Integrated Pathway Clusters with Coherent Biological Themes for Target Prioritisation

Yi-An Chen*, Lokesh P. Tripathi*, Benoit H. Dessailly[¤a], Johan Nyström-Persson[¤b], Shandar Ahmad, Kenji Mizuguchi*

National Institute of Biomedical Innovation, Ibaraki, Osaka, Japan

Abstract

Prioritising candidate genes for further experimental characterisation is an essential, yet challenging task in biomedical research. One way of achieving this goal is to identify specific biological themes that are enriched within the gene set of interest to obtain insights into the biological phenomena under study. Biological pathway data have been particularly useful in identifying functional associations of genes and/or gene sets. However, biological pathway information as compiled in varied repositories often differs in scope and content, preventing a more effective and comprehensive characterisation of gene sets. Here we describe a new approach to constructing biologically coherent gene sets from pathway data in major public repositories and employing them for functional analysis of large gene sets. We first revealed significant overlaps in gene content between different pathways and then defined a clustering method based on the shared gene content and the similarity of gene overlap patterns. We established the biological relevance of the constructed pathway clusters using independent quantitative measures and we finally demonstrated the effectiveness of the constructed pathway clusters in comparative functional enrichment analysis of gene sets associated with diverse human diseases gathered from the literature. The pathway clusters and gene mappings have been integrated into the TargetMine data warehouse and are likely to provide a concise, manageable and biologically relevant means of functional analysis of gene sets and to facilitate candidate gene prioritisation.

Citation: Chen Y-A, Tripathi LP, Dessailly BH, Nyström-Persson J, Ahmad S, et al. (2014) Integrated Pathway Clusters with Coherent Biological Themes for Target Prioritisation. PLoS ONE 9(6): e99030. doi:10.1371/journal.pone.0099030

Editor: I. King Jordan, Georgia Institute of Technology, United States of America

Received February 20, 2014; **Accepted** May 7, 2014; **Published** June 11, 2014

Funding: This work was in part supported by the Industrial Technology Research Grant Program in 2007 (Grant Number 07C46056a) from New Energy and Industrial Technology Development Organization (NEDO) of Japan, and also by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology (Grant Numbers 25430186, 25293079) and from the Ministry of Health, Labor, and Welfare ("The Adjuvant database project") to K.M. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: kenji@nibio.go.jp (KM); chenyian@nibio.go.jp (Y-AC); lokesh@nibio.go.jp (LPT)

These authors contributed equally to this work.

¤a Current address: Takeda Cambridge Ltd., Cambridge, United Kingdom
¤b Current address: Level Five Co. Ltd, Tokyo, Japan and Australia-China Centre for Wheat Improvement, Murdoch University, Perth, Australia

Introduction

There has been an exponential increase in the amount and complexity of biological data. Extracting meaningful biological insights from this vast array of data via functional analysis of the large resultant gene sets and to prioritise genes and gene sets for further experimental characterisation is a formidable challenge. Gene-set-functional-enrichment (GSFE) relies on a statistical analysis of the relative abundance of biological themes associated with a given gene set and identifies themes (and associated genes) that are overrepresented and therefore, likely to be more relevant to the biological conditions under study.

It is increasingly evident that gene and proteins do not function alone, but rather as a part of complex pathways where they interact with various biomolecules (such as proteins, nucleic acids and metabolites). Therefore, an accurate representation of biological pathway information is essential to understand the biological relevance of genes and proteins within specific biological contexts. An ever growing number of pathway databases, thereby, constitute an increasingly important component of any computational framework for the functional annotation of genes and proteins. However, the available pathway resources often differ widely in scope and content, which severely hampers a unified analysis and interpretation of high-throughput biological data using diverse pathway repositories [1–3]. In the absence of reasonable compatibility, a unified representation of gene function by leveraging the biological information stored in various pathway repositories remains a non-trivial task.

Integration of pathway repositories offers significantly attractive benefits in terms of more extensive and robust functional annotations, which in turn will contribute to a better understanding of gene function and regulation in complex biological systems. Furthermore, it also lends itself to providing a more concise and relatively discrete representation of enriched biological themes in combined GSFE studies (Figure 1). In recognition of these benefits, several efforts have been initiated to gather and integrate biological data, including pathway information from various biological databases. The DAVID gene functional classification tool employs a heuristic approach to grouping genes into modules based on similarities in the biological annotations [4]. IPAD defines inter-associations between pathways, disease, drugs and organ specificity based on the overlapping gene associations [5]. IntPath examines overlaps between genes, gene pairs and pathway names to integrate pathways within and across various databases [6]. PathwayAPI attempts to standardise the representation of genes and gene-gene relationships across pathways and merges them to infer more fortified pathway representations [1]. Pathway Distiller employs a holistic approach where pathways are consolidated into clusters either based on shared genes, gene ontology associations and protein-protein interactions (PPIs) or based on their associations (enriched and/or non-enriched) with specific gene sets under study [7]. Most of these tools, however, provide a standalone web interface and have not been integrated into a more general data-mining platform. Such a platform is often essential for prioritising genes for further characterisation in drug discovery and other applications.

Here we describe a new approach to integrating pathway data primarily for target prioritisation. While our method for pathway integration is simple and straightforward, the main novelty lies in its tight integration into the TargetMine data warehouse system [8]. We chose to combine data from three pathway repositories, KEGG [9], Reactome [10] and NCI-Nature curated PID [11]. These three are among the largest and most widely used curated pathway repositories and they employ different approaches to curating and compiling pathway information. For instance, the KEGG pathway repository consists of curated reference pathway maps, which are then mapped to genes within different organisms based on orthologous associations. Reactome compiles expert-curated molecular reactions associated with different biological processes, which are assembled into a biomolecular network to form pathways. NCI PID compiles expert-reviewed molecular interaction data from NCI-Nature curated data, BioCarta and Reactome into biomolecular pathways.

We will first show how the various pathways can be clusters based on shared gene content, on the premise that significant overlaps in gene content between the pathways should reflect overall functional congruity between them. This notion will be confirmed by the biological relevance of the integrated pathway clusters using semantic similarities between Gene Ontology (GO) biological process terms [12] (hereafter referred to as GO terms) annotated to the genes within each pathway. We will further demonstrate the usefulness of pathway clustering based on comparative GSFE analysis on diverse gene sets associated with pathogenesis, inflammatory responses and human diseases, gathered from the literature. A dedicated user interface connects the pathway clusters and gene mappings to TargetMine, for target prioritisation and early-stage drug discovery [8].

Results and Discussion

By integrating pathway data from KEGG, Reactome and NCI, we created Integrated Pathway Clusters (IPCs) for three organisms, *Homo sapiens* (human), *Mus musculus* (mouse) and *Rattus norvegicus* (rat) (Figures 1 and 2; Tables S1A, B and C). The human IPCs, consisting of a total of 1748 pathways associated with 8624

