Establishing Consensus Annotation for the Hallmarks of Cancer

Yi Chen (y.chen@liacs.leidenuniv.nl)
Leiden Institute of Advanced Computer Science(LIACS)
Fons Verbeek
Leiden Institute of Advanced Computer Science(LIACS)
Katherine Wolstencroft
Leiden Institute of Advanced Computer Science(LIACS)

Research Article

Keywords: Gene Ontology, the hallmarks of cancer, semantic similarity, Co-expression network

Posted Date: December 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-119639/v1

Version of Record: A version of this preprint was published at BMC Bioinformatics on April 6th, 2021. See the published version at https://doi.org/10.1186/s12859-021-04105-8.

RESEARCH

Establishing Consensus Annotation for the Hallmarks of Cancer

Yi Chen*, Fons. J. Verbeek and Katherine Wolstencroft

*Correspondence:
y.chen@liacs.leidenuniv.nl
The Leiden Institute of Advanced Computer Science(LIACS), Snellius Gebouw, Niels Bohrweg 1, Leiden, Netherland
Full list of author information is available at the end of the article

Abstract

Background: The hallmarks of cancer provide a highly cited and well-used conceptual framework for describing the processes involved in cancer cell development and tumourigenesis. However, methods for translating these high-level hallmarks concepts into data level-links between individual genes and individual cancer hallmarks varies widely between studies. When we examine different strategies for linking and mapping cancer hallmarks in detail, we see significant differences, but also consensus.

Results: Here we present the results of a comparison of hallmark mapping schemes from multiple studies. We show where there is consensus knowledge that can help us better understand the core biological processes, pathways and network that are associated with the hallmarks of cancer. We also show where the largest differences between mapping schemes occur, and which differences represent changes in our understanding of cancer, changes in our understanding of biological processes in the non-disease state, or the accumulation of more experimental evidence over time.

Conclusions: Mapping strategies rely on intermediate knowledge resources, such as, biological pathway databases, or the Gene Ontology. The structure and annotations of these intermediate resources also change over time. The results of this study therefore highlight the challenges of integrating distributed and changing biological knowledge in bioinformatics.
Contact:y.chen@liacs.leidenuniv.nl

Keywords: Gene Ontology; the hallmarks of cancer; semantic similarity; Co-expression network

Introduction

The hallmarks of cancer, presented initially in 2000 and updated in 2011 [1][2], provides a conceptual framework for describing the process of tumorigenesis. The hallmarks suggest all cancer cells should have 10 essential molecular characteristics: 1) Sustaining Proliferative Signaling, 2) Evading Growth Suppressor, 3) Resisting Cell Death, 4) Enabling Replicative Immortality, 5) Inducing Angiogenesis, 6) Activating Invasion & Metastasis, 7) Genome Instability and Mutation, 8) Tumor Promoting Inflammation, 9) Deregulating Cellular Energetic and 10) Avoiding Immune Destruction. Since the theory was proposed, it has been widely used for interpreting cancer research results, particularly in large-scale, big data studies where whole genome and transcriptome data are compared [3][4][5]. To date, the two Hallmarks of Cancer papers have been cited over 78000 times [6], showing the utility of the hallmarks concepts as a unifying framework for pan cancer analyses. In order to

analyze cancer research results in the context of the hallmarks, the high-level concepts presented as hallmarks need to be interpreted and mapped to a data level of associated biological molecules. As there are currently no hallmarks of cancer knowledge bases, researchers often use existing well-annotated intermediate knowledge resources, such as biological pathway databases like KEGG [7]and MSigDB[8] or the Gene Ontology[9][10], for this task. This is effective for a single point in time, but as our understanding of cancer and biological interactions rapidly change, the structure and annotations of the intermediate resources also change and therefore so do the associations. [11] [12][13].The use of biological ontologies and knowledge bases to help structure, cluster and compare research results is well-established in bioinformatics[14][15]. However, interrelating continuously changing knowledge from multiple sources is a larger challenge. Consequently, despite the widespread use of the hallmarks of cancer, it can be difficult to assess how comparable results and conclusions are between studies. When we examine different strategies for linking and mapping cancer hallmarks in detail, we see significant differences, but also consensus. Here we explore whether the consensus knowledge can help us better understand the core biological processes and pathways that are associated with the hallmarks of cancer, in order to allow for a better comparison between studies. We also explore the differences and whether these represent changes in our understanding of cancer, changes in our understanding of biological processes in the non-disease state, or the accumulation of more experimental evidence over time. In this study, we assess the differences and consensus between 5 different hallmark mapping schemes collected from the literature GO1,[16]GO2,[17] GO3,[18] GO4,[19] and PW1[4]. We compare the Gene Ontology and biological pathway terms selected to represent individual cancer hallmarks, both directly and by analysing their semantic similarity. In addition, we examine the differences between the sets of genes that are annotated with the selected GO terms and biological pathways, which we name 'Hallmark Genes'.

In order to assess the impact of the differences between Hallmark Gene sets, (and further our understanding of the consensus), we compare downstream results by performing network co-expression and enrichment analyses with prognostic cancer genes from the TCGA[20]. If the hallmarks of cancer represent the process of tumorigenesis, genes that show changes in expression that are prognostic for patient survival may have direct involvement in this process as 'drivers'[21], or may be closely associated 'passengers.' Genes that are classified as both prognostic and hallmark genes would be expected to play more important roles in co-expression networks, so the ratios of genes classified as prognostic, hallmark and prognostic-hallmark are compared using network topology and enrichment analyses of network clusters. Finally, we investigate if structural changes to the Gene Ontology hierarchy or changes to pathway or GO annotation could explain the differences between mapping schemes. The results of these analyses provide a consensus of hallmark annotation for each cancer hallmark, providing a common foundation for understanding the hallmark concepts at the data level. In doing so we highlight the challenges of integrating accumulated and distributed biological knowledge over time and the current limitations of semantic similarity measures, which assume a static underlying knowledge structure and the same annotation corpus.

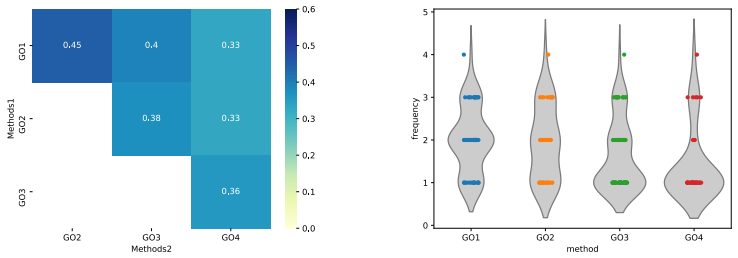
Results

Similarities between Gene Ontology terms

Semantic similarity

To investigate the consensus and divergence between different cancer hallmark mapping schemes, we used the Resnik method (in R GOSemSim[22]) to calculate the semantic similarity between GO terms selected to represent hallmarks. As shown in Figure 1(a), all pairwise semantic similarity scores were less than 0.5, but higher than 0.3, showing low semantic similarity. A manual check of the frequency of selection for each GO term showed that only one GO term, 'Negative Regulation of Cell Cycle' (GO:0045786), was selected by all 4 schemes. However, this term was mapped to different individual cancer hallmarks: in GO1 and GO2, it was mapped to 'Evading Growth Suppressor' while in GO3 and GO4, it was mapped to 'Sustaining Proliferative Signaling'. Similarly, Negative regulation of cell proliferation (GO:0008285), was selected by 3 mapping schemes but was allocated to 'Evading Growth Suppressor' in GO1 and GO2 schemes and allocated to 'Sustaining Proliferative Signaling' in GO3. In general, most GO terms were selected by only one or two mapping schemes. In GO3 and GO4, 57.9% and 77.1% of selected GO terms were unique to that scheme.(Figure 1(b)).

Figures



(a) (b)

Figure 1 (a)Semantic similarity scores of pairwise mapping schemes. Numbers in the heat map represent the Semantic similarity score between Gene Ontology terms selected in different mapping schemes.(b)Frequency of selection of GO terms from different mapping schemes. The Y-axis represents the frequency of selection per GO term and the X-axis represents different mapping schemes. Dots inside the shape represent single GO terms belonging to this mapping methods.

Consensus of GO terms for individual cancer hallmark

After establishing the global similarity between mapping schemes, we further examined consensus of GO term usage for individual hallmarks. For many individual hallmarks, a complete and consistent understanding on the linkage was not apparent. More than 50% of GO terms were selected by only one or two methods and only a few GO terms were simultaneously selected by more than 3 mapping schemes to annotate the same hallmark(Figure 2). For example, for the hallmark 'Tumor Promoting Inflammation', all mapping schemes define GO terms, but none were simultaneously selected by 3 schemes. In contrast, although only twelve different GO terms were selected for the hallmark 'Enabling Replicative Immortality'

across all methods, 3 were simultaneously selected by 3 mapping schemes, showing a greater consensus. For the hallmark 'Evading Growth Suppressor', GO3 and GO4 did not provide GO term mapping, but for GO1 and GO2, 4 of 6 terms were selected by both schemes. Despite the differences for each hallmark, there is also a degree of consensus that identifies a core of shared knowledge.

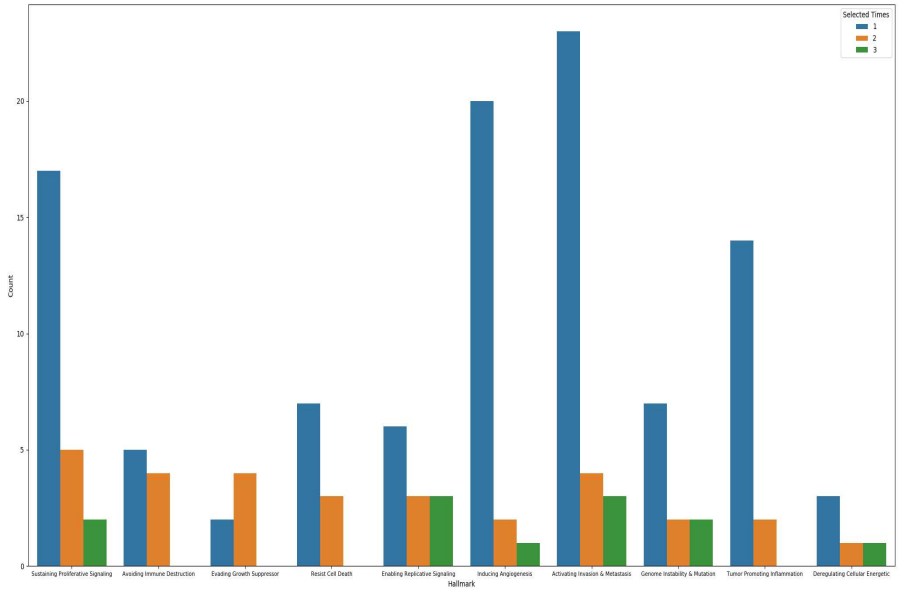


Figure 2 Frequency of selection of GO terms for individual cancer hallmarks. The Y-axis represents the number of GO terms. The X-axis represents individual cancer hallmarks.Bars colored in different color represent how frequently it was selected by mapping methods to annotate this cancer hallmark.

Analysing the divergence and consensus in hallmark gene sets

Four of the five hallmark mapping schemes used the Gene Ontology as an intermediate knowledge resource, but the fifth mapping scheme used biological pathways from KEGG[7] and MSigDB[8]. In order to directly compare annotations between these approaches, we either had to make use of existing mapping between GO and pathway resources, or we had to look at the intersection of the genes annotated with each resource. The former method is problematic due to the difference in granularity of annotation (as reactions in pathways are typically mapped to GO, rather than individual gene product functions in pathways). The latter is advantageous in this particular case because the authors of the pathway mapping method provided a full list of genes which were included in the biological pathways selected at the time of writing and could therefore be used for direct comparison.

The upset plot (Figure 3) shows the intersections and differences between hallmark gene sets. Only 836 Genes were identified as hallmark genes in all cases and there was a marked difference in size between the gene sets. 2171 genes were identified as hallmark genes in PW1, while the smallest gene set generated by GO mapping schemes, was 3078. The other three GO schemes generated sets of more than 9000 genes. This may indicate a greater specificity for PW1 and GO4, but

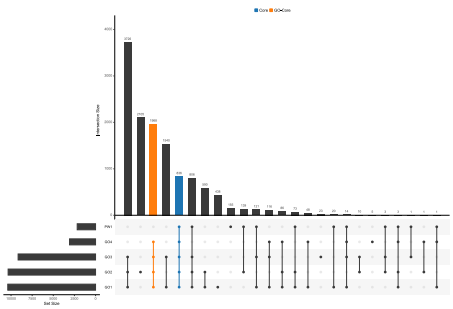


Figure 3 Hallmark Gene Set Comparison. The upset plot shows the number of genes in each hallmark gene set and their intersections. The orange and blue lines represent genes shared by GO mapping schemes and all mapping schemes respectively.

