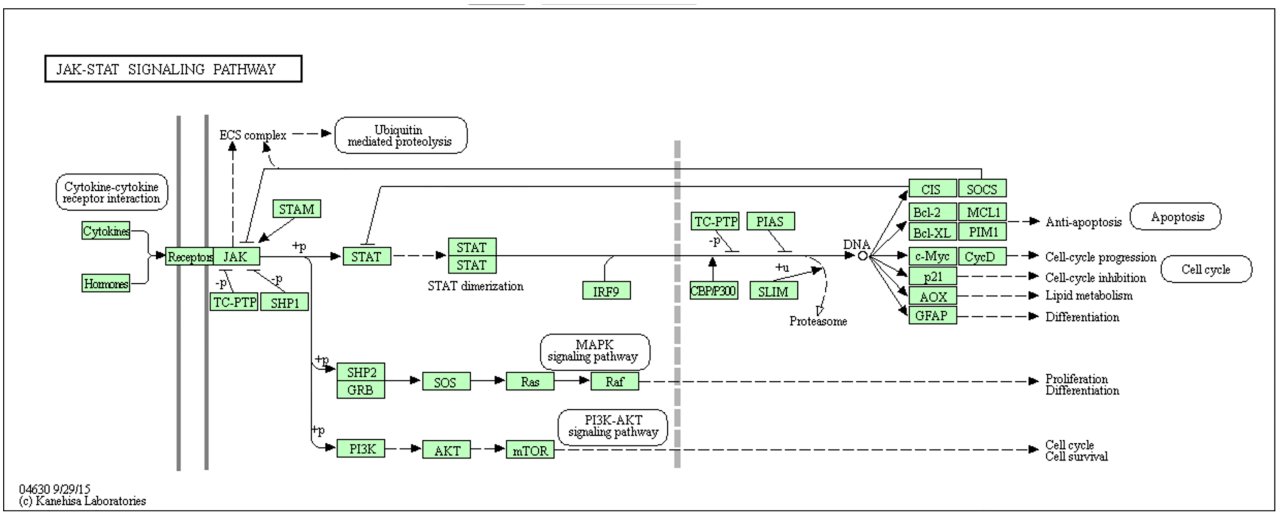

Fig 1. The JAK-STAT signaling KEGG pathway. The JAK-STAT signaling KEGG pathway shows the known interactions within the JAK-STAT signaling cascade.

https://doi.org/10.1371/journal.pone.0186004.g001

indicated the presence or not of each of the gene groups within each publication. Parentheses, spaces, dashes, and brackets were allowed at the beginning and end of the search terms, and periods, commas, colons, and semi-colons were allowed at the end only.

The columns of the resulting matrix consisted of the gene groups, and the rows were successfully searched publications. The value of the intersection of publication and gene group within the matrix was determined by whether the publication contained mention of the gene group within its title or abstract. If a member of the group was present, the value of the matrix at that row-column intersection was set to '1'. A value of '0' was given for intersections where the group was not present within the abstract. During Bayesian network generation this matrix was used to represent prior knowledge about the interaction among genes of interest; the presence of two or more genes in the same publication can indicate an interaction among them.

The first data set (DS1), used to examine the stability of our method across different topologies and sample sizes, was created from functional gene groups within the JAK-STAT signaling pathway only. Because each group can contain multiple genes, and each gene can have multiple aliases, a list of all genes and their associated aliases for each gene group was used to create a list of members associated with the gene group. To reduce the number of members to those which are most commonly found within literature, members were sorted by the number of times they were found within the title or abstract of any piece of biomedical literature within PubMed. The top five most commonly observed members from each functional gene group were selected for use (see Fig 3). PubMed initial searching yielded a list of 42,600 publications for which the title and abstract were downloaded and analyzed. Matrix creation yielded a prior knowledge set with 27 columns and 42,600 rows (see S1 File). The node numbers from the member list match each gene group to its respective column number within the prior knowledge matrix.

The second data set (DS2), used to examine the functionality and output of the method, was created using functional gene groups from the JAK-STAT signaling and Cardiac Muscle Contraction KEGG pathways as well as three randomly selected negative control genes: *SCF*, *NOS*, and *AC5*. Within this data set, each gene group's members were selected using the gene group name displayed on the KEGG pathway as well as the first (up to) five genes indicated by the KEGG database. No additional aliases for genes were included (see Fig 4). PubMed initial

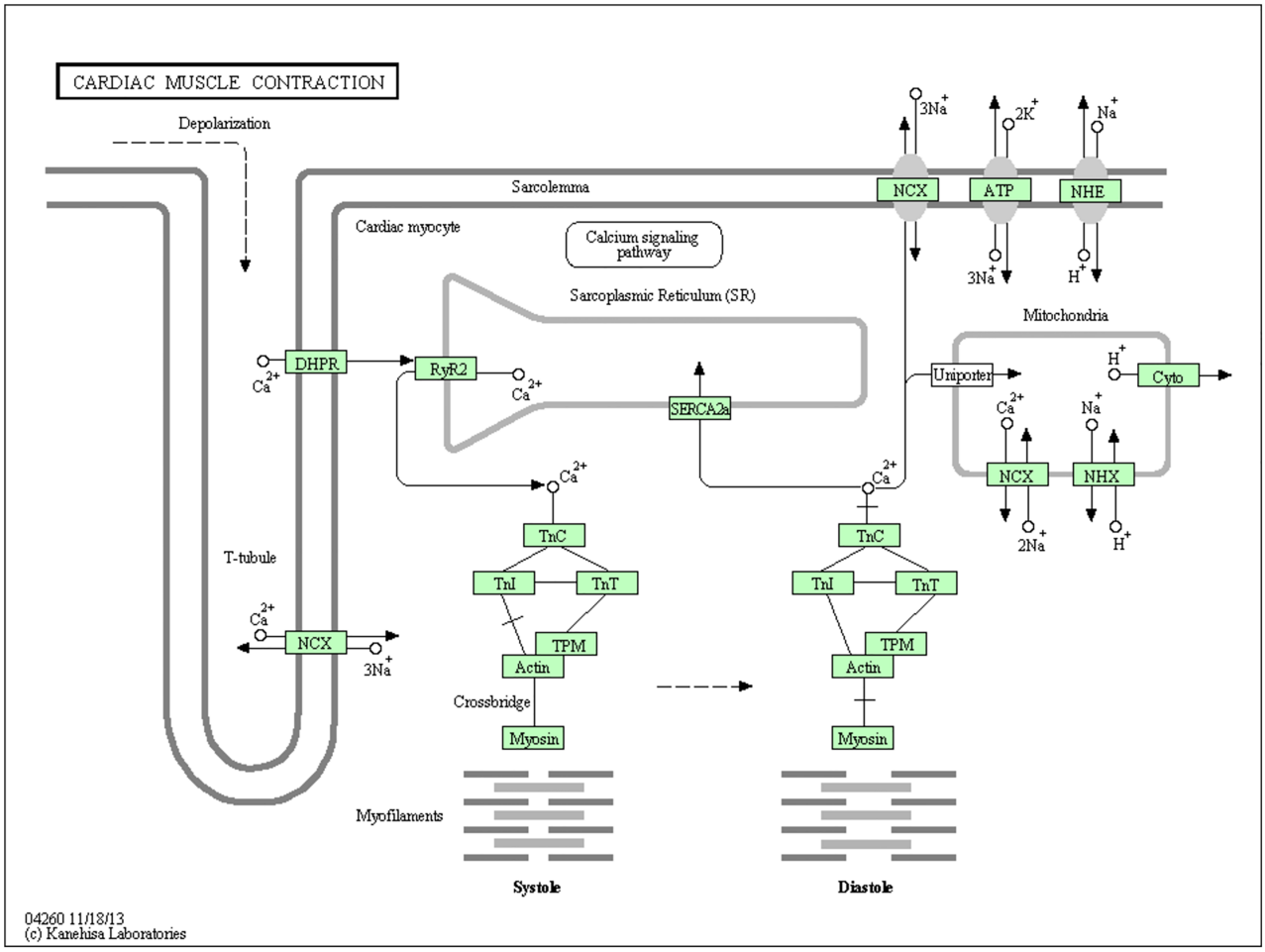

Fig 2. The Cardiac Muscle Contraction KEGG pathway. The Cardiac Muscle Contraction KEGG pathway shows how a ca2+ influx induces cardiac muscle contraction.

https://doi.org/10.1371/journal.pone.0186004.g002

searching yielded a list of 96,594 publications for which the title and abstract were downloaded and analyzed. Matrix creation yielded a prior knowledge set with 45 columns and 96,594 rows (see S2 File). As before, the node numbers from the member list match each gene group to its respective column number within the prior knowledge matrix.

