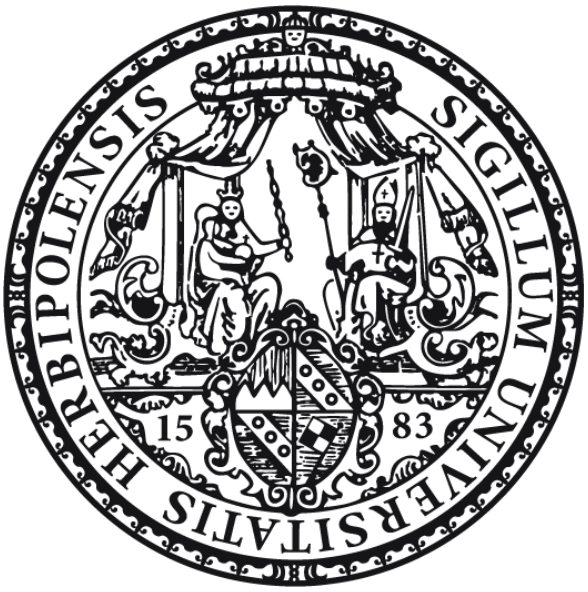


Machine-Learning-Based Identification of Tumor Entities, Tumor Subgroups, and Therapy Options

Bestimmung von Tumorentitäten, Tumorsubgruppen und Therapieoptionen basierend auf maschinellem Lernen

Doctoral thesis for a doctoral degree at the Graduate School of Life Sciences,

Julius-Maximilians-Universität Würzburg,

Section Biomedicine

submitted by

André Marquardt

from

Berlin-Mitte

Würzburg, **2023**

This document is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0): http://creativecommons.org/licenses/by-nc-nd/4.0 This CC license does not apply to third party material (attributed to another source) in this publication.

Submitted on: ……..……..

Office stamp

Members of the Thesis Committee

Chairperson: Prof. Dr. med. Thomas Dandekar

Primary Supervisor: Prof. Dr. rer. nat. Svenja Meierjohann

Supervisor (Second): Prof. Dr. rer. nat. Philip Kollmannsberger

Supervisor (Third): Dr. med. Markus Krebs

Date of Public Defence: ……..……

Date of Receipt of Certificates: ……..

Substantial parts of this thesis were published in the following open access articles:

1. **André Marquardt**, Antonio Giovanni Solimando, Alexander Kerscher, Max Bittrich, Charis Kalogirou, Hubert Kübler, Andreas Rosenwald, Ralf Bargou, Philip Kollmannsberger, Bastian Schilling, Svenja Meierjohann, Markus Krebs; Subgroup-Independent Mapping of Renal Cell Carcinoma-Machine Learning Reveals Prognostic Mitochondrial Gene Signature Beyond Histopathologic Boundaries, Front Oncol. 2021 Mar 15;11:621278, https://doi.org/10.3389/fonc.2021.621278

2. **André Marquardt**, Laura-Sophie Landwehr, Cristina L Ronchi, Guido Di Dalmazi, Anna Riester, Philip Kollmannsberger, Barbara Altieri, Martin Fassnacht, Silviu Sbiera; Identifying New Potential Biomarkers in Adrenocortical Tumors Based on mRNA Expression Data Using Machine Learning, Cancers 2021, 13, 4671, https://doi.org/10.3390/cancers13184671.

3. **André Marquardt**, Philip Kollmannsberger, Markus Krebs, Antonella Argentiero, Markus Knott, Antonio Giovanni Solimando, Alexander Kerscher; Visual Clustering of Transcriptomic Data from Primary and Metastatic Tumors – Dependencies and Novel Pitfalls; Genes 13, no. 8: 1335. https://doi.org/10.3390/genes13081335

TABLE OF CONTENTS

Table of Contents

1. SUMMARY

Molecular genetic analyses, such as mutation analyses, are becoming increasingly important in the tumor field, especially in the context of therapy stratification. The identification of the underlying tumor entity is crucial, but can sometimes be difficult, for example in the case of metastases or the so-called Cancer of Unknown Primary (CUP) syndrome. In recent years, methylome and transcriptome utilizing machine learning (ML) approaches have been developed to enable fast and reliable tumor and tumor subtype identification. However, so far only methylome analysis have become widely used in routine diagnostics.

The present work addresses the utility of publicly available RNA-sequencing data to determine the underlying tumor entity, possible subgroups, and potential therapy options. Identification of these by ML - in particular random forest (RF) models - was the first task. The results with test accuracies of up to 99% provided new, previously unknown insights into the trained models and the corresponding entity prediction. Reducing the input data to the top 100 mRNA transcripts resulted in a minimal loss of prediction quality and could potentially enable application in clinical or real-world settings.

By introducing the ratios of these top 100 genes to each other as a new database for RF models, a novel method was developed enabling the use of trained RF models on data from other sources.

Further analysis of the transcriptomic differences of metastatic samples by visual clustering showed that there were no differences specific for the site of metastasis. Similarly, no distinct clusters were detectable when investigating primary tumors and metastases of cutaneous skin melanoma (SKCM).

Subsequently, more than half of the validation datasets had a prediction accuracy of at least 80%, with many datasets even achieving a prediction accuracy of – or close to – 100%.

To investigate the applicability of the used methods for subgroup identification, the TCGA-KIPAN dataset, consisting of the three major kidney cancer subgroups, was used. The results revealed a new, previously unknown subgroup consisting of all histopathological groups with clinically relevant characteristics, such as significantly different survival. Based on significant differences in gene expression, potential therapeutic options of the identified subgroup could be proposed.

Concludingly, in exploring the potential applicability of RNA-sequencing data as a basis for therapy prediction, it was shown that this type of data is suitable to predict entities as well as

subgroups with high accuracy. Clinical relevance was also demonstrated for a novel subgroup in renal cell carcinoma. The reduction of the number of genes required for entity prediction to 100 genes, enables panel sequencing and thus demonstrates potential applicability in a real-life setting.

2. ZUSAMMENFASSUNG

Molekulargenetische Analysen, wie z. B. Mutationsanalysen, gewinnen im Tumorbereich zunehmend an Bedeutung, insbesondere im Zusammenhang mit der Therapiestratifizierung. Die Identifizierung der zugrundeliegenden Tumorentität ist von entscheidender Bedeutung, kann sich aber manchmal als schwierig erweisen, beispielsweise im Falle von Metastasen oder dem sogenannten Cancer of Unknown Primary (CUP)-Syndrom. In den letzten Jahren wurden Methylom- und Transkriptom-Ansätze mit Hilfe des maschinellen Lernens (ML) entwickelt, die eine schnelle und zuverlässige Identifizierung von Tumoren und Tumorsubtypen ermöglichen. Bislang werden jedoch nur Methylomanalysen in der Routinediagnostik eingesetzt.

Die vorliegende Arbeit befasst sich mit dem Nutzen öffentlich zugänglicher RNA-Sequenzierungsdaten zur Bestimmung der zugrunde liegenden Tumorentität, möglicher Untergruppen und potenzieller Therapieoptionen. Die Identifizierung dieser durch ML - insbesondere Random-Forest (RF)-Modelle - war die erste Aufgabe. Die Ergebnisse mit Testgenauigkeiten von bis zu 99 % lieferten neue, bisher unbekannte Erkenntnisse über die trainierten Modelle und die entsprechende Entitätsvorhersage. Die Reduktion der Eingabedaten auf die 100 wichtigsten mRNA-Transkripte führte zu einem minimalen Verlust an Vorhersagequalität und könnte eine Anwendung in klinischen oder realen Umgebungen ermöglichen.

Durch die Einführung des Verhältnisses dieser Top 100 Gene zueinander als neue Datenbasis für RF-Modelle wurde eine neuartige Methode entwickelt, die die Verwendung trainierter RF-Modelle auf Daten aus anderen Quellen ermöglicht.

Eine weitere Analyse der transkriptomischen Unterschiede von metastatischen Proben durch visuelles Clustering zeigte, dass es keine für den Ort der Metastasierung spezifischen Unterschiede gab. Auch bei der Untersuchung von Primärtumoren und Metastasen des kutanen Hautmelanoms (SKCM) konnten keine unterschiedlichen Cluster festgestellt werden.

Mehr als die Hälfte der Validierungsdatensätze wiesen eine Vorhersagegenauigkeit von mindestens 80% auf, wobei viele Datensätze sogar eine Vorhersagegenauigkeit von 100% oder nahezu 100% erreichten.

Um die Anwendbarkeit der verwendeten Methoden zur Identifizierung von Untergruppen zu untersuchen, wurde der TCGA-KIPAN-Datensatz verwendet, welcher die drei wichtigsten Nierenkrebs-Untergruppen umfasst. Die Ergebnisse enthüllten eine neue, bisher unbekannte

Untergruppe, die aus allen histopathologischen Gruppen mit klinisch relevanten Merkmalen, wie z. B. einer signifikant unterschiedlichen Überlebenszeit, besteht. Auf der Grundlage signifikanter Unterschiede in der Genexpression konnten potenzielle therapeutische Optionen für die identifizierte Untergruppe vorgeschlagen werden.

Zusammenfassend lässt sich sagen, dass bei der Untersuchung der potenziellen Anwendbarkeit von RNA-Sequenzierungsdaten als Grundlage für die Therapievorhersage gezeigt werden konnte, dass diese Art von Daten geeignet ist, sowohl Entitäten als auch Untergruppen mit hoher Genauigkeit vorherzusagen. Die klinische Relevanz wurde auch für eine neue Untergruppe beim Nierenzellkarzinom demonstriert. Die Verringerung der für die Entitätsvorhersage erforderlichen Anzahl von Genen auf 100 Gene ermöglicht die Sequenzierung von Panels und zeigt somit die potenzielle Anwendbarkeit in der Praxis.

3. INTRODUCTION

3.1 PERSONALIZED MEDICINE

The principles of "from bench to bedside" have manifested themselves in recent years, especially in oncological research, in the form of personalized and translational medicine (Figure 1). The goal here is to recommend therapies that are individually tailored to or most effective for an individual patient. The basis for today's personalized medicine is, in addition to the panel-based molecular genetic analysis of specific mutations or gene fusions, primarily (high-throughput) sequencing. This ultimately involves the comprehensive elucidation of a particular individual aspect, for example the transcriptome using RNA-sequencing, all mutations using whole exome sequencing (WES), or the entire genome using whole genome sequencing (WGS). In addition to the above-mentioned sequencing of the genome, there is also the possibility of resolving the methylome of a patient, indicating the methylation of given DNA segments. Unlike mutations, which can directly affect the activity of affected proteins, methylation modifications can impact the expression of genes. Ultimately, hyper- and hypomethylations can lead to deregulated protein expressions, having a similar impact on cells like mutations. With the help of Next Generation Sequencing (NGS) methods and corresponding studies, enormous progress has been made in the field of personalized medicine in recent years (1, 2). Certain mutations in the patient's genome can now be individually targeted and lead to improved survival. One example of the development of targeted therapy is the blockade of the MAP-Kinase (MAPK) signaling pathway in melanoma. This development is linked to BRAF V600E or V600K mutations in melanoma, which are present in almost half of all cutaneous melanomas. Inhibition via highly selective inhibitors for these mutations – such as FDA – (3) and EMA-approved vemurafenib (4) - showed significant survival benefits compared to dacarbazine treatment, with additionally fewer side effects (5, 6). However, treatment with specific inhibitors leads to the development of therapy-associated resistances in some patients, which may be caused by additional mutations or chromosomal aberrations (7). In these cases, reactivations of the MAPK signaling pathway are often observed, necessitating new solutions for inhibition (8–10). Further blockade was achieved by the development of MEK inhibitors, located downstream of BRAF, which prevents overactivation of BRAF at a different site (11). The combination of BRAF and MEK inhibitors shows an even better response in terms of survival than monotherapy with BRAF inhibitors (12–14), but also cause new resistance mechanisms (15) to happen.

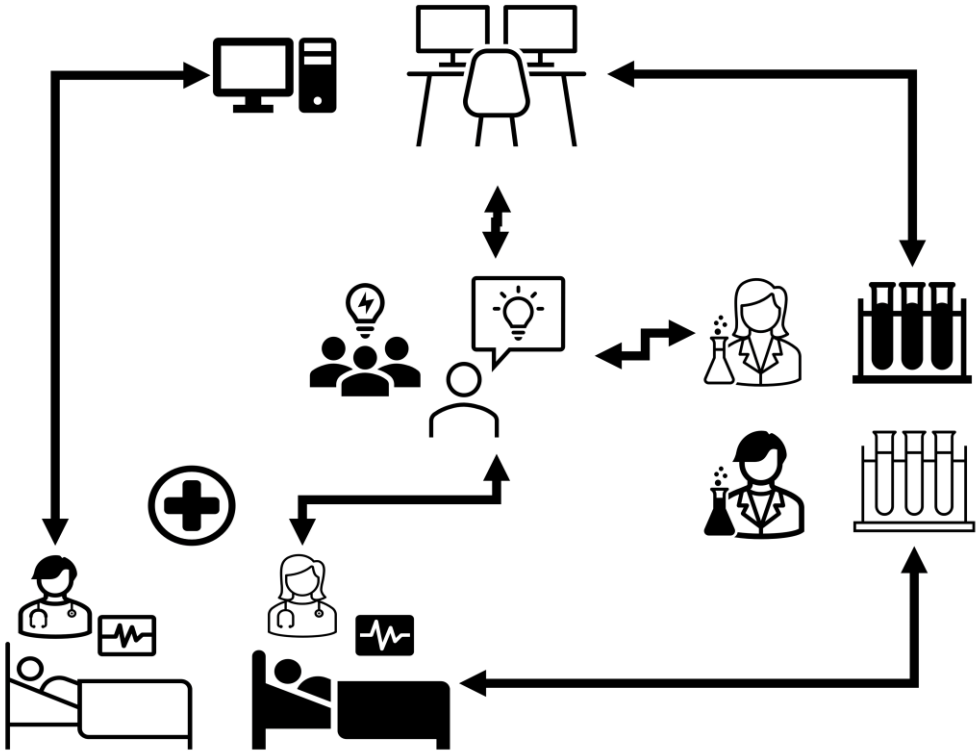


Figure 1: The principle of "from bench to bedside"

Schematic representation of the principle "from bench to bedside", including the exchange between clinicians (lower part), scientists (right part), and data analysts (upper part) in all directions.

Another notable example of the combination of personalized medicine and the use of sequencing techniques are the NTRK inhibitors larotrectinib and entrectinib. These specifically target the ETV6-NTRK3 fusion protein, but also other NTRK fusions, which results from chromosomal fusion of the two genes. The unique feature of larotrectinib is that it was the first European Union approved tumor-agnostic inhibitor, applicable independently of the underlying tumor entity, whereas entrectinib was approved by the Food and Drug Administration (FDA) if the United States. Thus, diagnosis of the specific fusion directly offers a therapy option for different tumor types (16–20).

A final example of personalized medicine using sequencing techniques is the application of immunotherapy in the presence of high tumor mutational burden (TMB). Here, antibodies against cytotoxic T-lymphocyte-associated antigen 4 (CTLA-4) and programmed cell death (PD1) and/or PD-ligand 1 (PD-L1) are used in mono- or combination therapy resulting in immune checkpoint blockade. The use of TMB as a biomarker was first demonstrated in non-small cell lung cancer (NSCLC) but became a valid marker for other entities as well (21, 22). Since then, this type of immunotherapy, based on TMB as a biomarker, has also been applied in other entities. Recent studies suggested that immunogenic neoantigens expressed by tumor cells can trigger a response to immunotherapy caused by a high TMB (23–26). The TMB itself

can be determined via Whole Exome Sequencing (WES) and provides insight into the total mutation number of the sample at hand as a numerical number. Since WES is not only cost-intensive and time-consuming to perform, but also requires a certain standard of sample material in terms of quality and quantity, it is not always feasible. Due to this, alternative approaches based on panel sequencing have been established, which determine a good approximate value for the TMB, based on far fewer sequenced genes, yet being pretty accurate in determining the TMB, as long as the proposed optimal panel size between 1.5 and 3 Mb is reached (27).

3.2 DNA-Sequencing Techniques

For DNA-sequencing several distinctions can be made. One example for differentiation could be between targeted and untargeted sequencing. The first important application of target sequencing is the so-called panel sequencing, often performed using either Illuminas bridge amplification method (Figure 2 lower right side) or the proton detection method as used by ThermoFisher (Figure 2 lower middle part). An example for a targeted gene panel would be a gene set enabling the identification of somatic DNA mutations for targetable genes, to clarify different therapy options in a quick and cost-efficient way. Nowadays, there are panels that combine these DNA analyses with various RNA analyses in order to also determine specific fusions or expression alterations. Furthermore, newly developed methods are able to calculate the TMB from panel sequencing to include another therapeutic option such as immune checkpoint inhibitors (ICI). Here, detection of microsatellite instability (MSI) markers also play an important role in panel sequencing, as MSI tumors of different entities display increased ICI sensitivity (28, 29).

The second possible targeted sequencing method is the Sanger sequencing, which is one of the oldest DNA-sequencing methods and is based on the so-called chain-termination synthesis. Based on dideoxyribonucleoside triphosphates (ddNTP) the elongation of the new DNA-strand is terminated after the usage of a ddNTP instead of a dNTP. Based on chance, the new strands get terminated at different positions and the separation of the different strands based on length displays the last integrated nucleotide. Due to its familiar workflow and the fast cost-effective sequencing the method still is widely used in modern approaches. Pyrosequencing (Figure 2 lower left side) can be seen as a third targeted sequencing option, often used to validate mutations because of its high accuracy. The low amount of required sample material also provides an advantage over Sanger sequencing, as sample material availability often is a limiting factor.

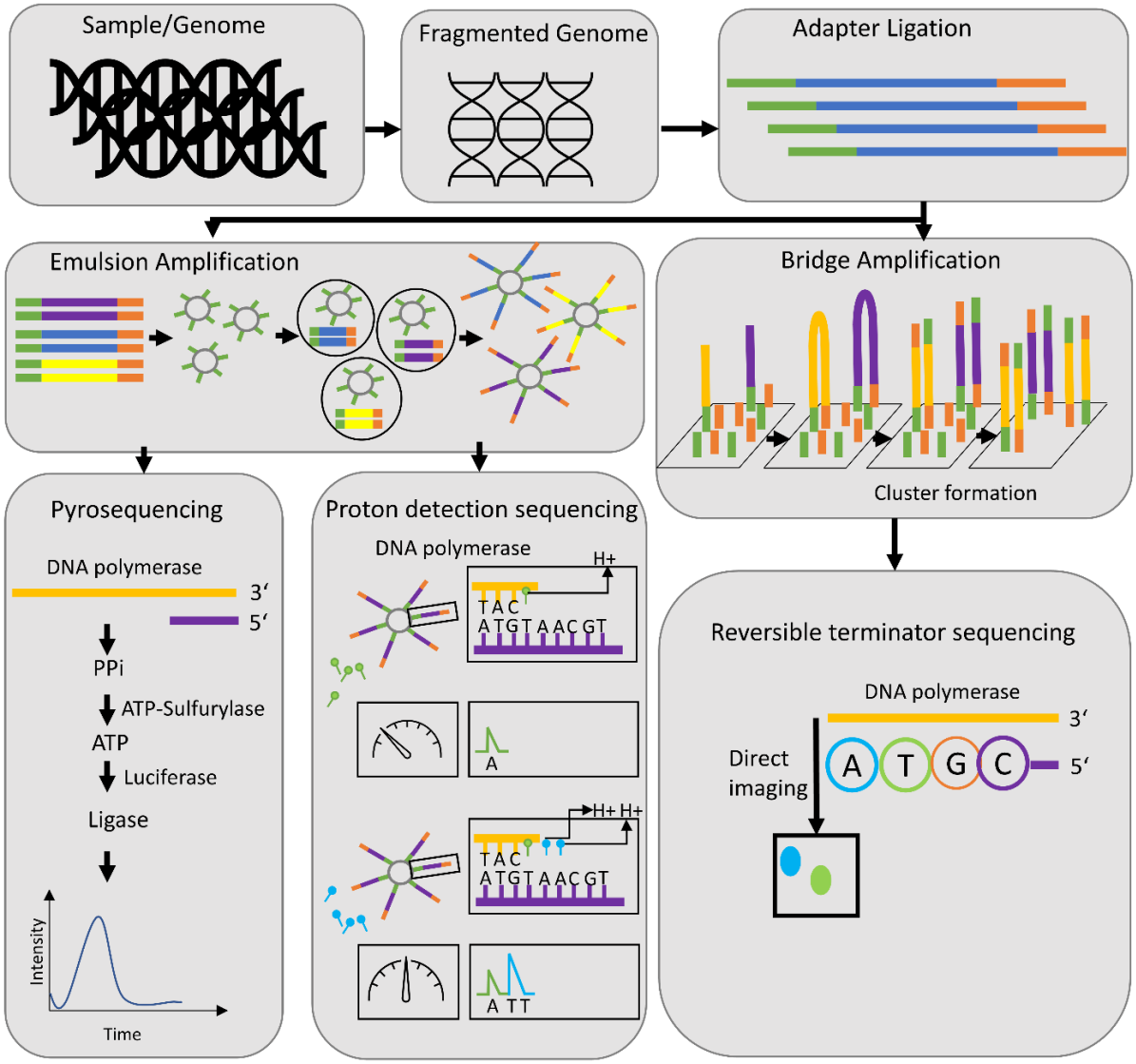


Figure 2:Different sequencing techniques

Simplified presentation of three of the major DNA-sequencing techniques routinely used in laboratories, starting from the sample's genome, with following fragmentation and adapter ligation step (top row). Depending on the aimed sequencing technique used, the clonal amplification is either done by emulsion (middle left part) or by bridge amplification (middle right part). Following the clonal amplification by emulsion technique, either pyrosequencing – based on the detection of light emission occurring with the base insertion – (left lower part) or the proton detection sequencing – based on the proton release occurring with the base insertion – (middle lower part) are possible. Based on the so-called bridge amplification the reversible terminator sequencing, where a direct imaging of fluorescently labeled nucleotides is possible, is performed (lower right part).

The panel sequencing is opposed by the extensive sequencing. These are methods that resolve a complete aspect of the patient in depth. The first method to be mentioned here is whole exome sequencing (WES). Based on DNA, all exons of protein-coding genes – the so-called exome – are selected, amplified, and finally sequenced. The use of WES aims at a comprehensive identification of all potential protein mutations in order to determine, for example, a tumor disease. Further applications – besides determining possible therapeutically

targetable mutations – include among others a more exact determination of the TMB but also of the MSI status of the patient.

Since WES only maps about 1% of the entire genome and only the currently known protein-coding genes, but not other influencing factors that are not encoded in exomes, it is sometimes necessary to perform a whole genome sequencing (WGS). Here, the entire genome is sequenced. Since this requires a certain quantity and quality of sample material, this is not always possible. Compared to WES, the sequencing is coverage usually is lower, due to the amount of information to be sequenced, meaning that the informative value of WGS could be argued to be lower than that of WES. However, this downside is compensated by the additional information gained in non-protein-coding regions. In addition to mutations, chromosomal aberrations can also be identified based on WGS, such as copy number changes, but also breakpoints and chromosome shortenings can be reliably determined using this technique.

To obtain information on whether such chromosomal aberrations have a real influence on the disease, additional information on the so-called transcriptome is required. This consists of all the mRNAs produced in the cell and consequently maps all the "blueprints" for the proteins to be produced. Investigations using RNA-sequencing are particularly important in the case of chromosome fusions, as this is the only way to determine whether the resulting transcript is actually converted and thus likely to become a protein. Important fusion proteins, for example, are BCR-ABL, causing the chronic myeloid leukemia (CML (30, 31)) and serves as confirmation of diagnosis, or PML-RARA, causing the acute promyelocytic leukemia (PML (32)).

Additional analysis of methylation assays and the associated elucidation of the patient's methylome provides even further and deeper insights into the pathogenesis of a disease. This is because inactivation (silencing), activation and also overactivation of genes can take place not only via mutations, gene amplifications or gene losses, but also through methylation of for example promoter binding sites (33). These methylations can ultimately change the binding to DNA, leading to more or less transcript and consequently protein.

Furthermore, there are many other specialized sequencing methods, such as ChIP-Seq (Chromatin ImmunoPrecipitation DNA-Sequencing – a biochemical method to determine protein-DNA interactions (34)), ATAC-Seq (Assay for Transposase-Accessible Chromatin using sequencing – to assess genome-wide accessibility of chromatin (35)), or TOMO-seq (RNA tomography sequencing – to obtain genome-wide expression data with spatial resolution (36)). In addition, some of the mentioned methods are also applicable on a single cell level, like single cell RNA-sequencing (scRNA-sequencing (37)), methylation sequencing (38) or scATAC-seq (39).

For a complete comprehensive identification of all possible causes of a disease and subsequent use in clinical practice, a combination of all previously mentioned methods would be necessary.

Even if in comparison the applications of panel sequencing seem limited, they are used more frequently in clinical routine compared to all-encompassing genomic and transcriptomic analyses. This is due to the costs incurred and the time required for evaluation, but also to the problem of limited tumor material. However, in comparison complete sequencing is usually more accurate and often provides more information, yet it is not of interest or importance for personalized medicine, which is why WES, WGS, RNA-sequencing or methylome analyses are mainly used in the field of research.

3.3 Origin and Condition of Patient Samples

Even though the present work deals with the theoretical processing of samples from tumor patients, it is still necessary in any analyses to understand the origin of the data at hand. For example, there are significant differences in obtaining sample material depending on the location of the tumor or the entity itself. Additionally, the further use of the sample material and the preservation and long-term storage play a certain role in the generation and finally also in the quality of the obtained data.

The first step in obtaining patient samples is the biopsy. Here, tissue is removed from the living organism using special tools. There are different types and possibilities, such as needle biopsy, fine needle biopsy, or punch biopsy. When obtaining patient material, certain complications must always be expected. Especially in tissue with a very good blood supply, such as kidney, lung or liver, bleeding or wound infections can occur, causing additional morbidity for the patient. Because of the possible complications, a thorough evaluation must be made in each situation as to whether a biopsy is useful and necessary. This question becomes interesting, for example, if a patient would need to be re-biopsied in order to obtain new or more tumor material.

This makes it even more important to make optimal use of the material obtained. The first possibility is to directly process the obtained material as fresh material. Although this type of material tends not to be used in everyday clinical practice, it provides probably the best results in terms of the quality of the DNA and RNA. This has a particularly positive effect on the evaluation of data, which is why very important analyses, for example for study inclusion, sometimes have to be performed on fresh material (Figure 3 right side).

If a direct further processing of the sample material is not possible the desired material can be fixed using formalin and then embedded in paraffin – Formalin Fixed, Paraffin Embedded (FFPE) – also enabling the long-term preservation of the material (Figure 3 left side). In addition to the good preservation of all cell structures and proteins – which are, however, no longer active – this method also results in certain denaturation and changes to the DNA. Another disadvantage of this method, besides the time required for sample embedding, is that the procedure itself is not necessarily standardized.

Due to these disadvantages, there is a newer method for specific applications such as RNA-sequencing or mutation analysis, which compensates for these disadvantages – the use of fresh frozen material. Here, the tissue is frozen at extremely low temperatures (-190°C) within a very short time after obtaining the material (Figure 3 middle). Even though no changes to the DNA occur and it is therefore – for example – the method of choice for sequencing, it has specific disadvantages, which is why it is not used as standard. Firstly, the durability of samples frozen in this way is limited to the fact that the samples remain permanently stored at low temperatures and can only be thawed once. This circumstance prevents re-analyses of tumor material. The necessary technical equipment, which has to be available directly on site, is also a certain disadvantage of this method.

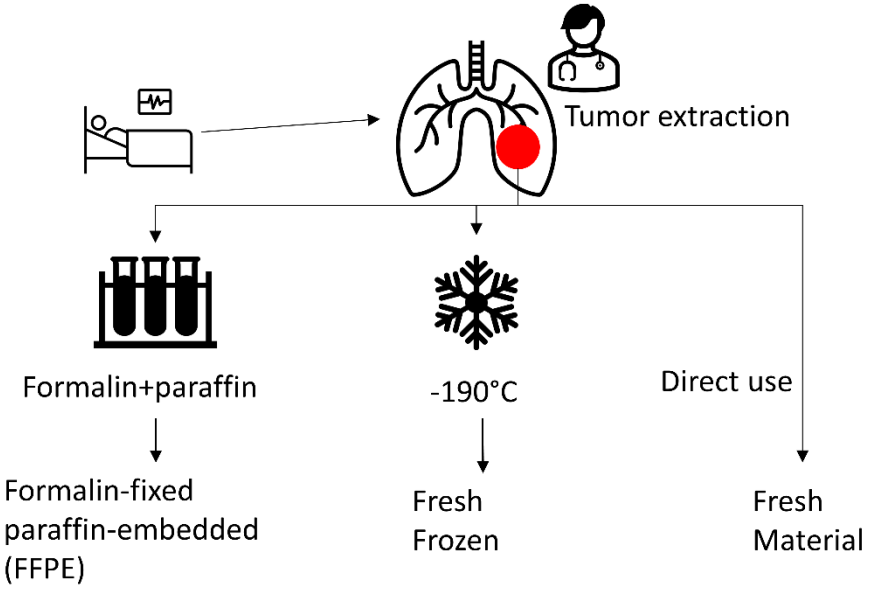


Figure 3: Overview of tumor sample types

Simplified overview of three of the most commonly used types of tumor samples for DNA-sequencing. After tumor extraction the material is either fixated using formalin and embedded in paraffin (left side), frozen at a very low temperature (usually -190°C – middle), or directly used (right side).

In comparison, the use of FFPE tissue offers many disadvantages at first glance, but recent work has shown that FFPE tissue can also be used for RNA-sequencing (40) and offers a good and cost-effective alternative for gene mutation panel analyses (41).

3.4 PUBLICLY AVAILABLE DATASETS

As acquisition of sample material and the resulting sequencing data is not always easy or feasible without problems, publicly accessible datasets are often used in (cancer) research.

Probably the most important public database is the one provided by The Cancer Genome Atlas (TCGA) consortium. The database currently comprises 72 projects of 67 different entities with a total of 86.046 included patients (release 34.0 of 27.07.2022 – accessed on 21.08.2022 at 4:38 pm). The most important features of this database are that the sample collection and its analyses as well as evaluations are standardized and consequently allow comparisons between the individual datasets, also applicable for the collected clinical data. In addition, at the time of sample collection, patients have not yet received any therapy. Furthermore, the database provides several different sequencing analyses for most of the samples, i.e. methylome, transcriptome (including miRNA), genome (WES and WGS) and protein expression data– also for normal tissue samples – allowing for comprehensive questions and analyses.

The International Cancer Genome Consortium (ICGC) with their corresponding database consisting of 86 projects, 22 entities, and a total of 24.289 donors (release 28 of 27.03.2019, accessed 21.08.2022 at 4:40 pm) is another well-established database. In contrast to the TCGA database, this is the collected sequencing from different institutions, which ultimately provides high quality data, but which may not necessarily be comparable with each other due to different sample collection and (data) processing. Usually, these are also samples taken before the start of therapy, but in comparison to the TCGA database, the data does not always have to be in a comparable form. Different from TCGA database, the ICGC database does not necessarily provide every type of data for each dataset, but generally spoken, it also contains methylation data, RNA- and DNA-sequencing data, and protein expression data. Of note, most of the TCGA datasets are also contained in the ICGC data portal.

Since the two databases consist of samples taken before therapy, only specific questions, such as the development of specific entities, can be answered, but in great detail. In contrast to the two databases mentioned, the so-called Gene Expression Omnibus (GEO) consists not only of therapy-naïve samples, but of many different samples, also from different organisms or

xenograft models. Since these are often also very specific datasets – for example for tumor development with and without therapy – other, questions can be answered with the help of these datasets. However, it should be noted that the datasets come from a wide variety of institutions with the possibility for every researcher to upload data to the GEO. One advantage of this database is the availability of a wide variety of datasets which allow comprehensive comparisons – also between different sequencing platforms and methods. Another important advantage is the presence of single cell sequencing datasets, which provide another point of view into the development of diseases. In contrast to TCGA and ICGC the GEO data portal does usually not have comprehensible data different sequencing analysis for each dataset and is comprised of single center analysis for a specific research question. Nevertheless, the GEO data portal harbours DNA and RNA analysis for nearly every question – not only for cancer – with the downside, that data does not necessarily has to comparable between the different datasets.

The last database to be mentioned is cbioportal, which, like the GEO, is a collection of datasets from different sources (42, 43). This dataset also contains RNA- and DNA analysis, but comparable to the GEO data portal, not every sequencing analysis is necessarily available for each dataset. Of note, the cbiportal also contains TCGA and ICGC data.

Even though publicly available datasets are crucial for research they, however, do have their very own drawbacks. For example, it is not possible to obtain raw data without prior application or approval, as patients' rights must be preserved. For this reason, one only gets access – at least publicly and freely accessible – to data that have already been analyzed. Particular care must be taken, as to which extent the available data has already been analyzed and whether the obtained data is comparable in the retrieved form.

If access to unprocessed data is desired, this is usually preceded by an application procedure which, depending on and in compliance with data protection, checks whether the applicants meet all the requirements for conscientious handling of the data. The so-called European Genome-Phenome Archive (EGA Archive) is one possibility for gaining access to this type of data. However, TCGA data are also available via a separate application procedure.

3.4.1 The TCGA-KIPAN Dataset

The determination of subgroups within tumor entities has been clinical and histopathologic practice for many decades, especially as different tumor subtypes could respond differently to given treatment. For example, myeloid leukemias are classified into different subgroups according to the French-American-British (FAB) classification system. Further known

examples of different subgroups are lung carcinomas, which can be divided not only into squamous cell carcinomas or adenocarcinomas but also into small cell or non-small cell carcinomas. Bioinformatics can also be of help in this area of research. In myeloid leukemia for example, there is a direct translation of molecular genetic properties and special features into subgroups (32, 44) and a direct link between molecular genetics and therapy response was shown (45), indicating the use and need of comprehensive molecular elucidation of patient samples. The use of certain markers to classify individual entities, for example BRCA mutations in breast and prostate cancer (46, 47), microsatellite instability (MSI) in colorectal carcinoma (48), EGFR alterations in lung cancer (49) or BRAF mutations in melanoma (50), massively increased recently. The possibilities offered by molecular genetic classification of patients´ tumor samples are by no means exhausted and can be further expanded, especially in the area of cross-entity or pan-cancer diagnostics and therapy using bioinformatics and especially ML.

Of special interest in this work are the Renal Cell Carcinoma (RCC) datasets of TCGA. These are three datasets representing the three major histopathological subgroups of RCC, that can be divided into two subgroups, starting from the site of origin of the disease or the cell of origin. Clear cell and papillary RCCs originate from the proximal tubule, whereas chromophobe RCCs originate from the cortical tubule.

The first dataset – TCGA-KIRC – is the largest subgroup of RCCs, the clear cell RCCs (ccRCC). This subgroup accounts for approximately 75-80% of RCCs and exhibits high tumor heterogeneity. The most common dysregulations are inactivation of the von Hippel-Lindau (VHL) gene together with PBMR1, SETD2 and BAP1 mutations (51). Additionally, there is a known subset in ccRCC which is driven by MTOR pathway alterations.

The second dataset – TCGA-KIRP – already consists of a special subgroup, the papillary RCCs (pRCCs). This subgroup is divided into type 1 and type 2. Type 1 can be attributed to alterations in the MET gene located on chromosome 7, where also germline mutations may occur. Type 2 can be further divided into three different subgroups. The first two groups are dependent on alterations on genomic level in either CDKN2A or the ETD2/BAP1/PBRM1 genes. The third group is characterized by a methylation alteration, the so-called CpG Island Methylator Phenotype (CIMP).

The third dataset – TCGA-KICH – represents the smallest and rarest subgroup of RCCs, chromophobe RCCs (chRCC). This subgroup is often based on aneuplodies of certain chromosomes (52). In addition, certain mutations can be observed more frequently such as in the genes TP53, PTEN, FAAH2, PDHB, PDXDC1 and NZF765.

Therapy basis for localized RCC disease is the surgical removal of the tumor for all three subgroups. If removal is no longer possible, various therapeutic options may be considered. In ccRCC – since it is the largest and best-studied subgroup – interferon-alpha immunotherapy or various multi- or tyrosine kinase inhibitors such as sorafenib, sunitinib or pazopanib are available. Other options include treatment with MTOR inhibitors – such as everolimus or temsirolimus – and on VEGFA-targeting and anti-angiogenic antibodies such as bevacizumab, which is relevant for a subset of ccRCCs. Standard chemotherapy is not used in RCC due to its inefficacy (53).

Since pRCCs and chRCCs are quite rare compared to ccRCCs, there are hardly any studies on the respective subgroups, yielding limited knowledge about the corresponding therapies. For pRCCs, however, it is recommended to proceed according to the guidelines of ccRCCs (54). For chRCCs, mitochondrial-directed therapy can be recommended, since this subgroup is known to be dependent on mitochondria (55–57), but guideline recommendations do not differ from the ccRCC guidelines.

In addition to these histopathological subgroups, the WHO has adapted its classification in recent years and subclassified RCC further to reflect the complexity of the disease, introducing groups such as clear cell papillary RCCs (58, 59). As the task of identifying the present RCC subgroup based on histopathological methods becomes increasingly difficult, identification based on molecular genetic characteristics is desirable. Bioinformatic analyses are the method of choice for these tasks and the combination of all three RCC datasets provides an optimal basis for further bioinformatic investigations.

3.5 Computational Biology and Bioinformatics

Bioinformatics has become indispensable in the field of personalized medicine. Bioinformatics and computational biology are disciplines representing a facet of data sciences and are used synonymously in this thesis or summarized under the term bioinformatics. Strictly speaking, bioinformatics deal with studies of large datasets (e.g. the analysis of genetic data) whereas computational biology is more concerned with finding solutions to problems arising from bioinformatic analyses. When looking at the applications, it is noticeable that personalized medicine is more an area of bioinformatics, whereas machine learning rather belongs to the area of computational biology.