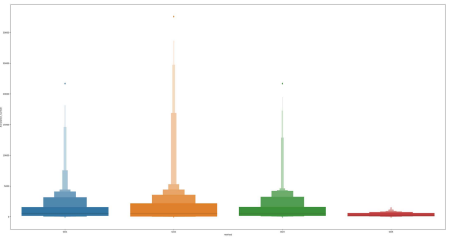


Figure 4 Gene Product Count. The box plot shows the number of gene products annotated to GO terms in different mapping methods.

as they are not subsets of the larger hallmark gene sets, this does not fully explain the discrepancy. The number of annotations for selected GO terms belonging to GO4 were significantly lower than the other 3 schemes (Figure 4), showing a greater specificity in GO term selection. GO1, GO2 and GO3 included GO terms with more than 20000 annotations while the maximal annotations belonging to a single term in GO4 was 1396. The average number of annotations for selected GO terms in GO4 was 347, which was significantly smaller than the others (GO1:1447, GO2:2032, GO:1380). It should be noted that GO3 and GO4 covered only 8/10 hallmarks ('Evading Growth Suppressor' and 'Deregulating Cellular Energetic' were not defined for GO3,'Avoiding Immune Destruction' and 'Evading Growth Suppressor' were not defined for GO4) [19]. For GO1, GO2 and G3, although they were similar in size, each had more than 2000 genes unique to that set, showing a mixed picture of consensus and difference in interpretation.

Individual cancer hallmarks

Figure 5 presents a comparison of the different mapping schemes for each individual cancer hallmark. The dots represent the number of genes belonging to individual cancer hallmarks in different sets and the bars represent the consistent gene count for each hallmark. GO1 and GO3 reach an agreement on the hallmark 'Resisting Cell Death', but large differences can be observed in other hallmarks. For example, for the hallmark 'Sustaining proliferative signalling', the number of genes varies from around 7000 in GO2 to 1336 in GO4, while in the other 2 sets, the numbers are 2925(GO1) and 3735(GO3). Similar results can be observed in other individual hallmarks, showing the different components of gene sets. It was not possible to include the PW1 in this comparison as the authors of the study did not explicitly state which biological pathway related to which cancer hallmark. We can infer this for many of the pathways and observe that multiple hallmarks may be represented in some pathways, but it is an interpretation of results, rather than a reuse of stated results and so was omitted.

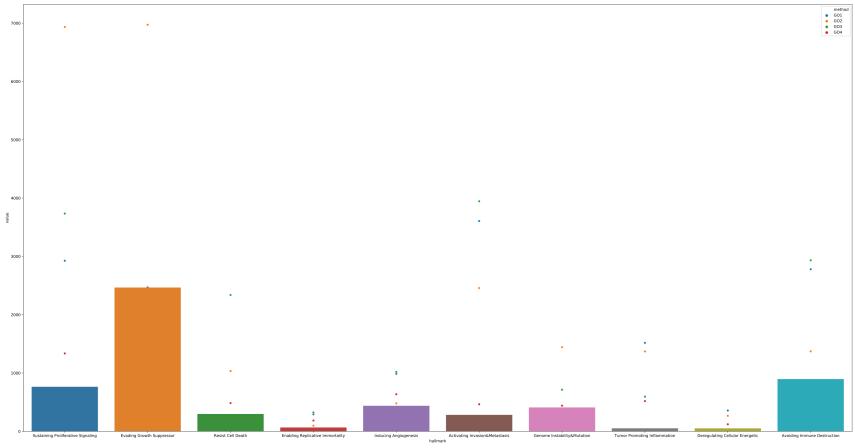


Figure 5 Gene Product count for individual cancer hallmarks. Nodes in the graph represent the count of gene products annotated to GO terms selected in different mapping schemes. Bars in the graph represent the number of gene products which are consistent across all the mapping schemes.

Impact of hallmark mapping strategies on downstream analyses

The results of comparing GO terms and hallmark genes have shown both consensus and difference. We therefore investigated the impact of using different mapping schemes on the results of downstream omics analyses.

Prognostic genes and hallmarks

Survival analysis is often used as an indicator of the importance of genes for specific cancer types. By studying large numbers of individual cases, genes whose expression are prognostic for an unfavourable outcome can be identified. As hallmark genes are involved in the biological activities that promote cancer development, it is expected that prognostic genes would either be hallmark genes themselves, or be co-expressed with hallmark genes [4]. As different mapping schemes would generate different hallmark gene sets, investigating the varying overlap in prognostic-hallmark genes would highlight the consequences of different hallmark definitions. Here we identified the overlap between prognostic and hallmark genes for 17 cancer types using different mapping schemes. The prognostic gene data was taken from PW1 [4]. Prognostic-hallmark genes shared between multiple cancer types were identified, where there were 5 or more shared genes. The impact of selecting different mapping schemes was assessed by pairwise comparisons. The Jaccard Index of prognostic-hallmark gene groups was calculated. Figure 6 shows the results.

Figure 6 shows the similarities between groups of prognostic-hallmark genes in different mapping schemes. As expected, none of the 6 comparisons showed completely consistent results, reflecting the differences between hallmark gene sets. Comparisons between GO4 and the other three schemes were the most dissimilar, with the absence of any genes shared between some cancers (e.g. between Renal Cancer, Stomach Cancer and Endometrial Cancer). GO1 and GO3 were grouped closest together, with the highest Jaccard index scores. Comparisons between GO1 and GO2, GO2 and GO3 produced a moderate scores. The full list of results can be found in additional file 2.

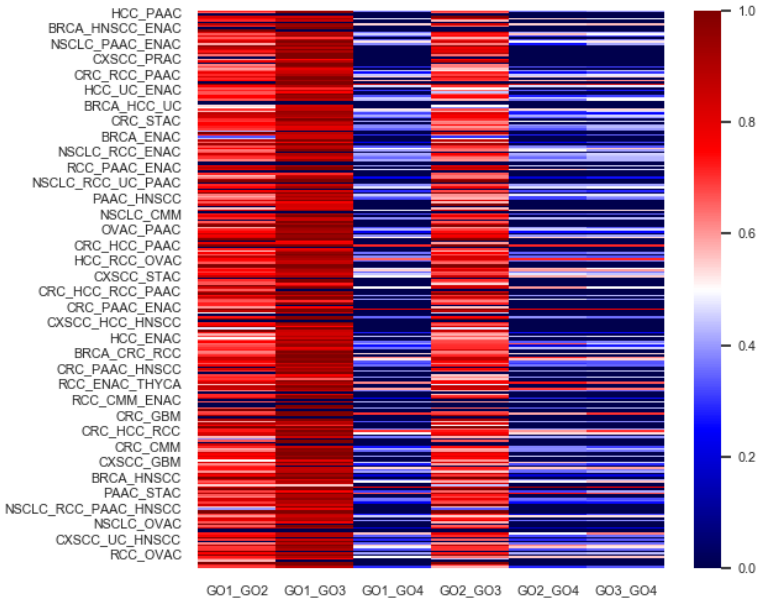


Figure 6 Pairwise comparison of groups of prognostic-hallmark genes identified in the same cancers between mapping schemes. Each block represents a group of prognostic-hallmark genes. The y-axis shows the combination of cancer types in which the corresponding group of genes are classified as prognostic-hallmark genes (abbreviations of cancer names can be found in additional file 1). The X-axis shows pairwise comparison of mapping schemes.The color of each block represents the similarity scores of the same group of genes using the Jaccard index. Darker red represents a higher score and darker blue a lower score.

Cancer hallmarks enrichment and co-expression networks

To further investigate the relationship between prognostic genes and cancer hallmark mapping schemes, we examined the co-expression networks of prognostic and hallmark genes and their enrichment in breast cancer. Based on the breast cancer transcriptome data from the TCGA Genomics Data Commons(GDC) data portal(https://portal.gdc.cancer.gov/), we calculated the Pearson correlation coefficient between pairwise prognostic genes of breast cancer, using the top 25% coefficient of pairwise genes. The co-expression network was constructed based on the coefficient value. By clustering prognostic genes, we investigated which individual hallmarks played a leading role in cancer development with an enrichment analysis using a hypergeometric test[23]. Genes were clustered using a multilevel algorithm and divided into four clusters[24]. Figure 7 shows the enrichment analysis results for GO1 to GO4. Breast cancer was selected for this study because there are more than 1000 breast cancer cases in the TCGA database (the 3rd largest), and it only has 582 prognostic genes. When constructing a co-expression network, a large number of cases can minimize correlation coefficient bias and a relatively small number of prognostic genes can help to minimise the number of prognostic genes that need to be excluded from the study due to an insufficient coefficient value. Figure 7 shows the differences between clusters of enrichment for individual hallmarks when using different mapping schemes. Consistent with previous findings, GO4 showed the most considerable difference compared to the other three mapping schemes, as all four network clusters were only enriched in prognostic genes. GO1 and GO3 were

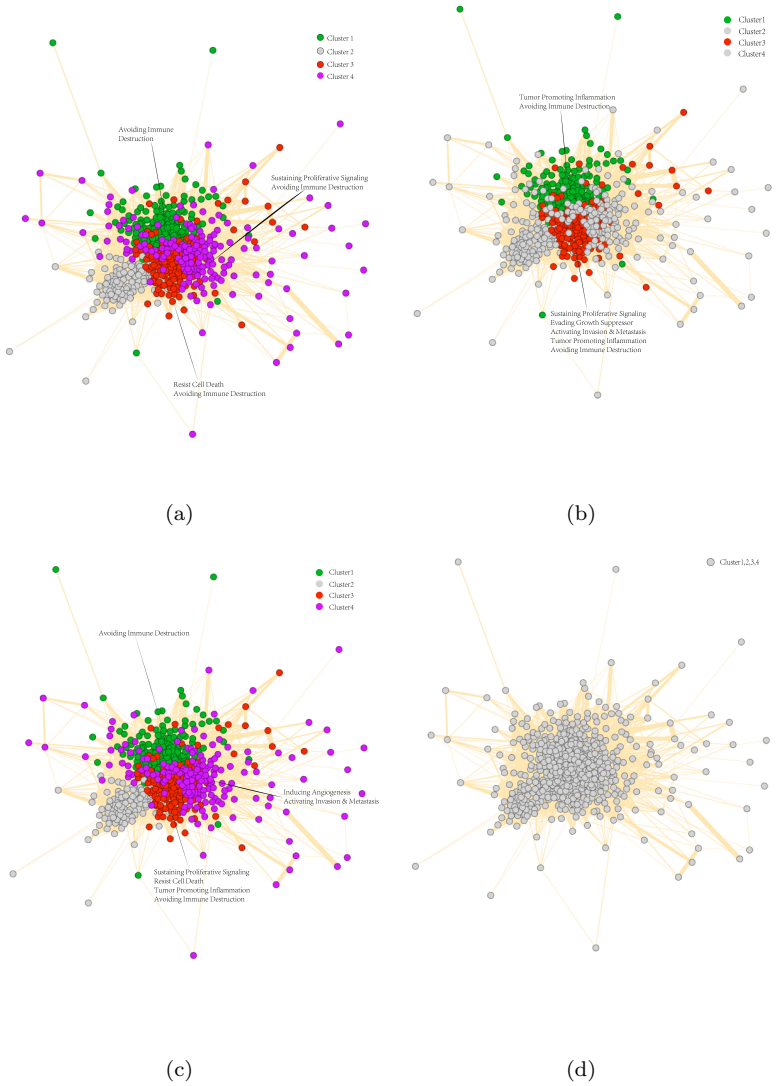


Figure 7 A,b,c,d represent co-expression networks constructed using prognostic genes and hallmark genes from GO1,GO2,GO3 and GO4 method,respectively.. Nodes coloured in grey represent a cluster only enriched in prognostic genes while nodes with different colours represent clusters enriched in both prognostic and hallmark genes. Dashed lines show which cancer hallmarks the clusters are enriched in.

the most similar to each other, but the enrichment in network clusters still showed considerable differences. For example, Cluster 2 was only enriched in the hallmarks 'Resisting Cell Death' and 'Avoiding Immune Destruction' in GO1, but it was additionally enriched in the hallmarks 'Tumor Promoting Inflammation' and 'Sustaining Proliferative Signaling' in GO3. Similar situations can be seen in cluster3. GO2 was more similar to GO1 and GO3 than to GO4, as cluster 1 and cluster 3 were both enriched in prognostic and hallmark genes. However, cluster 2 showed no enrichment in hallmark genes in the GO2 network. The inconsistencies identified in these networks would result in different conclusions when considering the involvement of individual cancer hallmarks in relation to prognostic genes. Combining these find-

ings with our previous results, it is clear that there are large downstream impacts on results as a consequence of using different hallmark mapping schemes.

GO evolution in relation to difference between mapping schemes

The hallmarks mapping schemes under comparison were developed over the period of 7 years and therefore were developed using different versions of the Gene Ontology and associated annotation. Understanding which differences between mapping schemes were the result of topological or annotation changes to GO could therefore help to further refine consensus and therefore make results and conclusions more comparable between studies.