Generation of Bayesian networks

Bayesian networks are comprised of a directed acyclic graph (DAG) in which the nodes represent random variables from the domain and an edge between two nodes represents a dependency between those variables. As the number of nodes within the network increases, the search space for Bayesian network learning grows exponentially. The K2 algorithm was used to reduce this search space. Given a specific starting node, the topological input used by K2 was created using a random permutation of the remaining nodes in the network. To score potential edges within the network, the marginal likelihood of the graph containing an edge was computed given the publication data within the input matrix. A maximum of five parents per node was used to reduce the density of edges within the graph. The resulting graphs were

NODE	MEMBERS					
1	C-Myc	bhlhe39	mrtl	mycc	c-myc	myc
2	AKT3	akt2	akt1	pkb	rac	akt
3	PIK3	mcm	pi3k			
4	SOS	gf1	sos2	ns4	sos1	hgf
5	GRB2/SHP2	shp2	grb2	ns1	cfc	ash
6	CIS	g18	cish	cis-1	socs	cis
7	Bcl-2	ppp1r50	bcl2	bcl-2		
8	SOCS	socs-1	socs-3	socs1	cis-4	socs3
9	CBP/P300	rsts	ep300	crebbp	cbp	p300
10	IRF9	isgf3g	irf-9	irf9	isgf3	p48
11	STAT	stat6	stat5	stat1	stat3	
12	PIAS	miz	pias3	pias1	gbp	dip
13	SHP1	sh-ptp1	ptpn6	shp1	shp-1	hcp
14	STAM	stam-1	stam1	stam2	stam	hbp
15	JAK	jak-3	tyk2	jak3	jak1	jak2
16	SLIM	slim1	fhl-1	slimmer	fhl1	slim
17	TC-PTP	ptn2	ptpt	tc-ptp	ptpn2	tcptp
18	MCL1	mcl1s	mcl1	mcl-1	eat	tm
19	P1M1	pim1	pim			
20	CycD	prad1	ccnd3	ccnd2	bcl1	ccnd1
21	p21	p21cip1	cdkn1a	waf1	cip1	p21
22	Bcl-XL	bclxl	bcl2l1	bcl-xs	bcl-x	bcl-xl
23	Ras	c-ha-ras1	hras1	c-h-ras	hras	p21ras
24	Raf	craf	raf1	ns5	c-raf	raf-1
25	Aox	aoh1	aox1	ao		
26	mTOR	frap2	raft1	frap1	frap	mtor
27	GFAP	alxdrd	gfap			

Fig 3. Consensus network creation. Using the functional gene groups from the JAK-STAT signaling pathway, PubMed searches were used to calculate the number of occurrences for each gene and its aliases within publication abstracts. The top five most commonly observed members from each functional gene group were selected for use as members of the data set. The node numbers match each gene group to a column number within its prior knowledge matrix (see S1 File).

https://doi.org/10.1371/journal.pone.0186004.g003

NODE	MEMBERS					
1	JAK	JAK1	JAK2	JAK3	TYK2	
2	STAM	STAM2				
3	TC-PTP	PTPN2				
4	SHP1	PTPN6				
5	STAT	STAT1	STAT2	STAT3	STAT4	STAT5A
6	IRF9	IRF9				
7	PIAS	PIAS3	PIAS4	PIAS1	PIAS2	
8	CBP/P300	CREBBP	EP300			
9	SLIM	FHL1				
10	CIS	CISH				
11	Bcl-2	BCL2				
12	Bcl-XL	BCL2L1				
13	c-Myc	MYC				
14	p21	CDKN1A				
15	AOX	AOX1				
16	GFAP	GFAP				
17	SOCS	SOCS4	SOCS7	SOCS1	SOCS2	SOCS3
18	MCL1	MCL1				
19	PIM1	PIM1				
20	CycD	CCND1	CCND2	CCND3		
21	SHP2	PTPN11				
22	GRB	GRB2				
23	SOS	SOS1	SOS2			
24	Ras	HRAS				
25	Raf	RAF1				
26	PI3K	PIK3CA	PIK3CB	PIK3CD	PI3KR1	PI3KR2
27	AKT	AKT3	AKT1	AKT2		
28	mTOR	MTOR				
29	DHPR	CACNG3	CACNG2	CACNA2D3	CACNG5	CACNG4
30	NCX	SLC8A1				
31	RyR2	RYR2				
32	SERCA2α	ATP2A2				
33	TnC	TNNC1				
34	TnI	TNNI3				
35	TnT	TNNT2				
36	TPM	TPM1	TPM2	TPM3	TPM4	
37	Actin	ACTC1				
38	Myosin	MYH6	MYH7	MYL2	MYL3	
39	ATPase	ATP1B4	ATP1A1	ATP1A2	ATP1A3	
40	NHE	SLC9A1				
41	NHX	SLC9A6				
42	CYTO	UQCR11	COX6B2	COX4I1	COX5B	
43	SCF	KITLG				
44	NOS	NOS1				
45	AC5	ADCY5				

Fig 4. Consensus network creation. Using the functional gene groups from the JAK-STAT signaling and Cardiac Muscular Contraction pathways as well as three randomly selected genes, *SCF*, *NOS*, and *AC5*, 45 genes were selected as members of the functionality data set. The node numbers match each gene group to its respective column number within the prior knowledge matrix (see S2 File).

https://doi.org/10.1371/journal.pone.0186004.g004

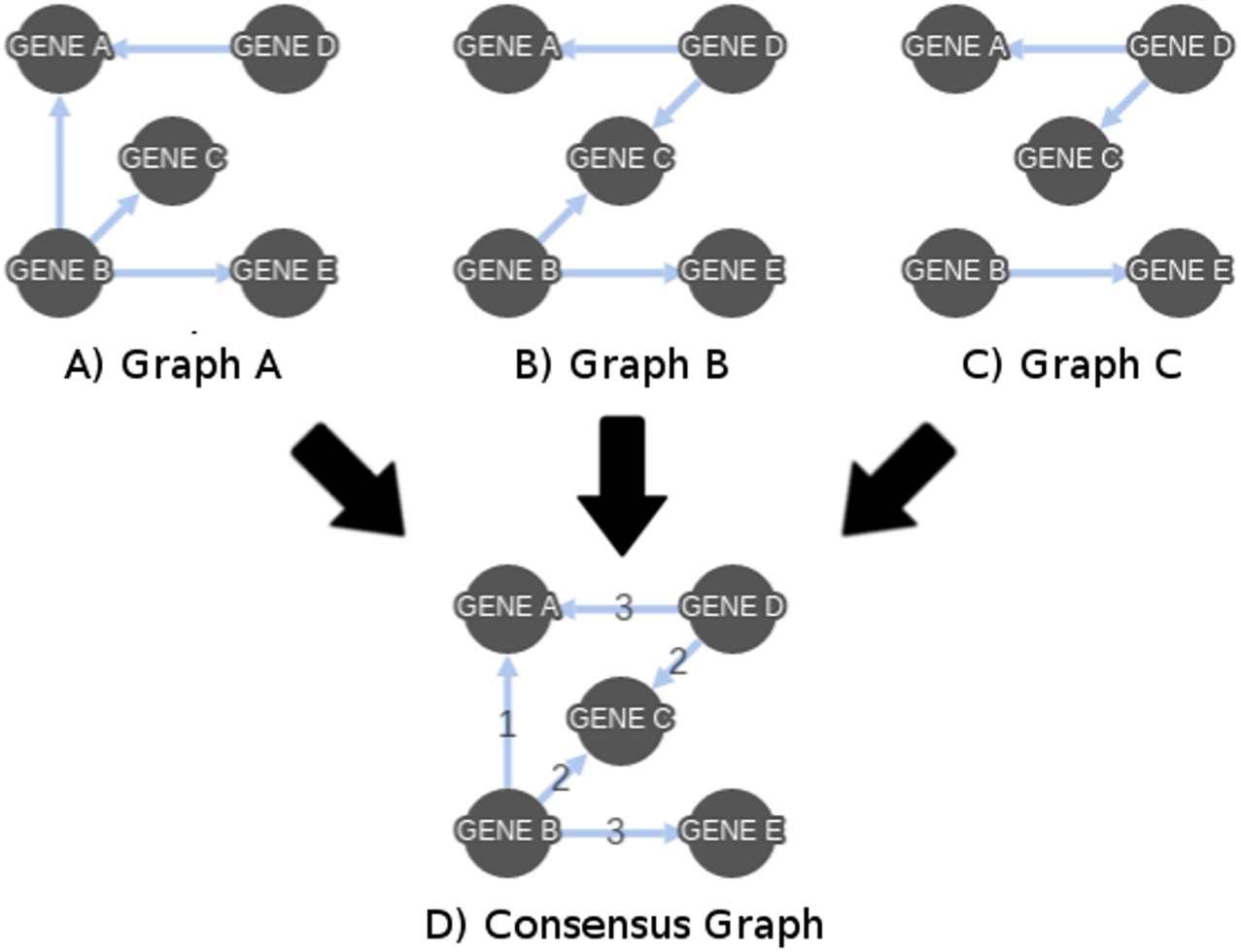

Fig 5. Consensus network creation. Graphs A, B, and C are used to create a consensus graph (D). The sum of the edge weights of each graph are used to determine the edge weight in the final graph.

https://doi.org/10.1371/journal.pone.0186004.g005

represented with adjacency matrices containing directed edges between nodes (see S3 and S4 Files).

Consensus network construction

Since each node given within a topological order is limited to the prior nodes when looking for potential parent nodes, the topological input for K2 has the potential to introduce bias into the generated Bayesian network. In order to reduce this bias, multiple Bayesian networks were generated using randomized topologies that begin from every node within the network. From DS1, for each of the 27 functional gene groups, a number of randomized topologies, ranging from 4 to 100, were created. This resulted in the number of Bayesian networks created to be between 108 and 2700. The Bayesian networks generated using these topologies and their resulting adjacency matrices were combined, summing the total number of graphs containing each edge and creating a single network where edge weights represent the collection of these summations (see Fig 5). By storing the resulting network within an adjacency matrix, the combination of directed graphs has the potential to have directed edges in both directions between two nodes. This bi-directionality was removed by adding the matrix to the transposition of