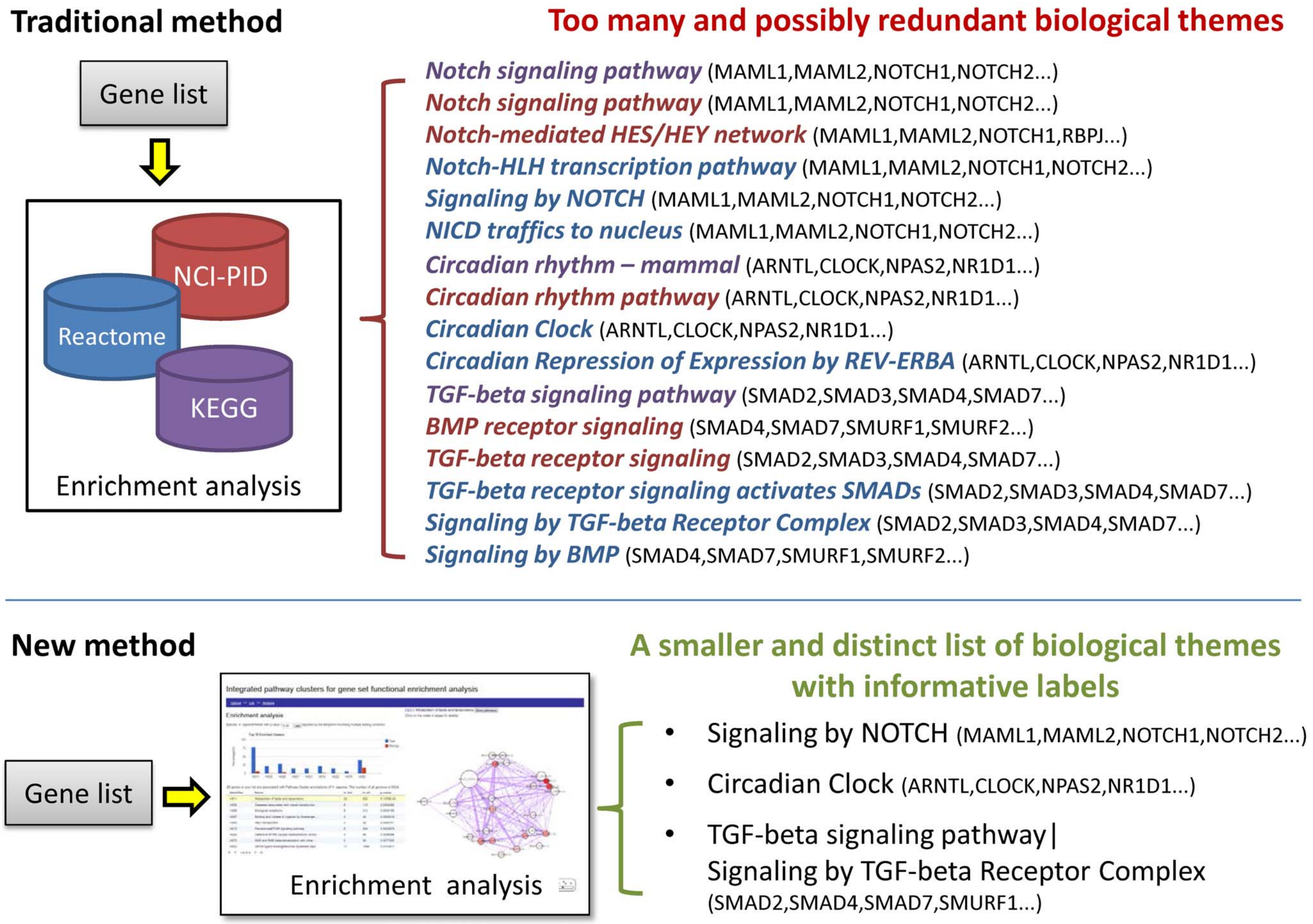
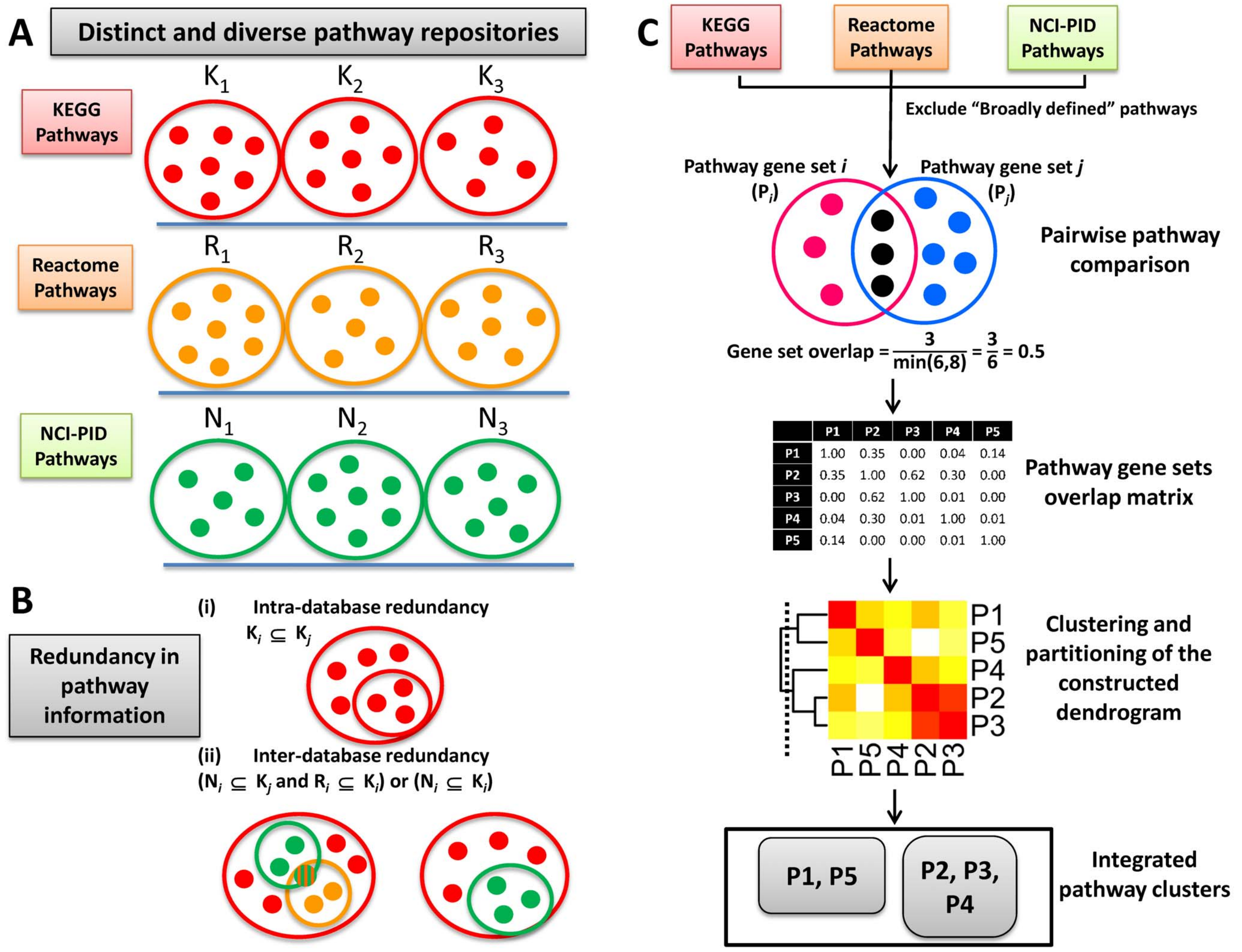

Traditional method
Too many and possibly redundant biological themes
Gene list
NCI-PID
Reactome
KEGG
Enrichment analysis
Notch signaling pathway (MAML1,MAML2,NOTCH1,NOTCH2...)
Notch signaling pathway (MAML1,MAML2,NOTCH1,NOTCH2...)
Notch-mediated HES/HEY network (MAML1,MAML2,NOTCH1,RBPJ...)
Notch-HLH transcription pathway (MAML1,MAML2,NOTCH1,NOTCH2...)
Signaling by NOTCH (MAML1,MAML2,NOTCH1,NOTCH2...)
NICD traffics to nucleus (MAML1,MAML2,NOTCH1,NOTCH2...)
Circadian rhythm – mammal (ARNTL,CLOCK,NPAS2,NR1D1...)
Circadian rhythm pathway (ARNTL,CLOCK,NPAS2,NR1D1...)
Circadian Clock (ARNTL,CLOCK,NPAS2,NR1D1...)
Circadian Repression of Expression by REV-ERBA (ARNTL,CLOCK,NPAS2,NR1D1...)
TGF-beta signaling pathway (SMAD2,SMAD3,SMAD4,SMAD7...)
BMP receptor signaling (SMAD4,SMAD7,SMURF1,SMURF2...)
TGF-beta receptor signaling (SMAD2,SMAD3,SMAD4,SMAD7...)
TGF-beta receptor signaling activates SMADs (SMAD2,SMAD3,SMAD4,SMAD7...)
Signaling by TGF-beta Receptor Complex (SMAD2,SMAD3,SMAD4,SMAD7...)
Signaling by BMP (SMAD4,SMAD7,SMURF1,SMURF2...)
New method
A smaller and distinct list of biological themes with informative labels
Gene list
Integrated pathway clusters for gene set functional enrichment analysis
Enrichment analysis
Enrichment analysis
• Signaling by NOTCH (MAML1,MAML2,NOTCH1,NOTCH2...)
• Circadian Clock (ARNTL,CLOCK,NPAS2,NR1D1...)
• TGF-beta signaling pathway| Signaling by TGF-beta Receptor Complex (SMAD2,SMAD4,SMAD7,SMURF1...)