Since this thesis is more concerned with computational biology, the concept of machine learning (ML) must be defined and distinguished from the concepts of deep learning (DL) and

artificial intelligence (AI). These three terms are often used more or less synonymously, with the term artificial intelligence being used in most cases.

Machine learning is a sub-unit or part of AI and is mainly used for things for which conventional algorithms would be too slow or the programming too cumbersome. A very early example of such specialized AI – as it could be called – is a spam filter for emails. One of the main discussions in the field of ML is not the choice of the algorithm itself, but often the data basis used or to be used. The principle of "garbage in – garbage out" applies here. This means, that any model developed is or can only be as good as the data on which it is based. Therefore, a certain data curation is normally inevitable in order to be able to present the data to the algorithm in the best possible way.

In general, the learning of algorithms in the field of ML can be divided into three classes: supervised, unsupervised, and semi-supervised. Supervised learning – in contrast to unsupervised learning – uses data which are already labeled, i.e. for which a specific class is known, for example the specific entity of a tumor which is indicated for the sequencing data. If such a representation as labeled dataset is possible, it often refers to so-called classification problems, where as a rule of thumb it should be possible to represent the number of labels in an understandable way in a dropdown menu. Problems where this is not reasonably possible are so-called regression problems. In contrast, unsupervised learning uses unlabeled data, i.e. data for which no class is known in advance. This is often a clustering problem, which is why so-called cluster algorithms are used for this purpose. These have the task to assign a label to the unknown data. Common clustering methods are for example k-means, Hierarchical Cluster Analysis (HCA) or Expectation Maximization. The combination of both extremes – all or no data points have labels – is the basis for the so-called semi-supervised learning. Here, some data points have specified labels, others do not. Using the data points with classes, the unknown labels of the other data points can be determined in order to apply classification algorithms to them.

Beside this classification, there are three different types of ML, depending on the task: Regression, Classification, and Clustering. Regression is the prediction of numerical values and can be compared to multivariate analysis in traditional statistics.

Classification refers to the prediction of categories, such as whether the email in question is spam or not. The data used for learning gets labeled and the predictions of the developed model can be compared with the actual labels. Possible algorithms for application to classification problems are, for example, decision trees, random forests, k nearest neighbor (KNN), or support vector machines (SVM). In general, the present data is split into two different sets. The selected algorithm uses the first data split – the so-called learning dataset – to learn on and is then tested or applied on the remaining data – the test dataset – to determine the

performance of the model. Subsequently, the model can also be used for data that were neither included in the learning nor in the test dataset, the so-called evaluation data.

Clustering is the only unsupervised learning method in the examples mentioned above. Basically, this involves grouping similar data points together and separating data points that are different from each other in order to form clusters of data points. Algorithms can then be used for classification. Possible algorithms for clustering are for example k-Means – or the k independent meanshift algorithm – Hierarchical Cluster Analysis (HCA), or Expectation Maximization. In the field of analysis of RNA-sequencing data, the methods of t-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) have become particularly popular.

The methods of ML are faced with the methods of DL, whereby DL can be described as a subunit of ML. Often the terms are used synonymously – also together with AI – although there are differences. DL is based on the basic idea of the perceptron and thus in a certain way on neural networks. A perceptron is the artificial assumption of a neuron and tries to represent the architecture of the brain. There is an input layer, which is linked to the next layer via so-called weights – i.e. how important is this particular property from the input layer. If this subsequent layer is actually the output layer, it is called a neural network. If there are at least two more layers – so-called hidden layers – between the input and the output layer, it is called a Deep Neural Network (DNN). If there are at least 10 additional layers, it is called very deep learning. The procedure of the neural network (NN) is simply spoken repetitive learning and adapting the weights until the largest possible amount of data points of the learning dataset is correctly predicted. There are many different algorithms such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short term memory networks (LSTMs), gated recurrent unit (GRUs), Variational Autoencoders (VAEs), and GANs. DL has been applied in the field of medicine and oncology mainly in the area of image analysis. The use of "artificial intelligence" was predicted to have a great future especially in the area of radiology and pathology (60, 61). The fact that the determination of, for example, tumor entities based on image data has not yet been completely taken over by AI shows how difficult the work and also the evaluations in this area really are.

In addition, there are already approaches that use biological data such as methylation data (62–64) or RNA-sequencing data (65–73) to predict tumor entities. However, for RNA-sequencing but also for methylation data, these ML based approaches have so far failed to gain broad acceptance, so that there are probably no standardized ML models in routine diagnostics and medical care to date, even though the underlying methods and also specific analysis (e.g. hypermethylation of hMLH1 for Lynch-syndrome or MGMT for glioblastomas) are already used in routine diagnostics. The reasons for this are of different nature, whereby

one main problem is the used and available data. Another important problem to mention is the so-called batch effect, which describes specific changes in the analysis of biological samples within one run or analysis. Due to the occurrence of the batch effect, biological data are very difficult or even impossible to compare with each other, especially when they were generated by different laboratories and with different methods or kits. Harmonization of biological data, but especially of sequencing data, remains a major problem and is still an unmet need.

3.6 AIM OF THE PROJECT

In recent years, bioinformatics has become increasingly important within clinical and also oncological research. In addition to research, this also includes direct clinical applications, since the analysis and interpretation of sequencing data would not even be possible without bioinformatics, and thus plays a special role in personalized medicine. Despite the increasing importance and the solid establishment of bioinformatics in some areas of translational and personalized medicine, many of the efforts have not yet been able to make their way into clinical practice. In addition to the high cost of the required data or procedures and the time needed, other factors also play a role, such as broad applicability, often being a limiting factor.

The present work deals with the question of the usability of sequencing data – especially RNA-sequencing data – in clinical routine. For this, besides the introductory question to be answered whether RNA-sequencing data is at all suitable for entity prediction in a larger and more comprehensive approach using machine learning (ML), there are further problems to be solved. Specifically, it must be clarified whether complete RNA-sequencing is necessary for entity prediction, or whether the amount of sequencing required can be reduced – for example, to the size of a reduced gene panel – when methods of ML are applied. Another important aspect that has not yet been adequately addressed is the prediction of the underlying tumor entity in case of metastases. This not only includes cancer of unknown primary (CUP) cases but also samples of cancer patients with uncertain secondary tumor or metastatic status. Previous works always assumed that their methods and models can also be applied to metastases, although this application has not yet been conclusively answered or worked up theoretically. Therefore, one aim of this work is to answer the question whether ML models are applicable to metastases of different resection sites.

Furthermore, the question of usability and applicability must be addressed, otherwise routine use, especially at other institutions, cannot be guaranteed. In order to create a possibility for the routine application of ML within cancer diagnostics, data sets from different sources were

analyzed in the present work using ML methods to enable the determination of tumor entities based on RNA-sequencing data - both in the metastatic and primary state.

As different subgroups within various entities are already known and described, another important question is which therapeutic consequences can be concluded from identified tumor subgroups. To tackle this question, exemplarily the three different histopathological subgroups of renal cell carcinoma were analyzed using unbiased visual clustering methods. Subsequent characterization of obtained clusters using machine learning was performed to gain new insights into the different clusters but also subgroups and ultimately to evaluate possible new therapy options. Since visual clustering itself can be based on different data – e.g. logarithmic or unprocessed data – it was also important to investigate this aspect of visual clustering and to analyze the influence on the clustering itself.

Consequently, based on the results of the identification of the subgroups, a possible therapy should then be suggested (Figure 4).

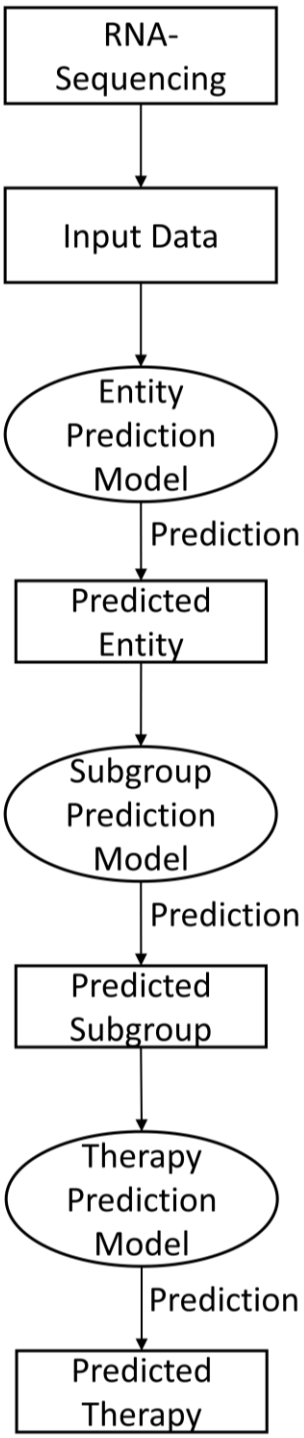


Figure 4: Proposed workflow for therapy prediction based on machine learning

Proposed workflow for tumor entity prediction with subsequent prediction of possible best suited personalized therapies starting with RNA-sequencing data as input data. From this point on the different steps of developing prediction models for entities and subgroups to therapy models and eventually a therapy prediction has to be performed.

4. MATERIAL AND METHODS

4.1 DATASETS

4.1.1 The Cancer Genome Atlas (TCGA)

As a consortium, The Cancer Genome Atlas (TCGA) aimed to characterize most cancer entities on molecular basis using NGS. For their studies, they performed whole exome sequencing (WES) to detect mutations, RNA-sequencing for transcriptomic differences and methylation-arrays for insights into the methylome. All these data are generated on tumor samples that have not yet been under treatment representing the tumor as it is. Additionally, the database contains further clinical information on each sample as well as corresponding histological samples – with respective information about them – making it the biggest and most comprehensive of its kind.

For this work, a total of 27 different tumor entities with a combined sample size of 9.260 specimen were used to generate the RNA-sequencing based model (Table 1). For one entity (Pheochromocytoma and Paraganglioma – PCpG), not only primary tumor samples but also the three additional groups – Additional New Primary PCpG, Solid Tissue Normal PCpG, and Metastatic PCpG – serving as a learning control, were used. In summary, the data basis serving as a learning cohort is comprised of 30 different labels. The data basis for each entity are the HTSEQ-FPKM files – representing the read counts normalized for sequencing depth and length of gene – that are publicly and freely available and are containing the sequencing results of 60,483 protein coding genes, non-protein coding genes (e.g. pseudo genes), microRNAs, and long-non-coding RNAs (lncRNAs).

TCGA-Identifier	Primary Site	Nr. of samples (RNA-Seq)	Nr. of samples (Methylation)
ACC (Adrenocortical Carcinoma)	• Adrenal gland	79	80
BLCA (Bladder Urothelial Carcinoma)	• Bladder	414	418
BRCA (Breast Invasive Carcinoma)	• Breast	1102	817
CESC (Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma)	• Cervix uteri	304	307
COAD (Colon Adenocarcinoma)	• Colon • Rectosigmoid junction	478	313
GBM (Glioblastoma Multiforme)	• Brain	156	140
ESCA (Esophageal Carcinoma)	• Esophagus • Stomach	161	185
HNSC (Head and Neck Squamous Cell Carcinoma)	• Base of tongue • Bones, joints and articular cartilage of other and unspecified sites • Floor of mouth • Gum • Hypopharynx	500	528

	• Larynx • Lip • Oropharynx • Other and ill-defined sites in lip, oral cavity and pharynx • Other and unspecified parts of mouth • Other and unspecified parts of tongue • Palate • Tonsil		
KIRC (Kidney Renal Clear Cell Carcinoma)	• Kidney	538	325
LAML (Acute Myeloid Leukemia)	• Hematopoietic and reticuloendothelial systems	151	140
LGG (Brain Lower Grade Glioma)	• Brain	511	516
LIHC (Liver Hepatocellular Carcinoma)	• Liver and intrahepatic bile ducts	371	317
LUAD (Lung Adenocarcinoma)	• Bronchus and lung	533	473
LUSC (Lung Squamous Cell Carcinoma)	• Bronchus and lung	502	370
MESO (Mesothelioma)	• Bronchus and lung • Heart, mediastinum, and pleura	86	87

OV (Ovarian Serous Cystadenocarcinoma)	• Ovary	374	10
PAAD (Pancreatic Adenocarcinoma)	• Pancreas	177	184
Primary Tumor PCpG (Pheochromocytoma and Paraganglioma)	• Adrenal gland • Connective, subcutaneous and other soft tissues • Heart, mediastinum, and pleura • Other and ill-defined sites • Other endocrine glands and related structures • Retroperitoneum and peritoneum • Spinal cord, cranial nerves, and other parts of central nervous system	178	179
Additional New Primary PCpG (Pheochromocytoma and Paraganglioma)	• See Primary Tumor PCpG	3	3
Solid Tissue Normal PCpG (Pheochromocytoma and Paraganglioma)	• See Primary Tumor PCpG	3	0
Metastatic PCpG (Pheochromocytoma and Paraganglioma)	• See Primary Tumor PCpG	2	0
PRAD (Prostate Adenocarcinoma)	• Prostate gland	498	502

READ (Rectum Adenocarcinoma)	• Rectal Adeno Carcinoma	0	98
SARC (Sarcoma)	• Bones, joints and articular cartilage of limbs • Colon • Connective, subcutaneous and other soft tissues • Corpus uteri • Kidney • Meninges • Other and unspecified male genital organs • Other and unspecified parts of tongue • Ovary • Peripheral nerves and autonomic nervous system • Retroperitoneum and peritoneum • Stomach • Uterus, NOS	259	261
SKCM (Skin Cutaneous Melanoma)	• Skin	103	104
STAD (Stomach Adenocarcinoma)	• Stomach	375	395
TGCT (Testicular Germ Cell Tumors)	• Testis	150	156
THCA (Thyroid Carcinoma)	• Thyroid gland	502	507

THYM (Thymoma)	• Heart, mediastinum, and pleura • Thymus	119	124
UCEC (Uterine Corpus Endometrial Carcinoma)	• Corpus uteri • Uterus, NOS	551	438
UVM (Uveal Melanoma)	• Eye and adnexa	80	80
Overall Sum		9260	8091

Table 1: Tumor entities used for machine learning by TCGA

Tumor Entities used as basis for machine learning with their assigned primary site, the corresponding TCGA-Identifier, and the number of samples available for either RNA-sequencing or methylation used in the learning process.

4.1.2 Further Evaluation Datasets for Entity Prediction

For evaluation purposes, publicly available FPKM data files provided either by the ICGC (https://dcc.icgc.org/projects/) (74) or gene expression omnibus (GEO – https://www.ncbi.nlm.nih.gov/geo/) (75) (Table 2) were used as an addition to four further TCGA datasets not included in the random forest learning approach. These additional datasets consist of 1999 samples in total.

Cohort Name	**Cohort Entity**	**Nr. of Samples (RNA-Seq)**
GSE135298(76)	Breast Cancer	93
GSE124535(77, 78)	Hepatocellular carcinoma (HCC)	35
GSE83533(79)	Acute Myeloid Leukemia (AML)	38
LIRI-JP	Liver Cancer	232
PRAD-CA	Prostate Carcinoma	144
RECA-EU	Clear Cell Renal Cell Carcinoma (ccRCC)	91

TCGA-READ	Rectal Carcinoma	166
TCGA-KIRP	Papillary Renal Cell Carcinoma (pRCC)	288
LICA-FR	Liver Cancer	161
BRCA-KR	Breast Cancer	50
TCGA-KICH	Chromophobe Renal Cell Carcinoma (chRCC)	65
ORCA-IN	Oral Cancer	40
GSE126975(80)	HNSC Cell lines	43
GSE92914(81)	Colon	12
PACA-AU(82)	Pancreas Carcinoma	91
OV-AU	Ovarial Cancer	93
PAEN-AU	Pancreas Carcinoma	32
PACA-CA	Pancreas Carcinoma	264
PRAD-FR	Prostate Carcinoma	25
Sum		1963

Table 2: Used evaluation cohorts

Datasets serving as evaluation cohorts for different steps in the validation of the random forest learning approaches, with their assigned primary site and the number of samples from RNA-sequencing present in the dataset.

4.1.3 Metastatic Datasets

Beside the analysis based on primary tumor samples, the intention of this work was also to introduce a model that can be applied to metastatic samples in order to predict their primary tumor. For this purpose, 14 further datasets were used for evaluation, of which ten are from TCGA database and four are from different sources. The latter are also used in further analysis regarding the dependency of transcriptomic features on the site of origin. The cohorts are comprising of 852 samples in total, from 5 different projects representing ten different primary sites of metastasis (Table 3).

Cohort Name	**Cohort Entity**	**Nr. of Samples (RNA-Seq)**
Metastatic TCGA-SKCM (83)	Skin Cutaneous Melanoma	367
Metastatic TCGA-THCA	Thyroid Carcinoma	8
Metastatic TCGA-SARC	Sarcoma	1
Metastatic TCGA-PRAD	Prostate Adenocarcinoma	1
Metastatic TCGA-PAAD	Pancreatic Adenocarcinoma	1
Metastatic TCGA-HNSC	Head and Neck Squamous Cell Carcinoma	2
Metastatic TCGA-ESCA	Esophageal Carcinoma	1
Metastatic TCGA-COAD	Colon Adenocarcinoma	1
Metastatic TCGA-CESC	Cervix Squamous Cell Carcinoma	2
Metastatic TCGA-BRCA	Breast Invasive Carcinoma	7
Dream Team (84)	Prostate Cancer	266
MBC-project	Breast Cancer	146
NEPC-WCM (85)	Neuroendocrine Prostate Cancer	49
Sum		852

Table 3: Used metastatic samples

Entities included in the random forest learning approach with metastatic samples used as evaluation cohorts for the best random forest model, with their assigned primary site and the number of samples from RNA-sequencing present in the dataset.

4.1.4 Renal Cell Carcinoma Dataset

The renal cell carcinoma (RCC) mostly is divided into its three major histopathologic groups – clear cell (ccRCC), papillary (pRCC), and chromophobe (chRCC) RCC. For model generation and RF analysis for cancer entity prediction only the largest subgroup, ccRCCs were used (Table 1). For the subsequent step of subgroup identification, the TCGA cohorts KIRP (n = 288, pRCC) and KICH (n = 65, chRCC) were additionally considered, forming the KIPAN (**KI**dney **PAN**cancer) dataset.

4.1.4.1 Further Renal Cell Carcinoma Evaluation Datasets

As an addition to the KIPAN dataset two further datasets served as external evaluation. The first dataset is the RECA-EU dataset (n=91), that also serves as an evaluation dataset in the cancer entity prediction model. The second dataset – GSE157256 (86) – consists of RCC that are caused by hereditary leiomyomatosis (hlRCC), which are also known as fumarate hydratase (FH)-deficient RCC (n=26).

4.2 MACHINE LEARNING

If not stated otherwise, all work was implemented in a Jupyter Notebook environment (version 7.5.0) with Python version 3.6.9. As additional libraries SciPy version 1.3.0 (87) and scikit-learn version 0.22.1 (88) were used.

The statistical analyses have been performed using SciPy stats module or the statannot module (version 0.2.2) using Kruskal-Wallis test (89). This test has been chosen due to the unknown behavior of biological sequencing data and tests, whether the two samples are from the same distribution or not.

For survival analysis the lifeline module (version 0.23.1) (90) for python was used, utilizing the KaplanMeierFitter.

4.2.1 Random Forest

The RF learning approaches, if not stated otherwise, in the respective result sections, were performed using the RF Classifier (RandomForestClassifier) of the sklearn.ensemble module for python. For learning and evaluation purposes a 50/50 split was utilized with 1.000 trees in the forest (n_estimators = 1000). This means, that the data was split into a learning cohort consisting of 50% of the samples and an evaluation cohort consisting of remaining 50% of the samples. The testing accuracy was based on the prediction results on the evaluation cohort, derived from the learning results of the learning cohort.

The RF itself is built on the principle of decision trees. In simple terms, RFs are hierarchically structured if/else questions and decisions. This principle is comparable with well-known games, in which one tries to find out which celebrity a person is, for example, with the help of questions that have to be answered with yes or no. The goal is always to find out with as few questions as possible.

A more concrete example would be to distinguish the geometric shapes of circle, ellipse, square and rectangle. For example, starting from the known properties of the shapes, one could begin by asking whether the shape looking for contains corners. If the answer to this question is "no", both the rectangle and the square are eliminated, whereas if the answer is "yes", they are the shapes that would remain. In the example with "no", one could then ask, whether all points on the line have the same distance to the center or not. If the answer to this question is yes, only the circle remains, if no, the decision falls on the ellipse. Continuing with the example of shapes with corners, a final question could be, whether all sides have the same length, which can only be answered with "yes" for the square, whereas the answer for the decision for the rectangle would have to be "no"(Figure 5).

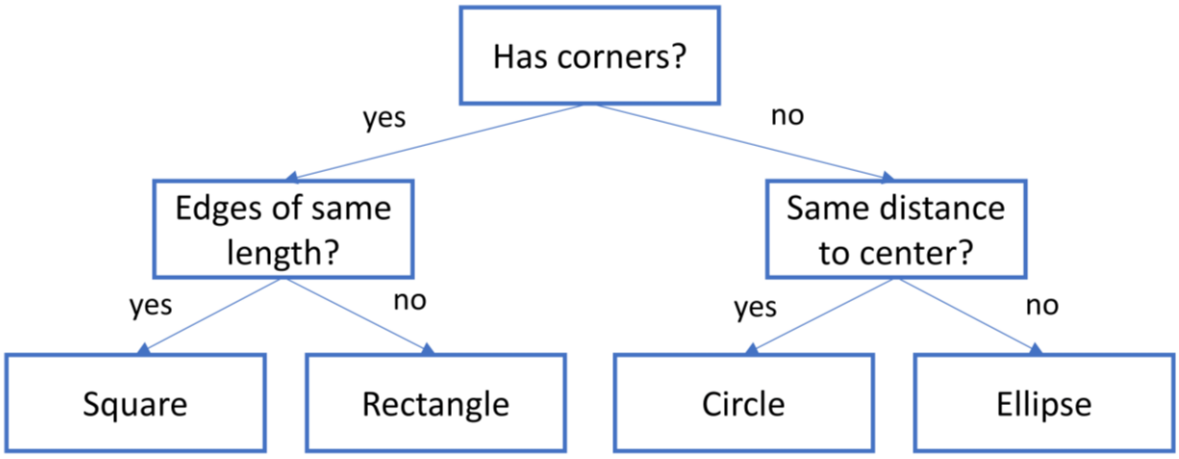


Figure 5: Basic decision tree example

Example of a basic decision tree, to decide on the shape of an unknown object, with indicated questions and the answer in the rectangles, and the possible answers to the questions indicated at the arrows.

As illustrated, the whole plot gets more complicated when adding for instance parallelograms, which would also need a question about the present angles of the corners.

Decision trees are generally very good at predicting the given training data, hence tend to overfitting, which is their major drawback, as they perform poorly at predicting new data. (91). The RF is one way to overcome this downside, as it is essentially just a collection of different decision trees, combined in one model (92). The general idea behind this ensemble method is, that if each tree tends to overfit on the different given data from the whole dataset, each tree overfits on only a part of the data and hence the obtained error gets minimized. Taken together, this results in completely different decision trees – the number of trees in the forest – that learn independently and finally make different decisions on the prediction of a sample.

To now obtain a prediction of the complete forest of decision trees – the RF – a prediction for each tree is performed and – due to the classification problems given in the presented work – subsequently a soft voting is used to obtain the final prediction. A soft voting in this context means, that each tree provides the probabilities for the possible output label of the model, which get averaged over all trees making the final decision the label with the highest probability.

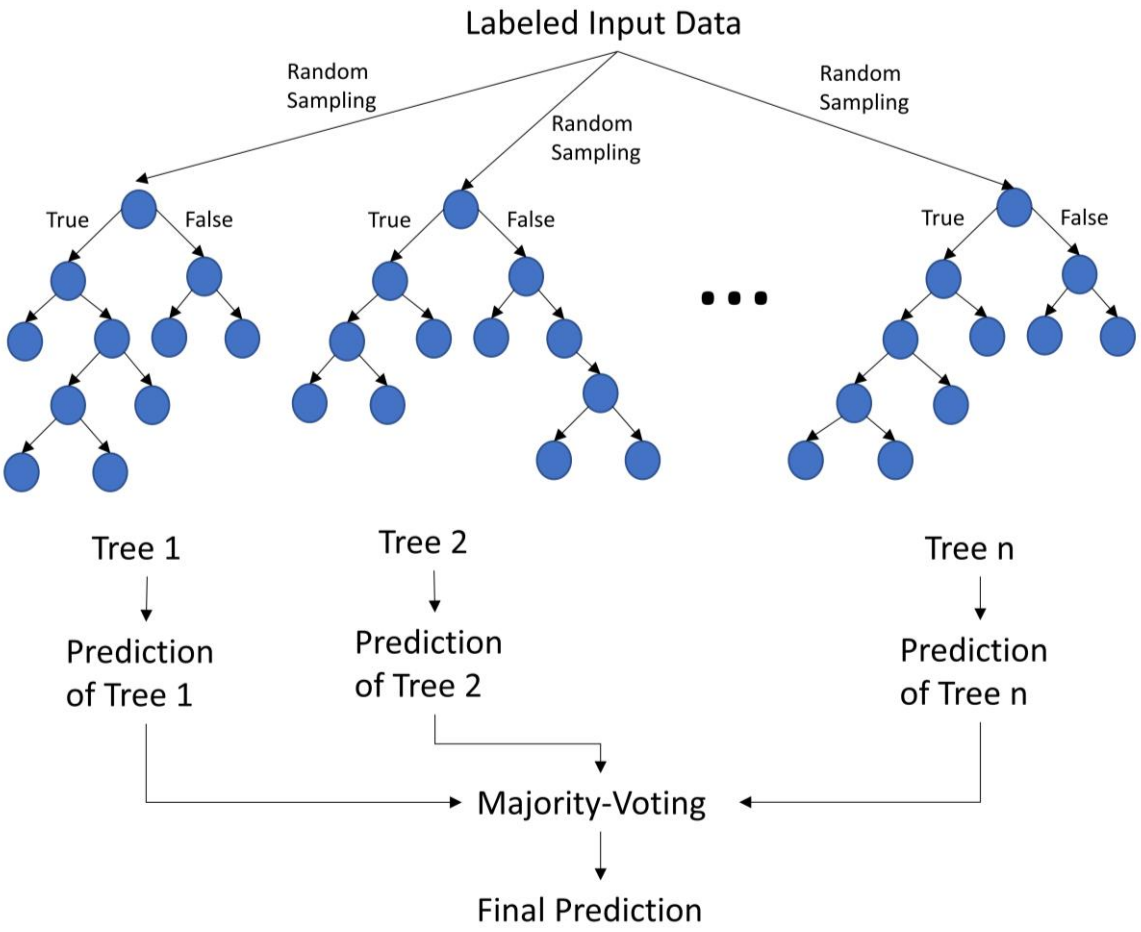


Figure 6: Exemplary presentation of a random forest

Example of the structure of a random forest consisting of n trees, depicting the random sampling fed into the respective decision trees with their prediction in the end. Subsequently, a majority voting is performed where the winner is the final prediction of the whole random forest.

To get further insight into the working of the trained model, the so-called features importances can be considered, representing the aggregated feature importances over all trees in the RF. It can be generalized, that features with higher values are more important for decision making of the RF than features with lower values (Figure 7).

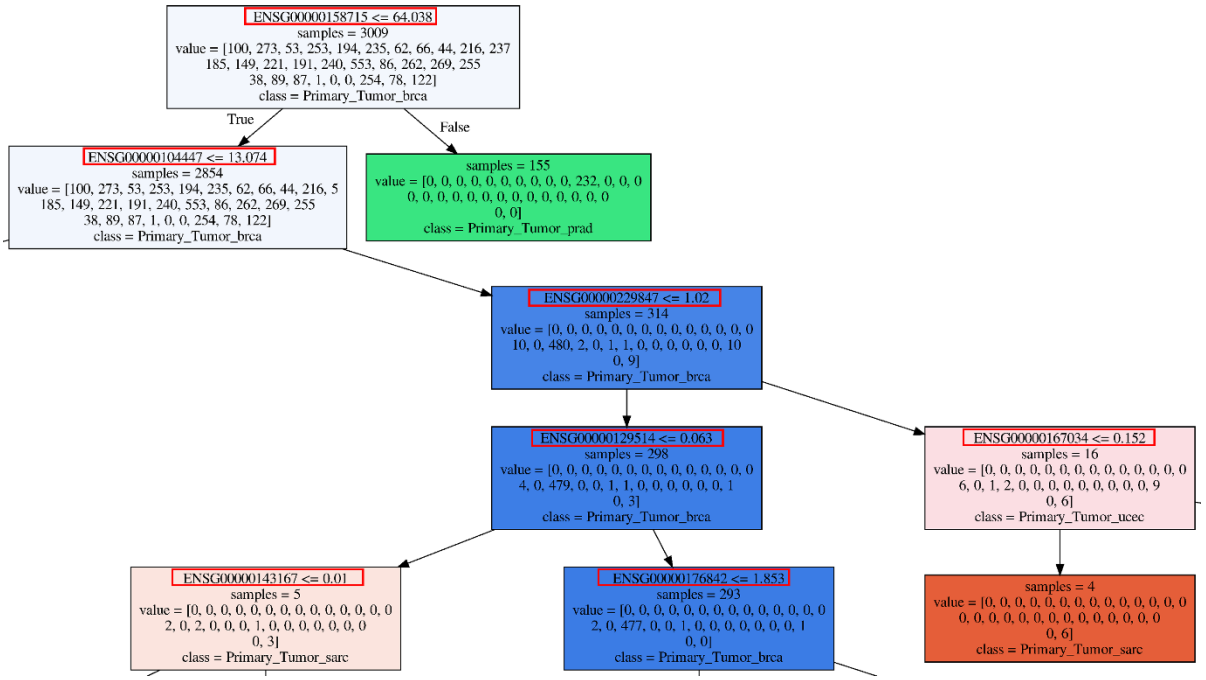


Figure 7: Decision making in a random forest tree

A section of an example of one tree in the forest of a generated random forest model showing the decisions made to get to a prediction based on the given features (marked in red). The different colours of the boxes indicate the possible classes.

4.2.2 Cross Validation

One possibility to test the performance of a model or the approach used is the so-called cross-validation in addition to the already mentioned testing accuracy under the respective split of the dataset into training and test dataset. In contrast to the multiple re-learning of the model based on the same split, the *k*-fold cross-validation gives a different insight into the quality of the model. Here, the dataset is divided into k (approximately) equal-sized, different partitions, the so-called folds are then divided differently and used to learn or evaluate the model. Frequently used parameter sizes for *k* are 5 or 10.

When using a 5-fold cross-validation, the available data would be divided into 5 sets. For each of these splits, the RF model is learned, which is then validated on the remaining set. Thus, in the first step, set 1 would serve as validation for the RF model learned on sets 2-5. In the second step, set 2 would serve as validation for sets 1, 3, 4, and 5. The learning and validation continues until each set has been used once for validation (Figure 2). Finally, *k* – in this example 5 – testing accuracies are obtained, which can be reported with standard deviation.

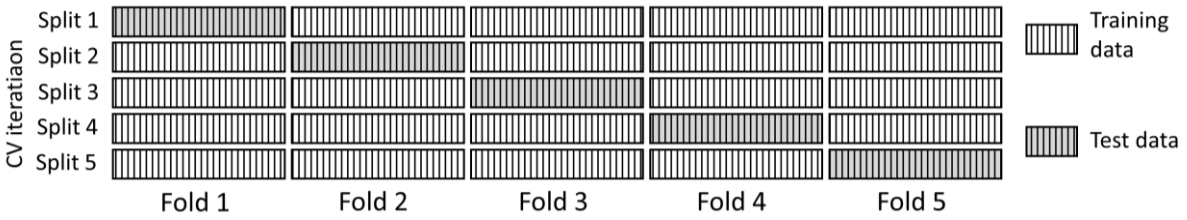


Figure 8: k-fold cross-validation

Schematic representation of a k-fold cross-validation using the example of k=5, with training data shown in white and test data shown in grey, for the different splits and the different data points.

One advantage of using cross-validation is the exclusion of randomness, since each sample is used exactly once for validation. This provides the advantage that supposedly difficult samples cannot fall out by chance, artificially raising the testing accuracy, making it unusually high. In the same way, however, samples that are supposedly easy to recognize are only used once in the validation cohort and therefore cannot falsify the testing accuracy upwards by chance. The advantage of this method goes along with the disadvantage of more computational resources needed, as *k* models are trained, also making this approach roughly *k* times slower.

One important downside of the cross-validation is the fact, that it can only be used for model evaluation purposes and not generating a model, making it unavailable to predict new data.

For datasets that are in ordered form and the different class sets have approximately the same size, a *k*-fold cross-validation may not be the best approach. This is due to the partitioning of the folds, which would lead to poor accuracies for the same number of classes relative to the chosen *k* under the circumstances mentioned above. The so-called stratified cross-validation can avoid this possible source of error. In comparison to the *k*-fold cross-validation, the ratios between the individual classes are included, whereby each set in the end reflects the proportions of the entire dataset (Figure 3). As the utilized scikit-learn module for python uses the stratified cross-validation by default, the mentioned *k*-fold cross-validation throughout the work are stratified cross-validation results (Figure 9).

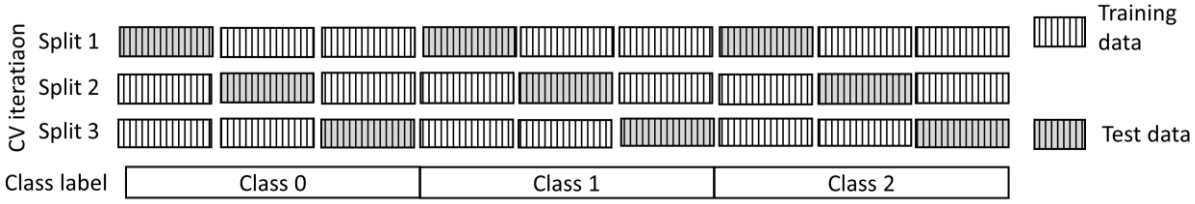


Figure 9: k-fold stratified cross validation

Schematic representation of a k-fold stratified cross-validation using the example of k=3, with training data shown in white and test data shown in grey, for the different splits and the different data points, additionally showing the class labels that are in an ordered manner for all data points.

4.3 Visual Clustering of High-Dimensional Data Using Dimension Reduction

Dimension reduction techniques are used to reduce the number of dimensions – also referred to as features – in high-dimensional datasets – such as RNA-sequencing data – in order to allow visualization. These methods have a wide range of applications, especially in the field of single-cell sequencing, where many data points or samples with tens of thousands of features (sequenced gene expressions) are acquired. The methods can be based on two basic techniques: either feature selection or feature extraction.

One of the most widely known methods for dimension reduction is the so-called principal component analysis (PCA), which is itself based on the principle of feature extraction. Other methods derived from this principle are, as variations of PCA, kernel PCA or graph-based kernel PCA. Other well-known methods like linear discriminant analysis (LDA) or generalized discriminant analysis (GDA) are based on this principle as well.

Procedures based on feature selection are for example backward feature elimination or forward feature selection. The latter is also used to a certain extent in this work to determine the most important features or genes within trained RF models.

Furthermore, there are also methods that combine both techniques, such as PCA or the already mentioned LDA. Additionally, methods such as canonical correlation analysis (CCA) or non-negative matrix factorization (NMF) can be mentioned here.

In addition to the methods and techniques, there are also methods based on projections, such as t-Distributed Stochastic Neighbor Embedding (t-SNE) or Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP), which will further be explained in greater detail as they are commonly used in the field of sequencing data and are used in the presented work.

In short, the high-dimensional data are projected on 2D or, if desired, on 3D and thus allow further analyzing of these data.

4.3.1 t-Distributed Stochastic Neighbor Embedding: The t-SNE Plot

If not stated otherwise, all t-SNE plots throughout this thesis are based on a PCA with 50 components rather than only 2 as it is commonly used, to explain a higher variance of the datapoints. For this purpose, the PCA of the sklearn.decomposite module was used and the

results were subsequently used as input for the t-SNE plotting utilizing the sklear.manifold module (93). Finally, 2D plotting (n_components=2) was performed using random initiation, perplexity=27, learning_rate=300, ni_iter=10.000, and for reproducibility random_state=0, as previously published (94).

The t-SNE plot – as a method for visualizing high-dimensional data – is an example of the class of visualization algorithms called manifold learning algorithms (MLA) usually applied after PCA, as PCA itself is only a good first approximation for the transformation of the data. Therefore, further and more complex methods need to be applied to this data afterwards (Figure 10A).

MLAs usually generate only two new features, as the main aim is to visualize the considered data. Additionally, they usually do not get used for intended supervised learning afterwards. MLAs can be very valuable for exploring new datasets and gaining a basic understanding of the dataset, whereas they usually cannot be used to predict new data. The t-SNE plot aims to represent the high-dimensional data in 2D. To achieve this, the algorithm transforms the equalities between the data points into joint probabilities and minimizes the Kullback-Leibler divergence (95). To do this, the algorithm starts with a random 2D representation for each data point. The t-SNE plot aims to bring points that are close together in the input data closer together and more distant data points further away. More emphasis is put on the distance between similar (close) points than on the distance between distant (far away) data points. As a result, a t-SNE plot can make statements about occurring clusters within the considered data because points within the clusters are similar, yet no global statements can be made based on the distance between two clusters or points (Figure 10B). Besides the non-global conservation of distances, the fact that t-SNE cannot work directly with high dimensional data and those results are different due to the convex cost function with different initializations can be seen as further disadvantages.

