Universiteit Leiden
The Netherlands

Opleiding Informatica

The relationship between prognostic genes in human and mouse cancer

Pascal Nuijten

Supervisor:
Dr. K.J. Wolstencroft

BACHELOR THESIS

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

22/5/2020

Abstract

With the ever evolving research techniques and growth of available resources past fundamental research results based on older methods risk to become outdated. Therefore a reconstruction of previous done research can significantly increase knowledge and enhance previous results. For the reconstruction of a part of the results from "A Pathology Atlas of the Human Transcriptome" the original data was processed again with updated insights and tools. Significant differences compared to the original graphs can be observed throughout this thesis. This thesis explores the value of data reuse and reanalysis in a swiftly moving field of science.

Contents

Chapter 1

Introduction

The ancient fight against one of the biggest diseases mankind knows, cancer, is yet not won. Although the cancer death rate in America declined by 1.6 % (2006-2015) still 606.880 people were victim of this disease. [1] This makes cancer still the second cause of death worldwide, right after heart disease. [2] Data shows that back in 2008-2014 only 65% of the patients diagnosed with cancer survived longer then 5 years. [1] These numbers are not very favorable especially when considering that cures for cancer are already studied for over several centuries. To find a solid cure for cancer the importance of understanding the way cancer is spreading and developing is big. In normal circumstances cell growth is limited by certain control mechanisms. But cancer succeeds in escaping this mechanism and resulting in a tumor by mutation in the so called oncogenes. [3] This pathway is why cancer is a system level disease, meaning that it disturbs regular cellular processes in a cell. [4] Most likely several somatic mutations have occurred when a tumor is induced. [3] These mutations are on DNA level and can be caused by tobacco, alcohol, food, UV radiation or yet unknown causes. [5][6] A human genome experiences 105 mutations a day, that normally would be recovered by different DNA repair systems. [6][7] All these different factors and aspects are making the understanding and therefore also treatment of cancer difficult. Right now the most commonly used way of collecting information and distinguishing a cancer type is by examining the tissue and cell type origin of the cancer. This still is a reliable way of retrieving some low dimensional information about a cancer. But since cancer in fact is several mutations in a cells DNA that causes unusual division to occur, there is way more valuable information extant at a molecular level. There exist numerous of different tumors that share interesting findings at this molecular level. This relatively new and growing type of looking at tumors is called, Pan-Caner analysis. [8] [9]

1.1 Related Work

1.1.1 Pan-Cancer research

There exist numerous methods to research cancer. A big advantage of Pan-Cancer analysis relative to other methods lies in the ability to obtain a better understanding of universal processes that drive tumour growth. Pan-cancer analysis is characterized by the use and comparison of data from a set of different cancer types instead of one. The identification of common as well as different genes present in diverse cancer types provide new insights that help in the understanding of cancer as a process. Pan-Cancer research demands big databases with molecular and clinical data. This relatively new vision on cancer research made room for large-scale studies such as The Cancer Genome Atlas (TCGA) [10] and the Pan Cancer atlas [11]. Databases like these are used frequently by various Pan-Cancer projects, a recent and renowned example is the paper : "A pathology atlas of the human cancer transcription" [12] which was published in the extensive journal *science*. In this article clinical as well as molecular data is combined to create a so called pathology atlas. The obtained data is retrieved from TCGA database [10] and the Human Protein Atlas [13] which are two big open access data collections. The combination of these databases resulted in the data of

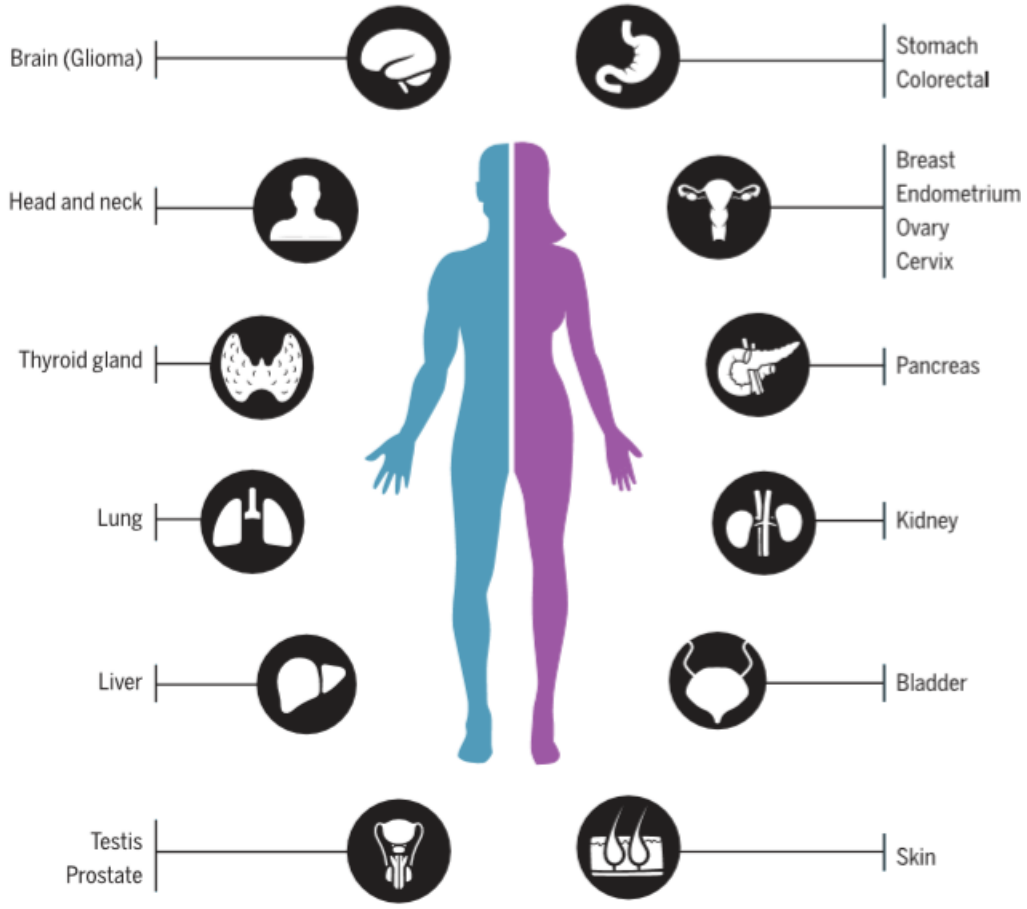

Figure 1.1: A schemetic overview of the 17 different cancers used by the paper : "A pathology atlas of the human cancer transcriptome". Source: "A pathology atlas of the human cancer transcriptome" [12]

8000 different patients spread over 17 different major cancer types (Figure 1.1).The clinical outcome of these patients is observed and linked to the expression levels of certain genes. When a correlation between a gene and the clinical outcome is found, a prognostic is established. A prognostic gene can either be favorable or unfavorable meaning that the patient has a good or bad survivability, respectively. One of the aims in the article is to create an atlas that allows an easy lookup for the prognostic role of protein coding genes in 17 different cancers (Figure 1.1). With the use of this newly created atlas the analysts found that the up-regulation of genes that are involved in the mitosis of a cell and the down regulation of genes involved in cellular differentiation both can be associated with a shorter patient survival. [12]

1.1.2 The Gene Ontology (GO)

The amount of identified genes linked to their functions is gaining momentum through the rapid progress in sequencing techniques such as nano-pore sequencing. [14] This gain of valuable data has a need for a well organized and structured way of managing. One of the big projects that succeeded in doing so is called the Gene Ontology, also known as GO. [15] [16] GO is an ontology that is assembled as a directed acyclic graph. This graph is showing the biological system of any eukaryote organism and can be branched into separate domains of gene functions: Biological Process, Molecular Function and Cellular Component. The nodes in this graph consist of so-called GO-terms and range from organism to molecular level. A GO-term contains a logical name and is linked to an actual biological process. These terms are connected in the acyclic graph by relationships that originate from nature. The Gene Ontology was originally based on the Flybase database [17], the mouse genome database [18], the Gene Expression Database for the Laboratory Mouse [19] and Saccharomyces Genome Database [20]. However in the last decade GO grew due to 28 new collaborations with projects such as GeneDB [21], InterPro [22] and Ensembl [23] contributed their annotations to the Gene Ontology. With the continuously growing identification of genes the use of a controlled vocabulary such as GO is considered as extremely useful. Not only does GO provide this huge network that represents a network of known processes and functions in eukaryotic cells, it also is a very helpful tool for researchers to screen their genes with. After running a GO-analysis an outline of a set of genes can be determined. [24] This features make GO a powerful and indispensable tool for bioinformaticians.

1.1.3 GO enrichment analysis

As mentioned in the previous section the Gene Ontology provides this powerful identification tool for gene sets from various organisms. There exist several GO "mapping" tools such as PANTHER [25], gProfiler [26], BiNGO [27] and GOrilla [28]. These tools use their own slightly differing way of calculating the P-value of over- or underrepresented genes in the specified gene set. This P-value indicates the probability that x number of genes out of the total annotated genes in a certain GO-term, given the proportion of the total genome are annotated to the GO-term. Due to this, a lower P-value is related to a more significant GO-term. With the use of a cut-off value a Gene Ontology profile can be determined. This profile gives a good sense of the aimed gene set.

1.1.4 The integration of more data and knowledge from other similar species

In genetics some organisms such as *Saccharomyces cerevisiae* (Baker's yeast), *Drosophila melanogaster* (Fruit fly), *Zea mays* (Corn) and *Mus musculus* (Mouse) are studied more in depth then others. [29] The reasons of preferring these organisms over others are mostly the convenience of fast breeding and the ethics of using them. Such organisms are also referred to as model organisms. Using such species can have benefits, mainly because data of organisms of interest often lack valuable information. The shared ancestry between genes in alternative model organisms is also called orthology. A good example organism for the use of such orthologues is the *homo sapien*, better known as the human. However big projects such as The Human Genome Project [30] have created a clearer picture of the human genome and its genes this yet only is a small part of the totality of all genes. Therefore a big gain information can be achieved by using a specie such as the mouse. Mice and human genes have the same mammals and therefore a lot of shared genes. A major advantage of using the mouse is that the studies on this specie are lot more common. The main reason for this is the clash with human ethics. [31] The method of using model organisms is used frequently and provides more information about genes and makes tests for the application of diseases easier. [32] [33] Using orthologue data can provide new insights about previously done assumptions or even expand existing data and associated conclusions.

1.2 Problem statement and research question

Because bioinformatics is such a rapidly evolving field the importance of keeping your work updated is big. Every day new research possibly provides new techniques, information or data. Therefore research that is repeated with the latest insights can result in different outcomes. Another aspect for interfering results are the tools that have been used for the concerned research. In bioinformatics there exist a lot of useful tools that are able to accomplish the same intended goal. In some cases tools will conclude the same outcome, however it is also very possible that tools with varying approaches will lead to differing results. Therefore results build on certain tools could differ from the results found when using a similar tool. This thesis is derived from the paper: "A pathology atlas of the human cancer transcriptome" [12] and its main objective is to verify and complement the GO-analysis done in the paper by using homologue species and the most recent version of GO. This thesis is aiming to answer to the following research question: "Can we enhance previous results from large-scale cancer omics by the reanalysis of the same data set with an updated version of the Gene Ontology and with the use of data from related model organisms?"