Figure 1. Benefits of using an integrated pathway repository for GSFE analysis.
doi:10.1371/journal.pone.0099030.g001


A
Distinct and diverse pathway repositories
KEGG Pathways
K_1
K_2
K_3
Reactome Pathways
R_1
R_2
R_3
NCI-PID Pathways
N_1
N_2
N_3
B
Redundancy in pathway information
(i) Intra-database redundancy
$K_i \subseteq K_j$
(ii) Inter-database redundancy
$(N_i \subseteq K_j$ and $R_i \subseteq K_j)$ or $(N_i \subseteq K_j)$
C
KEGG Pathways
Reactome Pathways
NCI-PID Pathways
Exclude "Broadly defined" pathways
Pathway gene set *i* (P_i)
Pathway gene set *j* (P_j)
Pairwise pathway comparison
Gene set overlap = $\frac{3}{\min(6,8)} = \frac{3}{6} = 0.5$
Pathway gene sets overlap matrix
Clustering and partitioning of the constructed dendrogram
P1
P5
P4
P2
P3
Integrated pathway clusters
P1, P5
P2, P3, P4

	P1	P2	P3	P4	P5
P1	1.00	0.35	0.00	0.04	0.14
P2	0.35	1.00	0.62	0.30	0.00
P3	0.00	0.62	1.00	0.01	0.00
P4	0.04	0.30	0.01	1.00	0.01
P5	0.14	0.00	0.00	0.01	1.00

Figure 2. Overcoming the challenges encountered in integrating diverse pathway data. A) Pathway data are compiled in varied repositories, which differ appreciably in scope and content and B) there exists significant redundancy within pathway definitions among different databases. C) An outline of our approach to integrating pathway data from KEGG, Reactome and NCI-PID databases.
doi:10.1371/journal.pone.0099030.g002

genes, included 6224 genes mapped to 253 pathways within KEGG, 6085 genes mapped to 1272 pathways within Reactome and 2573 genes mapped to 223 pathways within NCI PID (Table 1). Below we discuss our observations on the integration of pathway data, their functional coherence and applications to the analysis of sample gene sets. Unless specified, all observations below correspond to the human pathway data.

Pathways within and Across Pathway Databases Share a Large Number of Genes in Common

Gene products may participate in multiple biological processes and pathways. Different pathway databases employ different approaches to compiling pathway information, and therefore may significantly differ in content; however, there remain some redundancies in pathway definitions within and across different databases. Therefore, we first examined gene overlaps between the pathways within each pathway database and across the three pathway databases. The gene overlap index ($OI_{i,j}$) for a pair of pathways was determined by the ratio of the number of genes common to both the pathways to the number of genes within the smaller of the two pathways (see Methods).

A total of 242 pathways within KEGG, 202 pathways within NCI PID and 68 pathways within Reactome were examined for gene overlaps with each other in this manner (excluding pathways that were true subsets of one or more pathways). These included both intra-database (i.e., estimating gene overlaps between two pathways within a single database such as KEGG) and inter-database (i.e., estimating gene overlaps between two pathways from different databases) pathway comparisons. Amongst the intra-database pathway comparisons, we observed that 25 KEGG pathway pairs comprising 35 unique pathways (35 of 242; 14.4%) were remarkably similar (with $OI_{i,j}$ >0.8, i.e., the two pathways having 80% of their genes in common). Likewise, seven pathway pairs comprising eight unique pathways (8 of 202; 4%) and seven pathway pairs comprising 11 unique pathways (11 of 68; 16.1%) with $OI_{i,j}$ >0.8 were observed for NCI PID and Reactome databases, respectively (Table 2; Table S2). Amongst the inter-database pathway comparisons, for $OI_{i,j}$ >0.8, we observed 29 KEGG-Reactome pathway pairs comprising 21 unique KEGG and 19 unique Reactome pathways: 12 KEGG-NCI-PID pathway pairs comprising 12 unique KEGG and seven unique NCI-PID pathways: and nine Reactome-NCI-PID pathway pairs comprising

Table 1. The number of human genes and pathways from different databases, which were consolidated into clusters of related pathways.

	KEGG	Reactome	NCI PID
Pathways	253	1272	223
Genes	6224	6085	2573

doi:10.1371/journal.pone.0099030.t001

nine unique NCI-PID and six unique Reactome pathways (Table 2; Table S2).

The pathway comparisons highlighted significant overlaps between apparently similar pathways within and across pathway databases. For instance, KEGG pathway hsa00970 "Aminoacyl-tRNA biosynthesis" shared a significant number of genes with Reactome pathway REACT_15482 "tRNA Aminoacylation" (with $OI_{i,j} = 0.881$); likewise, Reactome pathway REACT_75790 "Cytokine Signaling in Immune system" shared a significant number of genes with NCI PID pathway il5_pathway "IL5-mediated signaling events" (with $OI_{i,j} = 0.8571$), among other examples. The above comparisons, however, also uncovered remarkable similarities between seemingly unrelated pathways; for instance, KEGG pathways hsa00190 "Oxidative phosphorylation" and hsa04966 "Collecting duct acid secretion" were found to have a significant number of genes in common (with $OI_{i,j} = 0.8519$), suggesting that our approach towards pathway comparisons may provide insights into the possible cross-talks between varied biological processes (Table S2).

These observations suggest considerable overlaps among genes that were mapped to certain biological processes and pathway definitions within and across the three pathway repositories. These overlaps in information offer a useful means of consolidating large amounts of heterogeneous pathway data into a more manageable number of complimentary, broad-based and yet coherent biological themes, which is likely to contribute to a more streamlined functional analysis of genes and gene sets.

Hierarchical Clustering of Pathways Based on Gene Overlap Indices

The gene overlap indices for all pairs of pathways were collated into a matrix, resulting in rows of overlap profiles. Based on these profiles, the pathways were then clustered to produce a dendrogram (Methods and Figure 3A).

Splicing the dendrogram at incrementally relaxed pairwise distance cutoffs generated a series of clusters; using cutoffs of 0.6, 0.65 and 0.7 yielded 105 (multi-member) clusters and 20 singletons, 84 clusters and 14 singletons and 67 clusters and 10 singletons, respectively.

After a visual inspection of the size distribution and the total number of multi-member clusters and singletons, we judged a cutoff of 0.7 to be optimal (likely to produce functionally congruent clusters while keeping the total number of clusters manageable) (Figure 3A). The pathway clusters thus generated varied in size from two pathways in clusters such as no18, no24 and no26 to 187 pathways in cluster no01 (Table S1A). The resulting clusters (hereafter referred to as IPCs) were further evaluated using a series of qualitative and quantitative measures to assess their functional congruency and biological relevance.

Pathways within a Cluster Share a Higher Fraction of Genes than Those from Different Clusters

To investigate whether the gene overlap-based distance metric resulted in well separable clusters, we assessed the overall $OI_{i,j}$ of pathways within and across the IPCs and compared the results with those from randomly generated pathway clusters.

The average intra-cluster $OI_{i,j}$ (0.175) was much higher than the average inter-cluster $OI_{i,j}$ (0.022). The former value was significantly higher than the corresponding value from the randomised dendrograms (0.045±0.003) with a *p*-value of <0.01, as this value was greater than the maximum (0.053) from 100 simulation runs (see Methods).

The above observations suggest that our approach groups together pathways, which have a high fraction of genes in common and are therefore likely to be functionally related.

Manual Inspection Revealed Selected Pathway Clusters Consisting of Functionally Related Pathways

A manual inspection of the pathway names within selected IPCs suggested that functionally similar pathways were grouped into clusters using our approach. For instance, cluster no27 "Metabolism of lipids and lipoproteins" included eight pathways (seven KEGG pathways and one Reactome pathway), all of which were associated with lipid metabolism (Figure 3B). Likewise, cluster no15 "Glycolysis/Gluconeogenesis | Lysine degradation | Valine, leucine and isoleucine degradation" included 10 pathways, most of which were associated with amino acid metabolism (Figure 3C). These observations suggest that pathway clusters generated by our approach are likely to include functionally related pathways and thereby likely to be biologically meaningful.

Validation of the Functional and Biological Relevance of the Constructed Pathway Clusters

We further performed a series of quantitative assessments to examine whether the IPCs consisted of functionally related pathways and were biologically coherent and suitable for gene set analysis and target prioritisation. Below, we individually describe our observations on these evaluations.

Benchmarking pathway clusters against reference (KEGG pathway) sub-types. We employed *purity* and *edit distance* measures [13] to assess how well the KEGG pathways belonging to a particular reference sub-type (defined as the functional "categories" and "sub-categories" defined in the KEGG pathway

Table 2. Pathway pairs within (intra-database) and across (intra-database) pathway datasets, with $OI_{i,j} > 0.8$.