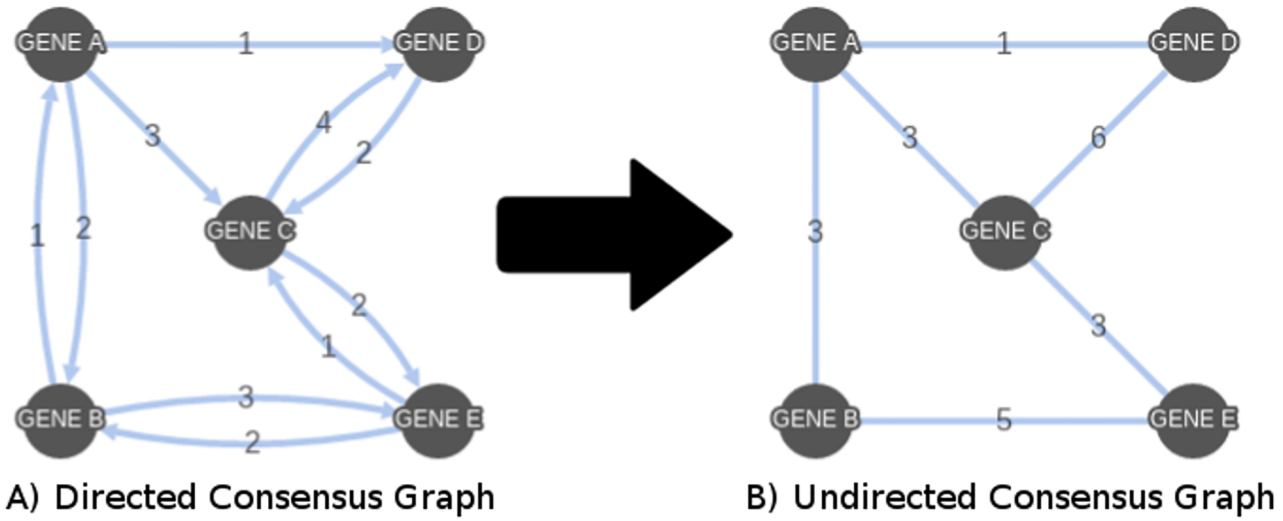

Fig 6. Removing bi-directionality from consensus networks. Bi-directionality is removed from directed consensus networks by adding the edge weights from both directions for each set of nodes in (A) and using an undirected edge with the new total in (B).

https://doi.org/10.1371/journal.pone.0186004.g006

itself and removing edges in the lower left triangle. The result, a consensus network, is a single, undirected graph with the weight of each edge being the number of individual Bayesian networks containing it (see Fig 6). The consensus network for DS2 was created in the same manner as above, using eight topologies only, utilizing 360 Bayesian networks.

Consensus network resolution

By combining multiple Bayesian networks into a single consensus network, an edge with a larger weight (meaning it was present in a significant number of individual Bayesian networks) can be interpreted as having a higher potential for representing a true interaction between nodes. Because of this, we introduce the concept of edge resolution. The resolution of an edge, ranging from 0 to 1, is calculated by dividing its weight by the largest weight of any single edge in the network. An edge resolution threshold can then be introduced, and removing edges below differing thresholds will yield different interpretations. For example, high-resolution networks will contain edges that were present in only a small number of Bayesian networks, while low-resolution networks will contain edges present in a large number of networks (Fig 7). All

	A	B	C	D	E
A	0	10	3	6	0
B	0	0	8	0	3
C	0	0	0	9	2
D	0	0	0	0	1
E	0	0	0	0	0

A) Adjacency Matrix

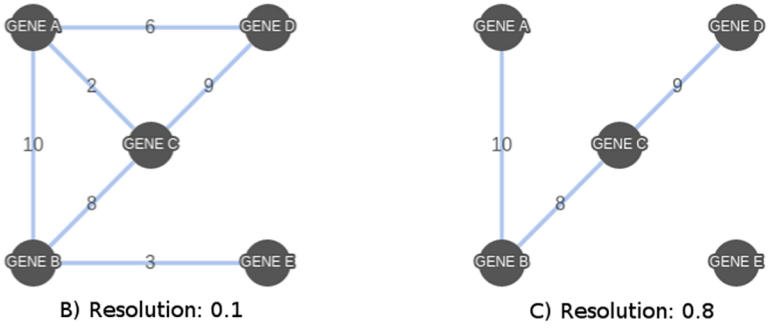

Fig 7. A consensus network at different resolutions. (A) The adjacency matrix for a consensus network created from 10 separate Bayesian networks. (B) The graph of the consensus network with a resolution of 0.1. (C) The graph of the consensus network with a resolution of 0.8. The resolution determines which edges to include by dividing the edge weight by the largest weight of any single edge in the network and removing those which do not exceed the cut-off.

https://doi.org/10.1371/journal.pone.0186004.g007

testing of consensus networks was done on a range of ten resolutions from 0.1 to 1.0 in order to characterize results across all resolutions.

Network stability across sample sizes

Depending on the search criteria used, sample sizes for publications retrieved from PubMed differ greatly in number. For consensus network generation via publication data to be considered stable across a broad range of sample sizes, we used Cohen's kappa coefficient to determine if networks created with relatively smaller sample sizes would infer the same relationships as those created with a larger dataset. Cohen's kappa coefficient is a statistic used to determine the chance-corrected proportion of agreement between judges classifying items into qualitative categories [37]. First, we randomly split DS1 into 10 sections of 4620. Second, for each test section, two consensus networks were generated: a test network created using the test section, and a predicted network was created using the remaining nine sections. Finally, for each pair of test and predicted networks, Cohen's kappa coefficient was calculated for each of the ten testing resolutions (0.1–1.0). The kappa calculation was used to determine if each pair of networks agreed on the set of edges that composed their consensus networks.

Stability across topological input sizes

The number of topological orders used to create a consensus network can be severely limited by the resources available with which to perform the calculations. The runtime of network creation grows linearly as the number of topologies increases. The question then arises, does a consensus network created with a smaller number of topologies differ, and by how much, than one created using more topologies? In order to test if consensus networks derived from differing numbers of topologies would yield the same consensus network, using DS1, a consensus network was built from differing numbers of random topologies, ranging from four to 100 by fours. For each successive pair of networks (four and eight, eight and 12, . . . 96 and 100) as well as the pair including four and 100, we calculated Cohen's kappa coefficient to measure agreement between each of the 10 test resolutions (0.1–1.0).

Functionality and inferred interactions

Using a prior knowledge data set that contains only the functional gene groups from a single KEGG pathway has the potential to make inferences within that pathway but lacks the ability to infer novel interactions with genes outside of the pathway. Having members from two different KEGG pathways as well as three negative controls, data set DS2 was used to verify the ability of our method to not only confirm existing intra-pathway interactions, but also infer pathway-pathway interactions, confirm a segregation of interactions between multiple pathways, and exhibit a lack of interactions among the three negative control genes. A consensus network was created using eight random topologies from each gene group, resulting in 360 Bayesian networks.

Implementation and source code

The methods described in this experiment were performed using the following R packages: RISmed [38], KEGGREST [39], and KEGGgraph [40]. All Bayesian and consensus network generation for DS1 was done on the Blue Waters petascale machine at the University of Illinois at Urbana-Champaign utilizing 1024 Cray CPU-only XE6 nodes, consisting of two 16-core AMD processors with 64 GBs of RAM. [41]. Network generation for DS2 was completed on a 32-core Intel machine with 64 GBs of RAM. The source code of the software presented is freely available in the web repository GitHUB: https://github.com/Timer/bayesian-learning.

Results

We analyzed resultant consensus networks generated by data mining publications from the PubMed database. The prior knowledge matrix for Consensus network creation for DS1 consisted of the following: 42,600 rows (representing the publications in which the abstract and/or title contained at least 2 of the members of the 27 functional gene groups in the JAK-STAT signaling KEGG pathway) and 27 columns (each representing the presence, or lack of, the functional gene group within the current row's publication). Using DS1 we analyze the stability of our method across differing sample sizes and the stability of our method with a varying number of topological orders. Using DS2 we analyze the functionality and the resulting inferred interactions. Consensus network creation for DS2 consisted of the following: a prior knowledge matrix consisting of 96,594 rows (representing the publications in which the abstract and/or title contained at least 2 of the members of the 45 functional gene groups within the JAK-STAT signaling and Cardiac Muscle Contraction KEGG pathways as well as three negative control genes) and 45 columns (each representing the presence, or lack of, the functional gene group within the current row's publication).

Network stability across sample sizes

Determining the stability of our method across different publication sample sizes was conducted by dividing the publication samples into 10 separate test sets and using them to create consensus networks, yielding 10 sets of edges representing inferred gene interactions. For each test case, the remaining publication samples were used to create a predicted consensus network, which would act as the network with which to test for agreement. For each of the 10 tests cases, Cohen's kappa coefficient was calculated for 10 network resolutions ranging from 0.1 to 1.0. Fig 8 shows the kappa coefficients for tests 1 through 10, across the range of network resolutions. The full set of results showing the kappa coefficient for each resolution within each of the 10 tests can be found in S5 File.

Stability across topological input sizes

With the high computational cost of generating Bayesian networks, and in turn consensus networks, it is expected that the number of topologies randomly generated for Bayesian network creation will vary depending on the resources available to the investigator. With this in mind, we tested the agreement between networks generated with different numbers of topological inputs. Topological input sizes ranging from four to 100 per node, increasing by four with each iteration, were created in order to compare their resulting consensus networks. Each network was compared with the network created by the next largest topological input size, and the networks created with the smallest and largest number of inputs were also compared. Cohen's kappa coefficients were calculated for 10 network resolutions, from 0.1 to 1.0, and were used to determine the agreement between each set of networks, for each level of resolution. Fig 9 contains the comparison sizes as well as the smallest, largest, and mean kappa coefficients for each set of comparisons. The full set of results, showing the kappa coefficient for each resolution within each input size comparison can be found in S6 File.

Discussion

With regard to network stability, if our method is stable across a wide range of sample sizes, the kappa coefficient should show agreement between each resolution for each test case. When testing the agreement between varying sample sizes, the kappa coefficients are uniform across all test and resolutions, with the lowest being 0.5802 at a network resolution of 0.5 (see S5 File).