1.3 Summary

Cancer is still one of the biggest diseases worldwide, and yet there does not exist a solid guaranteed cure. This explains why cancer is such a hot topic for researchers currently. For decades typical cancer research implied examining a specific tumor type on different scopes. But lately cancer research has been expanded and tumors are examined in various ways. One of these approaches stands out and is called pan-cancer research. [8] [9] Pan-cancer analysis differs from regular cancer research by comparing (meta)data from patients with various cancer types. Instead of only looking intratumoral also extratumoral data is taken into account which can create new insights. A good example of a recent Pan-cancer research is the article: "A pathology atlas of the human cancer transcriptome". [12] In this paper bioinformatics experts deal with clinical data from 8000 patients to create a map that helps predicting whether a gene has a positive or negative effect on the clinical outcome of a patient. The data was collected from the big databases: the Human Protein Atlas [13] and TCGA database [10]. By mapping genes to the Gene Ontology hierarchy [15] [16] a map of genes with their clinical outcome can be linked to their biological process. In order to succeed the researchers used certain methods and versions of tools that already ran out of date. In this thesis the main focus is on recreating some of the results of the paper by using maintained tools and up to date data. In addition to the paper this thesis is introducing the use of *Mus musculus* (mouse) genes which share homology with humans and therefore can create a increasement of data.

1.4 Overview

This thesis is a end product of the bachelor degree Bioinformatics at the university of Leiden. This project is supervised by Dr. K. J. Wolstencroft and part of the Leiden Institute of Advanced Computer Science (LIACS). This thesis contains the following chapters :

- **Chapter 1**, covers the introduction, the problem statement and the research question of the thesis.
- **Chapter 2**, this chapter provides information about the used tools and explains the approaches for this thesis.
- **Chapter 3**, presents and discusses the results published by this thesis.
- **Chapter 4**, wraps up this thesis by providing the conclusions and feasible further additions to this work.

Chapter 2

Materials & Methods

2.1 Materials

In this section a brief explanation of the tools that were applied in order to accomplish this thesis are provided.

2.1.1 Cytoscape

The visualization of biochemical reactions and gene transcriptions play a critical role in the understanding of system biology. [35] A widely known tool for this kind of visualizations is Cytoscape. [34] Cytoscape is a open source java build tool that allows users to create network graphs of molecular species and their intermolecular interactions. With plenty of possible environments and settings Cytoscape forms a powerful tool and is directly usable with the Gene Ontology [15], which is a directed acyclic graph. Over the years various Cytoscape plug-in modules were established. These modules extend the core functionality of Cytoscape in terms of biological semantics, additional network analyses and new algorithms. A, for this thesis, repeatedly used Cytsocape plug-in method is BiNGO [27].

2.1.2 BiNGO

BiNGO [27] is, as mentioned earlier, a Cytoscape plug-in method. Identical to Cytoscape it is a open-source Java tool. The developers created BiNGO as a response to the Gene Ontology [15] with the aim to perform a functional enrichment analysis with already existing or new Cytoscape nodes. Although the tool originates from 2005 the interface allows you to setup your own reference set, ontology file organism/annotations and the statistic test and its significance level. Statistics is applicable on everything around us. In the biology statistics has been given an indispensable role. In BiNGO the GO analysis is performed with the hyper-geometric test [36]. This thesis exclusively uses the BiNGO in combination with the hypergeometric test. The hyper-geometric test is a way to identify whether a population is over- or under-represented in a sample. BiNGO uses this test to answer the following question: What is the probability that x genes (obtained from the test set) belong to the same GO-term as n genes (obtained from the reference set). The hypergeometric test can be expressed as the following formula:

$$h(x; N, n, k) = \frac{[{}_kC_x] \cdot [{}_{N-k}C_{n-x}]}{[{}_NC_n]}$$

Where :

- x = The amount of genes in the input list that are classified as the target GO-term.
- N = The amount of all existing genes (in the reference set).
- n = The amount of genes in the in input list.
- k = The amount of genes in the reference set that are classified as the target GO-term.
- ${}_kC_x$ = Represents the possible combinations of k things taken x at the time.

While walking down the GO-hierarchy, BiNGO uses the formula presented above to confirm the GO-terms with the specified P-value. A GO-term is rejected when the P-value is above the defined threshold.

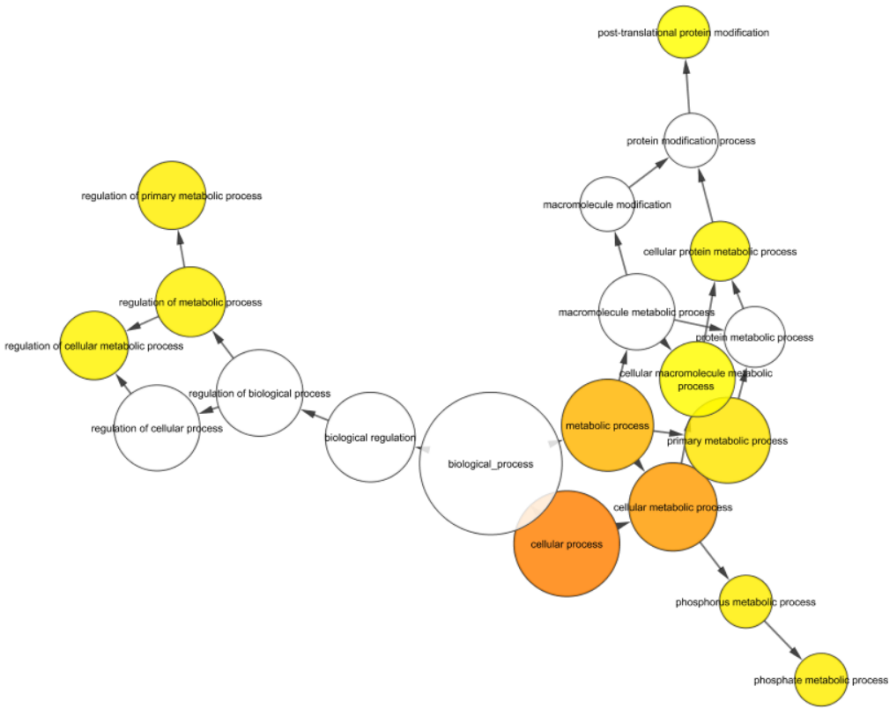

Figure 2.1: The Cytoscape representation of a BiNGO Gene Ontology enrichment of unfavorable lung cancer genes.

2.1.3 Plot.ly

The reconstructions of the aimed graphs that are shown in "A Pathology Atlas of the Human Cancer Transcriptome" [12] were made using the online tool Plot.ly [37]. Plot.ly provides a powerful online environment where data scientists can create and store their visualizations. The created graphs are still public accessible through: https://plot.ly/~PascalNuijten/. Every figure will contain a link to the original Plot.ly web environment.

2.1.4 The data

The goal of this thesis is to update and then reconstruct certain experiments from the paper "A Pathology Atlas of the Human Cancer Transcriptome" [12]. Therefore the data was acquired solely from the paper itself. As mentioned earlier the paper used genes obtained from 8000 different patients spread over 17 different major cancer types. This data is public and downloadable at: www.sciencemag.org. Table S6 from the supplementary tables contains the protein-coding genes that resulted from the survival analysis of the 17 major cancer types. For these genes the EnsemblID, Symbols, Mean Expression, Sample Numbers, Expression Cutoffs, Log-rank P and their Prognostic type were stored. The analysts of the paper performed the Gene Ontology enrichment analysis on all these genes with a cut-off of P = 0.001, also found in supplementary table S9. The results were then visualized in a bubble plot (figure 2.2). The x axes shows whether the enriched genes have a positive (favorable) or negative (unfavorable) effect on the clinical outcome. The y axes show the amount of different cancers where the GO-term was enriched.

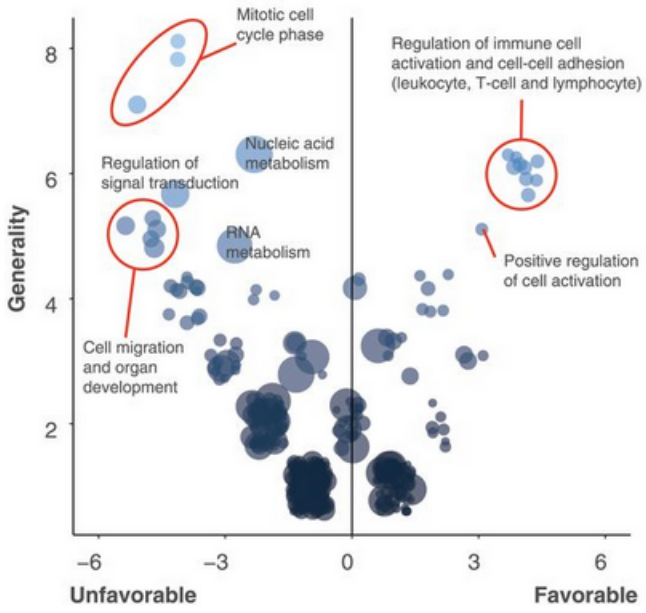

Figure 2.2: The common enriched Gene Ontology functions from 17 major cancer types obtained from the paper "A Pathology Atlas of the Human Cancer Transcriptome" [12].

2.1.5 Cosmic Data

The Catalogue Of Somatic Mutations in Cancer (COSMIC) [38] is a large data set that assists in the identification of the effects of somatic mutations in human cancer. In this collection genes are labeled with gene symbols, genome location, hallmark, tumor types, and many more. The COSMIC data is collected by scientific literature (over 26 000 publications) and the data sets is open accessible and easy to download.

2.1.6 The Orthologue search

As mentioned before a big gain of knowledge about a certain gene set can be achieved by the use of orthologue genes. The mapping of orthologue species can be done in various ways. Two different methods for orthologue search were observed during the research process of this thesis : BLASTp [40] and [41]. BioMart is a web-application which is released by Ensembl [43] and targets to unify biomedical databases. BioMart comes with a homologue search tool where the identifiers corresponded with the identifiers used in the BiNGO tool. With the homologue search tool the data of homologue genes is collected trough various genomic databases. When observing the homologue results a small percentage ($\tilde{5}\%$)of the homologue genes obtained a rather low mapping score (70%). To attempt and receive better homologue matches an other homologue mapping tool named BLASTp was introduced. BLASTp is released in 1990 by the National Center for Biotechnology Information (NCBI) [42] and refers to Basic Local Alignment Search Tool protein. Users can run an alignment between a manually updated reference set and their genes of interest. Although the BLASTp results looked promising the identification of the potential orthologues caused difficulties. The resulting identifiers of the orthologue genes were not translatable to the identifiers used by BiNGO.

