Implementation of an RNA-Seq pipeline and differential gene expression analysis of canine prostate cancer samples

Master Thesis

Melanie Himmelfreundpointner, BSc

to achieve the university degree of
Diplom-Ingenieurin
in Biomedical Engineering

submitted to

Institute of Biomedical Informatics,
Graz University of Technology
Stremayrgasse 16/I
Head: Univ.-Prof. Dr.rer.nat.habil. Leila Taher

Supervisor
Univ.-Prof. Dr.rer.nat.habil. Leila Taher

Graz, May 2021

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

11.05.2021

Date

Signature

Acknowledgement

This thesis would not have been possible without the support of several people and I would like to extend my sincere thanks to all of them:

First of all, I want to thank my supervisor Prof. Dr. Leila Taher for precious support, patience and immense knowledge and it has always been a privilege to work with her.
Furthermore, I want to thank the whole team of the Insitute of Bioinformatics for interesting presentations which enables me to expand my knowledge.

I would like to thank my partner Felix for supporting me endlessly and being my solid anchor. Most of all, I want to thank my family, especially my mother Barbara for giving me the opportunity for a good education and for being my lovely mom. Thank you for always encourage me to give my best in every situation!

Melanie Himmelfreundpointner
Graz, May 2021

Abstract

Keywords: RNA-seq, differential gene expression analysis, prostate adenocarcinoma, canine cell lines

Prostate cancer can affect both humans and animals, but the dog is the only other large mammal that spontaneously develop prostate cancer with aging. The dog is an adequate animal model of aggressive human prostate cancer, because of the similar anatomy of the prostate and pathological resemblances to men. Research into the molecular mechanisms of prostate cancer is facilitated by the use of *in vitro* model systems and therefore, in recent years many cell lines have been established from primary tissue sources. Cell lines are not only used for understanding molecular mechanisms in tumor progression, but also for designing new therapeutic strategies. However, cell lines can also differ from primary tumors in biologically significant ways and not every cell line may be a suitable model for their tumor type. Finding and interpreting the differences (DEGs) between canine prostate cell lines and canine prostate primary tumor tissues is the main goal in this thesis to find out if the cell lines are a suitable model for their tissue samples.
Therefore an RNA-sequencing (RNA-seq) analysis as well as a differential gene expression analysis was performed to identify genes or molecular pathways that are differentially expressed. In the first part of this thesis a RNA-seq pipeline was constructed for pre-processing the sequencing data as well as mapping to the reference transcriptome. This pipeline uses some of the most widely used tools in the field of RNA-seq analysis. The second part consists of the differential gene expression analysis by providing a complete computational workflow. The used data set is obtained by the group of *Prof. Dr. Ingo Nolte* from the *Tierklinik Hannover* and consists of canine prostate cancer cell lines and tissues. Pathways that were down-regulated in cell lines compared to tumor tissue included cell communication, immune processes, cell adhesion molecules and ECM-receptor interaction. Up-regulated pathways were generally associated with cellular growth and DNA repair. The results provide insights into the expression differences between cell lines and tissues. The strong changes which were observed should be considered when using cell lines as models for tissues. It is important to understand that differences exist for designing and interpreting experimental studies. Thus, the choice of an appropriate cell line for a specific project depends mainly on the goal and context of the study and the suitability of cell lines should be carefully validated before use.

Contents

List of Figures

List of Tables

List of Abbreviations

WHO	World Health Organization
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
RNA-seq	Ribonucleic acid sequencing
TCC	Transitional cell carcinoma
PAC	Prostate adenocarcinoma
cDNA	complementary DNA
GEA	Gene expression analysis
CPU	Central processing unit
SE	single-end
PE	paired-end
DEG	differentially expressed gene

1 Introduction

Besides cardiovascular diseases, cancer is the second leading cause of death. According to World Health Organization (WHO) there are over 1.8 million deaths and 3.90 million cases of cancer in Europe each year [1]. In light of this, it is of paramount importance to analyze cancer cells and cancerous tissue samples to understand the genetic processes behind it. Latest technologies and scientific advancements make it possible to better understand the genetics in the body, however many things are yet to be investigated and explored. In this thesis the main focus lies on creating a pipeline for RNA-seq analysis and finding the differentially expressed genes between cancer cell lines and cancerous tissue samples.

1.1 Cancer

Cancer is a disease, where cells grow and spread uncontrollably in the body [4]. This growth can be triggered either due to genetic factors or external influences [4]. The most important differentiation in cancer genetics is between benign and malignant tumors, an abnormal mass of tissue where cells grow and divide uncontrollably [135]. Benign tumors are non-cancerous, they can be easily demarcated from the surrounding tissue, they grow slowly and the structure of the tumor cells differ only little from the original cells [5]. A malignant tumor instead grows very quickly and aggressively invades surrounding tissues [5]. Malignant tumors often spread throughout the body along nerves or blood vessels and can be transported to other tissues of the body via the bloodstream [2]. These tumor cells again propagate and grow into a new tumor. Such types of tumors are called metastases [2]. Only malignant tumors are referred to as cancer and they can reappear after they have been removed [6]. However, what makes cancer so dangerous is their property to metastasize, which is the cause of 90% of all deaths regarding cancer, according to Gundem et al. [7]. Malignant and benign tumors are classified according to the type of tissue and the infested organ where it first was detected [3]. According to the National Cancer Institute [3] there are hundreds of different cancer types which can occur, but 90 percent of all cancer types fall into some main categories:

- Carcinoma
- Sarcoma
- Myeloma
- Leukemia
- Lymphoma
- Mixed Types

At least about 90 percent of all human cancer types are carcinomas, which are malignancies of epithelial cells [2]. Such epithelial tissues are common throughout the body like in the skin, gastrointestinal tract, prostate, liver, kidneys, urinary bladder and many more [8].

Carcinomas are divided into:

- **Basal cell carcinoma**
  According to Lanoue et al. [9] it is the most commonly occurring cancer in the world. Metastasis is extremely rare and originates from cells of the hair follicle, from epidermal stem cells or from cells of the interfollicular epidermis [10].
- **Squamous cell carcinoma**
  It is the second most common form of skin cancer and appears really often in sun-damaged skin [11].

- **Transitional cell carcinoma**
  It originates from the transitional epithelial cells of the urinary system, e.g. kidneys, bladder or accessory organs and is the most common bladder cancer in the United States [12]. Transitional cell carcinoma develops from the urothelium, which lines the renal pelvis, ureter and urinary bladder [13]. TCC of the renal pelvis or ureter is a relatively rare disease (less than 1%) in contrast to the urinary bladder (90%), as Ozsahin et al. [14] reported.

- **Adenocarcinoma**
  It originates in glandular cells, which can be found in tissues and organs which are able to produce secretion, mucus, digestive juices or seminal fluids [15]. Adenocarcinomas can occur in many parts of the body, for example in the breast, lung, prostate, gastrointestinal tract like colon, pancreas, stomach or esophagus and also make 70 % of cancer of unknown origin [16].

The main focus of this master thesis is on the adenocarcinoma and transitional cell carcinoma, related to the prostate.

1.1.1 Human prostate Cancer

Almost all prostate cancer types are adenocarcinomas and are globally the most often noncutaneous cancer in men [19]. The prostate is a gland found only in men where parts of the seminal fluid is produced. Prostate cancer is especially frequent in developed countries [19]. The causes of prostate cancer development are largely unknown, however age as well as genetics are high risk factors [16]. There are three stages in which human prostate cancer develops [86]:

1. **Intraepithelial neoplasia:**
   It is a pre-malignant form of prostate cancer, a hyperplasia, i.e. an overgrowth of tissue, of luminal cells [20]. Luminal cells coat the lumen and produce human prostate specific antigen (*PSA*) and androgen receptor (AR) [21]. Also an increasing loss of basal cells appears, leading many to speculate that luminal cells can be the cell source for prostate cancer and therefore driving force for tumor progression [21].

2. **Androgen-dependent adenocarcinoma:**
   Initial phase of prostatic cancer cells are androgen-dependent and they need androgen to proliferate [22]. Usually androgenes are removed to stop proliferation, which is often used in therapeutic approaches regarding prostate cancer [22]. A complete loss of basal cells is also associated with this stage [86].

3. **Androgen-independent adenocarcinoma:**
   After the early stage, many adenocarcinomas become more aggressive and thus resistant to hormone deprivation, which means that the carcinoma becomes androgen-independent. [23]. Unfortunately, the processes and reasons that control the transition to androgen-independent prostate cancer are not known yet[23].

In contrast to other types of cancer, prostate cancer has a low mutation rate, as well as few chromosomal losses or gains [24]. The most commonly focal loss is observed at the phosphatase and tensin homolog (*PTEN*) gene [24]. Advances in molecular biology allow a detailed investigation of the genomic events regarding prostate cancer development and emergence [26]. Prostate cancer is divided into two molecular groups [26]:

- **ETS rearrangements and features of chemoplexy:**
  About 50% of prostate cancers show gene fusions involving ETS transcription factors, whereby chromosomal rearrangements can create a *TMPRSS2-ERG* gene [25]. This fusion protein contributes to prostate oncogenesis activating a transcriptional program by upregulating the expression of enhancer of zeste 2 polycomb repressive complex 2 subunit (*EZH2*), SRY-box

transcription factor 9 (*SOX9*), MYC proto-oncogenes (*MYC*) and repression of NK3 homeobox 1 (*NKX3-1*) expression [86].

- **Absence of ETS rearrangements**
If no ETS transcription factors are involved, often mutation of the speckle type BTB/POZ gene (*SPOP*) or deletion of the cadherin 1 (*CDH1*) gene occur, whereby *SPOP* expression is often downregulated in prostate carcionomas [27].

The development of prostate cancer leads to certain mutations and deregulation of relevant transcription factors [31]. According to Labbe et al. [31], some studies revealed that the loss of the NK3 Homeobox 1 gene (*NKX3-1*) is a tumor-initiating event that ruptures normal prostate epithelial differentiation and can trigger or promote further oncogenic events [31]. Tan et al. [32], reported about that the *NKX3-1* gene promotes the survival of prostate cancer cells because it regulates the AR transcriptional network in such a way that it benefits the cancer cells [32]. It is a tumor suppressor for prostate cancer, but the functional loss of that gene results in a downregulation of other genes that are essential for the prostate, like *SPOP*, *TP53*, *FOXA1*, and *PTEN* [26, 86]. Sun et al. [30] has reported that the phosphatase and tensin homolog gene (*PTEN*) is the top common mutated genes in prostate cancer, which is a multi-functional tumor suppressor and regulates cell proliferation [30]. Kurfurstova et al. [91] observed a downregulation of NAD(P)H quinone dehydrogenase 1 (*NQO1*) together with loss of *PTEN* in an advanced prostate cancer stage which implies that NQO1 may have a tumor suppressive role and may promotes tumor growth [91]. In a healthy prostate, mouse double minute 2 homolog gen (*MDM2*) prevents the transcriptional activity of tumor protein p53 gene (*TP53*) and regulates DNA repair, apoptosis as well as cell cycle [86]. By disrupting these processes, mutations can accumulate in a cell increasing the risk of developing prostate cancer [86].

Cancer-Prostate Cancer Foundation Dream Team [28], also confirmed that the common recurrent gene alterations in prostate cancer included androgen receptor (AR) mutation (62.7%), *TP53* mutation or deletion (53.3%), *PTEN* deletion (40.7%), *RB1* loss (8.6%), *BRCA1* or *BRCA2* mutation or deletion (14.6%) and *CDK12* mutation (4.7%) [28].

A frequently observed down-regulated gene in prostate cancer is cytochrome P450 family 1 subfamily A member 1 (*CYP1A1*) [90]. This phase I enzyme can activate various compounds, which can subsequently lead to carcinogenesis [90]. Interestingly, collagen type I alpha 2 chain (*COL1A2*) gene has been reported to be overexpressed in metastatic prostate tumors [89]. This could be due to the fact that ectonucleoside triphosphate diphosphohydrolase 5 (*ENTPD5*) down-regulation cause morphologic and growth pattern changes and that they could be prevented by collagen [89].

1.1.2 Canine prostate cancer

The dog is the only large mammal that suffers spontaneously from prostate cancer with aging [35]. It is an adequate animal model because of the similar anatomy and pathological resemblances to men [92]. Therefore, canine carcinomas are a natural model for human cancer [92]. Transitional cell carcinoma (TCC) and prostate cell carcinoma (PAC) are the most predominantly diagnosed canine prostatic carcinomas [36]. PAC and TCC in dogs show highly invasive growth and present an aggressive behaviour as well as a poor prognosis [36]. The TCC resembles human invasive bladder cancer and metastatic tumors of closely neighboring organs [33]. The occurance of canine prostate carcinomas is rare, but both species, mens and dogs have similar histopathology and metastatic behaviour [33]. The canine prostate cancer share many characteristics with humans, for example similar morphology, presence of Prostatic Intraepithelial Neoplasia (PIN), Proliferative Inflammatory Atrophy (PIA), bone and lymph node metastasis [37, 38]. Keeping in mind, that the prostate cancer in dogs is androgen independent [37]. Canine prostate cancer shows alterations in *C-MYC*,*TP53* and *MDM2* expression and these discoveries pointed out that the genetic behaviour is similar in both species, dogs and humans [92]. According to Fonseca-Alves et al. [96] genetic and biological processes of canine and human prostate cancer are similar because there are studies in which both species show the loss of *NKX3-1* expression as well as increased *C-MYC* expression.

In canine prostate cancer a loss of androgen receptor (AR), NK3 homeobox 1 (*NKX3.1*) and phosphatase and tensin homolog (*PTEN*) expression occur [97]. Prostate carcinomas in dogs can develop from luminal, ductal, and urothelial cells [98]. Deregulation of cancer-related proteins like MDM2, PTEN, TP53, catenin beta 1 (CTNNB1) and cadherin 1 (CDH1) are common occurrences in dog and human prostate cancer [98].
The expression of prostaglandin-endoperoxide synthase 2 (*PTGS2* or *COX-2*) has been documented in canine prostate cancer as well as in human prostate cancer [99]. The connection between *COX-2* expression in dogs and the histologic type of tumor as well as the presence of inflammation is widely unknown and only a little is known about the mechanisms regulating the expression of *COX-2* [99]. However, Henry F. L'Eplattenier et al. reported that oncogenes, growth factors, cytokines, endotoxin and phorbol esters can up-regulate the expression of *COX-2* [99].
A noticeable difference between man and dog prostate cancer is the role of androgens [39]. In men the development and function of sex organs as well as the development of prostate cancer is dependent of androgens [97]. In contrast, the androgen receptor is present in normal canine prostatic tissue and is important for normal canine sex organ maintenance and function in dogs [39]. Contrary to humans, dogs have a rarely expression of the androgen receptor (AR) [97]. Due to the fact that castration does not decrease the incidence of prostate cancer and androgen receptor expression is not present in dogs, it is quite possible that androgens does not play an important role in the pathogenesis of canine prostate cancer as Chen-Li Lai et al. reported [39].
A marker which is often used in men is the prostate specific membrane antigen (PSMA) , but former studies reported that *PSMA* was not expressed in canine prostate cancer but recent work has shown the opposite [97].

1.1.3 Prostate cell lines

"Cancer cell lines represent useful tools to investigate tumorigenesis and to establish new therapeutic approaches. For research in rare but severe tumor entities like canine prostate cancer, cell lines are even more valuable, as access to tissue samples and primary cultures is limited. In general, cell lines are established from tumor-burdened individuals." (Packeiser et al. 2020 S.2) [42].

Cell lines are *in vitro* models which are used in medical research, especially in cancer research [43]. Cell lines can simulate the molecular reaction of a whole organism and provide a source of biological material [43]. Furthermore, under specific conditions cell lines retain a lot of genetic properties of the original cancer [88]. Cell lines also differ from primary tumors, which means the original or first tumor in the body, in biologically significant ways and not every cell line can be a adequate model for their tumor type [88]. Thus, finding differences between cell lines and primary tumor tissues is of great importance to assess if the cell line is a good model or not [44].
The advantages of cell lines are their inexpensiveness, easy growth, disposability and suitability for high-throughput screening [45]. However, cancer cell lines do not represent the heterogeneity of primary tumors and acquire specific properties during *in vitro* growth [45].
Cell lines are also of great importance in canine prostate cancer research but are currently limited to a rather small number [57]. All the seven used canine cell lines in this thesis show significant immunohistological differences compared to the respective canine tumor tissue samples [42]. This is due to the fact that the cell lines have undergone own variances induced by subculturing [42].

Overall seven different cell lines were used in this study and were described and characterized from Eva Maria Packeiser et al. [42] for the first time:

- **Adcarc 1258**
  Represents a multi-resistant canine male prostate adenocarcinoma cell line
- **Adcarc 0846**
  A prostate adenocarcinoma derived cell line that showed a significantly increased *CLDN-1* expression.

- **Adcarc 1508**
  A prostate adenocarcinoma cell line which tended to show a down-regulation of *miR-141* and *miR-375.*

- **Metadcarc 1511.2**
  A male metastasis-derived cell line of adenocarcinoma which is resistant to apoptotic effects induced by doxorubicin.

- **Metadcarc 1511.3**
  A male metastasis-derived cell line of adenocarcinoma and resistant to apoptotic effects induced by doxorubicin.

- **TCC 1509**
  Represents an extremely rare case of canine TCC existence in prostate tissue.

- **TCC 1506**
  A cell line from female bladder tissue which is classified as TCC after pathohistological examination of the initial tissues.

1.2 Gene expression analysis

The analysis and interpretation of transcriptomic data are the most complex issues scientists deal with [69]. It requires a high quality differential gene expression analysis, which means the identification of the differentially expressed genes between multiple sample groups or experimental conditions [46]. Gene expression is a regulated process in which the cell is able to react to its environment [48]. Moreover, gene expression is involved in the procedure of transferring genetic information from DNA fragments to RNA molecules further to build proteins [48]. In order to get better insights into the world of cellular reactions at a definite point in time, gene expressions are researched. By studying gene expression, i.e. in which cells under which environmental conditions a gene is active, the gene functions can be derived [69].

To find differentially expressed genes in terms of two or more experimental groups a computational analysis is necessary. Such a gene expression analysis is an essential technique for cancer research, discovering new drugs and treatment options [69]. With the increasing popularity and application of this analysis, several tools and methods have been developed [49]. The most commonly used methods to measure gene expression are Northern blotting, quantitative polymerase chain reaction (qPCR), DNA microarray and RNA-Seq [56]. In this study the used method for finding the differential expressed genes is RNA-seq.

1.3 RNA-sequencing

RNA-sequencing (RNA-seq) was first established in 2008 and from that moment on RNA-seq data has grown exponentially [50]. RNA-seq can discover and quantify complex biological processes, find expressed genes, analyse RNA–protein bindings, determine sequence variations, characterize transcriptomes or identify transcripts [40].

In an RNA experiment RNA is obtained from a tissue or cell and a classic workflow of a RNA-seq experiment consists of several steps (Figure 1). The experiment includes isolation of RNA from a cell or tissue sample, the library construction, the sequencing and the bioinformatic analysis [53]. Initially, the isolation and purification of the cellular RNAs is done [54]. The isolated RNAs have to be treated with DNase to remove contamination by genomic DNA [54]. Then purity and quantity checks are implemented to ensure the quality of the isolated RNAs [52]. In the library construction step the aim is to convert RNAs into a library of complementary DNAs (cDNAs) fragments and adapters because of the cDNAs superior chemical stability for further sequencing [51]. Due to the specific sequencing length of each platform, the DNA has to be split into fragments [55].