Previous research has shown large changes in GO[12]. From 2004 to 2015, the number of terms in GO increased by 2.5 fold (from 16139 to 40810) and the number of GO terms used for annotations of human genes increased 3.8 fold (from 2972 to 11403). Furthermore, the number of GO annotations for human genes changed from 19615 to 109152, increasing by 6.3 fold. The proportion of protein-coding human genes with at least 1 annotation changed from 32% to 65% and relationships between GO terms were also enriched and changed. 21998 connections became 78078 in the same period and 6833 relationships were removed. Other structural changes, such as, terms becoming obsolete or being merged into other terms also changed the hierarchical structure and information content of GO. According to the Archived data from the Gene Ontology Consortium (http://archive.geneontology.org/full/), 1086 GO terms were made obsolete at the same period. By constructing directed acyclic graphs (DAG) for each mapping scheme, based on selected GO terms and their neighbors, using Gene Ontology archived data at different time points, we evaluated the extent to which GO evolution contributed to the inconsistency between mapping schemes.

Structure differences between mapping schemes

GO2 is derived from a paper published in 2012, GO3 and GO4 were published in 2017, and GO1 was published in 2015. As none of the publications provided information on the version of GO used in their initial analyses, we selected evenly-spaced time points to study GO evolution across the whole period. To construct the directed acyclic graphs (DAGs), we first create the full GO hierarchical graph with the archived relationship data downloaded from Gene Ontology Consortium in Cytoscape[25]. For each mapping scheme, we generated a sub-graph of all GO terms selected and their neighbours, and made pairwise comparisons between each sub-graph. In Figure 8, we compared the GO hierarchy graphs of GO1 and GO2 (GO archive data from 2012 and 2016). Crimson nodes represent GO terms that existed in both time points but were only selected by GO2, while light red nodes represent GO terms which are obsoleted in 2016. Similarly, dark blue nodes represent GO terms that existed in both 2012 and 2016 but were only selected by GO1 and light blue nodes represent GO terms had not been created in 2012. Structural differences are shown in Figure 8, where 237 of 400 nodes are coloured, confirming the inconsistency between mapping schemes. Seven GO terms selected by GO2 were obsolete by 2016, and 20 terms selected by GO1 were not created in 2012. Similar observations were made in the comparison between GO3 and GO4 (additional file

3). These results show that although there were major structural changes to GO over the time period, the majority of terms were available for selection for each mapping scheme. This suggests structural differences were not the main factor and that differences in interpretation of the relationships between GO and the cancer hallmarks played a larger role. It is worth noting that the mapping scheme from GO2 is being used and maintained for further research. Despite all the changes to GO, there have been no major updates to this mapping scheme since the initial publication.

Gene product counts changes at different time points

The number of gene products annotated to GO terms also changes over time. We additionally investigated the number of gene products annotated to selected GO terms at different time points (figure 9). We find a huge increase for most GO terms when comparing their annotation gene product number from 2012 to 2016. Although the number alone cannot properly reflect the exact changes between annotations (as the number of deleted and added annotations to a single term could be the same), it still reflects a general increase in Gene Ontology annotation. Changes in GO annotations increase the number of genes in a hallmark gene set. Mapping schemes that favour less specific GO terms for hallmark definitions would therefore show a larger increase in hallmark gene set size and show a larger impact on downstream analyses.The comparison between Gene product count data in 2016 and 2019 further supports the idea that the Gene Ontology annotation data continuously changed and it still affects the composition of hallmark gene sets when using updated knowledge source.

Consensus between methods

The results of this analysis have shown inconsistencies between different mapping schemes, but they have also shown consensus. From this consensus, we can identify a common understanding of functional cancer hallmark mapping, which could contribute to the creation of a systematic mapping method. The goal is to combine current consensus knowledge and maintain an active integration with GO and pathway resources as they change in the future. To investigate the consensus from all mapping methods, we first combined GO and pathway results by identifying corresponding GO terms for the pathways selected by PW1. For pathways from MSigDB, GO and pathway mapping was extracted directly from MSigDB (Additional file 4). For pathways from KEGG, corresponding GO terms were not provided by KEGG directly. Therefore, corresponding GO terms were derived by identifying GO definitions that were the most similar to the description of KEGG pathways. 14 GO terms were identified to represent the KEGG pathways. After the correspondence between GO and pathways had been established, we used GO terms to describe and define the level of consensus across all five mapping schemes. For each cancer hallmark, GO terms selected by three or more mapping schemes were seen as consensus terms. In total, we identified 42 consensus GO terms across all hallmarks and we identified some degree of consensus for each hallmark (table 1). Figure 10 shows the visualization of the hallmark' Activating Invasion & Metastasis' where there were seven consensus terms. Nodes coloured in orange and yellow represent those selected by

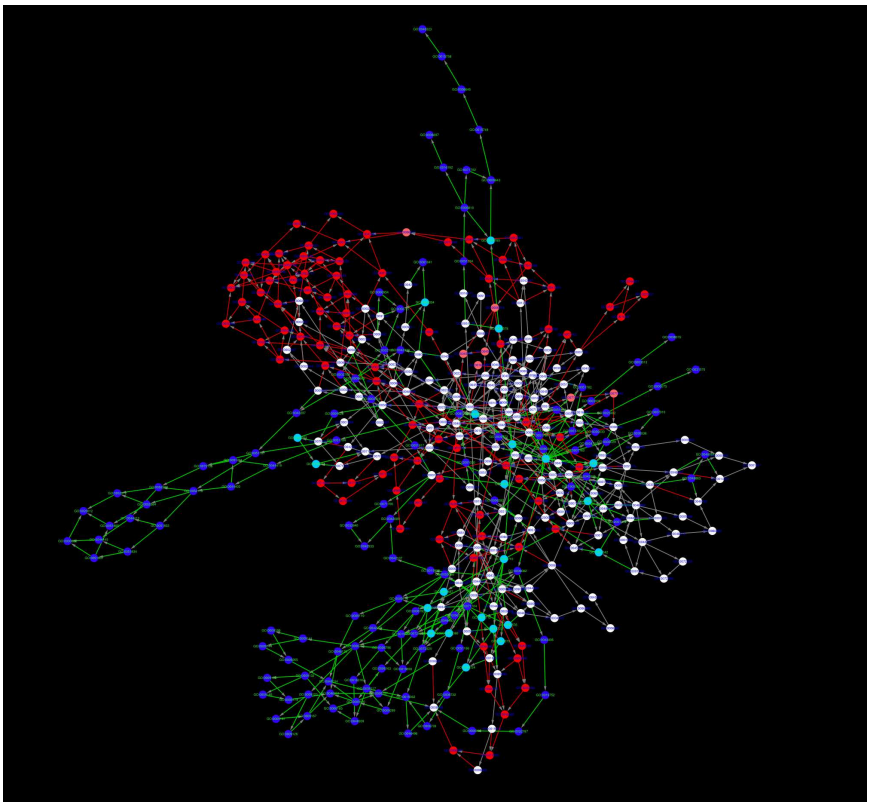


Figure 8 A comparison of the GO Biological Process topology in different years, constructed from all selected GO terms and their first neighbours. Crimson nodes represent GO terms that existed in both time points but were only selected by GO2, while light red nodes represent GO terms which are obsoleted in 2016. Similarly, dark blue nodes represent GO terms that existed in both 2012 and 2016 but were only selected by GO1 and light blue nodes represent GO terms had not been created in 2012.

3 and 4 methods respectively. Similarly, for the hallmark 'Sustaining Proliferative Signaling', there were also seven terms considered as consensus terms. These hallmarks show the most consensus. Other hallmarks show much less agreement. For example, for 'Inducing Angiogenesis', only two terms were consensus terms, while 23 different terms were chosen across all mapping schemes to annotate this hallmark. A similar situation can be seen in 'Tumor Promoting Inflammation' where only two terms were consensus terms out of 16 terms selected across all methods. These results indicate that a refinement is required in the hallmark definitions, in order to establish a better consensus.

Discussion

This study attempts to compare the conceptual similarities between cancer hallmark definitions from multiple time-points, developed for different types of analyses, in order to establish a consensus. Although the hallmarks are in widespread use, the numbers of publications that explicitly describe the association between the hallmark concepts and biological molecules and/or functional annotations are limited. For the publications that describe hallmark mapping in detail, other problems remain. Missing provenance information means that the exact versions of GO

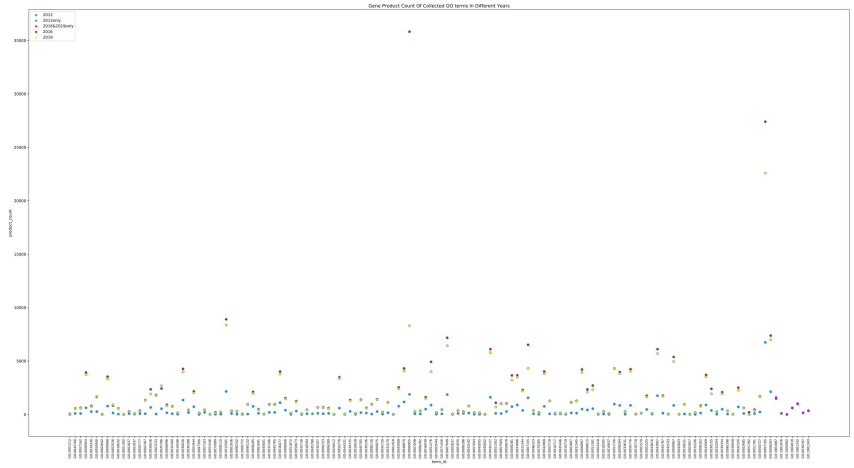


Figure 9 The number of gene product annotations in 2012 and 2016. Blue,red and yellow dots represent GO terms in 2012,2016 and 2019, respectively. green dots represent the GO terms that existed in 2012 but were obsolete in 2016 and 2019 and pink dots represent GO terms that haven't been created in June, 2012 but were created before July, 2016. The x-axis represents selected Gene Ontology terms; the y-axis represents the count of gene products annotated to a GO term and its descendants

Table 1 Consensus for individual cancer hallmarks

Hallmark	GO terms	Term name	Frequency of selection(without pathway)	Frequency of selection
Sustaining Proliferative Signaling	GO:0008283	cell population proliferation	2	3
Sustaining Proliferative Signaling	GO:0007049	cell cycle	2	3
Sustaining Proliferative Signaling	GO:0051301	cell division	2	3
Sustaining Proliferative Signaling	GO:0030307	positive regulation of cell growth	2	3
Sustaining Proliferative Signaling	GO:0045786	negative regulation of cell cycle	2	3
Sustaining Proliferative Signaling	GO:0008284	positive regulation of cell population proliferation	3	4
Sustaining Proliferative Signaling	GO:0045787	positive regulation of cell cycle	3	4
Evading Growth Suppressor	GO:0009968	negative regulation of signal transduction	2	3
Evading Growth Suppressor	GO:0045786	negative regulation of cell cycle	2	3
Evading Growth Suppressor	GO:0008285	negative regulation of cell population proliferation	2	3
Evading Growth Suppressor	GO:0030308	negative regulation of cell growth	2	3
Resist Cell Death	GO:0012501	programmed cell death	2	3
Resist Cell Death	GO:0043067	regulation of programmed cell death	2	3
Resist Cell Death	GO:0043069	negative regulation of programmed cell death	2	3
Enabling Replicative Immortality	GO:0032200	telomere organization	2	2
Enabling Replicative Immortality	GO:0001302	replicative cell aging	2	2
Enabling Replicative Immortality	GO:2000772	regulation of cellular senescence	2	2
Enabling Replicative Immortality	GO:0000723	telomere maintenance	3	3
Enabling Replicative Immortality	GO:0032204	regulation of telomere maintenance	3	3
Enabling Replicative Immortality	GO:0090398	cellular senescence	3	3
Inducing Angiogenesis	GO:0001525	angiogenesis	2	3
Inducing Angiogenesis	GO:0001570	vasculogenesis	2	2
Inducing Angiogenesis	GO:0045766	positive regulation of angiogenesis	3	4
Activating Invasion and Metastasis	GO:0034330	cell junction organization	2	3
Activating Invasion and Metastasis	GO:0030030	cell projection organization	2	3
Activating Invasion and Metastasis	GO:0030155	regulation of cell adhesion	2	3
Activating Invasion and Metastasis	GO:0007162	negative regulation of cell adhesion	2	3
Activating Invasion and Metastasis	GO:0007155	cell adhesion	3	4
Activating Invasion and Metastasis	GO:0016477	cell migration	3	4
Activating Invasion and Metastasis	GO:0001837	epithelial to mesenchymal transition	3	4
Genome Instability and Mutation	GO:0045005	DNA-dependent DNA replication maintenance of fidelity	2	2
Genome Instability and Mutation	GO:0031570	DNA integrity checkpoint	2	3
Genome Instability and Mutation	GO:0006281	DNA repair	3	4
Genome Instability and Mutation	GO:0006282	regulation of DNA repair	3	3
Tumor promoting Inflammation	GO:0002367	cytokine production involved in immune response	2	2
Tumor promoting Inflammation	GO:0050727	regulation of inflammatory response	2	2
Deregulating Cellular Energetic	GO:0071456	cellular response to hypoxia	2	3
Deregulating Cellular Energetic	GO:0006096	glycolytic process	3	3
Avoiding Immune Destruction	GO:0002418	immune response to tumor cell	2	2
Avoiding Immune Destruction	GO:0006955	immune response	2	3
Avoiding Immune Destruction	GO:0002837	regulation of immune response to tumor cell	2	2
Avoiding Immune Destruction	GO:0050776	regulation of immune response	2	3

and pathway resources are not always available and need to be inferred. Mapping definitions for individual hallmarks are missing in a number of cases, and, more generally, comparing functional annotations derived from changing knowledge resources is problematic. For the studies that documented their mapping methods, we found that mapping to GO as an intermediate knowledge resource was the most common approach. The fact that we only identified one example of a pathway-based approach that fitted our inclusion criteria, against four GO approaches, is a limitation for establishing pathway consensus. Nevertheless, both methods are