2.1.7 Python

In order to process and manage the data multiple scripts were written in the programming language Python 3 [44]. This high-level programming language allows its users to edit and manipulate all sorts of files and data. The choice of using Python as the primary code language for this project is based on the large amount of available extensions and its high code versus results ratio. Python exists since 1991 and is a frequent used tool for bioinformaticians and data scientists. Therefore a lot of good documentation and forums are present at this time. For this project the newest version of Python (3.7.6) available was used.

2.2 Methods

The work flow of this thesis is illustrated in figure 2.3. All the processes are sorted in chronological order and explained in the corresponding subsections below. In the next chapter the matching results are presented and clarified.

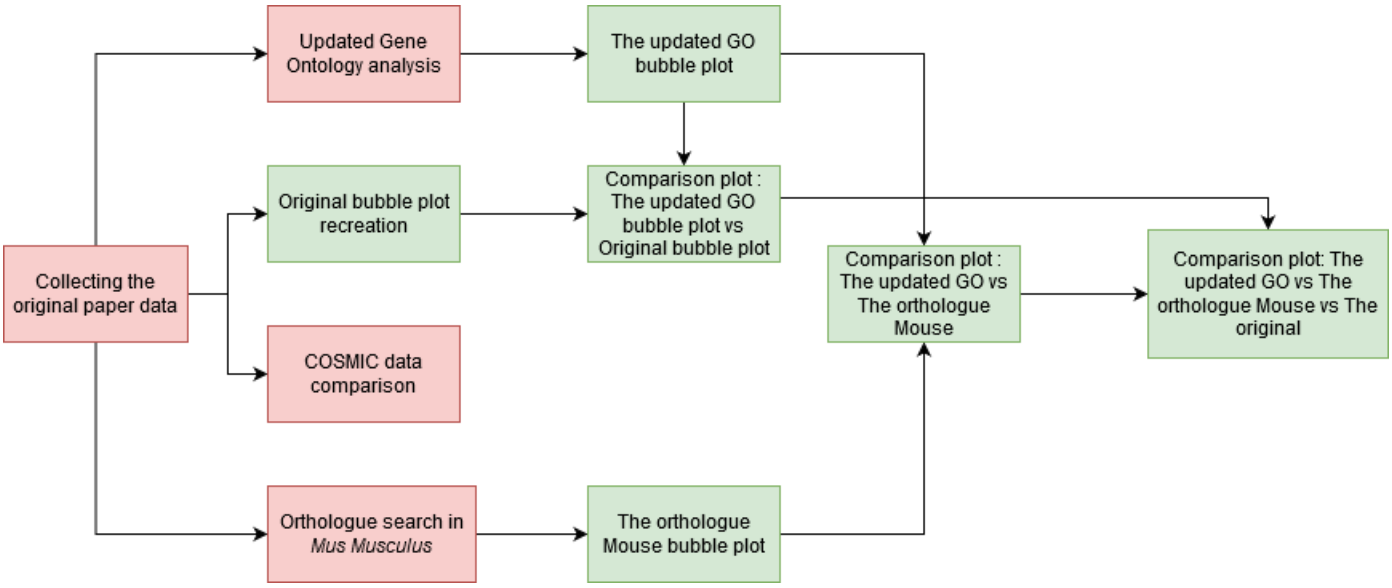

Figure 2.3: The thesis workflow illustrated in draw.io as a diagram. Red and green tiles represent the processing and visualization steps respectively.

2.2.1 Collecting the original paper data

The first step for this thesis was to acquire the required data. All data is exclusively obtained from the paper "A pathology Atlas for the Human Transcriptome"[12]. In the Figures & Data section the researchers provide the supplementary materials used for the paper. The 22 tables (S1-S21) found in this download folder are also chronologically introduced in the paper. For this thesis table S6 and S9 are the most interesting ones. Table S6 offers the expression cut-off for the genes received from the 17 major cancer types. The identification of the prognostic genes was done using this expression cut-off. The data in Table s6 is marked with the following tags:

- EnsemblIDs : An identifier used by Ensembl [43] to label genes.
- Symbols : The official gene symbol.
- Mean Expression : The observed expression level of the gene.
- Sample Number : This number identifies the cancer type in dispute.
- Expression Cutoffs : Subset of all RNA expressions values.
- Log-rank P : The degree to which the gene is (favorable or unfavorable) differentially expressed

- Prognostic Types : Represents the clinical results by being either Favorable or Unfavorable. Favorable genes are associated to a longer survival rate where unfavorable genes relate to a short patient survival rate

The data available in table S6 operates as the fundament for this thesis. Table S9 shows the statistics of the actual Gene Ontology enrichment analysis done by the paper. This table is used for the original bubble plot (figure 2.2) and hence interesting to be used as a reference for this thesis.

2.2.2 Original bubble plot recreation

For the comparison of the new obtained results it was important to stay as close to the original plot as possible. In the original paper no workflow or information about what tool was used to establish the bubble plot (figure 2.2). Therefore Plot.ly [37] was used to create all charts presented in this thesis. Considering that the original used data for the bubble plot in figure 2.2 was still available in supplementary table S9 the original plot was recreated using Plot.ly. Exactly as in the original plot the x-axis represents the wether the GO-term is favorable or unfavorable (positive or negative). The y-axis shows in how many (out of the 17) cancer types the GO-term was annotated. The bubble size visualizes for every GO-term the number of enriched genes, bigger bubbles contain more genes.

2.2.3 Updated Gene Ontology analysis

Where the original bubble plot uses DAVID (no version specified) [45] [46] to perform the Gene Ontology enrichment analysis this thesis explicitly chooses to use BiNGO [27]. Every version of DAVID uses an old version of the Gene Ontology to perform the enrichment analysis. For example, the latest version of DAVID (version 6.8 released in Oct. 2016) uses the Gene Ontology dated from Jan. 2010. To put this in perspective, only in the last 16 months (Sept. 2018 - Jan. 2020) the gene ontology database grew with 121,135 genes and 149437 annotations. An information gain of this size can not be neglected and hence the most recent version of GO should be applied. The interface of DAVID does not allow to specify the used GO version, therefore BiNGO was used. This Cytoscape plug-in performs a Gene Ontology Enrichment Analysis on a custom submitted GO version. The following parameters for the BiNGO analysis were used :

- Input : For every cancer and prognostic type available in supplementary table S6 a GO enrichment analysis was performed.
- Over- or under representation : To find a correlation between genes and cancer only the over represented genes are interesting.
- Statistical test : The Hypergeometric test. [36]
- Significance level : Two significance levels were tested : 0,001 and 0,05.
- Categories to be visualized : Over represented categories after correction.
- Reference set : The whole annotation is used as a reference set

- Ontology file : This step separates BiNGO from DAVID. Here a manual downloaded Gene Ontology can be inserted. Ontology files are monthly updated and accessible at : geneontology.org/docs/download-ontology/. For this thesis the Gene Ontology version from June 2019 is used.
- Name space : biological_process is the only interesting category for the the aiming results.
- Organism/annotation : This is the gene data set that is used as a reference. The data set is organism specific and also updated monthly by GO. It is accessible at : current.geneontology.org/products/pages/downloads.html. For this thesis the *homo sapiens* annotations for 2019-06-01 release were used.

2.2.4 COSMIC data comparison

This step mainly was conducted to obtain a better sense of the provided paper data. The Catalogue Of Semantic Mutations In Cancer (COSMIC) [38] is a database with genes and their impact in human cancer. For this thesis the Cancer Gene Census (CGC) was used. [39] The CGC is a list of genes with mutations that are causally implicated in human cancer. With the help of a custom written Python (3.7.6) [44] script all genes from the COSMIC CGC data table (available at : https://cancer.sanger.ac.uk/census) were matched with the annotated genes from the previous step

2.2.5 A reconstruction of the Uhlen et al GO enrichment plot

After the gene ontology enrichment analysis it was relevant to put the results in perspective with the reference graph acquired from the paper "A pathology atlas from the human transcriptome". To make a one on one comparison it is essential to use the same parameters and calculations. In the original plot the following parameters were determined:

- X-axis (Generality) $= \sum_{i=1}^{n}(N_{fav,i} + N_{unf,i})$
- Y-axis (Directionality) $= \sum_{i=1}^{n} N_{fav,i} - \sum_{i=1}^{n} N_{unfav,i}$
- Bubble size = The total amount of annotations to the GO-term.

i represents the individual cancer types and n is the total number of cancer types, which is 17. $N_{fav,i}$ and $N_{fav,i}$ are either 1 or 0 depending on the genes prognostic type (favorable or unfavorable). High level GO-terms are very global and do not add much information. Examples of these terms are : immune system process, cellular process, localization, cell killing, and many more. Therefore the first 3 layers of the Gene ontology hierarchy were discarded.

2.2.6 Comparison plot: The Original versus the updated GO bubble plot

Comparing the original with the new bubble plot side by side already uncovers a lot of variations. However the real differences can be spotted when these two plots are joint together. In this step the

goal was to create one bubble plot containing both the "old" as well as the "new" Gene Ontology enrichment data.

2.2.7 Orthologue search in *Mus Musculus*

As mentioned earlier the best method, for this thesis, for matching orthologue genes is the use of BiomMart [40]. For this process the following steps were processed :

1. Choose Ensembl Genes 99 as database and Human genes as data set at : https://www.ensembl.org/biomart/martview.
2. Add the Gene stable IDs for every cancer (favorable and unfavorable separated) as external reference IDs in the Filters function.
3. Choose the Mouse gene stable ID and Mouse gene name as Homologues in the Attributes sections (the following parameter gives an idea of how well the orthologue is matched :%id target gene identical to the mouse gene).
4. After processing a selection between different possible output files can be made.

With this steps for every 17 cancer types two lists with homologue cancer genes can be obtained, one favorable and one unfavorable.

2.2.8 The orthologue mouse bubble plot

After collecting the orthologue mouse data the next challenge was to visualize them. The same calculations as in the original bubble plot were repeated. Equally as in the updated GO bubble plot the first three levels of the Gene Ontology were deleted. To obtain an organized and good representation of the data as an addition the GO-terms annotated less then 150 times were deleted.

2.2.9 Comparison plot: The updated GO versus The orthologue mouse

Identical to the established comparison plot for the original plot and the updated GO plot the same ratio's and parameters were preserved.

2.2.10 Comparison plot: The updated GO versus the orthologue Mouse versus the original

To obtain a good overview of the data difference a combined bubble plot containing the original, the updated GO and the orthologue mouse data was established. The same parameters and calculations were used to derive all the bubble plots, therefore a fair comparison between them can be made.

Chapter 3

Results

In this chapter the corresponding results of this thesis are presented and discussed. All sections in the previous chapter that have led to results are reviewed in the following chronological sections.

3.1 The recreation of the Uhlen et al GO Enrichment Bubble Plot

To stay as close to the original bubble plot (figure 2.2) as possible the precise same clusters and bubbles are tagged in figure 3.1. When observing the two bubble plots one thing instantly stands out, the difference in cluster formation. Where the original plot shows three bubbles in the "mitotic cell cycle phase" cluster (in the top left) the recreated plot only shows two. Although only two GO-terms are visible there actually stack two bubbles on top of each other. By looking at the data a graph equal to figure 3.1 would be expected. But perhaps the authors of the paper tried to create a better visual effect by moving stacked bubbles slightly. Furthermore the graph ended up with exact the same amount of bubbles, bubble locations and bubble sizes.