	Intra-database			**Inter-database**		
	KEGG	**Reactome**	**NCI PID**	**KEGG-Reactome**	**KEGG-NCI PID**	**Reactome-NCI**
Pathway pairs	25	7	4	29	12	9
Unique pathways	35	11	8	21 KEGG, 19 Reactome	12 KEGG, 7 NCI PID	6 Reactome, 9 NCI PID

doi:10.1371/journal.pone.0099030.t002

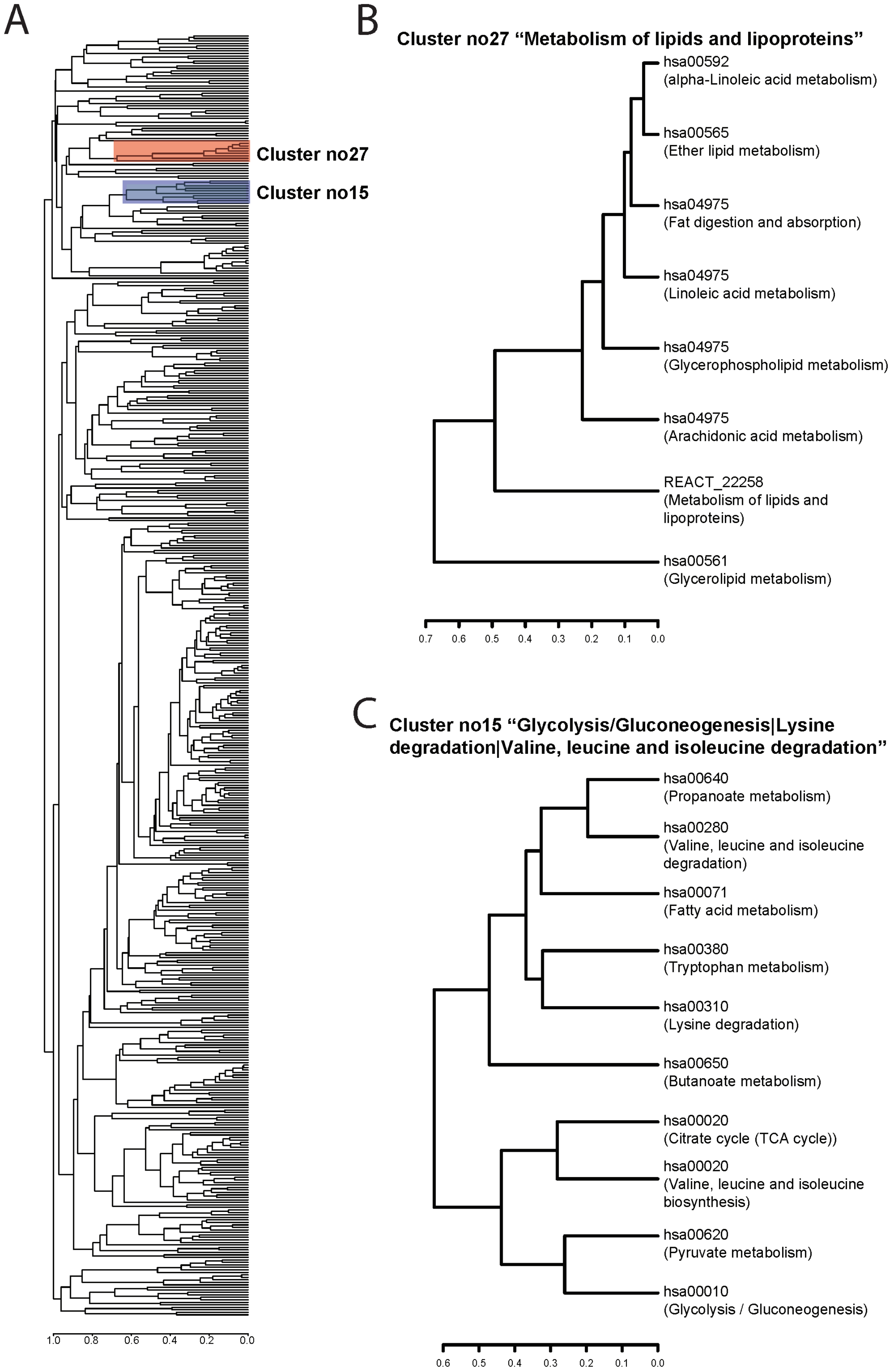

A
Cluster no27
Cluster no15
1.0
0.8
0.6
0.4
0.2
0.0
B
Cluster no27 "Metabolism of lipids and lipoproteins"
hsa00592
(alpha-Linoleic acid metabolism)
hsa00565
(Ether lipid metabolism)
hsa04975
(Fat digestion and absorption)
hsa04975
(Linoleic acid metabolism)
hsa04975
(Glycerophospholipid metabolism)
hsa04975
(Arachidonic acid metabolism)
REACT_22258
(Metabolism of lipids and lipoproteins)
hsa00561
(Glycerolipid metabolism)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
C
Cluster no15 "Glycolysis/Gluconeogenesis|Lysine degradation|Valine, leucine and isoleucine degradation"
hsa00640
(Propanoate metabolism)
hsa00280
(Valine, leucine and isoleucine degradation)
hsa00071
(Fatty acid metabolism)
hsa00380
(Tryptophan metabolism)
hsa00310
(Lysine degradation)
hsa00650
(Butanoate metabolism)
hsa00020
(Citrate cycle (TCA cycle))
hsa00020
(Valine, leucine and isoleucine biosynthesis)
hsa00620
(Pyruvate metabolism)
hsa00010
(Glycolysis / Gluconeogenesis)
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Figure 3. Hierarchical clusters of functionally related pathways based on gene overlap profiles. A) Dendrogram generated by using a matrix of gene overlap indices for all pairwise pathway comparisons. Specific pathway clusters no27 (red) and no15 (blue) are highlighted. B) Cluster no27 "Metabolism of lipids and lipoproteins" included eight pathways, all associated with lipid metabolism. C) Cluster no15 "Glycolysis/ Gluconeogenesis | Lysine degradation | Valine, leucine and isoleucine degradation" consisted of ten pathways, most of which were associated with amino acid metabolism.
doi:10.1371/journal.pone.0099030.g003

database) were clustered together in non-singleton IPCs (see Methods for more details).

The *purity* scores for the IPCs when compared with the KEGG pathway categories and sub-categories were 0.48 and 0.22, respectively; these values were significantly higher than the average of the *purity* scores computed for randomised dendrograms (0.09 and 0.01, respectively) (Table 3) and were even higher than the maximum values (0.31 and 0.09, respectively) from 100 simulation runs; yielding a *p*-value of <0.01.

Likewise, the collective *edit distance* scores for all the IPCs were 93 and 137, respectively, which were much lower than the average of the collective *edit distance* scores computed for randomised dendrograms (158 and 249, respectively) (Table 3) and were even lower than the minimum values (133 and 233, respectively) from 100 simulation runs; yielding a *p*-value of <0.01.

The above observations suggest that IPCs described above correspond more closely to the reference sub-types as defined in the KEGG pathway database than randomised clusters and are thus, likely to represent biologically meaningful themes for functional annotation.

GO term semantic similarity-based evaluation of pathway clusters. Gene ontology (GO) annotations are one of the most useful and widely used means to estimate functional similarity between gene products. Semantic similarity is an approach to estimating the similarity or likeness between two terms of a given ontology (such as GO) [14]. In this study, the semantic similarity measure by Pesquita *et al.* [15] was extended to assess the functional similarity between a pair of (non-identical) pathways within an IPC or those found in different IPCs, thereby estimating the functional coherence of IPCs.

We first performed all-against-all pairwise pathway comparisons based on the GO term semantic similarities (GOSS) between their constituent genes (see Methods) and then we examined the overall functional similarity scores (*FS*) within and across IPCs (intra- and inter-cluster *FS*, respectively).

The median *FS* within an IPC was significantly higher than the median *FS* across IPCs (0.47 and 0.32, respectively; $p = 2.2\times10^{-16}$ by the two-sided Mann-Whitney-Wilcoxon test, $W = 114426.5$) (Figure 4), thereby suggesting that the pathways within a cluster were functionally more closely related than the pathways in different clusters. We further compared *FS* observed within and across IPCs with those observed within randomised dendrograms. The median *FS* within an IPC was much higher than the average of median *FS* within a pathway cluster in the randomised dendrograms (0.47 and 0.33, respectively). Furthermore, the average *FS* within and across the clusters, 0.32 and 0.34, respectively, were statistically indistinguishable within randomised dendrograms.

Taken together, our observations suggested that IPCs were comprised of pathways, which shared an overall higher functional similarity with each other than with pathways from different clusters. Therefore, the pathway clusters were biologically meaningful and likely to represent coherent biological themes.

Gene set functional enrichment analysis. To assess the effectiveness of the IPCs in target prioritisation, we performed GSFE analysis on different sets of genes, which were known to be associated with hepatitis C virus (HCV) pathogenesis [16,17], lung tumourigenesis in mice [18] and non-immune human diseases [19]. Below, we discuss the three case studies involving GFSE analysis using IPCs.

