Universiteit Leiden
The Netherlands

Opleiding Informatica

Bioinformatics analysis of the core hallmarks of cancer genes

Metin Yarim

Supervisors:
Katherine Wolstencroft

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

11/06/2020

Abstract

Research into cancer has led to a broad field of studies and so there are many different ways to do research on cancer. The Hallmarks of Cancer is such a way as it provides a framework to improve the study into the different mechanisms and types of cancer. However, due to different methods into the mapping of genes to the hallmarks, different results can be obtained. Based on the results of [5] a core set of 793 genes is used that has been mapped consistently to hallmarks.

The aim of this thesis is therefore to analyse this core set to the different hallmarks and the mapping of the genes. With the use of bioinformatics methods, like functional enrichment analysis, the core set and the different study methods were compared. This thesis shows that the core set of genes represent the different hallmarks. However, this representation of the hallmark is slim and it varies much between the hallmarks.

Contents

1 Introduction

Cancer research is a broad field of study and a lot of work has been put into to it. Still much work has to be done. One of the main reasons for this is that cancer is a complex disease with all kinds different variations. There are many causes and mechanisms that makes it hard to find a general cure or treatment of cancer, if there even is one. This also created many studies that researched different types of cancer using different techniques and different information. Of course, different types of cancer require different techniques. However, a lot similarities can be found between the different types of cancer and how they develop and behave. As a result, the hallmarks of cancer were found by Hanahan and Weinberg [6] and they were later updated with more hallmarks [7]. This provided in a framework to study the different mechanisms and types of cancer more efficiently.

The Hallmarks of cancer are made of eight capabilities and two characteristics of tumors and its environment that enables tumors to develop and grow. In 2000, the first six hallmarks were introduced: 1) Sustaining proliferative signaling, 2) evading growth suppressors, 3) resisting cell death, 4) enabling replicative immortality, 5) inducing angiogenesis, and 6) activating invasion and metastasis. Later in 2011 two emerging hallmarks were added to the set: 1) Reprogramming energy metabolism and 2) evading immune destruction. Next to those two hallmarks, two enabling characteristics were added: 1) Genome instability and mutation and 2) tumor-promoting inflammation. Those are called so because those characteristics enable tumor cells to acquire the different hallmarks. With the help of genome instability tumor cells are more easily able to have mutations in the genome and that in turn can create cells that show those different hallmarks. This is the same with tumor-promoting inflammation characteristic. For simplicity, during this study the two characteristics will be regarded as a hallmark. So, this brings it to a total of ten hallmarks that will be studied.

1.1 Related Work

With the help of the hallmarks many studies have been done into finding genes belonging to the different hallmarks and also for finding differences between various types of cancer. However, currently there are several main methods on how genes can be linked to a hallmark. In this paper two of those methods are compared. The first method is mapping genes to hallmarks using known biological pathways. This was used in the study by Uhlen et al [4]. There the transcriptomes of protein-coding genes of 17 types of cancer where compared. In order to identify hallmark genes they mapped various pathways to the hallmark. This resulted in 14 pathways of which their genes were seen assigned as being hallmark genes.
The second method is mapping with the Gene Ontology(GO) terms and it is the most used method. In order to link genes to a hallmark Gene Ontology terms were selected for each hallmark. Based on those terms genes can be found that have one of those terms and can be linked to the hallmark. The used papers with this mapping method are Kiefer et al.(2017)[2], Hirsch et al.(2017)[1], Knijnenburg et al.(2015)[3], and Plaisier, Pan and Baliga(2012)[10]. They are referred to as GO1-4 respectively for this study.

1.2 Data & Problem

A great part of the studies with the Hallmarks of Cancer involves the categorising of genes and mapping them to the hallmarks for which they are involved in. This information can then be used for further research and for better and more specific treatment of human cancers. However, this step can be quite challenging for mainly two reasons:
The first reason is the complexity of the genes used for the hallmarks itself. In general, genes can have different functions and locations and so can play a role in multiple processes for multiple hallmarks. It is also possible that the process it is involved in is a general process, so that it is used in used different places and circumstances. A gene or a mutated form of it might also be (more) involved in only several types of cancer. This is also visible in the study by done Uhlen et al [4]. There, 17 types of cancer were compared and the results showed a huge variety of hallmark genes between the different types of cancer. Some genes were associated to only one type of cancer but others to multiple types. This shows that a gene can easily be linked to multiple hallmarks and that can also differ between different types of cancer.
The second reason is that the results can be quite diverse depending on the method that is used. There are several methods for mapping genes to hallmarks as described earlier. The most used method is mapping with the Gene Ontology terms. Those terms are statements that describe either a molecular function, a cellular component or a biological process and can be structured hierarchically using different types of relations within a set. These terms are connected to genes by all sorts of studies. In addition to the term an evidence code is added to state through what kind of way this term is connected to the genes. During a study for hallmark genes, a group terms are associated with the selected hallmark. Genes with one of those terms can then mapped to that hallmark. In addition to this evidence codes can be used to further specify the search. As the hallmarks describe different capabilities or processes, terms that describe the biological processes of the genes are mostly used, so the annotation set Biological Process. Another method is to use known biological pathways. In this method pathways that can be associated with a hallmark are used and the genes in the pathways can then be mapped to the hallmark. As these methods are quite different and so can the results of it be different. That can be seen in the study done by [5] where large differences between the mapping methods are found. Five different papers were compared where four of them used Gene Ontology terms (GO1-4) and the other one used gene pathways ([4]). Even between the four papers using Gene Ontology terms there was a lot of differences in the agreement of annotating genes to hallmarks. There was also a major difference in the total amount of genes being mapped to a hallmark between the methods, where one method gave almost 10000 genes while another one only gave little more than 2000 genes as results. In the end there were only 793 genes that were annotated to a hallmark by all five papers.

The Hallmarks of Cancer are useful to study with, but these reasons show that the mapping of genes to hallmarks is challenging. Especially in the results between different mapping methods this is visible. In order to find causes of the differences between the mapping methods, the set of 793 genes from [5] is analysed in this study. This set will be called the core set in this study. While using several bioinformatics methods, the core set is studied for the different hallmarks and how they are represented in the core set. One of those methods is enrichment analysis for GO terms. With this analysis GO terms are found that are over-represented in a selected set of genes.

The aim of this study is to analyse the core set of 793 hallmark genes. This set has been mapped multiple times but with different methods with different results. Therefore, in this study this set is compared with the help of bioinformatics methods. To search how the mapping is done between the hallmark and how the results represent the different hallmarks. This is done with the following research question:

Does the gene set that has been consistently mapped to the hallmarks, represent specific cancer processes, or is it simply bias in the literature?

In section 2 the used methods are explained. In section 3 the results are shown and they are further discussed in section 4.