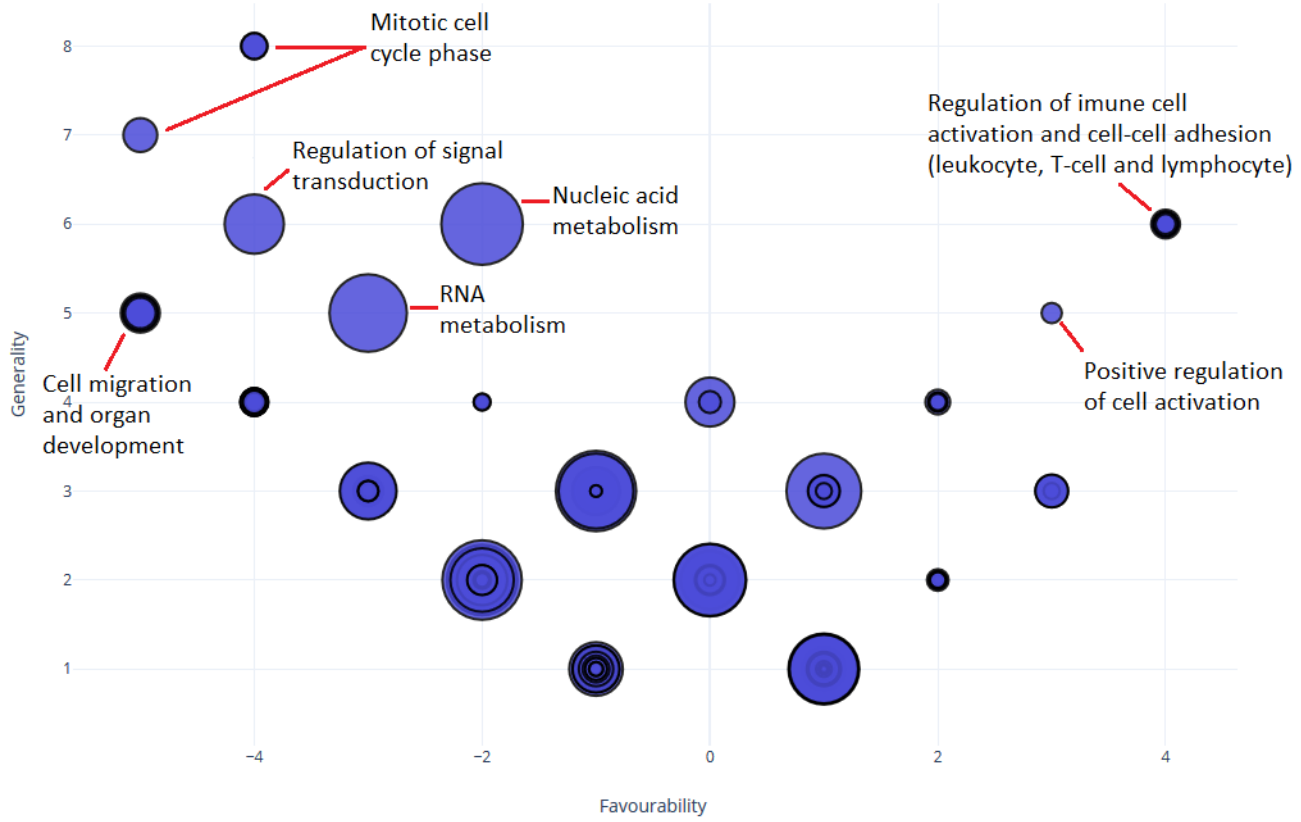

Figure 3.1: The recreation of the original bubble plot (fig 2.2) with the original data in Plot.ly. The axes remain identical to the original plot, therefore the x axis indicates wether the GO-term is annotated in more favorable or unfavorable runes. The Y axis shows in how many of the 17 major cancer types the GO-term is annotated. The bubble size represents the total annotations in the 17 cancers combined. Thicker bands or bands within other circles indicate more GO-terms stacking at the same location. Plot is online available at: https://plot.ly/~PascalNuijten/30

3.2 The reconstruction of the Uhlen et al GO enrichment Bubble Plot

The first objective was to determine the right P-value. Hence two value's were tested: 0.001 and 0.05. For every individual cancer type the total enriched GO-terms were recorded and presented in Figure 3.2. As the graph shows, three cancer types do not annotated any GO-terms at P-value = 0.001. But even when increasing this value only stomach cancer starts annotating GO-terms. One of the aims in this thesis is to increase data, therefore the p-value of 0.05 would fit better. However a big downfall of this relatively high P-value is the reduction of significance. In Figure 3.3 the Cytoscape networks created by the BiNGO Gene Ontology analysis with different p-values are compared for liver cancer in Figure 3.3. The comparison in Figure 3.3 shows a significant gain of GO-terms. However, the core of the graph remains the same and only small (white) bubbles were attached at the border of the graph. These bubbles are more specific GO-terms but also matched with less certainty. Therefore to preserve reliable results and keep as close to the paper as possible the identical P-value of 0.001 was chosen.

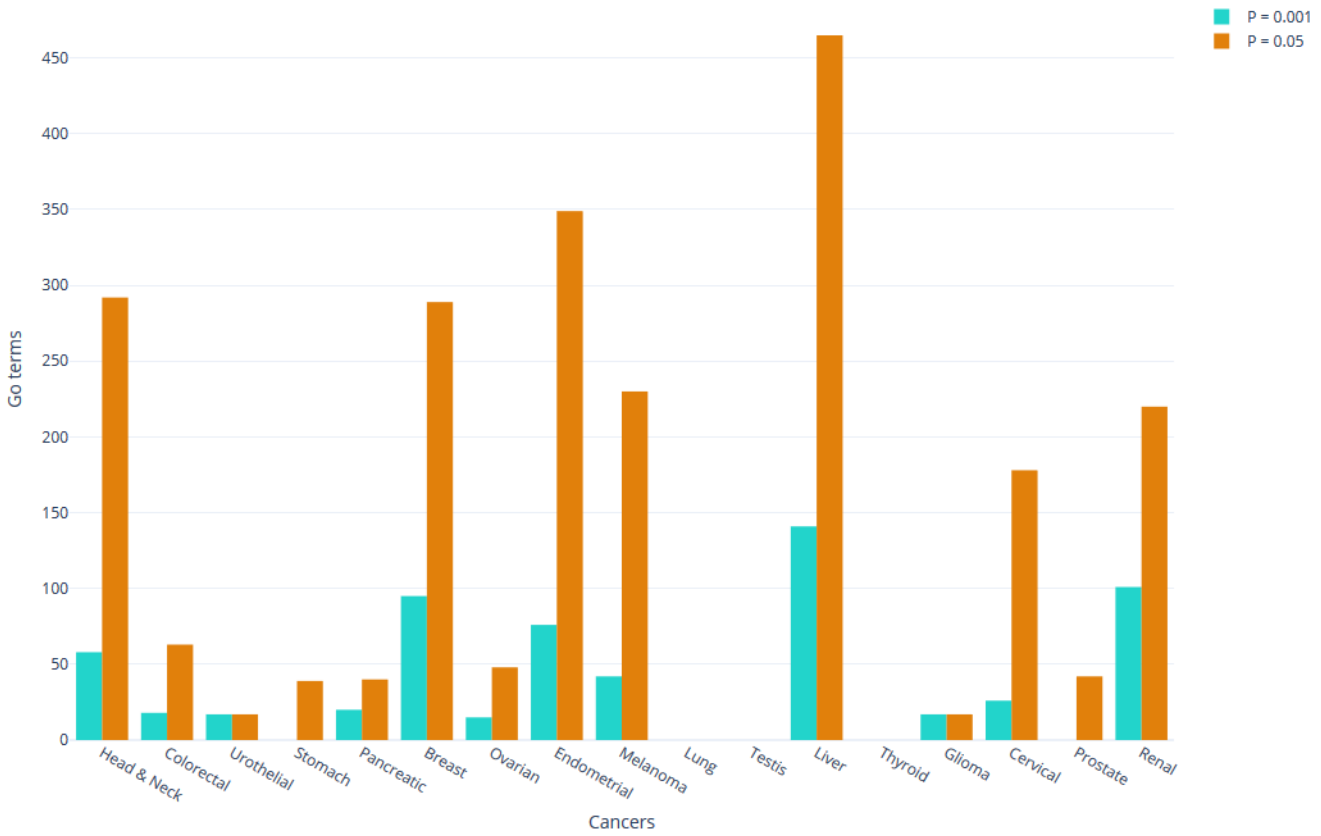

Figure 3.2: Difference in enriched GO-term for all major favorable cancer types used in "A Pathology Atlas of the Human Cancer Transcriptome"[12] with P = 0.001 and P = 0.05. The chart is also visible at: https://plot.ly/~PascalNuijten/36/

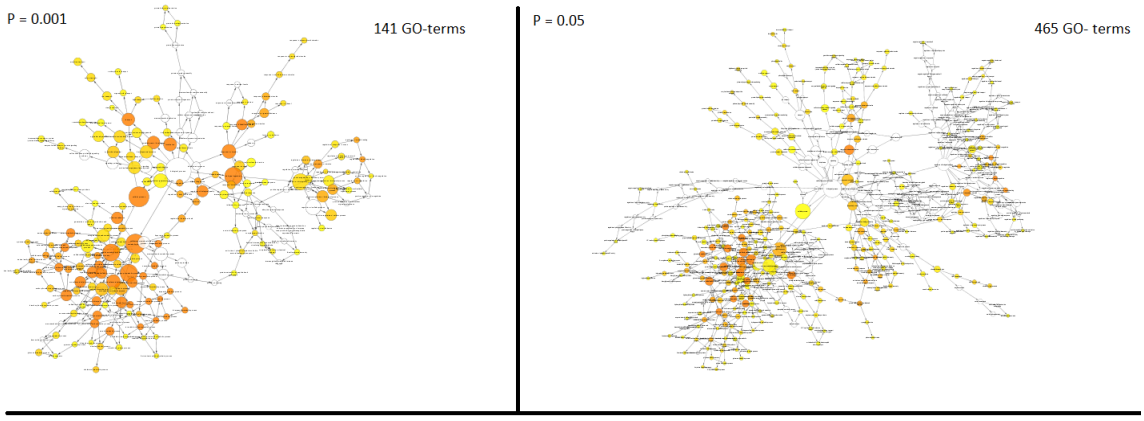

Figure 3.3: The BiNGO GO enrichment analysis for favorable genes in Liver cancer with different P-values. P-value 0.001 contains 141 GO-terms (left) and P-value 0.05 contains 465 GO-terms (right).