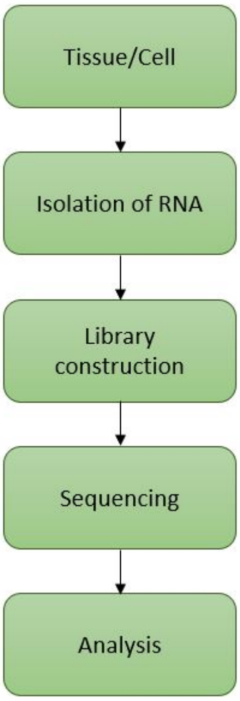


Figure 1: **The classic RNA-seq workflow [52]**. Such an experiment starts with the isolation of RNA from a specific cell or tissue and the constructed RNA-seq library is sequenced. After sequencing, a computational analysis is required.

The aim in sequencing is to receive the nucleic acid sequence from the cDNA library [52]. Each molecule is sequenced with a high-throughput sequencing technology to produce millions of short sequence reads [55]. With these it is possible to analyze the transcriptome in a high qualitative way [55]. There are two methods of sequencing:

- single-end sequencing
- paired-end sequencing

DNA fragments can be sequenced either on one end, called single-end sequencing, or both ends, called paired-end sequencing (Figure 2). The paired-end sequences have the benefit of increasing the randomisation of fragments and it could be, that short fragments overlap, which means additional information in comparison with single-end sequences [52].

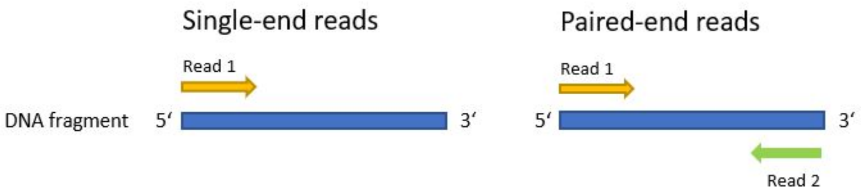


Figure 2: **Single-end sequencing and paired-end sequencing.** Paired-end sequencing allows sequencing of both ends of the DNA fragment and single-end means sequencing only one end.

High-throughput sequencing technologies can produce millions of short sequence reads and it is therefore possible to analyze the transcriptome in a quantitative way using RNA-seq [52].

1.3.1 Processing RNA-seq data

After the sequencing, the RNA-seq data has to be processed (Figure 3). Initially a quality check of the input data should be performed in order to look at the overall quality of the millions of reads [58]. In the next pre-processing step, the quality of the reads have to be improved, hence removing low-quality bases and artifacts such as adapter or library construction sequences from raw reads, called trimming is done [52]. This step can consist of several steps like quality trimming, adapter trimming and removing experimental artefacts [58]. After the alignment of the pre-processed reads to a reference genome, the differential expressed genes can be calculated and listed with a differential gene expression analysis. For all these steps, several tools and software was developed.

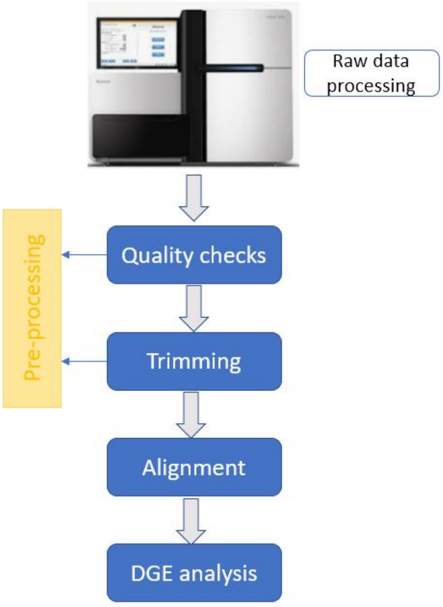


Figure 3: **Processing RNA-seq data**. Initially pre-processing of the raw reads is important, including the quality control and trimming step. Furthermore, the reads have to be aligned to the reference genome. Furthermore a differential gene expression (DGE) analysis is required to find the differential expressed genes.

However, each of these steps use different bioinformatic tools and it is very common to implement a pipeline including all those steps and tools. An in-house RNA-seq pipeline allows full control of the version of tools used, having only a few restrictions on input data types and also allowing to reproduce every single step [59].

1.3.2 RNA-seq pipeline tools

Due to the multitude of different applications and analysis scenarios of RNA-Seq, there is no optimal pipeline for every application. The selection of analysis strategies and software tools regarding RNA-seq depend on the organism to be examined and the respective research goal [101]. RNA-seq pipelines can be distinguished by methods, software, different annotations, run-time parameter values and normalization methods [101]. When RNA-seq is used in a differential gene expression analysis project the reads have to go through several steps (Figure 3) [52]. In order to execute all these steps automatically or partially automatically, so-called workflow management systems are widely used [100]. To ensure reproducibility and reusability worfklow engines help to automate pipelines [100]. Workflow management systems like Biopipe [102], Galaxy [103], GeneProf [104] or PegaSys [105] are characterized by the easy-to-use graphical user interface [61]. GXP Make [108] is quite simliar to Snakemake [73], which is inspired by the build system GNU Make [60, 110]. For all of these, the workflow consists of numerous rules with input and output files [60]. The syntax of Snakemake is close to the pseudocode of the Python language [100]. One significant advantage of Snakemake over e.g. PegaSys is the cooperation with any installed tool, software or obtainable web service with clear-defined input and output files, thus it is more flexible [61]. After all, Snakemake is one of the first systems which supports file name inference with several named wildcards in rules [61].

The most time-consuming step in such a workflow is the alignment to the reference genome [100]. Alternatively, a pseudo alignment to a transcriptome is performed, which recently gained in popularity because of the higher speed and accuracy [100].

There are many appropriate tools available for each step of a RNA-seq pipeline, but there is no gold standard regarding the optimal RNA-seq analysis pipeline. A big number of RNA-seq pipelines were developed because of the wide range of applications that RNA-seq has. Researchers use different RNA-seq pipelines depending on the research goal. Known pipelines like RNAflow [75] or the pipeline provided by the nf-core community (nf-core/rnaseq) uses Nexflow as workflow management system. RNAflow pipeline focuses on easy installation, execution and reproducibility and provide reasonable results with a minimal set of input parameters usable also for non-experts [75]. The pipeline OneStopRNAseq [76] offers a web application and the back-end is implemented within the Snakemake framework [76].The desired type of analysis and data required by the user, can be selected or changed in the analysis path.

Another common pipeline is VIPER [77], where the underlying framework enables easy and efficient rerunning of the analysis. Further feature is the graphical summary report that allows quick summarization of experimental results. Furthermore, the integrated batch-correction, Virus detection and Immunology modules are unique in this pipeline compared to others [77].

The QuickRNASeq [78] pipeline is designed for simplicity and visual interactivity and is applied for large-scale RNA-seq analysis of complex data sets [78].

The structure and procedures of those different pipelines are often similar, mostly different tools are used for the respective steps.

1.3.3 Pre-processing tools

In order to create a quality report in the beginning or after the mapping step several tools like FastQC [74], which is the standard tool and HTQC [111] were developed. A pretty typical tool for bioinformatic pipelines is MultiQC [81], it is used to combine all reports into one large report. Another tool for quality control and reports is NGSQC [112].

For removing low-quality reads and trim adaptor sequences software tools such as Trimmomatic [82], FASTX-Toolkit [113] and rCorrector [114] are popular [69].

1.3.4 Alignment Tools

One of the most popular aligners for RNA-seq reads is HISAT [115], a extremely efficient software for aligning reads from RNA sequencing experiments. For detecting non-canonical splice junctions, a aligner such as STAR [83] or GEM [116] are used because of their enormous accuracy. Pertea et al. [62] emphasized that HISAT is one of the fastest aligners among the rest. But according to Baruzzo et al. [63] STAR gives better performance than HISAT on human data sets, consisting of a higher percentage of correct aligned reads and a higher precision rate on a simulated human data set [63].

The length and type of reads, strandedness of the RNA-seq library and the length of sequenced fragments are crucial variables to define which aligner fits best[69]. Due to the fact that the reads

are short and a high accuracy is needed, the used aligner is STAR.
For generating a quality report after aligning the reads Qualimap [117] or samtools [118] can be used.

1.3.5 Differential gene expression analysis tools

A number of tools using the negative binomial model such as edgeR [119], DESeq2 [127] or baySeq [120] have been developed in order to identify the genes that are differentially expressed [69]. Other tools like NOIseq [121] or SAMseq [122] adopt non-parametric techniques [69]. However, a big difference in such differential gene expression tools is that some can only perform a pair-wise comparison, but others can perform multiple comparisons including covariates and analyzing time-series data like limma-voom [125], DESeq [126], DESeq2 [127], and maSigPro [128, 69].
Normalization methods which ignore highly expressed features or highly variables are TMM [129], DESeq [126] or PoissonSeq [130].
In this further analysis the tool DESeq2 is used, a method for identification of differentially expressed genes of count data. DESeq2 uses a negative binomial distribution and this model is defined as [127]:

$$Y_{it} = NB\sim\ (\mu_{it}, \phi_t)$$

$$\mu_{it} = s_i \cdot q_{it}$$

$$log(q_{it}) = X_i \cdot \beta_t$$

Where $\mathrm{X_i}$ is the design matrix, μ_{it} the fitted mean and a gene-wise dispersion parameter ϕ_{t}.
β_{t} is a vector with the coefficients, giving the $\log_2$ fold-changes for gene i for each column of the model matrix X [127]. $\mathrm{q_{it}}$ is proportional to the true concentration of fragments and by normalizing the the raw counts, the scaling factors ($\mathrm{s_{it}}$) are calculated [64]. By calculating a median fold change between samples for all genes with positive expression, compared to a geometric mean $\mathrm{Y_i}^{\mathrm{R}}$, the scaling factor is estimated [64]:

$$\tilde{s}_{\mathrm{i}} = \frac{Y_{it}}{Y_i^r}$$

$$Y_i^R = (\prod_{i=1}^{m} Y_i^R)^{1/m}$$

m = total number of samples
DESeq2 model needs to estimate the coefficients β_{t} and the dispersion parameter ϕ_{t} from the mean model [127]. The estimation process consists of three steps, beginning with a transcript-wise dispersion estimation , involving fitting a negative binomial generalized linear model (GLM) to the count data [64]. This GLM utilizes a dispersion that is estimated by method-of moments and is needed to obtain a fitted mean μ_{it} for every transcript [64]. Due to the fact that genomic studies are often very complex, linear models are used [127]. DESeq2 will test if a model parameter differs significantly from zero using a Wald test, dividing the estimated log fold change by its standard error [127].

1.4 Aims and Objectives

The aim of these master thesis is to construct a pipeline to pre-process RNA-seq data and perform a differential gene expression analysis of a canine prostate cancer dataset generated by the group of Prof. Dr. Ingo Nolte from the Tierklinik Hannover. Finding the differentially expressed genes between canine tissues and cell lines of different patients with either a diagnosis of an prostate adenocarcinoma or a transitional cell carcinoma is the purpose.
Particularly, this master thesis has the following objectives:

1. Pre-process the RNA-seq data
2. Analyze an RNA-seq dataset consisting of canine cancer samples and perform a differential gene expression analysis
3. Interpret results and compare with findings in the literature

2 Materials and Methods

Many methods and software have been developed recently to pre-process the raw data and to find out the differentially expressed genes in the course of RNA-seq.

2.1 Data

The data used in this master thesis was obtained by the group of *Prof. Dr. Ingo Nolte* from the Tierklinik Hannover [41] and consists of overall 14 used data samples. Isolation of total RNA and library preparation of those sample are described by Eva-Maria Packeiser as follows:
"For cell line Adcarc1258 and tissue samples P1 and P3, sequencing data of previously published experiments were included (Gene Expression Omnibus database accession identifier [65] GSE122916, samples PT-1 for P1 and PT-6 for P3). For cell line Adcarc1258, sequencing data of triplicates that served as solvent control (0.15 % V/V DMSO) were used from a previous study (data submitted for publication). Apart from the DMSO treatment of Adcarc1258, all cell pellets including those from previous studies were treated equally concerning sampling, storage, RNA isolation, library preparation and sequencing, as were the tissue samples. RNA from cell pellets was isolated using the RNeasy Mini Kit (Qiagen, Hilden, Germany), in accordance with manufacturer's protocols. For tissue samples, the AllPrep DNA/RNA/miRNA Universal Kit (Qiagen) was utilized as previously described [65]. RNA was quantified photometrically on a take 3 plate in a Synergy2 plate reader (BioTek, Bad Friedrichshall, Germany). Samples with RNA integrity numbers £5.2 measured with RNA 6000 Nano LabChip on an Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, USA) were further processed for library preparation using the NEBNext Ultra RNA preparation kit (New England Biolabs, Ipswich, USA). Tissue sample P3 was excluded due to low RNA integrity numbers in three independent trials of RNA-isolation. Single read sequencing was conducted on an Illumina NextSeq500 platform (Illumina, San Diego, USA) with a read length of 75bp."(Eva Maria Packeiser, 2019)

In the used dataset six canine PAC cell lines and tissue samples as well as two canine TCC cell lines and tissue samples from a total of six patients (dogs) were used (Table 1). For each patient a tissue sample as well as a cell line were used because of the aim to find differentially expressed genes between each cell line type to its tissue of origin while controlling for differences across patients. The cell lines were profiled by immunophenotype in comparison to respective original tumor tissues, which were taken from neoplastic sections [42].
From patient no.4 two cell lines and two tissue samples of lymph notes metastasis were obtained. Cell lines such as Metadcarc1511.2 and Metadcarc1511.3 offer the excellent option to analyse the cellular characteristic regarding PAC metastasis because the cell lines were derived from metastasis and from primary tumor of the same patient [42].
Patient 5 is missing on the list due to the patient's owner denying the animals necropsy [42].

The Ensembl annotation (Release 101) (Hunt et al. 2018) for canis lupus familiaris and the DNA (Release 101) (Hunt et al. 2018) (assembly CanFam3.1) from the Ensembl FTP server, were downloaded in GTF format and in FASTA format, respectively [66].

2.2 Software

The execution and implementation of the pipeline was done on a machine with 64 CPU cores running the Linux (x86_64) operating system. The generated pipeline for RNA-Seq analysis consists of several software (Table 2). Furthermore RStudio (v.1.3) [85] was used for running R Scripts (v.3.6.3) for differential gene expression analysis and visualization.

Table 1: **Used canine data set with overall 14 different samples.** The whole data set consists of six different patients (dogs) with at least one cell line and a tissue sample. The diagnosis of these patients are either trasitional cell carcinoma (TCC), prostate adenocarcinoma (PAC) or prostate adenocarcinoma metastasis in lymph nodes.

Patient	Sample ID	Sample type	Diagnosis	Original Tissue	Sequencing method
P1	P1	Tissue	PAC	Prostate	paired-end
P1	Adcarc1258	Cell line	PAC	Prostate	paired-end
P2	P2	Tissue	PAC	Prostate	single-end
P2	Adcarc0846	Cell line	PAC	Prostate	single-end
P3	P3	Tissue	PAC	Prostate	paired-end
P3	Adcarc1508	Cell line	PAC	Prostate	single-end
P4	Ln4.2	Tissue	PAC metasthasis	Lymph nodes	single-end
P4	Metadcarc1511.2	Cell line	PAC metasthasis	Lymph nodes	single-end
P4	Ln4.3	Tissue	PAC metasthasis	Lymph nodes	single-end
P4	Metadcarc1511.3	Cell line	PAC metasthasis	Lymph nodes	single-end
P6	P6.1	Tissue	TCC	Prostate	single-end
P6	TCC1509	Cell line	TCC	Prostate	single-end
P7	B7	Tissue	TCC	Bladder	single-end
P7	TCC1506	Cell line	TCC	Bladder	single-end

Table 2: **Used Software in the RNA-seq pipeline.** For the pipeline five different tools are used and their version and availability are listed.

Software	Version	Availability
Snakemake [73]	5.10.0	https://snakemake.readthedocs.io/en/stable/getting_started/
FastQC [74]	0.11.9	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trimmomatic [82]	0.39	http://www.usadellab.org/cms/?page=trimmomatic
STAR [83]	2.7.6a	https://github.com/alexdobin/STAR
Anaconda [72]	1.7.2	https://www.anaconda.com/

2.2.1 Snakemake

Snakemake (v.5.10.0), a workflow management system, useful for creating reproducible and scalable data analysis [73]. The so-called *Snakefile* is a simple text-file which is compiled and executed with the Snakemake compiler from shell [73]. A *Snakefile* consists of several rules, including input files, output files and the corresponding shell commands:

```
rule name_of_rule:
    input:
        "path/to/sample/.fastq.gz"
    output:
        "path/to/output/sample.fastq.gz"
    shell:
        "shell command {input} {output}"
```

Rules can be executed in parallel for saving time and increasing efficiency. The recommended installation of Snakemake happens via Anaconda (v.1.7.2), which is an open source distribution of the programming languages R and Python for data science applications [72].

2.2.2 FastQC

The used tool for checking the quality of the raw data is FastQC (v0.11.9) [74]. It generates a *.html report for each sample containing a modular set of analysis, which give a quick overview to assess the data [80]. A standard report consists of several analysis and statistics like:

1. General Statistics
2. Sequence Counts
3. Sequence Quality Histograms
4. Sequence Quality Scores
5. Base Sequence Content
6. Sequence GC Content
7. Per Base N Content
8. Sequence Length Distribution
9. Sequence Duplication Levels

2.2.3 Trimmomatic

Trimmomatic (v.0.39) [82] is a command line tool for trimming and croping the raw reads in FASTQ format. This tool performs quality trimming if the quality is poor at the beginning or at the end of each read and removes too short reads. Again, single-end reads and paired-end reads have to be treated differently. For single-end reads one input file and one output file is defined. For paired-end reads two input files and two output files are specified. The trimming step can be removed from the pipeline if the quality of the data is really good, thus no quality trimming has to be done.

2.2.4 STAR

The used tool for mapping the RNA-seq data was STAR (v.2.7.6a)[83], 'Spliced Transcripts Alignment to a Reference', an ultra fast universal RNA-seq aligner [83]. The main goal was to discover the origin location of each read on a given reference. Alignment is divided into two major steps:

1. Generate genome index: first the reference genome must be indexed so that reads may be quickly aligned.
2. Align reads to reference genome

In the first step the reference genome sequences in FASTA format as well as the annotation files were used to generate genome indexes, allowing a quick retrieval of the positions in the reference genome. When the positions is established, the reads can be mapped to the reference genome. Again, distinguishing between single-end and paired-end reads was necessary, based on the different number of FASTQ files.
After mapping several output files were generated [83]:

1. ReadsPerGene.out.tab:
   in this file the counted number of reads per gene while mapping is reported and a read is counted if it overlaps one gene.

2. SJ.out.tab:
   those files include high confidence collapsed splice junctions, which means each splicing is counted and summerized in SJ.out.tab.

3. Log.progress.out:
   the current job progress statistics are reported. Statistics such as percentage of mapped reads or the number of processed reads are updated in one minute intervals.

4. Log.out:
   it is the main log file with a lot of information if mapping has completed successfully. Log.out-files are mostly used when there are issues with the alignment or for troubleshooting and debugging.

5. Log.final.out:
   After alignment to a reference is completed, a summary of mapping statistics is generated. Such statistics are very useful for quality control and contain the number of input reads, the number of uniquely mapped reads, the number of splices, the mismatch rate per base, the percentage of unmapped reads and much more. For each read such statistics were calculated and summarized over all reads, considering STAR counts a paired-end read as one read.

6. Aligned.sortedByCoord.out.bam:
   For saving time, STAR can output alignments directly in binary BAM format instead of SAM format. BAM files can be unsorted or sorted by coordinates, which is required by many downstream applications. Aligned paired-end reads as well as multiple alignments of a read are always adjacent.