2 Methods

In order to analyse the core set of 793 genes that was mapped by all the papers to a hallmark, the results were stripped down with only the genes of that are in the core set. As described in the introduction four of the five paper methods were performed with Gene Ontology(GO) terms, GO1-4. This resulted then in lists of genes per hallmark for each method which used GO terms. The other paper used several pathways for the different hallmarks, so here each pathways was stripped down to only the genes belonging in the core set. However, the results of that paper gave pathways, not hallmarks and those pathways could not be easily linked to a specific hallmark and it is not mentioned by the paper or anywhere else how the pathways are linked to the hallmarks. Because of that, most of the methods used in this study are performed only with the four paper results that used GO terms as they were more comparable. As described earlier, this study only used GO terms belonging to the annotation set biological process. In Figure 1 the map is visible of the methods used in this study. Table 1 shows the abbreviations that are used here for the hallmarks to name them easily.

Hallmark	Abbreviation
Sustaining proliferative signaling	*Signaling*
Evading growth suppressors	*Growth*
Resisting cell death	*Cell death*
Enabling replicative immortality	*Immortality*
Inducing angiogenesis	*Angiogenesis*
Activating invasion and metastasis	*Metastasis*
Genome instability and mutation	*Mutation*
Tumor-promoting inflammation	*Inflammation*
Deregulating cellular energetics	*Energetics*
Avoiding immune destruction	*Immune*

Table 1: The abbreviations for each hallmark

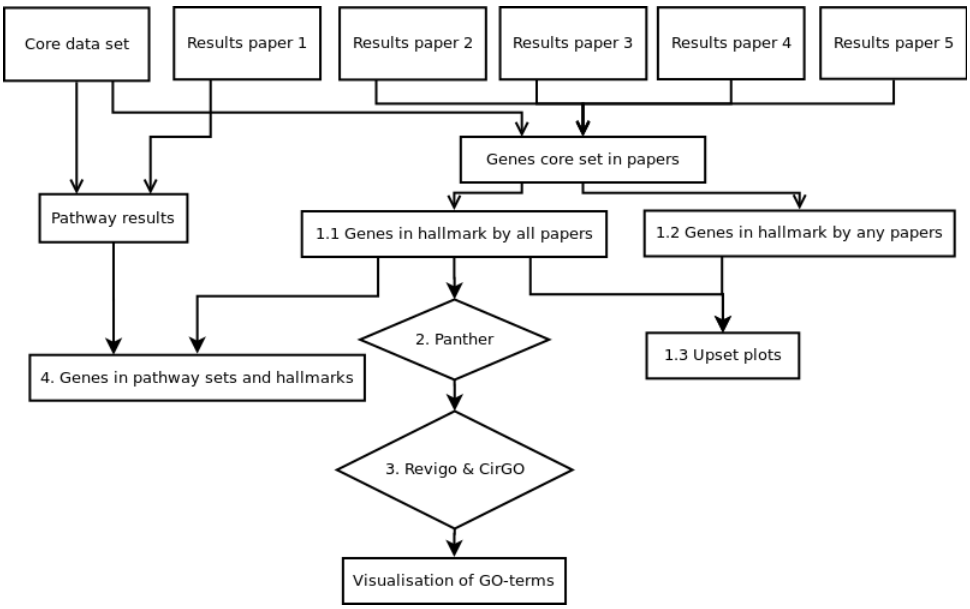

Figure 1: Map of methods

2.1 Comparisons

The first method involves different comparisons between the different results of the papers. With the results from the GO terms two different sets of comparisons were made using excel sheets and a few simple python programs.

2.1.1

The first comparison is about finding the agreed genes for each hallmark. This was done by checking for each hallmark with the results of the GO paper methods. If every method mapped a gene to a hallmark, that gene was then counted as being linked to that hallmark. This can give insight for each hallmark how many times it is linked by all paper methods using GO terms and how this is divided between the different hallmarks.

2.1.2

The second comparison determined if a gene was linked to a hallmark by any method. So if at least one method mapped a gene to a hallmark then the gene was counted as being linked to that hallmark. By doing this, it could give insight into not only which hallmarks were linked a gene even only once but also, together with the first comparison, it shows how much of an agreement there is between the different paper methods. This is caused by the fact that a lot of genes are linked to a hallmark but not by all methods. This agreement, when it is linked by all methods, cam differ greatly between hallmarks. It also shows how the genes are linked in the two different ways of comparing.

2.1.3

In order to research the comparisons more, several UpSet plots were made. This is done in RStudio with the UpSetR package. An upset plot uses different sets of information to visualize the combinations of intersections between the sets. For this study the sets contained the genes of a hallmark or of a paper method in a particular situation. The upset plots then shows intersecting groups of genes between the different sets. These plots was made with the linking for a specific hallmark across the different paper methods.
It was also done for sets of genes that all were linked a specific amount of times to a hallmark agreed by all the paper methods. It can tell how the mapping is done for each hallmark, like if the mapped genes are consistently to one or more other hallmarks. But also how genes that are linked to a specific hallmark are linked to other hallmarks.

2.2 Functional Enrichment

To analyse the results further the program Panther [8] was used to perform functional enrichment analysis. This was done on the results of the first comparisons per hallmark, using the overrepresentation test with Panther version 14.1 released on 12/3/2019. The gene set was compared against all human genes in the database with the PANTHER GO-Slim Biological Process as the annotation data set using Fisher test and the false discovery rate as correction. The GO-Slim database was used to get a broad set of results. However, this also affected the results in the way that the resulted terms were less specific.
This produced lists of GO terms for every hallmark. Based on the input genes these lists contains GO terms that are over-represented in the set of genes with a certain P-value. This can give a view of what kind GO terms are linked to that hallmark based. These analyses are performed for GO biological processes as they are used the most in general and provides the most important information.
In addition to the standard results per hallmark, unique GO terms were determined for each hallmark from the Panther results that were not found in other hallmarks.

2.3 Visualisation Functional Enrichment results

In order to visualize the enrichment results from Panther the program CirGO [9] was used. The lists from Panther were first pre-processed into the right format. This is be done with Revigo [11]. This program uses GO terms and a corresponding value, for example their P-value, to summarize the list. It uses a clustering algorithm that is based on semantic similarity. The similarity is measured by how close terms are in the GO hierarchy, for example if two terms are siblings or parent and child in the hierarchy. With the similarity and the P-value of each term certain terms can be cut off and clusters can be made with a single representative GO term. In this study the GO terms from the Panther results were given as the input with their raw P-value while using medium similarity and the SimRel semantic similarity measure. After that step a treemap with a two-level hierarchy is made.
With the treemap produced from Revigo, a visualisation is made with the program CirGO. CirGO calculates and organises the data and values from the treemap and a two-layer hierarchical circular structure is made. It contains two rings that show groups of GO terms and their proportion in the

whole. In the inner ring are parent GO terms stated. In the outer ring can child terms be added based on their size and significance within the parent group. As a result, visualisation was made for the standard results for every hallmark and also for a few lists with unique GO terms.