3.3 A reconstruction of the Uhlen et al GO Enrichment Bubble Plot

The acquired GO-terms with the GO enrichment analysis from the previous section were placed in to perspective by the establishment of an identical bubble plot. For the bubble plot in Figure 3.4 the same calculations and scales as the original plot were used. The most interesting and important GO-terms are labeled and linked to their bubbles. Furthermore three clusters were marked (left to right) : Cellular component organization, Protein containing complex subunit and Leukocyte differentiation. The establishment of such a cluster is done with the help of QuickGO [47]. This web-based tool visualizes selected GO-terms as a hierarchy. For all bubbles that appear close to each other such a hierarchy was created to verify wether a cluster can be determined. An hierarchy example of the cluster Cellular Component Organization (left cluster in Figure 3.4) is shown in Figure 3.5.

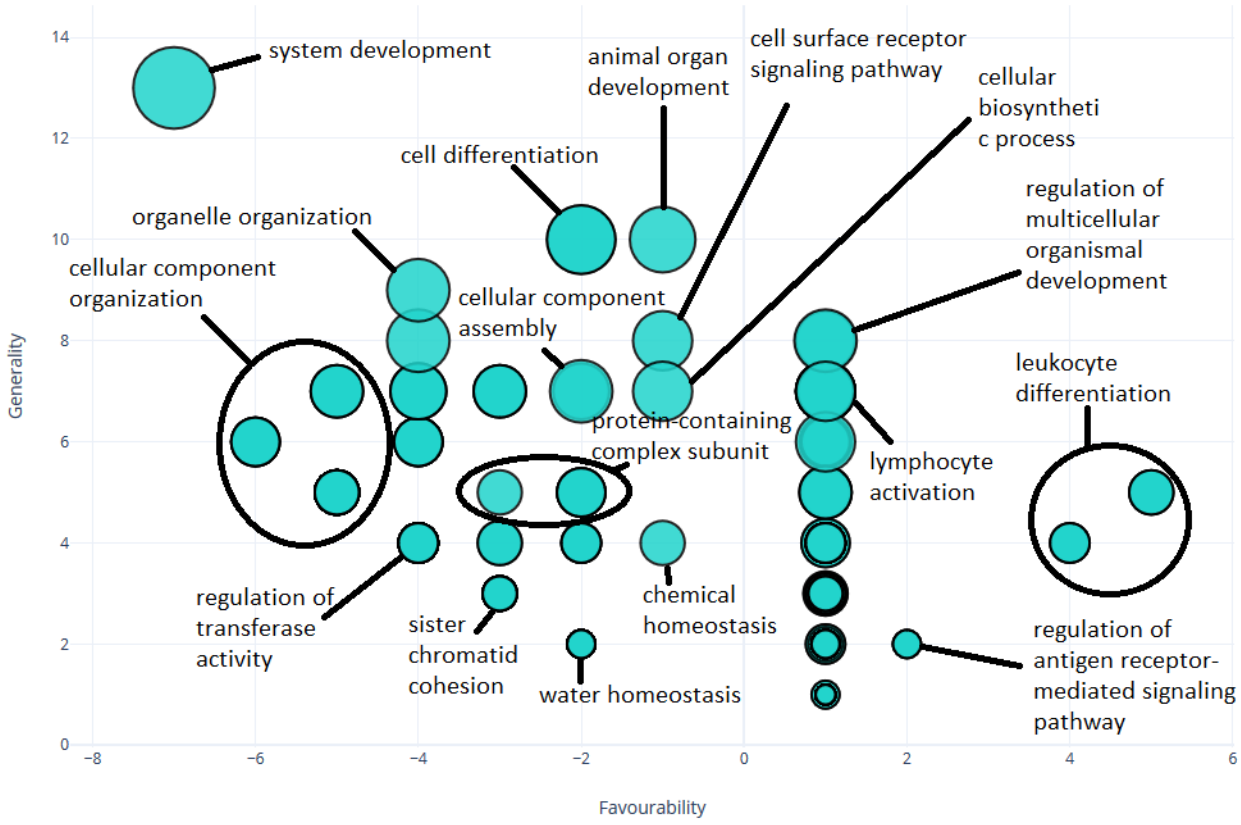

Figure 3.4: The updated Gene Ontology enrichment analysis bubble plot. The x axis shows in how many cases the GO-term was favorable or unfavorable. The y axis represents the amount of cancers (out of 17 major cancers) the GO-term was present. The Bubble sizes symbolize the number of annotated genes to a GO-term. For a better overview of the graph visit the online version: https://plot.ly/~PascalNuijten/15/

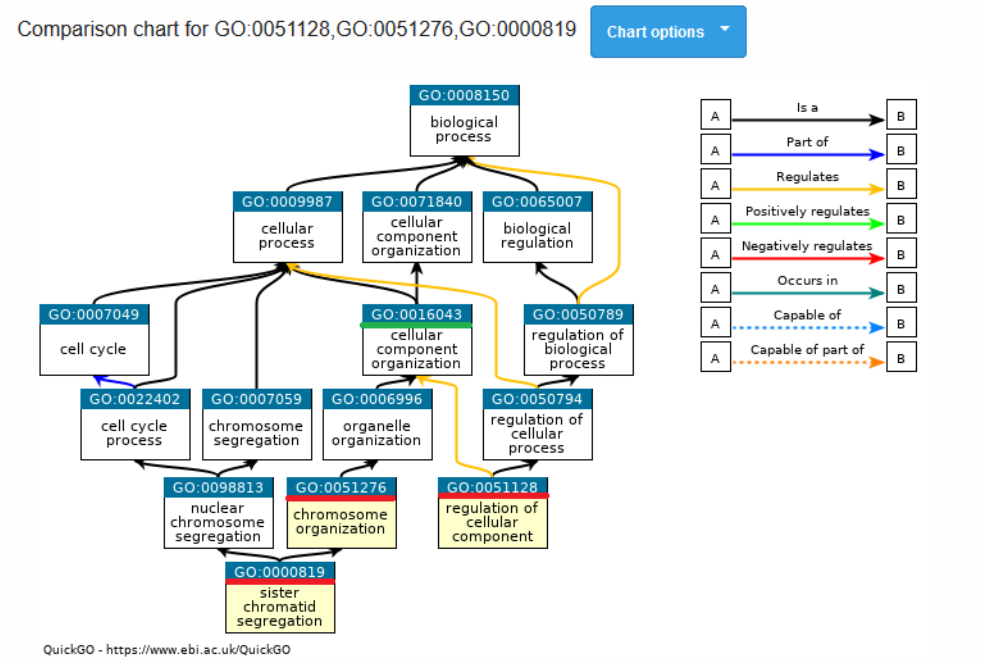

Figure 3.5: The Gene Ontology hierarchy of the cluster: Cellular Component Organization (visible in Figure 3.4) obtained by QuickGO [47]. QuickGO is a nice way of visualizing the relationships between the terms in the clusters. The red marked GO-terms represent the separate bubbles in the bubble plot. The green marked GO-term is the overarching term and therefore the identity of the cluster. Relations of the GO-terms are clarified in the legend

3.4 COSMIC data comparison

As already mentioned in the previous chapter this step was conducted to obtain a better sense of the data. Surprisingly only 24 out of 672 prognostic genes (visible in Figure 3.4 matched with the COSMIC genes. This is an approximate 3.6%. Which is a rather low percentage, and therefore opposing the expectations. Although this is against expectations it still can be explained. The COSMIC gene census is a collection of 793 genes that have mutations directly linked to one or more cancers. Some of these mutations show differential expression and therefore appear as prognostic genes. However, the prognostic genes obtained from the Uhlen paper most likely also represent downstream effects of mutations in interacting genes and proteins. Therefore the genes of the Uhlen paper are not necessarily directly linked to cancer but could also be downstream effects of mutations.

3.5 Comparison plot: The original versus the updated GO plot

Figure 3.6 shows the combination of the previous two bubble plots (Figure 3.1 and Figure 3.4). To get a fair comparison all scales and calculations to obtain the bubble plot are equivalent. When looking at this graph the difference in the amount of bubbles stands out immediately. This gain of GO-terms is the result of the increase of gene knowledge in the last decade. More genes have been

successfully identified and linked to their GO-term, this concludes in a growth of GO-annotations. Another interesting observation is the relative growth in generality and unfavorability. The GO-term *system development* and the cluster *cellular component organization* create a good example of the shifting bubbles. The growth in generality can be explained by the growing gene knowledge, more genes in different cancers are annotated to popular GO-terms. The original bubble plot is just as the new bubble plot skewed to the left. However, the new discovered cluster for example is not found by the original paper. Therefore conclusions based on the original bubble plot may differ due to the new insights granted by the updated gene ontology enrichment analysis.

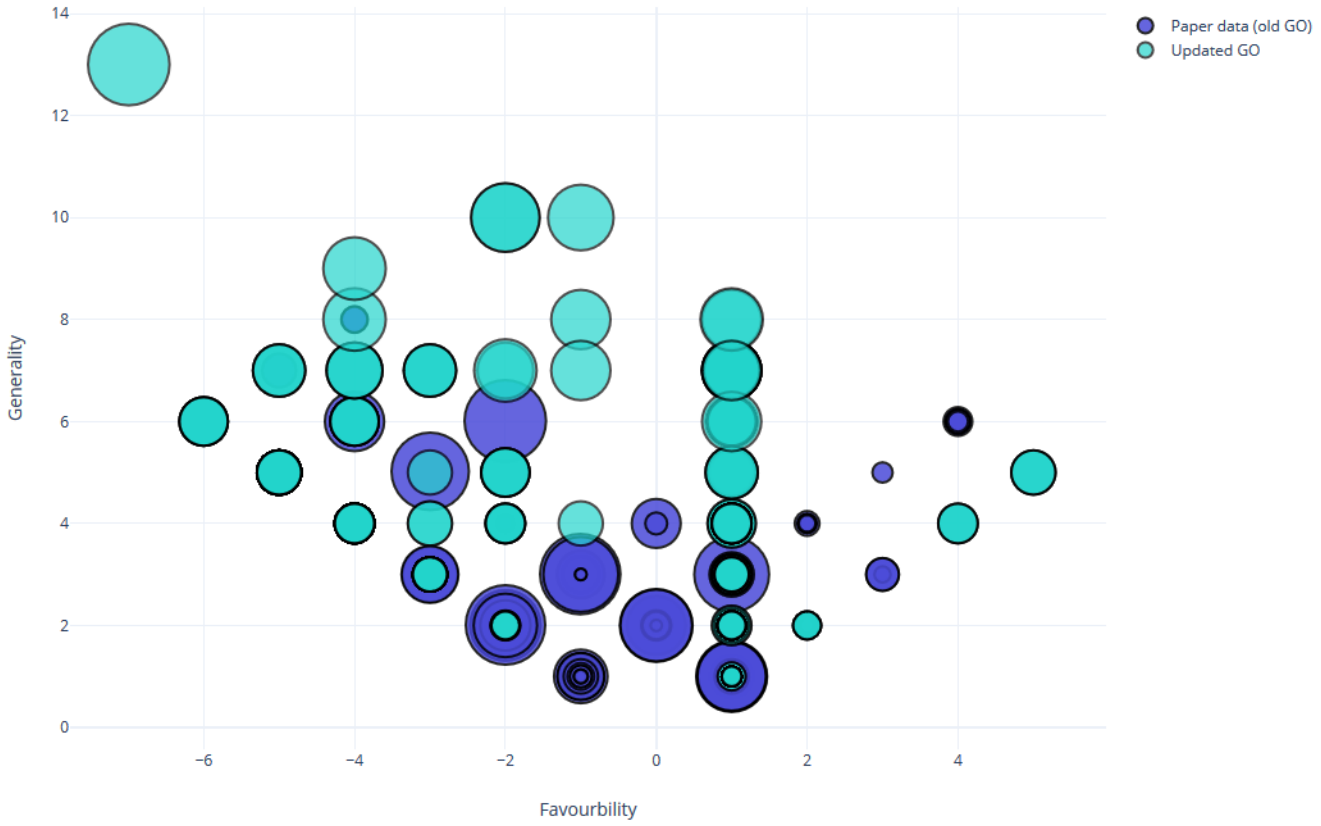

Figure 3.6: The bubble plot combining the original and the updated gene ontology enrichment analysis. The x axis shows in how many cases the GO-term was favorable or unfavorable. The y axis represents the amount of cancers (out of 17 major cancers) the GO-term was present. The Bubble sizes symbolize the number of annotated genes to a GO-term. The bubble plot is online accessible at: https://plot.ly/~PascalNuijten/27/