2.3 Differential gene expression analysis

After pre-processing steps of the raw FASTQ-files a count matrix was obtained with all samples and corresponding number of sequence fragments assigned to each gene in csv-format. The counts of the reads should be un-normalized. The whole gene expression analysis was created in R with the DESeq2 (v.1.30.0) [127] package from Bioconductor (v.3.10), providing methods for finding expressed genes by use of negative binomial models. Additionally, a samples file, a table with sample information and conditions to compare, is required to map a treatment condition to each sample. In this analysis the main focus was on finding differentially expressed genes between cell lines and tissues for each diagnosis (PAC, TCC and PAC metastasis). Thus, three analysis were implemented for the three types of diagnosis to find the differentially expressed genes.

This analysis started with the count matrix, the samples table and the design formula in order to use the DESeqDataSetFromMatrix() method to create the DESeq2 object. The design formula indicates how to model the samples, in this case the effect of cell lines and tissue samples for each patient:

$$design =\sim \text{Patient} + \text{Sample type}$$

The design formula means that DESeq2 will test the effect of the sample type (the last factor), controlling for the effect of patients (the first factor), so that the algorithm returns the fold-change result only from the effect of sample type.

Before runing DESeq2, it is essential to choose an appropriate reference level, a baseline level of a factor. This can be done by the *relevel-* function in DESeq2. In this particular case the reference was the tissue. Prefiltering, which means removing rows in which there are very few reads should

be done. In this script a minimal pre-filtering is performed to keep only rows that have at least one read total.
When generating the DESeq2 object, first, a internal normalization of the matrix is performed, which means calculating the geometric mean for each gene across all samples [127]. DESeq2 uses shrinkage estimation for dispersions and fold-changes, hence the dispersion value is estimated for each gene using a model fit process [127]. For fitting, a negative binomial generalized linear model is used for each gene and for significance testing the Wald test is used [127]. Count outliers are automatically detected and removed from the analysis. To get the final results table, the results function from DESeq2 was used providing a DESeq2 results object, which is a simple subclass of a data frame. This data frame contains the columns: *baseMean*, *log2FoldChange*, *lfcSE*, *stat*, *pvalue* and *padj* for each gene. The *lfcSE* represents the standard error of the $\log_2$ fold-change, which means how much a quantity changes between two conditions. For instance a $\log_2$ fold-change of 2.5 signifies that the gene's expression is increased by a multiplicative factor of $2^{2.5}$. The base mean is the mean of normalized counts of all samples and is used only for estimating the dispersion of each gene. The stats column describes the difference between a chi-squared distribution and the deviance of the full model to the reduced model, generating the P-value [67]. This parameter indicates the probability that a fold-change can be seen by the null hypothesis [67]. The adjusted P-value (padj) is the smallest significance level in which a certain comparison will be indicated statistically significant.
Only genes with an adjusted P-value below 0.05 and an absolute $\log_2$ fold-change value greater than one were acknowledged as differentially expressed.

Furthermore, the only annotation in the results table of the differentially expressed genes is the Ensembl Gene ID, which is not very informative. There are a lot of ways to add annotations, like gene names, transcript types or chromosome names. The package which was used for adding annotations is an R package called biomaRt (v.2.46.0) [133]. The Biomart database which was used is the Canis lupus familiaris gene ensembl (CanFam3.1). The information e.g columns which are added to the results table is the gene name and the transcript type.

2.4 Visualization

All the resulting graphs and plots of the analysis were created in R with RStudio.[85] For visualization the packages DEGreport (v.1.26.0) [94] and ggplot2(v.3.3.2) [95] were used.

2.5 Functional enrichment analysis

To gain greater biological insight of the differentially expressed genes a gene ontology (GO) functional enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was performed using the R package ClusterProfiler [134]. This tool compares a list of differentially expressed genes to a reference set, to assess the significance of enrichment for previously annotated and defined processes [145]. It aims to understand which pathways the differentially expressed genes are implicated in and the Gene Ontology (GO) categories describe the cellular component (CC), molecular function (MF) and the biological process (BP). Determine whether there is enrichment of known biological functions, pathways or networks.

3 Results

This chapter describes the constructed pipeline for pre-processing RNA-seq data and a gene expression analysis of canine prostate cancer samples.

3.1 RNA-seq pipeline

The implemented RNA-seq pipeline contains all the listed steps (Figure 4), starting with the raw reads and finishing with the count matrix, a table containing the number of mapped reads for all samples with the corresponding genes.

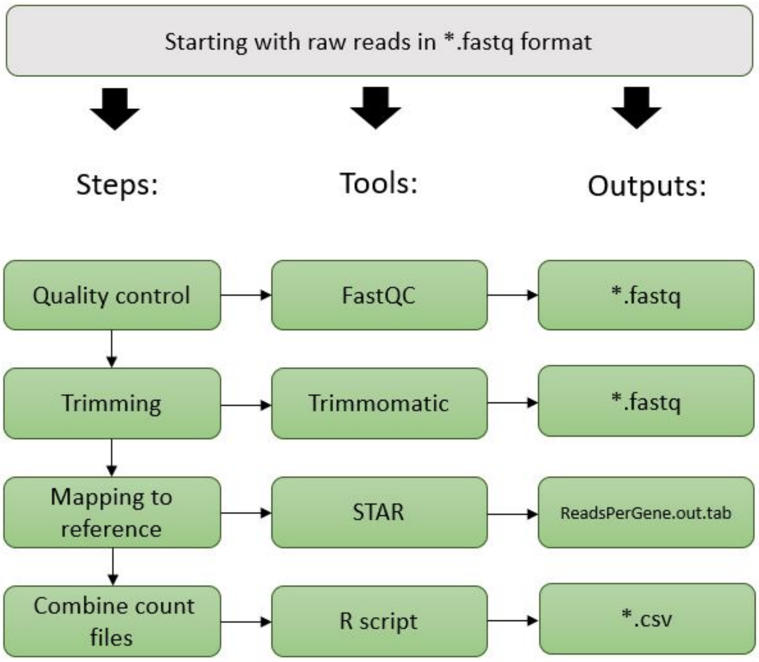


Figure 4: **Workflow of the RNA-seq pre-processing pipeline.** Starting with the raw data, some quality checks are performed. Trimmomatic performs quality trimming if the quality is poor at the beginning or at the end of each read and removes too short reads. With the popular STAR [83] aligner the reads are mapped to a reference genome and counted how many reads are mapped to a specific gene. The final output of the pipeline is a table containing the number of reads in each sequencing library mapped to each gene. This table is further referred to as ”count matrix”.

3.1.1 Configuration

A configuration file is available to set the parameters of the analysis and the user can change storage paths, name files, parameters and options only in this file without adjusting the pipeline (appendix).

3.1.2 Quality control

The raw reads are stored in FASTQ format and before analysing, some quality checks have to be done, to ensure no biases or problems in the data. Checking hundreds of sample reports can be enormously time consuming, therefore the tool MultiQC (v1.9) is used [81]. It merges the quality reports of each sample from FastQC tool into a single report. Due to the fact that single-end

libraries consists of one FASTQ file and paired-end libraries of two files, as a result they have to be analysed separately.
The mean quality score (Figure 5) of all single-end sequences is a quick overview of the range of quality across all samples, in which the y-axis shows the quality score and the x-axis the position (bp). The green area signifies very good quality, a phred score between 28-40, the orange one reasonable quality, a phred score between 20-28 and the red area means poor quality, a phred score between 0-20.
A phred quality score indicates the probability of the base being called correctly and the scores generally range from 0 to 40 with higher scores indicating greater confidence in the call, here an average phred score of 35 (Figure 6) was established.

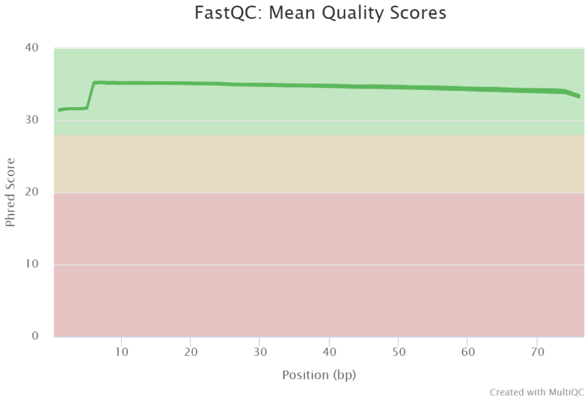


Figure 5: **Mean quality score of single-end reads of the canine data set.** The used tool for checking the quality is FastQC and for merging all quality reports in one plot MultiQC is used. For each single-end library there is one green line in the mean quality score plot.

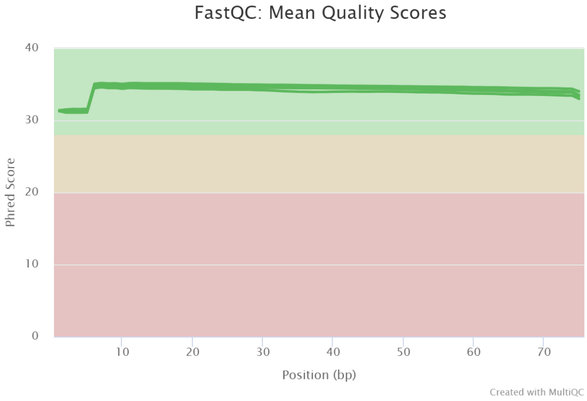


Figure 6: **Mean quality score of paired-end reads.** The used tool for checking the quality is FastQC and for merging all quality reports in one plot MultiQC is used. For each paired-end library there is one green line in the mean quality score plot.

Each file contains between 18 - 30 million sequences, the median across all single-end files is 25,9 million sequences and across all paired-end files it is 22,75 million sequences.

3.1.3 Mapping to the reference genome

One of the most important metric in alignment is the percentage of the reads that have been uniquely mapped to the reference transcriptome (Table 3).

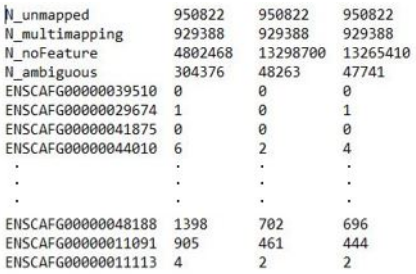
Table 3: **Results after mapping step.** The number and percentage of uniquely mapped reads to the reference genome for each sample

Patient	Sample type	Average read length	Uniquely mapped reads in %	Uniquely mapped reads
P1	Tissue	149	92,12%	21,976,280
P1	Cell line	149	88,25%	59,631,632
P2	Tissue	75	88,28%	20,511,330
P2	Cell line	75	88,42%	50,973,287
P3	Tissue	149	91,15%	20,385,207
P3	Cell line	75	89,26%	71,949,494
P4	Tissue	75	87,06%	19,810,641
P4	Cell line	75	89,82%	70,317,952
P4	Tissue	75	87,84%	19,847,040
P4	Cell line	75	89,58%	69,178,384
P6	Tissue	75	88,60%	25,793,133
P6	Cell line	75	89,60%	68,014,141
P7	Tissue	75	88,22%	20,858,108
P7	Cell line	75	89,34%	67,871,610

The mean in percentage of all reads which have been aligned to the reference genome is 89,11% and the median is 89,03%. The high amount of mapped reads points to successful trimming and alignment steps as well as data generation.

3.1.4 Counting the reads mapped to each gene

With the *–quantMode GeneCounts* command from STAR in the pipeline, a file called ReadsPerGene.out.tab is generated for each sample, counting how many reads are mapped to a specific gene (Figure 7). If one, only one read overlaps (1nt or more) one gene, it is counted.



```
N_unmapped          950822   950822   950822
N_multimapping      929388   929388   929388
N_noFeature         4802468  13298700 13265410
N_ambiguous         304376   48263    47741
ENSCAFG00000039510  0        0        0
ENSCAFG00000029674  1        0        1
ENSCAFG00000041875  0        0        0
ENSCAFG00000044010  6        2        4
 .                  .        .         .
 .                  .        .         .
 .                  .        .         .

ENSCAFG00000048188  1398     702      696
ENSCAFG00000011091  905      461      444
ENSCAFG00000011113  4        2        2
```

Figure 7: **Excerpt of reads per gene file.** This file, generated from STAR, contains of four columns with the different strandedness options and the gene ids.

This file consists of 4 columns with the different strandedness options listed in the *STAR Manual* [84]. The strandedness is dependent upon the cDNA library construction method, hence a stranded RNA-seq library involves information on which strand the messenger RNA (mRNA) originated, while in unstranded RNA-seq libraries the information is lost:

1. column 1: gene ID
2. column 2: number of mapped reads for unstranded RNA-seq
3. column 3: number of mapped reads for the 1st read strand aligned with RNA
4. column 4: number of mapped reads for the 2nd read strand aligned with RNA [84]

The first four rows consist of:

1. N_unmapped: the number of reads which are unmapped to the genome
2. N_multimapping: the number of reads which are multimapping to the genome
3. N_noFeature: the number of reads which are mapped to the genome but do not belong to a feature
4. N_ambiguous: the number of reads which belong to more than one feature

Selecting the correct strandedness for the specific data is important, otherwise incorrect results can arise in further analysis. The used data set comprises unstranded reads, hence column two. An R Script can be found in the appendix where the particular read counts of each sample are merged to one file.

3.1.5 Count matrix

For this data set, the count matrix consists of 30,424 genes (rows) and 14 samples (columns). 25,140 (82,63%) of 30,424 genes have values other than zero for at least one sample. In all, 19,947 genes are protein-coding genes, where 18,843 protein-coding genes have values other than zero.

3.2 Gene expression analysis

3.2.1 Overview of the data set

The whole dataset consisted of fourteen samples:, seven cell lines and seven tissue samples from overall six different patients. The principal component analysis (PCA) is a statistical procedure that can be used to reduce the dimensions of a data set and computes a description of the data set with a reduced number of variables. PCA plot is useful for visualizing the overall effect of experimental covariates and batch effects. In the PCA plot the differences between the cell lines and tissues can be seen (Figure 8). For this study, the tissue samples grouped together like the cell line samples. The differences between the two groups of samples inidcate that the expression profiles of the two groups are dissimilar.

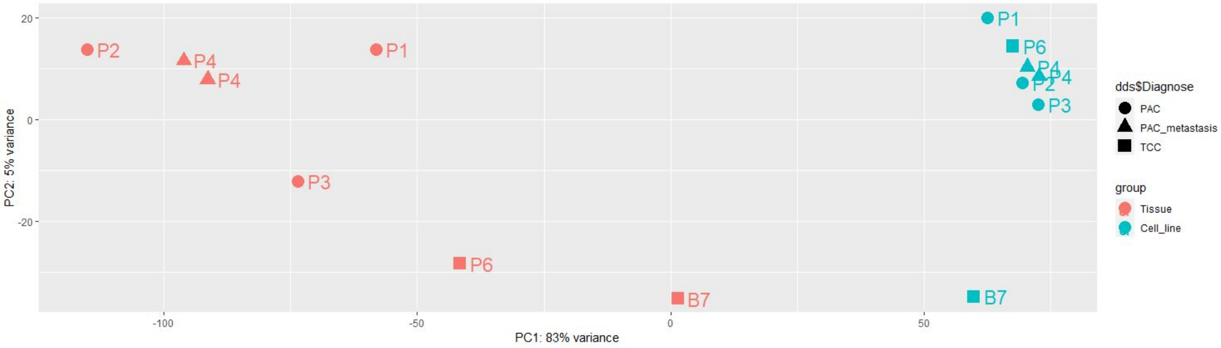


Figure 8: **PCA plot of the data set with 14 samples.** Each point represents a sample, either a tissue sample (red) or a cell line sample (blue) with the respective diagnosis and patient. Similar samples are usually seen grouped close together.

In the data set patients with three different diagnosis (PAC, TCC and PAC metastasis) can be found. Finding differentially expressed genes between cell lines and tissue samples was done seperately for each diagnosis. As a result, three differential gene expression analysis were performed and for each the same methods and steps were executed:

1. Before running the analysis, filtering of the data was done for all samples by excluding genes which have no values other than zero.
2. DEGs were identified between cell lines and tissue samples using the threshold for adjusted P-value smaller than 0.05 and $\log_2$ fold-change greater than or equal to one.
3. Up- and downregulated genes were identified.
4. Hierachical clustering
5. GO functional enrichment analysis
6. KEGG pathway enrichment analysis

3.2.2 PAC

To detect important alterations of a cell line compared to the relevant tissue sample a differential gene expression analysis was performed for the diagnosis PAC. Three tissue samples and three cell lines corresponding to three patients were included in this analysis. Overall, 22,979 genes were quantified with one or more reads/fragments/ in each of the six samples. Before running the analysis, filtering of the data was done by excluding genes which have not values other than zero. Thus, the filtered count matrix regarding the samples with diagnosis PAC consisted of overall 22,979 genes.

Using the threshold of 0.05 for adjusted P-value and $\log_2$ fold-change greater than or equal to one, 2,690 genes were found to be significantly differentially expressed including 129 up-regulated genes and 2,561 (95.2%) down-regulated genes (Table 4, Figure 9). In this comparison 2,535 differentially expressed genes (94,23%) were protein-coding genes. In the following evaluations and statistics, only differentially expressed protein-coding genes were considered.

The most up-regulated gene regarding the highest $\log_2$ fold-change of the canine data samples with diagnosis PAC was the *CYP1A1* gene (Table 5), while the most down-regulated gene regarding the lowest $\log_2$ fold-change *C1QC* gene was found (Table 6).

Table 4: **Number and percentage of differentially expressed genes for patients with diagnosis PAC.** A total of 2,690 genes were found to be differentially expressed between the cell line and the tissue samples, including 129 up-regulated genes and 2,561 down-regulated genes.

Genes	number	percent of all DEGs
Up-regulated	129	4.8%
Down-regulated	2561	95.2%
Overall	2690	100%

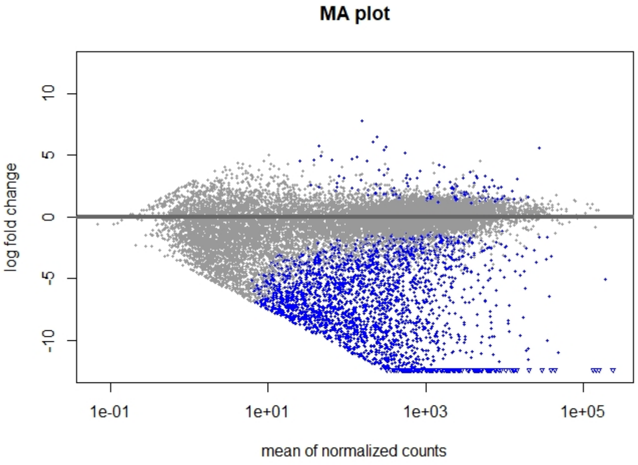


Figure 9: **MA plot of all genes of patients with diagnosis PAC.** A positive $\log_2$ fold-change means a gene is up-regulated in a cell line compared with the tissue sample, a negative $\log_2$ fold-change implies a gene is down-regulated of the cell line relative to the tissue sample. The blue dots show differentially expressed genes, grey dots indicate not differentially expressed genes.

Table 5: Ten up-regulated genes of the cell lines compared to the tissue samples in descending order of the $\log_2$ fold-change.