2.4 Pathways sets

Besides the results from the GO paper methods, results were obtained in the study by [5] that were based on the pathways used in [4]. These pathways included genes that were assigned as hallmark genes. In order to compare these pathways, each pathway was checked for which genes are part of the core set of genes. This can give a view for each pathway how many of the genes were also seen as hallmark genes by the GO paper methods.
After this step, each pathway was compared with the results from the first comparison, the genes for each hallmark linked by all four paper methods. This resulted in list for each pathway with the amounts of genes that were linked to each hallmark based on the GO paper results.

3 Results

The results of this study show that there is some agreement between the paper methods but that much as one might hope to get. As already mentioned earlier Uhlen et al. used pathways and the other papers used GO terms. This resulted in the fact that most of the results are based on the GO terms methods results.

3.1 Comparisons

Prior to comparing several notions could be made. An important one is that some hallmarks were not mapped by all the paper methods. The *Growth* hallmark is not mapped by GO3 and GO4. For the *Energetics* hallmark results are not provided by GO3 and for the *Immune* hallmark results are not provided by GO4.

3.1.1

With the first comparison the linking of a gene to a hallmark is done if it is mapped by all four GO terms paper methods, so that there is an agreement by all the those methods. In figure 2 the results are visible with the blue bars. It shows that there are some differences between several hallmarks. *Signaling* and *Growth* have high amounts of genes mapped to it while the other hallmarks have much less genes linked it. Between the old, the first six, and newer, the last four, hallmarks differences are not quite visible, aside from the first two hallmarks. However, the *Growth* hallmark is only mapped by two of the four paper methods, so that influences the agreement between the methods a lot.

3.1.2

The second comparison yielded all the links made by the GO terms paper methods to all hallmarks. So it is linked when it is mapped by at least one paper method. The results of this are shown in figure 2 with the orange bars. From it is visible that the amounts of genes being linked to a hallmark

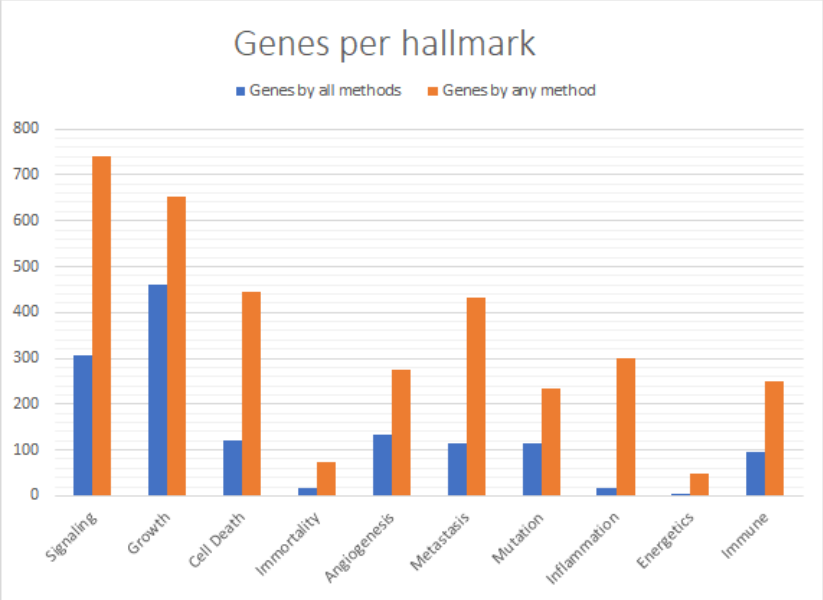

Figure 2: The quantities of genes linked to a hallmark by all paper methods and by any paper method

vary quite a lot. Especially the older hallmarks, the first six, have higher quantities, except the *Immortality* hallmark.
With these results combined with the results of the first comparison an agreement rate was made. Table 2 shows the percentages of genes that are linked to the hallmark by all the paper methods out of the total amount of genes that are linked to the hallmark by any paper method. The agreement rate between the hallmarks varies much, even between the older and newer hallmarks. Also in this case it can be seen that *Growth* only is mapped by two paper methods as the agreement is much higher than the other hallmarks.
Besides the agreement, the genes were also compared on the amounts of links to the different hallmarks. Figure 3 shows this in the case of linking by all the paper methods, in blue, and in the

Hallmark	Percentage
Signaling	41,4
Growth	71,0
Cell Death	27,5
Immortality	23,0
Angiogenesis	49,3
Metastasis	26,4
Mutation	49,8
Inflammation	6,0
Energetics	10,4
Immune	38,6

Table 2: The percentages of agreement per hallmark

case of linking by any methods, in orange, to a hallmark. The amounts when there is an agreement show that most of the genes are linked to one or two hallmarks but there are some that are linked to even four or five while other genes aren't linked to any hallmark. Linking by any paper method shows that the most genes are linked to several hallmarks, some even to eight or nine hallmarks.

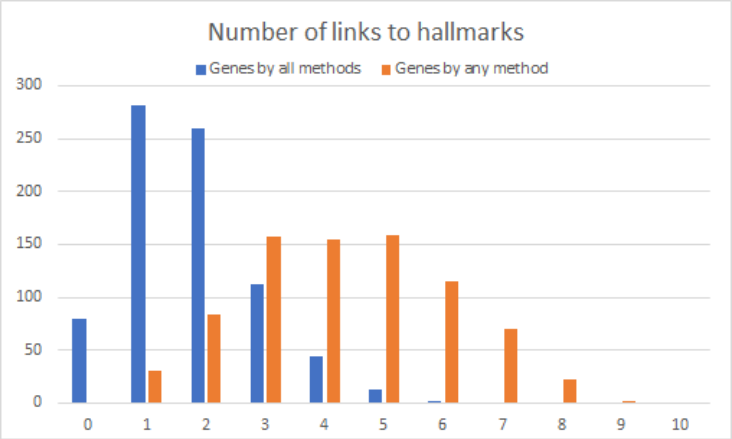

Figure 3: The number of times a gene is linked to a hallmark in the different ways of comparing.

3.1.3

After comparing the mapping of the paper methods, it can be seen that the agreement the different methods varies a lot between the different hallmarks. Percentages go from six until 50 (*Growth* has much higher results but is only mapped by two methods). In order to examine it further UpSet plots were made for some of the hallmarks with low agreement percentages. These show well how the linking of the genes for a hallmark is divided between the different paper methods. Figure 3.1.3 shows four hallmarks that have low agreements across the paper methods. Overall, the results of the hallmarks are quite spread out different sets and the set where all the paper methods agree on is not always largest. This can especially been seen for hallmarks b, c and d. Interestingly sets of genes without the results from GO4 are relatively higher than other sets with GO4 results or even the highest set. For *Signaling* this not entirely the case but here are still a lot of different sets that causes it to get a low agreement across all the paper methods. The spreading out is especially high for the *Inflammation* hallmark, where there are many sets with small quantities.