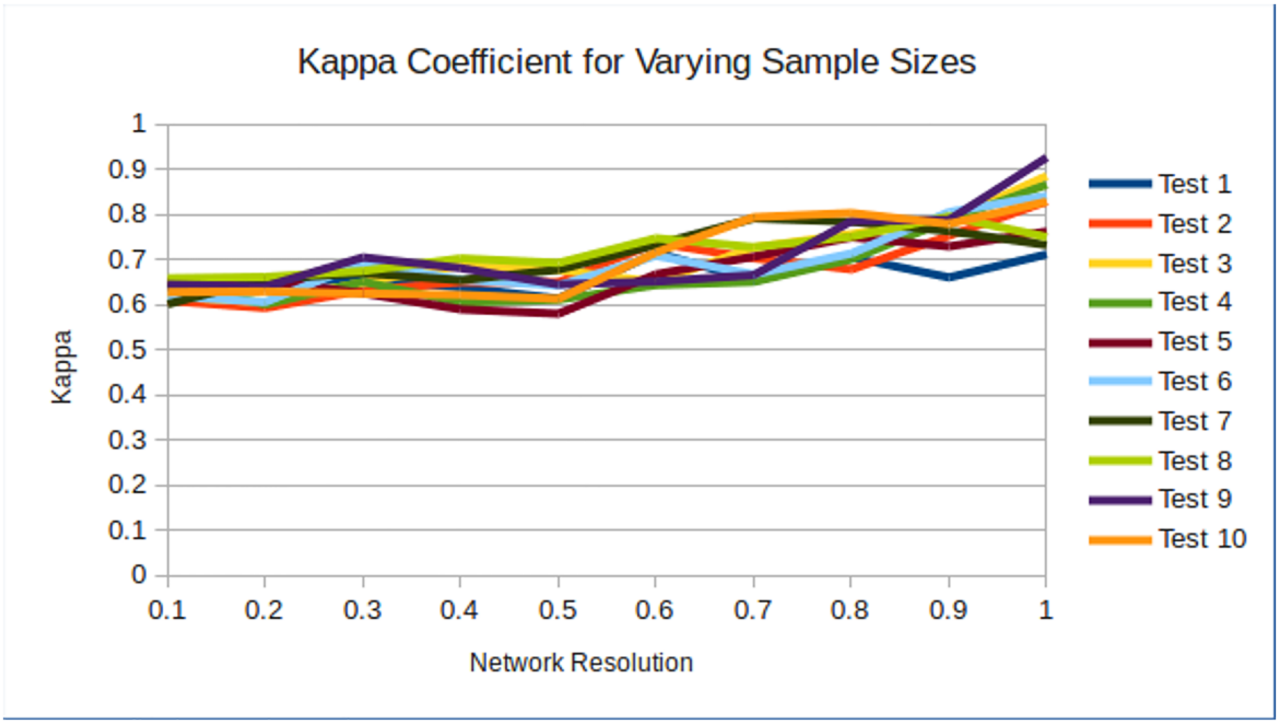

Fig 8. Kappa coefficients across network resolutions for differing sample sizes. When comparing consensus networks of differing sample sizes, for each of the 10 tests cases, Cohen's kappa coefficient was calculated for 10 network resolutions ranging from 0.1 to 1.0.

https://doi.org/10.1371/journal.pone.0186004.g008

The nature of a lower network resolution allows for less consensus being needed to include an edge as an inferred interaction, therefore, the upward trend of the kappa coefficients mirrors the expectation of growth as network resolution increases. Using Cohen's suggested interpretation of the kappa coefficient, all test cases show at least moderate agreement (0.41 to 0.60), all tests for resolutions above 0.5 show substantial agreement (0.61 to 0.80) or higher, and all tests at resolution of 1.0 are considered to have almost perfect agreement (0.81 to 0.99). These results indicate that at lower resolutions a smaller sample size will have at least a moderate agreement on which interactions are inferred, and as the resolution increases, so will the agreement between small and large sample sizes.

With the high computational cost of generating Bayesian networks, and in turn consensus networks, it is expected that the number of topologies randomly generated for Bayesian network creation will vary depending on the resources available to the investigator. With this in mind, we tested the agreement between networks generated with different numbers of topological inputs. In all cases, including the one between the smallest and largest number, the kappa coefficient is in the 'almost perfect' agreement range (0.81 to 0.99) and the max kappa among resolutions in five of the tests was considered perfect (1.0) (see S6 File). This indicates that the inferred interactions will be extremely similar regardless of the number of topological inputs. A slight upward trend can be observed in the mean kappa coefficient as the number of topological inputs increases, indicating there is a slight increase in stability as more Bayesian networks are utilized.

A driving focus of empirical research is to derive a causal relationship between a stimulus and a downstream molecule or receptor that can then drive a behavioral response, physiological change, and or produce an epigenetic alteration [42]. Paramount to driving the next generation in bioinformatics is the use of computational models to confirm current molecular

COMPARISON SIZES	MIN KAPPA	MAX KAPPA	MEAN KAPPA
4 – 100	0.8969	0.9784	0.9437
4 – 8	0.8625	0.9730	0.9207
8 – 12	0.9077	0.9837	0.9508
12 – 16	0.8930	1.0000	0.9545
16 – 20	0.9033	0.9893	0.9563
20 – 24	0.9492	1.0000	0.9740
24 – 28	0.9595	1.0000	0.9752
28 – 32	0.9548	1.0000	0.9758
32 – 36	0.9591	1.0000	0.9779
36 – 40	0.9641	1.0000	0.9857
40 – 44	0.9618	1.0000	0.9806
44 – 48	0.9618	1.0000	0.9806
48 – 52	0.9548	0.9953	0.9790
52 – 56	0.9548	1.0000	0.9785
56 – 60	0.9618	1.0000	0.9815
60 – 64	0.9652	1.0000	0.9870
64 – 68	0.9618	1.0000	0.9804
68 – 72	0.9556	1.0000	0.9846
72 – 76	0.9605	1.0000	0.9750
76 – 80	0.9618	1.0000	0.9836
80 – 84	0.9618	1.0000	0.9853
84 – 88	0.9618	1.0000	0.9856
88 – 92	0.9501	1.0000	0.9776
92 – 96	0.9262	1.0000	0.9731
96 – 100	0.9646	1.0000	0.9853

Fig 9. Kappa coefficients across network resolutions for differing sample sizes. The minimum, maximum, and mean kappa coefficients for each pair of networks created with varying sample sizes. It can be seen that variations in the sample size inputs for Bayesian network construction results in similar consensus networks.

https://doi.org/10.1371/journal.pone.0186004.g009

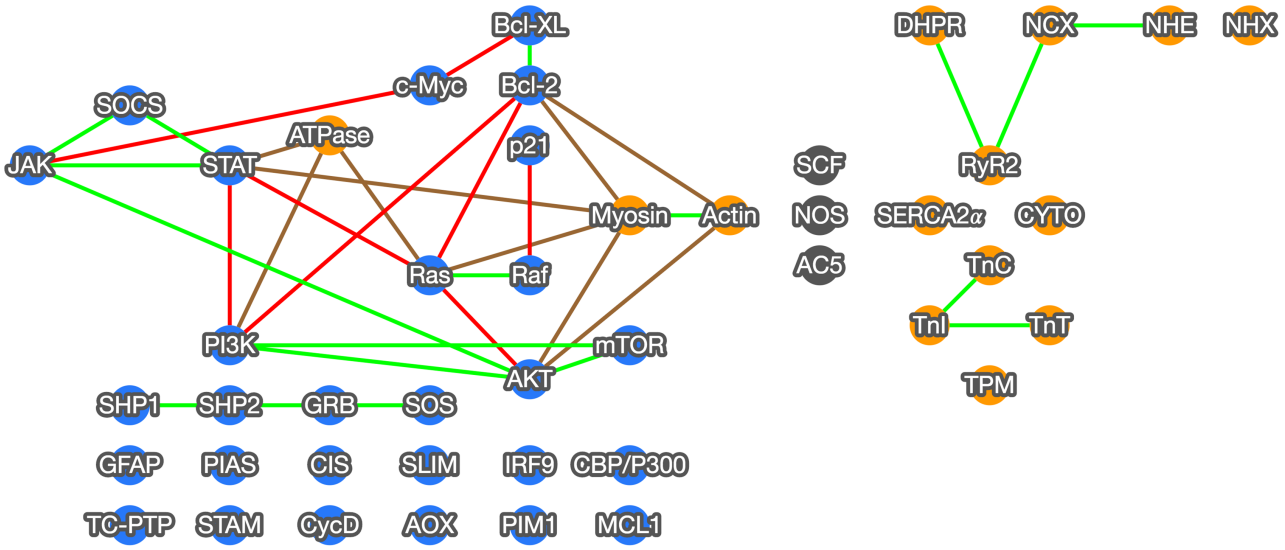

Fig 10. Consensus network output. Mining genetic information from abstracts cited within PubMed, our method generated a consensus network using large numbers of Bayesian networks representing potential genetic interactions within the JAK-STAT signaling (blue nodes) and Cardiac Muscle Contraction (orange nodes) KEGG pathways. Gray nodes are randomly selected negative controls. This graph shows the consensus network at a resolution of 0.9. Edges between genes indicate an inferred relationship. Green lines indicate a conserved interaction between the consensus network and the KEGG pathway, red lines indicate a novel interaction within the pathway, and brown lines indicate pathway-pathway interactions.

https://doi.org/10.1371/journal.pone.0186004.g010

pathways but also to help derive putative functional relationships previously unknown [43]. The data gathered within this study has focused on leveraging consensus networks to derive causal relationships that aim to uncover pathway interactions beyond KEGG curated pathways. Specifically, we tested our method on the current state of biochemical signaling knowledge on JAK-STAT signaling with an aim to distill them to core relationships without the use of significant computational resources. As shown in Fig 9, the optimal strategy for the number of topological inputs is essentially four (which is far less than 100+) and results in a significant reduction in computational power and in turn a reduction in computational time.

As shown in Fig 10, using a resolution of 0.90, 35 inferred interactions are present within the consensus network. The nodes are displayed to mirror their biochemical pathway positioning within the KEGG database for JAK-STAT signaling and Cardiac Muscle Contraction pathways. Illustrated in blue, the JAK-STAT signaling KEGG pathway is signaled through cytokines. Once bound to its receptor, Janus Kinase (*JAK*) based signaling is initiated resulting in subsequent downstream signaling. Illustrated in orange, The Cardiac Muscle Contraction and three randomly chosen negative control genes, shown in gray, are included to illustrate correct functioning of the algorithm and if pathway-pathway interactions are present. Shown in green, we found 18 similar interactions between the two KEGG pathways and our consensus network, with eight putative novel direct interactions shown in red. Nine pathway-pathway interactions were discovered as shown in brown.