Case study I: Hepatitis C virus (HCV) pathogenesis. We examined four gene sets associated with HCV pathogenesis for enriched IPC associations. These included three gene sets comprising PPI networks constructed from differentially abundant proteins in transgenic mouse models of HCV pathogenesis (CoreTGvsWT, PA28$\gamma^{-/-}$CoreTGvsWT and PA28$\gamma^{-/-}$CoreTGvsCoreTG, respectively; see [16] for details) and a fourth gene set (NS5A infection network), which comprises genes associated with the cellular networks involved in interactions between HCV NS5A protein and human host factors [17]. Functional analysis of the CoreTGvsWT, PA28$\gamma^{-/-}$CoreTGvsWT and PA28$\gamma^{-/-}$CoreTGvsCoreTG gene sets highlighted enriched associations with 28, 28 and 29, IPCs respectively. These figures were much lower than the number of individually enriched KEGG (91, 99 and 98) Reactome (416, 372 and 433) and NCI PID (134, 120 and 113) pathways associated with the above gene sets (Table S3A).

Among specific examples, the IPC no64 "SNARE interactions in vesicular transport" was highly enriched in all of the first three gene sets ($p = 0.01$, $p = 6.25\times10^{-5}$ and $p = 2.86\times10^{-4}$, respectively) (Table S3A). In our previous analysis [16], after inspecting a much larger number of enriched pathways (as shown above),

Table 3. *Purity* and *Edit distance* scores for the IPCs (*PD* = 0.7) when benchmarked against the KEGG pathway sub-types either at the top level (Main class) or the second level (Sub class) were much higher than those of the randomised dendrograms.

PD	*Purity*		*Edit distance*	
	Main class	**Sub class**	**Main class**	**Sub class**
0.7	0.48	0.22	93	137
Randomised dendrograms	0.09	0.01	158	249
0.6	0.54	0.28	123	153
0.65	0.52	0.29	109	149
0.75	0.48	0.24	88	134

Purity and *edit distance* scores at different *PD* cutoffs (0.6, 0.65 and 0.75) are included for comparison.
doi:10.1371/journal.pone.0099030.t003

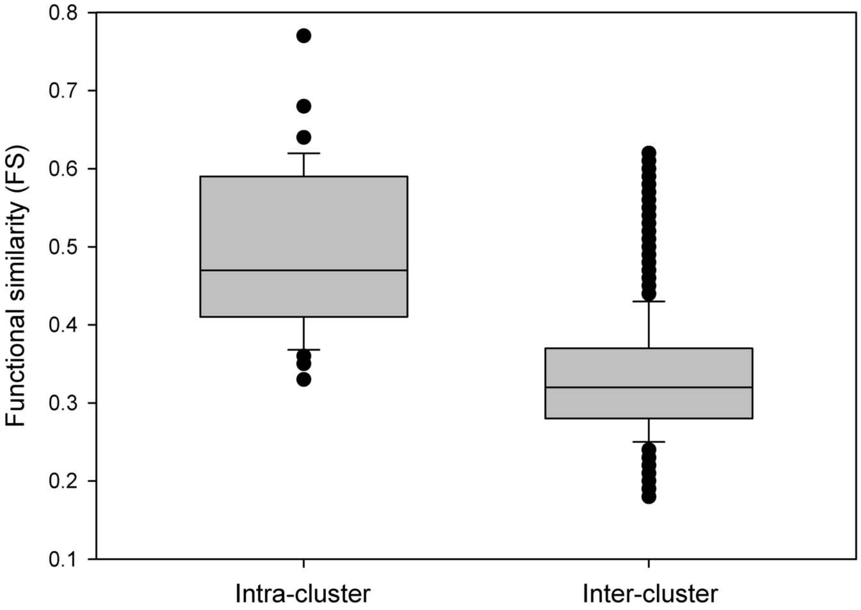


Figure 4. Functional similarity (*FS*) scores within a pathway cluster (Intra-cluster) were much higher than scores across pathway clusters (Inter-cluster).
doi:10.1371/journal.pone.0099030.g004

including those with relatively weak *p*-values, we demonstrated experimentally the involvement of vesicular transport proteins in HCV lifecycle. This finding would have been achieved more easily with the IPC analysis.

The NS5A infection network was associated with 25 enriched IPCs; this figure was much lower than the number of individually enriched KEGG (98), Reactome (488) and NCI PID (119) pathways associated with the NS5A infection network. The enriched IPCs included no23 "Endocytosis | Tight Junction" ($p = 6.84\times10^{-18}$) (Table S3A). Cluster no23 includes genes and pathways associated with cell adhesion and communication and cellular transport, some components of which had been strongly implicated in facilitating HCV lifecycle and tumourigenesis in HCV-induced hepatocellular carcinoma (HCC) [17]. In general, the enriched IPCs included all the biological themes that we had identified previously from a much larger list of relevant pathways and subsequently validated experimentally.

In some enriched IPCs, genes in the original gene set were mapped to two or more pathways, which were not enriched individually. For instance, within the PA28$\gamma^{-/-}$CoreTGvsCoreTG network, two autophagy associated factors GABARAPL1 and GABARAPL2 were mapped to enriched IPCs no011 "Cytokine Signaling in Immune system| Cytokine-cytokine receptor interaction | Herpes simplex infection | Tuberculosis" and no012 "GPCR ligand binding | Neuronal System| Neuroactive ligand-receptor interaction" ($p = 2.61\times10^{-28}$ and $p = 2.18\times10^{-8}$, respectively). These two genes would not have been identified by the standard pathway analysis, because they were mapped to two KEGG pathways hsa04140 "Regulation of autophagy" and hsa04727 "GABAergic synapse", which were components of no011 and no012, respectively and neither of which showed significant association with the original gene set ($p = 0.7395$ and $p = 0.0867$, respectively) (Table S3B). Recent studies have implicated autophagy response to HCV-induced endoplasmic reticulum stress in impairing Type I interferon production in HCV infection [20]. Therefore the analysis using IPCs was able to identify a novel biological theme not identifiable by previous methods.

Case study II: Lung tumourigenesis. We also performed a functional analysis of genes involved in the function of transcription factor Stat3 in carcinogen-induced lung tumourigenesis in mice [18]. Two gene sets examined (Stat3-upreg and Stat3-downreg, respectively) corresponded to PPI networks constructed from differentially expressed genes in Stat3 knockout mice. The Stat3-upreg and Stat3-downreg gene sets were associated with seven and six enriched IPCs, respectively. Among specific examples, Stat3-upreg was mapped to enriched pathway cluster mmu045 "TGF-beta signalling pathway" ($p = 0.003$) and Stat3-downreg was associated with an enriched pathway cluster mmu021 "Rheumatoid arthiritis" ($p = 0.019$) (Table S4A). Furthermore, within the Stat3-upreg gene set, complement activation-associated factor Cfh was mapped to enriched IPC no005 "Hemostasis | Disease | Adaptive Immune System | Pathways in cancer | HTLV-I infection | MAPK signaling pathway" ($p = 4.73\times10^{-7}$). The above association would not have been identified by the standard pathway analysis, because Cfh was mapped to three Reactome pathways REACT_86987 "Innate Immune System", REACT_144679 "Regulation of Complement cascade" and REACT_103920 "Complement cascade", which were components of no005 but individually, none of the three pathways showed significant association with the original gene set ($p = 0.3206$, $p = 0.3430$ and $p = 0.6079$, respectively) (Table S4B). Our observations appear to be consistent with previous studies, which have shown that the human orthologue of mouse Cfh is associated with the early stages of lung tumourigenesis [21,22]. These results demonstrate the relative ease of identifying enriched biological processes previously shown to play critical roles in Stat3-dependent carcinogen-induced lung tumourigenesis [18] and the ability of our approach to identify a novel biological theme not identifiable by previous methods.

Case study III: Non-immune human diseases. An IPC representing a general biological theme of the immune system (no1; Adaptive Immune System | Hemostasis | Developmental

Biology | Pathways in cancer | Innate Immune System) was enriched in all of the gene sets above. To confirm that this result was not an artefact of the clustering method, we performed a functional analysis of gene sets associated with non-immune human diseases, Atherosclerosis, Hypercholesterolemia and Pancreatitis. Our enrichment analysis revealed an enrichment of five IPCs for each of the three gene sets, respectively. These figures were much lower than the number of individual enriched KEGG (31, 18 and 48) Reactome (51, 23 and 43) and NCI PID (9, 1 and 13) pathways associated with the above gene sets (Table S5).