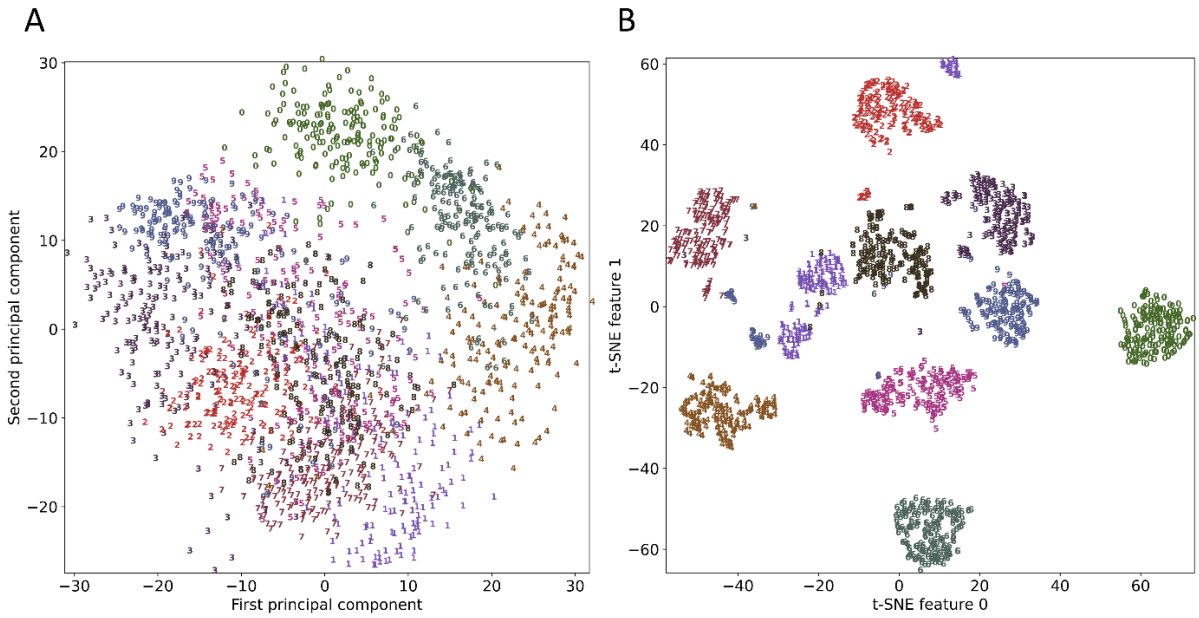


Figure 10: Difference of PCA and t-SNE by the example of MNIST dataset

(A) Principle component analysis with 2 components and (B) after being used for t-SNE plotting, utilizing the load_digits() function of the sklearn.datasets module to obtain the data for the handwritten numbers of the MNIST dataset.

4.3.2 UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

To overcome the problems of the t-SNE plot, the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) method can be used (96).

If not stated otherwise, the UMAPs presented here are based on an altered approach of the UMAP, not using a specific library for this, as it has shown to tend to better clustering results with less computational time and better scaling for bigger datasets (97). The approach has already been published previously (98). In short: the squared pairwise Euclidean distance for the distance of sample pairs is used for the initial dataset. On this, rho – the local connectivity parameter – and the first nearest neighbor are based and the sum of probabilities in the high-dimensional space is calculated – the so-called distance matrix. After that, the entropy – combining information about the nearest neighbors and the probabilities for each entry – and the optimal rho– based on a binary search for a fixed number of 15 nearest neighbors – are computed. The symmetry condition gets computed in a different, more simplified way, by dividing the sum of the probabilities by 2. The low-dimensional probabilities were built using mind_dist = 0.25, and cross-entropy was used as a cost function, utilizing a normalized Q

parameter. Lastly, the gradient of it was fed into the regular gradient descent learning, with 2 dimensions and 50 neighbors.

The UMAP is based on three assumptions about the data:

(The following was cited from https://umap-learn.readthedocs.io/en/latest/, accessed on 12.10.201 at 23:20)

1. *That the data is uniformly distributed on Riemannian manifold;*
2. *The Riemannian metric is locally constant (or can be approximated as such);*
3. *The manifold is locally connected.*

The UMAP algorithm employs some major advantages in contrast to the t-SNE plot. First, UMAP uses exponential probability distribution in high dimensions, which are not normalized. Next, the local connectivity parameter rho ensures the local connectivity of the manifold, giving a locally adaptive exponential kernel, finally making the distance matrix different for each point to point. Without the normalization, the UMAP outperforms t-SNE for large datasets, e.g. for single cell sequencing datasets. Additionally, UMAP does not apply a random normal initialization but rather uses Graph Laplacian to initiate the low-dimensional coordinates, making the results of UMAP more reproduceable. Instead of the perplexity parameter used by t-SNE, UMAP does use the number of nearest neighbors in addition to a symmetrization of high-dimensional probabilities, that is different and defined as the subtraction of the product of the probability and the transposed probability from the sum of the probability and the transposed probability:

$$p_{ij} = p_{i|j} + p_{j|i} - p_{i|j}p_{j|i}$$

Compared to t-SNE, the UMAP does use a binary cross-entropy instead of the KL-divergence, enabling UMAP to conserve global distances between datapoints. This makes it easier to interpret the results and draw relations between different observable clusters, which is why UMAP gets preferably used for single cell sequencing data.

5. RESULTS

5.1 CANCER ENTITY PREDICTION

5.1.1 COMPARISON OF DIFFERENT ALGORITHMS FOR PREDICTION

To get an overview of how well different machine learning algorithms perform in classifying the different considered 27 tumor entities (Table 1), based on RNA-sequencing data, six different algorithms were tested at the beginning. As a first test, to determine the further used approach, only the first 200 samples of each entity were used, if the dataset consisted of more than 200 samples.

The different algorithms tested were logistic regression (LogisticRegression), RF (RandomForestClassifier), Gaussian Naive Bayes (GaussianNB), k-nearest neighbor (kNeighborsClassifier), Linear Support Vector Classification (LinearSVC), and SupportVectorClassification (SVC). For the k-nearest neighbor algorithm, four different values for the parameter k were chosen: 1, 5, 10, and 30. Beside this adaption of parameters, all approaches have been performed using the default settings, without any in-depth hyperparameter scan.

To determine the quality for each algorithm and for comparison reasons, the F1 measurement was used. The F1 measurement uses the results of the prediction, such that the quality is based on the harmonic mean of precision and recall. As it can be seen, there are quite some differences between the tested algorithms, using the F1 measurement as quality score (Figure 11). The worst performing approach in the chosen setup was the Support Vector Classification with a F1 testing score below 0.1. The k-nearest neighbor approach was performing a little bit better with a F1 testing score of around 0.6 independent of the chosen value for k. Nevertheless, there were small differences, with an increasing F1 testing score from k=1 over k=5 to k=10. Interestingly, the F1 testing score dropped for the highest k. The four best performing models, regarding the automated calculated F1 testing score, were the linear regression, RF, Gaussian Naïve Bayes and the linear SVC with scores above 0.8, with the RF outperforming all other algorithms.

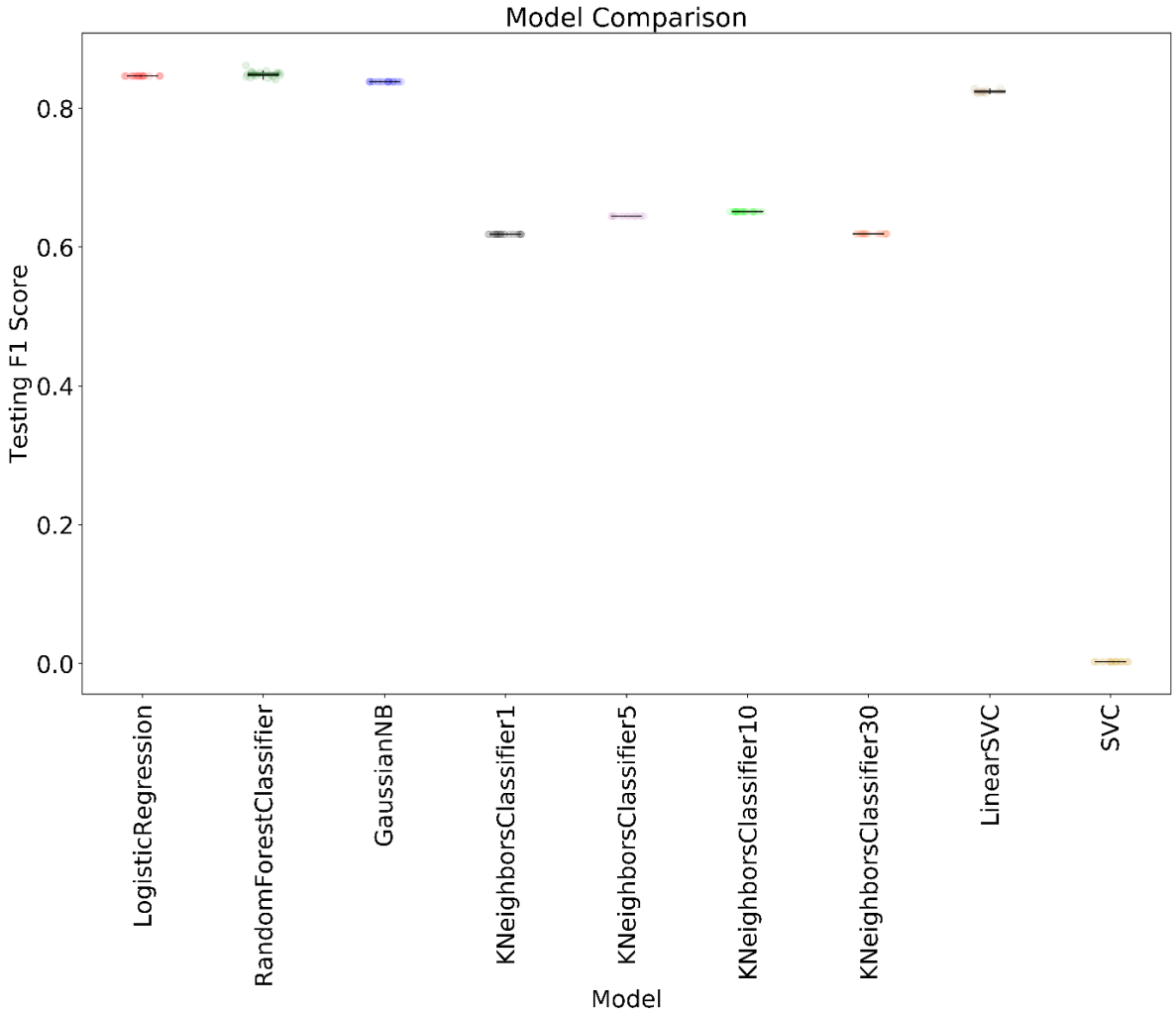


Figure 11: F1 Score comparison of different algorithms for entity prediction

Boxplots of the testing F1 score for different algorithms based on the learning and testing of the top 200 samples per entity downscaled RNA-sequencing dataset; NB: Naïve Bayes, SVC: linear vector classifier.

5.1.2 Methylation vs. RNA-Sequencing Data

After testing different ML approaches based on RNA-sequencing data, it was also necessary to select the underlying type of data used in for model generation. The two data types in question are either methylation data -derived from methylation arrays – or RNA-sequencing data. For RNA-sequencing data different approaches already exist, but for various reasons they are not yet established in clinical routine work. Models generated based on methylation data are probably the best currently existing, making investigations on the possible input data necessary.

To make a statement about the quality of the chosen learning approach for both cases, several random forest models were trained individually utilizing a 50/50 split.

Using all of the 8091 samples with methylation data (Table 1), the 541 different trained models resulted in a mean testing accuracy of 93.71% (min. 92.73%, max. 94.00%).

For RNA-sequencing, 100 different models were trained using all of the available 9260 samples (Table 1) as basis, resulting in a mean testing accuracy of 96.14% (min. 95.58%, max. 96.76%).

Due to the RNA-sequencing outperforming methylation data with the mean testing accuracy being over 2 percent points better, the decision was made to use RNA-sequencing data in combination with RF analysis as the basis of this work.

5.1.3 Utilizing the Best Performing RF Model to Predict Tumor Entities

It was furthermore of particular interest, whether the generated models were able to predict the correct corresponding tumor entity for RNA-sequencing samples that were not involved in the training of the actual model, but were generated analogous to the training data shown in Table 1. For this purpose, the additional datasets TCGA-READ (rectal adenocarcinomas), TCGA-KIRP (papillary RCCs – pRCCs), and TCGA-KICH (chromophobe RCCs – chRCCs) were considered. For evaluation purposes the best performing RF RNA-sequencing model was used to predict the analogous entities of those three datasets, namely, TCGA-COAD (colon adenocarcinomas) and TCGA-KIRC (for both pRCC and chRCC). For all these datasets the underlying entity are of the equal tissue of origin as the considered one. The results also confirmed this, as all of the 166 (100%) TCGA-READ samples, 283 of 288 (98.26%) of the TCGA-KIRP samples, and 63 of 65 (96.92%) of the TCGA-KICH samples were predicted correctly (Table 4).

Dataset	Entity	Approach	Number samples	Correctly Predicted	Percentage [%]
TCGA-READ	Rectal adenocarcinoma	All genes	166	166	100.00
TCGA-KIRP	Papillary Renal Cell Carcinoma	All genes	288	283	98.26
TCGA-KICH	Chromophobe Renal Cell Carcinoma	All genes	65	63	96.92

Table 4: Evaluation cohorts by TCGA for best performing random forest model

The three different cohorts from TCGA used for further evaluation of the best performing random forest RNA-sequencing model, using all genes to predict one of the 30 underlying entities. Besides the respective entity, the number of samples and the number of the assumed to be correct entity predictions with the additional percentage of correctly predicted samples are given. For the datasets READ, KIRP, and KICH the expected outcome entities were COAD, KIRC, and KIRC, respectively.

5.1.4 Analysis of Wrong Tumor Entity Assignments

Additionally, not only the overall accuracy of the best performing model, but also false predictions were of interest to get a better understanding of the model. To address the issue of false predictions and to see whether there are entities that are more commonly false predicted than others, the best performing model was used to make a prediction for *every* sample in the underlying dataset. Overall, this prediction showed that the model correctly identified the corresponding tumor entity for 9113 of 9260 (98.41%) samples, meaning that 147 (1.59%) samples were predicted incorrectly.

A Sankey plot for all predictions (n = 9260, Figure 12 A), especially for false predictions made by the developed model (n = 147, Figure 12 B), indicates that most of the false predictions occurred among closely related entities like lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) or esophageal adenocarcinomas (ESCA) and stomach adenocarcinomas (STAD), respectively . In general – beside the aforementioned LUSC/LUAD and STAD/ESCA problems – the Sankey plot indicates the major problems in correctly predicting entities like head and neck squamous cell carcinomas (HNSC), breast cancer (BRCA), sarcomas (SARC), and uterine corpus endometrial carcinoma (UCEC). Furthermore, in the case of squamous cell carcinomas such as HNSC, LUSC or CESC, a large overlap with BLCA, CESC, HNSC and LUSC was apparent, displaying a certain variety of predicted entities. This contrasts with the false predictions in adenocarcinomas, such as STAD or LUAD, which were only predicted as ESCA or LUSC, respectively, not showing any variety in prediction. A

certain exception were the false predictions of sarcomas (SARC), which had predictions in a total of 11 different entities.

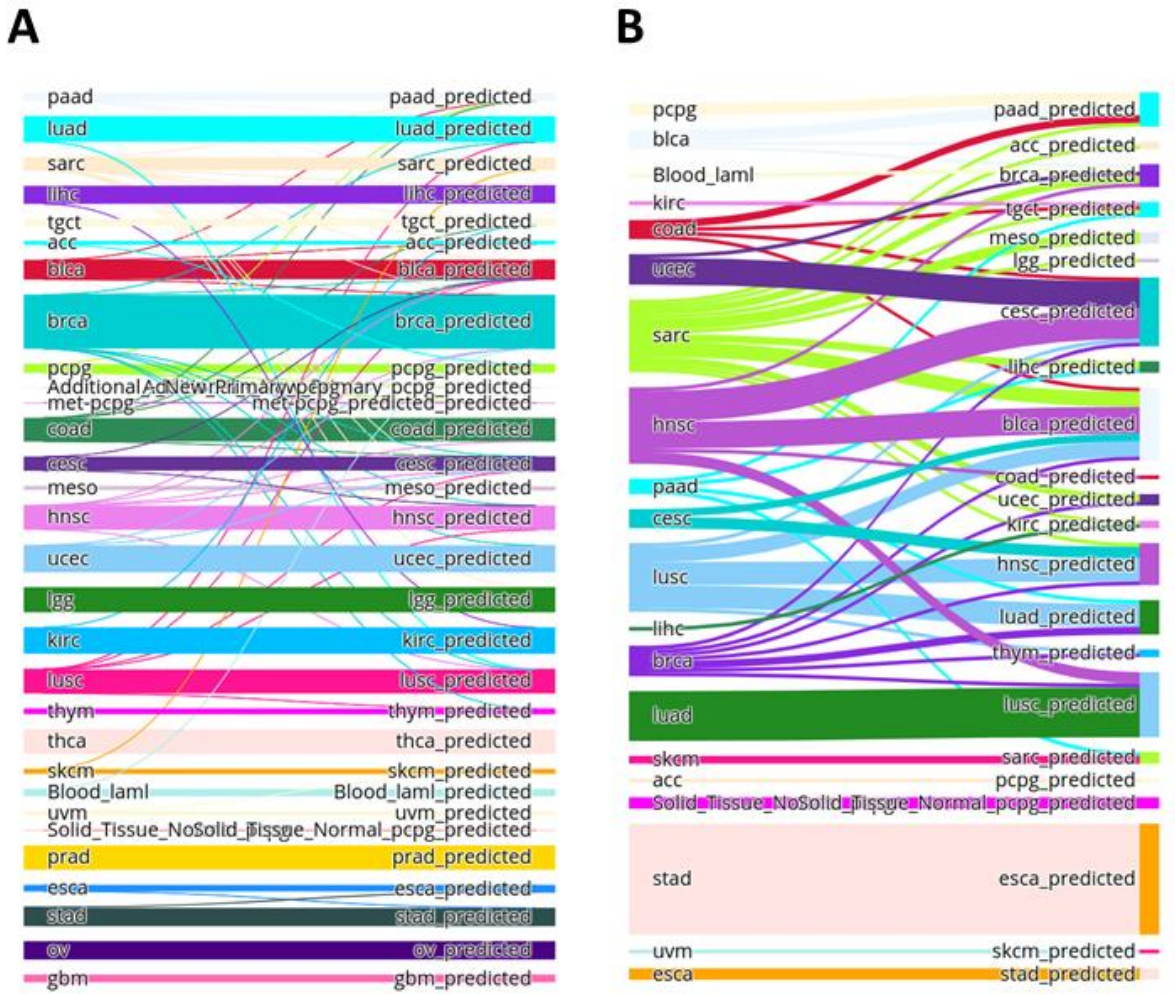


Figure 12: Sankey plot representation of random forest prediction results

Sankey plot representing (A) all entities with their respective TCGA identifier as given in Table 1,their sample size given as boldness of the line (n = 9260) on the left side and all predictions made by the best performing random forest (RF) RNA-sequencing model on the right side. Non-horizontal lines indicate false predictions. (B) represents only the false predictions (n = 147) made by the best performing RNA-sequencing RF model, again showing the respective TCGA identifier on the left side with their respective false predictions on the right side.

5.1.4.1 Improving the Results of the Random Forest

After analyzing the false predictions (Figure 12 B), histopathologically closely related entities were combined, as it has already been done before (62). Following this approach, the entities ESCA and STAD – both derived from the upper gastrointestinal tract – and LUSC and LUAD – both representing lung cancer – were combined to the broader entities “stomach” and “lung”, respectively.

Accordingly, the combination increased the accuracy from 98.41% to 98.96% by adding 51 correct predictions, reducing the false predictions from 147 to 96 (Figure 12).

5.1.5 Reducing the Number Genes for Prediction

The initial analysis comprised all genes contained in the dataset, utilizing the results of RNA-sequencing. Since RNA-sequencing is most commonly not available in the clinical everyday life, the important next step was to test whether the reduction of the gene number to a minimum was possible to make the approach routinely available – as panel for example – in order to save time and money. In a first approach, the number of genes under consideration was reduced to 100, which is as much as 0.165% of the initial number of genes.

Since feature selection is a very time-consuming procedure, a simplified selection approach was utilized. The feature values obtained by the best performing model using all genes were used as starting point, followed by deleting every feature with a value equal to zero, subsequently starting the actual feature selection process. Finally, the number of occurrences of a gene in the top100 transcripts with the highest feature values were counted and combined for all models, only using the 100 genes that had the highest top100 counts. Based on these 100 genes, 20 new RF models were trained with a mean testing accuracy of 93.80% (min. 93.33 max. 94.14%), which is only 2.34 percent points in mean worse than using all genes.

Approach	Minimum [%]	Maximum [%]	Mean [%]
All genes	95.58	96.76	96.14
Top100 genes	93.33	94.14	93.80

Table 5: Testing accuracy comparison between all genes and top100 genes

Comparison between the two different tested approaches, either using all or only a selected 100 transcripts for random forest analysis, depicting minimum, maximum, and the mean top1-testing accuracy.

Because of this result, another model was trained on all genes, this time without the usage of feature selection steps, again utilizing the above-mentioned ensemble technique. The genes that were in the top100 the most often overlapped in 98.00% with the top100 transcripts obtained with feature selection steps. This showed that there is little to no difference between the two analyzed feature selection approaches. In all subsequent analyses, the top100 genes obtained utilizing feature selection – rather than the ensemble method – were used (Table 6).

SN	ENSG ID	Type	HGNC	Description	Diagnostic Usage
1	ENSG00000042832	pep	TG	thyroglobulin	used as Thyro-globulin
2	ENSG00000142515	pep	KLK3	kallikrein related peptidase 3	used as PSA
3	ENSG00000158715	pep	SLC45A3	solute carrier family 45 member 3	used as prostein
4	ENSG00000014257	pep	ACPP	acid phosphatase, prostate	used as PAP
5	ENSG00000073282	pep	TP63	tumor protein p63	used as p63
6	ENSG00000112499	pep	SLC22A2	solute carrier family 22 member 2	used as OCT2
7	ENSG00000131400	pep	NAPSA	napsin A aspartic peptidase	used as napsin A
8	ENSG00000115705	pep	TPO	thyroid peroxidase	used as MSA
9	ENSG00000138792	pep	ENPEP	glutamyl aminopeptidase	used as CD249
10	ENSG00000165556	pep	CDX2	caudal type homeobox 2	used
11	ENSG00000091831	pep	ESR1	estrogen receptor 1	used
12	ENSG00000129514	pep	FOXA1	forkhead box A1	used
13	ENSG00000107485	pep	GATA3	GATA binding protein 3	used
14	ENSG00000136352	pep	NKX2-1	NK2 homeobox 1	used
15	ENSG00000125618	pep	PAX8	paired box 8	used
16	ENSG00000077498	pep	TYR	tyrosinase	used
17	ENSG00000167034	pep	NKX3-1	NK3 homeobox 1	used
18	ENSG00000165409	pep	TSHR	thyroid stimulating hormone receptor	
19	ENSG00000113494	pep	PRLR	prolactin receptor	
20	ENSG00000253563	ncrna	NKX2-1-AS1	NKX2-1 antisense RNA 1	
21	ENSG00000110484	pep	SCGB2A2	secretoglobin family 2A member 2	

22	ENSG00000167751	pep	KLK2	kallikrein related peptidase 2	
23	ENSG00000167749	pep	KLK4	kallikrein related peptidase 4	
24	ENSG00000197308	ncrna	GATA3-AS1	GATA3 antisense RNA 1	
25	ENSG00000164736	pep	SOX17	SRY-box 17	
26	ENSG00000135903	pep	PAX3	paired box 3	
27	ENSG00000113722	pep	CDX1	caudal type homeobox 1	
28	ENSG00000178919	pep	FOXE1	forkhead box E1	
29	ENSG00000101076	pep	HNF4A	hepatocyte nuclear factor 4 alpha	
30	ENSG00000122852	pep	SFTPA1	surfactant protein A1	
31	ENSG00000185303	pep	SFTPA2	surfactant protein A2	
32	ENSG00000205002	pep	AARD	alanine and arginine rich domain containing protei	
33	ENSG00000183747	pep	ACSM2A	acyl-CoA synthetase medium chain family member 2A	
34	ENSG00000066813	pep	ACSM2B	acyl-CoA synthetase medium chain family member 2B	
35	ENSG00000148513	pep	ANKRD30A	ankyrin repeat domain 30A	
36	ENSG00000160862	pep	AZGP1	alpha-2-glycoprotein 1, zinc-binding	
37	ENSG00000120903	pep	CHRNA2	cholinergic receptor nicotinic alpha 2 subunit	
38	ENSG00000143578	pep	CREB3L4	cAMP responsive element binding protein 3 like 4	
39	ENSG00000127377	pep	CRYGN	crystallin gamma N	

40	ENSG00000225362	pep	CT62	cancer/testis antigen 62	
41	ENSG00000035664	pep	DAPK2	death associated protein kinase 2	
42	ENSG00000115468	pep	EFHD1	EF-hand domain family member D1	
43	ENSG00000170370	pep	EMX2	empty spiracles homeobox 2	
44	ENSG00000150667	pep	FSIP1	fibrous sheath interacting protein 1	
45	ENSG00000143167	pep	GPA33	glycoprotein A33	
46	ENSG00000159184	pep	HOXB13	homeobox B13	
47	ENSG00000176842	pep	IRX5	iroquois homeobox 5	
48	ENSG00000009765	pep	IYD	iodotyrosine deiodinase	
49	ENSG00000153822	pep	KCNJ16	potassium voltage-gated channel subfamily J member 16	
50	ENSG00000197705	pep	KLHL14	kelch like family member 14	
51	ENSG00000136944	pep	LMX1B	LIM homeobox transcription factor 1 beta	
52	ENSG00000007952	pep	NOX1	NADPH oxidase 1	
53	ENSG00000167332	pep	OR51E2	olfactory receptor family 51 subfamily E member 2	
54	ENSG00000072042	pep	RDH11	retinol dehydrogenase 11	
55	ENSG00000164265	pep	SCGB3A2	secretoglobin family 3A member 2	
56	ENSG00000229415	pep	SFTA3	surfactant associated 3	
57	ENSG00000168878	pep	SFTPB	surfactant protein B	
58	ENSG00000168484	pep	SFTPC	surfactant protein C	

59	ENSG00000188467	pep	SLC24A5	solute carrier family 24 member 5	
60	ENSG00000104154	pep	SLC30A4	solute carrier family 30 member 4	
61	ENSG00000124664	pep	SPDEF	SAM pointed domain containing ETS transcription factor	
62	ENSG00000109436	pep	TBC1D9	TBC1 domain family member 9	
63	ENSG00000089225	pep	TBX5	T-box 5	
64	ENSG00000118526	pep	TCF21	transcription factor 21	
65	ENSG00000134490	pep	TMEM241	transmembrane protein 241	
66	ENSG00000184012	pep	TMPRSS2	transmembrane serine protease 2	
67	ENSG00000186854	pep	TRABD2A	TraB domain containing 2A	
68	ENSG00000124900	pep	TRIM51	tripartite motif-containing 51	
69	ENSG00000104447	pep	TRPS1	transcriptional repressor GATA binding 1	
70	ENSG00000178935	pep	ZNF552	zinc finger protein 552	
71	ENSG00000235584	ncrna	AC008268.1	novel transcript	
72	ENSG00000228650	ncrna	AC008940.1	novel transcript	
73	ENSG00000228835	ncrna	AC012123.1	novel transcript, antisense to KLHL14	
74	ENSG00000259793	ncrna	AC013726.1	novel transcript, antisense to DGKD	
75	ENSG00000223808	ncrna	AC044784.1	novel transcript	
76	ENSG00000259725	ncrna	AC106738.2	novel transcript	

77	ENSG00000234918	ncrna	AL157387.1	novel transcript	
78	ENSG00000224842	ncrna	AL161908.1	novel transcript, antisense to LIM1B	
79	ENSG00000276888	ncrna	AL512624.2	novel transcript, antisense to POTEM	
80	ENSG00000270090	ncrna	AL590235.1	novel transcript	
81	ENSG00000275563	ncrna	AL929601.3	novel transcript, antisense to POTEG	
82	ENSG00000229847	ncrna	EMX2OS	EMX2 opposite strand/antisense RNA	
83	ENSG00000228262	ncrna	LINC01320	long intergenic non-protein coding RNA 1320	
84	ENSG00000258586	ncrna	LINC02274	long intergenic non-protein coding RNA 2274	
85	ENSG00000236130	ncrna	PTCSC2	papillary thyroid carcinoma susceptibility candidate 2	
86	ENSG00000259104	ncrna	PTCSC3	papillary thyroid carcinoma susceptibility candidate 3	
87	ENSG00000252621	ncrna	RF00019	-	
88	ENSG00000257520	ncrna	SFTA3	surfactant associated 3	
89	ENSG00000255399	ncrna	TBX5-AS1	TBX5 antisense RNA 1	
90	ENSG00000274310	cdna	AC091076.1	SRY (sex determining region Y)-box 17 (SOX17) pseudogene	

91	ENSG00000240237	cdna	AF305872.1	ribosomal protein L21 (RPL21) pseudogene	
92	ENSG00000214313	cdna	AZGP1P1	alpha-2-glycoprotein 1, zinc-binding pseudogene 1	
93	ENSG00000179362	cdna	HMGN2P46	high mobility group nucleosomal binding domain 2 pseudogene 46	
94	ENSG00000235687	cdna	LINC00993	long intergenic non-protein coding RNA 993	
95	ENSG00000242899	cdna	RPL7P16	ribosomal protein L7 pseudogene 16	
96	ENSG00000249661	cdna	TNRC18P1	trinucleotide repeat containing 18 pseudogene 1	
97	ENSG00000236313	cdna	VN1R53P	vomeronasal 1 receptor 53 pseudogene	
98	ENSG00000240800				
99	ENSG00000243350				
100	ENSG00000227869				

Table 6: Identified top100 genes

The 100 genes with the highest feature values as identified by the best performing random forest model referred to as top100 transcripts. All the top100 transcripts are listed according to a serial number (SN) with their unique ENSG identifier, the type of the encoded transcript (pep =peptide/protein-coding, ncrna = non-coding RNA, cdna = transcribed pseudogenes – either processed or unprocessed), their corresponding HUGO gene symbol (HGNC), and their description or rather their corresponding trivial name. The genes that are either already used in pathological routine work or are closely related to those do have an additional comment beside their description.

The results showed that the use of the top100 transcripts is almost as good as the use of all genes in predicting their potential entities based on RNA-sequencing. A closer look at these 100 genes showed that a large part of the genes identified in this way are already sufficiently known as marker genes and are routinely used by pathologists. Of these 100 identified transcripts, the proteins of at least 17 are already used by pathologies (examples shown in Figure 13 A). Incidentally, among the top100 transcripts were also two non-coding-RNAs (ncRNA) namely NKX2-1-AS1 and GATA3-AS1, which are antisense RNAs to histopathological known genes NKX2-1 (Thyroid Transcription Factor 1 – used for differential

diagnosis of lung and thyroid carcinomas) and GATA3 (GATA Binding Protein 3 – a valuable marker for breast cancer diagnosis (99, 100)). Beside these already known transcripts, there are several protein-coding transcripts that could potentially serve as biomarker for certain entities (Figure 13 B), as they show expression mainly for one specific entity. Of note, the data basis also made it possible to analyze non-coding elements, like (long) non coding RNAs but also micro RNAs (miRNAs), which could also serve as marker transcripts for particular entities (Figure 13 C).

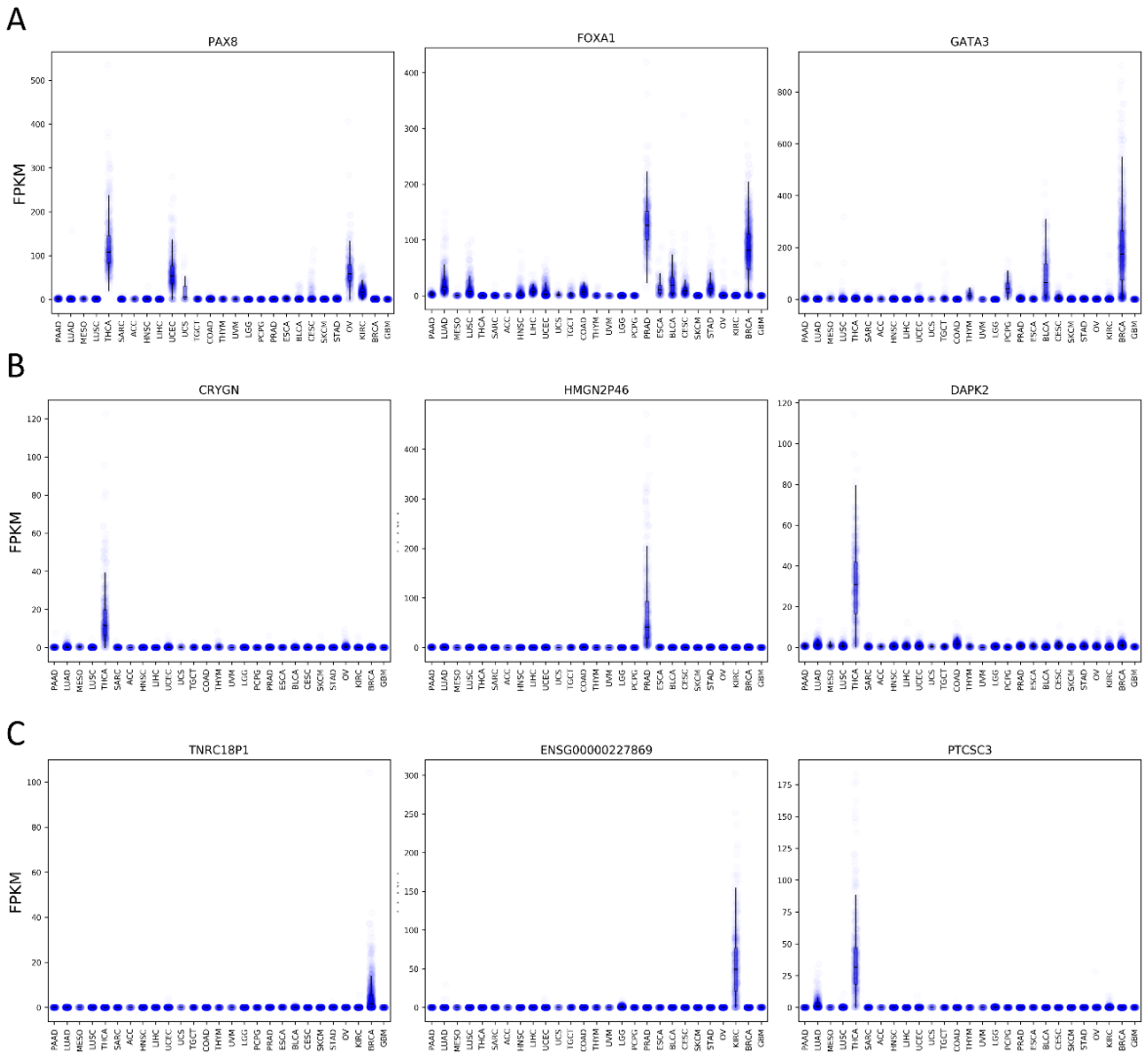


Figure 13: RNA-expression boxplots of bona fide top100 candidate genes

Boxplots for each entity with their respective TCGA identifier as given in Table 1, representing the FPKM expression. (A) shows three representaive genes of the top100 transcripts, that are already known and established for use by pathologists; (B) represents four manually selected proteincoding genes of the top100 transcripts that show expression in only one entity without established use; (C) showing four genes of the top100 transcripts that are not proteincoding but only show expression in one respective entity and could therefore serve as potential new biomarkers.

5.1.5.1 Validation and Comparison of the Reduced RNA-Sequencing Gene Set

To further validate the top100 transcripts as a specific 100 gene signature, the obtained testing accuracy – again utilizing RF learning – for a gene signature with 100 randomly selected genes was analyzed. In total, 10.000 different models were generated, resulting in a mean testing accuracy of 87.81% (min. 80.20% max. 92.23%). The best performing random gene model already showed how much power the RF has in predicting the correct entity, as 100 random genes are good enough to have a testing accuracy of more than 92% in the best case. Taking a closer look at the randomly selected genes of the 10 best performing random models showed an overlap of only one single gene for all models.

In comparison, when using the best performing model with the selected 100 genes to predict all samples, 9003 out of the 9260 samples (97.22%) were predicted correctly.

Again, combining LUSC and LUAD and ESCA and STAD, respectively, increased the correct results by 67 (overall accuracy of 97.95%), only 1.01 percent points worse than using all transcripts utilizing the entity combination.

Using the best performing RF model based on the top100 transcripts to predict the tumor samples of the three evaluation datasets (READ, KIRP, KICH) only little differences are noticeable when comparing the results with the ones obtained utilizing all available transcripts. For the KIRP cohort as well as for the KICH cohort, there is no difference in the number of correctly predicted samples, whereas for the READ cohort, two samples were predicted wrong (Table 7).

These results underline and further evaluate the quality of the selected top100 transcripts as an individual relevant gene signature.

Dataset	Entity	Approach	Amount samples	Accordingly Predicted	Percentage [%]
TCGA-READ	Rectal adenocarcinoma	Top100	166	164	98.80
TCGA-KIRP	Papillary Renal Cell Carcinoma	Top100	288	283	98.26
TCGA-KICH	Chromophobe Renal Cell Carcinoma	Top100	65	63	96.92

Table 7: Prediction evaluation of TCGA evaluation cohorts using top100 genes

The three different cohorts from TCGA that were used to further evaluate the best performing random forest model using the top100 transcripts to predict one of the 30 underlying entities. The respective entity is given as well as the number of samples and the number of correct predicted ones with the additional percentage of correctly predicted samples. For the datasets READ, KIRP, and KICH the expected outcome entities were COAD, KIRC, and KIRC, respectively.

To further evaluate and test the performance of the selected top100 transcripts, a further screening against random combinations of transcripts was indispensable. An additional analysis using increasing amounts of random transcripts was performed, using between 100 and 1000 different random transcripts. For each number for transcripts – beside 100 – 1000 models were trained based on random transcript combinations. For the number of 100 random transcripts, 10.000 different models were trained, to put more emphasis on this number of genes due to the intended use of the top100 transcripts.