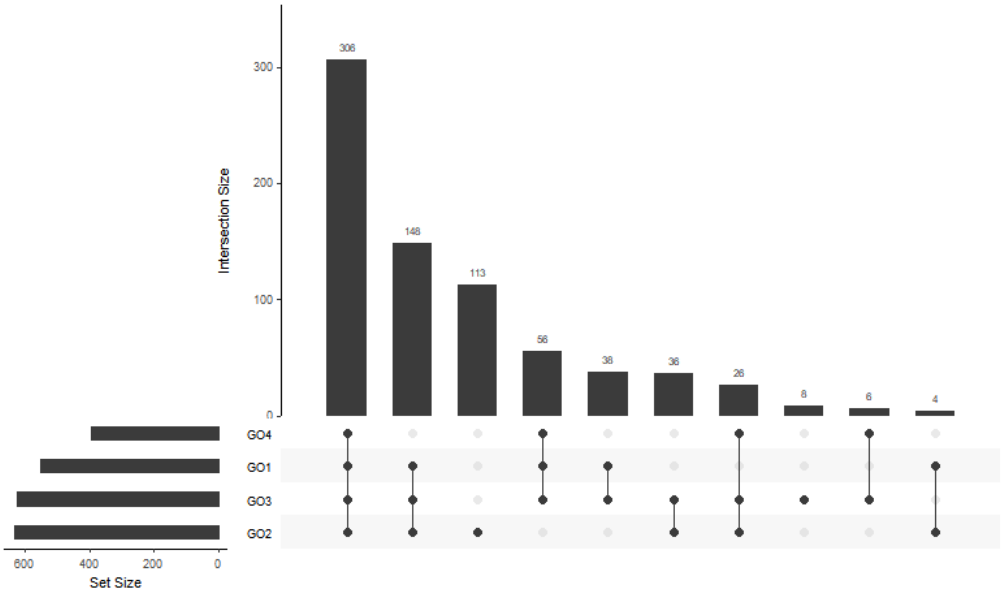

Figure 4: Upset plot for the *Signaling* hallmark with different linkings to the four GO paper methods.

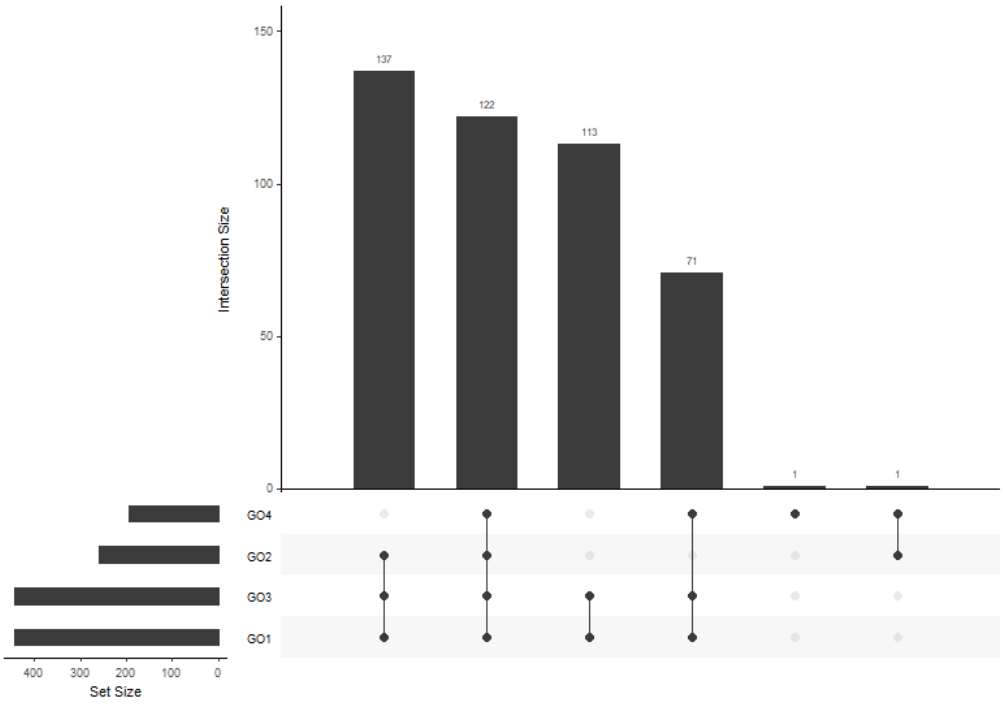

Figure 5: Upset plot for the *Cell death* hallmark with different linkings to the four GO paper methods.

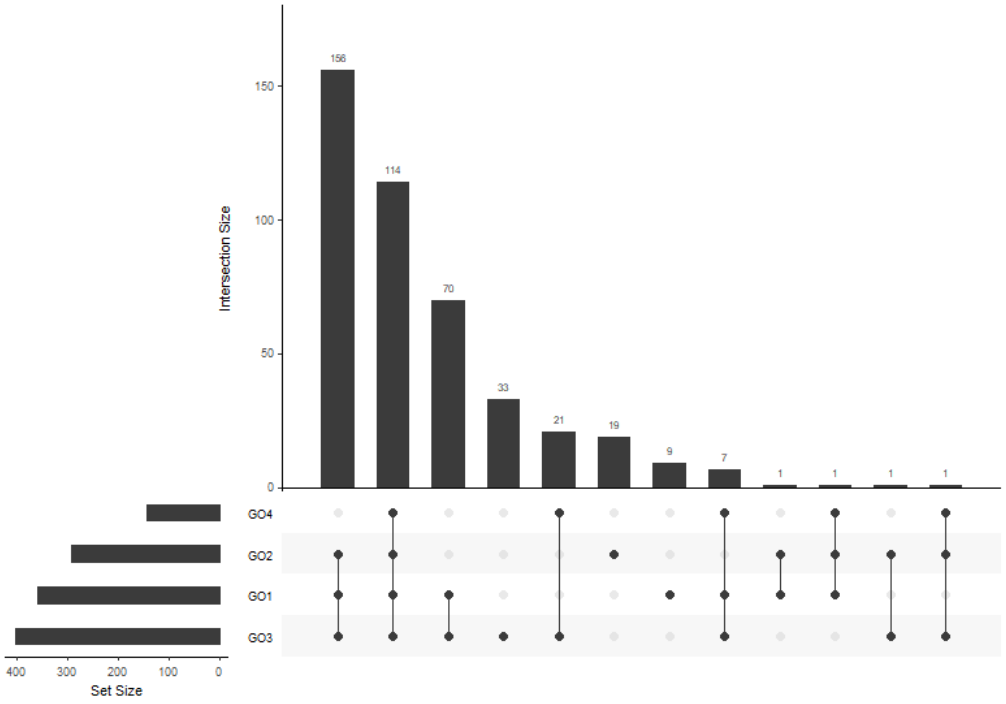

Figure 6: Upset plot for the *Metastasis* hallmark with different linkings to the four GO paper methods.