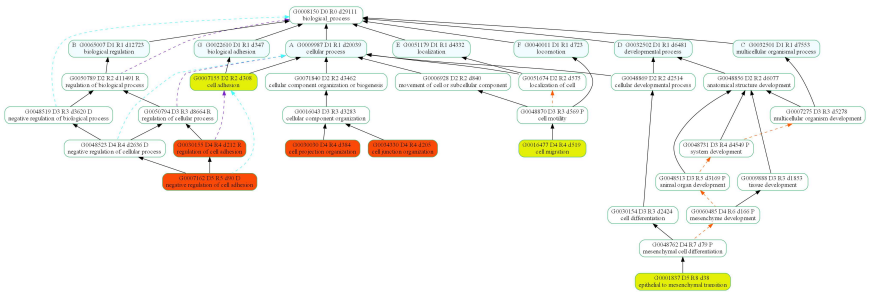


Figure 10 Visualization of the consensus in individual cancer hallmark Activating Invasion and Metastasis(include pathway mapping method). Nodes colored in red and yellow represent GO terms are selected or enriched by genes in 3 and 4 methods, respectively.

being used and should therefore be considered here. The GO-generated hallmark gene sets are much larger than those of the pathway example. Gene Ontology terms selected for mapping typically contained a mixture of terms from different levels in the GO hierarchy. Any terms selected from higher levels have a large number of genes annotated to them, but they also have a low information content and lack specificity. For example, terms such as, Immune Response (GO:0006955) or Cell Cycle (GO:0007049), were selected to define cancer hallmarks. The inclusion of such terms explains the large numbers of genes in GO1 and GO2 in particular. Substituting these general terms with more specific descendant terms could make the GO term set more informative and reduce the overall number of gene products. However, the differences in levels GO terms cannot explain the differences between mapping schemes. Mapping schemes with smaller hallmark gene sets are not defined by descendants of the more general terms selected by mapping schemes with larger hallmark gene sets. This would result in a high semantic similarity when comparing mapping schemes. The results showed that all pairwise semantic similarity scores were less than 0.5, meaning that definitions of at least some hallmarks show a lack of general consensus. A major limitation for analysing the semantic similarity between mapping schemes, however, is that available semantic similarity methods assume the same underlying knowledge structure. In reality, in this case and many others, we are comparing annotations derived from different versions of the knowledge structure. For a fair comparison, we need to know if the same GO terms were available for each different research group to select, or if some were only introduced later or became obsolete between time-points. Our analysis of the GO topological structure revealed that most terms were available through the whole time-period, but that the researchers simply did not select the same or similar. We also identified rearrangements in higher level terms from the biological process hierarchy that meant that relationships between many terms were altered significantly between time points. For example, the term 'death' (GO:0016265) was a descendant of term 'Biological process', and was therefore one of the highest level terms in GO hierarchy. It only had one descendent, 'cell death'(GO:0008219). In March 2016, this term was made obsolete. In our network comparison between GO1 and GO2 using archived GO data from 2012 and 2016, this term existed in the GO2 network but was obsolete in the GO1 network. In addition, the descendant term

'Cell death' was connected to the term 'cellular process'(GO:0009987) in the 2012 GO2 network, while in the 2016 GO1 network, the linkage to 'cellular process' had been removed and substituted with a linkage to the term 'single-organism cellular process(GO:0044763)', which was itself created after 2012. These alterations at high levels in the GO structure affect everything at lower levels and therefore would have a large impact when calculating semantic distance or information content between the two sets of GO terms. Current semantic similarity measures do not take such changes into account. Despite the aforementioned problems with analysing the similarities and differences, there are some clear examples of consensus between the mapping schemes. The core gene set is small in comparison to the whole collection of hallmark genes across all methods, but the core offers important information about where there is shared understanding. Genes in the core gene set are not all annotated with a small number of hallmarks, but spread across all ten hallmarks. This means that there is a partial shared understanding across all mapping schemes. Similarly, when examining the consensus GO terms, there are consensus terms for every hallmark, although the amount of consensus varies. For the hallmark 'Activating Invasion & Metastasis', seven GO terms are considered as consensus terms and three of them are selected by four mapping schemes, which shows a remarkable consensus. In contrast, the hallmark 'Inducing Angiogenesis', only has two consensus terms out of 23 terms selected across all mapping schemes. The consensus for this hallmark is therefore not sufficient to define all biological hallmark activities and should be further refined. Although we were able to establish some consensus, we also observed differences between mapping methods and a lack of consensus in some definitions. Our topological and information content analysis of GO at different time points showed that although the same terms were available, researchers often selected different terms in later studies. This indicates that the largest effect was a difference in understanding of the hallmark concepts. This could represent a change in our collective understanding of cancer from the accumulation of evidence, or it could simply be bias from the research experiences of the individuals involved in the selected studies. A broader community-wide discussion about the hallmarks and their definitions is required to determine this. In addition to the approaches identified here, the COSMIC data resource (Catalogue of Somatic Mutations in Cancer) has undertaken a manual annotation approach to hallmark identification. For each gene in the COSMIC Gene Census, curators are manually extracting evidence for cancer hallmark involvement from the literature. To date, approximately 300 genes have been annotated out of 700. The process is slow and will take a number of years to complete, but COSMIC expects to identify hallmark activity for almost all genes in the census. When we compare the consensus GO terms from this analysis with those enriched in the Gene Census, the majority (31/42) are present and enriched, indicating that the consensus GO terms we have identified represent important consensus knowledge.

The consensus GO terms identified for each cancer hallmark show where there is a shared understanding of the hallmarks of cancer. They could therefore be the foundation for a more systematic approach to mapping cancer hallmarks to data via intermediate knowledge resources. In the longer term, the consensus could be the starting point for a broader community discussion about how to identify hallmark

activity in omics data sets and a community-wide ontology development process. A similar community approach has proven successful for other research communities, with the Gene Ontology itself being a prime example [9]. If we can achieve a shared understanding of the cancer hallmarks and what they mean at the data level, we could gain more insight into comparisons between and within different cancer research results.

Methods

This study is based on the analysis of cancer hallmark mapping schemes from publications that explicitly describe the mapping between intermediate knowledge bases and the hallmarks. We considered publications dating from after the second cancer hallmarks review paper [2], in order to consider all 10 hallmarks. We identified a large number of publications that describe a link to the hallmarks, but very few that explicitly describe the mapping procedure in sufficient detail to reproduce the process. Four publications and one peer-reviewed poster satisfied these criteria [18][19] [16][17][4]. The motivations for hallmark mapping were different between these publications, but all attempted to identify a general hallmark representation. There were two common ways to link the hallmarks of cancer to biological molecules and data; (1) mapping to Gene Ontology terms and (2) mapping to biological pathways. Both types of annotation allow a direct link to individual genes. Ulhen et al mapped hallmarks to biological pathways from KEGG and MSigDB. We refer to this mapping scheme as PW1. The other four mapping schemes are based on the association between gene ontology(GO) terms and cancer hallmarks and are referred to as GO1, GO2, GO3, GO4, respectively. The mapping scheme created by Plaisier (GO2)[17] was also used and cited in 2016[26] and Thorsson[27]. Although the GO mapping scheme is maintained by the authors at https://github.com/baligalab/sygnal/blob/master/R/goSimHallmarksOfCancer.R, it did not appear to change between these time-points.

Study workflow

The analysis process of this study is shown in two workflows (Figure 11 and 12). Figure 11 presents the workflow of comparison between different mapping methods and the investigation of the impact of using different mapping schemes on downstream omics analyses. Figure 12 shows the process of investigating the impact of GO evolution on the differences between mapping schemes. The hallmarks of cancer mapping schemes from each publication were developed using different methodologies and in accordance with the authors' background knowledge and aims. Table 2 summarises the main differences. For PW1, the genes annotated to each selected biological pathway were recorded by the authors at the time of publication and were therefore used directly, in addition to being updated with new information. For GO1-GO4, genes annotated with the selected terms (or their descendants) were not included in the publication data and were therefore reconstructed from current information. The genes currently annotated by selected Gene Ontology terms were

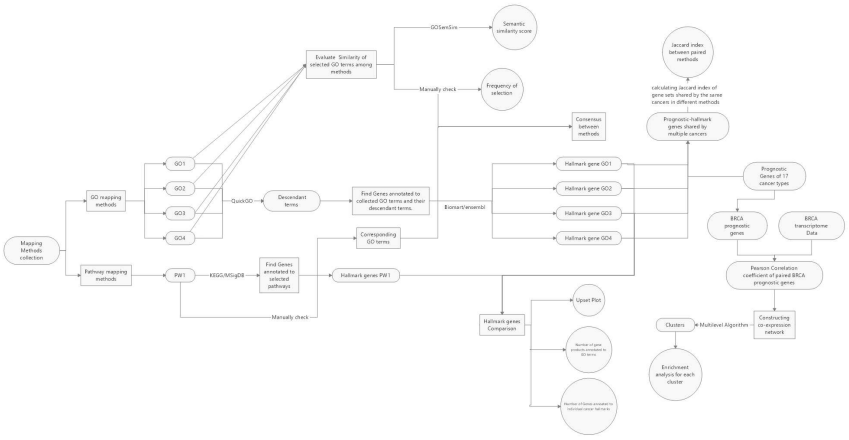


Figure 11 Workflow of analysis process. Rectangular nodes represent methods or steps generating data. capsule-shaped nodes represent data and circles represent graphs or visualization output.

Table 2 Summary of 5 mapping methods

Methods	Source	Experts involved	Size(pathway/GO)	Gene Number	Hallmark Included
GO1	Gene Ontology	Yes	57	10386	10 hallmarks
GO2	Gene Ontology	Unknown	40(1 obsolete)	10346	10 hallmarks
GO3	Gene Ontology	Yes	67	9164	8 hallmarks
GO4	Gene Ontology	Unknown	35	3078	8 hallmarks
PW1	MSigDB,Kegg	Yes	14	2172	unknown

identified using the human ENSEMBL database (version 99)[28] via biomart[29] and descendant terms were identified using QUICKGO [30]. The list of GO terms selected by each method and identified for each individual hallmark are available in additional file 5. The Full list of genes annotated by selected Gene Ontology terms and biological pathways are available in additional file 6.

Semantic similarity

The semantic similarities between the GO terms selected by the 4 GO mapping schemes were calculated in order to determine if there were close or distant relationships between terms included in different sets. The R package GOSemSim[31] was used with a best match average strategy (BMA). The annotation data used was the Genome wide annotation for the Human from Bioconductor[32]. The Resnik method[22] was used to assess semantic similarity. We performed pairwise comparisons for the four Gene Ontology term sets.

Hallmark gene comparison

To examine the overlap between Hallmark Genes from different mapping schemes, an upset plot was constructed by using the R package UpsetR[33]. In addition, the number of genes annotated to individual cancer hallmarks with different GO mapping schemes was also investigated. Bar chart was created using the Python package Seaborn[34].

Prognostic genes

The definition and data for prognostic genes was taken from the PW1 publication [4] as they provided a list of prognostic genes for 17 different types of cancer types in detail (additional file 7).

Prognostic-hallmark genes in multiple cancers

Genes labelled as both prognostic genes and hallmark genes are named prognostic-hallmark genes. For each method, We classified genes into different groups based on the number of cancer types where they were exclusively labelled as prognostic-hallmark genes. Then for each group, we further classified them into subgroups based on cancer types where they were labelled as prognostic-hallmark genes. Subgroups with less than 5 genes were removed. For each subgroup, if it existed in at least one method, it was included. For existing subgroups, we calculated the Jaccard Index to determine how different the subgroups were in pairwise comparisons of the mapping schemes. If a subgroup did not exist in 1 of 2 compared mapping schemes, the score was 0. Results were visualized by heatmap using the Python Package Seaborn[34]. Blocks colored in red represent high Jaccard index while blocks colored in blue represent low Jaccard index or 0.

Gene co-expression and clustering

To construct gene co-expression networks, we used public RNA-Seq data from TCGA with HTSeq-FPKM value[20]. 1222 breast cancer samples were downloaded. The co-expression coefficient of two different genes was calculated using Pearson correlation coefficient, using the following formula:

$$R_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

x_i and y_i represent the FPKM value of a gene in one patient.$\bar{x}$ and $\bar{y}$ represent the average FPKM value.

In this study, we used prognostic genes of breast cancer to build a co-expression network. To filter out low-correlated pairs, we selected the top 25% coefficient of pairwise genes. Based on the coefficient value, we constructed a co-expression network of prognostic genes using igraph in python[35]. Genes were clustered based on a multilevel algorithm[24]. For every cluster, we performed a cumulative distribution function(CDF) to identify if they were enriched in individual cancer hallmarks[23]. The clustering results can be found in additional file 8.