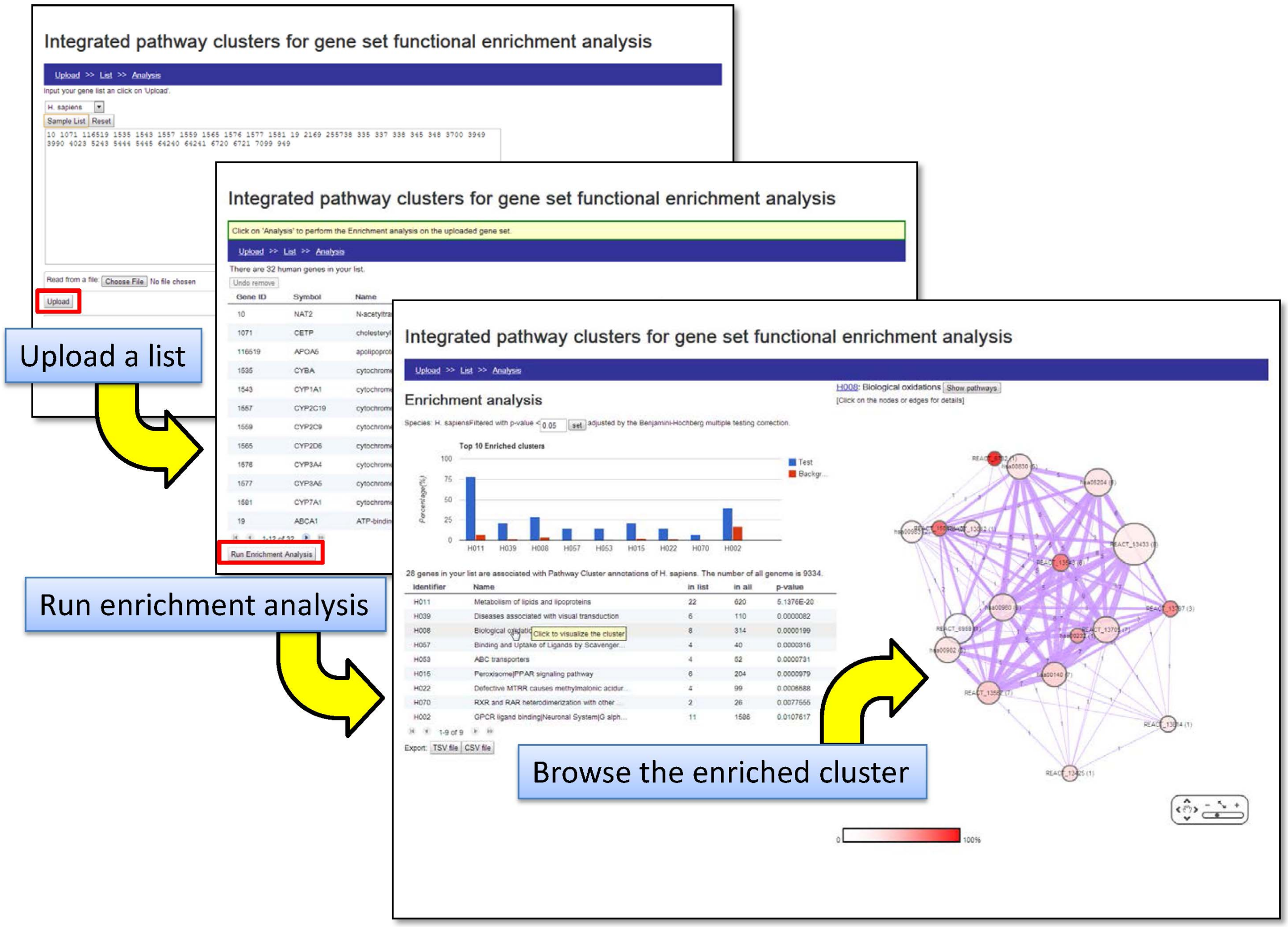
Amongst the most significant associations, IPC no027 "Metabolism of lipids and lipoproteins" was associated with the Hypercholesterolemia gene set ($p = 1.13\times10^{-21}$) (Table S5), which is consistent with the perturbations in lipid metabolism in this disease [23]; enriched IPC no010 "Dilated cardiomyopathy | ECM-receptor interaction | Integrin cell surface interactions" and no027 "Metabolism of lipids and lipoproteins" were associated with the Atherosclerosis gene set ($p = 6.96\times10^{-7}$ and $p = 0.008$, respectively) (Table S5), which is consistent with the pathology of the cardiovascular disease [24]; IPC no014 "Biological oxidations | Metabolism of xenobiotics by cytochrome P450" and no040 "Glutathione metabolism" were associated with the Pancreatitis gene set ($p = 7.66\times10^{-8}$ and $p = 1.58\times10^{-5}$, respectively), which is consistent with the xenobiotic stress and glutathione depletion associated with chronic pancreatitis [25].

The above examples suggest that our IPCs were able to provide a relatively quick, manageable and meaningful approximation of biological themes associated with diverse gene sets.

Data Visualisation and Accessibility

A web interface, tightly connected to TargetMine, was developed for visualising the IPCs and performing GSFE (http://targetmine.nibio.go.jp/pathclust/). It allows a user to upload a list of candidate genes (such as a list of differentially expressed genes, or a set of genes whose protein products interact with a given protein) to TargetMine and create a gene list. The user can then retrieve enriched IPCs and examine their pathway and gene content. Further analysis of these genes and pathways may be performed using TargetMine with its query builder or pre-defined templates.

Each IPC is visualised as a network graph, with the nodes representing the pathways and the edges representing gene overlaps between them. The size of each pathway node reflects the number of genes within that pathway and the thickness of the edges connecting individual pathways reflect the extent of gene overlaps between the connected pathway nodes. A mouse over function allows the user to highlight individual pathways within a cluster; the gene content of each pathway may also be displayed with mouse clicks (Figure 5).



Integrated pathway clusters for gene set functional enrichment analysis

Upload >> List >> Analysis

Input your gene list an click on 'Upload'.

H. sapiens

```
10 1071 116519 1535 1543 1557 1559 1565 1576 1577 1581 19 2169 255738 335 337 338 345 348 3700 3949
3990 4023 5243 5444 5445 64240 64241 6720 6721 7099 949
```

Read from a file: Choose File No file chosen

Upload a list

Integrated pathway clusters for gene set functional enrichment analysis

Click on 'Analysis' to perform the Enrichment analysis on the uploaded gene set.

Upload >> List >> Analysis

There are 32 human genes in your list.

Gene ID	Symbol	Name
10	NAT2	N-acetyltra
1071	CETP	cholesteryl
116519	APOA5	apolipoprot
1535	CYBA	cytochrome
1543	CYP1A1	cytochrome
1557	CYP2C19	cytochrome
1559	CYP2C9	cytochrome
1565	CYP2D6	cytochrome
1576	CYP3A4	cytochrome
1577	CYP3A5	cytochrome
1581	CYP7A1	cytochrome
19	ABCA1	ATP-bindin

1-12 of 32

Run enrichment analysis

Integrated pathway clusters for gene set functional enrichment analysis

Upload >> List >> Analysis

Enrichment analysis

Species: H. sapiensFiltered with p-value < 0.05 set adjusted by the Benjamini-Hochberg multiple testing correction.

28 genes in your list are associated with Pathway Cluster annotations of H. sapiens. The number of all genome is 9334.

Identifier	Name	in list	in all	p-value
H011	Metabolism of lipids and lipoproteins	22	620	5.1376E-20
H039	Diseases associated with visual transduction	6	110	0.0000082
H008	Biological oxidatio Click to visualize the cluster	8	314	0.0000199
H057	Binding and Uptake of Ligands by Scavenger...	4	40	0.0000316
H053	ABC transporters	4	52	0.0000731
H015	Peroxisome|PPAR signaling pathway	6	204	0.0000979
H022	Defective MTRR causes methylmalonic acidur...	4	99	0.0006688
H070	RXR and RAR heterodimerization with other ...	2	26	0.0077555
H002	GPCR ligand binding|Neuronal System|G alph...	11	1586	0.0107617

1-9 of 9

Export: TSV file CSV file

H008: Biological oxidations Show pathways

[Click on the nodes or edges for details]

Browse the enriched cluster

Figure 5. The online user interface allows the users to query and visualise integrated pathway clusters and perform GSFE analysis with the supplied list of genes.

doi:10.1371/journal.pone.0099030.g005

Comparison with Related Resources

The availability of the IPCs within a data warehouse environment makes our approach different from most other integrated pathway repositories such as IntPath [6], IPAD [5], PathwayAPI [1] and Pathway Distiller [7], as well as more general gene function annotation tools such as DAVID [4]. None of these tools provide seamless links with biological data types, other than the integrated biological themes available within these repositories. In contrast, our data model for the IPCs enables the users to link up these functional associations with diverse biological data types stored in TargetMine, such as disease phenotypes, protein structural domains and drug–target associations.

Some of these integrated pathway repositories employ more complex approaches than ours and/or include additional pathway and biological datatypes to infer integrated pathway clusters. Our method is simple and fast and IPCs can be updated automatically. It can also be extended to a larger number of pathway databases or even to other biological data types such as GO annotations.