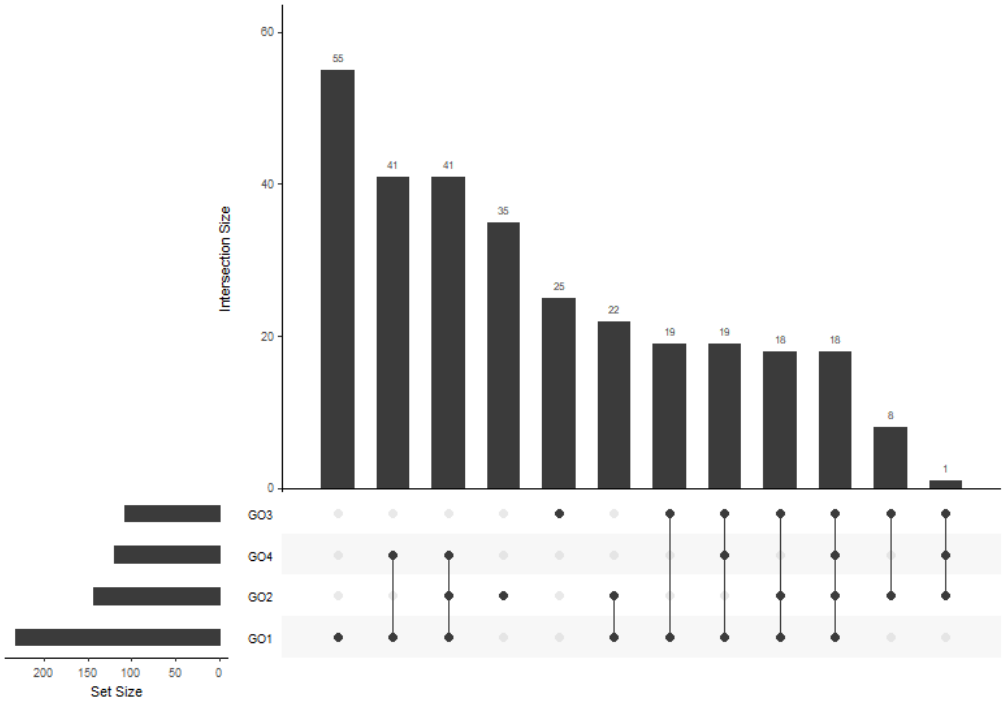

Figure 7: Upset plot for the *Inflammation* hallmark with different linkings to the four GO paper methods.

Figure 3 shows that there is a high amount of genes that are agreed by all methods for one or two hallmarks. But there are also quite some genes that are not linked by all the methods to a hallmark. For those sets of genes UpSet plots were made. The set of genes being mapped once in figure 8 contains amounts that differ much per hallmark. Most of them have expected high amounts like *Growth* and *Signaling* or expected low amounts like *Energetics* and *Immortality* hallmarks. However, there are some interesting features. The amount of *Mutation*, for instance, is much higher than other hallmarks that have equal amounts of agreed genes, see also figure 2. So for *Mutation* much of the genes are not mapped for other hallmark and this is not the case for other similar hallmarks.
In the set of genes that are mapped twice, figure 9, the *Growth* and *Signaling* hallmark together are the most mapped. *Growth* is also connected much with other hallmarks, while other hallmarks have few mappings. In this set the amounts of the different hallmarks are as expected compared with the total amounts of agreed genes. The usage of the *Growth* hallmark by only two GO paper methods can also be seen in these sets as it is the most mapped in every case and with many hallmarks.
In addition to the sets that have genes that are linked by all GO paper methods there were also genes that have no agreement. Instead an UpSet plot was made with this set of genes that shows how the set is divided across the hallmarks, see figure 10. In this case a gene is mapped to a hallmark even if it linked by only one GO paper method. However, while this does not show how many times a gene is linked to a hallmark, it still can indicate how well a hallmark is represented in this set. As visible in figure 10, *Signaling* has the most links, even more than *Growth*. This shows that the genes are easily linked to the *Signaling* hallmark, once or more, but are not agreed by all the GO paper methods. Figure 2 also indicates that a great part of the 793 genes are linked to *Signaling* but that not even half of it is agreed upon by all methods.

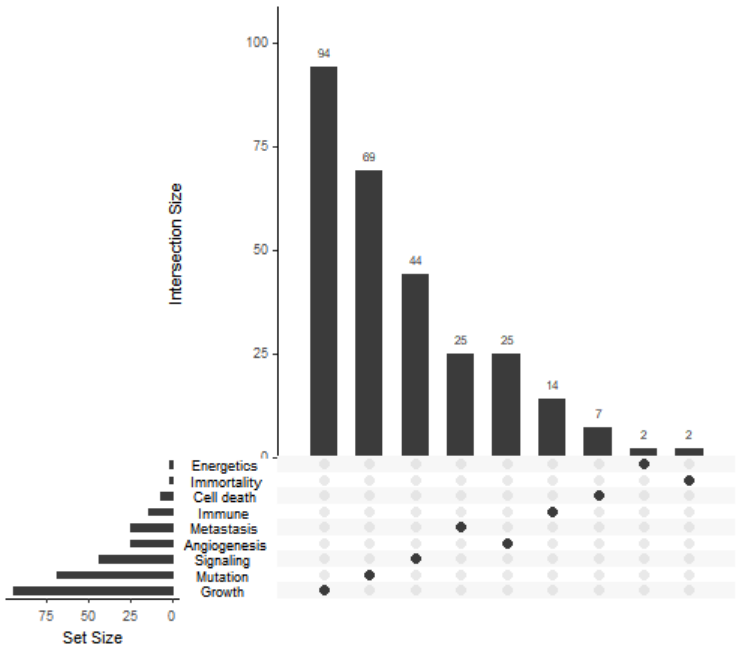

Figure 8: UpSet plot with genes mapped once by the four GO paper methods.

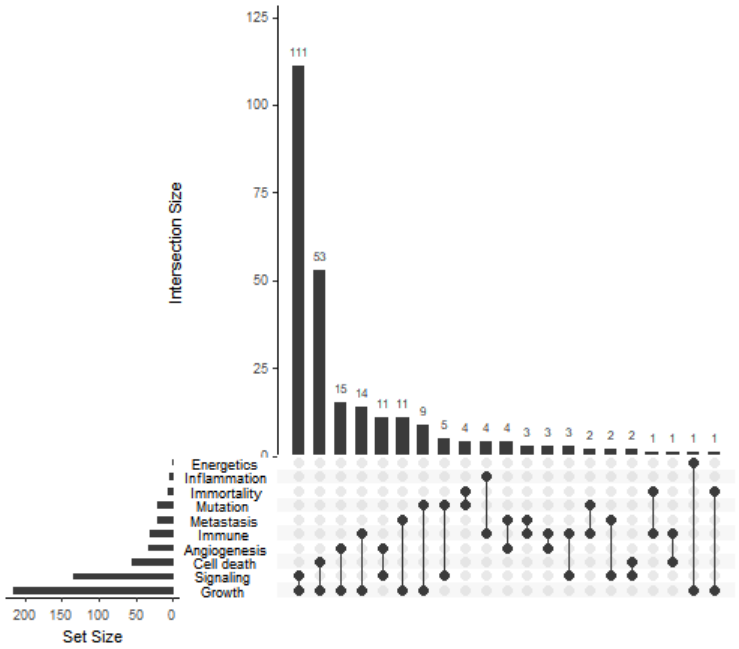

Figure 9: UpSet plot with genes mapped twice by the four GO paper methods.