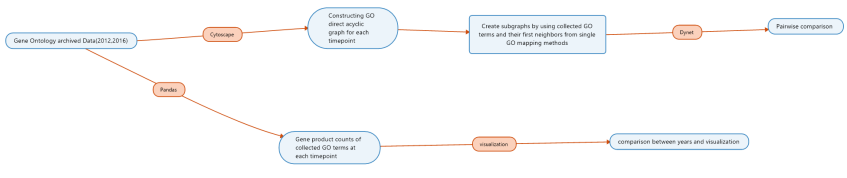


Figure 12 Workflow of investigating possible reasons of inconsistency. Nodes colored in blue represent data or processing steps while node colored in orange represent tools used in comparison.

GO evolution

TO investigate the impact of changes to the structure of GO , we used the Gene Ontology Biological Process hierarchy and associated annotations at three time points(June 2012, June 2016,July 2019). Data including GO term relationships and gene product count was downloaded from the Gene Ontology Consortium http://archive.geneontology.org/ [9].

Comparison between different Versions of GO

We constructed two GO hierarchy graphs in Cytoscape[25] based on archived relationship data downloaded from the Gene Ontology Consortium corresponding to different time points. For GO1, GO3 and GO4 mapping schemes, we create 3 sub-graphs by selecting selected GO terms of GO mapping schemes, their first neighbors and the edges between them based on the 2016 graph, while for GO2, we create a sub-graph in similar way based on the 2012 graph. The comparisons between sub-graphs was performed using DyNet[36]. White nodes and edges represent GO terms and relationships which are included in both networks while nodes and edges with color represent GO terms or relationships which are included in 1 of 2 networks. Dot plot shows the number of gene products annotated to different GO terms. They are created by using python package Seaborn[34].

Consensus between methods

For individual cancer hallmarks, a consensus GO term is one that was selected by more than 3 methods. Visualization of consensus terms belonging to hallmarks was performed using GOA-tools [37]. GO terms selected by 3 and 4 schemes are colored in orange and yellow respectively. Corresponding GO terms for MSigDB pathways are identified and mapped by the MSigDB database and are therefore used directly in our comparison. The KEGG pathway database does not provide an equivalent mapping to corresponding GO terms, so we derived these mappings by examining the most similar GO term definitions and KEGG pathway definitions.

Data provenance

This study aims to compare mapping schemes and data from multiple time points, using multiple knowledge and data resources. In order to make the work here transparent and reproducible, the provenance of all data and tools are listed. The descendants of selected GO terms were identified using QuickGO API and downloaded in May 2020. Genes annotated to selected GO terms and their descendants were identified using Biomart with the Ensembl 99 dataset. The version of GOsemsim used for semantic similarity was 2.14.0 and the annotation dataset in GOsemsim was Homo Sapien from OrgDb, version 3.10. The underlying GO version for each of these tools was declared to be the latest version at the time of analysis, although the exact version number was not provided by tool documentation. The classification of prognostic genes for 17 cancer types was taken directly from Ulhen et al, 2017 [4]. RNA-Seq data for breast cancer co-expression network construction was downloaded from TCGA, v23.0, and published on the 7th April, 2020. Gene Ontology data for 2012 and 2016 was taken from the Gene Ontology archive, published in June 2012 and June 2016 respectively. Pathway data from PW1 was from MSigDB version 5.2.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Yi Chen and Katherine Wolstencroft designed the research. Yi Chen collected the data and performed the experiment. Yi Chen and Katherine Wolstencroft analyzed the data and interpreted the results. Yi Chen, Katherine Wolstencroft and Fons.J.Verbeek wrote and revised the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
No ethics approval was required for the study.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable

Availability of data and materials
All data generated or analysed during this study are included in this published article and its supplementary information files

Acknowledgements
We would like thank the authors of Uhlen et al, and Sunjae Lee in particular, for his helpful advice and quick responses to queries about the reuse of his data.

Funding
This work has been supported by Chinese Scholarship Council through Leiden University

References

1. Hanahan, D., Weinberg, R.A.: The hallmarks of cancer. Cell **100**, 57–70 (2000)
2. Hanahan, D., Weinberg, R.A.: Hallmarks of cancer the next generation. Cell **144**, 646–674 (2011)
3. Aran, D., Sirota, M., Butte, A.J.: Systematic pan-cancer analysis of tumour purity. Nature communications **6**(1), 1–12 (2015)
4. Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., et al.: A pathology atlas of the human cancer transcriptome. Science **357**(6352) (2017)
5. Deng, Y., Luo, S., Zhang, X., Zou, C., Yuan, H., Liao, G., Xu, L., Deng, C., Lan, Y., Zhao, T., *et al.*: A pan-cancer atlas of cancer hallmark-associated candidate driver lnc rna s. Molecular oncology **12**(11), 1980–2005 (2018)
6. Google: Citation Statistics.
7. Kanehisa, M., Goto, S.: Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research **28**(1), 27–30 (2000)
8. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences **102**(43), 15545–15550 (2005)
9. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.*: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25–29 (2000)
10. Consortium, G.O.: The gene ontology resource: 20 years and still going strong. Nucleic acids research **47**(D1), 330–338 (2019)
11. Huntley, R.P., Sawford, T., Martin, M.J., O'Donovan, C.: Understanding how and why the gene ontology and its annotations evolve: the go within uniprot. GigaScience **3**(1), 2047–217 (2014)
12. Tomczak, A., Mortensen, J.M., Winnenburg, R., Liu, C., Alessi, D.T., Swamy, V., Vallania, F., Lofgren, S., Haynes, W., Shah, N.H., *et al.*: Interpretation of biological experiments changes with evolution of the gene ontology and its annotations. Scientific reports **8**(1), 1–10 (2018)
13. Dameron, O., Bettembourg, C., Le Meur, N.: Measuring the evolution of ontology complexity: the gene ontology case study. PLoS One **8**(10), 75993 (2013)
14. Livingston, K.M., Bada, M., Baumgartner, W.A., Hunter, L.E.: Kabob: ontology-based semantic integration of biomedical databases. BMC bioinformatics **16**(1), 1–21 (2015)
15. Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., *et al.*: Toward interoperable bioscience data. Nature genetics **44**(2), 121–126 (2012)
16. Knijnenburg, T.A., Bismeijer, T., Wessels, L.F., Shmulevich, I.: A multilevel pan-cancer map links gene mutations to cancer hallmarks. Chinese journal of cancer **34**(3), 48 (2015)
17. Plaisier, C.L., Pan, M., Baliga, N.S.: A mirna-regulatory network explains how dysregulated mirnas perturb oncogenic processes across diverse cancers. Genome research **22**(11), 2302–2314 (2012)
18. Kiefer, J., Nasser, S., Graf, J., Kodira, C., Ginty, F., Newberg, L., Sood, A., Berens, M.E.: A systematic approach toward gene annotation of the hallmarks of cancer. AACR (2017)

19. Hirsch, T., Rothoeft, T., Teig, N., Bauer, J.W., Pellegrini, G., De Rosa, L., Scaglione, D., Reichelt, J., Klausegger, A., Kneisz, D., *et al.*: Regeneration of the entire human epidermis using transgenic stem cells. Nature **551**(7680), 327–332 (2017)
20. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Network, C.G.A.R., *et al.*: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10), 1113 (2013)
21. Van De Vijver, M.J., He, Y.D., Van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., *et al.*: A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine **347**(25), 1999–2009 (2002)
22. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007 (1995)
23. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E.W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., Contributors, S...: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods **17**, 261–272 (2020). doi:10.1038/s41592-019-0686-2
24. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), 10008 (2008)
25. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research **13**(11), 2498–2504 (2003)
26. Plaisier, C.L., O'Brien, S., Bernard, B., Reynolds, S., Simon, Z., Toledo, C.M., Ding, Y., Reiss, D.J., Paddison, P.J., Baliga, N.S.: Causal mechanistic regulatory network for glioblastoma deciphered using systems genetics network analysis. Cell systems **3**(2), 172–186 (2016)
27. Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Yang, T.-H.O., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., *et al.*: The immune landscape of cancer. Immunity **48**(4), 812–830 (2018)
28. Hunt, S.E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., Parton, A., Armean, I.M., Trevanion, S.J., Flicek, P., et al.: Ensembl variation resources. Database **2018** (2018)
29. Kinsella, R.J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al.: Ensembl biomarts: a hub for data retrieval across taxonomic space. Database **2011** (2011)
30. Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'donovan, C., Apweiler, R.: Quickgo: a web-based tool for gene ontology searching. Bioinformatics **25**(22), 3045–3046 (2009)
31. Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., Wang, S.: Gosemsim: an r package for measuring semantic similarity among go terms and gene products. Bioinformatics **26**(7), 976–978 (2010)
32. Carlson, M.: org. Hs. eg. db: Genome Wide Annotation for Human. R package version 3.8.2 (2019)
33. Conway, J.R., Lex, A., Gehlenborg, N.: Upsetr: an r package for the visualization of intersecting sets and their properties. Bioinformatics **33**(18), 2938–2940 (2017)
34. Waskom, M.: seaborn: statistical data visualization. Python 3.6
35. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal **Complex Systems**, 1695 (2006)
36. Goenawan, I.H., Bryan, K., Lynn, D.J.: Dynet: visualization and analysis of dynamic molecular interaction networks. Bioinformatics **32**(17), 2713–2715 (2016)
37. Klopfenstein, D., Zhang, L., Pedersen, B.S., Ramírez, F., Vesztrocy, A.W., Naldi, A., Mungall, C.J., Yunes, J.M., Botvinnik, O., Weigel, M., *et al.*: Goatools: A python library for gene ontology analyses. Scientific reports **8**(1), 1–17 (2018)

Figures

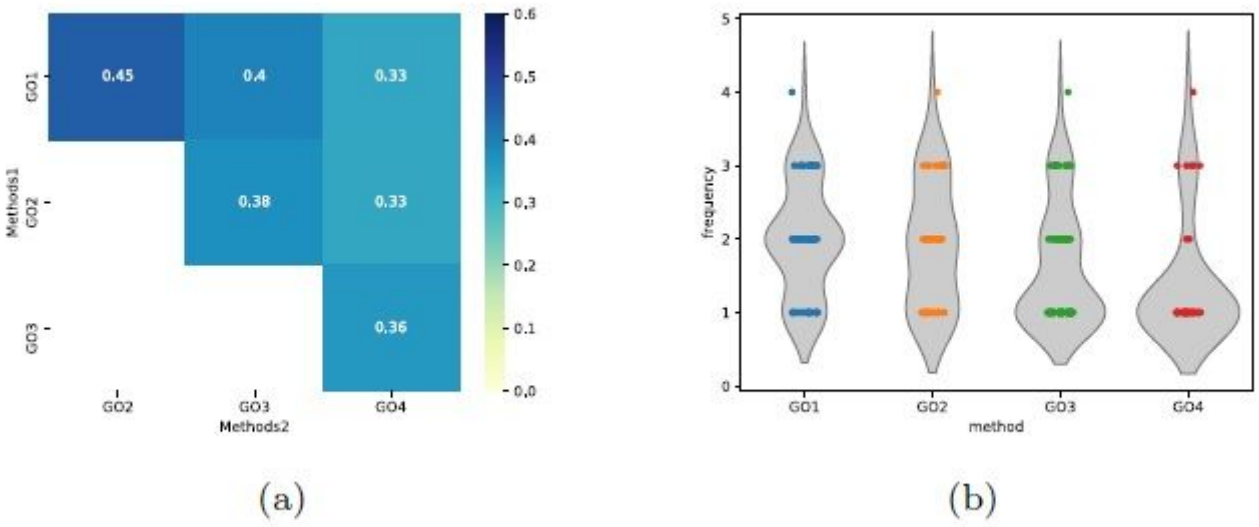


Figure 1

(a)Semantic similarity scores of pairwise mapping schemes. Numbers in the heat map represent the Semantic similarity score between Gene Ontology terms selected in different mapping schemes. (b)Frequency of selection of GO terms from different mapping schemes. The Y-axis represents the frequency of selection per GO term and the X-axis represents different mapping schemes. Dots inside the shape represent single GO terms belonging to this mapping methods.