Among the existing integrated repositories, the Human Pathway Database (HPD) is the closest to our approach in that it integrates pathway data from KEGG, Reactome, NCI PID and BioCarta based on gene/protein overlaps and provides a standalone web interface to query large gene sets for human pathways within a data warehouse [26]. Its data warehouse framework is a less comprehensive system than TargetMine and HPD only considers the extent of gene/protein overlap between pathways to estimate pathway similarity, whereas our approach considers not only gene overlaps but also the similarity of the gene overlap profiles.

hiPathDB adopts a full integration approach where individual pathways are consolidated into a unified derivative superpathway based on shared components. This method provides a holistic and a concise view of biological processes including cross talks between different signalling pathways, but it also results in a loss of information at the molecular level [27]. Our IPCs are designed to complement the existing functional annotations and our data model allows the users to revisit the underlying gene-pathway associations in their original form.

Other differences between our IPCs and the clusters (groups of functionally related genes) defined in the popular DAVID gene functional classification tool include 1) automatically assigned informative names for the IPCs (in contrast to the DAVID clusters with no representative names), and 2) visualisation of IPCs as network graphs to allow the users to examine connections and relationships between constituent pathways.

Conclusions

We describe our approach to integrating pathway information from public repositories based on shared gene content into functionally coherent pathway clusters. The resultant IPCs provided a convenient way to identify broad functional categories relevant to the biological phenomenon under study and thereby enabled swift candidate gene prioritisation. Since our approach relies only on gene overlap between pathways, its inherent flexibility ensures that data from additional pathway repositories (and even non-pathway gene sets) can be readily accommodated to expand the content and coverage of the IPCs.

We assessed the quality of the IPCs using multiple independent measures, including the agreement with the reference sub-types defined in the KEGG database and intra- and inter-cluster semantic similarity scores. With the help of these measures, we established that the IPCs were functionally coherent and biologically meaningful. We further demonstrated the ease of employing the IPCs to analyse large gene sets extracted from the literature.

Our fully automated approach has been integrated into the TargetMine data warehouse and enhanced its ability to investigate complex biological systems for better target discovery. It has also enabled seamless updates of the IPCs synchronised with TargetMine updates, which are scheduled every month in general.

Materials and Methods

Pathway Data

An overview of our approach to overcoming the challenges encountered in integrating diverse pathway data is shown in Figure 2. In the present analysis, pathway associations for the genes within the human, mouse and rat genomes were extracted from KEGG (retrieved on 16/06/2012), Reactome (release date 26/06/2012) and NCI-Nature curated Pathway interaction database (retrieved on 05/04/2012) repositories. The non-IEA (Inferred from Electronic Annotation) GO annotations for the corresponding genes above were retrieved using the TargetMine data warehouse [8]. Some pathways are broadly defined and include many genes (for example, KEGG pathway "Metabolic pathways"). Since these pathways are uninformative for gene prioritisation purposes, we set an arbitrary cut-off of 700 and excluded eight such pathways with more than this number of genes from the subsequent analysis.

Estimating Agreements Across Pathways Based on Gene Composition

Next, we examined the agreement between different pathways within and across pathway repositories based on the overlaps of their gene composition. For each pathway P_i in the dataset (where $i = 1,\ldots,N$ and N is the total number of pathways), let G_i be the set of genes in the pathway. For a pair of pathways, P_i and P_j, $OI_{i,j}$ was defined as

$$OI_{i,j} = \frac{|G_i \cap G_j|}{\min(|G_i|,|G_j|)}$$

where $G_i \bigcap G_j$ is the set of genes shared by G_i and G_j and $|\ldots|$ is the number of genes in the set [28]. An $OI_{i,j}$ of 1 indicates that either the two pathways are identical in size and gene composition or that one pathway is a true subset of the other. To simplify the computation, pathways that were true subsets of larger pathways ("fully contained pathways") were excluded from the subsequent pairwise comparisons (but they were reintroduced into the final clustering results, as will be described later). Likewise, an $OI_{i,j}$ of 0 indicates that the two pathways have no genes in common.

Pathway Clustering Based on Gene Overlap Indices

The collective gene overlap indices for each pathway were then collated to generate corresponding gene overlap profiles. We defined each row o_i of the matrix $OI_{i,j}$ as the gene overlap profile of pathway P_i. The Pearson correlation coefficient R was calculated for each pair of gene overlap profiles (o_i and o_j) and transformed into pairwise distances PD as $PD = 1 - R$. With this distance metric, average-linkage clustering was performed using the hclust function of the R statistical package (www.r-project.org). The dendrogram was partitioned at incremental PD cutoffs and the resulting clusters of related pathways were manually examined to select the most suitable cutoff (see Results).

Once the pathway clusters were established, the fully contained pathways were reintroduced into the clusters that included their "parent" pathways. (If a fully contained pathway had more than one parent and these parent pathways belonged to different clusters, the fully contained pathway was assigned to all these clusters.).

Pathway Cluster Naming

To provide the pathway clusters with informative labels, we examined the gene composition of each pathway cluster and identified the pathways that collectively contributed ≥50% of the genes within a cluster. Their entry names in the original database were then assigned to the corresponding cluster. (In case of two or more pathways contributing ≥50% genes within a cluster, their entry names were concatenated to assign cluster names.).

Assessing the Functional Homogeneity within the Pathway Clusters Based on KEGG Pathway Sub-types

We adapted the *purity* and *edit distance* measures as defined by Brown *et al.* [13] to assess the efficacy of the pathway clustering approach. These scores were used to assess the consistency between our pathway clusters and reference sub-types of pathways as defined in KEGG; KEGG classifies its pathways into "categories" (at the top level such as Metabolism, Cellular Processes and Human Diseases) and "sub-categories" (at the second level such as Energy metabolism, Cell growth and death and Immune system). Only KEGG pathways within a cluster were evaluated in this manner. Clusters containing only one pathway (singletons) were excluded from the following analysis.

In this study, *purity* was defined as the fraction of the constructed pathway clusters that consisted entirely of KEGG pathways belonging to a single reference sub-type ("category" or "sub-category") and were therefore, "functionally homogenous". Here, *purity* reflects the efficacy of our approach in resolving the pathway clusters into functional categories corresponding to KEGG pathway sub-types; a *purity* score of 1 indicates that all KEGG pathways within each pathway cluster were mapped to a single KEGG sub-type, whereas a *purity* score of 0 indicates that none of the pathway clusters were functionally homogenous.

Likewise, in this study, *edit distance* was computed as the minimal number of split and/or merge operations, which were required to transform individual pathway clusters into a KEGG pathway sub-type. For instance, if the pathways corresponding to the KEGG sub-type "Immune system" are distributed across two clusters, each containing other KEGG pathways, two split and one merge operations would be sufficient to transform the two clusters into a single cluster containing all pathways within the KEGG "Immune system" sub-type. The *edit distance* for this process would be 3.

To assess the statistical significance of these measures, 100 randomised dendrograms were generated by shuffling the pathways across the clusters in a manner such that the number of pathway clusters and the number of pathways within a given cluster were preserved. The randomised dendrograms were used to create 100 random sets of pathway clusters.

We defined the *p*-value of the significance of these two observations (*purity* and *edit distance*) using the fraction of the *purity* and *edit distance* scores amongst the randomised conditions that was greater than the actual *purity* and *edit distance* score of the constructed pathway clusters.

Functional Similarities of Pathways and Pathway Clusters

We extended the GOSS [14] defined between a pair of genes to those between a pair of pathways and used this measure to assess the functional similarities within a pathway cluster (intra-cluster coherence) or between pathway clusters (inter-cluster separation). First, the algorithm of Wang *et al.* [29] was employed via an in-house Scala/Java implementation to estimate the GOSS between a pair of genes. This method takes into account both the properties of the annotated GO terms including their parent and child terms and the types of relationships between them (such as "is_a" and "part_of", which are assigned semantic contribution weights of 0.8 and 0.6, respectively).

Next, we defined functional similarity (*FS*) between a pair of pathways, P_i and P_j, as fellows. *FS* may be naturally defined by calculating all possible pairwise GOSS values between genes in G_i and G_j. However, such a measure would simply reflect the amount of overlap between G_i and G_j, which was already taken into account in our clustering algorithm. Since we wished to assess the quality of our pathway clusters based on non-trivial functional similarities between the constituent pathways, we needed to remove contributions from the overlapping genes.