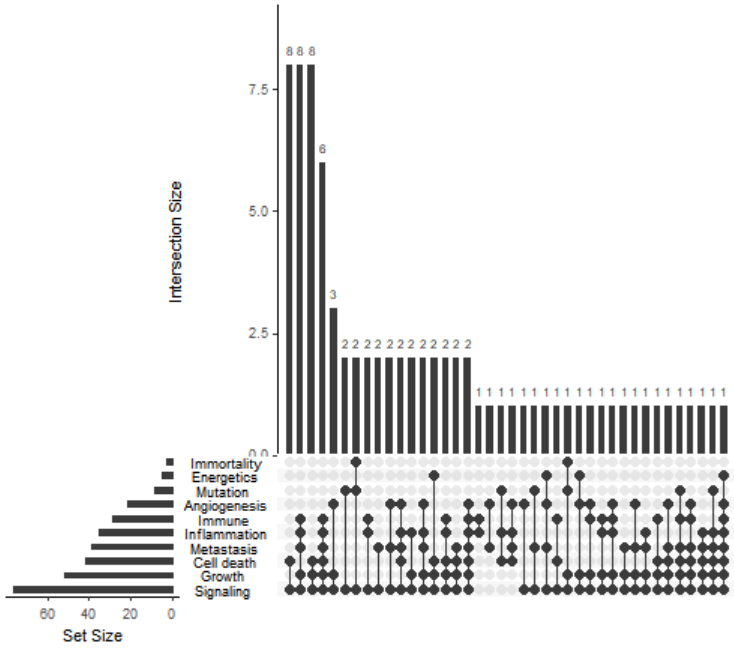

Figure 10: UpSet plot with genes without agreement across the four GO paper methods.

3.2 Functional Enrichment

For the mapping of hallmark genes the GO paper methods made a selection of GO terms for each hallmark. Genes with those GO terms can then be found and eventually can be selected as a hallmark gene for that hallmark. This process resulted in the lists of genes being mapped the hallmarks for every GO paper method. In order to check the mappings done by these paper methods, every hallmark is analysed for the GO terms. For every hallmark an enrichment analysis was performed on the genes that were mapped by all the paper methods to that hallmark, so the resulting genes from the first comparison. With the use of the Panther program over-represented terms were found. Those enriched terms can then be used to associate what kind of genes or processes are present in the set of genes. As described earlier the program only searched for GO terms for biological processes. This is done for all the hallmarks. However, because the amount of genes with agreement by all four GO paper methods for the *Energetics* hallmark was too low, only five genes, no significant results could be obtained with Panther and visualisation of it was then not possible. This can already show that this hallmark may not be used as well as the other hallmarks by the paper methods.

3.3 Visualisation Functional Enrichment results

After finding the enriched GO terms for each hallmark with Panther, visualisation of it was performed using Revigo and CirGO. The Revigo program produced a two-leveled treemap for each case. That was then used to make a circular plot as described earlier. This is done for every hallmark, except for the *Energetics* hallmark as there were no Panther results for it. the Figure 11 shows the plot made for the *Sustaining* hallmark. In it are different sets of GO terms with the processes belonging to the *Signaling* hallmark according to the enrichment analysis. Based

on the given GO terms that visible in figure 11 most are indeed related to signaling or to various regulations where signaling is needed. In this case the variety of GO terms can be expected as signaling is done throughout all the processes. The *Signaling* hallmark has a high amount of agreed genes across all methods but also the other hallmarks with lower amounts show expected results. For example, the *Inflammation* hallmark, figure 12, that has a low amount of agreed genes still show expected GO terms that are linked to it.
In addition to the original results from Panther other CirGO plots were made. These were obtained by looking for the unique GO terms found for a hallmark. This excludes GO terms that are be found in multiple hallmarks and also a lot of generic GO terms that do not point to a specific process. One case of it is visible in figure 13 for the *Signaling* hallmark. Comparing with the original results, the unique GO terms are lower in count but they are more specific. It can show how refined the hallmark is or how the hallmark is overlapping with other hallmarks.

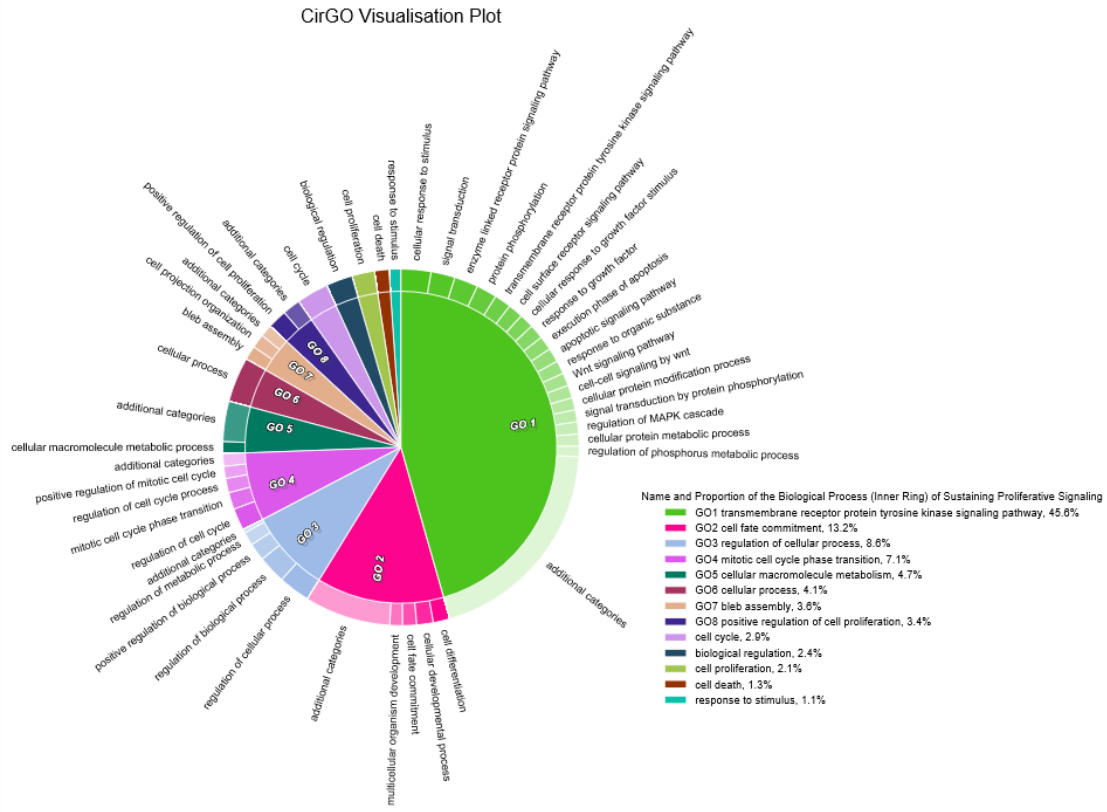

Figure 11: CirGO plot for the *Signaling* hallmark.