3.6 Orthologue search in *Mus Musculus*

To visualize the information gain obtained by this step the graph in Figure 3.7 was established. The graph reveals that the amount of GO-terms for some cancer types has more then doubled relatively to the updated Gene Ontology enrichment analysis. For *stomach*, *lung*, *thyroid* and *prostate* cancer there were no enriched GO-terms for the human search but but only for the homologue mouse one. This is serious gain of information, where previously no sense of the data was obtained for human genes there now emerges a better understanding of the data due to the orthologue mouse genes. This difference in annotations can be devoted to the amount of research that is done in mice. Where ethics oppose human research the research in mouse continues. Therefore a popular method for result verification and a better sense of biological data is the check of homologue species.

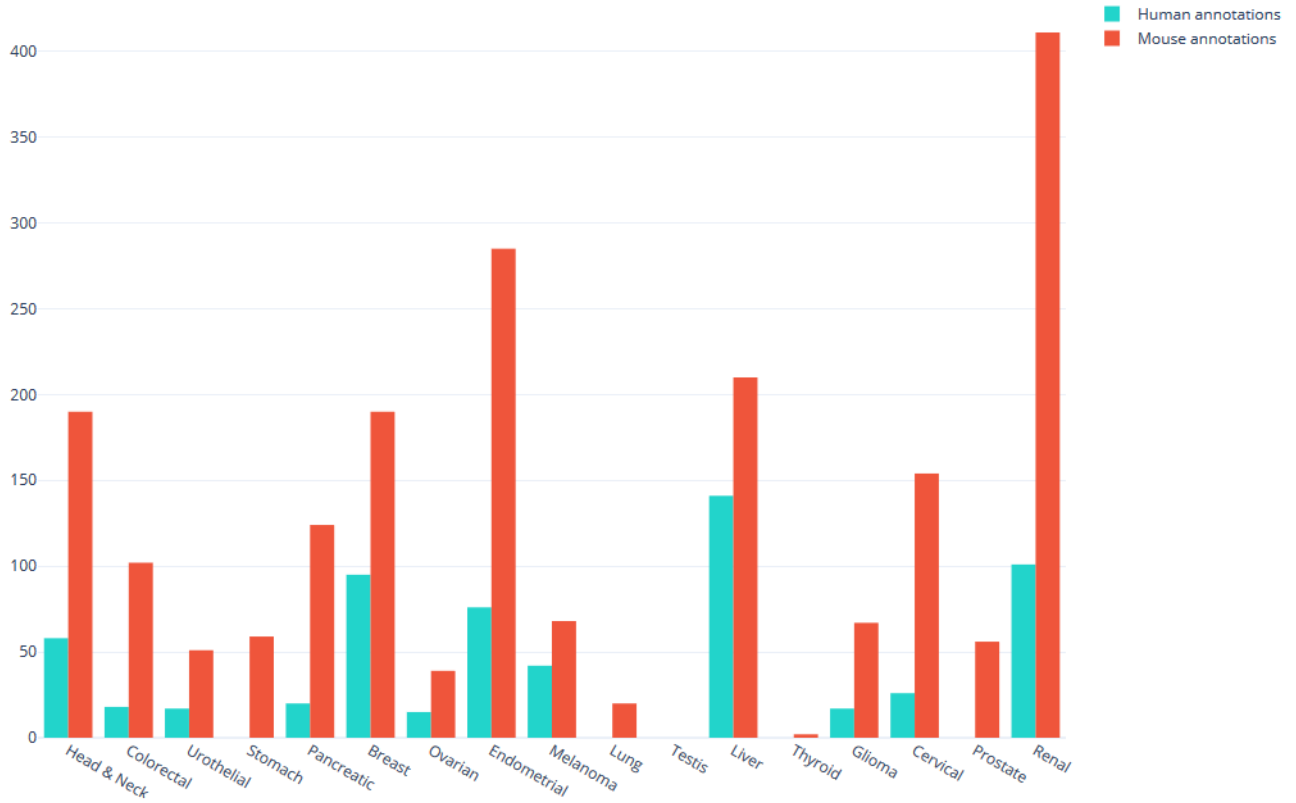

Figure 3.7: Difference in enriched GO-term for all major favorable cancer types for human and mouse (both P=0.001). The chart is also visible at: https://plot.ly/~PascalNuijten/38/

3.7 The orthologue Mouse bubble plot

With the results obtained from GO enrichment analysis in the previous section another bubble plot was created. As mentioned, the method for creating the plot (Figure 3.8) is the same as the previous bubble plots and is described in Section 2.2.5 to the earlier discussed bubble plots. When observing the mouse bubble plot a lot of interesting bubbles can be distinguished. Where not only the amount of bubbles increased significantly also several big bubbles showed up on the border of bubble plot. Because of the size and position, a strong GO-term signal can be correlated to these bubbles. For example, GO-terms presented as a big bubble in combination with a high generality such as multicellular organismal process, organ development and the metabolic process cluster evidence a strong relation with cancer. Wether this relation has a positive or negative impact on the clinical outcome is indicated by the favorability, which is positioned on the X-axis.

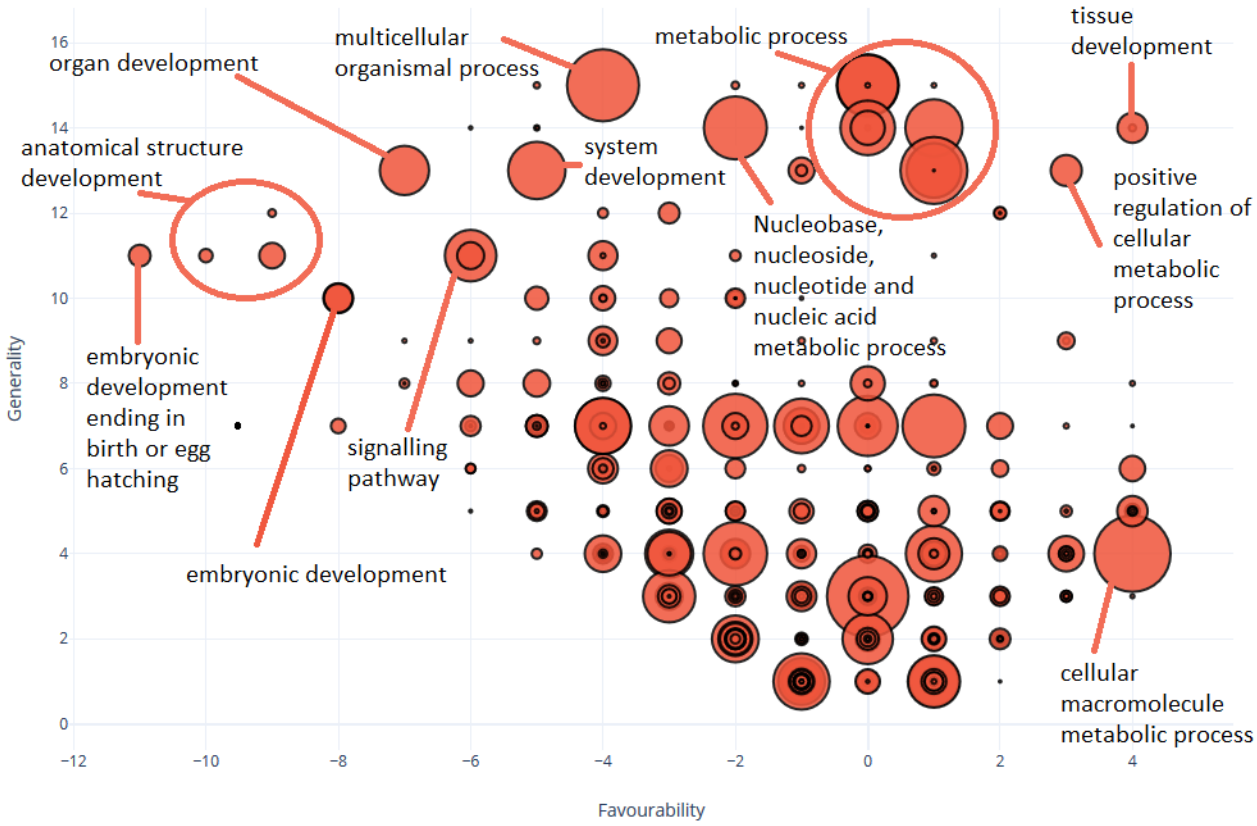

Figure 3.8: Bubble plot containing GO-terms obtained by the mouse gene ontology enrichment analysis. GO-terms appearing on the first three levels of the GO hierarchy were discarded. Interesting bubbles and clusters are labeled. The x axis shows in how many cases the GO-term was favorable or unfavorable. The y axis represents the amount of cancers (out of 17 major cancers) the GO-term was present. The Bubble sizes symbolize the number of annotated genes to a GO-term. The bubble plot is online accessible at: https://plot.ly/~PascalNuijten/21/

3.8 Comparison plot: The updated GO versus The orthologue mouse

To get a good impression of the information gain, created by the orthologue mouse data, another united bubble plot was established (Figure 3.9). The most important and interesting overlapping GO-terms are highlighted in both data set. As expected a fairly large number of new bubbles are introduced in the mouse data. Noteworthy is the fact that a lot of these new bubbles appear with a quite strong signal in the top half of the graph (above generality: 8) where the human data only shows *system development*. This shift and increase of bubbles indicates a large gap between the present knowledge in human and mouse. Furthermore in some cases there is a significant difference in position and size of GO-terms such. For example, where *cell differentiation* is presented as a quite important GO-term in the human data it only appears on the bottom of the graph in the mouse annotation analysis. Knowing this the reliability of only using human graph is questioned.