Gene	$\log_2$ fold-change	adj. P-value
CYP1A1	7.81	8.18×10^{-8}
FGF21	6.47	5.15×10^{-7}
IGFL3	6.04	6.18×10^{-4}
NQO1	5.62	7.22×10^{-6}
CFAP77	4.92	1.60×10^{-3}
TFF1	4.90	3.14×10^{-4}
CFAP65	4.71	1.75×10^{-3}
C15H4orf45	4.53	9.53×10^{-4}
FADS6	4.51	1.01×10^{-2}
TRIB3	4.02	1.20×10^{-6}

Table 6: Ten down-regulated genes of the cell lines compared to the tissue samples in descending order of the $\log_2$ fold-change.

Gene	$\log_2$ fold-change	adj. P-value
C1QC	-17.14	2.41×10^{-37}
LAMA4	-17.07	3.34×10^{-29}
COL1A1	-16.58	5.97×10^{-55}
BPI	-16.21	1.62×10^{-27}
C1S	-16.17	2.34×10^{-11}
SMPDL3A	-16.17	1.65×10^{-27}
C1QB	-16.14	1.65×10^{-27}
THY1	-15.82	1.46×10^{-25}
COL6A3	-15.81	3.66×10^{-52}
MMP2	-15.69	5.84×10^{-31}

Most of the adjusted P-values of all differentially expressed genes are less than 0.001. High bars at adjusted P-value smaller than 0.05 with a steep slop to baseline P-value levels are also a function of either uniformity of the data or a large sample size or both (Figure 10).

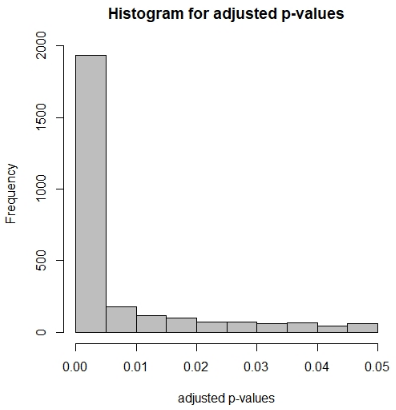


Figure 10: Histogram of the adjusted P-value < 0.05 of all differentially expressed genes

In the PCA plot of the samples with diagnosis PAC the differences between the cell lines and tissues can be seen (Figure 11). The cell lines were grouped together as well as the tissue samples grouped toward the positive end of the plot denoting similar inter-sample correlation.

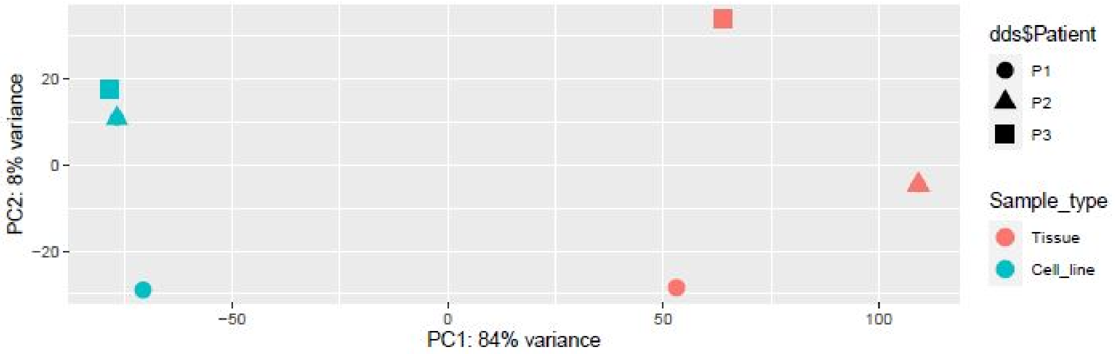


Figure 11: **PCA plot of differentially expressed genes between tissues and cell lines from patients with diagnosis PAC.**

Similar to PCA, clustering is also a method used to identify strong patterns in a data set. The correlation of gene expression for all pairwise combinations of samples is shown in a heatmap. All differentially expressed genes were analyzed by hierarchical cluster analysis based on their expression profiles (Figure 12). This heatmap showed that cell line samples as well as tissue samples cluster together, which suggests that they were similar (Figure 12).

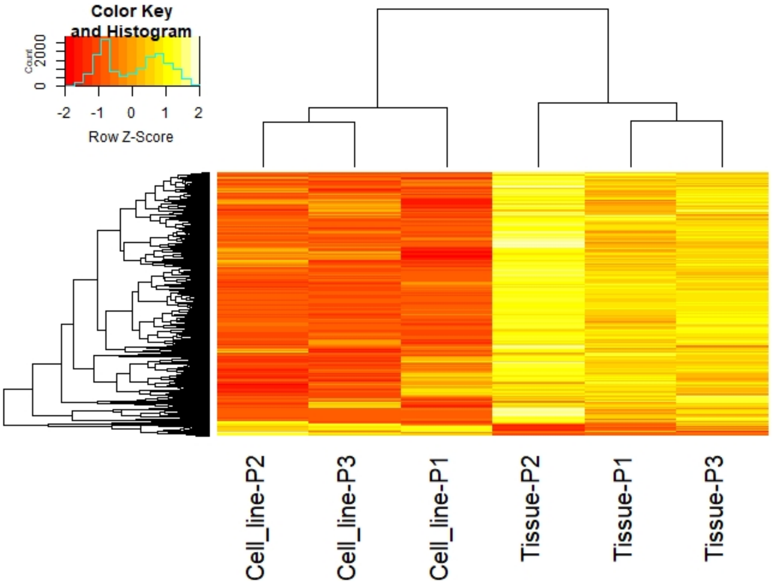


Figure 12: **Heatmap of the six samples and differentially expressed genes between cell lines and tissue samples with diagnosis PAC.** A hierarchical clustering on the regularized log (rlog) transformed expression matrix subsetted by all differentially expressed genes was performed. The aim of the rlog transformation is to remove the dependence of the variance on the mean. The hierarchical tree indicates seperation between cell line and tissue samples based on the normalized gene expression values. Each column represents a cell line or tissue sample from a specific patient (P1, P2 or P3) and each row represents a gene. Differences in expression were shown in different colors and negative numbers indicate down-regulation and a positive number means up-regulation.

To better understand which pathways the differentially expressed genes are implicated in a GO functional enrichment analysis was performed (Figure 13, 14, 15). The results showed that a total of 2535 differentially expressed genes were annotated into 758 GO terms, including 669 biological processes, 20 cellular components, and 69 molecular function annotations. Functional enrichment results suggested the differentially expressed genes were relevant to the biological processes of immune response, regulation of immune system process and defense response (Table 7). Enriched GO terms under the molecular functions with the smallest adjusted P-value were Transmembrane signaling receptor activity, immune receptor activity and cytokine binding (Table 7). The enriched GO terms regarding cellular components with the smallest adjusted P-value were extracellular matrix, cell surface and external side of plasma membrane (Table 7).

To know the signal pathways where the genes were distributed, a KEGG pathway enrichment analysis was performed and 80 enriched pathways were found, including cell adhesion, cytokine-cytokine receptor interaction and calcium signaling pathway, in order of increasing adjusted P-value (Figure 16).

Table 7: **Functional enrichment analysis representing five GO terms of each category (CC, MF, BP) of differentially expressed genes in the cell lines ordered by the adjusted *P*-value.**

ID	Description	Gene ratio	adjusted P-value	Count
Cellular components (CC)				
GO:0031012	Extracellular matrix	92/1259	3.72×10^{-33}	92
GO:0009986	Cell surface	134/1259	4.92×10^{-33}	134
GO:0009897	External side of plasma membran	87/1259	7.23×10^{-33}	87
GO:0098552	Side of membran	99/1259	1.19×10^{-24}	99
GO:0062023	Collagen-containing extracellular matrix	56/1259	9.65×10^{-24}	56
Molecular functions (MF)				
GO:0004888	Transmembrane signaling receptor activity	152/1189	1.76×10^{-17}	152
GO:0140375	Immune receptor activity	44/1289	1.34×10^{-15}	44
GO:0019955	Cytokine binding	43/1289	6.10×10^{-14}	43
GO:0004896	Cytokine receptor activity	38/1189	2.11×10^{-12}	38
GO:0005539	Glycosaminoglycan binding	36/1189	2.67×10^{-10}	36
Biological Processes (BP)				
GO:0006955	Immune response	201/1151	7.37×10^{-44}	201
GO:0002682	Regulation of immune system process	180/1151	5.74×10^{-33}	180
GO:0006952	Defense response	177/1151	2.96×10^{-31}	117
GO:0001775	Cell activation	136/1151	5.16×10^{-27}	136
GO:0022610	Biological adhesion	174/1151	2.96×10^{-26}	174

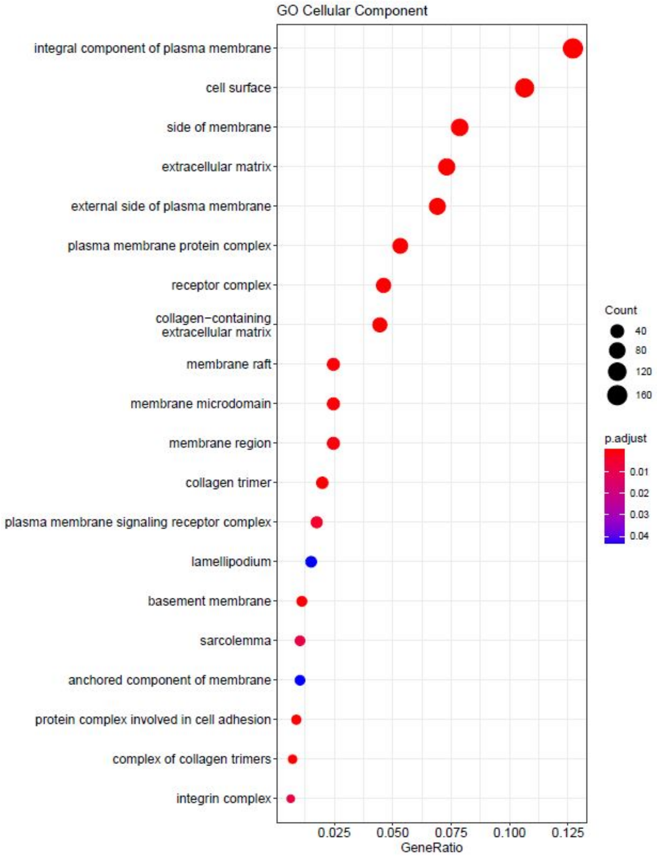


Figure 13: **Enriched GO terms of the differentially expressed genes of diagnoses PAC ordered by the number of counts.** This plot represents the enriched cellular components.

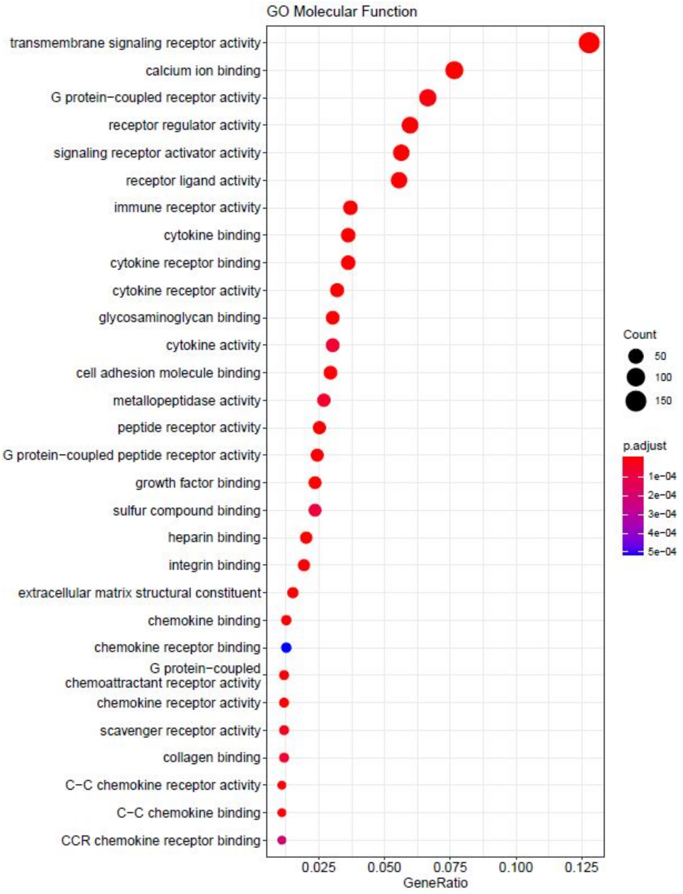


Figure 14: **Enriched GO terms of the differentially expressed genes of diagnoses PAC ordered by the number of counts.** This plot represents the enriched molecular functions.

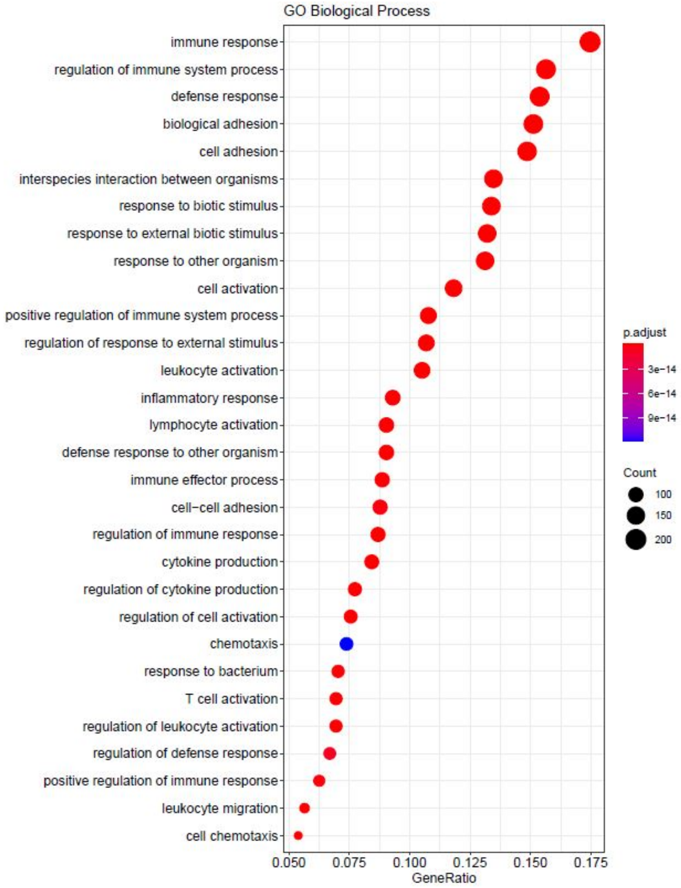


Figure 15: **Enriched GO terms of the differentially expressed genes of diagnoses PAC ordered by the number of counts.** This plot represents the enriched biological processes.

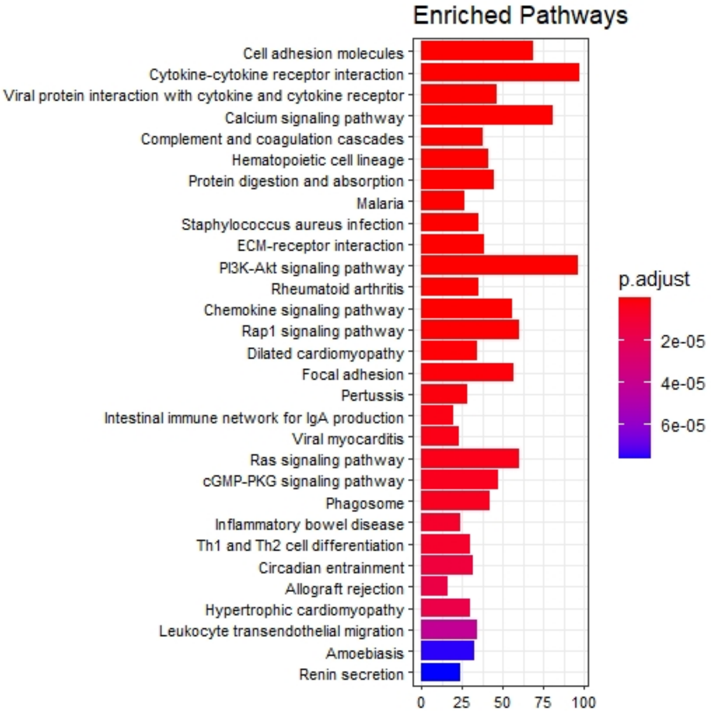


Figure 16: **KEGG pathway enrichment analysis of the differentially expressed genes of diagnosis PAC.** The thirty enriched pathways ordered by increasing adjusted *P*-value are mentioned.

3.2.3 TCC

The second comparison was between cell lines and tissue samples from patients with the diagnosis TCC. Overall there are two patients with a cell line and a tissue sample regarding TCC, hence four samples were used. The filtered count matrix regarding the samples with diagnosis TCC consisted of overall 22,075 genes. Overall, 723 genes were found as significantly differential expressed, indicating 67 up-regulated (9.3%) genes and 656 down-regulated (91%) genes of the cell lines compared to the tissues (Table 7, Figure 17). In this comparison 705 differentially expressed genes (97.5%) were protein-coding genes. In the following evaluations and statistics, only differentially expressed protein-coding genes were considered. The gene *SPINK5* was found to be a up-regulated gene with the highest $\log_2$ fold-change and the *C1QC* gene was found to be a down-regulated gene with the lowest $\log_2$ fold-change (Table 8 and 9).

Table 8: **Number and percentage of differentially expressed genes for patients with diagnosis TCC.** A total of 723 genes were found to be differentially expressed of the cell line comparing to the tissue samples, including 67 up-regulated genes and 656 down-regulated genes

Genes	number	percent
Up-regulated	67	9.3%
Down-regulated	656	91%
Overall	723	100%

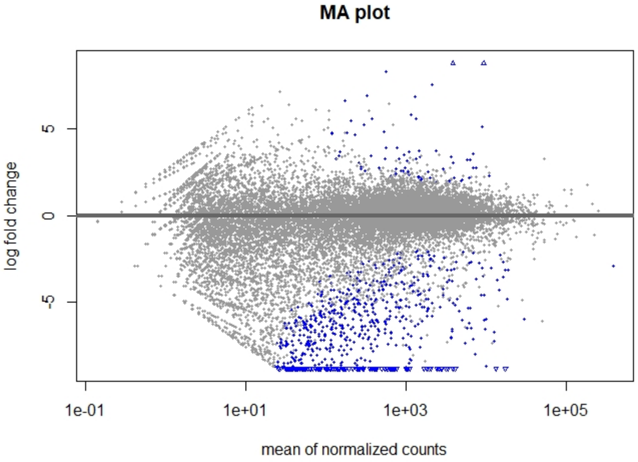


Figure 17: **MA plot of all genes of patients with diagnosis TCC.** A positive $\log_2$ fold-change means a gene is up-regulated in a cell line compared with the tissue sample, a negative log $_2$ fold-change implies a gene is down-regulated of the cell line relative to the tissue sample. The blue dots show differentially expressed genes, grey dots indicate not differentially expressed genes.

In the histogram most of the adjusted P-values of all differentially expressed genes of patients with diagnosis TCC were less than 0.001. There is a sharp spike near zero that drops off into a somewhat uniform distribution, which indicates that some significant differences were found between cell line and tissue samples (Figure 18) .

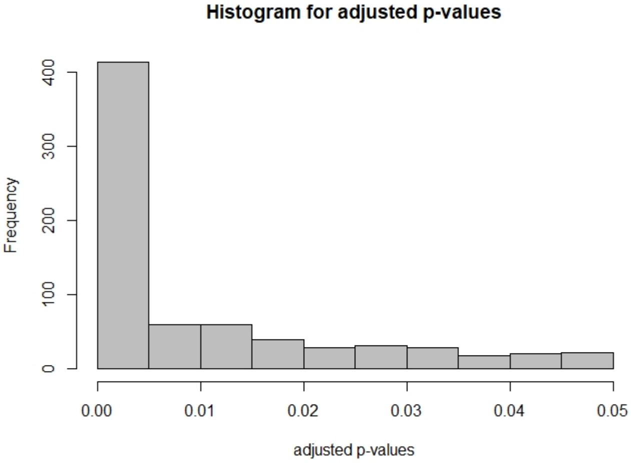


Figure 18: Histogram of the adjusted P-values from differentially expressed genes with diagnosis TCC

Table 9: Ten up-regulated differentially expressed genes of the cell lines compared to the tissue samples for diagnosis TCC ordered by $\log_2$ fold-change.