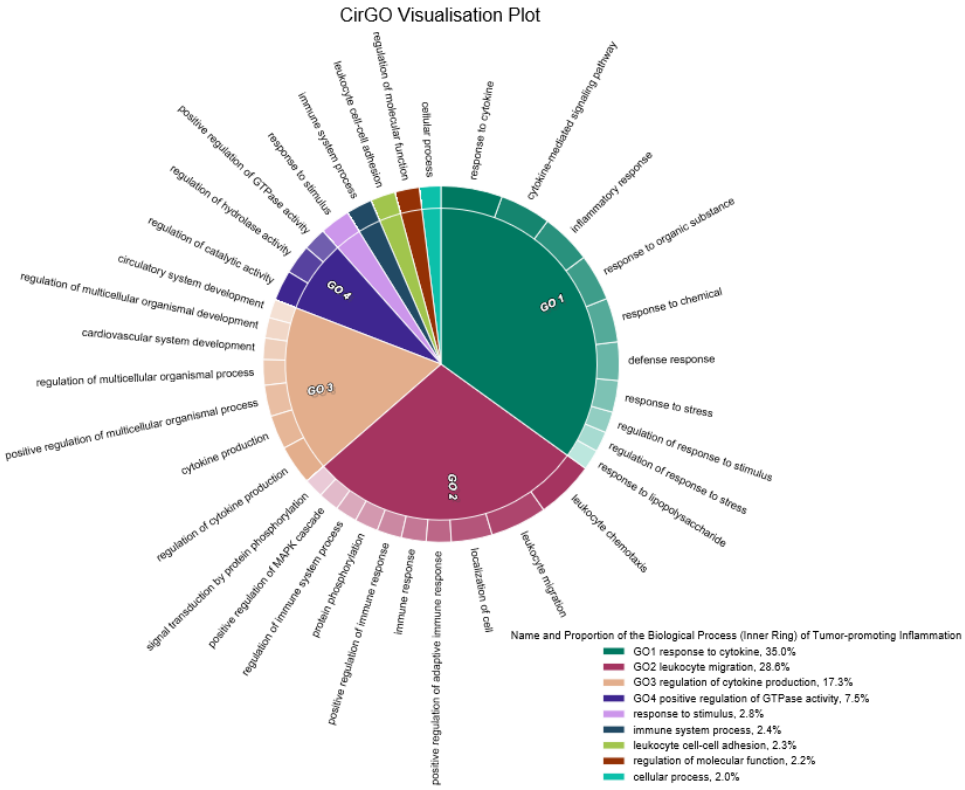

Figure 12: CirGO plot for the *Inflammation* hallmark.

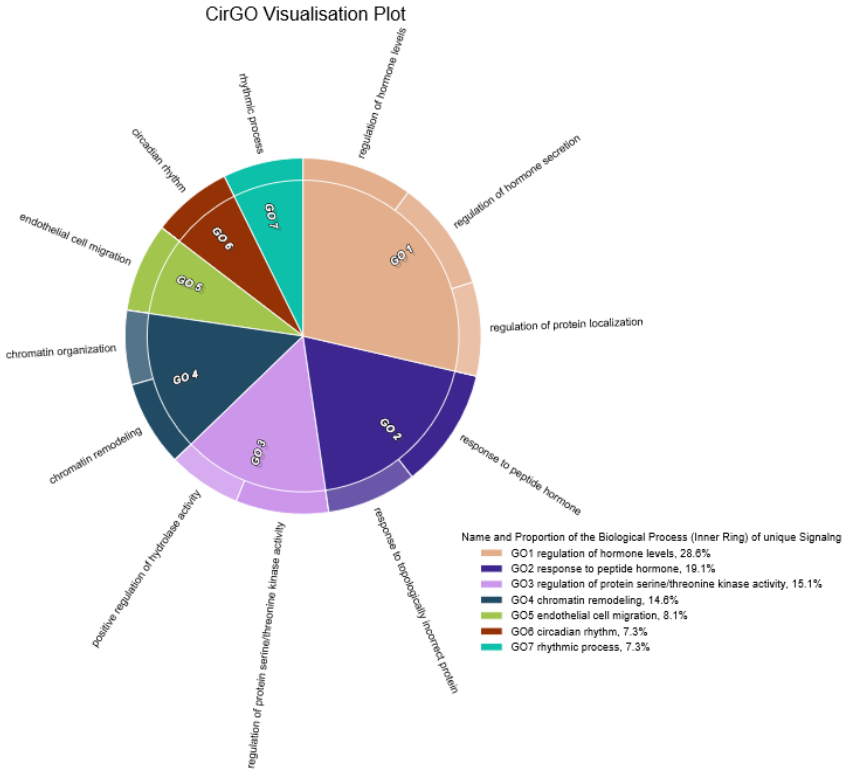

Figure 13: CirGO plot for the *Signaling* hallmark with unique results.

3.4 Pathway sets

The previous results were based on the four GO paper methods. However, the core set of genes from [5] is based on the GO paper methods and on the pathways from [4]. In order to analyse the complete results the pathways were checked on whether the genes were in the core set. The next step was comparing those genes with the linked hallmark genes from the first comparison. The table with the results are visible in figure 14. The pathways are quite diverse and don't always seem to be linked to a specific hallmark. The amounts of genes being in the core set also differs a lot, especially the ones with a high total amount don't always have a high amount in the core set. Many pathways have expected amount for various hallmarks. For example, the immune response pathway has much links with *Immune* and the angiogenesis pathway with *Angiogenesis*. However, some pathways also have other hallmarks linked with high amounts or don't have a clear single high hallmark. This is of course affected by the fact that some hallmarks have high amounts of linked genes and others don't as is seen in the first comparison (see figure 2). *Signaling* and *Growth* both have high amounts in most of the pathways where as *Inflammation* and *Energetics* only have low amounts.