The mammalian JAK-STAT signaling pathway has been extensively studied for the past twenty years [44] in mammals through gene knockout models, crystallography studies, and the use of antibodies for detection of specific isoforms and pharmokinetic studies to unravel disease pathologies [45]. JAK-STAT signaling is intimately associated with cell membrane cytokine receptors and at its core signaling, there are over 35 different cytokine receptor combinations with over 37 cytokines they respond to [46]. Specifically in mammals, there are four *JAKs* and seven *STATs* that are used in combination by over 50 cytokines [47]. These

redundant, yet varying combinations allow for tissue specific responses. Essentially most all cytokine receptors signal through *JAKs* and their subsequent downstream effector molecules, signal transducer of activation (*STAT*) proteins, while others signal through *MAP* Kinase cascade. Nonetheless, we focused this study within the proximal JAK-STAT signaling pathway.

As JAK-STAT pathways are involved in a plethora of Metazoan biochemical processes, most are focused in cellular growth, differentiation, survival, and resistance to cell death [48]. Interestingly, we did include ubiquitin mediated proteolysis which is involved in tagging cells for degradation, as well as *MAPK*, which can manipulate cellular proliferation and growth through alteration of cellular transcription [49]. Additionally, the *PI3K-AKT* signaling pathway is also involved in cytokine signaling, involved in modifying cell survival, proliferation, and even glucose metabolism [50]. Thus the JAK-STAT pathway has a multi-pronged approach to modifying cellular behavior and our consensus network identified three distinct interactions, and their corresponding nodes were identified as conserved between the KEGG pathway and our consensus network; *JAK-STAT-PI3K-AKT-mTOR*, *JAK-STAT-SOCS*, and *SHP1-SHP2-GRB-SOS* (Fig 10).

Conserved pathway interactions

One of the largest conserved interactions between our consensus network and KEGG was *JAK-STAT-PI3K-AKT-mTOR* Fig 10. This pathway is responsible for a significant number of interactions which will be described below. With *JAK* responsible for phosphorylating STAT, and *STAT* dimerization translocated to the nucleus to cause transcriptional changes, JAK and *STAT* are intimately involved. *STAT* also has an interaction with PI3K signaling [51]. PI3K is a lipid kinase that controls signaling and cell regulation [52]. For example, in human tumor cell lines, a proteomic study found a significant link between *PI3K* and *STAT3* in human cancer causing enhanced phosphorylation, with its effects reversible with PI3K inhibitors [53]. Despite being responsible for cell regulation, *PI3K-AKT* is also related to tumorigenesis [54]. With its concordant discovery in the early 1980's, the *PI3K-AKT* pathway is a result of *PI3K* enzyme activation phosphorylating membrane inositol lipids, impacting cellular signal transduction [55, 56]. Once activated, this drives *AKT- PKB* kinases to change their conformation and phosphorylation. Activated *AKT* then moves to the nucleus, activating target sequences involved in cell proliferation, survival, growth, and angiogenesis [55]. One hormone that can specifically modulate this pathway is insulin, which modifies glucose uptake, among others [57], which is critical to cell growth. As tumor detection and mortality rates have slowly changed over the past decade, an enhanced vigor in therapeutic targets have revolved around the *PI3K-AKT* pathway for cancer drug discovery [58–60], specifically with the addition of genomic studies [61]. Ultimately, *mTOR* (mechanistic target of rapamycin) is the end kinase complex in our conserved interaction that is responsible for protein synthesis, autophagy, and cell survival which interestingly can be modulated by nutrients and growth factors as previously described above [62]. Collectively this cascade is critical to cell function and an inability to keep cellular proliferation and signaling in check can lead to disease and is why these genes act together.

Our consensus network also found a conserved signaling cascade between *JAK-STAT* and *SOCS*. Much of the JAK-STAT signaling pathway is related to internal cellular signaling, therefore, having signaling molecules in place to provide negative feedback is critical to halt runaway responses. The suppressor of cytokine signaling (SOCS) proteins include approximately eight intracellular proteins (*SOCS1-7*, and *CIS*) and function as E3 ubiquitin ligases that mediate the degradation of proteins [63, 64]. *SOCS* proteins were initially coined as repressors of cytokine signaling and loss of function or disruption in signaling lead to chronic inflammation

and unchecked cellular growth [65, 66]. Thus, it is of no surprise that *JAK-STAT* and *SOCS* developed a conserved interaction in our consensus network as fidelity in signaling is critical to regulated cytokine signaling.

A third conserved interaction is *SHP1-SHP2-GRB-SOS*. *SHP1* and *SHP2* play critical roles in cellular growth, differentiation and cellular chemotaxis, but have also been implicated in cancer, neurodegeneration, and metabolic disorders such as diabetes [67]. *SHP1* and *SHP2* are cytoplasmic protein tyrosine phosphatases that work in concert but also oppose cellular cascades, with *SHP1* a negative transducer and *SHP2* a positive signal transducer [68]. Of note, a reduction in *SHP1* is observed in lymphomas and luekemias [69], whereas a reduction in *SHP2* or knockout of *SHP2* gene causes death mid gestation in mice [70], overall reduced *STAT* activation [71], and even leads to Noonan syndrome [72]. *SHP2* activity is a well-known binder of GRB2 [73], and *GRB2* binds to Son of Sevenless (*SOS*). This overall binding to a guanine nucleotide exchange factor is the mode by which reactivation of *RAS* is accomplished. This allows re-engagement of *RAS*, an oncogene, and *RAF* [74, 75]. Mutations in either *RAS* or *RAF* have been linked to several human cancers [76]. *SOS* helps to bind GTPases which are essentially molecular switches for cellular activity that hydrolyzes *GTP* to *GDP* plus phosphate that induces *RAS* activation and ultimately cellular proliferation and differentiation. Mutations in *SOS* can lead to hereditary gingival fibromatosis type 1 [53]. It is not surprising that some of the most talked about and well-known pathways have a very direct relationship and easily measurable tangibles as shown in these three examples. However, we also observed other interactions in our consensus network that can not be superimposed over the KEGG database pathway, which represent significant correlations within JAK-STAT.

Novel pathway interactions

One of the most significantly novel JAK-STAT network interactions was *RAS-Bcl-2*. These two cell signaling genes are involved in cell fate with the former modulating *Bcl-2* activity and *Bcl-2*, involved in anti-apoptosis. Despite these two not being in a linear relationship within the JAK-STAT signaling cascade, previous functional relationships have been observed between *Bcl-2* and members of the Ras superfamily [77–79]. For example, though *Bcl-2* can be a ruler of anti-apoptosis and thus cancer cell survival, cross-talk interactions between small GTPases which are pharmacologically inhibited actually reverted death phenotypes of *Bcl-2* expressing cells, essentially modifying anti-apoptosis and thus providing an additional target for drug resistant cancers [78]. These poorly understood crosstalk interactions have only recently been identified [80] and our use of consensus networks has reaffirmed and also highlighted significant correlative protein interactions. Upstream of *Bcl-2* and *RAS* is a negative repressor of JAK-STAT signaling, (*SOCS*), which works to halt runaway activation of *JAK*.

Another unique novel interaction found within our consensus network was *RAS-AKT*. *RAS*, ras viral oncogene homolog, is a small GTPase that when bound with *GTP* is activated and used to transduce a signal primarily through *MAPK* signaling pathways (for example *RAF*) as shown in Fig 10. *AKT* (v-akt murine thymoma viral oncogene homolog) is a known oncogene and can cause rare genetic diseases such as Proteus Syndrome [81] and Cowden Syndrome 6 [82]. Other studies have found in *AKT*-null mice an increased sensitization to other cytokines like tumor necrosis factor and even genotoxic effects induced by gamma or ultraviolet radiation [83]. As both *AKT* and *RAS* lie within cell cycles responsible for cell survival and proliferation, respectively, it is not surprising that they may have an interaction. Wang and colleagues [84] found a putative functional relationship between *AKT-mTOR* and *RAS-MAPK* pathways in liver cancer. They identified a concomitant increase in *AKT* and *RAS* resulted in enhanced liver tissue carcinogenesis [85] whereas suppression of gene expression

reduced cell growth. Thus these two pathways play an intimate co-functional role in cell survival and proliferation.

Lastly, another interaction identified within our Consensus network as novel within the confines of the KEGG pathway is *c-Myc* and *AKT*. *C-Myc* is an oncogene that is responsible for modulating transcription which impacts cell-cycle proliferation but is also critical to programmed cell death [86]. It is modulated by mitogenic signals and repressed by growth inhibitory signals [87]. Disruption in *c-Myc* results in significant reduction of cell growth [88] and has been linked to several cancers despite a half-life of 20–30 minutes [89]. A few studies have identified an inverse relationship between *AKT* and *c-Myc* [90, 91], with some finding that inhibition of *c-Myc* with MadMyc suppressed heptaocarcinoma development with a corresponding *AKT* induction [92]. These studies collectively identify that the significant interplay between unchecked cell proliferation and survival is what can ultimately lead to carcinomas. Just as additional research is needed to better unravel the shift from asymptomatic cancerous tissue to nutrient deprived and stressed cellular machinery that regulate cell apoptosis, cell proliferation, and other collective processes, these consensus networks point to conserved pathways and also uncover putative novel protein-protein interactions that can be further validated via genome-wide association studies, RNA-seq, and microarray analyses.

Pathway-pathway interactions

In addition to the novel, community interactions inferred using our method, we also added three randomly chosen genes (Fig 10, highlighted in gray) as well as the Cardiac Muscle Contraction KEGG pathway, to test whether our algorithm could recapitulate conserved community interactions that are defined within KEGG and also resist the addition of randomly inserted genes and pathways. Based on data shown within Fig 10, our consensus network found several conserved, as well as unique, interactions within the KEGG pathways. With the addition of the Cardiac Muscle Contraction pathway, three gene groups found a significant pathway-pathway interaction. *ATPase* is an ancestral ATPase Na^+/K^+ pump that maintains an electrochemical gradient within muscle cells [93]. Much of this interaction is widely debated but more specifically it binds to transcription regulators like *SNW1*; a transcription regulator *SKIP* (Ski- interacting protein); *SMAD* family members like *SMAD7*, which is a known antagonist of *TGF-β* (Transforming growth factor-β); and even *PDPK1* kinase activity, which is responsible for activation of *AKT*, which is downstream of *PI3K*. Disruptions in Na^+/K^+ -ATPase function can lead to thyrotoxic periodic paralysis [94] and McArdle disease [95]. *Atpase* has a novel interaction with *PI3K* [96] and interestingly ouabain-induced signaling can impact the *EGFR-Src-RAS-ERK* pathway and also *PI3K1A-PDK-AKT* pathway [92, 97], with the latter resulting in hypertrophy solely in differentiated cardiac myocytes [98].