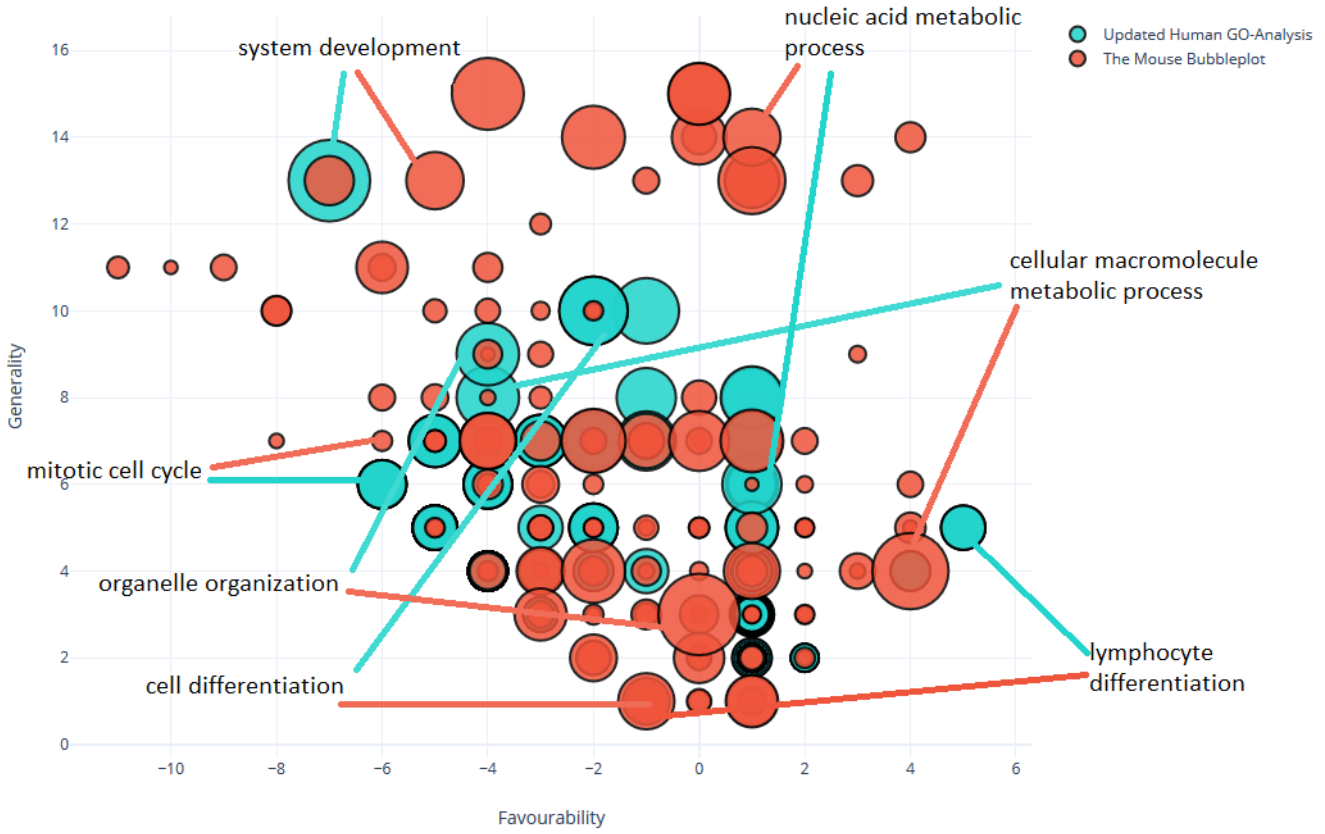

Figure 3.9: Bubble plot presenting the new obtained human GO enrichment analysis data and the same enrichment analysis done with orthologue mouse data. Interesting bubbles that represent the same GO-term in both datasets are labeled to distinguish the difference in position and size. The x axis shows in how many cases the GO-term was favorable or unfavorable. The y axis represents the amount of cancers (out of 17 major cancers) the GO-term was present. The Bubble sizes symbolize the number of annotated genes to a GO-term. The bubble plot is online accessible at: https://plot.ly/~PascalNuijten/73/

3.9 Comparison plot: The updated GO versus the orthologue mouse versus the original

The previous step was al about presenting the data differences in methods and time. Therefore it is making sense to compare all these different data sets in one last bubble plot. In this bubble plot the original as well as the updated human and mouse data are visualized (Figure 3.10). It is remarkable that except for the overall skew to the left no other common characteristic seem to exist. There for it is fair to assume that the original plot did not provide a complete picture. Beside the fact that more data is introduced and the signals are overall stronger also the location of GO-terms vary through the in this thesis presented steps. The paper "A pathology Atlas of the Human Cancer Transcriptome" [12] originally concluded from Figure 2.2 that : "many of the common unfavorable genes are related to cell proliferation, including mitosis, cell cycle regulation and nucleic acid metabolism". Looking at the graph in Figure 3.10 a lot compared to the original plot did change. The conclusion of the Uhlen paper partly still holds, for example mitosis, cell cycle regulation and nucleic acid metabolism still show a rather strong signal. However a lot of new, even more interesting, bubbles such as *system development*, *organelle organization* and more rise up and should be added to the conclusion.

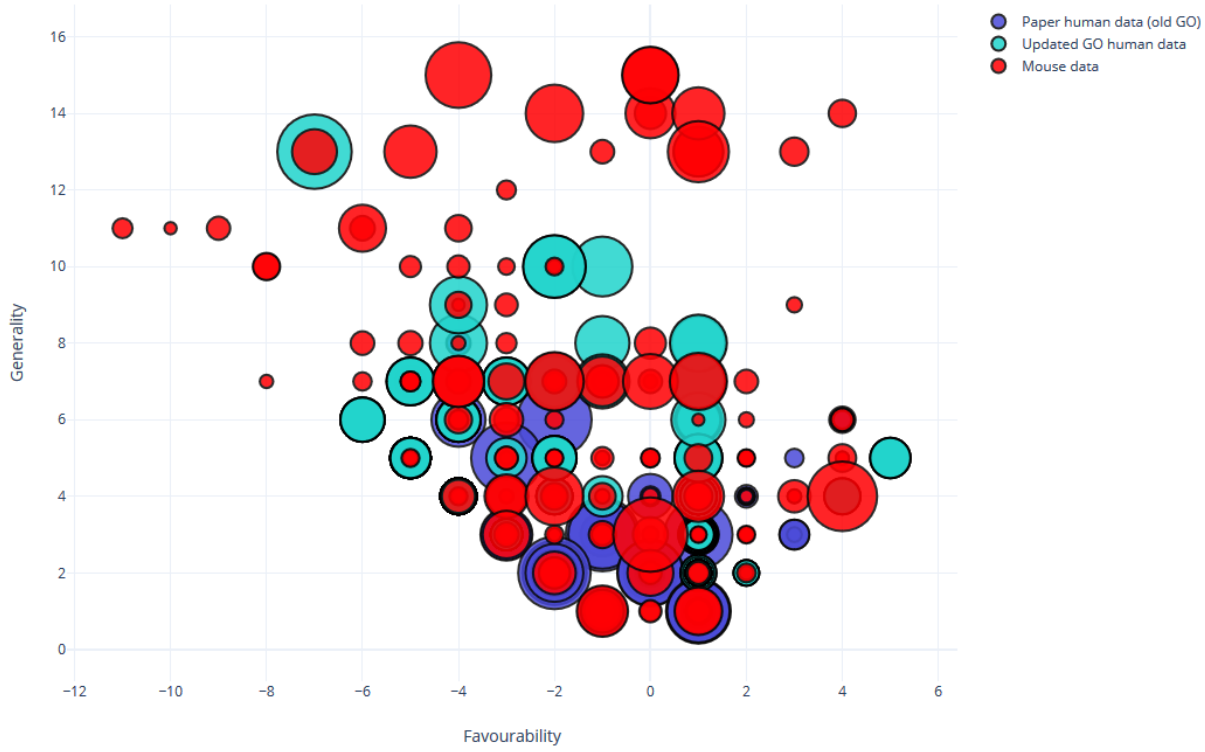

Figure 3.10: The original human, the updated human and the mouse bubble plot presented in one graph. All data is processed through the same methods and scales and hence the data one on one comparable. The x axis shows in how many cases the GO-term was favorable or unfavorable. The y axis represents the amount of cancers (out of 17 major cancers) the GO-term was present. The Bubble sizes symbolize the number of annotated genes to a GO-term. The bubble plot is online accessible at: https://plot.ly/~PascalNuijten/40/

Chapter 4

Conclusions and Further Research

4.1 Conclusions

As the previous chapter already implied, a rather large gain of information can be achieved by updating or switching up data resources. Where the original bubble plot counts 437 GO-terms, the new plot already contained 1250 GO-terms. This is an increasement of 186 %, which is an enormous amount of data. Not only is this thesis showing that older assumptions in this rapid innovating field of science easily outdate, it also proves that data verification by the comparison of orthologue species can lead to different or even new insights. In this thesis the use of homologue species shows a significant growth of data and therefore also provided new bubbles such as *signaling pathway*, *anatomical structure development*, and many more. The comparison plot in Figure 3.9 shows that certain mouse bubbles obtained a significant differing position regarding to the human data in the graph. This means that the coordinates from the corresponding GO-terms vary in value. As mentioned in chapter 2.2.5, the x-axis and y-axis are derived from the difference in favorability and the number of cancers with genes mapped to the GO term, respectively. Since the human genes are one on one translated to orthologue mouse genes the shifting bubbles can be devoted to either a gain or reduction of genes linked to a GO-term. A gain or reduction of mapped genes can be declared by the increase of knowledge about the genes. Where the human gene might lack evidence, the orthologue mouse gene might be mapped to an additional, or substitute, GO-term. This effect will causes bubbles to obtain a different location. More prove for the gap between human and mouse data can be found in Figure 3.8. This graph immediately reveals a gain of bubbles in comparison with the human bubble plot. Nevertheless some of the interesting additional bubbles are correlated to processes that occur in early stage of life. Example of these GO-terms are : organ development, embryonic development and embryonic development ending in birth or egg hatching. These additional bubbles can be devoted to the difference in embryonic research, where ethics clash with the research of human fetus other species such as mice are less complicated to examine. Therefor a gap of information is found. When looking at the bubble plots in Figure 3.4 - Figure 3.10 a skew to the unfavorable site (left) side can be observed. This means that the greater part of the genes annotated to the skewed GO-terms are correlated to a negative, unfavorable, clinical outcome. As mentioned, a negative clinical outcome is linked to genes that are found in patient with a, in proportion, short survival rate. This trend towards the unfavorable side is already found in the original data. The original paper, "A Pathology Atlas of the Human Cancer Transcriptome" [12] shares 18470 prognostic cancer genes. Out of these genes 8254 genes were marked favorable

and 10216 unfavorable. This difference partly explains the skew towards the unfavorable side, more unfavorable genes will likely result in a bigger set of annotated GO-terms. However when examining Figure 2.2 and Figure 3.10 the difference in favorability is indicating either a bigger variance in gene set size or another factor. An alternative factor is the fact that regular cancer research usually just aims for the unfavorable genes. Therefore there is more known about these genes and their annotations. Where favorable genes lack knowledge and hence annotations, more annotations are found for unfavorable genes. In Figure 2.2 the GO-term *system development* scores a high generality and low favorability. Which means that genes correlated to the GO-term *system development* were found across almost all cancers with an unfavorable clinical outcome. The bubble is also found in the mouse bubble plot (Figure 3.9). However *system development* is the outstanding top term in the human graph it is not quite as remarkable in the mouse plot due to comparable bubbles. To feedback the problem stating and research question, we indeed can conclude that re-doing older analysis could lead to different results and therefore differ in conclusion. Figure 3.6 is showing that the reconstruction of a research in a rapid evolving field such as bioinformatics leads to very different results. An other goal of the thesis was to see what the effect of orthologue species would have on the data set. As Figure 3.7 already implies a big gain of data is obtained when looking at mouse data. The bubble plot in Figure 3.9 visualizes this data difference and hence proves that a lot of new and interesting data has been added. To obtain a clear display of the data shifts Figure 3.10 is putting all the data changes in perspective.