To achieve this goal, in considering a pair G_i and G_j, let $G'_i = G_i - G_j$, i.e., a set of genes in G_i but not in G_j. Similarly, let $G'_j = G_j - G_i$. By adopting a best-match average approach analogous to that of Pesquita *et al.* [15], the best match functional similarity score $S(g_1)$ for each g_1 in G'_i was defined as

$$S(g_1) = \max{}_{g_2}(GOSS(g_1,g_2)), g_2 \in G'_j$$

where the maximum was taken over all g_2 in G'_j. The GOSS for a pair of genes, $GOSS(g_1,g_2)$, was defined by [15] as:

$$GOSS(g_1,g_2) = \frac{Ave_{t_1}(\max_{t_2} SS(t_1,t_2)) + Ave_{t_2}(\max_{t_1} SS(t_1,t_2))}{2}$$

where $SS(t_1,t_2)$ is the GO semantic similarity between two terms t_1 and t_2, Ave_{t_1} means taking the average over all the terms t_1 that were assigned to gene g_1 and the maximum was taken over all the terms t_2 that were assigned to gene g_2. In other words, this measure represents the average similarity between each term assigned to g_1 and its most similar term among those assigned to g_2, averaged with its reciprocal to obtain a symmetric score.

Finally, by using $S(g)$ above, the functional similarity for a pathway pair, $FS(P_i,P_j)$, was defined as:

$$FS(P_i,P_j) = \frac{Ave_{g_1}(S(g_1)) + Ave_{g_2}(S(g_2))}{2}, g_1 \in G'_i, g_2 \in G'_j$$

where Ave_{g_1} means taking the average over all g_1 in G'_i.

$FS(P_i,P_j)$ was computed for all pathway pairs within a cluster (intra-cluster) and for all pathway pairs across different clusters (inter-cluster).

To assess the statistical significance of the intra- and inter-cluster *FS* scores, 100 randomised dendrograms were generated (as described in the previous section) and used as controls. The intra- and inter-cluster $FS(P_i,P_j)$ was computed for each randomised dendrogram and these collective scores were then compared with the intra- and inter-cluster $FS(P_i,P_j)$ of the constructed pathway clusters.

Functional Enrichment Analysis of Gene Sets using Pathway Clusters

Functional enrichment analysis was performed on human and mouse gene sets extracted from the literature. These included gene sets associated with HCV pathogenesis [16], lung tumourigenesis in mice [18] and non-immune disease-related gene sets [19]. The above gene sets were mapped to the IPCs and the enrichment of specific functional categories was estimated by performing Fischer's exact test. The inferred *p*-values were further adjusted for multiple test correction to control the false discovery rate using the Benjamini and Hochberg procedure [30,31] and the annotations/pathways were considered significant if the adjusted $p \leq 0.05$.

Visualisation and Web Interface

The visual representation of the pathway clusters was implemented with JavaScript libraries including jQuery and Cytoscape Web.

Supporting Information

Table S1 Integrated pathway clusters A) Human. B) Mouse. C) Rat.
(XLSX)

Table S2 Pathway pairs which share $OI_{i,j} \geq 0.8$.
(XLSX)

Table S3 A) Enriched IPC associations for the HCV pathogenesis-associated datasets. B) Enriched IPCs associated with genes within the HCV pathogenesis-associated datasets, which were mapped to one or more non-enriched pathways.
(XLSX)

Table S4 A) Enriched IPC associations for the Lung tumourigenesis-associated datasets. B) Enriched IPCs associated with genes within the Lung tumourigenesis-associated datasets, which were mapped to one or more non-enriched pathways.
(XLSX)

Table S5 Enriched IPCs associated with genes within the Non-immune human diseases-associated datasets, which were mapped to one or more non-enriched pathways.
(XLSX)

Acknowledgments

The authors thank the members of the Mizuguchi laboratory for their critical review of the study and the manuscript.

Author Contributions

Conceived and designed the experiments: KM Y-AC LPT BHD SA. Performed the experiments: Y-AC LPT. Analyzed the data: Y-AC LPT SA KM. Contributed reagents/materials/analysis tools: Y-AC LPT BHD JN. Wrote the paper: Y-AC LPT KM.

References

1. Soh D, Dong D, Guo Y, Wong L (2010) Consistency, comprehensiveness, and compatibility of pathway databases. BMC Bioinformatics 11: 449.
2. Stobbe MD, Houten SM, Jansen GA, van Kampen AH, Moerland PD (2011) Critical assessment of human metabolic pathway databases: a stepping stone for future integration. BMC Syst Biol 5: 165.
3. Stobbe MD, Jansen GA, Moerland PD, van Kampen AH (2012) Knowledge representation in metabolic pathway databases. Brief Bioinform.
4. Huang da W, Sherman BT, Tan Q, Collins JR, Alvord WG, et al. (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 8: R183.
5. Zhang F, Drabier R (2012) IPAD: the Integrated Pathway Analysis Database for Systematic Enrichment Analysis. BMC Bioinformatics 13 Suppl 15: S7.
6. Zhou H, Jin J, Zhang H, Yi B, Wozniak M, et al. (2012) IntPath–an integrated pathway gene relationship database for model organisms and important pathogens. BMC Syst Biol 6 Suppl 2: S2.
7. Doderer MS, Anguiano Z, Suresh U, Dashnamoorthy R, Bishop AJ, et al. (2012) Pathway Distiller - multisource biological pathway consolidation. BMC Genomics 13 Suppl 6: S18.
8. Chen YA, Tripathi LP, Mizuguchi K (2011) TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery. PLoS One 6: e17844.
9. Aoki-Kinoshita KF, Kanehisa M (2007) Gene annotation and pathway mapping in KEGG. Methods Mol Biol 396: 71–91.
10. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37: D619–622.
11. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, et al. (2009) PID: the Pathway Interaction Database. Nucleic Acids Res 37: D674–679.
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
13. Brown DP, Krishnamurthy N, Sjolander K (2007) Automated protein subfamily identification and classification. PLoS Comput Biol 3: e160.
14. Pesquita C, Faria D, Falcao AO, Lord P, Couto FM (2009) Semantic similarity in biomedical ontologies. PLoS Comput Biol 5: e1000443.
15. Pesquita C, Faria D, Bastos H, Ferreira AE, Falcao AO, et al. (2008) Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 9 Suppl 5: S4.
16. Tripathi LP, Kambara H, Moriishi K, Morita E, Abe T, et al. (2012) Proteomic Analysis of Hepatitis C Virus (HCV) Core Protein Transfection and Host Regulator PA28gamma Knockout in HCV Pathogenesis: A Network-Based Study. J Proteome Res 11: 3664–3679.
17. Tripathi LP, Kambara H, Chen YA, Nishimura Y, Moriishi K, et al. (2013) Understanding the Biological Context of NS5A-Host Interactions in HCV Infection: A Network-Based Approach. J Proteome Res 12: 2537–2551.
18. Ihara S, Kida H, Arase H, Tripathi LP, Chen YA, et al. (2012) Inhibitory roles of signal transducer and activator of transcription 3 in antitumor immunity during carcinogen-induced lung tumorigenesis. Cancer Res 72: 2990–2999.
19. Chen J, Xu H, Aronow BJ, Jegga AG (2007) Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinformatics 8: 392.
20. Chandra PK, Bao L, Song K, Aboulnasr FM, Baker DP, et al. (2014) HCV Infection Selectively Impairs Type I but Not Type III IFN Signaling. Am J Pathol 184: 214–229.
21. Amornsiripanitch N, Hong S, Campa MJ, Frank MM, Gottlin EB, et al. (2010) Complement factor H autoantibodies are associated with early stage NSCLC. Clin Cancer Res 16: 3226–3231.
22. Cui T, Chen Y, Knosel T, Yang L, Zoller K, et al. (2011) Human complement factor H is a novel diagnostic marker for lung adenocarcinoma. Int J Oncol 39: 161–168.
23. Watts GF, Juniper A, van Bockxmeer F, Ademi Z, Liew D, et al. (2012) Familial hypercholesterolaemia: a review with emphasis on evidence for treatment, new models of care and health economic evaluations. Int J Evid Based Healthc 10: 211–221.
24. Jiang XC, Goldberg IJ, Park TS (2011) Sphingolipids and cardiovascular diseases: lipoprotein metabolism, atherosclerosis and cardiomyopathy. Adv Exp Med Biol 721: 19–39.
25. Wallig MA (1998) Xenobiotic metabolism, oxidant stress and chronic pancreatitis. Focus on glutathione. Digestion 59 Suppl 4: 13–24.
26. Chowbina SR, Wu X, Zhang F, Li PM, Pandey R, et al. (2009) HPD: an online integrated human pathway database enabling systems biology studies. BMC Bioinformatics 10 Suppl 11: S5.
27. Yu N, Seo J, Rho K, Jang Y, Park J, et al. (2012) hiPathDB: a human-integrated pathway database with facile visualization. Nucleic Acids Res 40: D797–802.
28. Camargo LM (2011) Use of Computational Methods and Protein-Protein Interactions to Understand the Aetiology of Neurological Disorders: University of Cambridge.
29. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics 23: 1274–1281.
30. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate- A practical and powerful approach to multiple testing. J R Statist Soc B 57: 289–300.
31. Noble WS (2009) How does multiple testing correction work? Nat Biotechnol 27: 1135–1137.