Gene	$\log_2$ fold-change	adj. P-value
SPINK5	8.82	1.14×10^{-15}
FGF21	8.32	1.24×10^{-8}
FDCSP	6.91	1.66×10^{-2}
CA4	6.87	1.34×10^{-14}
GABBR2	5.84	1.40×10^{-7}
GPC3	5.60	5.39×10^{-7}
KRTDAP	5.60	3.31×10^{-3}
CYP2A13	5.51	8.20×10^{-3}
DNM3	5.23	2.50×10^{-3}
PSAT1	5.15	4.20×10^{-9}

Table 10: Ten down-regulated differentially expressed genes of the cell lines compared to the tissue samples for diagnosis TCC ordered by $\log_2$ fold-change.

Gene	$\log_2$ fold-change	adj. P-value
C1QC	-15.15	1.34×10^{-2}
C1QB	-14.83	1.45×10^{-2}
TNFSF11	-14.36	1.99×10^{-2}
C1QA	-14.27	2.96×10^{-2}
DLA-DRA	-13.68	8.30×10^{-30}
DLA-DQA1	-13.53	5.40×10^{-26}
RGS1	-12.99	1.27×10^{-13}
VCAM1	-12.70	2.01×10^{-11}
DLA-79	-12.60	9.40×10^{-11}
CD34	-12.55	1.40×10^{-10}

All differentially expressed genes were analyzed by hierarchical cluster analysis based on their expression profiles (Figure 19), indicated a similarity of cell line samples and tissue samples due to the clustering.

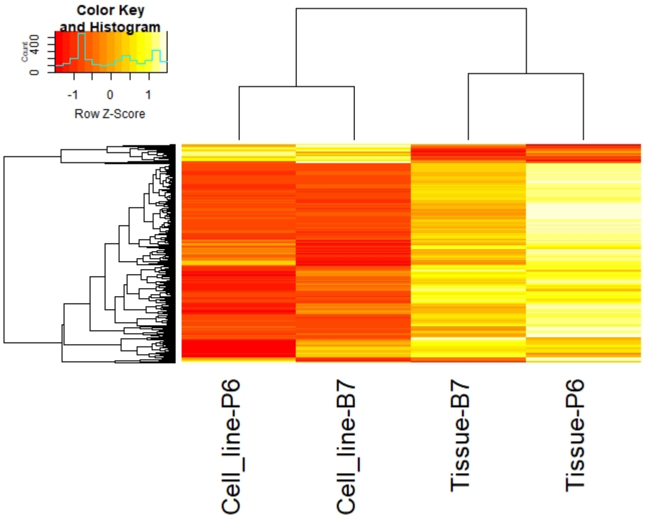


Figure 19: **Heatmap of the differentially expressed genes between cell line and tissue samples regarding diagnosis TCC.** Each column represents a cell line or tissue sample from a specific patient(P6 or P7) and each row represents a gene. Differences in expression were shown in different colors and negative numbers indicate down-regulation and a positive number means up-regulation.

To further explore the biological functions of the 705 differentially expressed genes a GO functional enrichment analysis was performed. The results showed that a total of 705 differentially expressed genes were annotated into 343 GO terms, including 289 biological processes, 17 cellular components, and 37 molecular function annotations. Functional enrichment results demonstrated the differentially expressed genes were relevant to the biological processes of defense response, immune response, as well as biological and cell adhesion (Table 11). The main functional terms of molecular function were calcium ion binding, signal receptor activator activity and receptor regulator activity (Figure 21, 22). Integral component plasma membran, extracellular matrix and cell surface were significantly enriched in cellular components (Figure 20).
To know the signal pathways where the genes were distributed, a KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analysis was performed and overall 44 enriched pathways were found, including cell adhesion molecules, PI3K signaling pathway and ECM-receptor interaction (Figure 23).

Table 11: **Functional enrichment analysis representing five GO terms of each category (CC, MF, BP) ordered by tje adjusted P-value of all differentially expressed genes in the cell lines of diagnosis TCC.**

ID	Description	Gene ratio	adjusted P-value	Count
Cellular components (CC)				
GO:0031012	extracellular matrix	39/360	3.97×10^{-17}	39
GO:0062023	collagen-containing extracellular matrix	26/360	1.12×10^{-13}	26
GO:0005581	collagen trimer	17/360	1.12×10^{-13}	17
GO:0009986	cell surface	36/360	1.21×10^{-6}	36
GO:0098644	complex collagen trimers	6/360	1.17×10^{-5}	6
Molecular functions (MF)				
GO:0005509	calcium ion binding	35/332	6.97×10^{-5}	35
GO:0061135	endopeptidase regulator activity	15/332	6.97×10^{-5}	15
GO:0005201	Extracellular matrix	9/332	6.97×10^{-5}	9
GO:0004866	Endopeptidase inhibitor activity	14/332	1.07×10^{-4}	14
GO:0030414	Peptidase inhibitor activity	14/332	1.25×10^{-4}	14
Biological Processes (BP)				
GO:0006952	Defense response	64/320	2.64×10^{-14}	64
GO:0006955	Immune response	64/320	5.11×10^{-14}	64
GO:0032101	Regulation of response to external stimulus	47/320	2.28×10^{-11}	47
GO:0002684	Positive regulation of immune system process	45/320	3.51×10^{-11}	45
GO:0022610	Biological adhesion	59/320	7.19×10^{-11}	59

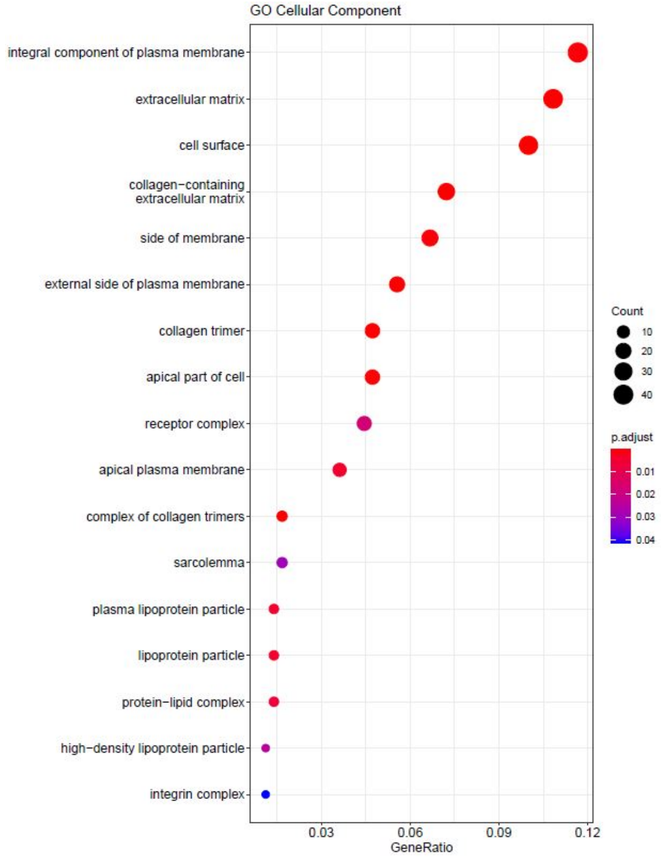


Figure 20: **Enriched GO terms of the differentially expressed genes of diagnoses TCC ordered by the number of counts.**This plot represents the enriched cellular components.

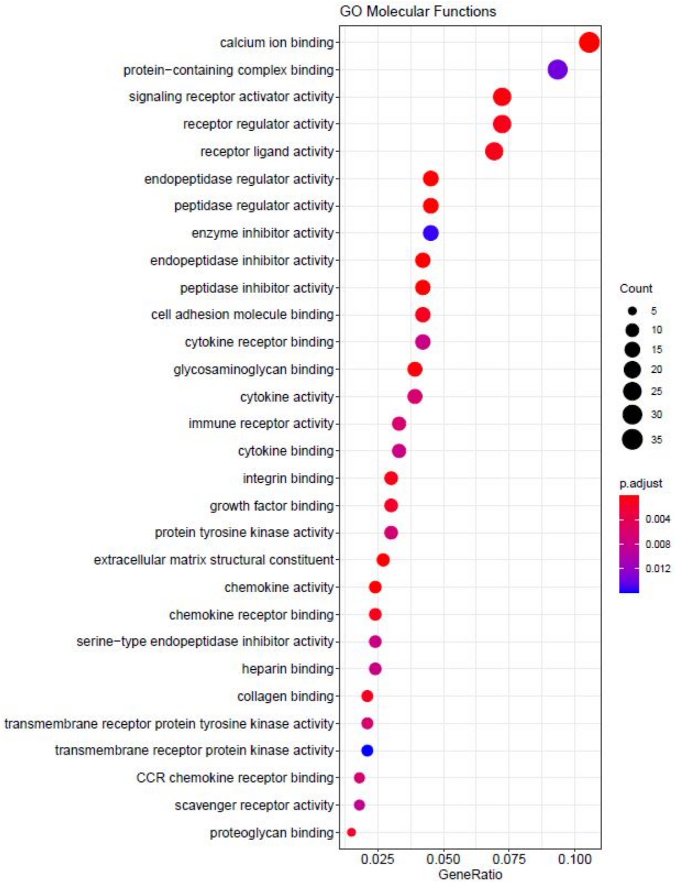


Figure 21: **Enriched GO terms of the differentially expressed genes of diagnoses TCC ordered by the number of counts.**This plot represents the enriched molecular functions.

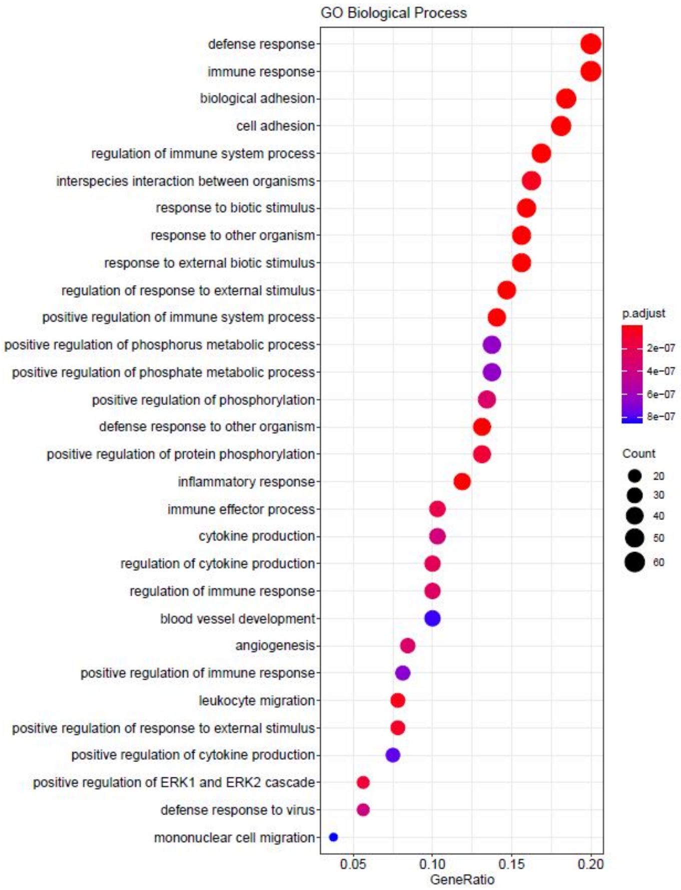


Figure 22: **Enriched GO terms of the differentially expressed genes of diagnoses TCC ordered by the number of counts.**This plot represents the enriched biological processes.

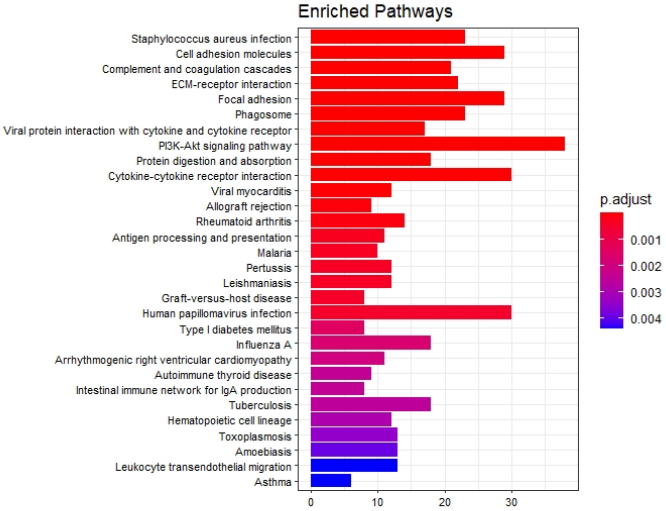


Figure 23: KEGG pathway enrichment analysis for the differentially expressed genes of diagnosis TCC ordered by the adjusted P-value.

3.2.4 PAC metastasis

The third comparison was between cell lines and tissues with the diagnosis PAC metastasis. Four samples, two cell lines and two tissues, from the same patient with the diagnosis PAC metastasis of the canine cancer data set were used. The filtered count matrix regarding the samples with diagnosis PAC metastasis consisted of overall 22,321 genes. Summarizing, 5,658 were found as differentially expressed with 1,783 up-regulated and 3873 down-regulated genes of the cell line compared to the tissue samples (Table 10). For further analysis, only the 5,330 differentially expressed genes with transcript type protein-coding were used. As well as the other two previous analyses, there were more down-regulated genes than up-regulated genes in the cell lines compared to the tissues (Figure 24). The up-regulated gene with the highest $\log_2$ fold-change and down-regulated genes of the canine data samples with diagnosis PAC metastasis were arranged in descending fold-change (Table 11 and 12).

Table 12: **Number and percentage of differentially expressed genes for patients with diagnosis PAC metastasis.** A total of 5,656 genes were found to be differentially expressed between the cell line and the tissue samples, including 1,783 up-regulated genes and 3873 down-regulated genes.

Genes	number	percent
Up-regulated	1783	32%
Down-regulated	3873	68%
Overall	5656	100%

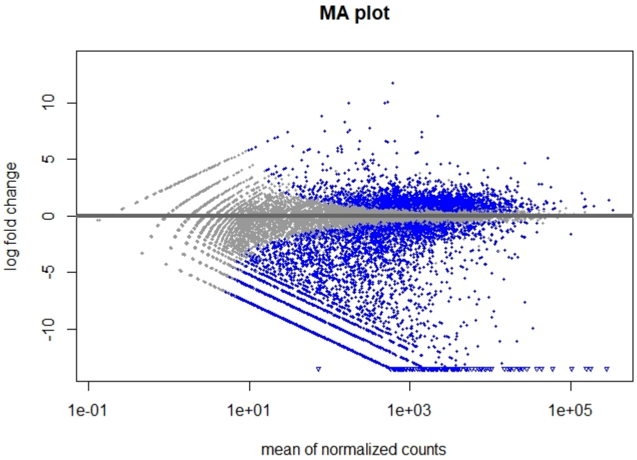


Figure 24: **MA plot of all genes of patients with diagnosis PAC metastasis.** The blue dots show differentially expressed genes, grey dots show not differentially expressed genes.

Furthermore, the differentially expressed genes were analyzed by hierarchical cluster analysis where genes with similar expression profiles clustered (Figure 29). This result suggested a similarity between cell line samples as well as tissue samples.

Another GO functional enrichment analysis was performed, to explore the differentially expressed genes concerning their biologial, molecular and cellular function. The results showed that 5,330 differentially expressed genes were annotated into 703 GO terms, including 643 biological processes,

Table 13: Ten down-regulated differentially expressed genes of the cell lines compared to the tissue samples for diagnosis PAC metastasis ordered by $\log_2$ fold-change

Gene	$\log_2$ fold-change	adj. P-value
DACT2	-23.92	3.45×10^{-6}
IGHM	-19.57	5.32×10^{-14}
IGKC	-18.81	1.19×10^{-10}
DCN	-18.62	2.04×10^{-6}
SFRP2	-17.59	4.38×10^{-7}
EFEMP1	-17.47	1.69×10^{-6}
MMP2	-17.17	9.50×10^{-12}
COL1A1	-17.04	4.18×10^{-30}
SPARC	-16.72	2.45×10^{-27}
Q1QC	-16.43	2.69×10^{-25}

Table 14: Ten up-regulated differentially expressed genes of the cell lines compared to the tissue samples for diagnosis PAC metastasis ordered by $\log_2$ fold-change.

Gene	$\log_2$ fold-change	adj. P-value
SALL3	8.35	1.91×10^{-7}
TFF1	7.58	1.35×10^{-8}
RAB3C	7.54	2.27×10^{-3}
PLPPR4	7.47	4.95×10^{-2}
FGF21	7.48	2.20×10^{-23}
GDPD4	7.03	1.01×10^{-3}
WNT7A	6.96	1.58×10^{-38}
COL10A1	6.63	4.45×10^{-7}
A3GALT2	6.62	6.48×10^{-39}
SCNN1G	6.60	1.85×10^{-11}

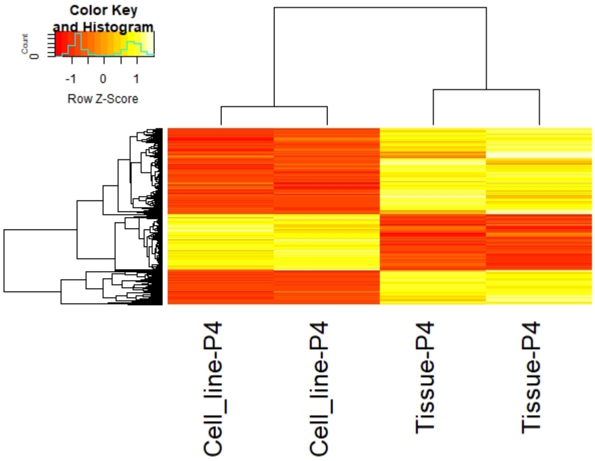


Figure 25: **The clustered heatmap of all differentially expressed genes in cell line and tissue samples with diagnosis PAC metastasis.** A separation can be observed between those two groups. Higher expression is shown in yellow, lower expression in red.

16 cellular components and 44 molecular functions (Table 15). As in the analysis of diagnosis PAC, the main biological processes were immune and defense response, as well as biological adhesion. Also similarities between the diagnosis PAC and PAC metastasis were found regarding molecular functions (Figure 26, 27). Cytokine binding as well as cytokine receptor activity is significant in both. Ordered by number of counts, again, immune response, biological adhesion and cell adhesion were found as enriched terms concerning biological processes (Figure 28).