As it can be seen, the mean but also the maximum and minimum testing accuracy increased when increasing the number of random transcripts to 200 and then stagnated around 93-94% in mean (min. 90.00-93.00%, max. 95.00%). Additionally, the usage of only 100 random transcripts results in the lowest minimum testing accuracy being 15.38% worse than the worst result using all available transcripts (Table 8).

Number Random Genes	Minimum Testing Accuracy [%]	Maximum Testing Accuracy [%]	Mean Testing Accuracy [%]
100	80.20	92.23	87.81
200	88.82	93.48	91.55
300	90.59	94.62	92.75
400	91.52	94.87	93.33
500	92.13	95.24	93.69
600	92.21	95.55	93.96
700	92.54	95.43	94.14
800	92.92	95.70	94.31
900	93.00	95.64	94.41
1000	93.29	96.01	94.51
60,483	95.58	96.76	96.14

Table 8: Testing accuracy of random genes

Comparison of the minimum, maximum and mean testing accuracy for different random gene amounts.

When comparing the best random transcripts selection with the best performing random forest model based on all 60,483 possible transcripts, the minimal difference in testing accuracy is 0.75%, achieved with 1000 random transcripts. When looking at the maximum testing accuracy for all of the tested transcript amounts, there were only slight increments of about 0.3% between 300 to 1000 random transcripts with a plateau reached at 500 random transcripts, whereas the differences between 100, 200, and 300 random transcripts were above 1 percent point for each step.

5.1.5.2 Reducing the Number of CpG Sites

Due to the satisfactory results of the reduced RNA-sequencing gene set, a reduction of the methylation data became of interest, to get another insight into the comparison between the use of RNA-sequencing data and methylation data, also retrieved from TCGA. Analogous to the determination of the top100 transcripts from RNA-sequencing, the top500 CpG-sites were determined in the methylation analysis, which were most frequently in the top500 per RF model.

Compared to the RNA-sequencing results, the reduction of CpG-sites – in 2623 different models each – showed more influence on the testing accuracy.

The use of the top100 CpG-sites resulted in a reduction of the mean testing accuracy to 81.76% (min 79.63%, max. 84.03%), almost 12 percent points worse than using all CpG-sites. The mean testing accuracy increased to 91.57% (min 89.99%, max 92.98%) when using 500 CpG-sites (Table 9). Compared to the complete 485577 CpG-sites, this is only 0.001% of the original dataset with around 2.3% inferior results.

Amount CpG Sites	Mean Testing Accuracy [%]	Min Testing Accuracy [%]	Max Testing Accuracy [%]	STD Testing Accuracy [%]
100	81.76	79.63	84.03	0.61
200	85.69	83.56	87.42	0.58
300	88.91	86.70	90.53	0.50
400	89.93	88.31	91.50	0.47
500	91.57	89.99	92.98	0.43

Table 9: Testing accuracies using reduced CpG-sites

Reduction of the number of used CpG-sites for random forest model training with the resulting mean, minimum, maximum and standard deviation of the testing accuracy.

Furthermore, three feature selection processes were performed, which were automatically terminated when the testing accuracy fell below 92.5%, comparable to the RNA-sequencing approaches. Only 22 of the specified entities were used to get a first impression of the results.

The best testing accuracies were 97.89%, 97.86% and 97.95%. These were reached during the course of the feature selection process, i.e. after a reduction of the CpG-sites. It is noticeable that a steady improvement of the testing accuracy during the process takes place before the accuracy rapidly dropped at a certain point (Figure 14). The maximum testing accuracy was reached at different steps during the process (16734 of 20492 steps, 17998 of 20643 steps or after 19424 of 20645 steps), but always with the same number of remaining CpG-sites (3775 CpG Sites), which also completely overlapped between the individual approaches. Additionally, the number of CpG-sites that last exceeds the selected lower threshold of 92.5% – serving as criterion for termination of the process – was identical each time. For each of the three individual processes the number of remaining CpG-sites before dropping below the threshold of 92.5% was 18, again completely overlapping between the different approaches.

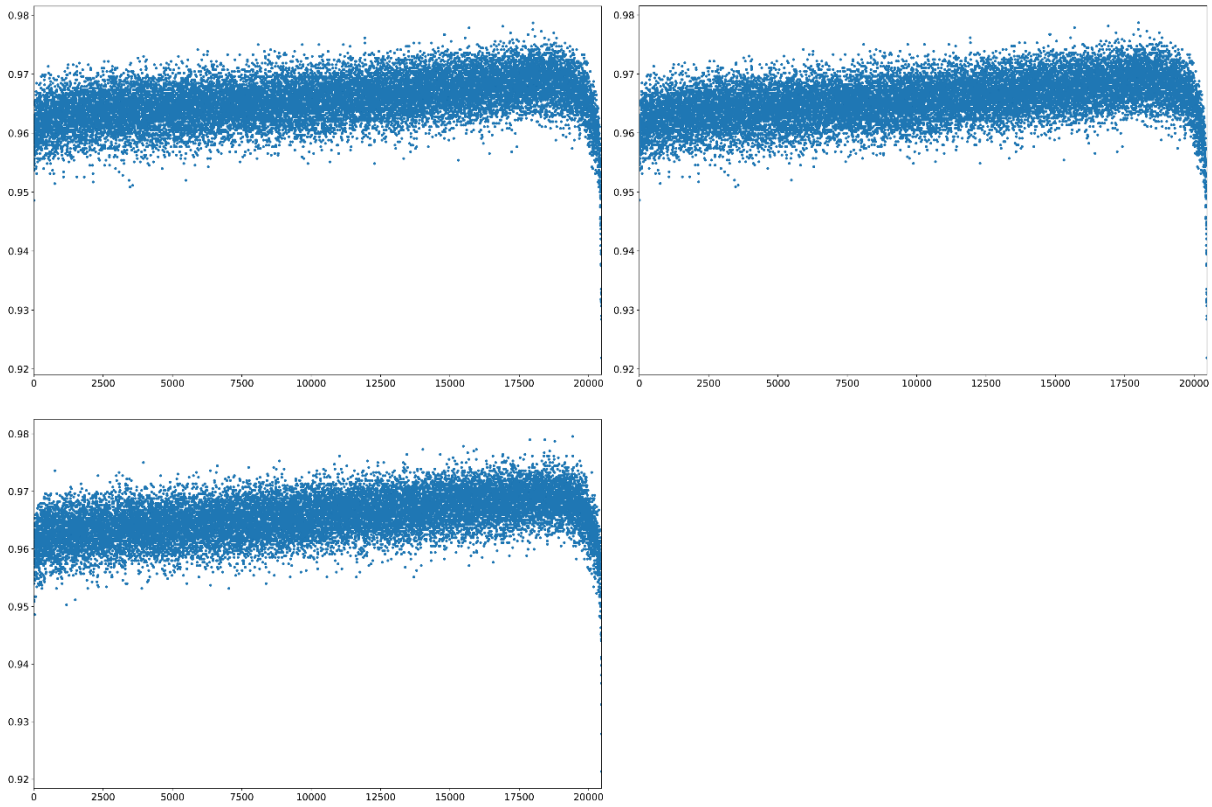


Figure 14: Scatterplot of feature selection process for CpG-sites

Scatterplot representation of the three individual performed feature selection processes on a reduced set of 22 entities based on CpG methylation sites. Each dot represents the testing accuracy of the according feature selection process step.

In summary, it can be stated that the reduction of CpG-sites lead to a certain increase in testing accuracy. In the three scenarios considered, the maximum testing accuracy was about 1% greater than the maximum compared to the use of all transcripts. However, it should be noted that the results are not fully comparable, since they used different approaches and slightly different data sets. Considering the intended usability in a real-life setup and the reduction of transcripts or CpG-sites, the RNA sequencing approach showed clear superiority of about 1% to 10% in testing accuracy over the methylation data.

In conclusion, the use of RNA sequencing data showed more reliable and interpretable results in the prediction of tumor entities than methylation data.

5.1.6 Harmonizing RNA-Sequencing Data

Despite convincing prediction results achieved with all genes as well as with a greatly reduced set of genes using RNA-sequencing data, the problem remains that this type of data from

different sources or laboratories can usually not be directly compared with each other. So far, there are already some ideas to solve this problem, such as the basic reprocessing of all raw data to adjust the bioinformatic analyses, which have a great influence on the final data or mathematical approaches to normalize the data (101–104). However, there still is the problem that the solutions are complex and difficult to understand, or even alter the data, which is the reason why no method has been able to establish itself as a standard in the field of RNA-sequencing.

To introduce a new approach for harmonizing RNA-sequencing data, the relationships (quotients) of all the beforehand mentioned top100 transcripts to each other (except for itself) were introduced, as the so-called quotient method. Subsequently, new RF models based on these newly created 9900 transcript relationships (quotients) were generated. The mean testing accuracy for the 20 models trained resulted in a mean testing accuracy of 93.80% (min. 93.52%, max. 94.66%), thereby confirming the results of the 100 transcripts alone by having the same mean testing accuracy.

5.1.6.1 VALIDATION OF RNA-SEQUENCING HARMONIZATION

To further test and evaluate the newly introduced approach for harmonizing RNA-sequencing data, datasets from ICGC and other laboratories depositing data in the Gene Expression Omnibus (GEO) were used. Further 18 datasets were added to the four TCGA cohorts, totaling in 22 evaluation datasets.

Of these evaluation datasets, eleven were taken from the ICGC database, four were derived from within TCGA and seven were gathered from different sources deposited in the GEO. Overall, these evaluation datasets consisted of 1999 samples. For 1580 (79.04%) of these 1999 samples, the correct entity was predicted by the best performing model based on the quotients of the previously determined top100 transcripts. For six datasets the model was able to predict all (100%) samples correctly. For further six datasets, the percentage of correct predicted samples were above 95%, with additional two datasets over 80% correct predictions. The last eight datasets had prediction accuracies below or near 60%. For the PRAD-FR dataset no correct predictions could be made, and for the PACA-CA dataset only 3.79% of the samples were correctly classified, being the worst among the evaluation datasets.

As the model only used a very specific set of transcripts, also with non-coding RNAs (Table 6) among them, the overlap between this set of transcripts and the provided transcripts of the evaluation datasets were also of interest. Across all 22 datasets, the mean overlap was 88.73

transcripts, ranging from 30 to 100. Excluding all the datasets provided by TCGA the mean overlap for the remaining 18 cohorts was 86.22 transcripts, again ranging from 30 to 100 (Table 10).

Cohort	**Cohort Entity**	**Approach**	**Amount Samples**	**Correctly Predicted**	**%**	**overlap top100 [%]**
PRAD-CA	Prostate Carcinoma	Quotients	144	144	100.00	100
GSE124535	Hepatocellular Carcinoma	Quotients	35	35	100.00	97
RECA-EU	Clear Cell Renal Cell Carcinoma	Quotients	91	91	100.00	96
GSE83533	Acute Myeloid Leukemia	Quotients	38	38	100.00	73
GSE135298	Breast Cancer	Quotients	93	93	100.00	68
LIRI-JP	Liver Cancer	Quotients	232	232	100.00	67
TCGA-READ	Rectal Carcinoma	Quotients	166	164	98.80	100
TCGA-KIRP	Papillary Renal Cell Carcinoma	Quotients	288	283	98.26	100
LICA-FR	Liver Cancer	Quotients	161	158	98.14	97
BRCA-KR	Breast Cancer	Quotients	50	49	98.00	65
TCGA-KICH	Chromophobe Renal Cell Carcinoma	Quotients	65	63	96.92	100
ORCA-IN	Oral Cancer	Quotients	40	38	95.00	77
GSE126975	Head and Neck Squamous Cell Carcinoma Cell lines	Quotients	43	38	88.37	97

GSE92914	Colon Carcinoma	Quotients	12	10	83.33	30
PACA-AU	Pancreas Carcinoma	Quotients	91	56	61.54	96
OV-AU	Ovarian Cancer	Quotients	93	49	52.69	96
PAEN-AU	Pancreas Carcinoma	Quotients	32	8	25.00	96
PACA-CA	Pancreas Carcinoma	Quotients	264	10	3.79	100
PRAD-FR	Prostate Carcinoma	Quotients	25	0	0,00	97

Table 10: Prediction accuracies for evaluation cohorts using quotient approach

All entities that were used to evaluate the performance of the introduced data harmonization applying the best performing random forest model based on the quotients of the calculated top100 transcripts. For each cohort, the correlating entity, the number of samples and correctly predicted number of samples with calculated percentage are displayed. Additionally, for each dataset the number of genes that were overlapping with the top100 transcripts are given. The cohorts are sorted descending based on the percentage of correctly predicted samples.

In this analysis, a high number of overlapping transcripts was not a guarantee to get correct predictions, as it can be seen for PRAD-FR, or PACA-CA with 97 and 100 genes overlap and prediction accuracies of only 0% and 3.79%, respectively. On the other hand, there were also examples of datasets with only 67 or 68 overlapping transcripts and an accuracy of 100% (LIRI-JP and GSE135298).

To now verify the effect of the introduced transformation and harmonization of data from different sources on the visual clustering level, t-SNE plots and UMAPs were used.

Due to the large number of different cohorts and datasets – and the resulting number of different samples – a manual selection of datasets was performed. The cohorts included four different entities (breast cancer, prostate cancer, colon cancer, and renal cell carcinoma) from all three databases (TCGA, GEO, and ICGC), totaling in 3970 samples from 18 cohorts. Although the cohorts were chosen manually, the selection itself aimed to represent different qualities in prediction accuracy.

To be able to use the unprocessed data from the different data sources in one approach, all samples had to be adjusted to a common base. Considering that all trained models were based on TCGA data, the genes present in the TCGA files served as basis for further adjustment. Specifically, for non TCGA samples, all genes not included in the data, but present in the TCGA, were set to 0 (not expressed). Genes that were not included in the TCGA data were omitted.

The results of the t-SNE plot, based on the unprocessed and adjusted data showed on the one hand certain datasets that were far distant to all other datasets and are represented in distinct clusters “alone”, whereas on the other hand, there was one cluster containing all other datasets. Even though there might be possible clusters visible inside this main cluster, the nature and calculation of the t-SNE plot did not allow to discriminate between certain distinct clusters inside it, as the distances inside a cluster are not globally comparable (Figure 15 A).

For the UMAP representation of the generated dataset, no obviously distinct clusters were visible. When looking closer, also taking into consideration that UMAP allows to interpret distances globally, several distinct clusters were present, for example the PRAD-CA dataset on the top middle of the plot, the TCGA-PRAD (Primary Tumor PRAD) cohort on the bottom of the plot, or the MBC-project dataset on the right side of the plot (Figure 15 B).

Both plots therefore showed no distinct clustering based on the underlying tissue or the data sources.

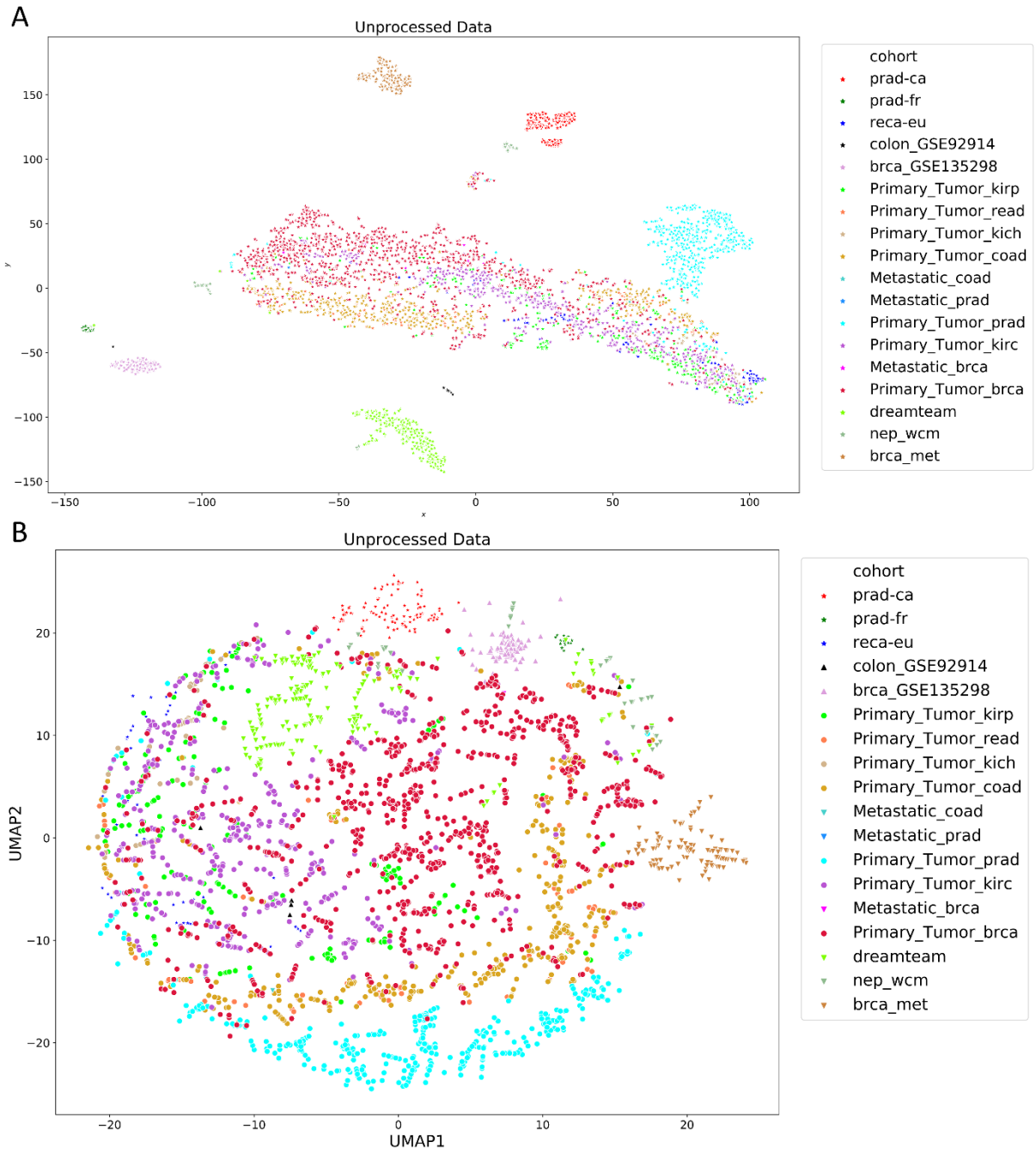


Figure 15: Unprocessed visual clustering of different tissues using t-SNE and UMAP

Visual cluster representation of the 18 manually selected cohorts consisting of 3970 samples of four different tissues based on the unprocessed FPKM values, using (A) t-SNE plot and (B) UMAP.

When considering the harmonized data based on the quotients of the top100 transcripts as identified by the RF models, the t-SNE plot (Figure 16 A) as well as the UMAP (Figure 16 B) did show certain clustering based on the underlying tissue. Even though not all samples cluster correctly or form distinct clusters when compared to the main tumor entity cluster (especially for breast cancer samples in the t-SNE plot), most of the samples are visually clustered together according to the expected tissue. Especially for the cohorts PRAD-CA and NEPC-WCM (Figure 16 red-circles), two prostate cancer cohorts, it is noticeable that the samples of

these cohorts are co-clustering for both approaches but are very far distant located to the main prostate cluster. Taken together, the newly introduced harmonization method for RNA-sequencing data based on the quotients of the previously determined top100 genes reflected clustering according to the underlying entities, as shown by t-SNE and UMAP analysis. This is an improvement compared to the usage of the unprocessed RNA-sequencing data, which mainly clustered based on cohorts rather than tumor entities.

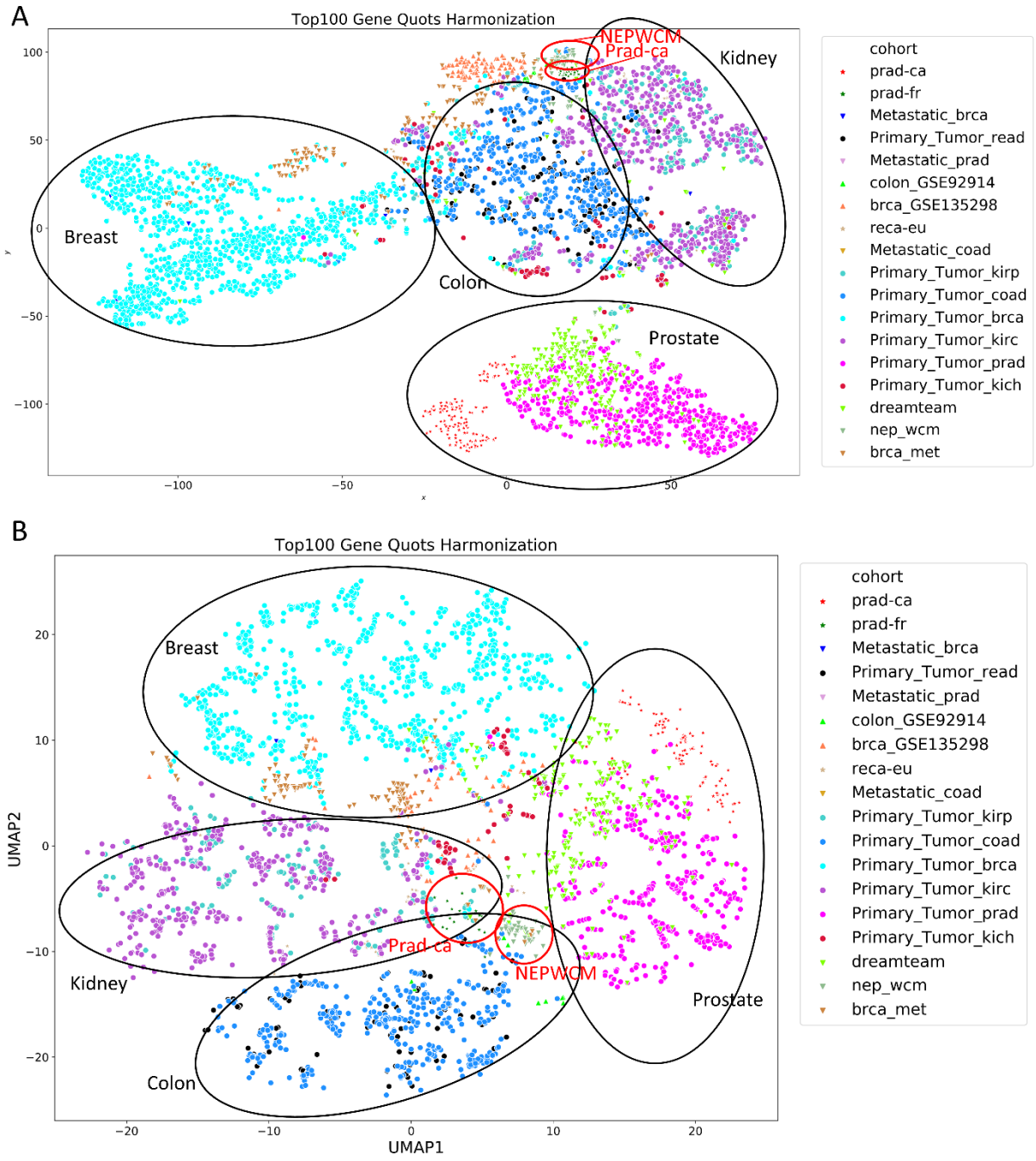


Figure 16: Harmonized visual clustering of different tissues using t-SNE and UMAP

Visual cluster representation of the 18 manually selected cohorts consisting of 3970 samples of four different tissues based on the top100 quotients harmonized data, using (A) t-SNE plot and (B) UMAP. Clusters based on underlying

entity are shown with black circles and are labelled accordingly. The two prostate cancer cohorts, that are not clustering accordingly – PRAD-CA and NEPC-WCM – are marked with red circles.

5.2 METASTASIS AND CANCERS OF UNKNOWN PRIMARY

In addition to the prediction of tumor entities, there are other important clinical questions in the field of malignancies that need to be addressed. One of these questions deals with metastatic tumors and as a special case the so-called Cancer of Unknown Primary (CUP) syndrome. For both scenarios, it is crucial to identify the primary tumor or the possible location of it. However, before the developed model could be applied to data from metastases, it had to be clarified whether there is a difference between metastatic samples at the transcriptional level depending on their resection-site, in order to exclude false predictions based on this. To test this, four different metastatic datasets containing three different primary sites with different local and distant resection sites were used.

5.2.1 VISUAL CLUSTERING OF RESECTION SITES

To approach the question how the resection site of metastatic tumors influences the transcriptomic features of a sample, the visual clustering approaches t-SNE and UMAP were used. Additionally, to get a more comprehensive analysis, log10 and log10+1 data transformations were used alongside the unprocessed FPKM values.

For the first dataset in this analysis, consisting of prostate cancer metastases (PRAD-SU2C, Dream Team), the t-SNE plot showed no clear cluster formation when using the unprocessed FPKM values (Figure 17 A), whereas it appears to show up to three different clusters, depending on the used log transformation (Figure 17 B and C). The results for the UMAP approach were very similar to the unprocessed data not being informative at all and the transformed data showing three clearly distinct clusters. It appears, that a certain bone and liver cluster are among the formed clusters, notably most visible within the UMAPs but also within the t-SNE plots. Taking the other samples into account, not all liver samples were clustering together, and the bone cluster only contained a small proportion of all bone samples.

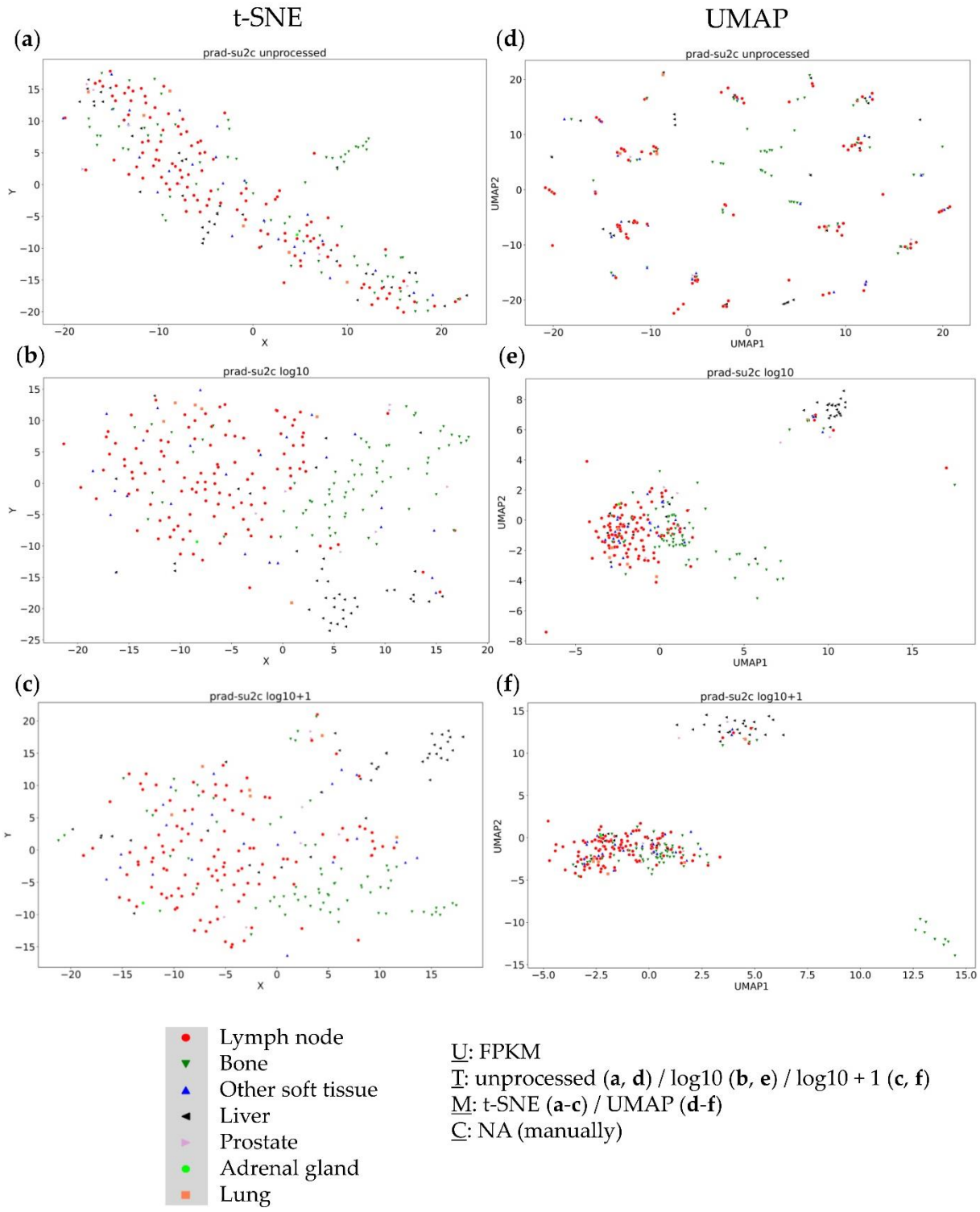


Figure 17: Different visual clustering approaches of the PRAD Su2C (Dream Team) dataset

Clusterings of the PRAD Su2C (Dream Team) dataset, consisting of metastatic prostate cancers, for t-SNE clustering approach using (A) unprocessed, (B) log10 transformed, and (C) log10+1 transformed FPKM values and for UMAP clustering approach using (D) unprocessed, (E) log10 transformed, and (F) log10+1 transformed FPKM values. Figure taken from (105).

The second dataset consisted of neuroendocrine prostate cancers (NEPC WCM), which does not show any clear cluster formation regardless of the used data when considering the t-SNE

plots (Figure 18 A to C). This statement remained the same, when looking at the results of the UMAP with unprocessed data (Figure 18 D). In contrast, the log10 (Figure 18 E) and log10+1 (Figure 18 F) transformed data showed three distinct clusters each. Compared to the Dream Team dataset, also containing metastatic prostate cancer samples, no bone or liver cluster, or any other cluster specific to resection site, was observable regardless of the approach or the underlying dataset.

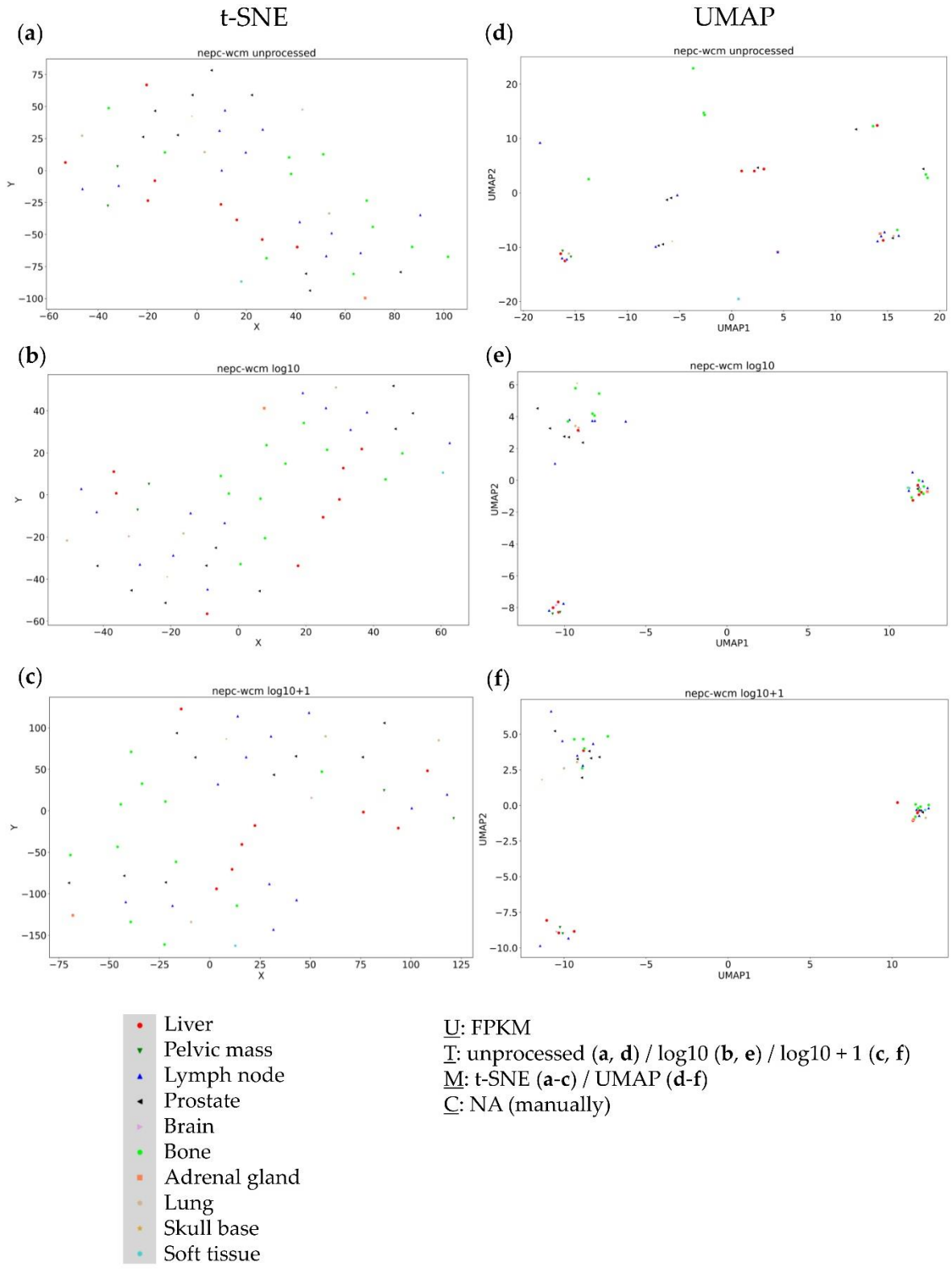


Figure 18: Different visual clustering approaches of the NEP WCM dataset

Clusterings of the NEPC WCM dataset, consisting of neuro endocrine metastatic prostate cancers, for t-SNE clustering approach using (A) unprocessed, (B) log10 transformed, and (C) log10+1 transformed FPKM values and for UMAP clustering approach using (D) unprocessed, (E) log10 transformed, and (F) log10+1 transformed FPKM values- Figure taken from (105).

The third dataset contains the metastatic samples of the TCGA-SKCM dataset. For this dataset, no clusters could be detected using t-SNE plots (Figure 19 A to C). When looking at the UMAPs for this dataset, the unprocessed FPKM values did not reveal any distinct clusters (Figure 19 D), whereas the log10 (Figure 19 E), and log10+1 (Figure 19 F) transformed datasets showed either a large cluster with only a few outliers or three different clusters, respectively. Looking at the UMAPs, no connection between cluster formation and resection site was evident.

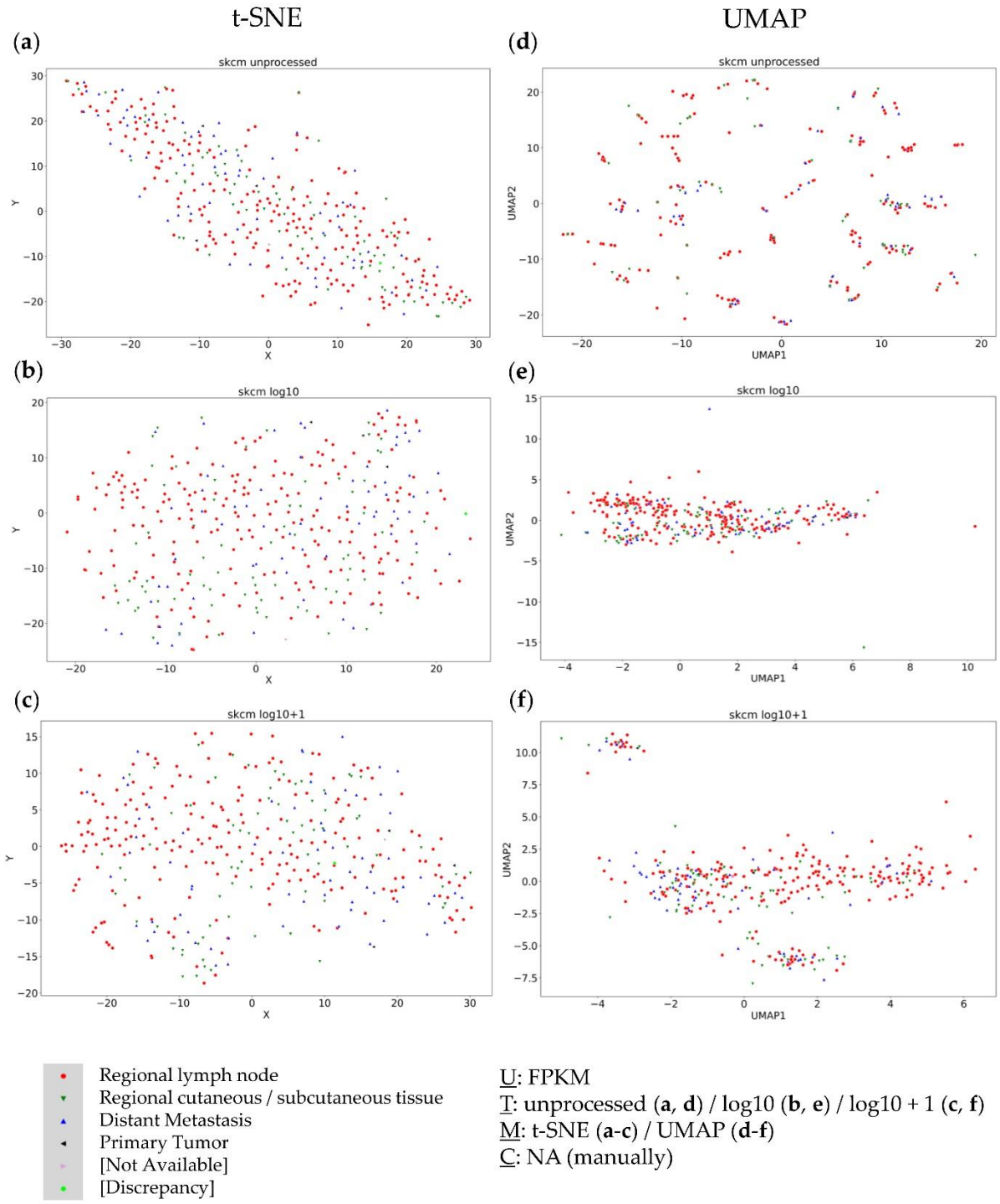


Figure 19: Different visual clustering approaches of the metastatic TCGA-SKCM dataset

Clustering of the metastatic samples of the TCGA-SKCM dataset, for t-SNE clustering approach using (A) unprocessed, (B) log10 transformed, and (C) log10+1 transformed FPKM values and for UMAP clustering approach using (D) unprocessed, (E) log10 transformed, and (F) log10+1 transformed FPKM values – Figure taken from (105).