In addition to a Na^+/K^+ -*ATPase*, *Myosin* and *Actin* were also found to be present within the JAK-STAT portion of the consensus network. *Actin* is the most abundant protein present within the human body and is also one of the most conserved across a plethora of Metazoans. *Actin* and its isoforms are responsible for several functions, however it is most well-known for its cytoskeleton nature; within cardiac muscle contraction, thin filament *actin* is the scaffold upon which *myosin* contractions work [99]. As such, it is no surprise *actin* and *myosin* produce an interaction, however there are also interactions with *AKT* and *Bcl-2*. *Bcl-2*, as indicated previously, is a protein responsible for cellular metabolism, cell fate, mitochondrial function and ultimately cell proliferation with disruptions in its signaling causing B-cell follicular lymphoma [100]. Bcl-2 is also localized to the endoplasmic reticulum and regulates the control of Ca^{2+} levels which also influences *actin* cytoskeleton development and provides a cue in apoptosis [101]. Interestingly, *Bcl-2* is also found to enhance *F-actin* and *myosin* polymerization which

results in inhibition of cell-cell adhesion and motility [102, 103]. Thus *Bcl-2's* link to *myosin* and *actin* underline the importance of metastasizing cancers and reiterate the interrelatedness of the internal milieu within the human body. Collectively, these new pathway-pathway interactions reiterate the functional interrelatedness of biochemical pathways that are missed when just looking at a specific pathway alone and though our algorithm does not provide a causal relationship of interaction, the interaction is a powerful tool to identify significant community relationships among proteins.

Conclusion

Our method of inferring gene interactions can be helpful to investigators in multiple ways. With our method, the confirmation of currently accepted gene interactions as well as the hypotheses of new interactions is possible. By using consensus networks created with multiple, randomly generated topological inputs, inferences can be made about gene interactions without prior knowledge of the gene functions themselves. Any group of genes can be examined for possible interactions, regardless of biological system, and even across organisms. The method of consensus network generation with randomized topological inputs could also be adapted for use with other sources of data: microarray, RNA-Seq, and mutation data. The more prior knowledge that is applied to the initial gene selection, the more precise and informative the results will be. Additionally, the current version of our text searching and parsing is very rudimentary; the addition of natural language processing could greatly increase the effectiveness of the entire process, even allowing the detection of specific interaction types. In its current form, our method uses R and Matlab to query and pre-process data, with a large parallelized, petascale system to generate Bayesian networks. We are moving towards releasing a web-driven front end that would allow investigators to use this system with their own data, integrated KEGG pathway data, and data from large experiment databases such as the GDC.

Supporting information

S1 File. Prior knowledge matrix for data set 1. A matrix representing the prior knowledge input used with creating our Bayesian networks. The matrix has 28 columns and 42600 rows. The rows consist of publications which contain at least two of the functional gene groups we are interested in. The first column is the PubMed ID of the publication represented by the current row. The remaining 27 columns represent the functional gene groups given in Fig 3. The value of these 27 columns are '0' if the publication did not contain the column's genes group, and '1' if the group was present within the publications abstract.
(CSV)

S2 File. Prior knowledge matrix for data set 2. A matrix representing the prior knowledge input used with creating our Bayesian networks. The matrix has 46 columns and 96,594 rows. The rows consist of publications which contain at least two of the functional gene groups we are interested in. The first column is the PubMed ID of the publication represented by the current row. The remaining 45 columns represent the functional gene groups given in Fig 4. The value of these 45 columns are '0' if the publication did not contain the column's genes group, and '1' if the group was present within the publications abstract.
(CSV)

S3 File. Consensus network adjacency matrix for data set 1. The adjacency matrix representing the consensus network generated with 42,600 publications, 27 functional gene groups, and 100 random topologies per group resulting in 2700 total Bayesian network inputs. Each row and column index represents the matching group from Fig 3. The numbers within the matrix

represent the number of Bayesian networks that contained the edge connecting the group with the column group.
(CSV)

S4 File. Consensus network adjacency matrix for data set 2. The adjacency matrix representing the consensus network generated with 96,594 publications, 45 functional gene groups, and eight random topologies per group resulting in 360 total Bayesian network inputs. Each row and column index represents the matching group from Fig 4. The numbers within the matrix represent the number of Bayesian networks that contained the edge connecting the group with the column group.
(CSV)

S5 File. Sample size kappa coefficients. The Cohen's kappa coefficient results for each pair of consensus networks used to test stability across sample sizes. There are 10 different tests in the file. Each test contains 10 network resolutions, each having its own set of four values used to calculate the kappa coefficient.
(XLS)

S6 File. Topological kappa coefficients. The Cohen's kappa coefficient results for each pair of consensus networks used to test stability across topological input sizes. There are 25 different tests in the file. Each test contains 10 network resolutions, each having its own set of four values used to calculate the kappa coefficient.
(XLS)

Author Contributions

Conceptualization: Anthony Deeter, Zhong-Hui Duan.

Data curation: Anthony Deeter.

Formal analysis: Anthony Deeter, Mark Dalman.

Investigation: Anthony Deeter.

Methodology: Anthony Deeter, Joseph Haddad.

Project administration: Zhong-Hui Duan.

Software: Anthony Deeter, Joseph Haddad.

Supervision: Zhong-Hui Duan.

Validation: Anthony Deeter.

Visualization: Anthony Deeter.

Writing – original draft: Anthony Deeter, Mark Dalman.

Writing – review & editing: Anthony Deeter, Mark Dalman, Joseph Haddad, Zhong-Hui Duan.

References

1. PubMed; 2016. Available from: http://www.ncbi.nlm.nih.gov/pubmed [cited 10/20/2016].
2. Plikus MV, Zhang Z, Chuong CM. PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. BMC Bioinformatics. 2006; 7:424. https://doi.org/10.1186/1471-2105-7-424 PMID: 17014720

3. Muin M, Fontelo P, Liu F, Ackerman M. SLIM: an alternative Web interface for MEDLINE/PubMed searches—a preliminary study. BMC Med Inform Decis Mak. 2005; 5:37. https://doi.org/10.1186/1472-6947-5-37 PMID: 16321145
4. Ding J, Berleant D. MedKit: a helper toolkit for automatic mining of MEDLINE/PubMed citations. Bioinformatics. 2005; 21(5):694–695. https://doi.org/10.1093/bioinformatics/bti087 PMID: 15509599
5. Ding J, Hughes LM, Berleant D, Fulmer AW, Wurtele ES. PubMed Assistant: a biologist-friendly interface for enhanced PubMed search. Bioinformatics. 2006; 22(3):378–380. https://doi.org/10.1093/bioinformatics/bti821 PMID: 16332704
6. Perez-Iratxeta C, Bork P, Andrade MA. XplorMed: a tool for exploring MEDLINE abstracts. Trends Biochem Sci. 2001; 26(9):573–575. https://doi.org/10.1016/S0968-0004(01)01926-0 PMID: 11551795
7. Perez-Iratxeta C, Perez AJ, Bork P, Andrade MA. Update on XplorMed: A web server for exploring scientific literature. Nucleic Acids Res. 2003; 31(13):3866–3868. https://doi.org/10.1093/nar/gkg538 PMID: 12824439
8. Goetz T, von der Lieth CW. PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts. Nucleic Acids Res. 2005; 33(Web Server issue):W774–778. https://doi.org/10.1093/nar/gki429 PMID: 15980583
9. Gene Expression Omnibus; 2016. Available from: http://www.ncbi.nlm.nih.gov/geo/ [cited 10/20/2016].
10. Genomic Data Commons; 2016. Available from: https://gdc.cancer.gov/ [cited 10/20/2016].
11. Petricoin EF, Hackett JL, Lesko LJ, Puri RK, Gutman SI, Chumakov K, et al. Medical applications of microarray technologies: a regulatory science perspective. Nat Genet. 2002; 32 Suppl:474–479. https://doi.org/10.1038/ng1029 PMID: 12454641
12. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995; 270(5235):467–470. https://doi.org/10.1126/science.270.5235.467 PMID: 7569999
13. Reis-Filho JS. Next-generation sequencing. Breast Cancer Res. 2009; 11 Suppl 3:S12. https://doi.org/10.1186/bcr2431 PMID: 20030863
14. Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, et al. Annotating genomes with massive-scale RNA sequencing. Genome Biol. 2008; 9(12):R175. https://doi.org/10.1186/gb-2008-9-12-r175 PMID: 19087247
15. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009; 55(4):641–658. https://doi.org/10.1373/clinchem.2008.112789 PMID: 19246620
16. Lancashire LJ, Lemetre C, Ball GR. An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies. Brief Bioinform. 2009; 10(3):315–329. https://doi.org/10.1093/bib/bbp012 PMID: 19307287
17. Yang ZR. Biological applications of support vector machines. Brief Bioinformatics. 2004; 5(4):328–338. https://doi.org/10.1093/bib/5.4.328 PMID: 15606969
18. Pearl J. Probabilistic Reasoning in Intelligent Systems. San Mateo, CA: Morgan Kaufmann; 1988.
19. Bielza C, Larranaga P. Bayesian networks in neuroscience: a survey. Front Comput Neurosci. 2014; 8:131. https://doi.org/10.3389/fncom.2014.00131 PMID: 25360109
20. Stapley BJ, Benoit G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pac Symp Biocomput. 2000; p. 529–540. PMID: 10902200
21. Zhou D, He Y. Extracting interactions between proteins from the literature. J Biomed Inform. 2008; 41(2):393–407. https://doi.org/10.1016/j.jbi.2007.11.008 PMID: 18207462
22. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics. 2004; 5:147. https://doi.org/10.1186/1471-2105-5-147 PMID: 15473905
23. Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, et al. Annotating cancer variants and anticancer therapeutics in reactome. Cancers (Basel). 2012; 4(4):1180–1211. https://doi.org/10.3390/cancers4041180
24. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014; 42(Database issue):D472–477. https://doi.org/10.1093/nar/gkt1102 PMID: 24243840
25. Ingenuity Pathway Analysis; 2016. Available from: http://www.ingenuity.com [cited 10/20/2016].
26. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016; 44(D1):D457–462. https://doi.org/10.1093/nar/gkv1070 PMID: 26476454

27. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28 (1):27–30. https://doi.org/10.1093/nar/28.1.27 PMID: 10592173

28. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic acids research. 2006; 34(suppl_1):D535–D539. https://doi.org/10.1093/nar/gkj109 PMID: 16381927

29. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, et al. IntAct—open source resource for molecular interaction data. Nucleic acids research. 2006; 35(suppl_1):D561–D565. https://doi.org/10.1093/nar/gkl958 PMID: 17145710

30. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic acids research. 2005; 33(suppl 1):D428–D432. https://doi.org/10.1093/nar/gki072 PMID: 15608231

31. Werhli AV, Husmeier D. Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. Stat Appl Genet Mol Biol. 2007; 6:Article15. PMID: 17542777

32. Sriram S. Predicting Gene Relations Using Bayesian Networks; 2011.

33. Cooper GF, Herskovits E. A Bayesian method for constructing Bayesian belief networks from databases. In: Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc.; 1991. p. 86–94.

34. Holland BR, Huber KT, Moulton V, Lockhart PJ. Using consensus networks to visualize contradictory evidence for species phylogeny. Molecular Biology and Evolution. 2004; 21(7):1459–1461. https://doi.org/10.1093/molbev/msh145 PMID: 15084681

35. Huson DH, Dezulian T, Klopper T, Steel MA. Phylogenetic super-networks from partial trees. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2004; 1(4):151–158. https://doi.org/10.1109/TCBB.2004.44 PMID: 17051697

36. Peña JM. Finding consensus Bayesian network structures. Journal of Artificial Intelligence Research. 2011; 42:661–687.

37. Cohen J. A coefficient of agreement for nominal scale. Educ Psychol Meas. 1960; 20:37–46. https://doi.org/10.1177/001316446002000104

38. Kovalchik S. RISmed: Download Content from NCBI Databases; 2015. Available from: http://CRAN.R-project.org/package=RISmed.

39. Tenenbaum D. KEGGREST: Client-side REST access to KEGG; 2016.

40. Zhang JD, Wiemann S. KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor. Bioinformatics. 2009; 25(11):1470–1471. https://doi.org/10.1093/bioinformatics/btp167 PMID: 19307239

41. Koonin EV, Altschul SF, Bork P. BRCA1 protein products: functional motifs. Nat Genet. 2000; 13:266–267.

42. Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, Carlin DE, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. Nature methods. 2016; 13(4):310. https://doi.org/10.1038/nmeth.3773 PMID: 26901648

43. Poultney CS, Greenfield A, Bonneau R. Integrated inference and analysis of regulatory networks from multi-level measurements. Methods Cell Biol. 2012; 110:19–56. https://doi.org/10.1016/B978-0-12-388403-9.00002-3 PMID: 22482944

44. Stark GR, Darnell JE. The JAK-STAT pathway at twenty. Immunity. 2012; 36(4):503–514. https://doi.org/10.1016/j.immuni.2012.03.013 PMID: 22520844

45. Rawlings JS, Rosler KM, Harrison DA. The JAK/STAT signaling pathway. J Cell Sci. 2004; 117(Pt 8):1281–1283. https://doi.org/10.1242/jcs.00963 PMID: 15020666

46. Murray PJ. The JAK-STAT signaling pathway: input and output integration. The Journal of Immunology. 2007; 178(5):2623–2629. https://doi.org/10.4049/jimmunol.178.5.2623 PMID: 17312100

47. Villarino AV, Kanno Y, Ferdinand JR, O'Shea JJ. Mechanisms of Jak/STAT signaling in immunity and disease. The Journal of Immunology. 2015; 194(1):21–27. https://doi.org/10.4049/jimmunol.1401867 PMID: 25527793

48. Igaz P, Toth S, Falus A. Biological and clinical significance of the JAK-STAT pathway; lessons from knockout mice. Inflamm Res. 2001; 50(9):435–441. https://doi.org/10.1007/PL00000267 PMID: 11603847

49. Seger R, Krebs EG. The MAPK signaling cascade. FASEB J. 1995; 9(9):726–735. PMID: 7601337

50. Schultze SM, Hemmings BA, Niessen M, Tschopp O. PI3K/AKT, MAPK and AMPK signalling: protein kinases in glucose homeostasis. Expert Rev Mol Med. 2012; 14:e1. https://doi.org/10.1017/S1462399411002109 PMID: 22233681

51. Liongue C, Ward AC. Evolution of the JAK-STAT pathway. JAKSTAT. 2013; 2(1):e22756. https://doi.org/10.4161/jkst.22756 PMID: 24058787

52. Cantley LC. The phosphoinositide 3-kinase pathway. Science. 2002; 296(5573):1655–1657. https://doi.org/10.1126/science.296.5573.1655 PMID: 12040186

53. Hart TC, Zhang Y, Gorry MC, Hart PS, Cooper M, Marazita ML, et al. A mutation in the SOS1 gene causes hereditary gingival fibromatosis type 1. Am J Hum Genet. 2002; 70(4):943–954. https://doi.org/10.1086/339689 PMID: 11868160

54. Hemmings BA, Restuccia DF. Pi3k-pkb/akt pathway. Cold Spring Harbor perspectives in biology. 2012; 4(9):a011189. https://doi.org/10.1101/cshperspect.a011189 PMID: 22952397

55. Sarker D, Reid AH, Yap TA, de Bono JS. Targeting the PI3K/AKT pathway for the treatment of prostate cancer. Clinical Cancer Research. 2009; 15(15):4799–4805. https://doi.org/10.1158/1078-0432.CCR-08-0125 PMID: 19638457

56. Vivanco I, Sawyers CL. The phosphatidylinositol 3-kinase—AKT pathway in human cancer. Nature Reviews Cancer. 2002; 2(7):489–501. https://doi.org/10.1038/nrc839 PMID: 12094235

57. Rafalski VA, Brunet A. Energy metabolism in adult neural stem cell fate. Prog Neurobiol. 2011; 93 (2):182–203. https://doi.org/10.1016/j.pneurobio.2010.10.007 PMID: 21056618

58. Chang F, Lee J, Navolanic P, Steelman L, Shelton J, Blalock W, et al. Involvement of PI3K/Akt pathway in cell cycle progression, apoptosis, and neoplastic transformation: a target for cancer chemotherapy. Leukemia. 2003; 17(3):590–603. https://doi.org/10.1038/sj.leu.2402824 PMID: 12646949

59. Hennessy BT, Smith DL, Ram PT, Lu Y, Mills GB. Exploiting the PI3K/AKT pathway for cancer drug discovery. Nature reviews Drug discovery. 2005; 4(12):988–1004. https://doi.org/10.1038/nrd1902 PMID: 16341064

60. Osaki M, Oshimura M, Ito H. PI3K-Akt pathway: its functions and alterations in human cancer. Apoptosis. 2004; 9(6):667–676. https://doi.org/10.1023/B:APPT.0000045801.15585.dd PMID: 15505410

61. Smit L, Berns K, Spence K, Ryder W, Zeps N, Madiredjo M, et al. An integrated genomic approach identifies that the PI3K/AKT/FOXO pathway is involved in breast cancer tumor initiation. Oncotarget. 2015;.

62. Hay N, Sonenberg N. Upstream and downstream of mTOR. Genes Dev. 2004; 18(16):1926–1945. https://doi.org/10.1101/gad.1212704 PMID: 15314020

63. Kubo M, Hanada T, Yoshimura A. Suppressors of cytokine signaling and immunity. Nat Immunol. 2003; 4(12):1169–1176. https://doi.org/10.1038/ni1012 PMID: 14639467

64. Yoshimura A. Regulation of cytokine signaling by the SOCS and Spred family proteins. Keio J Med. 2009; 58(2):73–83. https://doi.org/10.2302/kjm.58.73 PMID: 19597303

65. Jo D, Liu D, Yao S, Collins RD, Hawiger J. Intracellular protein therapy with SOCS3 inhibits inflammation and apoptosis. Nat Med. 2005; 11(8):892–898. https://doi.org/10.1038/nm1269 PMID: 16007096

66. Croker BA, Kiu H, Nicholson SE. SOCS regulation of the JAK/STAT signalling pathway. Semin Cell Dev Biol. 2008; 19(4):414–422. https://doi.org/10.1016/j.semcdb.2008.07.010 PMID: 18708154

67. Bialy L, Waldmann H. Inhibitors of Protein Tyrosine Phosphatases: Next-Generation Drugs? Angewandte Chemie International Edition. 2005; 44(25):3814–3839. https://doi.org/10.1002/anie.200461517 PMID: 15900534

68. Wang N, Li Z, Ding R, Frank GD, Senbonmatsu T, Landon EJ, et al. Antagonism or synergism. Role of tyrosine phosphatases SHP-1 and SHP-2 in growth factor signaling. J Biol Chem. 2006; 281 (31):21878–21883. https://doi.org/10.1074/jbc.M605018200 PMID: 16762922