Table 15: **Functional enrichment analysis representing five GO terms of each category (CC, MF, BP) of differentially expressed genes in the cell lines of diagnosis PAC metastasis ordered by the adjusted *P*-value.**

ID	Description	Gene ratio	adjusted *P*-value	Count
Cellular components (CC)				
GO:0009986	Cell surface	187/2624	1.73×10^{-26}	187
GO:0009897	External side of plasma membrane	106/2624	6.26×10^{-23}	106
GO:0098552	Side of membrane	134/2624	8.35×10^{-19}	134
GO:0031012	Extracellular matrix	93/2624	2.63×10^{-11}	93
GO:0062023	Collagen-containing extracellular matrix	56/2624	3.80×10^{-9}	56
Molecular functions (MF)				
GO:0140375	Immune receptor activity	61/2451	2.43×10^{-17}	61
GO:0004896	Cytokine receptor activity	55/2451	3.55×10^{-15}	55
GO:0019955	Cytokine binding	58/2451	9.90×10^{-13}	58
GO:0005539	Glycosaminoglycan binding	48/2451	8.32×10^{-9}	48
GO:0005126	Cytokine receptor binding	66/2451	2.56×10^{-6}	66
Biological Processes (BP)				
GO:0006955	Immune response	276/2396	4.67×10^{-29}	276
GO:0022610	Biological adhesion	261/2396	2.56×10^{-20}	261
GO:0006952	Defense response	249/2396	4.23×10^{-20}	249
GO:0007155	Cell adhesion	257/2396	7.67×10^{-20}	257
GO:0006954	Inflammatory response	146/2396	1.76×10^{-18}	146

To know the signal pathways where the genes were distributed, a KEGG pathway enrichment analysis was performed and overall 117 enriched pathways were found. The most enriched KEGG pathway terms in which the differentially expressed genes in the cell lines were significantly enriched were cytokine receptor activity, cell adhesion molecules and ECM-receptor interaction (Figure 29).

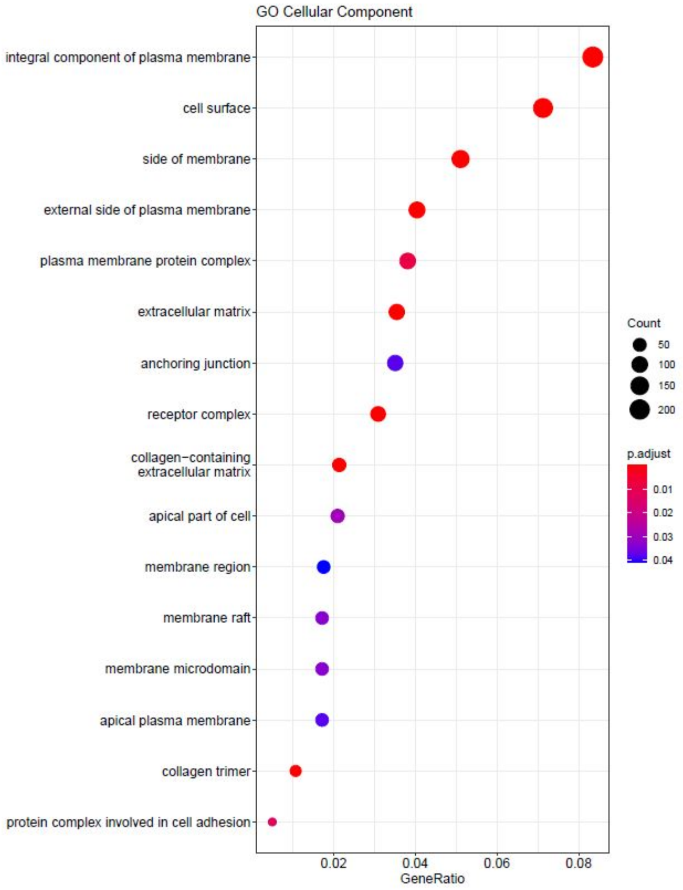


Figure 26: **Enriched GO terms of the differentially expressed genes of diagnoses PAC metastasis ordered by the number of counts.**This plot represents the enriched cellular components.

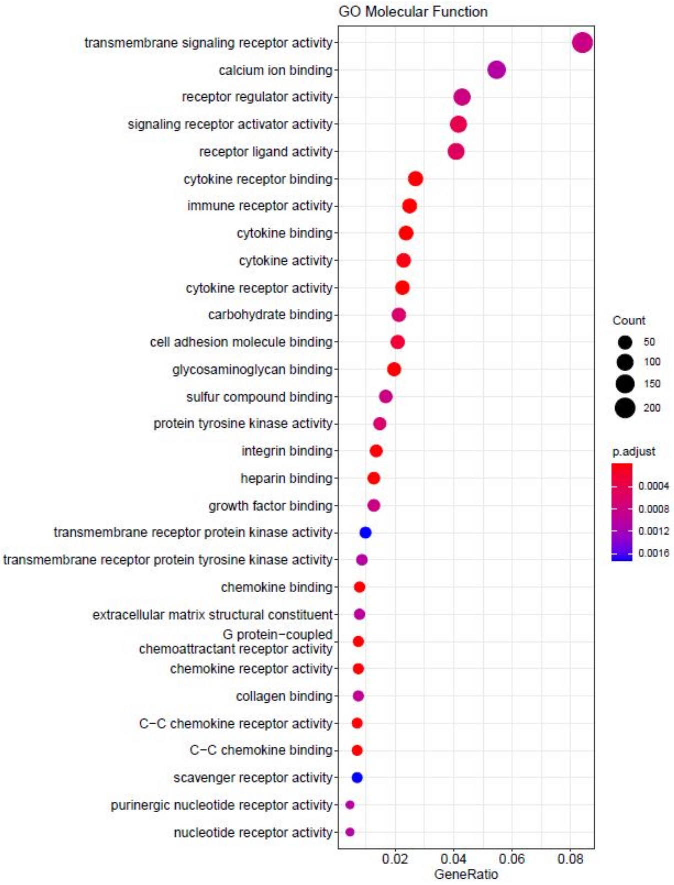


Figure 27: **Enriched GO terms of the differentially expressed genes of diagnoses PAC metastasis ordered by the number of counts.**This plot represents the enriched molecular functions.

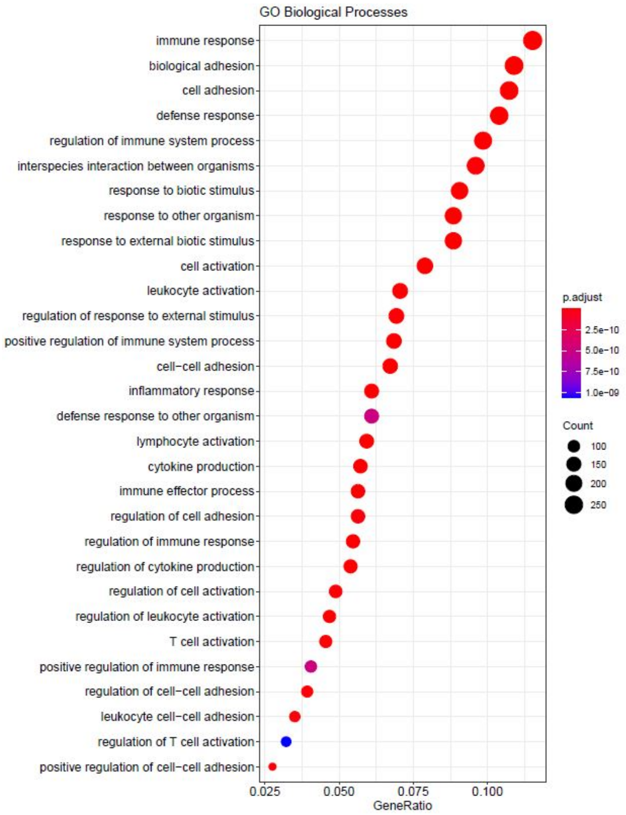


Figure 28: **Enriched GO terms of the differentially expressed genes of diagnoses PAC metastasis ordered by the number of counts.**This plot represents the enriched biological processes.

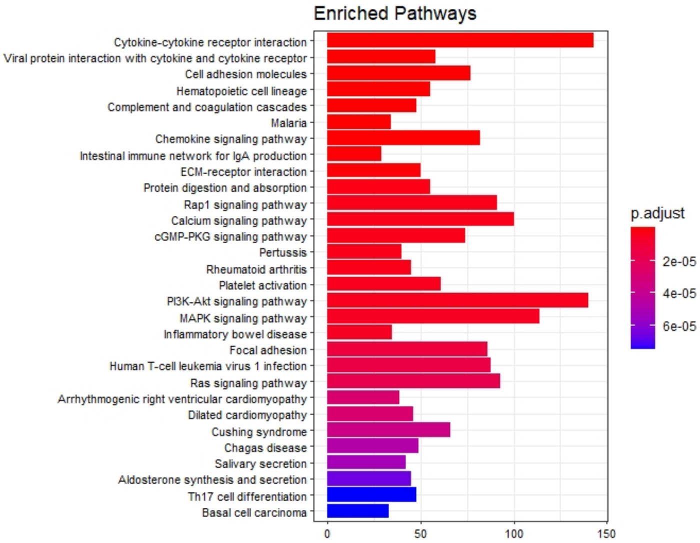


Figure 29: KEGG pathway enrichment analysis for differentially express genes with diagnosis PAC metastasis ordered by adjsuted P-value.

3.2.5 Comparison of the three diagnoses

To find out how many of the differentially expressed genes were specific to each diagnosis, the overlap between all differentially expressed genes in TCC, PAC and PAC metastasis. 82 genes were included exclusively in TCC, 382 genes were included exclusively in PAC and 3105 genes were included exclusively in PAC metastasis (Figure 30. Interestingly, 479 genes were common in all three groups, and even 1638 genes between the groups PAC and PAC metastasis (Figure30.

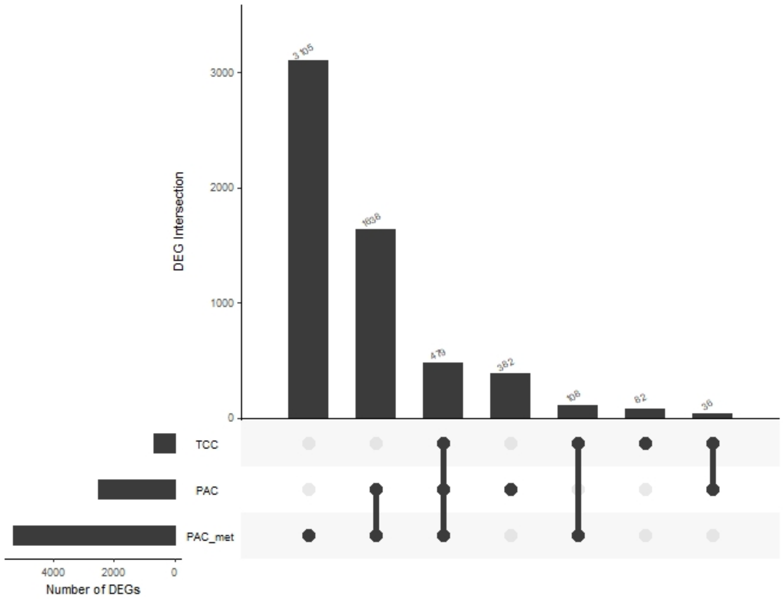


Figure 30: **The upset plot of differentially expressed genes of all diagnosis.** Intersection of DEGs, where each column corresponds to a diagnosis or set of diagnosis (the dots connected by lines below the x-axis) containing the same DEGs. The number of genes in each group can be found above the bars with the diagnosis on the left.

The GO functional enrichment analysis was performed for only the DEGs which were expressed in all three diagnosis. The results showed that the differentially expressed genes were annotated into 270 GO terms, including 236 biological processes, 8 cellular components and 26 molecular functions (Table 16).

Table 16: **Functional enrichment analysis representing GO terms of each category (CC, MF, BP) of differentially expressed genes in the cell lines which appear in each diagnosis.**

ID	Description
Cellular components (CC)	
GO:0009986	Cell surface
GO:0009897	External side of plasma membrane
GO:0098552	Side of membrane
GO:0031012	Extracellular matrix
GO:0062023	Collagen-containing extracellular matrix
Molecular functions (MF)	
GO:0140375	Immune receptor activity
GO:0004896	Cytokine receptor activity
GO:0050839	Cell adhesion molecule binding
GO:0004888	Transmembrane signaling receptor activity
GO:0005126	Cytokine receptor binding
GO:0005125	Collagen binding
GO:0030545	Signaling receptor regulator activity
GO:0008009	Chemokine activity
Biological Processes (BP)	
GO:0006955	Immune response
GO:0022610	Biological adhesion
GO:0006952	Defense response
GO:0007155	Cell adhesion
GO:0002682	Regulation of immune system process
GO:0001775	Cell activation
GO:0002252	Immune effector process
GO:0001816	Cytokine production
GO:0006954	Inflammatory response

4 Discussion

4.1 RNA-seq pipeline

In last few years RNA-seq has become the standard method for gene expression analysis and rapidly emerged as a replacement for microarray because of the higher sensitivity and dynamic range as well as lower technical fluctuations. The first aim of the master thesis was to construct a bioinformatic pipeline for RNA-seq data. The power of the pipeline is based on reproducibility and the inclusion of all important steps concerning RNA-seq. It allows the user to perform a full RNA-seq analysis starting from raw FASTQ-files sequencing data to a count matrix of the mapped reads, ready for differential gene expression analysis. Designed to be run easily and requiring only little configuration, the efficacy of the pipeline was validated on real sequencing data from the Tierklinik Hannover from dogs with prostatic cancer. The pipeline is portable and can be used for individual canine data analysis, if other species or genomes want to be used, the files can be changed.

In recent years there was a considerable effort of developing RNA-Seq tools and software. The immense number of obtainable tools has made the choice of a solid and stable computational pipeline difficult. The pipeline presented in this thesis is only using five tools and is characterized by simplicity. To assess sequencing data and decide about additional analyses or data processing the standard tool FastQC is used. The quality control tool FastQC returns two files for each samples, in this case 28 output files. It would be a burden to go through all the files for each sample, therefore the MultiQC tool was used to summarize all these output files to only one file. Poor quality of raw sequencing reads could be detected from this step. The removal of poor quality sequences is performed when trimming is desired. However, the user can specify the options for trimming with the tool Trimmomatic, but it is recommended to keep the numbers of removed bp as small as possible because the gene expression estimates can be changed by shorter reads. The user can also apply own thresholds or parameters regarding Trimmomatic, but this requires general knowledge about the trimming options. The same goes for the alignment step with STAR. Additional parameters can be added, but the default parameters are optimized and mostly sufficient for RNA-seq data sets. As mentioned in chapter 3.1.4, after the mapping step, the ReadsPerGene.out.tab files were generated for each sample consisting of 4 different columns with the different strandedness options. Here it is important that the user knows if the data is stranded or unstranded. The correct column of strandedness can be changed easily in the enclosed R script.

The limitations of the pipeline are the specific requirement for the input data. The pipeline is able to handle both, single-end and paired-end reads from RNA-seq dataset. Before using the pipeline, the user should be aware of which data is applied, because single-end and paired-end libraries can not be used at the same time. First, single-end and paired-end libraries must be saved in two separate folders and then the pipeline for the specific sequencing method can be used.

RNA-seq gives a better understanding of the cell transcriptome which has a big benefit in genetic research. In the next future, massive analysis of transcriptomes will become a routine not just in cancer research.

4.2 Differential gene expression analysis

The aim of the second part of the master thesis was to identify the genes that are differentially expressed between the canine cell lines and their tissues of origin, for each of the three diagnosis, using the count matrix from the RNA-seq pipeline.

When performing PCA, there was a clear clustering of samples from cell lines and tissues. However, performing differential gene expression analysis on each of the three diagnoses using a $\log_2$ fold-change equal or greater than one and an adjusted P-value smaller than 0.05 revealed specific patterns of gene expression.

The majority of all found differentially expressed genes were down-regulated compared to the tissue samples in all diagnoses. Down-regulated genes in all three diagnoses were the complement component 1q *C1Q* genes, which belong to Tumor Necrosis Factor super familiy. C1q can perform a lot of immune and non-immune functions and is able to induce apoptosis and growth supression of prostate cells [139]. According to Hong et al. C1q sustains the activation of tumor suppressor *WOX1*, which is needed for blocking cancer cell proliferation [139].

Another highly down-regulated gene was the group of collagens (*COL1A1*, *COL6AB*, *COL1OA1*), which have been reported to be over-expressed in metastatic prostate tumors [136, 137]. Collagen represents an obstacle to migration and facilitates the invasion and proliferation of cells [140]. According to Shuaishuai Xu et al. collagen is the major component of the tumor microenvironment and can influence tumor cell behavior through integrins, discoidin domain receptors, tyrosine kinase receptors, and some signaling pathways [141]. Furthermore collagens appear to be a tumor immunity regulator, a metastasis promoter as well as increasing tumor tissue stiffness [141]. This observed down-regulation of collagens in the cell lines could be due to the fact that the cell lines did not grow in the tumor microenvironment but *in vitro*.

Dapper homolog (DACT) 2, a member of *DACT* gene family, was also down-regulated in PAC metastasis cell lines. Shibao Li et al. suggested that *DACT-2* may be a potential tumor suppressor gene involved in the occurrence and development of tumors [142]. Shibao Li et al. observed a down-regulation of *DACT-2* in the cell lines too and claimed that this gene was frequently silenced by its promoter hypermethylation in prostate cancer, implying that the transcriptional silencing of DACT-2 may be one of the essential factors in the progression of prostate cancer [142]. The expression of this gene can be associated with promoter methylation, inferring that DNA methylation is the main regulatory mechanism of DACT2 inactivation [142].

Matrix metalloproteinase 2 (MMP2), a member of the MMP gene family, is known as a endcoder of the zinc-dependent enzymes capable of cleaving components of the extracellular matrix (ECM) and the molecules involved in signal transduction [138]. Here, in all analysis *MMP2* was down-regulated in cell lines compared to the specific tissue samples. Kun Liu et al. [138] reported the crucial role of the *MMP2* gene in the pathogenesis of the initiation, invasion, and metastasis of various tumors, such as lung and prostate cancer. It has been suggested that the MMP2 gene is associated with the affection of cell growth, the production of cell junction proteins, such as collagens, and the pathogenesis of metastasis and invasion [138].

Less genes were significantly up-regulated in cell lines across the tissues analyzed. Interestingly, the fibroblast growth factor 21 (*FGF21*) gene was up-regulated in all diagnoses. The *FGF21* gene is an important part of glucose regulations, lipid metabolism as well as cell growth. Overexpressing *FGF21* promote cell growth and migration considerable [143]. Therefore, that increasing *FGF21* expression in prostate cancer tissue or increasing circulating *FGF21* level can have an important impact on the proliferation and apoptosis of prostate cancer cells, and may become a new target for a specific treatment [143].

Furthermore, a Gene Ontology analysis on the differentially expressed genes was performed and the gene sets involved in immune processes and response. In cellular processes often the extracellular matrix is involved, which is often more collagen-rich and of increased stiffness. Altering extracellular matrix can support cancer growth and metastasis. As cell adhesion is related to cancer metastasis, we also found that it is significant in the analysis of PAC metastasis. Genes,

regarding cell adhesion molecules and membrane signaling proteins were down-regulated in cell lines compared with tissue sample.
However, gene sets were also involved in cell-cycle related pathways, which are continuously up-regulated in cell lines, perhaps reflecting *in vitro* culturing conditions. Characteristic cell culture medium is replete with cytokines, metabolites and growth factors with the difference that tissue cells in the body have to compete for.