In conclusion, the analysis of three different metastases datasets using visual clustering methods in combination with data transformations could not show any resection site dependencies. Yet, the influence of primary tumors onto respective metastases of the same entity – in combination – must be addressed.

To get insight into the transcriptomic relationship of primary tumors and respective metastases, the aforementioned methods and data transformations were used and applied to the complete TCGA-SKCM dataset – consisting of metastasis as well as primary tumors of skin cutaneous melanoma – and the dataset of the Metastatic Breast Cancer (MBC) project – containing primary and metastatic breast cancer samples.

The analysis of the complete TCGA-SKCM dataset showed no clustering based on the tumor status, and therefore no clustering based on resection site at all. Only in the UMAP log10+1 approach (Figure 20 F), a small cluster in the upper left corner of the plot was apparent, but without a homogenous clustering of one of the tumor status subgroups – either metastatic or primary. In general – regardless of approach chosen – there was a certain accumulation of primary (PM) tumors on one site, whereas the metastatic (MET) samples were located on the other side, even though no distinct cluster could be identified. Without previous knowledge about the tumor status, no conclusion could be drawn, indicating the transcriptional proximity of both, the PM and MET samples (Figure 20 A-F).

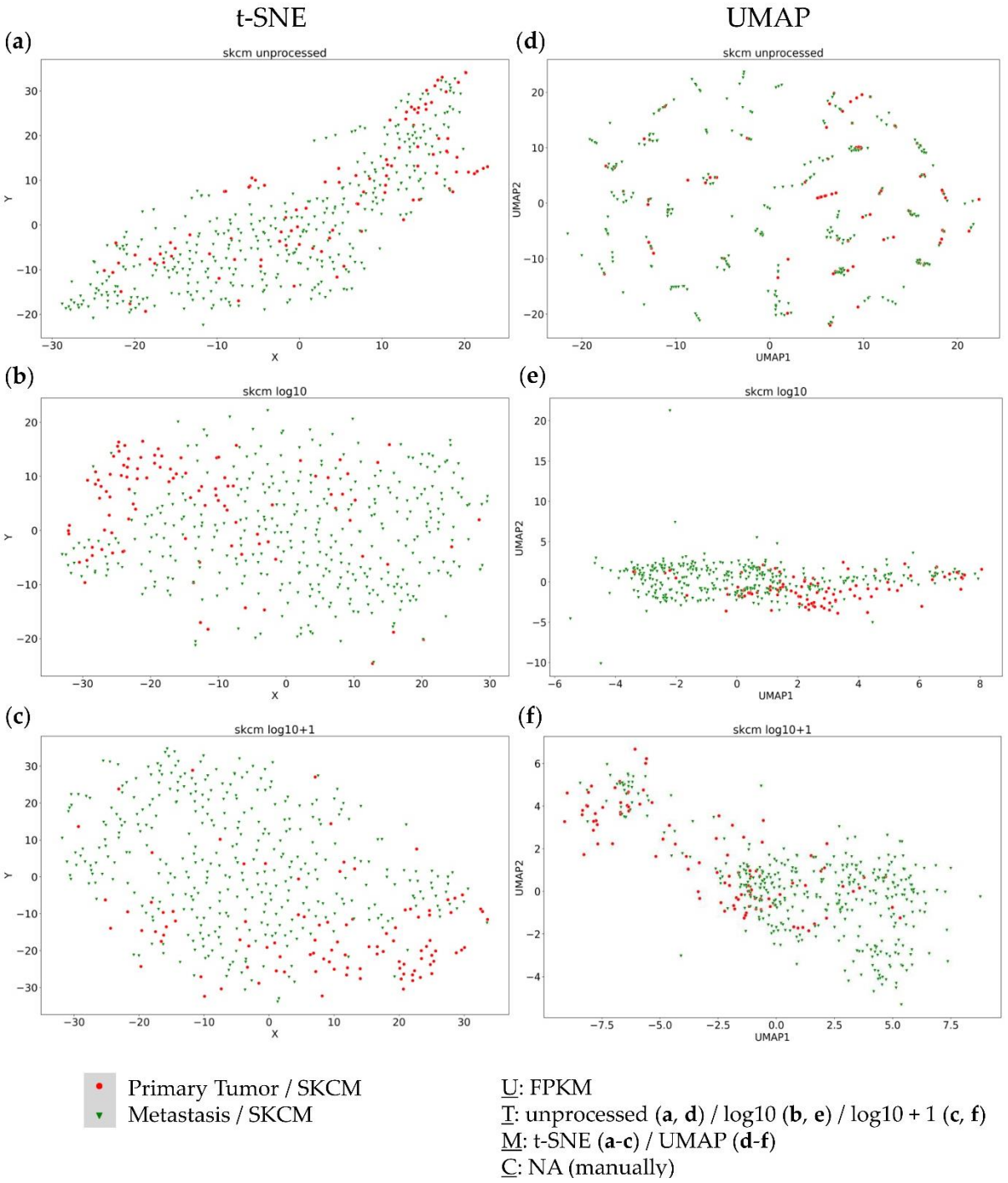


Figure 20: Different visual clustering approaches of the whole TCGA-SKCM dataset

Clustering of the whole TCGA-SKCM dataset including metastatic (MET) and primary (PM) tumor samples, for t-SNE clustering approach using (A) unprocessed, (B) log10 transformed, and (C) log10+1 transformed FPKM values and for UMAP clustering approach using (D) unprocessed, (E) log10 transformed, and (F) log10+1 transformed FPKM values -Figure taken from (105).

As a validation of the results of the complete TCGA-SKCM dataset, the MBC-project dataset was analyzed accordingly. Performing t-SNE analysis on this dataset did not result in any clusters (Figure 21 A, C, E). Unprocessed FPKM values in combination with UMAP did not give any insight into the clustering of the dataset (Figure 21 B). Logarithmic transformation of

the data did not form any distinct clusters in the UMAP approach (Figure 21 D+F), making this the only dataset with no cluster formations independent of the underlying methods or applied data transformation.

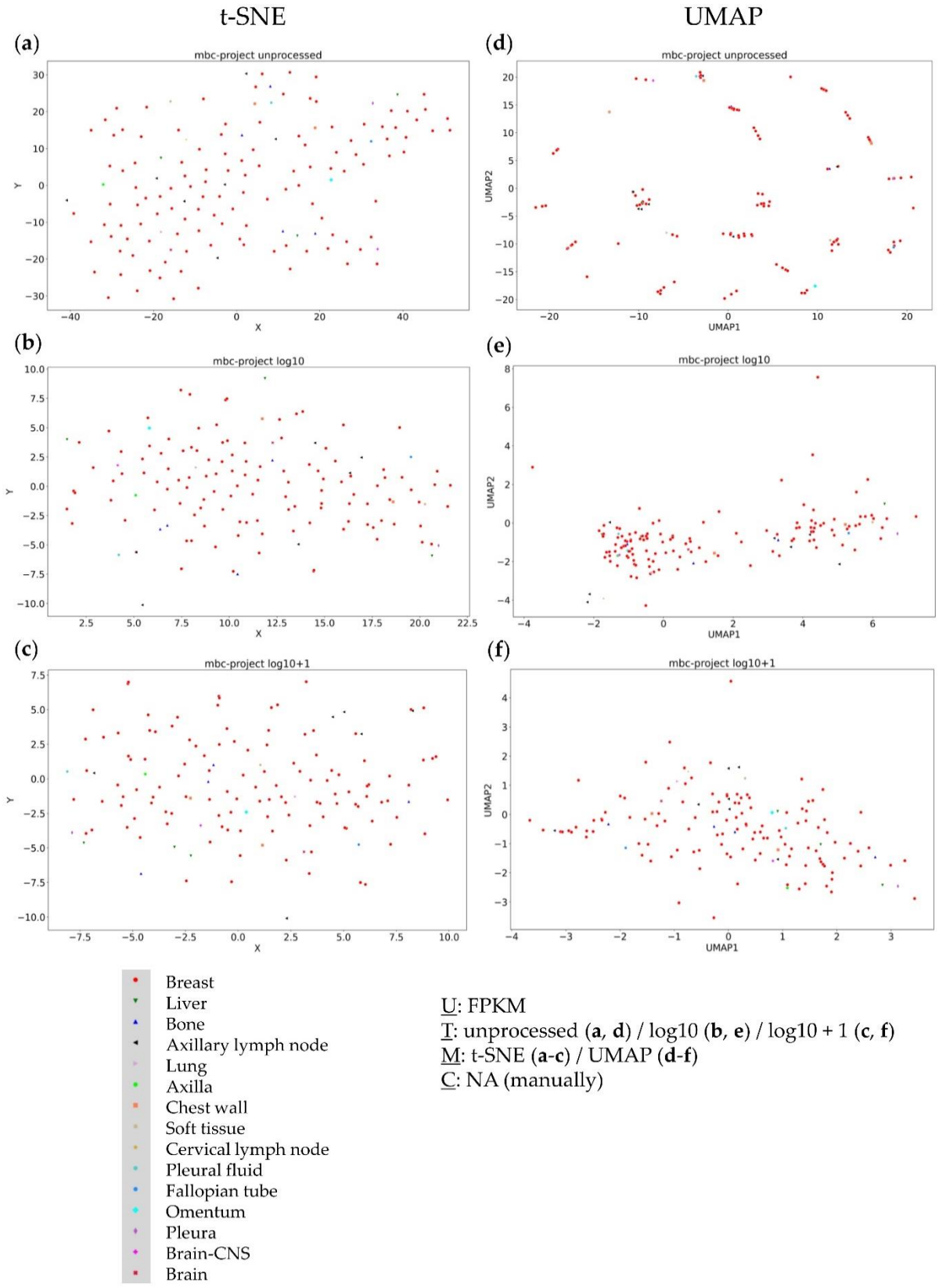


Figure 21: Different visual clustering approaches of the MBC-project dataset

Clusterings of the MBC-project (Metastatic Breast Cancer) dataset, consisting of metastatic breast cancers, for t-SNE clustering approach using (A) unprocessed, (B) log10 transformed, and (C) log10+1 transformed FPKM values and for UMAP clustering approach using (D) unprocessed, (E) log10 transformed, and (F) log10+1 transformed FPKM values – Figure taken from (105).

Taken together, there was no evidence of cluster formation for metastases based on resection site using two different data dimension reduction methods with three different data transformations, for all three considered different datasets. Furthermore, there was no evidence of distinct cluster formations in two additional datasets combining primary and metastatic tumors.

5.2.2 Predicting Tumor Entity Affiliation for Metastatic Samples

Since the previous performed analysis did not show any clear evidence for a resection-site dependent cluster formation based on the transcriptomic features, neither for metastatic nor for primary and metastatic disease combined, it was concluded, that the differences between --primary and metastatic tumors are only marginal regarding RNA-sequencing data. Based on this, the prediction of metastatic samples should be possible, independent of resection site, if the entity of the primary tumor is included in the trained RF model.

In a first analysis, only datasets of TCGA were considered, due to probable batch effects in other datasets. Furthermore, the analyzed top100 transcripts overlapped completely, allowing a comparison between the developed approaches: usage of a) all genes, b) top100 selected transcripts, and c) quotients of the selected top100 transcripts. For this analysis, TCGA offers an additionally 392 metastatic samples originating from ten different primary tumor sites, contained in the developed models. For most of these datasets – CESC, ESCA, PRAD, COAD, SARC, HNSC, and PAAD – a quality evaluation comparison could not be performed adequately, as the number of samples was very small, often only consisting of only one or two samples. Only three entities, namely SKCM, BRCA, and THCA contained more samples, with 362 of 368 (98.37%), 6 out of 7 (85.71%), and 8 out of 8 (100%) correct predictions, respectively. In combination, the best performing model using all genes was able to correctly predict 382 out of 392 (97.45%) of the metastatic TCGA samples (Table 11).

As before, evaluation of the top100 transcripts model was also necessary for the metastasis samples.

Compared to the model using all genes, there were only minor changes, such as the correct prediction of both metastatic HNSC samples (all genes correct = 1), the incorrect prediction of the one metastatic PAAD and ESCA sample (all genes correct = 1) and three fewer correct predictions for the metastatic SKCM samples (all genes correct = 362). For the other smaller cohorts, the predictions remained identical (Table 11), totaling in 378 correct predictions (378 of 382, 96.43%).

Due to the good results, the RF model trained on the quotients of the top100 transcripts was used in the next step for the prediction of the metastatic samples.

Compared to the usage of the RF model based on the top 100 genes, there were no differences in the number of correctly predicted samples in the metastatic TCGA cohorts (378 of 382, 96.43%), therefore showing equal results for both approaches, again.

Due to the newly introduced RNA-sequencing harmonization method, the prediction of three further metastatic datasets, not obtained from TCGA, was also possible. Two of the additional datasets consisted of 266 prostate cancer (Dream Team) samples, also including 49 neuroendocrine ones (NEPC-WCM), whereas the third one contained 146 breast cancer samples.

The addition of these three datasets extended the number of samples to 852 of which 736 (86.38%) got correctly classified according to their diagnosed primary tumor sites, with the percentage of correctly predicted samples for each dataset ranging from 0% to 100%. Of note, seven of the datasets only consisted of only one or two samples and the percentage therefore was not as meaningful as for other datasets. For the three datasets consisting of more than 100 samples, MBC-project, metastatic TCGA-SKCM, and Dream Team, the percentage of correct predictions were 93.84%, 97.82%, and 83.08%, respectively. The additional dataset NEPC-WCM was the only dataset with more than one sample for which no correct predictions could be made (Table 11). Comparing these results with the visual analysis done beforehand, the prediction accuracies for these datasets were in line with the visually observable results (Figure 16), showing the NEPC-WCM cohort far distant to the prostate cancer main cluster.

Regarding the top 100 transcripts, MBC-project overlapped with 79 transcripts, Dream Team overlapped with 68 transcripts, and NEPC-WCM overlapped with 75 transcripts, again showing that a higher overlap does not necessarily correlate with greater prediction accuracies (Table 11).

Cohort	**Entity**	**Nr. of Samples**	**Correctly Predicted (Accuracy [%])**			**overlap top100 [%]**
			All Genes	**Top100**	**Top100 Quotients**	
Metastatic TCGA-SKCM	Skin Cutaneous Melanoma	368	362 (98.37)	359 (97.82)	359 (97.82)	100
Metastatic TCGA-BRCA	Breast Invasive Carcinoma	7	6 (85.71)	6 (85.71)	6 (85.71)	100

Metastatic TCGA-CESC	Cervix Squamous Cell Carcinoma	2	1 (50.00)	1 (50.00)	1 (50.00)	100
Metastatic TCGA-ESCA	Esophageal Carcinoma	1	1 (100.00)	0 (0.00)	0 (0.00)	100
Metastatic TCGA-PRAD	Prostate Adenocarcinoma	1	1 (100.00)	1 (100.00)	1 (100.00)	100
Metastatic TCGA-COAD	Colon Adenocarcinoma	1	1 (100.00)	1 (100.00)	1 (100.00)	100
Metastatic TCGA-SARC	Sarcoma	1	0 (0.00	0 (0.00)	0 (0.00)	100
Metastatic TCGA-HNSC	Head and Neck Squamous Cell Carcinoma	2	1 (50.00)	2 (100.00)	2 (100.00)	100
Metastatic TCGA-THCA	Thyroid Carcinoma	8	8 (100.00)	8 (100.00)	8 (100.00)	100
Metastatic TCGA-PAAD	Pancreatic Adenocarcinoma	1	1 (100.00)	0 (0.00)	0 (0.00)	100
MBC-project	Breast Cancer	146	-	-	146 (93.84)	79
Dream Team	Prostate Cancer	266	-	-	221 (83.08)	68
NEP-WCM	Neuroendocrine Prostate Cancer	49	-	-	0 (0.00)	75

Table 11: Metastatic sample prediction accuracy comparison of the different developed models

Prediction of metastatic samples of primary tumor sites that are included in the trained random forest model using RNA-sequencing data. Next to the dataset correlating entity, the number of samples and the correct predictions separated according to approach used together with the calculated percentage of correct predictions are displayed. The last column contains the overlap of the in the dataset contained transcripts with the identified top100 transcripts.

5.3 Prediction of Entity-Specific Subgroups

All previous analyses confirmed the main assumption, that tumor entities can be reliably predicted based on RNA-sequencing data using a RF model. Furthermore, a novel harmonization method based on only 100 transcripts was introduced, enabling reliable prediction of tumor entities from different sources. Additional analysis based on visual clustering of metastatic samples indicated the possible application of the developed RF models also on metastatic samples. Applying the different developed models to metastatic samples again showed high prediction accuracies of usually above 80 or 90%. However, since survival of tumor patients is often not only determined by the original tumor entity, but also by the specific subgroup the original tumor is affiliated to, determination of the subgroup may be equally important, although this may depend on the tumor entity in question.

5.3.1 Clustering of the Renal Cell Carcinoma Dataset

After introducing RNA-sequencing data as suitable for entity prediction using RF models, it was necessary to clarify whether subgroups can also be identified with this data basis. To approach this question, visual clustering was tested as a subgroup determination tool, as specific distinct clustering could be observed within the considered metastatic datasets. For this testing purpose, the so-called TCGA-KIPAN dataset was used, consisting of the three largest histopathological subgroups of renal cell carcinoma (RCC) – clear cell (ccRCC), papillary (pRCC), and chromophobe (chRCC) RCC. The following sections will demonstrate different methods to obtain different results using dimensionality reduction methods, of the available high-dimensional gene expression data. Subsequently, the obtained clusters were further analyzed, characterized, and finally evaluated for one selected case using RF learning.

5.3.2 Clustering of the RCC Dataset Using UMAP

Adapted from the approaches for identifying clusters in metastases datasets, the first method to analyze the data was to use UMAP. Since the analysis of sequencing data is generally based on log transformed data, but so far, no evidence for the correctness for the exclusive use of log transformed data is available, both possibilities were considered here – with and without log transformation.

5.3.2.1 UMAP and Applied Log Transformation

The results using the log10 transformed KIPAN data utilizing UMAP (Figure 22) showed compact clusters for each entity, with some data points lying outside the histopathological subgroups. It was notable, that there were several ccRCC samples located within the chRCC cluster, comparable to the results obtained using t-SNE plotting.

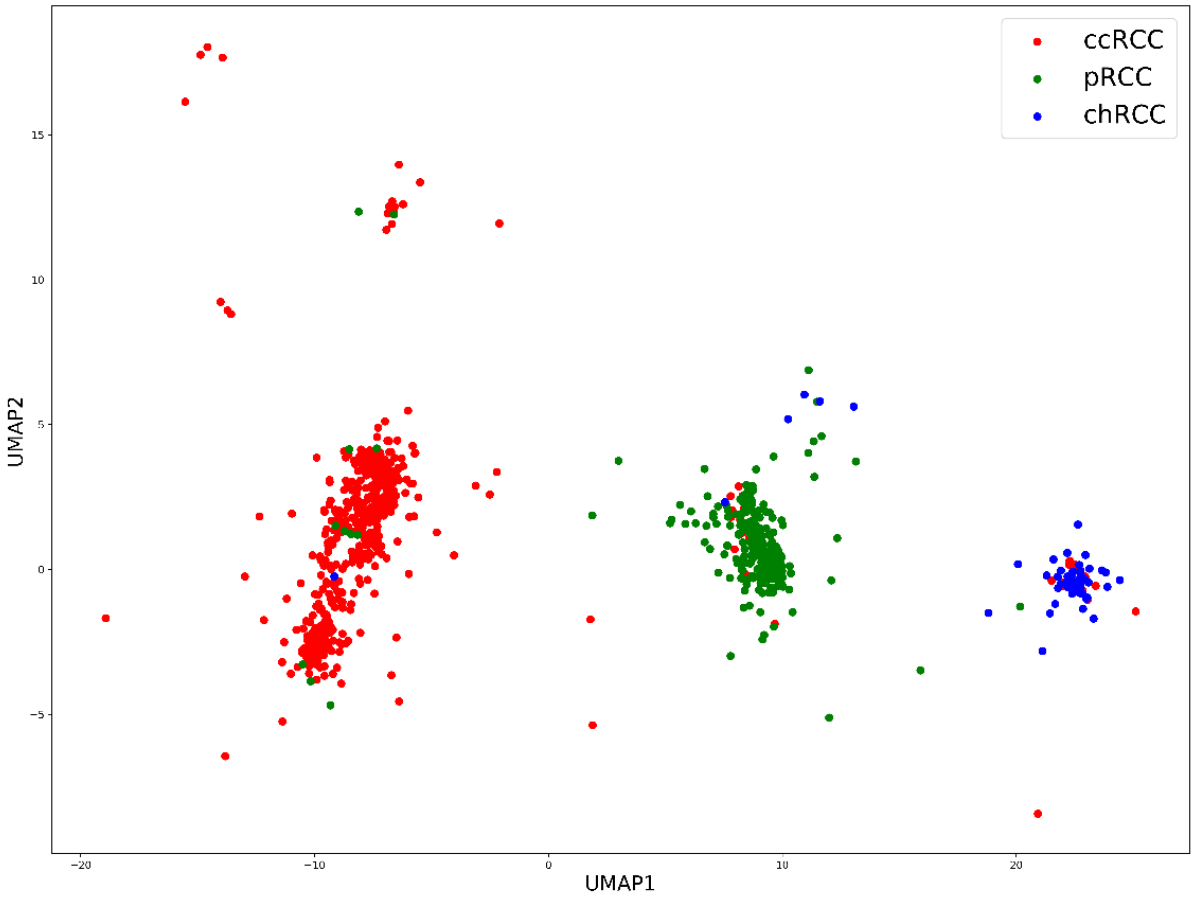


Figure 22: UMAP of log10 transformed TCGA-KIPAN dataset

UMAP for RNA-sequencing data of all three types of renal cell carcinoma (RCC) within the TCGA database using log10 transformed data. The different specimens are clear cell RCC (ccRCC – red), papillary RCC (pRCC – green) and chromophobe RCC (chRCC – blue).

Using log10+1 transformation (Figure 23), the clusters were again compact and represented the three major histopathologic subgroups, with the difference, that the only few notable outliers accumulated close to the chRCC cluster on the right side of the plot.

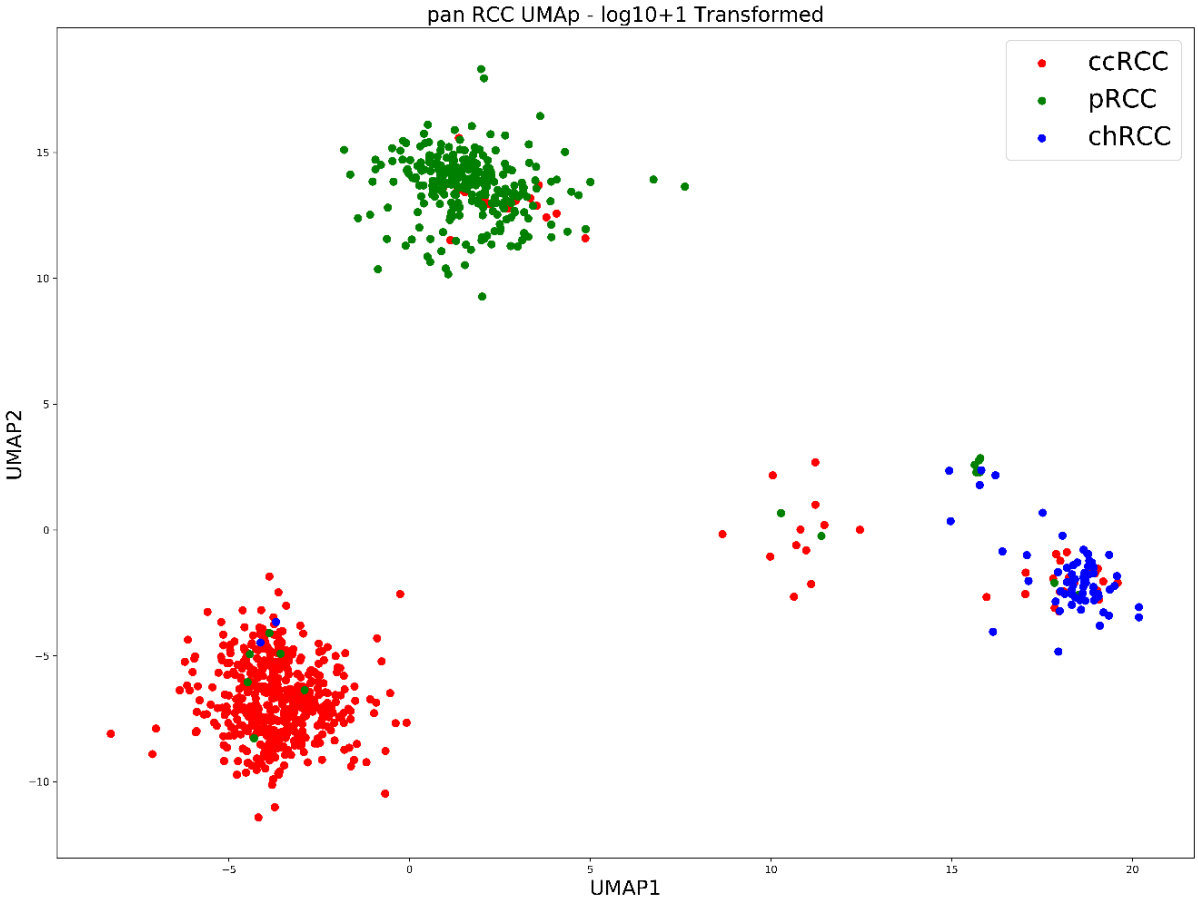


Figure 23: UMAP of log10+1 transformed TCGA-KIPAN dataset

Figure UMAP for RNA-sequencing data of all three types of renal cell carcinoma (RCC) within the TCGA database using log10 +1 transformed data. The different specimens are clear cell RCC (ccRCC – red), papillary RCC (pRCC – green) and chromophobe RCC (chRCC – blue).

5.3.2.2 UMAP and Untransformed Data

Due to the function of the UMAP, the unprocessed FPKM values did not lead to detectable distinct clusters (Figure 24). However, it appears that more ccRCC samples were located on the left side of the plot and the pRCC samples in the middle, with chRCC samples accumulating on the top right side.

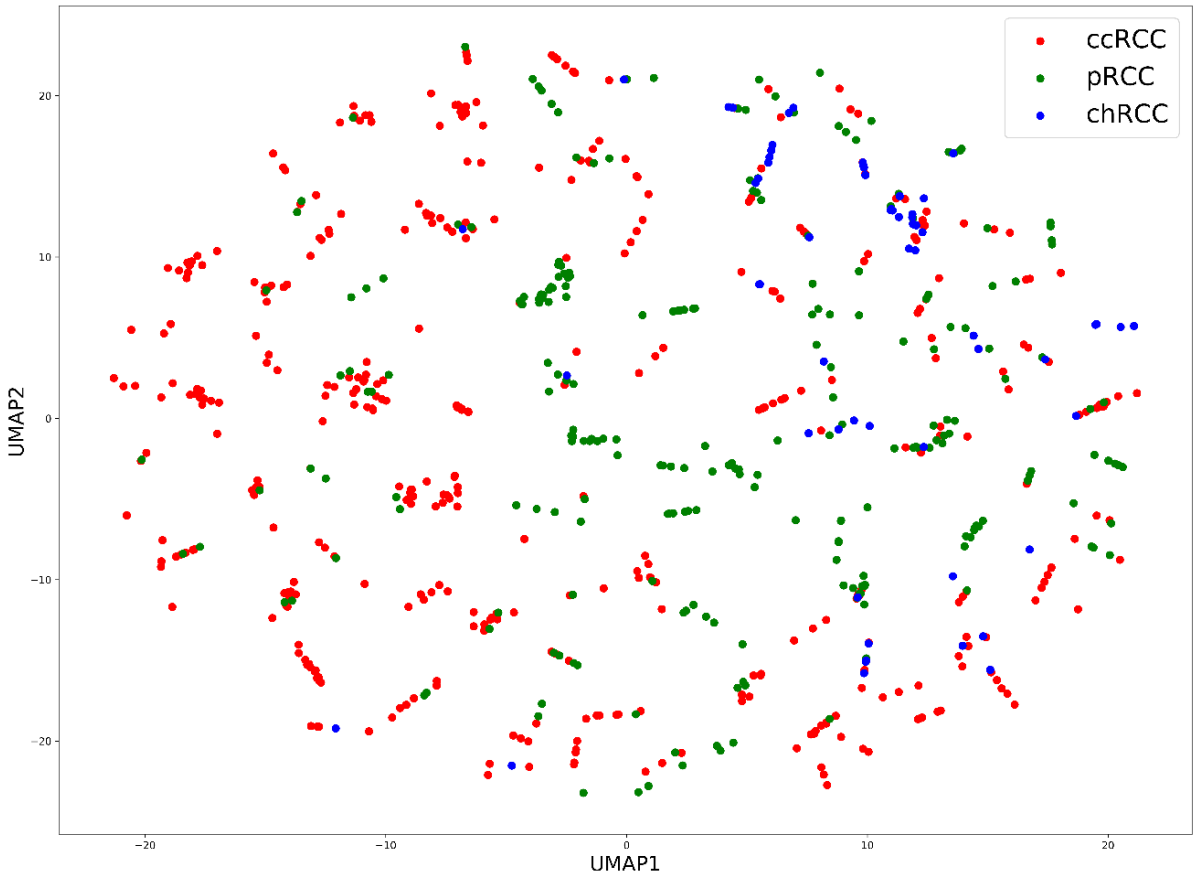


Figure 24: UMAP of unprocessed TCGA-KIPAN dataset

UMAP for RNA-sequencing data of all three types of renal cell carcinoma (RCC) within the TCGA database using unprocessed FPKM data. The different specimens are clear cell RCC (ccRCC – red), papillary RCC (pRCC – green) and chromophobe RCC (chRCC – blue).

Altogether, based on the analyses of the TCGA-KIPAN dataset by using UMAP as method of choice to visually describe clusters, it is apparent that histopathological subgroup clustering is largely observable when using log10 and log10+1 transformed data. In contrast, the untransformed data contained hardly any directly identifiable and distinct clusters.

5.3.3 Clustering of the RCC Dataset Using t-SNE Plotting

The second method that was used to analyze the data was the t-SNE plot, which has already been introduced and described in detail in the methods section and was also used beforehand to analyze the metastatic datasets. The basic idea of this method is to reduce the dimension of the input data by applying principal component analysis (PCA) in order to determine the distances between the data points based on their principal components in a subsequent machine learning step.

5.3.3.1 t-SNE and Applied Log Transformation

For the first evaluation of the used RCC transcriptome data, a logarithmic transformation to the base 10 (log10) of the FPKM values was used.

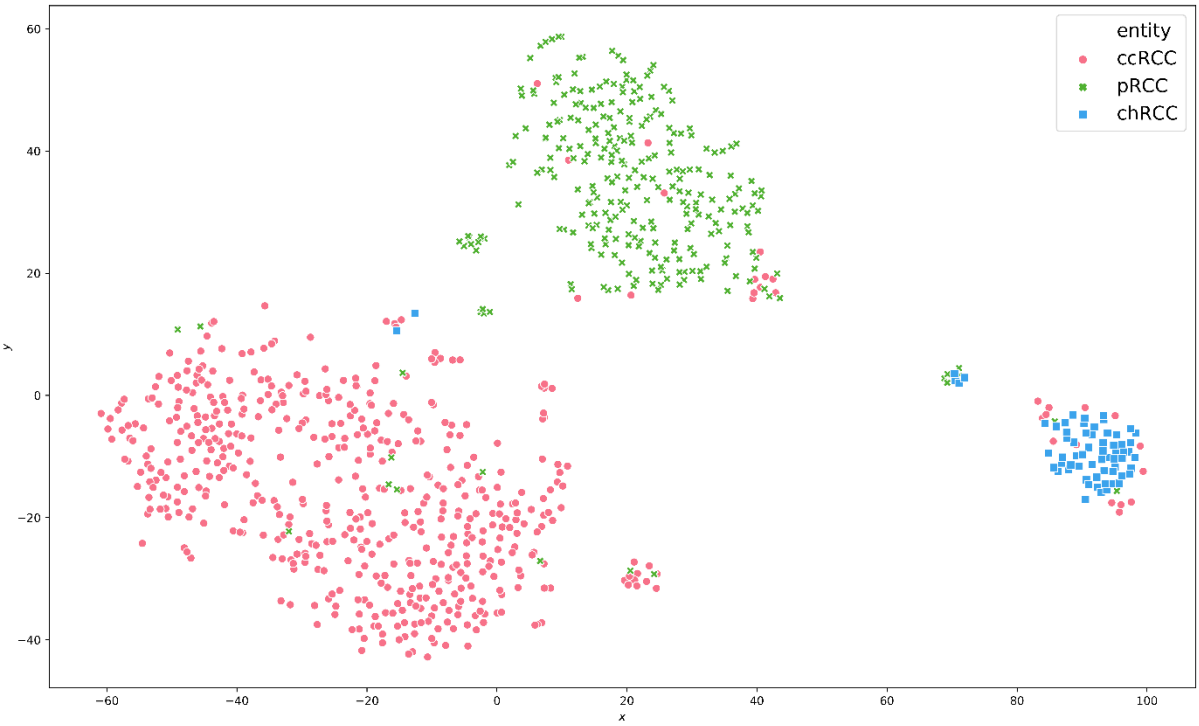


Figure 25: t-SNE plot of log10 transformed TCGA-KIPAN dataset

t-SNE plot for RNA-sequencing data of all three types or renal cell carcinoma (RCC) within the TCGA database using log10 transformed data. The different specimens are clear cell RCC (ccRCC – red), papillary RCC (pRCC – green) and chromophobe RCC (chRCC – blue).

The here used log10 transformation – *without* +1 – showed a clustering of the three entities mainly according to their histopathologic subgroup (Figure 25). However, there were some ccRCC (red) samples inside the chRCC (blue) cluster as well as a small cluster containing chRCC and pRCC (green) a little bit beside the main chRCC cluster. Similar small clusters were observable for the ccRCC and pRCC main cluster as well as some datapoints that were histopathologically not belonging to the assigned cluster.

In comparison, the log10+1 transformation – *adding* +1 to each expression value prior to logarithmization – also displayed a clustering according to the three histopathologic main subgroups (Figure 26). Again, several samples were not clustering according to their assigned histopathologic subgroup, with samples of the ccRCC subgroup clustering together with the chRCC cluster. Additionally, there were smaller subclusters for each histopathologic subgroup as well.

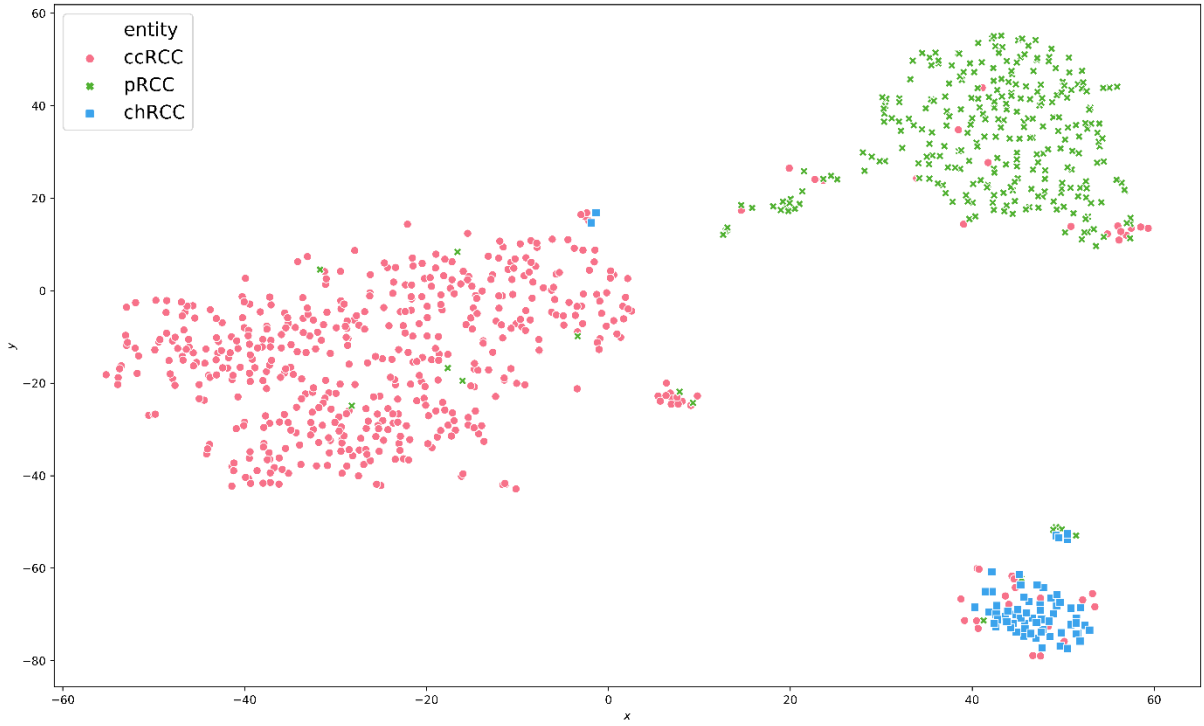


Figure 26: t-SNE plot of log10+1 transformed TCGA-KIPAN dataset

t-SNE plot for RNA-sequencing data of all three types or renal cell carcinoma (RCC) within the TCGA database using log10 +1 transformed data. The different specimens are clear cell RCC (ccRCC – red), papillary RCC (pRCC – green) and chromophobe RCC (chRCC – blue).