Pathway	Pathway total	In core set	Signaling	Growth	Cell death	Immor	Angio	Metas	Mutation	Inflam	Energetics	Immune
Apoptosis	430	212	84	152	64	4	28	31	16	5	0	38
DNA repair	125	104	24	28	5	9	4	3	89	0	0	7
Cell cycle process	193	117	68	76	3	6	3	6	32	0	0	4
Cell proliferation	512	222	109	158	22	3	44	29	15	6	1	23
Cell cell signaling	404	92	32	53	13	1	24	18	2	6	0	11
Angiogenesis	48	42	15	20	2	1	31	12	0	0	0	2
Immune response	235	69	16	39	13	0	12	11	1	7	1	30
Generation of precursor metabolites and energy	123	21	3	16	7	0	4	3	0	1	2	2
Epithelial to mesenchymal transition	10	7	5	6	2	1	3	2	1	0	0	1
p53 signaling pathway	68	45	27	37	14	3	3	3	7	0	0	0
Wnt signaling pathway	150	64	39	44	13	2	11	9	4	1	1	4
TGF beta signaling pathway	85	50	33	35	9	2	14	12	2	1	1	6
PI3K-AKT signaling pathway	339	152	71	85	32	3	43	38	2	2	2	25
Ras signaling pathway	228	87	39	44	14	2	34	27	1	0	0	13

Figure 14: The pathways with the amounts of genes being in the core set and the amounts linked to a hallmark based on GO papers results.

4 Discussion

The results of this study show that although the hallmarks are represented in the core set of genes, this representation is slim and inconsistent. There are many differences in how the core set of genes was linked to the various hallmarks and paper methods. These differences can be seen in various cases and may be caused in different ways.

Firstly it can be seen in the comparisons. Figure 2 shows that the amounts of mapped genes varies a lot between the hallmarks, even when every link is counted. The agreement rates of the hallmark, figure 2, shows great inconsistency of mapping between the hallmarks and the paper methods. Some hallmarks may have many genes linked to it but only a small part is linked by all the paper methods. Between the older and newer hallmarks the amounts also vary slightly. The

older hallmarks are more linked in total but the agreement rates varies much between the hallmarks. The newer hallmarks are less linked but still has some high agreements.
The inconsistency is shown well in the UpSet plots in figures 4 – 7 between the GO paper methods. It shows that the agreement between the paper methods varies between the hallmarks and that some other combinations of paper methods give higher amount of agreeing genes than all four. The total amount genes linked to a hallmark by the paper also varies. GO4 for example has a low amount of results for many hallmarks. GO4 is from 2012 while the other papers are from 2015 or 2017. This can suggest that the databases probably have had less information back then. It shows a limitation of this study that the paper methods may have used different database versions or other outdated information which could have led to unequal comparisons between the results of the different methods.
Figures 8 – 10 shows well how the genes are mapped to one or more hallmarks and how some hallmarks are more connected than others. The *Signaling* and *Growth* hallmarks are mapped a lot and to many other hallmarks. This is less for other hallmarks.
The results of the CirGO visualisation do show that the mapped genes are related to the expected processes in the hallmarks, especially when using unique results. However, because the GO-SLIM database was used for the enrichment analysis with Panther the results were less specific. This is visible in the CirGO plots as the GO terms are not precise and may not show many specific processes but more generic terms and processes. The differences between the various CirGO plots can be used in further research into the classification of the hallmarks. It can also be useful for searching and showing how the hallmarks are connected or overlapping with each other.
Figure 14 shows how the pathways can be compared with the GO paper results. Most of the pathways are indeed connected to the expected hallmark(s) but there is not always a clear distinction of it. The different amounts of linked genes per hallmark that was visible from the first comparison seem to be visible here as *Signaling* and *Growth* have high amounts in a lot of pathways. This shows that in a pathway some genes may also be involved with other functions than what the whole pathway does. This may be the case with the genes that are linked to *Signaling*, as signaling is done in all sorts of pathways.

The different results show that there is great inconsistency between the mapping of the different hallmarks. Some hallmarks have more genes mapped to it than others or do not have much agreement across several mapping methods. This can suggest that the hallmarks or the used (GO) terms for the hallmarks may need to be more defined to get more distinct results as there is much overlap between the results. However, during this study mainly the results of the four GO paper methods were used. The core set was also based on the results of the Uhlen et al. paper using pathways as mapping method. This was because comparing between the two main method results is difficult. An attempt of this was done here, in section 3.4, which showed that comparing it is difficult and that a specific linking was not always well visible. So further research should be done to compare the core set of genes using all the results more. This can be done by visualising the results in a better way. The results from the different papers also originated from different years with probably slightly different or smaller databases. This could have affected the results as GO4 consistently had lower amounts of mappings to hallmarks. Research with the most recent databases with GO terms and other information can then give better and updated results to compare with. This can be done with the databases that are checked but also with the terms or other info that used to check the databases with. For future studies the GO terms that are used could than be more defined. This

can be done for example with more use of evidence codes, which can help to define the hallmarks more.

As stated the hallmarks are not all represented equally in the core set. Several hallmarks are more visible than others. The newer hallmarks also seem less visible. Between the different GO paper methods there is not always much consensus which also lowers the amount of agreed hallmark genes. The agreed genes do show the processes that the hallmarks represent. So this can give information to use for further research although some hallmarks may be more representable than others depending on how many (agreed) genes are found.

References

[1] Hirsch T. et al. Regeneration of the entire human epidermis using transgenic stem cells. *Nature. Nature Publishing Group*, 551(7680):327–332, 2017. doi: 10.1038/nature24487.

[2] Kiefer J. et al. Abstract 3589: A systematic approach toward gene annotation of the hallmarks of cancer. *Cancer Research*, 77(13 Supplement):3589–3589, 2017. doi: 10.1158/1538-7445.AM2017-3589.

[3] Knijnenburg T.A. et al. A multilevel pan-cancer map links gene mutations to cancer hallmarks. *Chinese Journal of Cancer. BioMed Central*, 34(10):1–11, 2015. doi: 10.1186/s40880-015-0050-6.

[4] Uhlen M. et al. A pathology atlas of the human cancer transcriptome. *Science*, 357(6352), 2017. doi: 10.1126/science.aan250.

[5] Yi Chen et al. Mapping cancer hallmarks to genes: Comparing the impact of different approaches for pancancer analyses.

[6] D Hanahan and Weinberg RA. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000.

[7] D Hanahan and Weinberg RA. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011.

[8] Dustin Ebert Xiaosong Huang Huaiyu Mi, Anushya Muruganujan and Paul D. Thomas. Panther version 14: more genomes, a new panther go-slim and improvements in enrichment analysis tools. *Nucl. Acids Res.*, 2019. doi: 10.1093/nar/gky1038.

[9] Siira S.J. Kuznetsova I, Lugmayr A and et al. Cirgo: an alternative circular way of visualising gene ontology terms. *BMC Bioinformatics 20*, 84, 2019. doi: 10.1186/s12859-019-2671-2.

[10] C. L. Pan M. Plaisier and Baliga N.S. A mirna-regulatory network explains how dysregulated mirnas perturb oncogenic processes across diverse cancers. *Genome research. Cold Spring Harbor Laboratory Press*, 22(11):2301–2314, 2012. doi: 10.1101/gr.133991.111.

[11] Škunca N Supek F, Bošnjak M and Šmuc T. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, 6(7):e21800, 2011. doi: 10.1371/journal.pone.0021800.