In recent years, prostate cancer research has benefited greatly from the establishment and use of cell lines and their drug-resistant sub-lines. Also a growing effect regading isolating and establishing animal cell lines appeared.
However, all herein characterized cell lines also displayed some immunohistochemical differences to the respective tumor tissue which indicates that they have undergone individual changes through subculturing [34].
The results lead to conclude that the canine cell lines are dissimilar to the tissue samples and might be a limited tool for understanding canine prostate cancer biology, but it is important to understand that differences exist always using cell line models, since cell lines are not grown in a complex micro environment as tumor cells in vivo. The effects of the microenvironment are determined by myofibroblasts and some key cells of the immune system, which are likely the leading causes of expression differences, but studying these effects can allow an assessment of their relative importance for different drug responses [144]. Furthermore, cytokines should be studied when using cell lines because they can be the effectors of the tumor environment [144]. Thus, the choice of an cell line for a specific study depends mainly on the target and context of the project. It depends on factors like particular genomic alterations of interest as well as growth characteristics, for example maximal molecular similarity to tissue samples is desirable investigating drug sensitivity. Here, the cell lines only mirror a few of the molecular properties of primary tumors.
But the advantages of such cell lines are highly controlled conditions, homogeneity, discovery of molecular mechanisms, reproducibility as well as understanding the pathogenesis of prostate cancer [146].
The main limitations of cancer cell culture is the selection of phenotypic and genotypic cells during adaptation to *in vitro* conditions [146]. Furthermore the accumulation of mutations in cells over time in culture, a homogeneous population of cells as well as the isolation of cells from the tumor microenvironment are limiting factors [146].
Due to often similar interactions between cell lines and tissue samples, *in vitro* cancer models have gained a worldwide acceptance for a lot of therapeutic and diagnostic applications. Nonetheless, cancer cell line models have been, and will continue to be, the model for cancer studies, because cell lines with known genetic alterations will be also helpful in screening the efficacy of drugs with a particular genetic background. Thus, the present study can provide the basis for further research on prostate cancer and drug research.

5 Conclusion

The generation of cancer cell line models is currently still challenging, nevertheless, it has preclinical relevance, as cell lines have benefits for drug development and its usage. Due to the increasing use of sequencing technologies more information for each disease can be provided. Here, cell lines will be important translating all the sequenced data into new therapies and diagnostic tests. The best selection of the culture conditions and the best mimic of the microenvironment *in vivo* will be a catalyst to obtain new drugs and therapies for diseases.

References

[1] World Health Organization Regional Offer for Europe.
`https://www.euro.who.int/en/health-topics/noncommunicable-diseases/cancer/data-and-statistics` *Assessed on 04.11.2020*

[2] Geoffrey M Cooper. *The Cell: A Molecular Approach. 2nd edition.* Sinauer Associates; 2000. ISBN-10: 0-87893-106-6

[3] SEER Training Modules, Module Name. U. S. National Institutes of Health, National Cancer Institute.04.11.20 https://training.seer.cancer.gov/disease/categories/classification.html

[4] Seyed Hossein, Hassanpour Mohammadamin Dehghani. Review of cancer from perspective of molecular; Journal of Cancer Research and Practice

[5] Tarini Sinha et al. Tumors: Benign and Malignant. Cancer Therapy Oncology; Inernational Journal, 2018

[6] Leslie SW, Soon-Sutton TL, Sajjad H, et al. Prostate Cancer. 2020 Oct 28 In: StatPearls [Internet]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470550/

[7] Gunes Gundem, Peter Van Loo, Barbara Kremeyer, Ludmil B. Alexandrov el al. The Evolutionary History of Lethal Metastatic Prostate Cancer; Europe PMC Funders Group Author Manuscript.

[8] Tomas Ganz. Epithelia: Not just physical barriers; Departments of Medicine and Pathology, School of Medicine, University of California, Los Angeles, CA 90095

[9] Julien Lanoue, Gary Goldenberg. Basal Cell Carcinoma: A Comprehensive Review of Existing and Emerging Nonsurgical Therapies; Icahn School of Medicine at Mount Sinai Hospital, Department of Dermatology, New York, New York

[10] Mareike Alter, Uwe Hillen, Ulrike Leiter, Michael Sachse, Ralf Gutzmer. Current diagnosis and treatment of basal cell carcinoma; Journal of the German Society of Dermatology

[11] Howell JY, Ramsey ML. Squamous Cell Skin Cancer.; 2020 Aug 8; In: StatPearls

[12] Margaret C. Metts, MD, Jobe C. Melts, MD, Stephen J. Milito, MD, and Charles R. Thomas, Jr., MD. BLADDER CANCER: A REVIEW OF DIAGNOSIS AND MANAGEMENT; JOURNAL OF THE NATIONAL MEDICAL ASSOCIATION

[13] Brittany Katsinis, BS, RDMS. A Challenging Case of Poorly Differentiated Transitional Cell Carcinoma of the Kidney; Journal of Diagnostic Medical Sonography 2018, Vol. 34(4) 282–284

[14] Mahmut Ozsahin, Gamze Ugurluer, Abderrahim Zouhair. Management of transitional-cell carcinoma of the renal pelvis and ureter ; SWISS MED WKLY 2009;139(25–26):353–356

[15] SEER Training Modules, Module Name. U. S. National Institutes of Health, National Cancer Institute.04.11.20; https://www.cancer.gov/publications/dictionaries/cancer-terms/def/adenocarcinoma

[16] Mullangi S, Lekkala MR. Adenocarcinoma.; [Updated 2020 Sep 5]. In: StatPearls
https://www.ncbi.nlm.nih.gov/books/NBK562137/

[17] Cancer Research UK. https://www.cancerresearchuk.org/what-is-cancer/how-cancer-starts/types-of-cancer#carcinomas, 06.11.2020

[18] Guocan Wang, Di Zhao,Denise J. Spring and Ronald A. DePinho. Genetics and biology of prostate cancer; ISSN 0890-9369/18; www.genesdev.org

[19] SY Song, SR Kim, G Ahn HY Choi. Pathologic characteristics of prostatic adenocarcinomas: a mapping analysis of Korean patients; Prostate Cancer and Prostatic Diseases (2003) 6, 143–147

[20] Kang Cui, Xiangnan Li et al. Chemoprevention of prostate cancer in men with high-grade prostatic intraepithelial neoplasia (HGPIN): a systematic review and adjusted indirect treatment comparison; Oncotarget, 2017, Vol. 8, (No. 22), pp: 36674-36684

[21] Dingxiao Zhang, Shuhong Zhao et al. Prostate Luminal Progenitor Cells in Development and Cancer; Trends Cancer. 2018 November ; 4(11): 769–783. doi:10.1016/j.trecan.2018.09.003.

[22] Anne F. Fribourg, Karen E. Knudsen, Matthew W. Strobeck, Clint M. Lindhorst, and Erik S. Knudsen. Differential Requirements for Ras and the Retinoblastoma Tumor Suppressor Protein in the Androgen Dependence of Prostatic Adenocarcinoma Cells; Department of Cell Biology, University of Cincinnati College of Medicine, Cincinnati, Ohio 45267-0521

[23] Justin G. Mygatt, Adit Singhal, Gauthaman Sukumar, Clifton L. Dalgard and Johnan A.R. Kaleeba. Oncogenic Herpesvirus HHV-8 Promotes Androgen-Independent Prostate Cancer Growth; DOI: 10.1158/0008-5472.CAN-12-4196, Published September 2013

[24] Catherine S. Grasso, Yi-Mi Wu et al. The Mutational Landscape of Lethal Castrate Resistant ProstateCancer; Nature. 2012 July 12; 487(7406): 239–243. doi:10.1038/nature11125.

[25] Gerhauser et al.Molecular evolution of early onset prostate cancer identifies molecular risk markers and clinical trajectories; Cancer Cell. 2018 December 10; 34(6): 996–1011.e8. doi:10.1016/j.ccell.2018.10.016.

[26] The Cancer Genome Atlas ResearchNetwork. The molecular taxonomy of primary prostate cancer; Cell. 2015 November 5; 163(4): 1011–1025. doi:10.1016/j.cell.2015.10.025.

[27] Shyh-Han Tan, Gyorgy Petrovics and Shiv Srivastava. Prostate Cancer Genomics: Recent Advances and the Prevailing Underrepresentation from Racial and Ethnic Minorities; Int. J. Mol. Sci. 2018, 19, 1255; doi:10.3390/ijms19041255

[28] Dan et al. Integrative clinical genomics of advanced prostate cancer; Cell. 2015 May 21; 161(5): 1215–1228. doi:10.1016/j.cell.2015.05.001.

[29] Yiji Liao, Kexin Xu. Epigenetic regulation of prostate cancer: the theories and the clinical implications; Asian Journal of Andrology (2019) 21, 279–290; doi: 10.4103/aja.aja_53_18; published online: 7 August 2018

[30] Yiji Liao, Kexin Xu. Identification of key pathways and genes in PTEN mutation prostate cancer by bioinformatics analysis; BMC Medical Genetics (2019) 20:191 https://doi.org/10.1186/s12881-019-0923-7

[31] David P. Labbé and Myles Brown. Transcriptional Regulation in Prostate Cancer; Cold Spring Harb Perspect Med 2018;8:a030437

[32] Peck Yean Tan, Cheng Wei Chang et al. Integration of Regulatory Networks by NKX3-1 Promotes AndrogenDependent Prostate Cancer Survival; 0270-7306/12/$12.00 Molecular and Cellular Biology p. 399 – 414

[33] Melani AM Fork, Hugo Murua Escobar, Jan T Soller, Katharina A Sterenczak, Saskia Willenbrock, Susanne Winkler, Martina Dorsch, Nicola Reimann-Berg, Hans J Hedrich, Jörn Bullerdiek and Ingo Nolte. Establishing an in vivo model of canine prostate carcinoma using the new cell line CT1258; Published: 15 August 2008, doi:10.1186/1471-2407-8-240

[34] Eva-Maria Packeiser, Marion Hewicker-Trautwein, Heike Thiemeyer, Annika Mohr, Johannes Junginger, Jan Torben Schille, Hugo Murua Escobar, Ingo Nolte. Characterization of six canine prostate adenocarcinoma and three transitional cell carcinoma cell lines derived from primary tumor tissues as well as metastasis; PLoS ONE 15(3):e0230272. https://doi.org/10.1371/journal.pone.0230272

[35] Jill M. Keller, George R. Schade et al. A novel canine model for prostate cancer; 2013, https://doi.org/10.1002/pros.22642

[36] Masanori Kobayashi et al. MicroRNA expression profiling in canine prostate cancer; J. Vet. Med. Sci. 79(4): 719–725, 2017 doi: 10.1292/jvms.16-0279

[37] Bostwick, D.G. and J. Qian. High-grade prostatic intraepithelial neoplasia.; Mod. Pathol., 17: 360-379., 2004

[38] Bell, F.W., J.S. Klausner, D.W. Hayden, D.A. Feeney and S.D. Johnston, 1991. Clinical and pathologic features of prostatic adenocarcinoma in sexually intact and castrated dogs: 31 cases.; (1970-1987). J. Am. Vet. Med. Assoc., 199: 1623-1630.

[39] Chen-Li Lai, Rene van den Ham, Jan Mol, Erik Teske. Immunostaining of the androgen receptor and sequence analysisof its DNA-binding domain in canine prostate cancer; The Veterinary Journal 181 (2009) 256–260

[40] Fatih Ozsolak and Patrice M. Milos. RNA sequencing: advances, challenges and opportunities; Nat Rev Genet. 2011 February ; 12(2): 87–98. doi:10.1038/nrg2934

[41] Tierklinik Hannover, Group of Prof. Dr. Ingo Nolte; https://www.tiho-hannover.de/kliniken-institute/kliniken/klinik-fuer-kleintiere/profil-und-struktur/team/innere-medizin/prof-dr-ingo-nolte/

[42] Eva-Maria Packeiser, Marion Hewicker-Trautwein, Heike Thiemeyer, Annika Mohr, Johannes Junginger, Jan Torben Schille, Hugo Murua Escobar, Ingo Nolte. Characterization of six canine prostate adenocarcinoma and three transitional cell carcinoma cell lines derived from primary tumor tissues as well as metastasis; PLoS ONE 15(3): e0230272. https://doi.org/10.1371/journal.pone.0230272

[43] Gurvinder Kaur and Jannette M. Dufour. Cell lines: Valuable tools or useless artifacts; Spermatogenesis 2:1, 1–5; January/February/March 2012; G 2012 Landes Bioscience

[44] Ms. Vipul Chaudhary, Dr. Pamela Singh. CELL LINE: A REVIEW; International Journal of Advanced Research in Science and Engineering Vol.No.6, Issue No.4, April 2017

[45] Allison Warren, Andrew Jones et al., 2020; Global computational alignment of tumor and cell line transcriptional profiles; bioRxiv preprint doi: https://doi.org/10.1101/2020.03.25.008342;

[46] Adam McDermaid, Brandon Monier, Jing Zhao, Bingqiang Liu and Qin Ma. Interpretation of differential gene expression results of RNA-seq data: review and integration; Briefings in Bioinformatics, 20(6), 2019, 2044–2054

[47] Akram Mohammed , YanCui, Valeria R. Mas Rishikesan Kamaleswaran. Diferential gene expression analysis reveals novel genes and pathways in pediatric septic shock patients; (2019) 9:11270 — https://doi.org/10.1038/s41598-019-47703-6

[48] Jiannan GUO et.al. Transcription: the epicenter of gene expression; Guo / J Zhejiang Univ-Sci B (Biomed Biotechnol) 2014 15(5):409-411

[49] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics; Nat Rev Genet. 2009 January ; 10(1): 57–63. doi:10.1038/nrg2484.

[50] Clarissa M. Koch, Stephen F. Chiu, Mahzad Akbarpour, Ankit Bharat, Karen M. Ridge, Elizabeth T. Bartom and Deborah R. Winter. A Beginner's Guide to Analysis of RNA Sequencing Data; Originally Published in Press as DOI: 10.1165/rcmb.2017-0430TR on April 6, 2018

[51] Taylor Francis Group, LLC. RNA-seq Data Analysis A Practical Approach; International Standard Book Number-13: 978-1-4665-9501-9 (eBook - PDF)

[52] Korpelainen E, Tuimala J et al. RNA-seq Data Analysis A Practical Approach; Taylor Francis Inc, 2014.

[53] Alicia Oshlack, Mark D Robinson and Matthew D Young. From RNA-seq reads to differential expression results; Genome Biology 2010, 11:220

[54] Malachi Griffith, Jason R Walker, Obi Li Griffith, Benjamin J. Ainscough. Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud; PLoS Computational Biology · August 2015 DOI: 10.1371/journal.pcbi.1004393 Source: PubMed

[55] H.P.J. Buermans, J.T. den Dunnen. Next generation sequencing technology: Advances and applications; Biochimica et Biophysica Acta 1842 (2014): 1932–1941.

[56] Komal P. Singh et.al. Mechanisms and Measurement of Changes in Gene Expression; Biological Research for Nursing 2018, Vol. 20(4) 369-382

[57] Liu et al. Establishment and characterization of stable red, far-red (fR) and near infra-red (NIR) transfected canine prostate cancer cell lines; Cancer Cell Int (2020) 20:139

[58] Hussain Ahmed Chowdhury , Dhruba Kumar Bhattacharyya , and Jugal Kumar Kalita. Differential Expression Analysis of RNA-seq Reads: Overview, Taxonomy, and Tools; IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 17, NO. 2, MARCH/APRIL 2020

[59] Refael Kohen, Jonathan Barlev, Gil Hornung, Gil Stelzer, Ester Feldmesser, Kiril Kogan, Marilyn Safran and Dena Leshkowitz. UTAP: User-friendly Transcriptome Analysis Pipeline; BMC Bioinformatics (2019) 20:154

[60] Taura et al. Design and implementation of GXP make — A workflow system based on make; Future Generation Computer Systems 29(2013) 662-672

[61] Johannes Köster and Sven Rahmann. Building and Documenting Workflows with Python-Based Snakemake; Digital Object Identifier 10.4230/OASIcs.GCB.2012.49

[62] Pertea et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie, and Ballgown; Nat Protoc. 2016 September ; 11(9): 1650–1667. doi:10.1038/nprot.2016.095.

[63] Baruzzo, G. Hayer, K.E., Kim, E.J., Di Camillo, B., FitzGerald, G.A., Grant, G.R. Simulation-based comprehensive benchmarking of RNA-seq aligners.; Nat. Methods 2016, 14, 135.

[64] Tim Meese et.al. Filtering and data-driven hypothesis weighting for transcript level and rna-seq data analysis.; Ghent University

[65] Thiemeyer H, Taher L, Schille JT, Harder L, Hungerbuehler SO, Mischke R, et al. Suitability of ultrasound-guided fine-needle aspiration biopsy for transcriptome sequencing of the canine prostate.; Sci Rep. 2019; 9:13216. doi: 10.1038/s41598-019-49271-1

[66] Sarah E Hunt, William McLaren, Laurent Gil, Anja Thormann, Helen Schuilenburg, Dan Sheppard, Andrew Parton, Irina M Armean, Stephen J Trevanion, Paul Flicek, Fiona Cunningham. Ensembl variation resources; Database Volume 2018 doi:10.1093/database/bay119

[67] Michael Love, Simon Anders, Wolfgang Huber. Beginner's guide to using the DESeq2 package; bioRxiv (2014). doi:10.1101/002832

[68] Illumina Sequencing; Assessed on 05.11.2020; https://emea.support.illumina.com/bulletins/2016/04/fastq-files-explained.html

[69] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang and Ali Mortazavi; A survey of best practices for RNA-seq data analysis; Conesa et al. Genome Biology (2016) 17:13 DOI 10.1186/s13059-016-0881-8

[70] Ensembl, Canis Lupus Familiaris DNA (FASTA);Assessed on 29.09.2020; ftp://ftp.ensembl.org/pub/release-101/fasta/canis$_l upus_f amiliaris/dna/$

[71] Ensembl, Canis Lupus Familiaris Gene Sets;Assessed on 30.09.2020; ftp://ftp.ensembl.org/pub/release-101/gtf/canis_lupus_familiaris/ Canis_lupus_familiaris.CanFam3.1.101.chr.gtf.gz

[72] Anaconda3-2020.07-Linux; Assessed on 23.07.2020;https://repo.anaconda.com/archive/Anaconda3-2020.07-Linux-x86$_6$4.sh

[73] Köster, Johannes and Rahmann, Sven, Bioinformatics 2012; Snakemake - A scalable bioinformatics workflow engine;Assessed on 22.07.2020; https://snakemake.readthedocs.io/en/stable/

[74] FastQC;Assessed on 15.09.2020 https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc$_v$0.11.9.zip

[75] Marie Lataretu and Martin Hölzer. RNAflow: An Effective and Simple RNA-Seq Differential Gene Expression Pipeline Using Nextflow; Genes 2020, 11, 1487; doi:10.3390/genes11121487

[76] Rui Li 1, Kai Hu, Haibo Liu 1, Michael R. Green and Lihua Julie Zhu. OneStopRNAseq: A Web Application for Comprehensive and Efficient Analyses of RNA-Seq Data; Genes 2020, 11, 1165; doi:10.3390/genes11101165

[77] MacIntosh Cornwell et.al. VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis; BMC Bioinformatics (2018) 19:135

[78] Shanrong Zhao et.al. QuickRNASeq lifts large-scale RNA-seq data analyses to the next level of automation and interactive visualization; BMC Genomics (2016) 17:39 DOI 10.1186/s12864-015-2356-9

[79] FastQC Manual. Assessed on 15.09.2020, https://dnacore.missouri.edu/PDF/FastQC$_M anual.pdf$

[80] Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[81] Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller. MultiQC: Summarize analysis results for multiple tools and samples in a single report; Bioinformatics (2016), doi: 10.1093/bioinformatics/btw354, PMID: 27312411, Assessed on 12.11.2020

[82] Bolger, A. M., Lohse, M., Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170; Assessed on 20.09.2020, http://www.usadellab.org/cms/?page=trimmomatic

[83] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner; Bioinformatics, 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25.

[84] Alexander Dobin. STAR Manual;
https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/lecture$_{n}otes/STARmanual.pdf$

[85] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL: http://www.rstudio.com/.