4.2 Future Research

This thesis provides new insights by repeating previously done research with updated data and different tools. These new insights typically raise new questions. In this section these questions will be highlighted and then put in perspective. Down below a summation of the potential interesting future research steps can be found.

- Because the GO-term *system development* is so remarkable when observing Figure 2.2 and Figure 3.9 it would be interesting to look closer at the genes mapped to this term. A typical future question would be: are the same system development genes found across multiple cancers and do these genes have the same prognostic type? Answering for example this question could possibly give a good background of frequent occurring cancer genes.

- As mentioned before different tools might have different approaches for the same problem and therefore vary in results. This makes the use of even more tools interesting. An example could be the use of an alternative orthologue search tool or even repeat the orthologue search with other orthologue species. Results obtained with these approaches will tell more about the certainty of the previous results.

- Since the Gene Ontology is structured as a big tree a lot of terms are direct or indirect related. There exist various tools such as GOGO [48] which alow to measure the semantic similarities between GO-terms. It is potentially interesting to calculate the semantic difference between GO-terms which cluster or appear close to each other in the bubble plots. With these scores the significance of existing clusters and even new ones can be found.

- In 2000 Douglas Hanahan and Robert A.Weinberg introduced the six hallmarks of cancer [49], and came with an additional four hallmarks in 2011 [50]. These hallmarks describe cancer as a system level disease where every hallmark is covering a part of cancer. Every gene that is involved in the evolution of cancer can be linked to a hallmark. These hallmarks make a powerful tool to identify a set of cancer genes. Therefore the comparison of mapping to hallmarks with the gene enrichment analysis to GO-terms would be an interesting addition to the work done in this thesis.

4.3 Acknowledgements

This thesis is build on results obtained from "A Pathology Atlas of the Human Cancer Transcriptome" [12] by Mathias Uhlen et al. Special thanks to Katy Wolstencroft for her knowledge and advice on the subject.

Bibliography

[1] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal, "Cancer statistics 2019," *Statistics*, vol. 69, pp. 7-34, 2 2019.

[2] "Worlds health statistics 2018: monitoring health for the SDGs, sustainable devlopment goals," *World Health Organization*, vo1. 1, pp. 4-12, 2018.

[3] Neil A. Campbell, et al. "Biology," *Pearson Benjamin Cumminbs*, ed. 10 pp. 439-444, 2014.

[4] Jacqueline Boultwood, and Carrie Fidler, "Molecular Analysis of Cancer," *Humana Press*, pp. 1-6, 2002.

[5] Richard Rothenberg, "The Causes of Cancer, Revisted," *Annals of Epidemiology*, vol. 25, no. 3, pp. 215, 2015.

[6] Suzanne Clancy, "DNA Damage Repair: Mechanisms for Maintaining DNA Integrity," *Nature Education*, vol. 1, pp. 103, 2008.

[7] Alessandro Torgovnick, and Bjoern Schumacher, "DNA repair Mechanisms in Cancer Decelopment and Therapy," *Frontiers In Genetics*, vol. 6, pp. 157, 2015.

[8] Aran Dvir, et al, "Systematic Pan-Cancer Analysis of Tumour Purity," *Nature Communications*, vol. 6, no. 1, pp. 8971, 2015.

[9] Prashant Bavi, et al. "Developing a pan-cancer research autopsy programme," *Journal of Clinical Pathology*, vol. 72, no. 10, pp.689-695, 2019.

[10] John N. Weinstein, et al. "The Cancer Genome Atlas Pan-Cancer Analysis Project," *Nature Genetics*, vol. 45, no. 10, pp. 1113-1120, 2013.

[11] Li Ding, et al, "Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics, " *Cell*, vol. 173, no. 2, pp. 305-320, 2018.

[12] Mathias Uhlen, et al, "A Pathology Atlas of the Human Cancer Transcriptome," *Science*, vol. 357, no. 6352, pp. 660, 2017.

[13] Mathias Uhlen, et al, "Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, 2015.

[14] Miten Jain, et al, "The Oxford NAnopore MinION: delivery of nanopore sequencing to the genomics community," *Genome Biology*, vol. 17, no. 1, 2016.

[15] Michael Ashburner, et al, "Gene Ontology: Tool for the Unification of Biology," *Nature Genetics*, vol. 25, no. 1, pp. 25-259, 2000.

[16] G. O. Consortium, "The Gene Ontology resource: 20 years and still going strong," *Nucleic Acids Research*, vol. 47, no.D1, pp. D330-D-338, 2019.

[17] W M Gelbart, et al, "The FlyBase Database of the Drosophila Genome Projects and Community Literature," *Nucleic Acids Resarch*, vol. 27, no. 1, pp. 85-88, 1999.

[18] Blake, et al, "The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse," *Nucleic Acids Research*, vol. 28, no. 1, pp. 108-111, 2000.

[19] M Ringwald, et al, "GXD: a Gene Expression Database for the Laboratory Mouse: Current status and Recent Enhancements. The Gene Expresison Database Group," *Nucleic Acids Research*, vol. 28 no. 1, pp. 115-119, 2000.

[20] Ball, et al, "Integrating Functional Genomic Information into the Saccharomyces Genome Database," *Nucleic Acids Research*, vol. 28, no. 1, pp. 77-80, 2000.

[21] Logan-Klumpler, et al. "GeneDB–an Annotation Database for pathogens," *Nucleic Acids Research*, vol. 40, no. D1, pp. 98-108, 2012.

[22] Alex L Mitchell, et al. "InterPro in 2019: Improving Covarage, Classification and Access to protein Sequence Annotations," *Nucleic Acids Research*, vol. 47, no. D1, pp. 351-360, 2019.

[23] Zerbino, et al. "Ensembl 2018," *Nucleic Acids Research*, vol. 46, no D1, pp 754-761, 2018.

[24] Esmaeil Ebrahimie, et al. "Gene Ontology-Based Analysis of Zebrafish Omics Data Using the Web Tool Comparative Gene Ontology," *Zebrafish*, vol. 14, no. 5, pp. 492-494, 2017.

[25] Huaiyi Raudvere , et al. "PANTHER version 14: More Genomes, a New PANTHER GO0Slim and Improvements in Enrichement Analysis Tools," *Nucleic Acids Research*, vol. 47, no. D1, pp. D419-D426, 2019.

[26] Uku Raudvere, et al. "g:Profiler: a Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 update)," *Nucleic Acids Research*, vol. 47, no. W1, PP.W191-W198, 2019.

[27] Steven Maere, et al. "BINGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks," *Bioinforamatics*, vol. 21, no. 16, pp. 3448-3449, 2005.

[28] Eran Eden, et al. "GOrilla: a Tool for Discoverey and Visulaization of Enriched GO Terms in Ranked Gene Lists," *BMC Bioinformatics*, vol. 10, no.1, p. 48, 2009.

[29] Robert King, et al. "Model Organism," *A Dictionary of Genetics*, pp A Dictionary of Genetics, 2013.

[30] E. Lander, L. Linton, B. Birren, et al. "Initial Sequencing and Analysis of the Human Genome," *Nature*, vol. 409, no. 6822, pp. 860-921, 2001.

[31] "The European Code of Conduct for Research Integrity Revised Edition," *ALLEA - All European Academics*, 2017.

[32] Kylie Yong, et al. "Humanized Mice as Unique Tools for Human-Specific Studies," *Archivum Immunologiae Et Therapiae Experimentalis*, vol. 66, no. 4, pp. 245–266, 2018.

[33] Ryoji Ito, et al. "Humanized Mouse Models: Application to Human Diseases," *Journal of Cellular Physiology*, vol. 233, no. 5, pp. 3723–3728, 2018.

[34] Paul Shannon, et al. "Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Research*, vol.13, no. 11, pp. 2498-2504, 2003.

[35] Endy Drew, and Brent Roger. "Modelling Cellular Behaviour," *Nature*, vol.409, no.6818, pp. 391-395, 2001.

[36] Anwar Joarder. "Hypergeometric Distribution and Its Application in Statistics," pp. 641-643, 2011.

[37] Plotly Technologies inc. Collaborative data science publisher: Plotly Technologies Inc. Accessible at http://www.plot.ly, 2015.

[38] John G Tate, et al. "COSMIC: the Catalogue Of Somatic Mutations In Cancer," *Nucleic Acids Research*, vol. 47, no. D1, pp. D941-D947, 2019.

[39] Zbyslaw Sondka, et al. "The COSMIC Cancer Gene Census: Describing Genetic Dysfunction across All Human Cancers," *Nature Reviews. Cancer*, vol. 18, no. 11, pp. 696-705, 2018.

[40] Damian Smedley, et al. "The BioMart community portal: an innovative alternative to large, centralized data repositories," *Nucleic Acids Research* vol. 43, pp W589-W598, 2015.

[41] Stephen F Altschul, et al. "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403-410, 1990.

[42] National Center for Biotechnology information (NCBI). Bethesda: National Library of Medicine (US), Accessible at https://ncbi.nlm.nih.gov/, 1988.

[43] Cunningham, et al. "Ensembl 2019," *Nucleic Acids Research* vol. 47, no. D1, 2019, pp. D745-D751, 2019

[44] Python Software Foundation. Python Language Reference, version 3.7.6 Available at http://www.python.org, 2008.

[45] Da Wei Huang, et al. "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat protoc*, vol. 4, pp. 44-57, 2008.

[46] Da Wei Huang, et al. "Bioinformatics enrichment tools: paths towards the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1-13, 2009.

[47] David Binns, et al. "QuickGO: a web-based tool for Gene Ontology searching," *Bioinformatics*, vol. 25, no. 22, 2009, pp. 3045-3046, 2009.

[48] Zhao Chenguang and Zheng Wang. "GOGO: An Improved Algorithm to Measure the Semantic Similarity between Gene Ontology Terms," *Scientific Reports*, vol. 8, no. 1, pp. 15107, 2018.

[49] Douglas Hanahan and Robert A.Weinberg. "The Hallmarks of Cancer," *Cell*, vol. 100, pp. 57-70, 2000.

[50] Douglas Hanahan and Robert A.Weinberg. "Hallmarks of Cancer: The Next Generation," *Cell*, vol. 144, pp. 646-674, 2011.