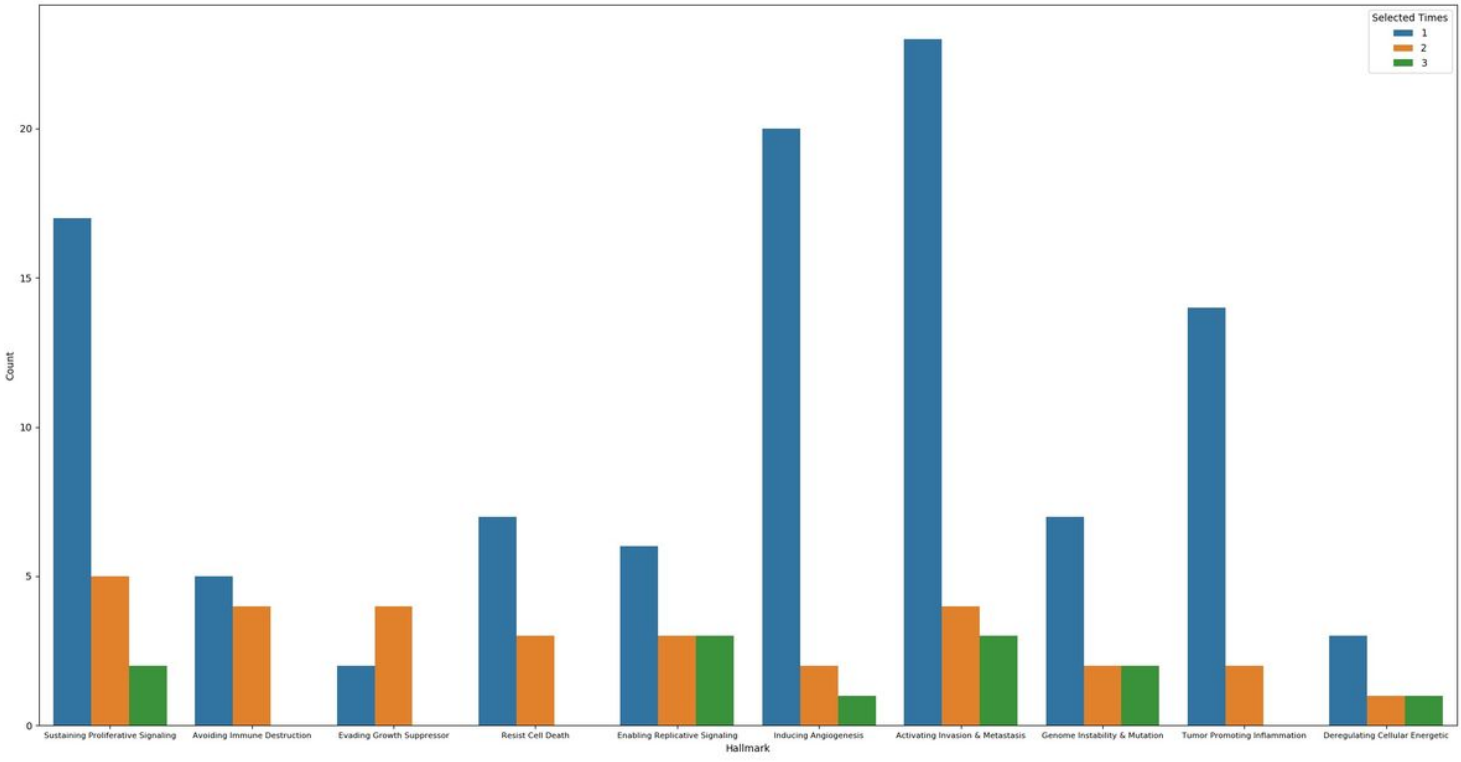


Figure 2

Frequency of selection of GO terms for individual cancer hallmarks. The Y-axis represents the number of GO terms. The X-axis represents individual cancer hallmarks. Bars colored in different color represent how frequently it was selected by mapping methods to annotate this cancer hallmark.

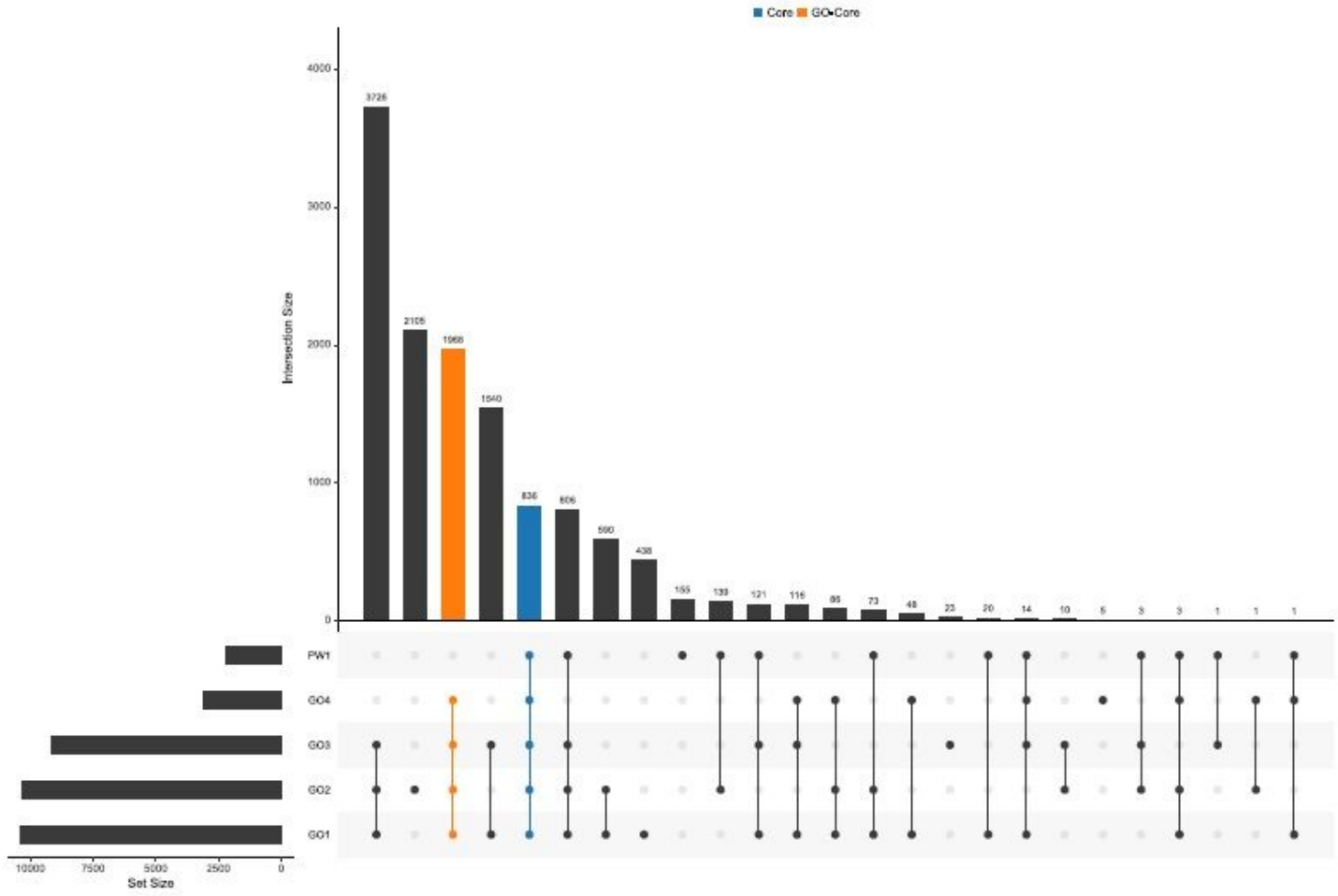


Figure 3

Hallmark Gene Set Comparison. The upset plot shows the number of genes in each hallmark gene set and their intersections. The orange and blue lines represent genes shared by GO mapping schemes and all mapping schemes respectively.

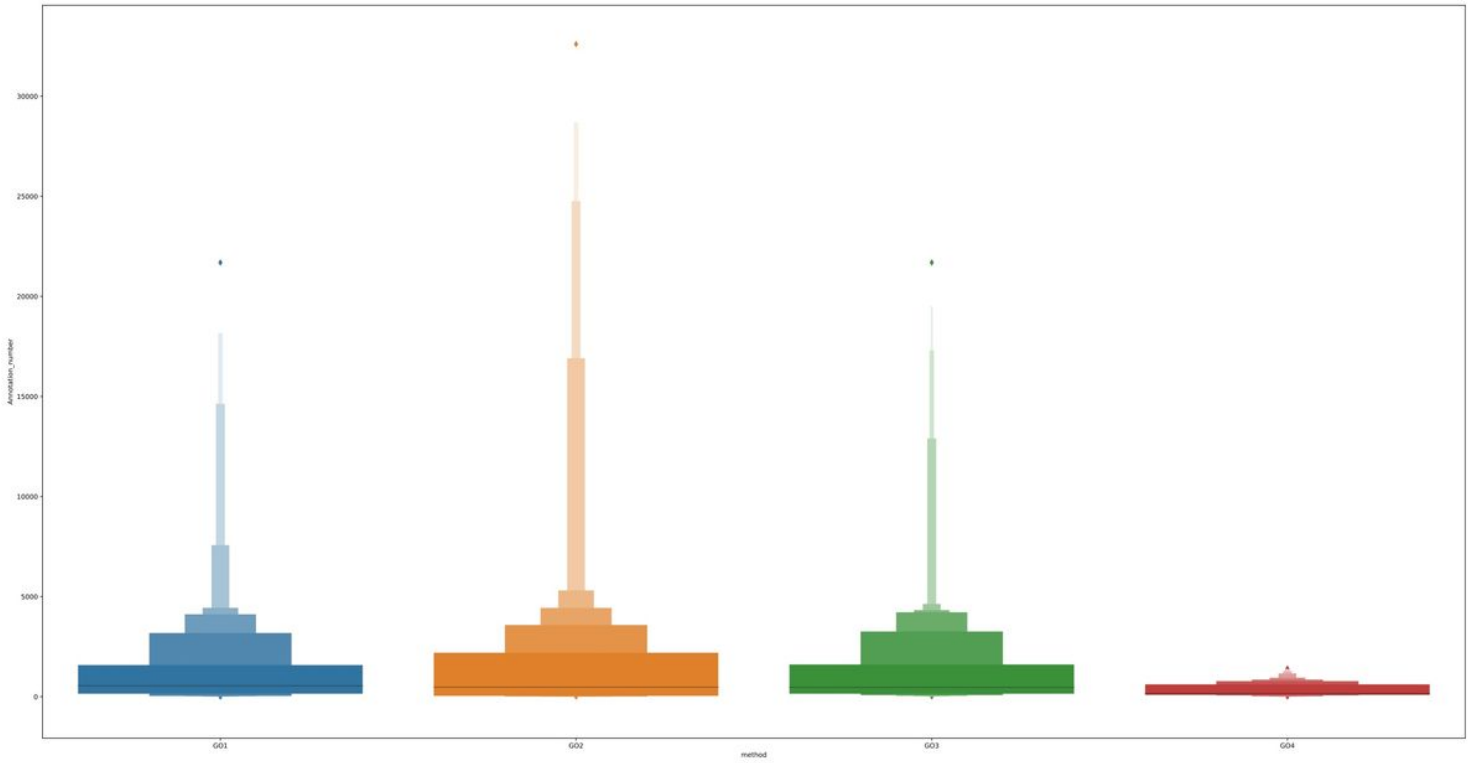


Figure 4

Gene Product Count. The box plot shows the number of gene products annotated to GO terms in different mapping methods.

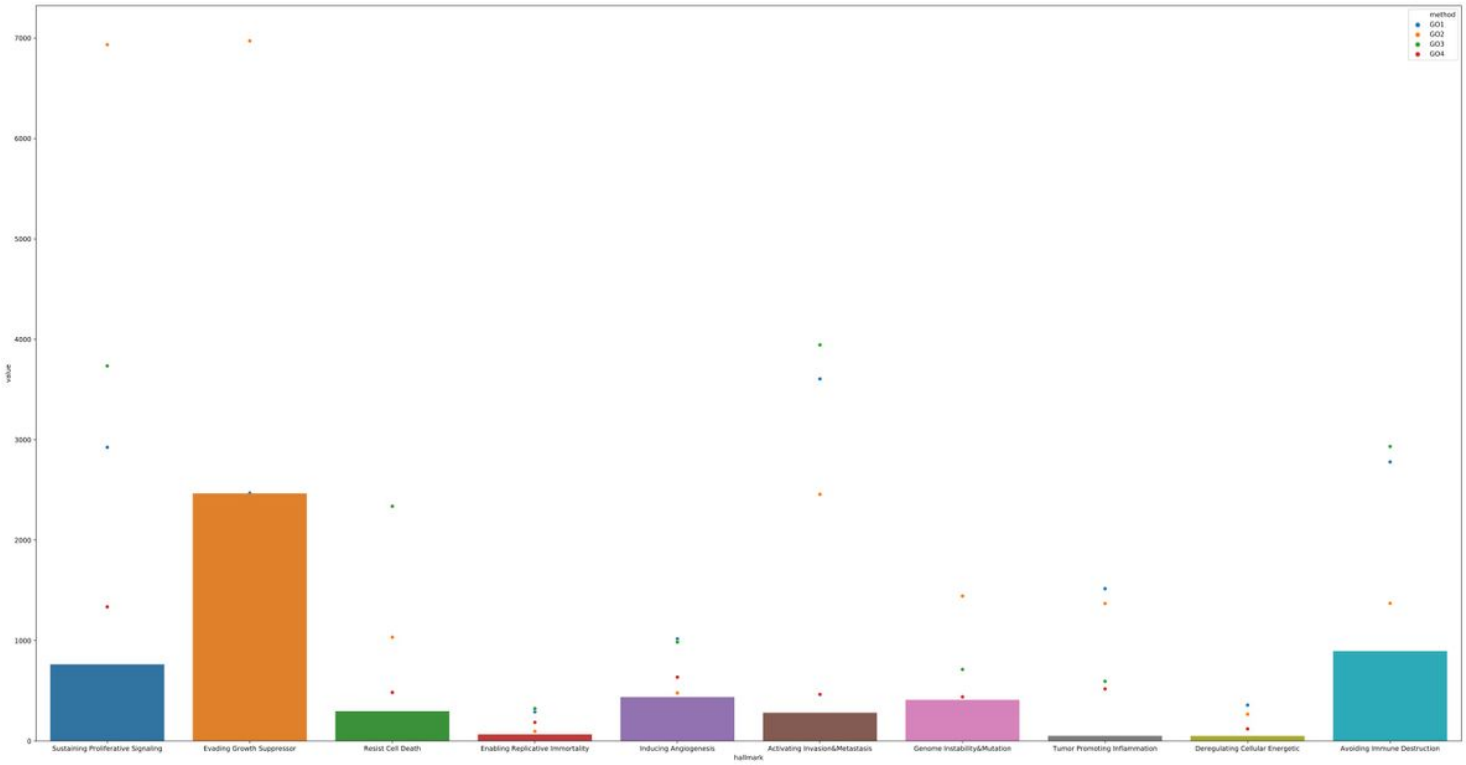


Figure 5

Gene Product count for individual cancer hallmarks. Nodes in the graph represent the count of gene products annotated to GO terms selected in different mapping schemes. Bars in the graph represent the number of gene products which are consistent across all the mapping schemes.