5.3.3.2 t-SNE and Untransformed Data

In addition to the analysis based on logarithmic transformed values, obtained FPKM values were also used in an unprocessed manner (Figure 27). Using this approach there were three different pRCC clusters surrounding a long-drawn ccRCC cluster. In the upper right corner, there was a prominent chRCC cluster, which was also mixed with ccRCC and pRCC samples.

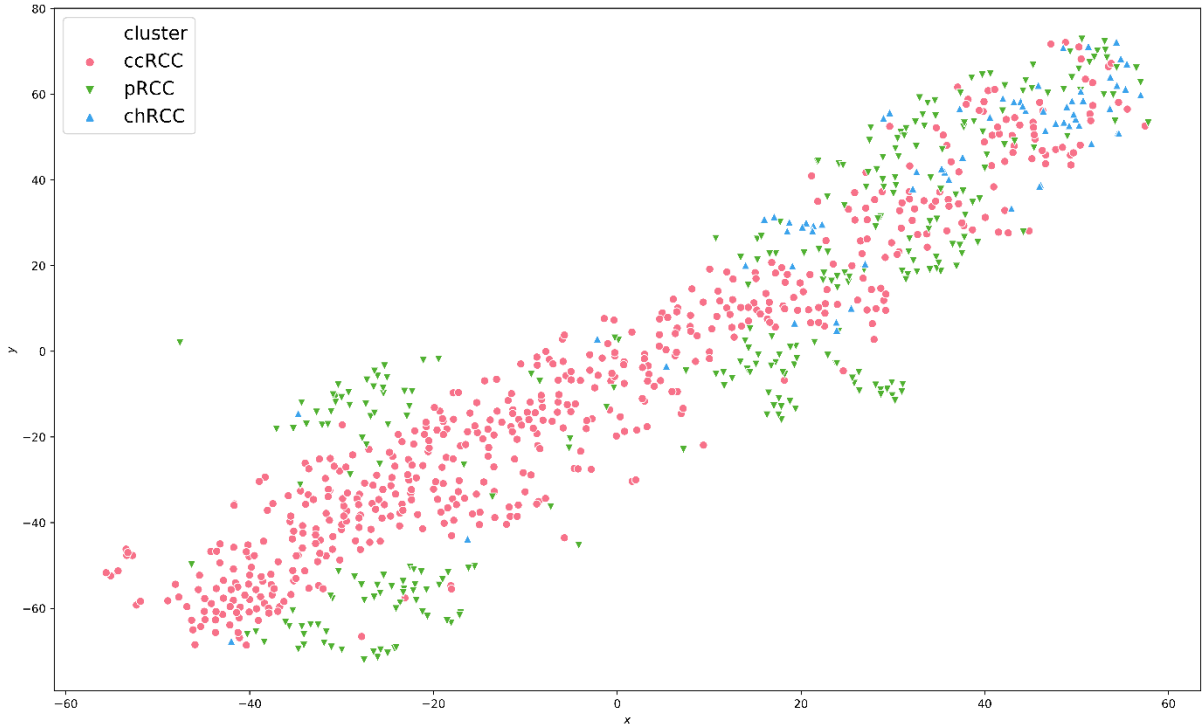


Figure 27: t-SNE plot of unprocessed TCGA-KIPAN dataset

t-SNE-plot for RNA-sequencing data from ccRCC (red), pRCC (green) and chRCC (blue) specimen within the TCGA database. Figure modified from (94).

Overall, analysis of the log10 and log10+1 transformed transcriptome of the TCGA-KIPAN data using t-SNE revealed a similar outcome to that observed when using UMAP: a separation mostly by histopathological subgroups. The use of the untransformed data was a certain exception here, since the t-SNE plot in particular shows a less distinct separation between the subgroups than the transformed data yet indicated a certain clustering for ccRCCs and pRCCs. The UMAP also showed similar results, but there were generally no distinct clusters, which is why this evaluation must be rated as inconclusive. Interestingly, the group of chRCCs furthermore showed an overlap with the other two subgroups in the t-SNE analysis. To determine if this observation is biologically relevant, additional analysis were conducted in the following.

5.3.4 Further Characterization of t-SNE Plot Without Log Transformation Results

As mentioned at the beginning of this chapter, all further in-depth analyses are limited to one of the data reductions results – the approach of the t-SNE plot *without* log transformation. One of the main reasons for this decision was the fact, that it was the only approach that allowed further clustering of the data and did not entirely reflect the three histopathological groups of the data already known so far. Thus, this approach had the highest probability to give new

insights not only into the subgrouping of RCCs but also into the usage of transformed and untransformed data in visual clustering approaches. Furthermore, an unprocessed view of the data could be closest to biological reality since no further changes were made to the measured values.

5.3.4.1 Manual Annotation and Characterization of Clusters

Based on the results with unprocessed data in the t-SNE plot (Figure 27) clusters were manually introduced, based on histopathological subgroups, to demonstrate better identifiability for further analysis and characterization (Figure 28 A). Since no distinguishable chRCC cluster was identifiable, the upper right corner of the plot – containing over 80% of chRCC samples (Figure 28C) – was merged into one new subgroup, the so-called *mixed subgroup (*Figure 28 *B)*, as it also contained large proportions of the other two RCC subgroups.

The newly defined *mixed subgroup* consisted of 102 (18.96%) ccRCC, 106 (36.81%) pRCC, and 53 (81.54%) chRCC samples, for a total of 216 (29.29%) of all samples (Figure 28 C).

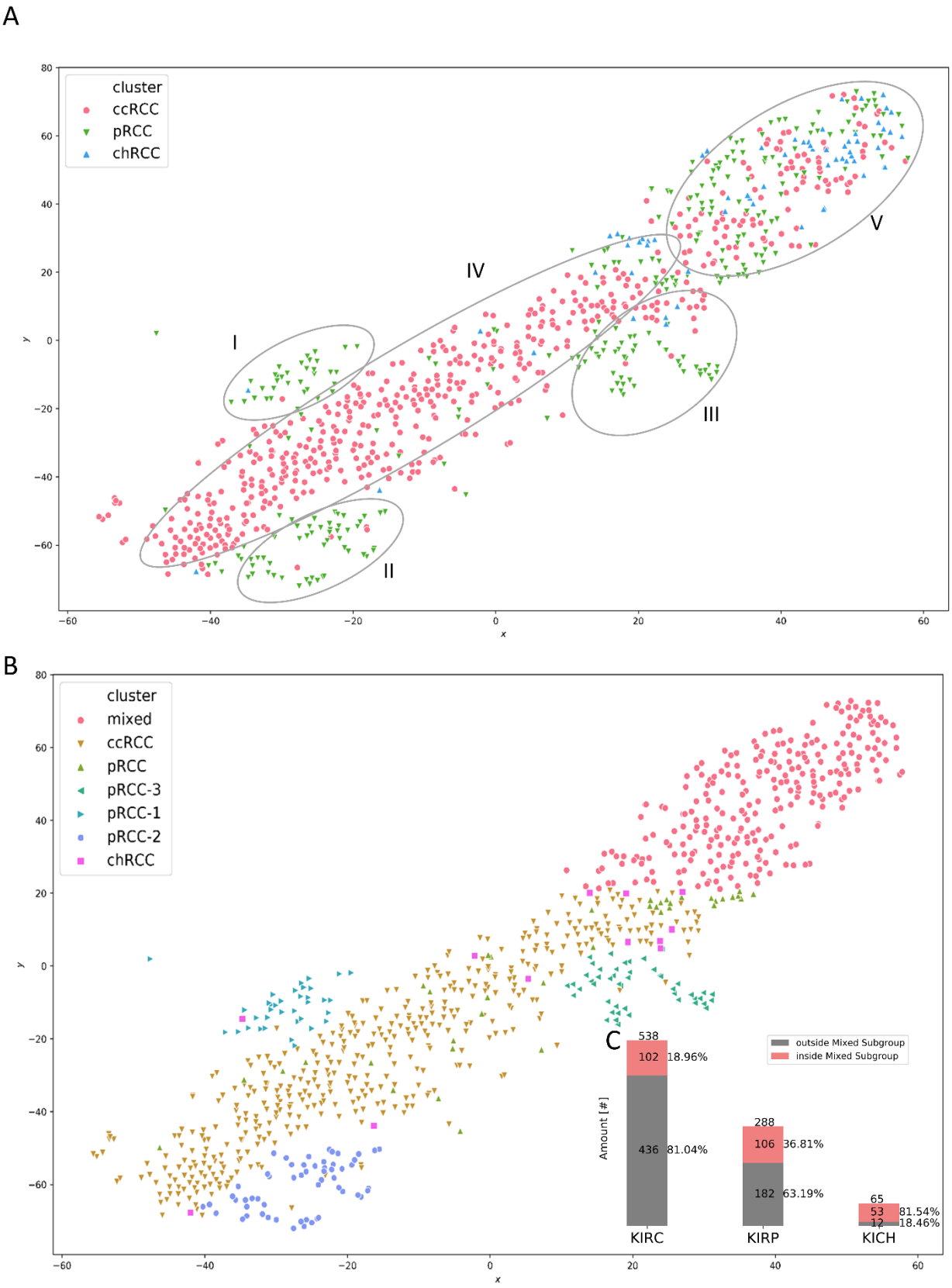


Figure 28: t-SNE plot analysis of unprocessed TCGA-KIPAN dataset

t-SNE-plot for RNA-sequencing data from all three RCC specimen (A) with circles and numbers indicating respective manually introduced clusters. Clusters I, II; and III represent papillary renal cell carcinoma (RCC) clusters, IV represents the clear cell RCC cluster, and V the newly introduced "mixed subgroup" containing all histopathological RCC subgroups. (B) Visually identified clusters with corresponding coloring for each cluster. (C)

Bar graph illustrating absolute numbers and the proportions of the RCC samples inside vs. outside the newly introduced mixed subgroup for each considered RCC subgroup. Figure modified from (94).

5.3.4.2 Clinical Characterization of Newly Defined Mixed Subgroup

To further confirm the obtained clusters using unprocessed FPKM values and t-SNE plotting for the TCGA-KIPAN dataset, statistical analyses including clinical characteristics were performed, to rule out their possible impact onto the clustering. The considered clinical parameters of interest were non tumor related like age and gender but also tumor related like the tumor stage, the grading (**G**) of the tumor (only available for ccRCCs) and different aspects of the tumor in question. These aspects are combined in the so-called TNM status, referring to the **T**umor, the lymph **N**odes, and the **M**etastases status of the tumor. Of note, the TNM status includes a numbering for each of the codes. For T the numbering ranges from 0 – no primary tumor detectable – to 4, indicating the increasing size or depth of penetration of the tumor. The Higher numbers of N status, ranging from 0 to 3, describe the amount and location of regional affected lymph nodes, with 0 indicating no infestation, and 1 to 3 indicating an increasing infestation of lymph nodes near the tumor. The M status is binary coded, where 0 indicates that no metastases are occurring, and 1 distant metastasis are present. The grading itself also ranges from 1 to 4 indicating different dimension of differentiation of the tumor. Here, 1 represents a well differentiated, 2 a moderately differentiated, 3 a poorly differentiated, and 4 a not differentiated tumor (https://www.uicc.org/resources/tnm). Accordingly, the clinical features of the histopathological groups lying inside of the *mixed subgroup* were compared with their counterparts lying outside the *mixed subgroup* (Table 12). Since the following analyses were focused on the comparison of samples inside vs. outside of the newly defined *mixed subgroup,* the different clusters of pRCCs were considered as one group and no further distinctions were made according to the manually defined clusters.

Using a Kruskal-Wallis test for $p < 0.05$, only five traits of the 19 considered ones were significant (Table 12). The one significant different characteristic for cRCC samples was the tumor grading, meaning that there could be a potential bias towards a G1/G2/G3 clustering, as the percentage of samples in the allocated groups is rather different. For pRCC samples the age and the gender are significantly different, therefore showing a potential bias based on age and the gender of the patient. For chRCC samples age and N status were significantly different between the two considered groups, again harbouring a potential bias based on age but also on the affected regional lymph nodes. All other traits like tumor stage T-status and M-status were not significant for any of the three histopathologic subgroups. Therefore, a systematic influence of the clinical parameters on the transcriptomic features – responsible for

the clustering result – could most likely be excluded, as especially the parameters important for tumors were not significantly different for all subgroups.

		ccRCC non-mixed n = 428	**ccRCC mixed** n = 102	**p-value**	**pRCC non-mixed** n = 182	**pRCC mixed** n = 105	**p-value**	**chRCC non-mixed** n = 12	**chRCC mixed** n = 53	**p-value**
Age	mean	60.2 ± 2.2	62.0 ± 11.8	0.196	59.6 ± 12.1	65.1 ± 10.9	**0.0002**	61.9 ± 12.9	49.6 ± 13.2	**0.012**
Gender	m	273 (63.79%)	71 (69.61%)	0.269	123 (67.58%)	89 (84.76%)	**0.001**	10 (83.33%)	29 (54.72%)	0.07
	f	155 (36.21%)	31 (30.39%)		59 (32.42%)	16 (15.24%)		2 (16.67%)	24 (45.28%)	
Tumor stage	I	223 (52.47%)	42 (64.29%)	0.166	108 (64,70%)	64 (38,32%)	0.419	2 (16.67%)	18 (33.96%)	0.1
	II	43 (10.12%)	14 (8.33%)		16 (9,60%)	4 (2,40%)		5 (41.67%)	20 (37.74%)	
	III	90 (21.18%)	33 (19.64%)		34 (20,30%)	16 (9,58%)		1 (8.33%)	13 (24.53%)	
	IV	69 (16.23%)	13 (7.74%)		9 (5,40%)	6 (3,60%)		4 (33.33%)	2 (3.77%)	
T	T1	228 (53.27%)	43 (42.57%)	0.092	119 (64.67%)	74 (70.48%)	0.463	2 (16.67%)	18 (33.962%)	0.14
	T2	53 (12.38%)	16 (15.84%)		22 (11.96%)	10 (9.52%)		5 (41.67%)	20 (37.74%)	
	T3	137 (32.00%)	41 (40.59%)		39 (21.20%)	20 (19.05%)		3 (25%)	15 (28.30%)	
	T4	10 (2.33%)	1 (0.99%)		4 (2.17%)	1 (0.95%)		2 (16.66%)	0 (0%)	
N	N0	192 (93.66%)	47 (94%)	0.929	29 (59.18%)	20 (71.43%)	0.21	4 (57.14%)	35 (94.60%)	**0.005**
	N1	13 (6.34%)	3 (6%)		16 (32.66%)	8 (28.57%)		2 (28.57%)	1 (2.7%)	
	N2	0 (0%)	0 (0%)		4 (8.16%)	0 (0%)		1 (14.29%)	1 (2.7%)	
M	M0	19 (90.48%)	3 (75%)	0.392	60 (63.16%)	35 (89.74%)	0.654	4 (80%)	3 (75%)	0.866
	M1	2 (9.52%)	1 (25%)		35 (36.84%)	4 (10.26%)		1 (20%)	1 (25%)	
Grading	G1	13 (3.06%)	13 (11.93%)	**0.014**						
	G2	195 (45.88%)	32 (29.36%)							
	G3	158 (37.18%)	48 (44.04%)							
	G4	59 (13.88%)	16 (14.67%)							

***Table 12*: Analysis of clinical characteristics of the TCGA-KIPAN dataset**

Clinical characteristics of RCC patients inside and outside the mixed subgroup. Except for age (mean ± standard deviation), all characteristics were presented as absolute values. p-values highlighted as bold were significant for $p < 0.05$. Table taken or adapted/modified from (94).

5.3.4.3 Random Forest-Based Transcriptional Analysis

After ruling out a systematical influence of clinical characteristics on the obtained clustering, an in-depth analysis of the manually annotated clusters could be performed.

To find out the differences between the annotated clusters and the newly defined *mixed subgroup,* a RF learning approach was used. For this analysis the RF model approach was slightly altered and 20 different models with a 70/30 split – 70% learning data, 30% evaluation data – with 1000 trees in the forest were trained. For this learning procedure pRCC samples that were not within one of the three pRCC clusters outside of the *mixed subgroup* were omitted.

With a testing accuracy of 92.06%, the best model outperformed the other 19 models, which had an average testing accuracy of 83.42% (min. 79.73%, max. 86.11%). A 10-fold cross-validation yielded a mean accuracy of 84.52% (± 4.58%), also showing the superior results of the best performing model. Compared with the results of tumor entity prediction shown previously, the mean results are relatively poor. One reason for this could be the manual classification of the clusters, which is not based on the histopathological groupings.

5.3.4.4 Differential Expression Analysis for Top Genes

Utilizing a RF learning model approach made it possible to check for the features that have the greatest impact on the generated model based on their feature values. When comparing the top 200 genes from the best performing model with the summed up most important features from the other 19 models, there was an overlap of only 92 genes (46%) (Supplementary Table 1).

To find out more about the top 200 genes with the highest influence in the top performing model, StringDB was used to identify possible interactions and clusters of identified genes (Figure 29). The network analysis was able to identify at least two big clusters of genes, that were indeed connected with each other. In red and blue, a significantly enriched cluster of mitochondrial genes present in the pathways “oxidative phosphorylation” (GO:0006119) and “respiratory electron transport chain” (GO:0022904) was identified. In yellow and green, StringDB revealed a significant accumulation of genes associated with “blood vessel development” GO:0001568) and “blood vessel morphogenesis” (GO:0048514), therefore representing an angiogenesis-related gene cluster.

Based on these StringDB results the genes represented in those two described clusters were further analyzed (Table 13). The top 10 of the identified top 200 genes in the best performing model are represented by 10 of the 13 mitochondrial genes obtained by the StringDB network analysis, with the top 3 genes being Mitochondrially Encoded Cytochrome B (MT-CYB – ENSG00000198727), Mitochondrially Encoded NADH:Ubiquinone Oxidoreductase Core Subunit 4 (MT-ND4 – ENSG00000198886) and Mitochondrially Encoded Cytochrome C Oxidase I (MT-CO1 – ENSG00000198804). Additionally, the other 3 mitochondrial genes not among the top 10, were ranked in the top 25 genes.

For the angiogenesis-related genes, TS Proto-Oncogene 1, Transcription Factor (ETS1 – ENSG00000134954) was the highest-ranking one in the RF model, placed at position 13, with all other genes ranking between 33 and 191. Only eight of the 15 angiogenesis-related genes were presented in the top 100 genes. Taken together, these results showed the high impact of mitochondrial genes onto the clustering and the RF learning beyond the well-known angiogenesis-related genes in ccRCC.

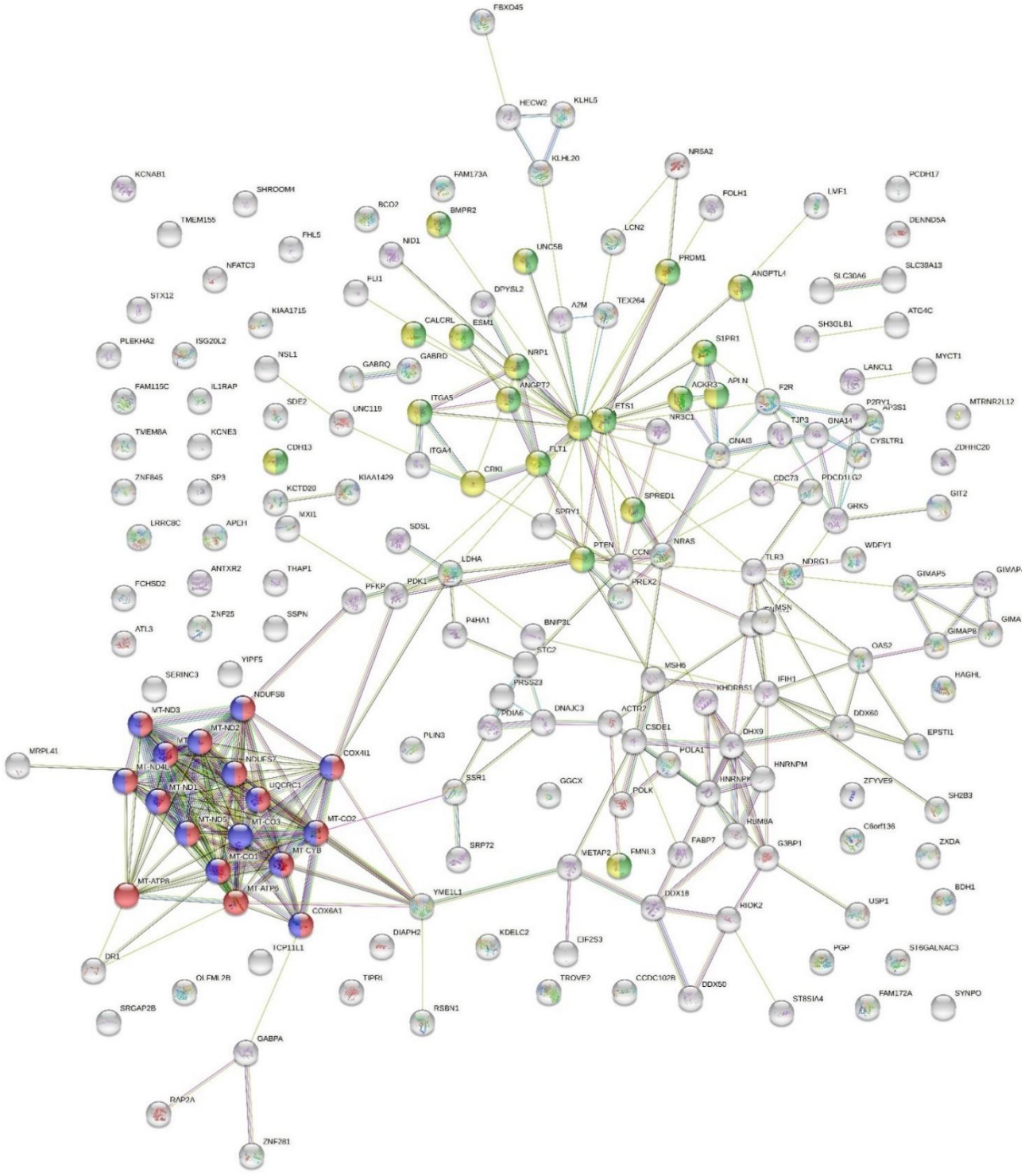


Figure 29: StringDB network analysis of random forest identified top200 genes

StringDB network of the top 200 genes identified as relevant classifiers for RCC sample clusters based on random forest learning. Genes affiliated with oxidative phosphorylation and respiratory electron transport chain are marked in red and blue, genes related to blood vessel morphogenesis and blood vessel development are marked in green and yellow. Figure taken from (94).

Mitochondrial Genes			**Angiogenesis-related Genes**		
HGNC Symbol	Ensembl gene ID	RF-Feature Position	HGNC Symbol	Ensembl gene ID	RF-Feature Position
MT-CYB	ENSG00000198727	1	**ETS1**	ENSG00000134954	13
MT-ND4	ENSG00000198886	2	**ANGPT2**	ENSG00000091879	33
MT-CO1	ENSG00000198804	3	**APLN**	ENSG00000171388	37
MT-CO3	ENSG00000198938	4	**FLT1**	ENSG00000102755	38
MT-CO2	ENSG00000198712	5	**CRKL**	ENSG00000099942	46
MT-ND4L	ENSG00000212907	6	**ITGA5**	ENSG00000161638	54
MT-ATP6	ENSG00000198899	7	**NRP1**	ENSG00000099250	56
MT-RNR1	ENSG00000211459	8	**PRDM1**	ENSG00000057657	93
MTATP6P1	ENSG00000248527	9	**PTEN**	ENSG00000171862	109
MT-ND1	ENSG00000198888	10	**VEGFA**	ENSG00000112715	112
MT-ND2	ENSG00000198763	20	**ACKR3**	ENSG00000144476	114
MT-ND3	ENSG00000198840	24	**CDH13**	ENSG00000140945	146
MT-RNR2	ENSG00000210082	25	**BMPR2**	ENSG00000204217	148
			CALCRL	ENSG00000064989	177
			ESM1	ENSG00000164283	191

Table 13: Selection of genes in clusters significantly overrepresented in the top200 genes

Gene families significantly overrepresented in the top 200 cluster classifying genes from Random Forest (RF) analysis. For each gene, HGNC symbol, Ensembl gene IDs, and the position in the RF model is shown. Table taken or adapted/modified from (94).

For further evaluation and verification, the gene expression of the individual groups – inside vs. outside of the *mixed subgroup* – of the identified mitochondrial and angiogenesis-related genes were investigated. A significant higher expression of the candidate mitochondrial genes – MT-CYB and MT-ND4 – was observed for all parts of the mixed subgroup from the three histopathological major subgroups compared to the clusters outside the respective *mixed subgroup* (Figure 30 A and B). Additionally, a differential expression for the different histopathologic subgroups could be observed when comparing inside vs. outside samples with normal tissue samples (Figure 30 C to E). For ccRCC (Supplementary Figure 1) and pRCC (Supplementary Figure 2), a significant lower expression of mitochondrial genes could be seen for the outside samples when compared to the normal tissue samples. Furthermore, a

significant higher expression for the comparison of inside vs. outside samples was present for all mitochondrial genes. This is also the case, when comparing normal tissue samples with the inside samples. In contrast, for chRCC (Supplementary Figure 3) samples, only small changes occurred for the comparison of samples outside the *mixed subgroup* vs. normal tissue. The significant higher expression for the comparisons of outside vs. inside *mixed subgroup* and the mixed group affiliated samples with the normal tissue was visible for chRCC as well.

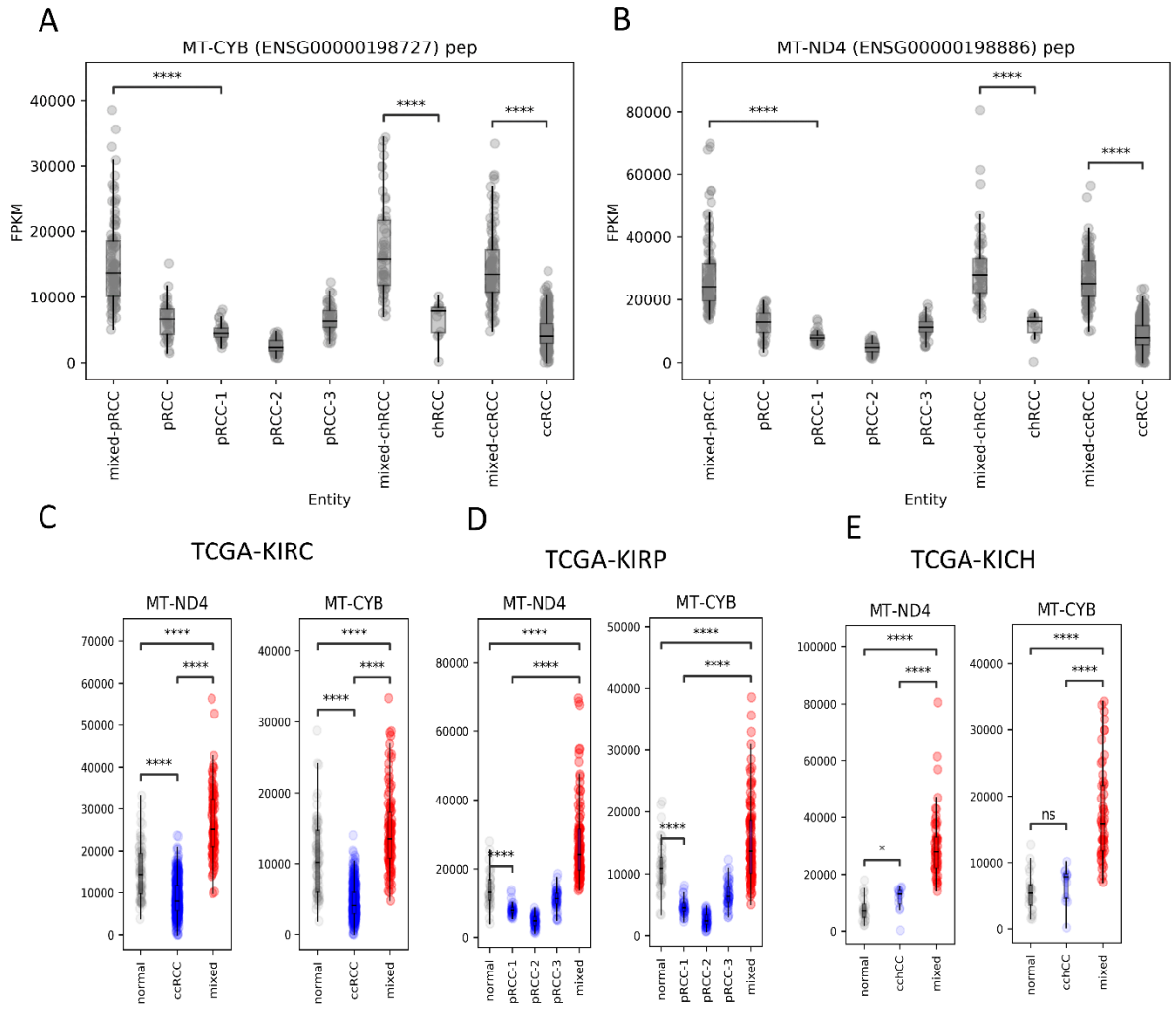


Figure 30: Expression comparison of unprocessed FPKM values for selected mitochondrial candidate genes

*(A) MT-CYB, (B) MT-ND4, as well as expression comparison to normal tissue samples of the candidate genes for (C) ccRCC, (D) pRCC and (E) chRCC samples. ns, not significant. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001. Figure adapted from (94).*

The results indicated that in addition to mitochondrial genes, responsible for oxidative phosphorylation, angiogenesis-related genes are also of importance. Upon closer examination, the differential expression of genes of the electron transport chain (ETC) could sindicate a change in the composition of the ETC, with lower expressions potentially occurring during hypoxia (106), also modifying metabolism of cancer cells (107). Additionally, it has been

shown, that tumor aggressiveness in RCCs correlates with a low mitochondrial respiratory chain content (108). These results could potentially indicate different oxygen conditions of the tumors.

Furthermore, highly significant changes for selected angiogenesis-related genes – FLT1 and ETS1 – occurred mainly within the ccRCC specimen. For the pRCC and chRCC subgroups, no high significances could be observed (Figure 31 A and B). When comparing to normal tissue samples, especially the ccRCC (Supplementary Figure 4) showed lower expression within the *mixed subgroup* samples. The pRCC (Supplementary Figure 5) and chRCC (Supplementary Figure 6) samples also showed a lower expression in this comparison for ETS1 (Figure 31 C to E). For the other identified angiogenesis-related genes, some did indeed show significant expression differences for the pRCC and chRCC as well – for example ETS1, PTEN or CRKL. These results again potentially point to the opposite metabolic effects of angiogenesis and oxidative phosphorylation, as a lack of angiogenesis could lead to hypoxia and therefore a potential alteration in the composition of the ETC.

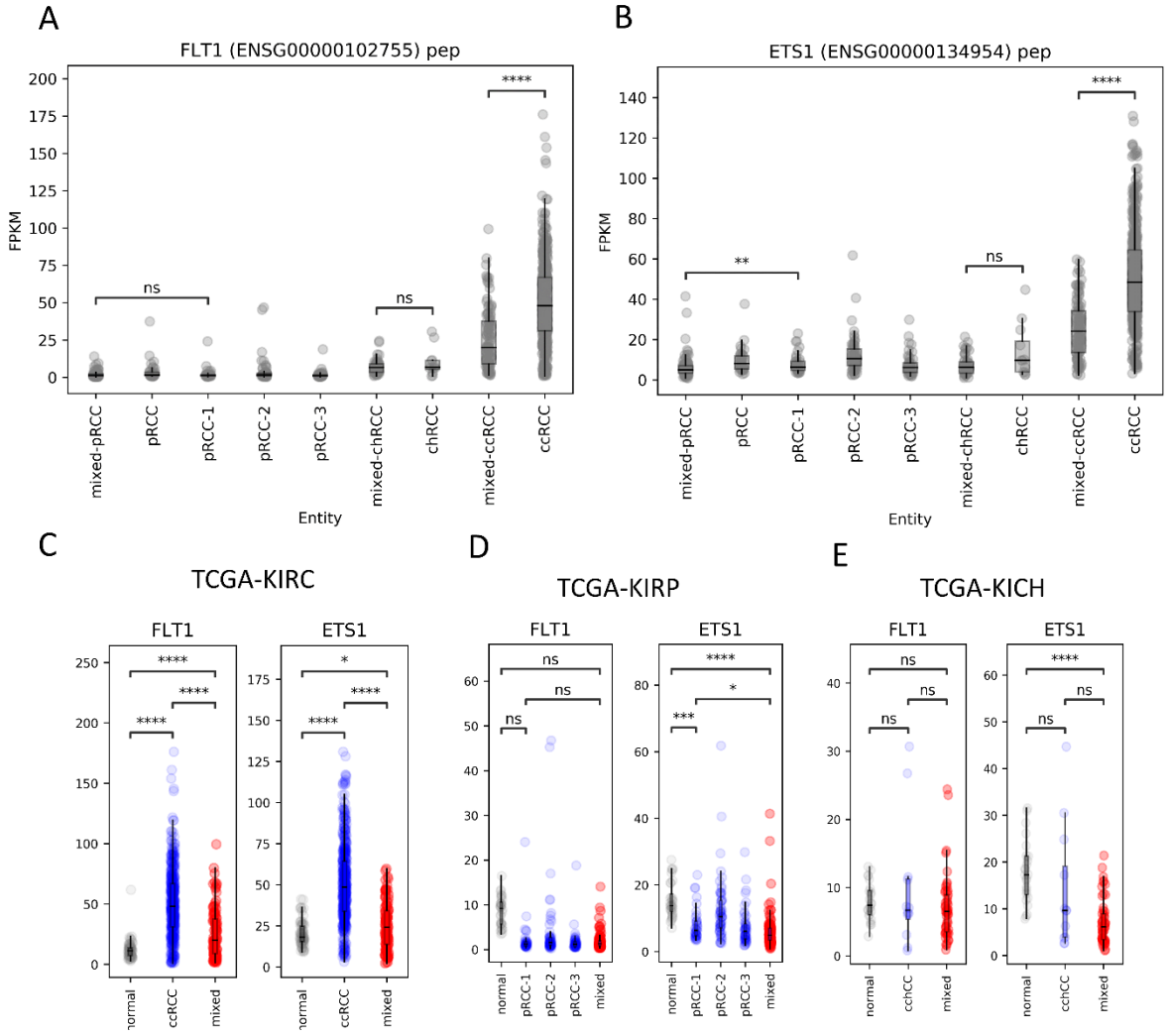


Figure 31: Expression comparison of unprocessed FPKM values for selected angiogenesis related candidate genes

*(A) MT-CYB, (B) MT-ND4, as well as expression comparison to normal tissue samples of the candidate genes for (C) ccRCC, (D) pRCC and (E) chRCC samples. ns, not significant. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001. Figure adapted from (94).*

Following these results, a possible correlation between the angiogenesis-related and the mitochondrial genes was of interest as they showed inverse expression patterns for at least the ccRCC samples. For this purpose, the sum of all mitochondrial and angiogenesis-related genes presented in Table 13 was calculated and subsequently normalized – between 0 and 1 – for all samples. Coloring all samples within the t-SNE plot according to their respective normalized expression showed an accumulation of high expressing mitochondrial samples in the upper right corner for all three considered specimens, gradually decreasing towards lower expression levels on the lower left corner (Figure 32 A). Differences in the normalized expression for the angiogenesis-related genes was only noticeable for the ccRCC samples – with a lower expression on the upper right corner, increasing to the lower left corner – whereas the pRCC contained only a few high expressing samples on the lower left corner and the

chRCC containing only a few samples highly expressing angiogenesis-related genes (Figure 32 B). To avoid a clustering biased by the high proportion of ccRCC samples within the analysis, a clustering of only ccRCC samples alone was performed as well, resulting in a mirrored t-SNE plot, comparable to the one before when looking into the normalized expression for mitochondrial and angiogenesis-related genes (Figure 32 C). These results led to the conclusion, that the clustering itself was not biased by the amount of ccRCC samples contained in the analysis.