69. Oka T, Yoshino T, Hayashi K, Ohara N, Nakanishi T, Yamaai Y, et al. Reduction of hematopoietic cell-specific tyrosine phosphatase SHP-1 gene expression in natural killer cell lymphoma and various types of lymphomas/leukemias: combination analysis with cDNA expression array and tissue microarray. Am J Pathol. 2001; 159(4):1495–1505. https://doi.org/10.1016/S0002-9440(10)62535-7 PMID: 11583976

70. Saxton TM, Henkemeyer M, Gasca S, Shen R, Rossi DJ, Shalaby F, et al. Abnormal mesoderm patterning in mouse embryos mutant for the SH2 tyrosine phosphatase Shp-2. EMBO J. 1997; 16 (9):2352–2364. https://doi.org/10.1093/emboj/16.9.2352 PMID: 9171349

71. Feng GS. Shp-2 tyrosine phosphatase: signaling one cell or many. Exp Cell Res. 1999; 253(1):47–54. https://doi.org/10.1006/excr.1999.4668 PMID: 10579910

72. Tartaglia M, Mehler EL, Goldberg R, Zampino G, Brunner HG, Kremer H, et al. Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. Nat Genet. 2001; 29 (4):465–468. https://doi.org/10.1038/ng772 PMID: 11704759

73. Cunnick JM, Dorsey JF, Munoz-Antonia T, Mei L, Wu J. Requirement of SHP2 binding to Grb2-associated binder-1 for mitogen-activated protein kinase activation in response to lysophosphatidic acid and

epidermal growth factor. J Biol Chem. 2000; 275(18):13842–13848. https://doi.org/10.1074/jbc.275.18.13842 PMID: 10788507

74. Shields JM, Pruitt K, McFall A, Shaub A, Der CJ. Understanding Ras: 'it ain't over 'til it's over'. Trends Cell Biol. 2000; 10(4):147–154. https://doi.org/10.1016/S0962-8924(00)01740-2 PMID: 10740269

75. Santarpia L, Lippman SM, El-Naggar AK. Targeting the MAPK-RAS-RAF signaling pathway in cancer therapy. Expert Opin Ther Targets. 2012; 16(1):103–119. https://doi.org/10.1517/14728222.2011.645805 PMID: 22239440

76. Downward J. Targeting RAS signalling pathways in cancer therapy. Nat Rev Cancer. 2003; 3(1):11–22. https://doi.org/10.1038/nrc969 PMID: 12509763

77. Braun F, de Carné Trécesson S, Bertin-Ciftci J, Juin P. Protect and serve: Bcl-2 proteins as guardians and rulers of cancer cell survival. Cell Cycle. 2013; 12(18):2937–2947. https://doi.org/10.4161/cc.25972 PMID: 23974114

78. Kang J, Pervaiz S. Crosstalk between Bcl-2 family and Ras family small GTPases: potential cell fate regulation. Front Oncol. 2012; 2(206.10):3389.

79. Velaithan R, Kang J, Hirpara JL, Loh T, Goh BC, Le Bras M, et al. The small GTPase Rac1 is a novel binding partner of Bcl-2 and stabilizes its antiapoptotic activity. Blood. 2011; 117(23):6214–6226. https://doi.org/10.1182/blood-2010-08-301283 PMID: 21474673

80. Murugan AK, Munirajan AK, Tsuchida N. Ras oncogenes in oral cancer: the past 20 years. Oral oncology. 2012; 48(5):383–392. https://doi.org/10.1016/j.oraloncology.2011.12.006 PMID: 22240207

81. Keppler-Noreuil KM, Lozier JN, Sapp JC, Biesecker LG. Characterization of thrombosis in patients with Proteus syndrome. Am J Med Genet A. 2017; 173(9):2359–2365. https://doi.org/10.1002/ajmg.a.38311 PMID: 28627093

82. Orloff MS, He X, Peterson C, Chen F, Chen JL, Mester JL, et al. Germline PIK3CA and AKT1 mutations in Cowden and Cowden-like syndromes. The American Journal of Human Genetics. 2013; 92(1):76–80. https://doi.org/10.1016/j.ajhg.2012.10.021 PMID: 23246288

83. Chen WS, Xu PZ, Gottlob K, Chen ML, Sokol K, Shiyanova T, et al. Growth retardation and increased apoptosis in mice with homozygous disruption of the Akt1 gene. Genes Dev. 2001; 15(17):2203–2208. https://doi.org/10.1101/gad.913901 PMID: 11544177

84. Wang C, Cigliano A, Delogu S, Armbruster J, Dombrowski F, Evert M, et al. Functional crosstalk between AKT/mTOR and Ras/MAPK pathways in hepatocarcinogenesis: implications for the treatment of human liver cancer. Cell Cycle. 2013; 12(13):1999–2010. https://doi.org/10.4161/cc.25099 PMID: 23759595

85. Ho C, Wang C, Mattu S, Destefanis G, Ladu S, Delogu S, et al. AKT and N-Ras co-activation in the mouse liver promotes rapid carcinogenesis via mTORC1, FOXM1/SKP2, and c-Myc pathways. Hepatology (Baltimore, Md). 2012; 55(3):833.

86. Amati B, Littlewood T, Evan G, Land H. The c-Myc protein induces cell cycle progression and apoptosis through dimerization with Max. The EMBO Journal. 1993; 12(13):5083. PMID: 8262051

87. Harrington EA, Bennett MR, Fanidi A, Evan GI. c-Myc-induced apoptosis in fibroblasts is inhibited by specific cytokines. EMBO J. 1994; 13(14):3286–3295. PMID: 8045259

88. Mateyak MK, Obaya AJ, Adachi S, Sedivy JM. Phenotypes of c-Myc-deficient rat fibroblasts isolated by targeted homologous recombination. Cell Growth Differ. 1997; 8(10):1039–1048. PMID: 9342182

89. Luscher B, Eisenman RN. c-myc and c-myb protein degradation: effect of metabolic inhibitors and heat shock. Mol Cell Biol. 1988; 8(6):2504–2512. https://doi.org/10.1128/MCB.8.6.2504 PMID: 3043180

90. Rohn JL, Hueber AO, McCarthy NJ, Lyon D, Navarro P, Burgering BM, et al. The opposing roles of the Akt and c-Myc signalling pathways in survival from CD95-mediated apoptosis. Oncogene. 1998; 17(22):2811–2818. https://doi.org/10.1038/sj.onc.1202393 PMID: 9879987

91. Hoffman B, Amanullah A, Shafarenko M, Liebermann DA. The proto-oncogene c-myc in hematopoietic development and leukemogenesis. Oncogene. 2002; 21(21):3414–3421. https://doi.org/10.1038/sj.onc.1205400 PMID: 12032779

92. Xie Z, Askari A. Na(+)/K(+)-ATPase as a signal transducer. Eur J Biochem. 2002; 269(10):2434–2439. https://doi.org/10.1046/j.1432-1033.2002.02910.x PMID: 12027880

93. Pestov NB, Ahmad N, Korneenko TV, Zhao H, Radkov R, Schaer D, et al. Evolution of Na,K-ATPase beta m-subunit into a coregulator of transcription in placental mammals. Proc Natl Acad Sci USA. 2007; 104(27):11215–11220. https://doi.org/10.1073/pnas.0704809104 PMID: 17592128

94. Kung AW, Lau KS, Cheung WM, Chan V. Thyrotoxic periodic paralysis and polymorphisms of sodium-potassium ATPase genes. Clin Endocrinol (Oxf). 2006; 64(2):158–161. https://doi.org/10.1111/j.1365-2265.2005.02442.x

95. Haller RG, Clausen T, Vissing J, Blomqvist CG. Reduced levels of skeletal muscle Na+K+ -ATPase in McArdle disease. Neurology. 1998; 50(1):37–40. https://doi.org/10.1212/WNL.50.1.37 PMID: 9443454

96. Wu Q, Lv T, Chen Y, Wen L, Zhang J, Jiang X, et al. Apoptosis of HL-60 human leukemia cells induced by Asiatic acid through modulation of B-cell lymphoma 2 family proteins and the mitogen-activated protein kinase signaling pathway. Molecular medicine reports. 2015; 12(1):1429–1434. https://doi.org/10.3892/mmr.2015.3534 PMID: 25815462

97. Liu L, Zhao X, Pierre SV, Askari A. Association of PI3K-Akt signaling pathway with digitalis-induced hypertrophy of cardiac myocytes. Am J Physiol, Cell Physiol. 2007; 293(5):C1489–1497. https://doi.org/10.1152/ajpcell.00158.2007 PMID: 17728397

98. Bai Y, Morgan EE, Giovannucci DR, Pierre SV, Philipson KD, Askari A, et al. Different roles of the cardiac Na+/Ca2+-exchanger in ouabain-induced inotropy, cell signaling, and hypertrophy. Am J Physiol Heart Circ Physiol. 2013; 304(3):H427–435. https://doi.org/10.1152/ajpheart.00462.2012 PMID: 23203972

99. Sugiura S. Actin-myosin interaction. Cardiovasc Res. 1999; 44(2):266–273. https://doi.org/10.1016/S0008-6363(99)00219-9 PMID: 10690303

100. Prudent J, Popgeorgiev N, Bonneau B, Gillet G. Bcl-2 proteins, cell migration and embryonic development: lessons from zebrafish. Cell Death Dis. 2015; 6:e1910. https://doi.org/10.1038/cddis.2015.286 PMID: 26469959

101. Desouza M, Gunning PW, Stehn JR. The actin cytoskeleton as a sensor and mediator of apoptosis. Bioarchitecture. 2012; 2(3):75–87. https://doi.org/10.4161/bioa.20975 PMID: 22880146

102. Ke H, Parron VI, Reece J, Zhang JY, Akiyama SK, French JE. BCL2 inhibits cell adhesion, spreading, and motility by enhancing actin polymerization. Cell Res. 2010; 20(4):458–469. https://doi.org/10.1038/cr.2010.21 PMID: 20142842

103. Callagy GM, Webber MJ, Pharoah PD, Caldas C. Meta-analysis confirms BCL2 is an independent prognostic marker in breast cancer. BMC Cancer. 2008; 8:153. https://doi.org/10.1186/1471-2407-8-153 PMID: 18510726