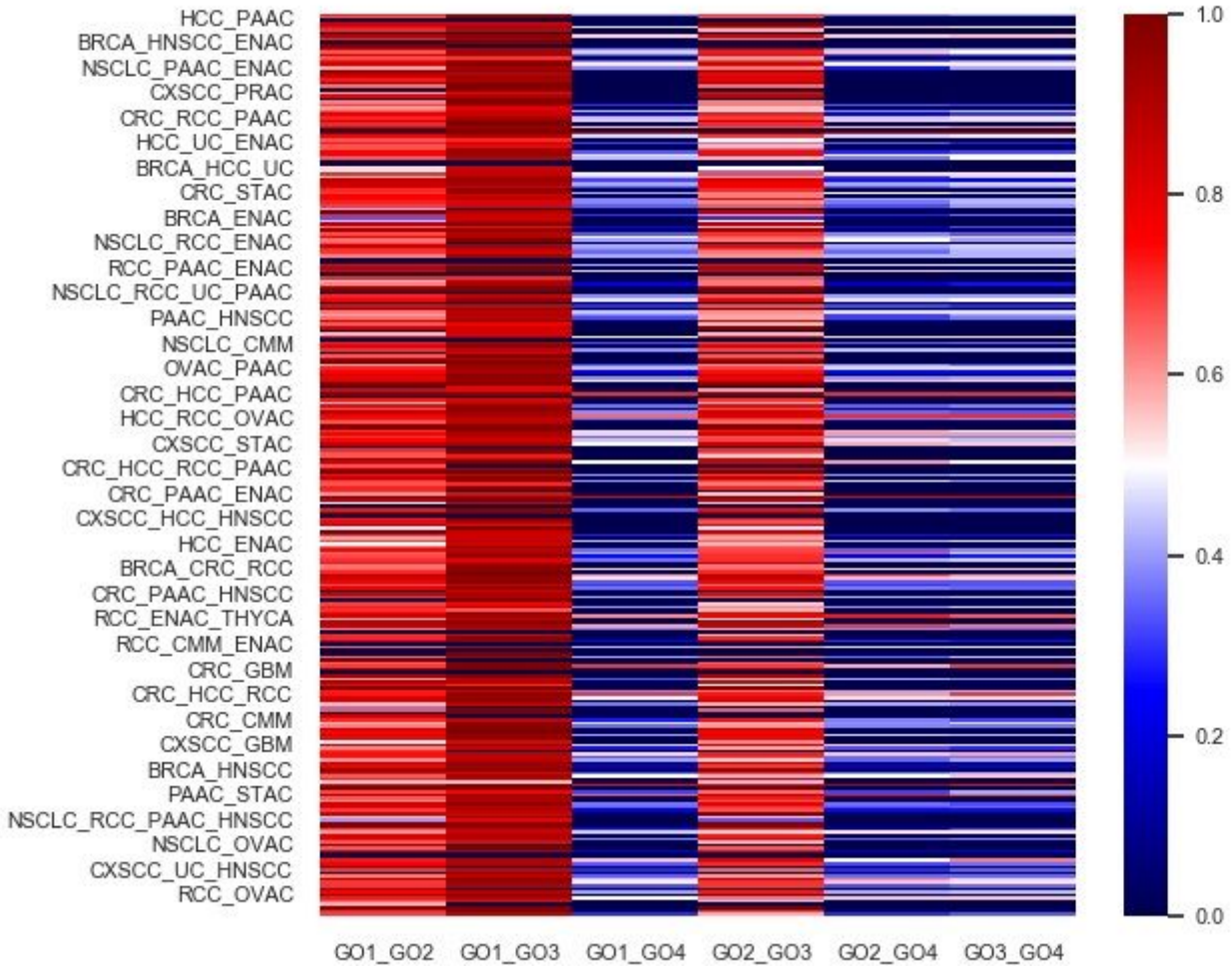


Figure 6

Pairwise comparison of groups of prognostic-hallmark genes identified in the same cancers between mapping schemes. Each block represents a group of prognostic-hallmark genes. The y-axis shows the combination of cancer types in which the corresponding group of genes are classified as prognostic-hallmark genes (abbreviations of cancer names can be found in additional file 1). The X-axis shows pairwise comparison of mapping schemes. The color of each block represents the similarity scores of the same group of genes using the Jaccard index. Darker red represents a higher score and darker blue a lower score.

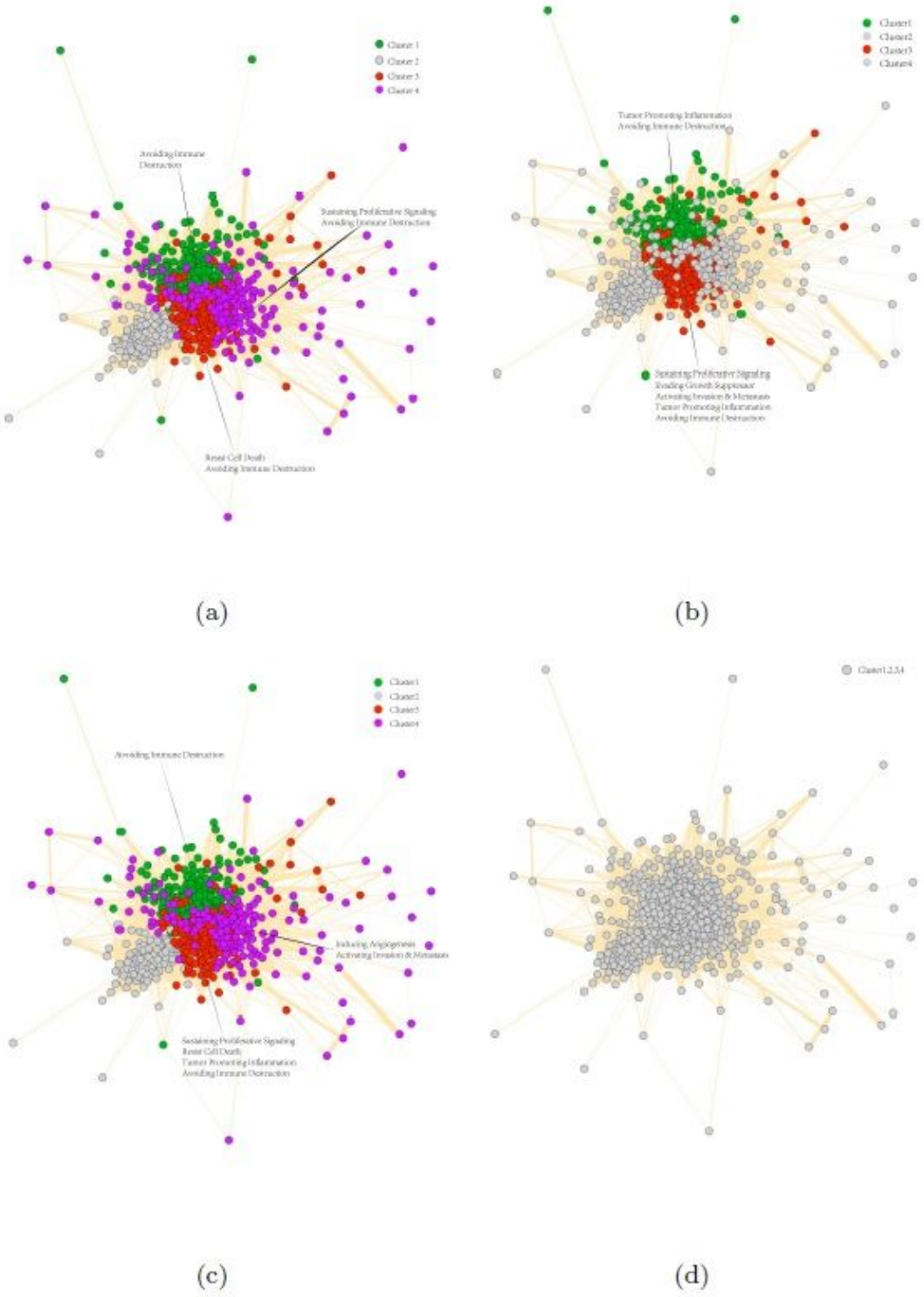


Figure 7

A,b,c,d represent co-expression networks constructed using prognostic genes and hallmark genes from GO1,GO2,GO3 and GO4 method,respectively.. Nodes coloured in grey represent a cluster only enriched in prognostic genes while nodes with different colours represent clusters enriched in both prognostic and hallmark genes. Dashed lines show which cancer hallmarks the clusters are enriched in.

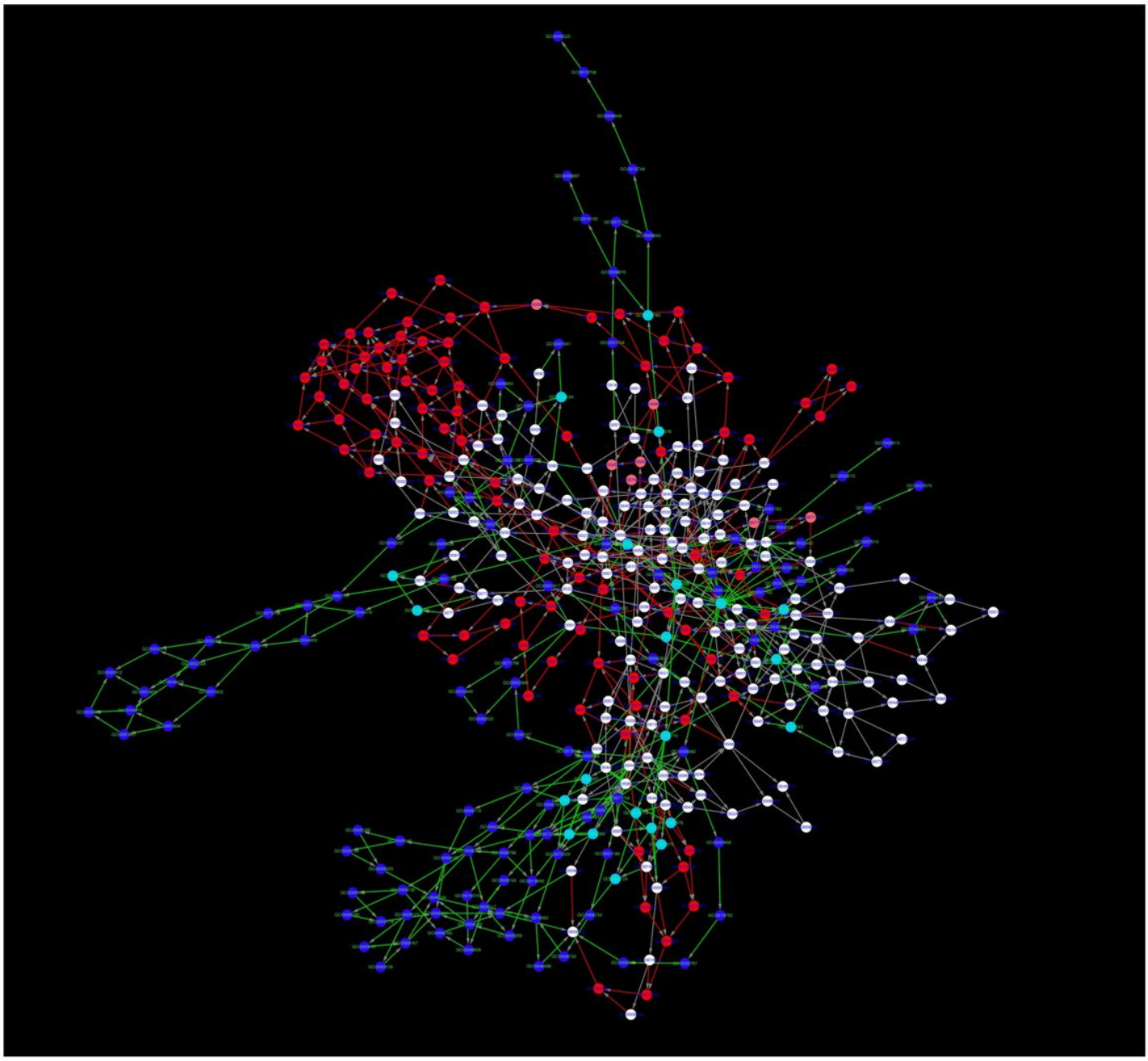


Figure 8

A comparison of the GO Biological Process topology in different years, constructed from all selected GO terms and their rst neighbours. Crimson nodes represent GO terms that existed in both time points but were only selected by GO2, while light red nodes represent GO terms which are obsoleted in 2016. Similarly, dark blue nodes represent GO terms that existed in both 2012 and 2016 but were only selected by GO1 and light blue nodes represent GO terms had not been created in 2012.