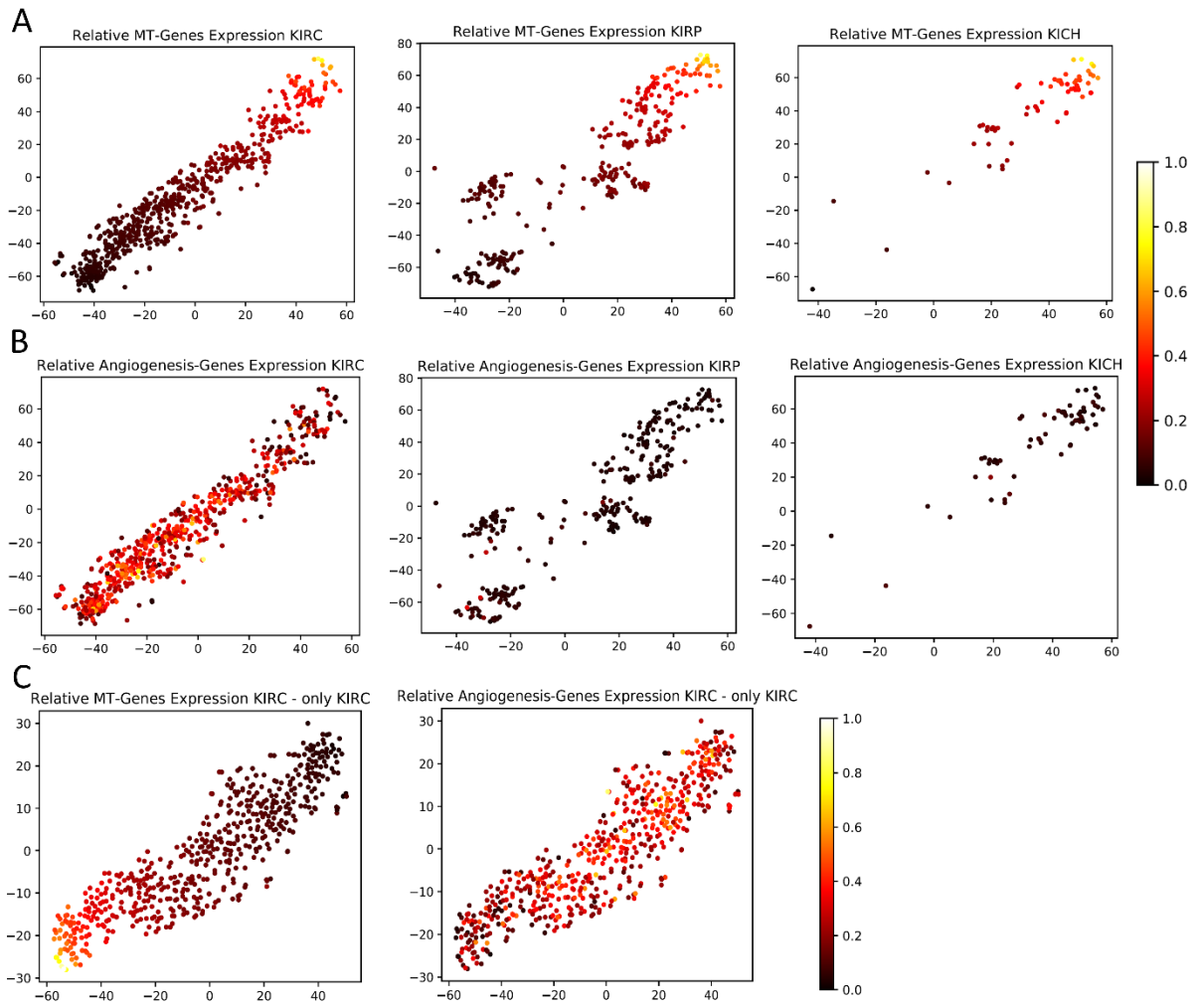


Figure 32: Normalized gene signature expression in TCGA-KIPAN cohorts

Normalized expression for the sum of (A) mitochondrial and (B) angiogenesis-related genes for the three considered cohorts KIRC (chRCC), KIRP (pRCC) and KICH (chRCC) within the t-SNE plot from the beginning of the analysis. (C) shows the normalized expression for both signatures when only considering the ccRCC cohort alone.

According to this conclusion, in-depth analyses of the correlation between mitochondrial and angiogenesis-related genes for all ccRCC samples were performed. The obtained results, utilizing the Pearson R for correlation measurement, confirmed the previous normalization analysis for this cohort (Figure 33 A) and confirmed the inverse correlation between the considered gene sets. To further validate this finding, the expression correlation of

mitochondrial and angiogenesis-related genes for three additional RCC cohorts – the ICGC-RECA-EU cohort (consisting of ccRCC samples), the CPTAC-3-Kidney cohort (consisting of not further specified RCC samples) and the GSE157256 cohort (consisting of fumarate hydratase-deficient RCC samples) – were calculated. All the considered validation cohorts showed an inverse correlation of mitochondrial and angiogenesis-related genes in the same manner as seen in the initial TCGA-KIRC cohort, even though, the ICGC RECA-EU dataset yields a slightly more negative Pearson R (Figure 33 B to D).

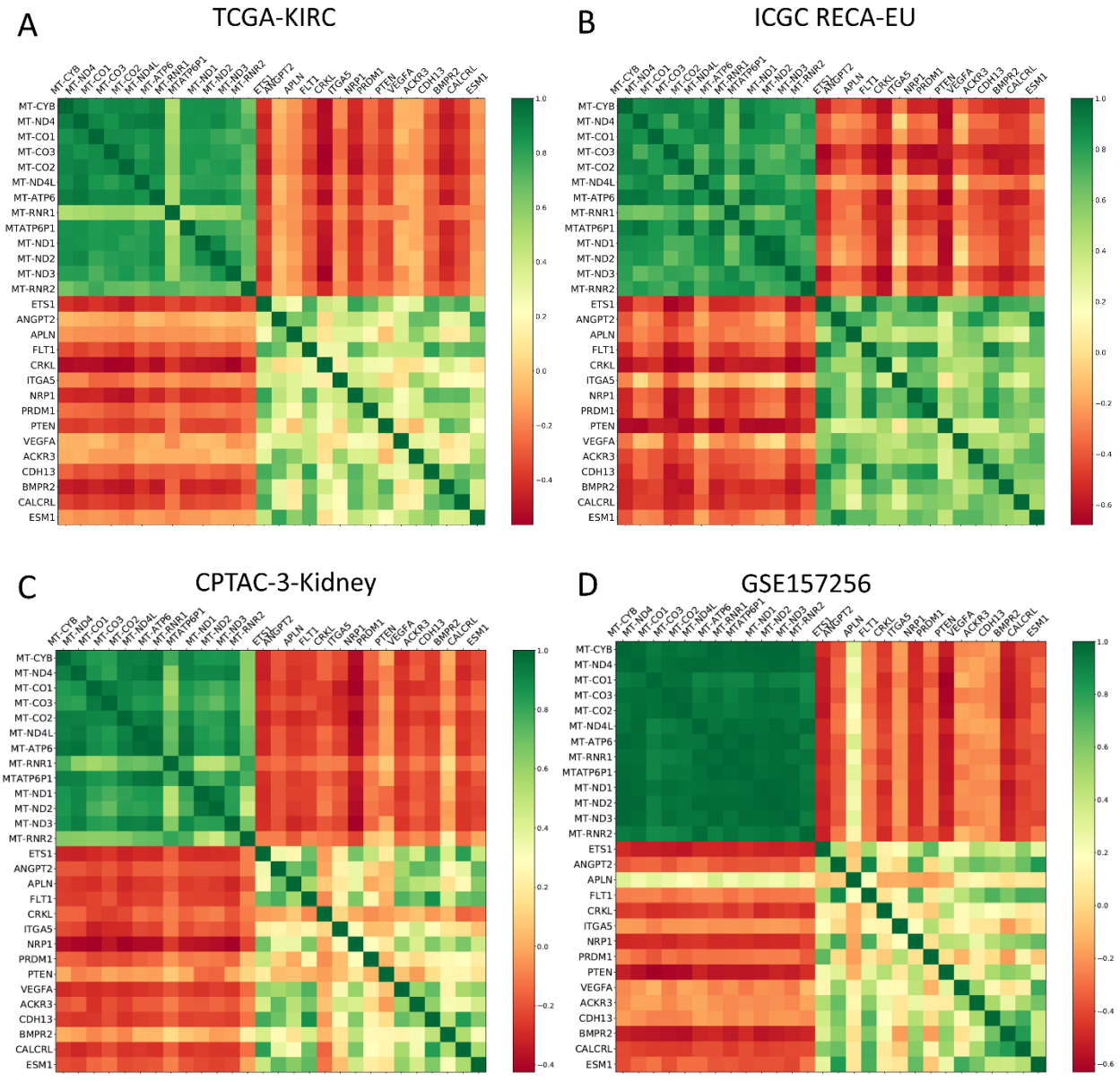


Figure 33: Pearson R correlation matrices for different renal cell carcinoma datasets

Color coded presentation of the Pearson R correlation matrix of the considered mitochondrial genes and angiogenesis associated genes for the (A) TCGA-KIRC cohort, the (B) ICGC RECA-EU cohort, the (C) CPTAC-3-Kidney cohort and the (D) GSE157256 cohort. Figure taken from (94).

5.3.4.5 Patient Survival Analysis for Mixed vs. Non-Mixed Subgroups

Following the significantly differential expression results for the genes obtained by the RF approach, mainly displayed in the two gene signatures – mitochondrial and angiogenesis-related genes – the question arose whether there is a clinical relevance, e.g. a survival difference. For this purpose, the Kaplan-Meier (KM) plots comparing the samples within the *mixed subgroup* with their counterparts not affiliated with the *mixed subgroup* for the different histopathologic RCC subgroups were analyzed.

For ccRCC (KIRC) patients survival was significantly reduced ($p<0.005$) in patients with tumors from the newly identified *mixed subgroup* (Figure 34 A), while the mixed subgroups was associated with improved survival in chRCC (KICH) patients (Figure 34 B). For pRCC patients, no significant were observed (Figure 34 C).

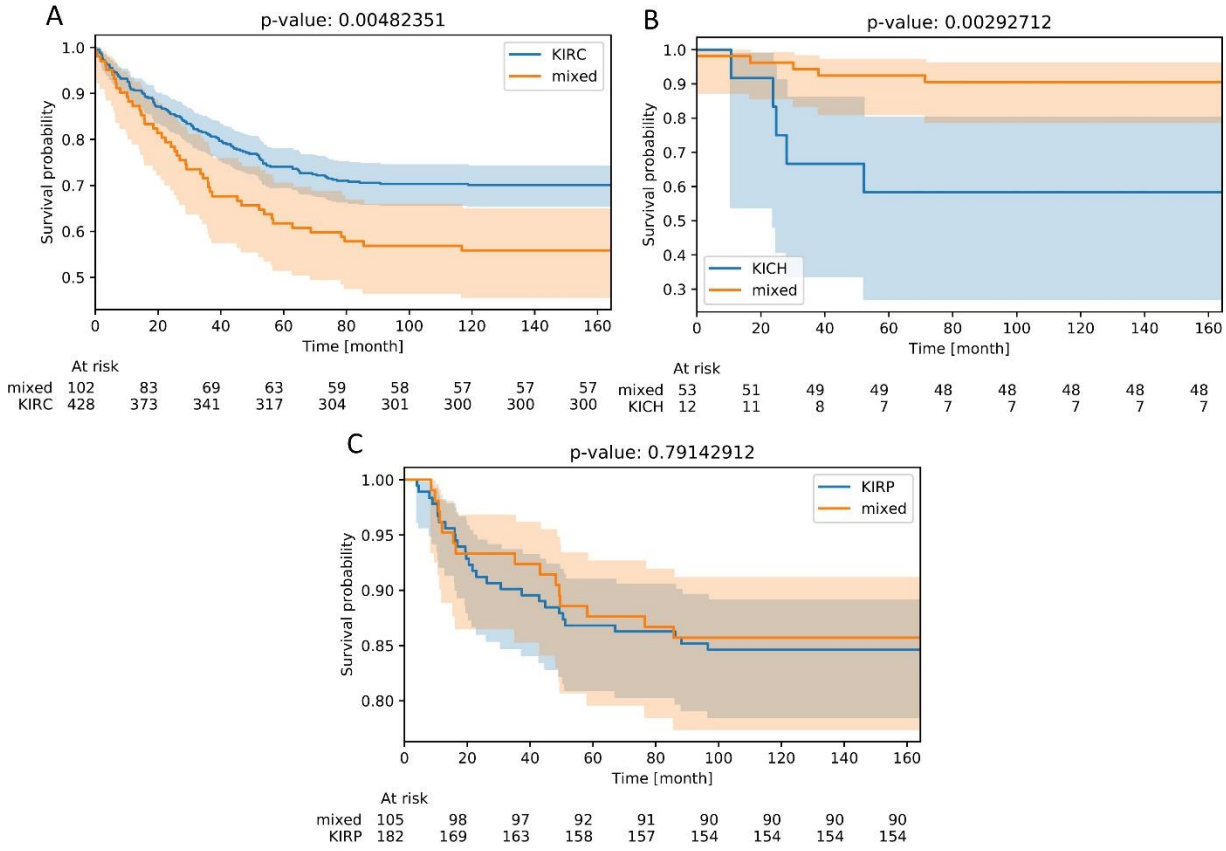


Figure 34: Kaplan Meier plot illustration of overall survival of ccRCC patients
(A), chRCC (B) and pRCC (C) from the TCGA database depending on mixed subgroup affiliation. Figure taken from (94).

5.3.4.6 Protein Expression Analysis

Since the here presented approach is based solely on mRNA data, statements can only be made about the transcriptional expression of genes. To overcome this problem, protein expression analysis of available bona fide gene candidates from different relevant pathways for each histopathologic RCC subtype were performed. Pathways and bona fide candidates were selected according to the obtained RF results as well as potential approved therapy options. For the considered mTOR/PI3K/Akt pathway, potentially treatable with mTOR-inhibition, a significant decrease in protein expression for TSC1 and PTEN for samples affiliated with the *mixed subgroup* was present for ccRCC and pRCC samples. Additionally, the *mixed subgroup* pRCC samples harbor a significant downregulation of MTOR. Regarding the angiogenesis-related genes, with potentially anti-angiogenesis therapy options, ccRCC samples as well as pRCC samples within the *mixed subgroup* present a slightly significant downregulation of protein expression. The immunotherapy associated protein PD-L1, relevant for treatment with immune checkpoint inhibitors, had a highly significant higher protein expression in ccRCC and pRCC samples affiliated with the *mixed subgroup*. The chRCC subgroup on the other hand does not show any significant protein expression differences for these genes at all.

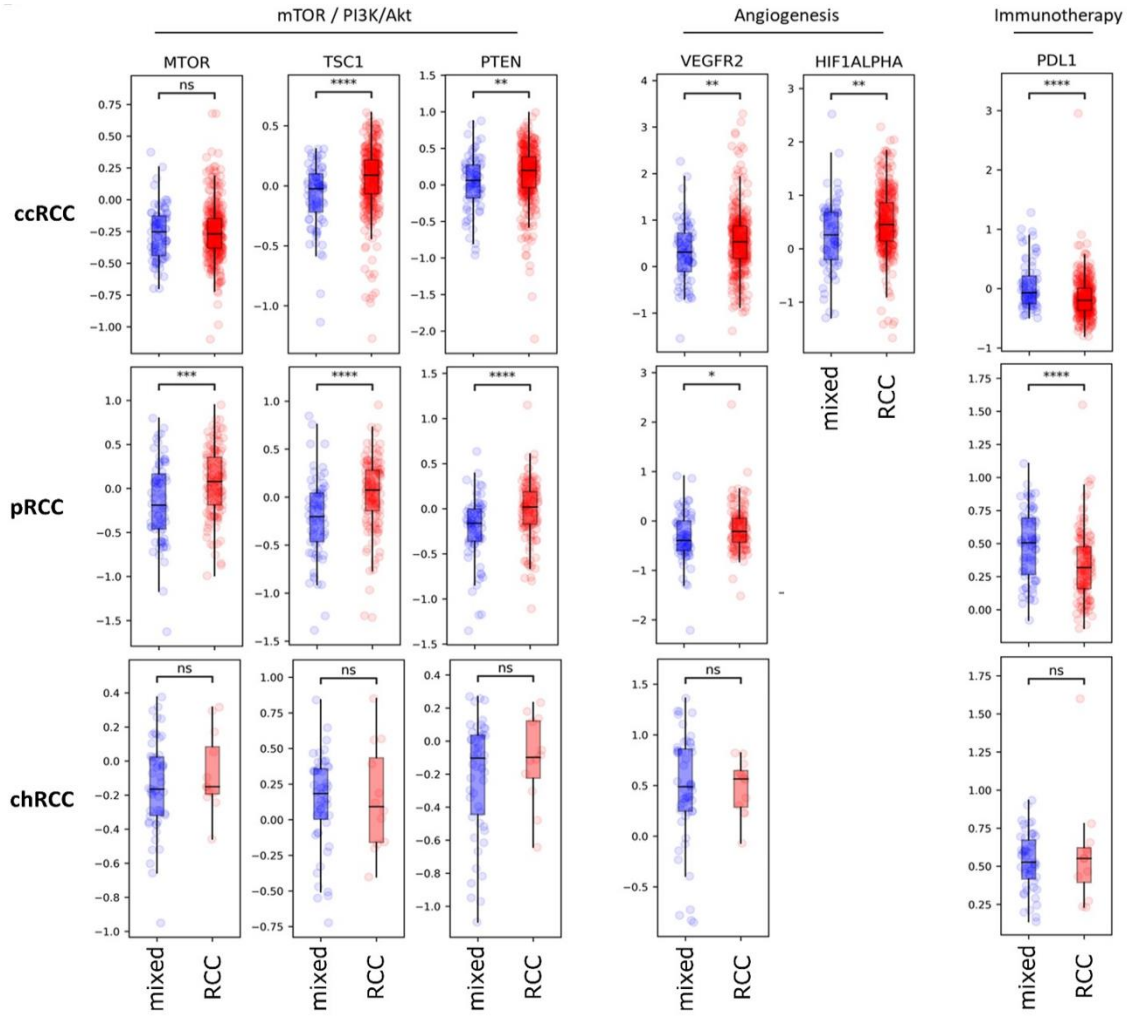


Figure 35: Protein expression levels of bona fide candidate genes

*Protein expression levels of bona fide candidate genes from mTOR-associated, angiogenesis-related and immune-related signaling for ccRCC, pRCC and chRCC samples inside (blue) and outside (red) the mixed subgroup (TCPA database). ns, not significant. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001. Figure modified from (94).*

Taken together, the combination of visual clustering and RF learning revealed significant transcriptomic differences between the investigated therapy naïve groups. These differences could also be partially confirmed at protein level and could therefore potentially be used for novel therapeutic targets. Furthermore, the results indicated that the survival benefit for patients affiliated with the newly identified *mixed subgroup* could at least be partially due to the subsequently received different therapies – mTOR inhibition for chRCCs and anti-angiogenesis compounds (for example tyrosine kinase inhibitors; TKIs) for ccRCC. As the results stated, the standard care for the different histopathological subgroups were potentially not suitable for the mRNA and correlating protein profiles of the patient, resulting in the observed survival benefit.

6. DISCUSSION

6.1 PREDICTION OF TUMOR ENTITIES BASED ON RNA-SEQUENCING DATA

At first, this thesis addressed the question whether the prediction of tumor entities based on RNA-sequencing is generally possible. Although there were already some publications addressing this question, no RNA-sequencing based study was yet able to demonstrate their essentialness for the determination of entities (65, 66, 71–73, 109). The reasons for this are different in each case, but can be roughly summarized into certain issues:

1. the (limited) amount and the quality of the data used

2. lacking/limited validation of the results

3. comprehensibility of the results from different sites/laboratories

4. feasibility of application of the results, especially in a real-life setup

6.1.1 THE INPUT DATA

The most important basis for a machine learning model is the underlying data for the model to learn on. Here, the principle of "garbage in – garbage out" applies in general, meaning that the results that can be obtained with data of low quality can themselves only be of low quality. Therefore, the selection of the underlying data is essential to be able to produce high-quality results. Having this principle in mind, data from TCGA database were chosen in the first steps of this work, as this database can generally be seen as one with data of good quality. One reason for choosing this database as the primary source is its general design. As a consortium database, for each of the entities and samples collected, not only the quality is determined in advance, but also – as it was important in this case – the processing and bioinformatic evaluation of the data produced in the laboratories were well-defined. For most of the patient samples follow-up data on survival and other information such as age and gender are also available, which is important for evaluation. In addition, there is additional basic information such as the TNM status of the tumor at hand, defining tumor, nodes, and metastases, which is of enormous importance for the subgroup determination in order to exclude possible biases that could originate from this source (e.g. clustering of samples according to tumor stage).

Another important reason for choosing TCGA as a database, besides the large number of different tumor entities and the fact that it is still actively maintained, is the fact that all available sequencing data originate from tumor samples *before* the start of the first therapy and thus represent the tumor sample in a therapy-naive state.

As already mentioned, TCGA database contains many different tumor entities and the exact number of entities used in the machine learning part is variable and should be selected according to the research question. Among the previous RNA-sequencing-based approaches, only one (65) had used a number of entities similar to the one used in this work, whereas all other previous approaches used fewer entities (66, 72, 73). By utilizing ML as method of choice, the decision of the investigated entities is somewhat arbitrary and should mostly be driven by the available amount of data. In theory, the cohorts used could be different but were intentionally chosen for this work. In very general terms, solid tumor entities consisting of at least 100 samples were selected. The only exceptions to this selection criteria are the uveal melanoma (UVM), adrenocortical carcinoma (ACC), and acute myeloid leukemia (LAML) cohorts, with the UVM and ACC cohorts consisting of less but nearly 100 samples (Table 1). The UVM cohort in particular was selected because of its proximity to the SKCM cohort (common progenitor cell: melanocytes). The ACC cohort is special in that it is a very rare tumor disease and the samples collected in this cohort represent the largest publicly available collection of RNA-sequencing data for this entity. The LAML cohort is thereby a certain exception in the selection, as this is a hematologic neoplasm and thus not a solid tumor. In a way, this entity serves as an internal validation group to better understand the results obtained based on the entities used. For a similar reason, namely to be able to directly internally evaluate the model trained at the end, the subgroups of the pheochromocytoma and paraganglioma (PCpG) cohort – Additional New Primary PCpG, Solid Tissue Normal PCpG, and Metastatic PCpG – were also introduced into the approach despite the fact that they only consist of 3 or 2 samples each.

Another critical point in data analysis, besides the samples and the data obtained from them, is the data actually used for analysis. In the pre-machine-learning era, the data under consideration were often reduced to a certain point before further analyses were performed to save computing time and computing capacity, as it was a limiting factor. A popular way to do this was to use, for example, the 5000 most variable genes within the cohort under consideration, where the genes had to be expressed in at least 75% of the samples, as it has been done by TCGA in their publications (110). In this example, the analysis is self-limited from the very beginning and introduces a certain bias, which ultimately affects the subsequent analyses and hence the results. Even the largest analysis to date for the prediction of tumor entities, although the goal of the work is the prediction of primary tumors – also for samples with CUP syndrome – uses a pre-analysis including differential gene expression analysis

between the entity under consideration and all other entities (65). Subsequently, the 40 genes for each entity considered to most significantly distinguish the groups are used. This type of pre-analysis finally biases the results since the machine learning model only uses the genes that have already shown the ability to significantly describe and separate the entities in the pre-analysis.
To be able to make a statement that is as unbiased as possible, also not introducing any bias due to possible data transformations, all 60,483 different transcripts provided by TCGA were used in an unprocessed manner in the analyses of this work.

In comparison to the data mentioned so far, which are all derived from bulk-RNA-sequencing, there is also the possibility of sequencing individual cells, the so-called single cell sequencing. With this method, it is possible to analyze individual cells from specific cell clusters or tissues – primarily with the aim of analyzing as many cells as possible – for example for expression analyses (scRNA-Seq) or for methylation studies. Compared to bulk sequencing, single cell sequencing enables the analysis of intra- and inter-tumor differences or commonalities (111, 112), but also the comparison of different cell types (113, 114) and the tumor micro environment (111, 115, 116). This applies to both RNA-sequencing and the determination of methylation. In particular, the analysis of tumor subtype ratios within a tumor can provide information on possible treatment strategies. Furthermore, the identification of the predominant clone within a metastasis in comparison to the primarius can provide profound insights such as therapy response (117–119) or whether the metastasis actually originated from the primarius.

There are also approaches for scRNA-sequencing with ML for entity determination, which provide very good results both in the comparison between normal and tumor tissue but also for entity classification (109, 120–122). The determination of tumor subgroups based on scRNA-sequencing has also already been successfully shown (123).

However, compared to bulk RNAseq, scRNA-sequencing also has some disadvantages, such as the higher cost of sequencing, or the reduced amount and sequencing depths of sequenced transcripts. There are also different established platforms, which – similar to bulk RNAseq – make comparisons between different data sets difficult. It is also possible that the selected method or conditions, such as the number of reads, may result in of gene dropouts, thus limiting sensitivity (124).

Taken together, the analysis of the feasibility of bulkRNA-sequencing seems to be most meaningful at this point in time, since direct translation into routine or clinical practice is more likely to be achievable.

6.1.2 Validation of Results

To this point there are only few in-depth ML analyses using bulk-RNA-sequencing techniques with a comparable number or even more different entities and samples than used in the present study. Two of the works gained their comparable higher number of samples by combining TCGA data with either their own generated data (methylation (62)) or by combining it with the ICGC database (65). Other approaches in comparison often used less entities and different approaches, also considering the evaluation of results, e.g. accessing different measurements of quality (66, 109, 121, 125). Due to these differences, in either data types or datasets, a direct comparison is not entirely possible between the different approaches or models. In case of methylation analyses, performed by Moran et al. (62), a positive prediction value of 88.6% was shown. This prediction value is comparably low to the achieved accuracies throughout the presented work in this theses or other published models. Additionally, Moran et al. presented a negative prediction value of 99.9%, showing the ability of their model to correctly identify tumor entities the patient is not affected of. However, the final evaluation of results should be part of future investigations, since the use of the parameters FP/TP/FN/TN – denoting false positives, true positives, false negatives, and true negatives, respectively – must be further evaluated in comparison to standard accuracy testing. Compared to the 88.6% positive prediction value from Moran et al., the resulting testing accuracies presented in this thesis were in any case comparable, most likely even better, even though a direct comparison is not possible due to the different used quality metrics. In the approach with a comparable amount of RNA-sequencing data, a top-1-accuracy (the accuracy obtained by using only the top predicted value/entity) of 98.54% in the cross validation and of 96.70% for the test dataset is shown. Also, in comparison to these results, the approach presented here is comparable for both approaches, before (98.41%) and after (98.96%) the combination of the LUSC/LUAD and STAD/ESCA cohorts, respectively. In total, the approach uses 817 different transcripts, for which an additionally testing against random genes (65) have not been performed. Comparing the results of the 800 random genes tested in the presented work, with the results obtained by the study, it is interesting to see, that the random results perform very similar to the selected 817 genes with a mean testing accuracy of 94.31% for the random genes and a maximum of 95.70%, showing the need of in-depth evaluation of ML results. Comparing the testing accuracies for the developed RF models with all genes and the prediction of all samples, an increase of 1.65% was notable (combination of LUSC/LUAD and ESCA/STAD showed a difference of 2.20%). Assuming the analysis using the 817 transcripts – stating that they used all available samples in their prediction – observed similar increases for their prediction accuracy of 98.54% for all samples, a testing accuracy of about 96.89% could be assumed, making this approach probably one percent point better than expected from the performed

random genes analysis. As a reminder, the best model based on only 100 genes had an overall prediction accuracy of 97.22%, or 97.95% after the entity combination. These results, especially in comparison, show that the chosen approach is competitive with regard to the used data.

A closer look at the false predictions of the newly introduced best performing RF model based on all available transcripts, as presented in this thesis, shows that squamous cell cancer entities such as LUSC, HNSC, CESC and BLCA account for a large proportion of the false predictions. In these cases, it is also noticeable that the incorrectly predicted entities almost always originate from those respective tissues. In addition, the cohort SARC is another entity with many false predictions, whereby the falsely predicted entities are spread over eleven different entities. In contrast to this, LUAD and LUSC, and STAD and ESCA, are cohorts which have false predictions only in the respective other entity. This confirms the usefulness of combining LUSC/LUAD or STAD/ESCA in a first step, but also shows that the prediction of these combined entities should be followed by another prediction specialized for exactly these cases. Furthermore, it also shows the clinical need for a tool that can determine tumor entities with very high accuracy to support pathologists. For example, one could imagine that the developed model could routinely be used and mainly serve as confirmation of the work of pathologists. However, if discrepancies between pathology and the machine learning model occur, the involved pathologist could directly confirm or exclude the predicted entity using specific markers. Especially for squamous cell tumors, this could be a considerable gain, as these entities are challenging to appoint. This can lead to incorrect diagnoses in some circumstances, which could hopefully be reduced by the additional use of the presented model. This will require some adjustment within pathologies itself, especially since RNA-sequencing would currently be required by default for the system to work, making the approach unfortunately unmanageable for current routine clinical use.

For two approaches one based on methylation data rather than RNA-sequencing data, a combination of data from different laboratories was performed. For the combination of TCGA and ICGC data a normalization of the data and thereby a certain harmonization was performed. In contrast to the performed harmonization in this work, where the success was shown using visual clustering approaches, it is not made fully clear that performed harmonization is technically allowed and useable as shown. This is especially noticeable for the ICGC data, because although they are collected on one database and were all generated by one consortium, the data were nevertheless generated differently depending on the site and subsequently bioinformatically analyzed differently. Comparing the prediction accuracies achieved with the RF models of the presented work for used ICGC datasets, it becomes evident, that there seem to be differences in the data, as some datasets get predicted perfectly, whereas other did not have a single correct predicted sample (e.g., PRAD-CA, PRAD-FR).

Normalization of the methylation data was also performed to harmonize the TCGA data with in-house generated data. Again, evidence that this normalization is valid, especially as for array data no combination or direct comparison between different data sources should be done (126, 127), is not provided.

The problem that arises from probably insufficient data harmonization or normalization can best be seen when evaluating trained machine learning models. When applied to other datasets, it often becomes evident that the model is "underperforming" and would hardly be suitable in a real-life setup, since the prediction accuracies of well over 90% cannot be reproduced and maintained with other than the training data. The so far largest data analysis showed only two additional datasets in the evaluation, having prediction accuracies of about 72% and 85%, respectively (65), compared to the six perfectly predicted datasets and several well over 80 and 90% accuracy in this presented work. This highlights the need for a universal applicable harmonization method, such that a standard evaluation of ML models can be established.

To obtain an applicability of the generated model in the present study for data of different sources, it was first necessary to develop an approach that harmonizes RNA-sequencing data from a wide variety of sources. To achieve this harmonization, the basic idea was that tumor entities would have to show certain recurring patterns independent of the sequencing facility and any batch effect, since they essentially represent the same entity. This basic assumption led to the idea that the ratios of all genes to each other should represent a stable system regardless of any individually introduced differences in data generation. However, since in this work 60,483 possible transcripts are available, this would mean that with the use of all ratios to each other 60,483*60,482 features – thus 3,658,132,806 – would have to be considered. Due to this enormous number of features and the resulting problems for computer capacity and computing time, the amount of data for the harmonization step had to be reduced.

To generate a set of genes usable for this purpose, a highly accelerated feature selection process was utilized. At each learning step, all features – equivalent to transcripts/genes in this case – with a feature value of 0 were deleted. In addition, the gene with the lowest value was deleted as well, to not stagnate at any given point. The process was carried out for 1761 steps, saving computing time and capacities. Finally, to identify the genes for further use, it was determined for each gene whether it was among the top100 genes with the highest feature values and the number of times it was present in the top100 features of the trained RF model was summed up and used as the final decision marker. However, to further validate the obtained top100 genes from the feature selection process, additional 100 models containing all available genes were generated, to counteract possible biases in the first selection steps of the used feature selection model. Again, for each gene, it was determined how often it occurred

in the top100 for each model. The result overlapped in 98 of the 100 genes, in a way showing the consistency of the RF, which is why the 100 genes from the simplified feature selection process were used for all further analysis. Furthermore, these analyses highlight the selected or learned features as they are not as random as assumed, although the initial starting point is different and random every time.

Taking a closer look at the so selected 100 genes it became apparent that a high proportion of those genes are already used as markers within pathological routine work. This also showed another advantage of the data used in the analysis, since these top 100 genes contain not only protein-coding but also non-coding RNAs, like long non-coding but also miRNAs, some of which are even completely uncharacterized at the present time, recognizable by their names as corresponding ENSG numbers. Since this transcript (ENSG00000227869 - Figure 13C) showed virtually only expression in the cohort KIRC, a verification in this entity should be done in the future. However, other transcripts also often showed expression in only one entity and could be evaluated as potential biomarkers in future research works, potentially extending the approved list of diagnostically used long non-coding RNAs (lncRNAs), like PCA3 in prostate cancer (128). In addition to cancer, the importance of lncRNAs has also been demonstrated for other diseases such as spinal muscular atrophy (129, 130), Angelman syndrome (131) and Dravet syndrome (132). Taken together, the results strengthened the emerging role of lncRNAs, not only from a mutational point of view (133), but also from the transcriptionally altered point of view (134, 135).

The subsequent evaluation of the 100 selected genes in terms of their ability to predict entities showed that these 100 genes were only 2.34% worse in mean testing accuracy compared to all genes. Using the best model and combining the LUSC/LUAD and STAD/ESCA cohorts, the 100 selected genes performed only 1.01% worse than all genes, predicting all samples with this model. Considering that only 0.17% of the original number of genes were used in this approach, this confirmed the quality of the selected genes but also of the chosen approach. An additional check of the 100 genes against randomly selected 100 genes also highlighted the selection, as for the 10 best random models only one of the 100 genes for all models overlapped with the selected top100 genes. In addition, the best model based on random 100 genes still only 1.1% worse in testing accuracy than the worst model based on the 100 selected genes, showing the general great performance of RF models built on biological data. Also, the range of prediction accuracies showed that the selected genes have a much better prediction quality than the random genes, due to a lower standard deviation. In conclusion, this last test showed that the selected genes performed measurably better than the random ones. Given the fact that only a fraction of the available genes was required, the prediction accuracy was very good and was still comparable to other approaches with their best models (62, 65). A further testing against random genes also showed that there was a strong increase in accuracy

between 100 and 200 or between 200 and 300 random genes, whereas there was only a small increase in accuracy at all further steps up to 1000 random genes. In the study by Yue Zhao et al. (65), which considered 817 genes, such a comparison against randomly selected genes was missing, meaning that the gain of the model compared to randomness cannot be accurately assessed. When comparing the obtained testing results for the different RF models either using 800 random transcripts or all available transcripts, only a small gain of accuracy was detectable. This result showed the need of in-depth analysis of selected gene sets to randomly selected genes, as otherwise no final evaluation of the quality of the model could be made.

Based on these positive results, the intended harmonization of the data was then carried out, whereby the quotients for all of the selected top100 transcripts were calculated for each sample. The newly obtained data basis reduced the number of quotients to only 9900 ratios of genes. Repeated individual learning of RF models based on these ratios again showed the consistency of the trained models, as the mean testing accuracy for the 100 genes alone had the same testing accuracy as the ones based on the 9900 ratios of the top100 genes. In conclusion, this showed that the use of the ratios themselves did not cause any loss in testing accuracy for entity prediction and thus can be used instead of the 100 genes alone.

6.1.2.1 Evaluation Datasets

Besides the testing accuracy, the predictive accuracy of the whole dataset and the test against randomness, the use of additional evaluation datasets is essential. For the evaluation of the model using all genes, only cohorts from the TCGA database were available in this study, whereas for the model using ratios also other datasets were useable. In total, 22 different cohorts from three different databases were used in this work to evaluate the generated models. The use of other TCGA cohorts – KICH, KIRP, and READ – served as a certain positive control or as a check of the transcriptomic proximity of the entities KIRC, KIRC, and COAD, respectively. For the cohorts KIRP and KICH, it appears that these two renal cell carcinoma entities, although originating from different cells of origin, show a certain transcriptomic proximity to each other, especially when compared to all other entities considered. Looking at the COAD and READ cohorts, it becomes apparent that both cohorts basically represent the same entities. Due to this, it was not surprising that the evaluation of the predictions for this cohort were this positive.

The most comprehensive evaluation of the developed method and model took place when using the harmonized data based on the top100 transcripts. For this purpose, 18 additional

datasets were selected from the ICGC database and the Gene Expression Omnibus (GEO). All these datasets are publicly available and contain data from human tumor samples – with at least 10 samples – included in the developed model and are already available as FPKM files. This allowed the data to be directly harmonized and then predicted with the appropriate model. In combination with the four datasets of TCGA, the evaluation of the generally applicable model consisted of a total of 1963 samples. The prediction of the specified entity was correct for 1580 (80.49%) samples. Taken individually, it appears as if the model performs at a mediocre level with respect to entity prediction. However, a closer look reveals that for 6 cohorts a perfect prediction could be made and for another 8 cohorts the accuracy was above 80%. For the cohort OV-AU, the results were particularly interesting, as in addition to the 49/93 (52.69%) samples correctly predicted as ovarian cancer further 43 were predicted as uterine cancer (UCEC), which have at least some spatial proximities to each other. The interesting thing about the result of this dataset is, that the TCGA-OV dataset, which serves as the basis of the model, is 100% correctly predicted in the model with all genes. Looking at the results of the model based on harmonized data, 9 samples were predicted incorrectly for the TCGA-OV dataset. Besides one sample which was predicted as STAD and one which was predicted as SARC, 7 more were predicted as UCEC, which thus overlaps with the results of the OV-AU dataset using the model based on ratios. A possible reason besides the sampling or the basic misclassification could be the underlying genes, which do not completely overlap with the genes available in the model and thus lead to a false prediction, because specific genes for e.g. subgroups are not available. Also, exactly these subgroups could be the cause for the false prediction since they might not be represented in the original model.

Another striking point is that 3 of the datasets with prediction accuracies below 80% were datasets with samples from pancreatic carcinomas. The reasons could be the same as for the OV-AU dataset, e.g. subgroups not represented in the underlying model.

The last and probably most striking dataset is the PRAD-FR dataset. Even though this dataset consists of only 25 samples, the predictions for each of these samples were wrong. STAD accounted for 14 of the predictions, SKCM for 9, LIHC for 1, and BLCA for 1. However, comparing the false predictions with the visual cluster results of harmonization, it was noticeable that this dataset cluster a priori distant to the other prostate carcinomas. Thus, the complete false prediction of all samples was not surprising, even though 97 of the 100 genes are present in the dataset. One possible reason for the separate clustering and the subsequent complete wrong predictions could be a specific subgroup of prostate carcinoma for which the corresponding markers were not captured within the model. The proximity to the NEPC-WCM dataset, which consists of neuroendocrine prostate carcinomas, leads to the assumption that the samples might also be affiliated to this subgroup, which only occurs to a small extent in the underlying dataset of the trained model. Strikingly, 10 samples were incorrectly predicted using

the harmonized model of the TCGA-PRAD dataset, with 4 predicted as BLCA or STAD and one each as COAD or SARC. For the NEPC-WCM dataset, an additional complicating factor is that only very few ratios could be calculated, since very many values – even if 75 out of 100 genes were present – show no expressions and thus lead to poor or incorrect predictions. The assumption of dataset inherent problems is strengthened by the results of the PRAD-CA dataset, since this one was predicted 100% correctly with 100 out of 100 genes overlap.