[86] Ugo Testa, Germana Castelli and Elvira Pelosi. Cellular and Molecular Mechanisms Underlying Prostate Cancer Development:Therapeutic Implications; Medicines 2019, 6, 82; doi:10.3390/medicines6030082

[87] Jan Torben Schille, Ingo Nolte, Eva-Maria Packeiser, Laura Wiesner, Jens Ingo Hein, Franziska Weiner, Xiao-Feng Wu, Matthias Beller, Christian Junghanss and Hugo Murua Escobar. Isoquinolinamine FX-9 Exhibits Anti-Mitotic Activity in Human and Canine Prostate Carcinoma Cell Lines; Int. J. Mol. Sci. 2019, 20, 5567; doi:10.3390/ijms20225567

[88] Peppino Mirabelli, Luigi Coppola and Marco Salvatore. Cancer Cell Lines Are Useful Model Systems for Medical Research; Cancers 2019, 11, 1098; doi:10.3390/cancers11081098

[89] Joaquın Villar, Marıa Isabel Arenas, Caitlin M. MacCarthy, Marıa Jose Blanquez, Oscar M. Tirado and Vicente Notario. PCPH/ENTPD5 Expression Enhances the Invasiveness of Human Prostate Cancer Cells by a Protein Kinase CD–Dependent Mechanism; Cancer Res 2007; 67: (22). November 15, 2007

[90] Yozo Mitsui, Inik Chang, Taku Kato. Functional role and tobacco smoking effects on methylation of CYP1A1 gene in prostate cancer; May 19, 2016 doi: 10.18632/oncotarget.9470

[91] Kurfurstova, D. et al. DNA damage signalling barrier, oxidative stress and treatment-relevant DNA repair factor alterations during progression of human prostate cancer; Mol. Oncol. 10, 879–894 (2016)

[92] Carlos Eduardo Fonseca-Alves, Priscila Emiko Kobayashi, Antonio Fernando Leis-Filho. E-Cadherin Downregulation is Mediated by Promoter Methylation in Canine Prostate Cancer; Front. Genet. 10:1242. doi: 10.3389/fgene.2019.01242

[93] Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550 (2014)

[94] Lorena Pantano (2020). DEGreport: Report of DEG analysis. R package version 1.26.0.,Assessed on 20.11.2020; http://lpantano.github.io/DEGreport/

[95] H. Wickham. ggplot2: Elegant Graphics for Data Analysis; Springer-Verlag New York, 2016, Assessed on 20.11.2020

[96] CARLOS E. FONSECA-ALVES, MARCELA M.P. RODRIGUES, VERIDIANA M.B.D. DE MOURA, SILVIA R. ROGATTO AND RENEE LAUFER-AMORIM. Alterations of C-MYC, NKX3.1, and E-Cadherin Expressionin Canine Prostate Carcinogenesis; 12 September 2013, https://doi.org/10.1002/jemt.22292

[97] Jessica K. Simmons, Said M. Elshafae, Evan T. Keller, Laurie K. McCauley and Thomas J. Rosol. Review of Animal Models of Prostate Cancer Bone Metastasis; Vet. Sci. 2014, 1(1), 16-39; https://doi.org/10.3390/vetsci1010016

[98] Renée Laufer-Amorim, Carlos Eduardo Fonseca-Alves, Rolando Andre Rios Villacis, Sandra Aparecida Drigo Linde, Marcio Carvalho, Simon Jonas Larsen, Fabio Albuquerque Marchi and Silvia Regina Rogatto. Comprehensive Genomic Profiling of Androgen-Receptor-Negative Canine Prostate Cancer; Int. J. Mol. Sci. 2019, 20, 1555; doi:10.3390/ijms20071555

[99] Henry F. L'Eplattenier, Chen Li Lai, René van den Ham, Jan Mol, Frederick van Sluijs, Erik Teske. Regulation of COX-2 expression in canine prostate carcinoma: increased COX-2 expression is not related to inflammation; J Vet Intern Med. Jul-Aug 2007;21(4):776-82. doi: 10.1892/0891-6640(2007)21[776:roceic]2.0.co;2.

[100] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine; Bioinformatics, Volume 28, Issue 19, 1 October 2012, Pages 2520–2522, https://doi.org/10.1093/bioinformatics/bts480

[101] Juliana Costa-Silva, Douglas Domingues, Fabricio Martins Lopes. RNA-Seq differential expression analysis: An extended review and a software tool; PLoS ONE 12(12): e0190152. https://doi.org/10.1371/journal.pone.0190152

[102] Hoon,S. et al. (2003). Biopipe: a flexible framework for protocol-based bioinformatics analysis. Genome Res., 13, 1904–1915.

[103] Goecks,J. et al. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol., 11, R86

[104] Halbritter,F. et al. (2011). GeneProf: analysis of high-throughput sequencing experiments. Nat. Methods, 9, 7–8.

[105] Shah,S.P. et al. (2004). Pegasys: software for executing and integrating analyses of biological sequences. BMC Bioinformatics, 5, 40.

[106] Goodstadt,L. (2010). Ruffus: a lightweight Python library for computational pipelines. Bioinformatics, 26, 2778–2779.

[107] Sadedin,S.P. et al. (2012). Bpipe: a tool for running and managing bioinformatics pipelines. Bioinformatics, 28, 1525–1526

[108] Taura,K. et al. (2010). Design and Implementation of GXP Make – A Workflow System Based on Make. In IEEE International Conference on eScience, IEEE Computer Society, pp. 214–221, Los Alamitos, CA, USA

[109] Tanaka,M. and Tatebe,O. (2010). Pwrake: a parallel and distributed flexible workflow management tool for wide-area data intensive computing.; HPDC '10: Proceedings of the 19th ACM International Symposium on High Performance Distributed ComputingJune 2010 Pages 356–359 https://doi.org/10.1145/1851476.1851529.

[110] Stallman,R.M. and McGrath,R. (1991). GNU Make—A Program for Directing Recompilation. http://wwwgnu.org/software/make/

[111] Yang X, Liu D et al. HTQC: a fast quality control toolkit for Illumina sequencing data. BMC bioinformatics 2013. 14(1): 33.

[112] Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM et al. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics. 2010;11 Suppl 4:S7

[113] Hannon, G.J. (2010).FASTX-Toolkit; http://hannonlab.cshl.edu/fastx$_{t}oolkit/$

[114] Song L and Florea L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaScience 2015. 4(1): 48.

[115] Daehwan Kim, Ben Langmead Steven L Salzberg. HISAT: a fast spliced aligner with low memory requirements, Nature Methods volume 12, pages357–360(2015)

[116] Marco-Sola S, Sammeth M, Guigó R, Ribeca P. et al. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods. 2012;9:1185–8.

[117] Okonechnikov K, Conesa A et al. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 2015. 32(October 2015): btv566.

[118] Li H, Handsaker B et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009. 25: 2078—-2079.

[119] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

[120] Hardcastle TJ, Kelly KA. et al. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics, 2010;11:422.

[121] Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 2015;43:e140

[122] Li J, Tibshirani R.Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res.2013;22:519–36

[123] Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013; 29:1035–43.

[124] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7:562–78.

[125] Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29.

[126] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106.

[127] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

[128] Nueda MJ, Tarazona S, Conesa A. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. Bioinformatics. 2014;30:2598 –602.

[129] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25.

[130] Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics. 2012;13:523–38

[131] Johnson WE, Rabinovic A, Li C. Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics. 2007;8:118–27.

[132] Nueda MJ, Ferrer A, Conesa A. ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments. Biostatistics. 2012;13:553–66.

[133] Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics, 21, 3439–3440

[134] Yu G, Wang L, Han Y, He Q (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology, 16(5), 284-287. doi: 10.1089/omi.2011.0118.

[135] Yu G, Wang L, Han Y, He Q (2012). The National Cancer Institute (https://www.cancer.gov)

[136] Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors., Nat Genet 2003;33:49–54.

[137] Stanbrough M, Bubley GJ, Ross K, et al. Increased expression of genes converting adrenal androgens to testosterone in androgen-independent prostate cancer., Cancer Res 2006;66:2815–25.

[138] Kun Liu, Shuo Gu et.al The MMP2 rs243865 polymorphism increases the risk of prostate cancer: A meta-analysis; Department of Urology, Huai'an First People's Hospital, Nanjing Medical University, Huai'an, China, 223300

[139] Hong Q, Sze C-I, Lin S-R, Lee M-H, He R-Y, et al. (2009) Complement C1q Activates Tumor Suppressor WWOX to Induce Apoptosis in Prostate Cancer Cells; PLoS ONE 4(6): e5755. doi:10.1371/journal.pone.0005755

[140] Harjanto D, Maffei JS, Zaman MH (2011) Quantitative Analysis of the Effect of Cancer Invasiveness and Collagen Concentration on 3D Matrix Remodeling; PLoS One. 2011; 6(9): e24891.

[141] Xu et al. The role of collagen in cancer: from bench to bedside; PLoS One. 2011; 6(9): e24891.; J Transl Med (2019) 17:309 https://doi.org/10.1186/s12967-019-2058-1

[142] Shibao Li et al. Downregulation of DACT-2 by Promoter Methylation and its Clinicopathological Significance in Prostate Cancer; J Cancer. 2019; 10(7): 1755–1763.

[143] Dai et al.FGF21 facilitates autophagy in prostate cancer cells by inhibiting the PI3K–Akt–mTOR signaling pathway; Cell Death and Disease (2021) 12:303 https://doi.org/10.1038/s41419-021-03588-w

[144] Jennifer L. Wilding and Walter F. Bodmer Cancer Cell Lines for Drug Discovery and Development; Published OnlineFirst April 9, 2014; DOI: 10.1158/0008-5472.CAN-13-2971

[145] Clarissa M. Koch et al. A Beginner's Guide to Analysis of RNA Sequencing Data; American Journal of Respiratory Cell and Molecular Biology Volume 59 Number 2 https://doi.org/10.1165/rcmb.2017-0430TR

[146] M. Cekanova, Kusum Rathore Animal models and therapeutic molecular targets of cancer: utility and limitations; DOI:10.2147/DDDT.S49584Corpus ID: 7371076

A Appendix

```
###################################################
# Configuration file for pipelines
###################################################

#This is the configuration file for using the pipeline:


READSPATH: /path/to/raw_reads
OUTPUTPATH: /output/path
SINGLEENDPATH: /path/to/single_end/files
PAIREDENDPATH: /path/to/paired_end/files
OUTPUTPATHFASTQC: /path/to/store/fastqc/report
OUTPUTPATHTRIM: /path/to/store/trimmed/files
ANNOTATIONFILE: /path/to/annotation/file
RSCRIPTPATH: /path/to/Rscript
STARINDEX: path/to/starindex
GTFFILE: path/to/gtf_file
COUNTMATRIX: path/to/store/countmatrix
OUTPUTSTAR: path/to/store/star/output


STAR: /path/toSTAR/STAR-2.7.6a/bin/Linux_x86_64/STAR
TRIMMOMATIC: path/totrimmomatic/trimmomatic-0.39-1/share
/trimmomatic-0.39-1/trimmomatic.jar
```

```
####################################################
# Snakemake pipeline for RNA-Seq (SINGLE-END reads)
####################################################

import os
import numpy as np
from os.path import join, basename, dirname
import glob

configfile: 'config.yaml'
input_path = config["READSPATH"]
output_path = config["OUTPUTPATH"]
single_end_path = config["SINGLEENDPATH"]
output_path_fastqc = config["OUTPUTPATHFASTQC"]
output_path_trim = config["OUTPUTPATHTRIM"]
trimmomatic = config["TRIMMOMATIC"]
merge_script_path = config["RSCRIPTPATH"]
STAR_INDEX = config["STARINDEX"]
gtf_file = config["GTFFILE"]
annotation_file = config["ANNOTATIONFILE"]
STAR = config["STAR"]
count_matrix_path = config["COUNTMATRIX"]
output_star = config["OUTPUTSTAR"]

SAMPLES_SE = []
SAMPLES_SE_fastq = []
SAMPLES_SE_fastq = os.listdir(single_end_path)


for i in SAMPLES_SE_fastq:
  SAMPLES_SE = list(map(lambda i: i[: -9], SAMPLES_SE_fastq))


rule qualityControl:
  input:
      single_end_path + "/{sample}.fastq.gz".format(sample=sample) for
      sample in SAMPLES_SE
  output:
      expand(output_path_fastqc + "/{sample}.html", sample=SAMPLES_SE),
      expand(output_path_fastqc + "/{sample}.zip", sample=SAMPLES_SE)
  shell:
      ./fastqc {input} -o {output}


rule multiqc:
  input:
    expand(output_path_fastqc + "/{sample}_fastqc.html", sample=SAMPLES_SE)
  output:
    report = output_path_fastqc + "/report_multiqc.html"
  params:
    path = output_path_fastqc
  shell:
    multiqc -o {params.path} {input}


rule trim:
  input:
    single_end_path + "/{sample}.fastq.gz".format(sample=sample)
    for sample in SAMPLES_SE
  output:
    out = expand(output_path_trim + "/{sample}_trimmed.fastq", sample=SAMPLES_SE)
  shell:
    trimmomatic SE -threads {6} -phred33 {input} - baseout {out}
    ILLUMINACLIP:TruSeq3-SE.fa:2:30:10
    LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50


rule create_index:
  input:
    fasta = annotation_file
    gtf = gtf_file
  output:
    STAR_INDEX
  shell:
    STAR
    --runThreadN 16 \
    --runMode genomeGenerate \
    --genomeDir {STAR_INDEX} \
    --genomeFastaFiles {input.fasta}\
    --sjdbGTFfile {input.gtf}\
    --sjdbOverhang 100


rule mapping:
  input:
    R1 = output_path_trim + "/{sample}_trimmed.fastq".format(sample=sample)
    for sample in SAMPLES_SE
    index = STAR_INDEX
  output:
    output_star + "/{sample}.bam"
  shell:
    STAR
    --runThreadN 10 \
    --genomeDir {input.index} \
    --readFilesIn {input.R1}\
    --outSAMtype BAM Unsorted\
    -- quantMode GeneCounts


rule merge:
  input:
    script = merge_script_path
  output:
    data = count_matrix_path
  shell:
    "Rscript␣{input.script}␣--out␣{output.data}"
```

```
####################################################
# Snakemake pipeline for RNA-Seq (PAIRED-END reads)
####################################################

import os
import numpy as np
from os.path import join, basename, dirname
import glob

configfile: 'config.yaml'
input_path = config["READSPATH"]
output_path = config["OUTPUTPATH"]
paired_end_path = config["PAIREDENDPATH"]
output_path_fastqc = config["OUTPUTPATHFASTQC"]
output_path_trim = config["OUTPUTPATHTRIM"]
trimmomatic = config["TRIMMOMATIC"]
merge_script_path = config["RSCRIPTPATH"]
STAR_INDEX = config["STARINDEX"]
gtf_path = ["GTFPATH"]
annotation_path = ["ANNOTATIONPATH"]
STAR = ["STAR"]
count_matrix_path = ["COUNTMATRIX"]
output_star = ["OUTPUTSTAR"]


SAMPLES_PE = []
SAMPLES_PE_fastq = []
SAMPLES_PE_fastq = os.listdir("/home/himmem/data1/merged/paired_end")

for i in SAMPLES_PE_fastq:
    SAMPLES_PE = list(map(lambda i: i[: -12], SAMPLES_PE_fastq))

SAMPLES_PE = list(dict.fromkeys(SAMPLES_PE))

rule qualityControl:
    input:
        R1 = paired_end_path + "/{sample}_R1.fastq.gz",.format(sample=sample)
        for sample in SAMPLES_PE,
        R2 = paired_end_path + "/{sample}_R2.fastq.gz".format(sample=sample)
        for sample in SAMPLES_PE
    output:
        expand(output_path_fastqc + "/{sample}_R1.html", sample=SAMPLES_PE),
        expand(output_path_fastqc + "/{sample}_R1.zip", sample=SAMPLES_PE),
        expand(output_path_fastqc + "/{sample}_R2.html", sample=SAMPLES_PE),
        expand(output_path_fastqc + "/{sample}_R2.zip", sample=SAMPLES_PE)
    shell:
        """
        ./fastqc {input.R1} -o {outout} &&
        ./fastqc {input.R2} -o {output}"
        """

rule multiqc:
    input:
        expand(output_path_fastqc + "/{sample}_fastqc.html", sample=SAMPLES_PE)
    output:
        report = output_path_fastqc + "/report_multiqc.html"
    params:
        path = output_path_fastqc
    shell:
        "multiqc {params.path} --filename {output.report}"


rule trim:
        input:
                R1 = paired_end_path + "/{sample}_R1.fastq.gz".format(sample=sample)
                for sample in SAMPLES_PE,
        R2 = paired_end_path + "/{sample}_R2.fastq.gz".format(sample=sample)
        for sample in SAMPLES_PE
        output:
                output_path_trim + "/{sample}.fastq",
                output_path_trim + "/{sample}.fastq"
        shell:
                """
                trimmomatic PE -threads {12} -phred33 {input.R1} {input.R2} {output}
                ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:keepBothReads
                LEADING:3 TRAILING:3 MINLEN:50
                """

rule create_index:
    input:
        fasta = annotation_file
        gtf = gtf_file
    output: STAR_INDEX
    shell:
        /home/himmem/data1/tools/STAR-2.7.6a/bin/Linux_x86_64/STAR
        --runThreadN 16 \
        --runMode genomeGenerate \
        --genomeDir {STAR_INDEX} \
        --genomeFastaFiles {input.fasta}\
        --sjdbGTFfile {input.gtf}\
        --sjdbOverhang 100


rule mapping:
    input:
        R1 = paird_end_path + "/{sample}.fastq.gz".format(sample=sample)
        for sample in SAMPLES_PE
        R2 = paired_end_path + "/{sample}.fastq.gz".format(sample=sample)
        for sample in SAMPLES_PE
        index = STAR_INDEX
    output:
        output_star + "/{sample}.bam"
    shell:
        STAR
        --runThreadN 10 \
        --genomeDir {input.index} \
        --readFilesIn {input.R1} {input.R2}\
        --outSAMtype BAM Unsorted\
        -- quantMode GeneCounts

rule merge:
    input:
        script = merge_script_path
    output:
        data = count_matrix_path
    shell:
        "Rscript {input.script} \
            --out {output.data}"
```

```r
# R script for merging all ReadsPerGene.out.tab files to one count matrix
## STAR count file format is
#column 1: gene ID
#column 2: counts for unstranded RNA-seq
#column 3: counts for the 1st read strand aligned with RNA
#column 4: counts for the 2nd read strand aligned with RNA

library(tools)
library(dplyr)

first_read <- read.table("output_star/ReadsPerGene.out.tab")
first_read[3:4] <-  list(NULL)

first_read <- first_read[-c(1,2,3,4), ]

# name of the target file
target_file_name <- "ReadsPerGene.out.tab"

# enter path of the main folder
path_target <- "output_star"
sub_folders <- list.files(path = path_target)

for (folder in sub_folders) {
  path_sub_folder <- file.path(path_target,folder)
  target_file <- file.path(path_sub_folder,target_file_name)
  if (file.exists(target_file)) {

    print("file found")
    print(folder)
    read <- read.table(target_file)
    read <- read[-c(1,2,3,4), ]
    read[3:4] <- list(NULL)
    names(read)[2] <- folder
    first_read <- inner_join(first_read, read, by= "V1")
    #print(first_read)
  } else {
    print("cant find file")
  }

}

first_read[2] <- list(NULL)
head(first_read)
names(first_read)[1] <- "Ensembl_gene_id"
ncol(first_read)

#write.csv(first_read, "counts_read.csv", row.names=FALSE)
write.csv(first_read, "counts_read_with_rownames.csv")
```