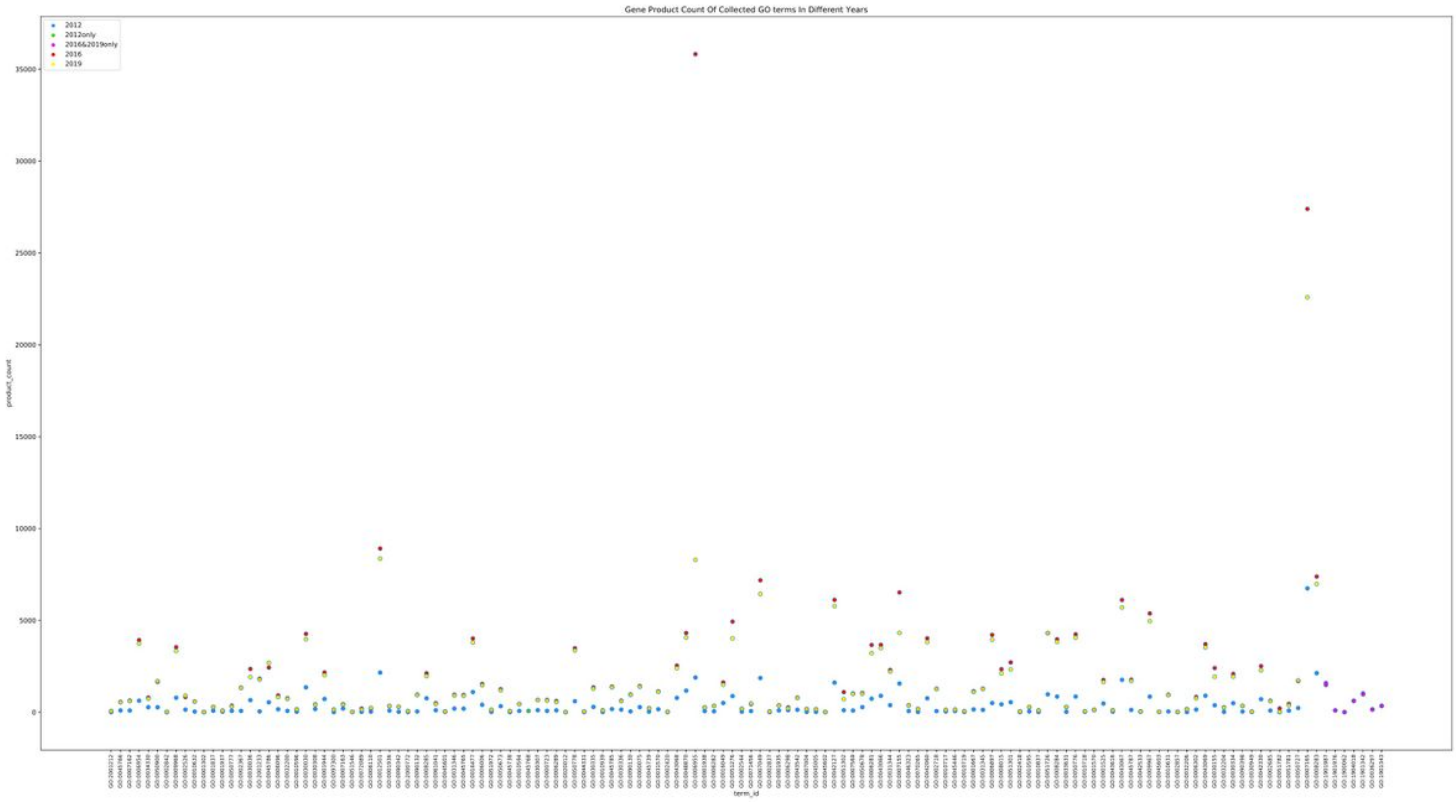


Figure 9

The number of gene product annotations in 2012 and 2016. Blue, red and yellow dots represent GO terms in 2012,2016 and 2019, respectively. green dots represent the GO terms that existed in 2012 but were obsolete in 2016 and 2019 and pink dots represent GO terms that haven't been created in June, 2012 but were created before July, 2016. The x-axis represents selected Gene Ontology terms; the y-axis represents the count of gene products annotated to a GO term and its descendants

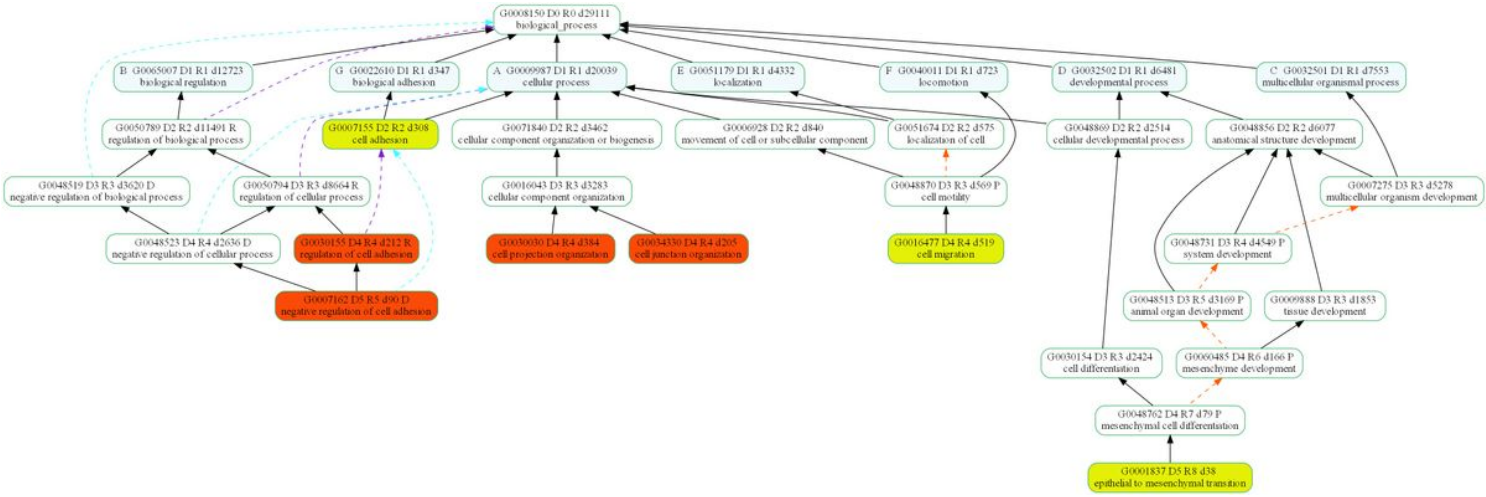


Figure 10

Visualization of the consensus in individual cancer hallmark Activating Invasion and Metastasis(include pathway mapping method). Nodes colored in red and yellow represent GO terms are selected or enriched by genes in 3 and 4 methods, respectively.

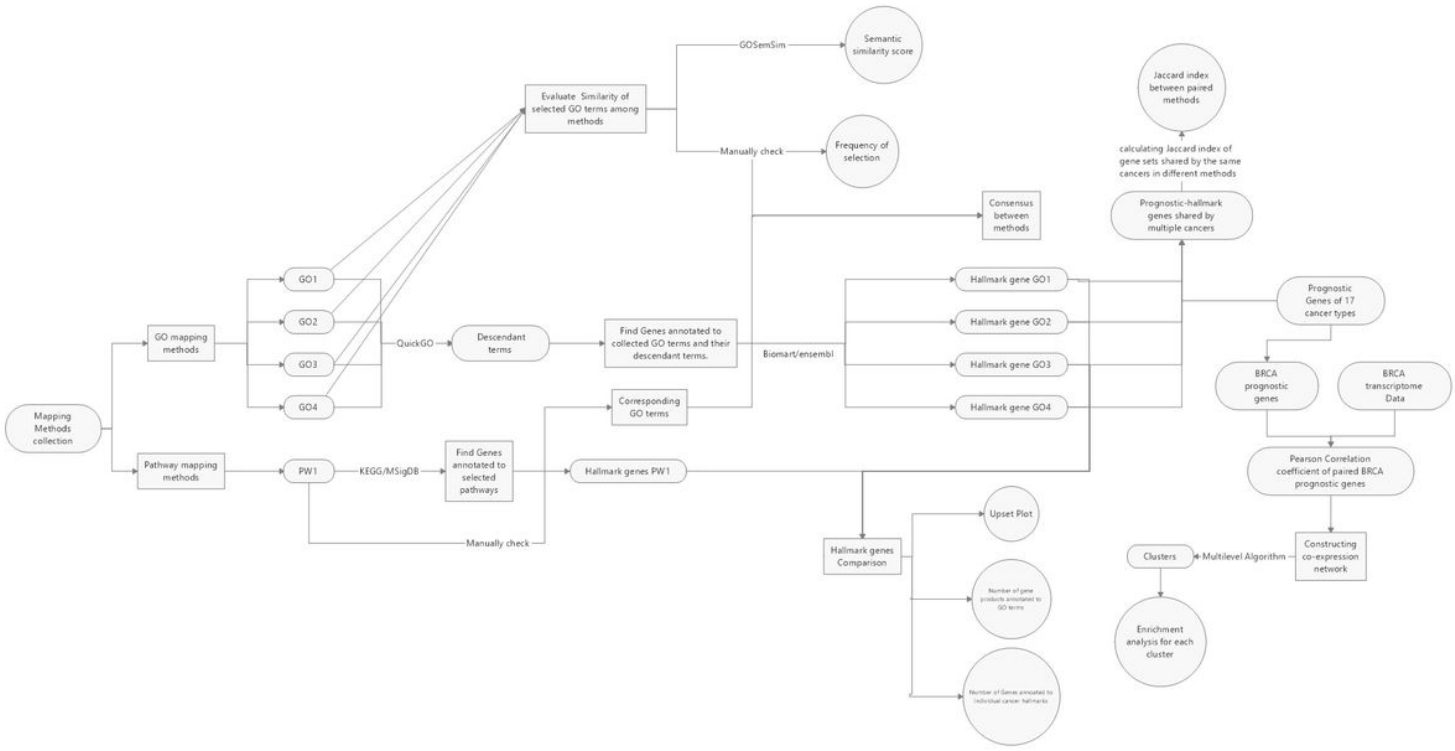


Figure 11

Workflow of analysis process. Rectangular nodes represent methods or steps generating data. capsule-shaped nodes represent data and circles represent graphs or visualization output.

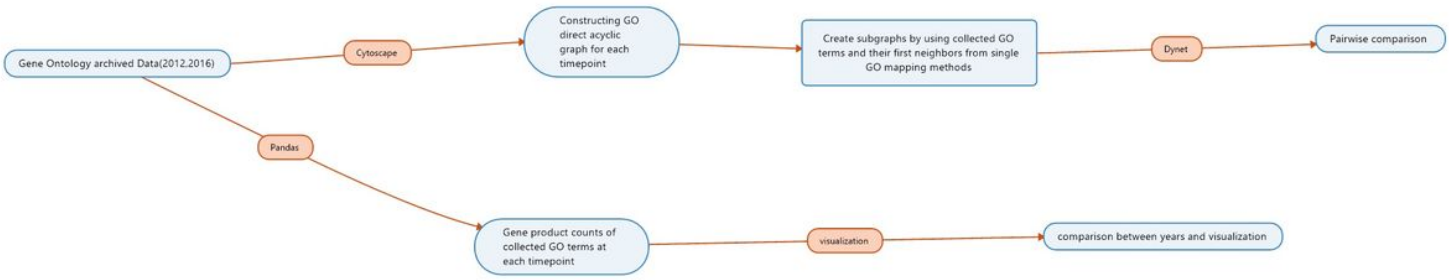


Figure 12

Workflow of investigating possible reasons of inconsistency. Nodes colored in blue represent data or processing steps while node colored in orange represent tools used in comparison.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.xlsx
- Additionalfile2.csv
- Additionalfile3.png
- AdditionalFile4.xlsx
- AdditionalFile5.xlsx
- AdditionalFile6.xlsx
- AdditionalFile7.xlsx
- AdditionalFIle8.csv
- Table1.xlsx
- Table2.xlsx