However, the influence of the 100 genes used depending on the entities considered becomes apparent in these analyses. For some entities, it seems that sufficient predictions can be done with only a handful of the 100 genes in the dataset, whereas the presence of almost all the genes considered could lead to poorer predictions if genes that are explicitly important for this entity in the model are missing.

Overall, this evaluation is probably one of the largest evaluations, in terms of number of different cohorts and databases, ever used in the field of tumor entity prediction using RNA-sequencing data. Considering this fact, the evaluation accuracies of other studies, and the fact that for more than half of the datasets used the predictions are very good, the harmonization can be considered a success. Unfortunately, no unrestricted recommendation for the use of the model trained on the harmonized data can be given and further evaluation of the model in routine clinical practice is mandatory to show the applicability when all 100 genes are always available. Improvement of the model with new data and special subgroups – based on harmonized data – may additionally lead to an improvement of the retrospective evaluation.

6.1.2.2 METASTATIC SAMPLES AND CUP PREDICTION

Probably the most important field of application for a model to predict tumor entities is not the determination of primary tumors, since this can already be done with sufficient quality by pathologists and physicians, but the determination of the primary tumors in cases of metastasis and furthermore the so-called CUP syndrome. The determination of the primary tumor is important due to the direct clinical relevance, particularly with regard to the best therapy. Another alternative application could be the differentiation between metastasis and secondary tumors. In the case of a metastasis, a palliative treatment would often be the result, whereas in the case of a second carcinoma, a targeted therapy for the second carcinoma would be appropriate. The possibly most complicated case is the presence of a CUP, as it can be seen as metastatic with an unknown primary tumor. However, the previous work of Moran et al. (62) showed that a targeted therapy designated for a predicted primary tumor compared to standard chemotherapy for the unknown primary tumor resulted in a significant survival benefit for

patients when compared to untargeted treatments (62). In this context, other studies often point out that their developed models also should be applicable to CUPs and metastasis, but almost always evidence for this statement to be true is lacking (62, 65, 66). Also, the dependency of metastases and the resection site often remains missing.

The correct classification of metastases to their corresponding primary tumor was addressed by analyzing three different datasets with included resection site of the metastatic specimen. The analysis using visual clustering showed, that there was no link between resection site and clusters harboring specific transcriptomic features. Therefore, the conclusion that the resection site has no influence on the prediction was made. Especially regarding data transformations, it becomes apparent that the data transformation used has a high impact on the clustering, but also does not lead to a dependency based on the resection site. In particular, only for the Dream Team dataset, a bone and liver cluster were found. However, it should be noted, that only the bone cluster persisted across all data transformations and clustering methods, whereas the liver cluster showed a dependence on transformation and method. Regarding different datasets, altered clusters showed up depending on methodology or data transformation. Since the bone cluster in question also consisted of only a part of the total bone samples, this cluster is assumed to be the only genuine cluster, since it is the only consistent one being always present. Causes for this could be, problems with the collection of the material, or problems with the preparation of the RNA, which is not uncommon for bone samples. Other cluster formations, such as the liver cluster mentioned above, could be due to insufficient tumor cell content and aberrant presence of the tissue from the metastatic tumor niche.

It could be shown in two different datasets – the complete TCGA-SKCM dataset and the MBC-project dataset – that no independent distinction could be made between metastatic and primary tumors using the applied methods. Furthermore, it became apparent, that knowledge about the classification in the respective group could lead to a clustering biased by this pre-assumption. Combined with the results of the analysis of the TCGA-KIPAN cohort chosen as the evaluation dataset, this showed that knowledge of existing or assumed subgroups may indeed have some influence on the final assessment of clusters. Thus, the question arises whether a clustering based on visual clustering procedures by dimension reduction should necessarily be done automatically and standardized – or at least by a fixed verification procedure – to prevent the introduction of a bias. Since no distinctions between primary tumor and metastasis can be detected in the bulk RNA-sequencing used here, the results can additionally be understood as a confirmation of the linear progression model (136). Considering the parallel progression model (137), one would rather have to assume different clusters with different transcriptomic properties. These findings in relation to the linear progression model are also in line with various studies that have so far not been able to make

a real distinction between primary tumor and metastasis, even at the single cell level (138, 139).However, it must be considered that in the classification of primary tumors and corresponding metastases, the differences may be less pronounced, or subgroups may be identified which are not yet known and therefore no clear (known) classification exists. Especially with bulk sequencing, the problem of plasticity and inter- and intratumoral heterogeneity cannot be fully resolved, making scRNA-sequencing possibly more suitable for elucidating intrinsic subgroups within a tumor, for example in gastric cancer (140) in melanoma (141) or the tumor microenvironment of breast cancer (142). These findings could also be transferred to metastases, potentially giving insights into metastasis formation (143). Therefore, scRNA-sequencing could be a potential tool to better reflect intratumoral heterogeneity and plasticity and thus determine the predominant or predominant subgroups. Overall, if scRNA-sequencing is generally feasible, machine learning will also become more useful for this application, thus allowing the identification of new subgroups that could subsequently be correlated with treatment response.

After showing that the resection site does not affect visual clustering for bulk- RNA-sequencing data, it was possible to reasonably predict metastatic samples on basis of bulk- RNA-sequencing data as well. Using all genes, only the metastatic samples of the TCGA database could be used due to the already mentioned different effects, like the batch effect, based on different laboratories. Additionally, only TCGA cohorts contained all of the genes used in model learning. Across all ten datasets, of which seven consisted of only one or two samples, with a total of 392 samples, 382 (97.45%) were correctly predicted. This shows that the model trained and used here can predict the primary tumor with a very high accuracy independent of the resection site. Additionally, using the top100 transcripts resulted in only small changes, so that the corresponding model correctly predicted a total of 378 (96.43%) samples. A peculiarity of this analysis is that for the dataset of metastatic HNSCs only one of two samples was correctly predicted using all genes, whereas using the 100 genes both samples were correctly predicted as HNSCs. Following this example, there might be cases where the usage of a reduced dataset could be of advantage – possibly reducing a bias in the data.

In combination with the data harmonization established, it was also possible to predict three additional cohorts of metastatic samples. For TCGA data, there were no differences between these two models, which again underlines the quality of the models but also the usability of the harmonization, as it again did not introduce any differences. For the datasets of Dream Team and MBC-project, there were also very good prediction accuracies with 221/266 (83.08%) and 137/146 (93.84%). For the metastatic samples of the NEPC-WCM dataset with 49 samples, not even one could be predicted correctly. However, this is hardly surprising when considering the visualization of the harmonization. Here it is noticeable that the NEPC-WCM dataset clustered far away from the prostate cluster and was additionally located next to the PRAD-FR

dataset, which had no correct predictions, either. This shows that the visualization methods used reflect the data very well and were able to identify datasets that were very likely to be predicted incorrectly. Looking further at the overlap of available genes to the top100 gene signature, it is again evident that no general statement can be made about the functionality of the model depending on the number of genes. For example, although the MBC-project with the highest accuracy also has the highest overlap with 79 genes, Dream Team with 68 has a lower overlap and a significantly better prediction quality than NEPC-WCM with 75 gene overlap. For future considerations, it would therefore be desirable if the specific top 100 gene set was always available for each investigated tumor sample, to enable a broad evaluation on the basis of all genes, allowing the eventual evaluation of the developed model.

In general, it can now be stated that – if CUPs are only metastases of unknown primaries and do not represent a completely separate, yet unknown entity – metastases and thus also CUPs can be predicted very accurately with the aid of the developed model on the basis of the quotients of only 100 genes. In combination with the evaluation data for primary tumors, this work represents the most comprehensive evaluation of a RF model – in terms of the number of different datasets from different sources – at the current time. The fact that for the metastatic dataset of the SKCM cohort of TCGA there is hardly any difference between the use of all genes and the top100 transcripts or their quotients shows that the models are virtually equally good, also for the analysis and prediction of metastases.

6.1.3 Application in Real-Life Setup

Considering the very good results from the predictions for both primary tumors and metastatic samples, it was of interest whether the developed model could actually have a place within clinical routine.

Currently, the model only requires 100 genes as an input, but – at the moment – these would have to be obtained from a complete RNA-sequencing experiment. To be applicable in a real-life setup a novel and specific array or panel of this newly discovered gene set should be developed. Considering the fact, that the sample material is one of the most important prerequisites, which is also not so easy to obtain, it may happen in some cases that performing RNA-sequencing is simply not possible. The sample quality would also have to be validated additionally, especially regarding the used model, since FFPE material is very often used and there could be differences compared to fresh material. However, if these two requirements are met for a sample, there is no reason the model could not be used in clinical practice. Nevertheless, RNA often is already obtained for other molecular genetic diagnostics, meaning

that only the RNA-sequencing is added as an additional step. Since molecular genetic diagnostics already involve sequencing, it is often not even necessary to purchase new expensive equipment for this purpose.

For the future, it is highly preferable to have specific panels or arrays for sequencing, depending on the 100 genes presented here, which can perform the sequencing for only the named 100 genes with a small amount of material, reducing the costs for the procedure. However, in-depth molecular genetic diagnostics including complete RNA-sequencing and generation of methylation data, are not routinely performed for every newly diagnosed cancer patient. In the case that a comprehensive elucidation of a tumor becomes routine at some point in the future, then, our presented model for entity prediction would only be a by-product and could be used without further costs and efforts.

Another point to be considered as critical in a real-life setup is the quantity of available sample material. As mentioned before, it could well be that the material used contains too few tumor cells and thus the analysis based on RNA-sequencing would fail or deliver poor or incorrect results. To prevent this, a preliminary step would be necessary, which first checks whether the material is likely to be tumor or normal tissue. If this step is performed by RNA-sequencing, one could again run into the problem of losing important material. One possibility here could be, for example, the automatic determination and evaluation of sections with the aid of machine learning, which could be developed specifically for this problem to increase the chances of success of RNA-sequencing and to save sample material.

6.2 Subgroup Analysis in RCC

Besides the determination of the correct tumor entity, the identification of subgroups is essential for the prediction of an appropriate therapy. To test the visual clustering methods for this application, the TCGA-KIPAN dataset was used – a combination of the different RCC subgroups of TCGA, KIRC/KICH/KIRP. For these entities, it was shown in the developed models that they can each be predicted as ccRCC and thus have some transcriptional proximity to each other. However, the important aspect to note here is that especially the ccRCC and chRCC subgroups have different therapeutic regimens, with pRCCs sharing the regiment of the ccRCC. Clear evidence-based therapy is mostly established for ccRCC addressing the angiogenesis pathway, but recently also utilizing immune checkpoint blockade. Due to the currently rather inefficient therapy options a further distinction following the prediction of the tumor as RCC is essential to initiate an appropriate therapy (53, 144–146).

As already done for the visual representation of the data harmonization and the investigation on the influence of the resection site of metastases on clustering, both t-SNE plots and UMAPs were used to initially analyze the general clustering of the RCCs. Again, as in the analysis of the metastasis data, three different data transformations were used for the KIPAN dataset – unprocessed, log10, and log10+1 – to compare the unprocessed approach previously followed with the standard logarithmic data transformation approach. As before, both methods showed a strong dependence of the clustering on the chosen data transformation. The logarithmic data transformations showed – with only a few exceptions – a clear and distinct cluster formation depending on the considered subgroup. Nevertheless, it is evident that the samples that clustered incorrectly occur mainly within the cluster of the chRCC subgroup and thus in a certain way formed a cluster that consists of all three subgroups. Since the UMAP approach based on unprocessed data unfortunately offered hardly any possibility for evaluation, as was already observed with the metastasis datasets, a comparison with the unprocessed data of the t-SNE approach was of particular interest. The latter did not show a complete distinct cluster formation based on the underlying subgroups, but nevertheless gave a hint of it, with a large ccRCC and several small pRCC clusters around the ccRCC cluster. Only the chRCC cluster formed a certain exception, since the chRCC samples themselves grouped together almost completely, but was completed by many ccRCC and pRCC samples. In a direct comparison between the individual approaches, there were two chRCC samples, which for log10 and log10+1 in the t-SNE plot and for the log10+1 approach in the UMAP were located outside the chRCC cluster and inside the ccRCC cluster. For the log10 UMAP approach, there was only one chRCC sample inside the ccRCC cluster, but 5 samples were inside the pRCC cluster. The unprocessed data from the t-SNE plot were selected and analyzed in more detail because, based on all the results, it could be concluded that this plot is the one that most closely represents biological reality and is therefore most useful for subgroup identification. Although, it could be argued, that the predetermined subgroups are more likely to be reflected in the other plots. However, most importantly, the two chRCC samples within the ccRCC cluster, but also the bone cluster shown previously within the Dream Team dataset, reinforced this decision, as natively occurring clusters should be recognizable within all data, regardless of data transformation. In particular, the log10+1 transformation deliberately introduces a small change to the expression values, which however leads to some change in genes that are not strongly expressed.

In the further in-depth analyses of the unprocessed t-SNE plot, it was shown that clustering can represent biological differences. For this purpose, a linear border was manually drawn within the t-SNE plot to represent a new cluster – consisting of all three subgroups – but mainly representing the chRCC cluster. In addition, the clusters for pRCC and ccRCC that could already be identified were added. The subsequent examination of the clusters for potential

biases due to clinical parameters such as TNM status or gender and age showed that, especially for the important clinical parameters such as TNM status but also grading, there were no statistically significant differences.

To be able to make a more precise statement about the transcriptomic properties of the clusters, several RF models were generated to predict the clusters of the individual samples. It should be noted that the chRCC samples outside the cluster designated as "*mixed subgroup*" were not used because they did not belong to an obvious cluster of their own and would most likely represent false classifications and hence lead to errors in the learned models. However, the separately identifiable pRCC clusters were treated as individual clusters. The testing accuracy of the individual models was consistently high, which is why an analysis of the most important transcripts was included. Analogous to the procedure for the RF models for entity prediction, the top200 genes were determined. The analysis of these genes showed only mitochondrial or mitochondrial-associated genes within the top10 genes, which is not surprising since a mitochondrial influence is already known for the chRCC subgroup. However, since the learning process did not focus on the chRCC subgroup alone, but on a mix of all three subgroups, the conclusiveness of the results to this extent was nevertheless surprising. A close examination of these mitochondrial genes, irrespective of the underlying histopathological subgroup, showed clear and significant differences between the samples affiliated with the newly defined *mixed subgroup* compared to their counterparts outside this *mixed subgroup*. However, because most of the data were derived from the ccRCC cohort, angiogenesis-related genes were also of interest. Here, no significant differences for the pRCC and chRCC subgroups were observabed. For the ccRCC cohort, however, there was a significantly lower expression of angiogenesis-related genes within the "*mixed subgroup* ". Prompted by these two new findings, it was subsequently investigated whether there was a relationship between angiogenesis-related and mitochondrial genes within the ccRCC cohort. In fact, it could be shown that there is a significant negative correlation between mitochondrial and angiogenesis-related genes within the defined top200 genes. This observation could also be made in two other independent clear cell renal cell carcinoma datasets and one FH-deficient dataset, again confirming the results, but more importantly the defined subgroup. This was further confirmed by protein expression analysis, showing significant differences between the clusters.

The results of the analyses performed for the unprocessed data are further confirmed by current research showing an advantage for some patients with ccRCC with additional administration of metformin (147), as it inhibits the mTOR-pathway (148) and the growth and migration of ccRCCs (149). In combination with the obtained clustering results, it could be concluded that a proportion of ccRCCs are more dependent on mitochondrial genes and mTOR-pathway, as shown by protein expression analysis, than on angiogenesis. Due to this,

mTOR inhibition alone or in combination with antiangiogenic tyrosine kinase inhibitors could be beneficial for some patients, as the underlying transcriptional profile is different from the majority of ccRCCs. Also the WHO itself recognizes certain grayscale cases such as clear cell chromophobe renal cell carcinoma (58).

Based on the presented subgroup and the confirmation of the transcriptional differences, which can also be shown in part at the protein level, assumptions could be made regarding possible therapies. As already mentioned, some clear cell renal cell carcinomas show improved survival with additional administration of metformin in recent studies (144, 147, 150). Transferred to our results, it is reasonable to assume that these are renal cell carcinomas that would be affiliated with the *mixed subgroup*. Since there is a therapeutic regimen for chromophobe renal cell carcinomas, it could be possible to be applied to the papillary and clear cell renal cell carcinomas within the newly defined *mixed subgroup* as well. The possible use of immunotherapy could also be given a new basis within RCC with the help of the results presented, as the PD-L1 expression is significantly higher in the defined *mixed subgroup*. This indicates a potentially better response to immune checkpoint blockade (ICB) therapy as it has already been shown and proposed for other entities, like triple negative breast cancer (151). However, it has to be considered that the results differ between studies and entities, also due to problems accessing the PD-L1 expression status by immune histological staining or the different used scoring methods (152). Therefore, it remains unclear to which extent the PD-L1 expression is a predictive biomarker for ICB (152, 153), for tumor types other than non-small-cell lung cancer (NSCLC) (154).

A further point to consider is the manual bias introduced in the analyses presented. The described subgroup and analyses are based on the manual separation into clusters. A risk classifier based on the obtained results and the resulting clustering would be indispensable for new patients to enable a classification of the new sample and thus a possible therapy recommendation. As this cannot be offered at this time.

In summary, the choice of using the unprocessed data for subgroup analysis could be justified, at least in RCCs shown here as an example. The assumption that unprocessed data based on t-SNE plots are the best choice for subgroup analyses was reinforced using the visual analyses of the metastatic data shown. Furthermore, our analyses confirm the assumption, that metastatic data do not cluster in dependence of the resection site, but in dependence of the underlying subgroup, which is also evident in the subgroup analysis of RCCs. The results obtained also show direct clinical relevance, as the newly found *mixed subgroup* not only shows a significant survival difference for ccRCCs and chRCCs, but also postulates a possible therapeutic regimen for this subgroup, which needs to be verified in future works.

6.2.1 Clinical Relevance of Subgroup Prediction

In comparison to the results combining ESCA and STAD or LUSC and LUAD, only the KIRC cohort was predicted in the subgroup approach. The prediction of the other two cohorts – KIRP and KICH – showed that these two subgroups are very likely to be reassigned to RCC with very high accuracy. However, all these examples leave out a possible clinical relevance. Especially in lung carcinoma, but also in RCCs, different therapeutic options arise depending on the subtype, which is why this determination is imminently important in addition to the pure localization. Furthermore, it was shown that an individual analysis of special subgroups is possible, opening up further therapeutic options, even some that were previously unknown. The procedure is also transferable to other entities and allows more precise subgroup determinations, which can and should be connected to the determination of the primary localization of the tumor. However, the retrospective analysis of the data also has the disadvantage that without risk stratification or without directly possible group classification, the individual benefit for new patients is initially low and further demonstrates the need for such classifications in everyday clinical practice.

7. CONCLUSION

All analyses performed throughout this thesis confirmed the main assumption: Tumor entities can be reliably predicted based on RNA-sequencing data. This prediction can be obtained using a RF model with up to nearly 99% prediction accuracy. Furthermore, a novel harmonization method based on the ratios of only 100 transcripts was introduced, enabling reliable prediction of tumor entities from different sources. The obtained testing accuracies were only about one percent point worse than using the initial full approach. Additional analysis based on visual clustering of metastatic samples indicated the possible application of the developed RF models also on metastatic samples. Finally, applying the different developed models to other datasets and datasets containing metastatic samples again showed high prediction accuracies of usually above 80 or 90%. These results further highlighted the potential of the newly developed harmonization method, enabling the combined used of data from different sources.

Additionally, it could be shown that data transformations were important for identification of subgroups within tumor entities. Using unprocessed data clinically relevant subgroups in RCCs could be identified, although distinct histopathologic separation was lost.

Despite all these new findings in the development of the models and the good overall applicability, the use of the models in a clinical routine setup still seems to be far away. Furthermore, the model presented is currently based on RNA-sequencing, which makes the approach itself quite expensive. The further development of the new potential biomarkers presented could possibly reduce the problem in the future. To achieve the goal of predicting therapy options, there is still a need for additional models that can determine the exact subgroup of the tumor and subsequently include other factors such as activating or resistance-mediating mutations.

Overall, it could be shown that even basic machine learning models can sufficiently support the growing and increasingly difficult work of molecular diagnostics in the future.

8. OUTLOOK

In the present work, some basic questions on the way to automatic prediction of appropriate tumor therapy using machine learning have already been answered. However, as the discussion and the results have shown, there are still many more questions that need to be answered before a therapy-option prediction model can be used in clinal routine work (Figure 36).

To improve the current RF model, other data, such as methylation data and miRNA data, could be incorporated with the goal of determining which data or data combinations are most predictive. These new models should additionally also be used to distinguish between normal and tumor tissue. Adding whole genome sequencing (WGS), possibly best suited to determine subgroups within tumor entities – based on karyotypes and copy number alterations (CNA) – combined with whole exome sequencing (WES) and already known therapy information such as response and resistance to the mentioned sequencing data, the prediction of tumor therapy could eventually be conceivable.

As shown, the identification of subgroups has a high clinical relevance and consequently a also for the choice of therapy. One possibility to further improve the developed RF model, for example, could be the determination of subgroups for all tumor entities used in the model, analogous to the shown RCC data.

Since a real-life setup should always be considered, it may well be that other datasets, e.g. derived from RNA-panel sequencing are more applicable, even though there might be some accuracy trade-off. Therefore, this kind of data should be tested as well to determine the best suited approach. In the present work, serious applicability based on harmonization of data with only 100 genes from RNA-sequencing has now been demonstrated for general use. In addition, for the first time, the number of genes is small enough for panel or array sequencing, also making Nanostring sequencing feasible for everyday clinical use.

Given the ever-changing landscape of therapeutic possibilities and options, it is necessary that the developed ML models are continuously re-trained and modified in order to achieve the best possible results.

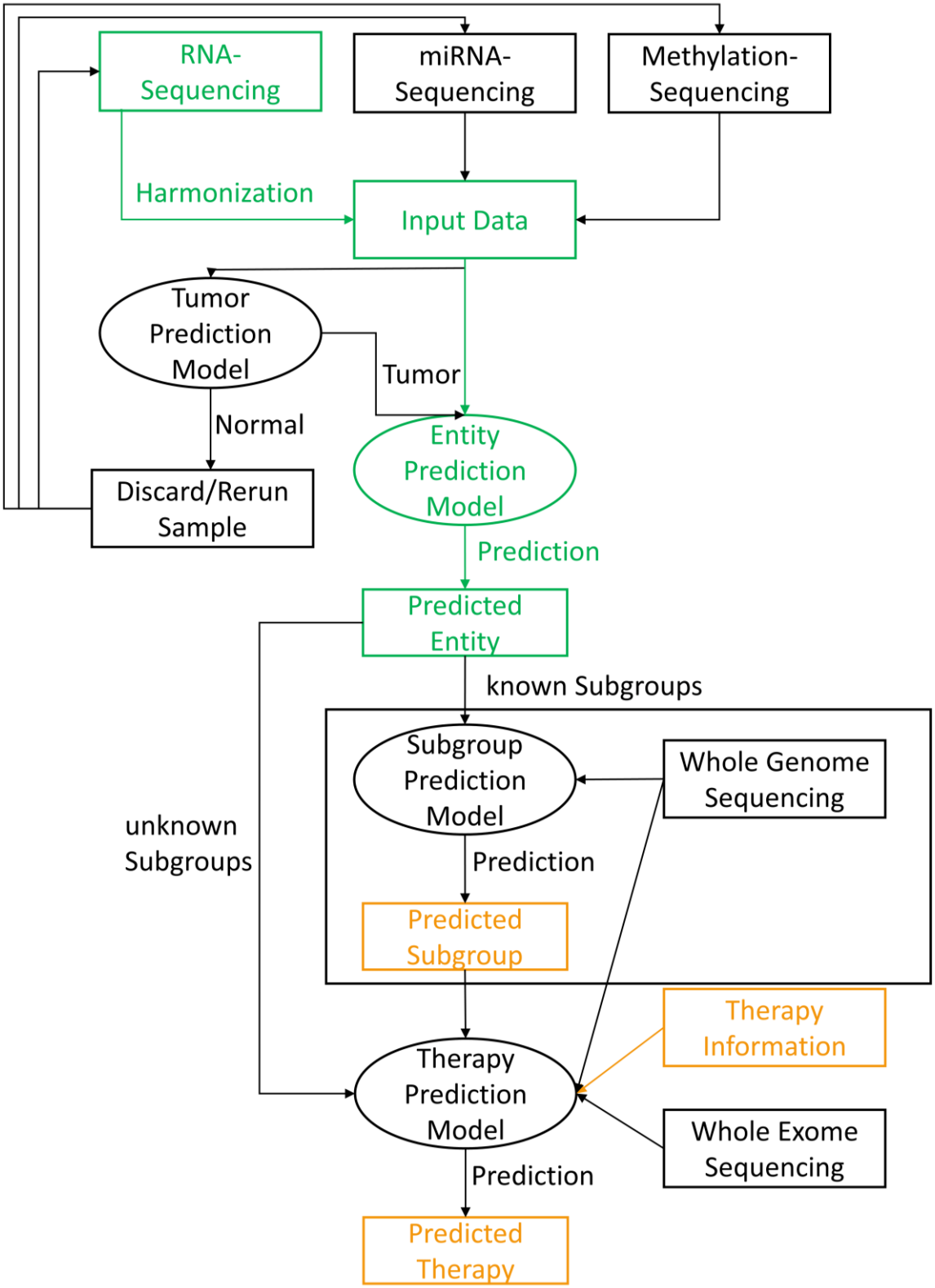


Figure 36: Updated proposed workflow for therapy prediction based on machine learning

Modified possible workflow to obtain a therapy prediction for patients with a tumor burden. Parts that have already been addressed and been examined in-depth are coloured in green. Objectives, that have only been addressed in part and have shown basic functionality but not in general, are marked in orange. All parts that could be needed in the workflow to predict a therapy option but have not yet been addressed are shown in black.

References

1. Yamamoto Y, Kanayama N, Nakayama Y, Matsushima N. Current Status, Issues and Future Prospects of Personalized Medicine for Each Disease. *J Pers Med* (2022) **12**. doi:10.3390/jpm12030444
2. Rodler S, Jung A, Greif PA, Rühlmann K, Apfelbeck M, Tamalunas A, et al. Routine application of next-generation sequencing testing in uro-oncology-Are we ready for the next step of personalised medicine? *Eur J Cancer* (2021) **146**:1–10. doi:10.1016/j.ejca.2020.12.024
3. Kim G, McKee AE, Ning Y-M, Hazarika M, Theoret M, Johnson JR, et al. FDA approval summary: vemurafenib for treatment of unresectable or metastatic melanoma with the BRAFV600E mutation. *Clin Cancer Res* (2014) **20**:4994–5000. doi:10.1158/1078-0432.CCR-14-0776
4. da Rocha Dias S, Salmonson T, van Zwieten-Boot B, Jonsson B, Marchetti S, Schellens JH, et al. The European Medicines Agency review of vemurafenib (Zelboraf®) for the treatment of adult patients with BRAF V600 mutation-positive unresectable or metastatic melanoma: summary of the scientific assessment of the Committee for Medicinal Products for Human Use. *Eur J Cancer* (2013) **49**:1654–61. doi:10.1016/j.ejca.2013.01.015
5. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med* (2011) **364**:2507–16. doi:10.1056/NEJMoa1103782
6. Flaherty KT, Puzanov I, Kim KB, Ribas A, McArthur GA, Sosman JA, et al. Inhibition of mutated, activated BRAF in metastatic melanoma. *N Engl J Med* (2010) **363**:809–19. doi:10.1056/NEJMoa1002011
7. Solit DB, Rosen N. Resistance to BRAF inhibition in melanomas. *N Engl J Med* (2011) **364**:772–4. doi:10.1056/NEJMcibr1013704
8. Nazarian R, Shi H, Wang Q, Kong X, Koya RC, Lee H, et al. Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature* (2010) **468**:973–7. doi:10.1038/nature09626
9. Kugel CH, Aplin AE. Adaptive resistance to RAF inhibitors in melanoma. *Pigment Cell Melanoma Res.* (2014) **27**:1032–8. doi:10.1111/pcmr.12264
10. Meierjohann S. Crosstalk signaling in targeted melanoma therapy. *Cancer Metastasis Rev* (2017) **36**:23–33. doi:10.1007/s10555-017-9659-z
11. Abe H, Kikuchi S, Hayakawa K, Iida T, Nagahashi N, Maeda K, et al. Discovery of a Highly Potent and Selective MEK Inhibitor: GSK1120212 (JTP-74057 DMSO Solvate). *ACS Med. Chem. Lett.* (2011) **2**:320–4. doi:10.1021/ml200004g

12. Flaherty KT, Infante JR, Daud A, Gonzalez R, Kefford RF, Sosman J, et al. Combined BRAF and MEK inhibition in melanoma with BRAF V600 mutations. *N Engl J Med* (2012) **367**:1694–703. doi:10.1056/NEJMoa1210093
13. Menzies AM, Long GV. Dabrafenib and trametinib, alone and in combination for BRAF-mutant metastatic melanoma. *Clin Cancer Res* (2014) **20**:2035–43. doi:10.1158/1078-0432.CCR-13-2054
14. Robert C, Karaszewska B, Schachter J, Rutkowski P, Mackiewicz A, Stroiakovski D, et al. Improved overall survival in melanoma with combined dabrafenib and trametinib. *N Engl J Med* (2015) **372**:30–9. doi:10.1056/NEJMoa1412690
15. Welsh SJ, Rizos H, Scolyer RA, Long GV. Resistance to combination BRAF and MEK inhibition in metastatic melanoma: Where to next? *Eur J Cancer* (2016) **62**:76–85. doi:10.1016/j.ejca.2016.04.005
16. Doebele RC, Drilon A, Paz-Ares L, Siena S, Shaw AT, Farago AF, et al. Entrectinib in patients with advanced or metastatic NTRK fusion-positive solid tumours: integrated analysis of three phase 1–2 trials. *The Lancet Oncology* (2020) **21**:271–82. doi:10.1016/S1470-2045(19)30691-6
17. Drilon AE, DuBois SG, Farago AF, Geoerger B, Grilley-Olson JE, Hong DS, et al. Activity of larotrectinib in TRK fusion cancer patients with brain metastases or primary central nervous system tumors. *JCO* (2019) **37**:2006. doi:10.1200/JCO.2019.37.15_suppl.2006
18. Drilon A, Laetsch TW, Kummar S, DuBois SG, Lassen UN, Demetri GD, et al. Efficacy of Larotrectinib in TRK Fusion-Positive Cancers in Adults and Children. *N Engl J Med* (2018) **378**:731–9. doi:10.1056/NEJMoa1714448
19. Hong DS, Kummar S, Farago AF, Lassen UN, Berlin J, Schilder RJ, et al. Larotrectinib efficacy and safety in adult TRK fusion cancer patients. *JCO* (2019) **37**:3122. doi:10.1200/JCO.2019.37.15_suppl.3122
20. Laetsch TW, DuBois SG, Mascarenhas L, Turpin B, Federman N, Albert CM, et al. Larotrectinib for paediatric solid tumours harbouring NTRK gene fusions: phase 1 results from a multicentre, open-label, phase 1/2 study. *The Lancet Oncology* (2018) **19**:705–14. doi:10.1016/S1470-2045(18)30119-0
21. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med* (2017) **9**:34. doi:10.1186/s13073-017-0424-2
22. Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, et al. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol Cancer Ther* (2017) **16**:2598–608. doi:10.1158/1535-7163.MCT-17-0386

23. Garofalo A, Sholl L, Reardon B, Taylor-Weiner A, Amin-Mansour A, Miao D, et al. The impact of tumor profiling approaches and genomic data strategies for cancer precision medicine. *Genome Med* (2016) **8**:79. doi:10.1186/s13073-016-0333-9
24. van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* (2015) **350**:207–11. doi:10.1126/science.aad0095
25. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* (2015) **348**:124–8. doi:10.1126/science.aaa1348
26. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med* (2014) **371**:2189–99. doi:10.1056/NEJMoa1406498
27. Buchhalter I, Rempel E, Endris V, Allgäuer M, Neumann O, Volckmar A-L, et al. Size matters: Dissecting key parameters for panel-based tumor mutational burden analysis. *Int J Cancer* (2019) **144**:848–58. doi:10.1002/ijc.31878
28. Le DT, Kim TW, van Cutsem E, Geva R, Jäger D, Hara H, et al. Phase II Open-Label Study of Pembrolizumab in Treatment-Refractory, Microsatellite Instability-High/Mismatch Repair-Deficient Metastatic Colorectal Cancer: KEYNOTE-164. *JCO* (2020) **38**:11–9. doi:10.1200/JCO.19.02107
29. Marabelle A, Le DT, Ascierto PA, Di Giacomo AM, Jesus-Acosta A de, Delord J-P, et al. Efficacy of Pembrolizumab in Patients With Noncolorectal High Microsatellite Instability/Mismatch Repair-Deficient Cancer: Results From the Phase II KEYNOTE-158 Study. *JCO* (2020) **38**:1–10. doi:10.1200/JCO.19.02105
30. Heisterkamp N, Groffen J. Molecular insights into the Philadelphia translocation. *Hematol Pathol* (1991) **5**:1–10. doi:Review
31. Groffen J, STEPHENSON J, Heisterkamp N, DEKLEIN A, BARTRAM C, GROSVELD G. Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. *Cell* (1984) **36**:93–9. doi:10.1016/0092-8674(84)90077-1
32. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* (2016) **127**:2391–405. doi:10.1182/blood-2016-03-643544
33. Robertson KD. DNA methylation and human disease. *Nat Rev Genet* (2005) **6**:597–610. doi:10.1038/nrg1655
34. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet* (2012) **13**:840–52. doi:10.1038/nrg3306

35. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* (2013) **10**:1213–8. doi:10.1038/nmeth.2688
36. Kruse F, Junker JP, van Oudenaarden A, Bakkers J. Tomo-seq: A method to obtain genome-wide expression data with spatial resolution. *Methods Cell Biol* (2016) **135**:299–307. doi:10.1016/bs.mcb.2016.01.006
37. Kanter I, Kalisky T. Single cell transcriptomics: methods and applications. *Front Oncol* (2015) **5**:53. doi:10.3389/fonc.2015.00053
38. Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep* (2015) **10**:1386–97. doi:10.1016/j.celrep.2015.02.001
39. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* (2015) **523**:486–90. doi:10.1038/nature14590
40. Bossel Ben-Moshe N, Gilad S, Perry G, Benjamin S, Balint-Lahat N, Pavlovsky A, et al. mRNA-sequencing whole transcriptome profiling of fresh frozen versus archived fixed tissues. *BMC Genomics* (2018) **19**:419. doi:10.1186/s12864-018-4761-3
41. Gao XH, Li J, Gong HF, Yu GY, Liu P, Hao LQ, et al. Comparison of Fresh Frozen Tissue With Formalin-Fixed Paraffin-Embedded Tissue for Mutation Analysis Using a Multi-Gene Panel in Patients With Colorectal Cancer. *Front. Oncol.* (2020) **10**:310. doi:10.3389/fonc.2020.00310
42. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* (2012) **2**:401–4. doi:10.1158/2159-8290.CD-12-0095
43. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* (2013) **6**:pl1. doi:10.1126/scisignal.2004088
44. Papaemmanuil E, Gerstung M, Bullinger L, Gaidzik VI, Paschka P, Roberts ND, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med* (2016) **374**:2209–21. doi:10.1056/NEJMoa1516192
45. Larkin JM, Ribas A, Flaherty K, McArthur GA, Ascierto PA, Dréno B, et al. Identifying prognostic subgroups for outcomes in BRAFV600 -mutated metastatic melanoma patients (pts) treated with vemurafenib (V) ± cobimetinib (C): A pooled analysis of BRIM-2, BRIM-3, BRIM-7 and coBRIM. *JCO* (2016) **34**:9536. doi:10.1200/JCO.2016.34.15_suppl.9536

46. Mateo J, Carreira S, Sandhu S, Miranda S, Mossop H, Perez-Lopez R, et al. DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. *N Engl J Med* (2015) **373**:1697–708. doi:10.1056/NEJMoa1506859
47. Atchley DP, Albarracin CT, Lopez A, Valero V, Amos CI, Gonzalez-Angulo AM, et al. Clinical and pathologic characteristics of patients with BRCA-positive and BRCA-negative breast cancer. *JCO* (2008) **26**:4282–8. doi:10.1200/JCO.2008.16.6231
48. Gelsomino F, Barbolini M, Spallanzani A, Pugliese G, Cascinu S. The evolving role of microsatellite instability in colorectal cancer: A review. *Cancer Treatment Reviews* (2016) **51**:19–26. doi:10.1016/j.ctrv.2016.10.005
49. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med* (2004) **350**:2129–39. doi:10.1056/NEJMoa040938
50. Vanella V, Festino L, Trojaniello C, Vitale MG, Sorrentino A, Paone M, et al. The Role of BRAF-Targeted Therapy for Advanced Melanoma in the Immunotherapy Era. *Curr Oncol Rep* (2019) **21**:76. doi:10.1007/s11912-019-0827-x
51. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* (2013) **499**:43–9. doi:10.1038/nature12222
52. Ciccarese C, Brunelli M, Montironi R, Fiorentino M, Iacovelli R, Heng D, et al. The prospect of precision therapy for renal cell carcinoma. *Cancer Treatment Reviews* (2016) **49**:37–44. doi:10.1016/j.ctrv.2016.07.003
53. Motzer RJ, Russo P. Systemic therapy for renal cell carcinoma. *J Urol* (2000) **163**:408–17. doi:Review
54. Ahrens M, Scheich S, Hartmann A, Bergmann L. Non-Clear Cell Renal Cell Carcinoma - Pathology and Treatment Options. *Oncol Res Treat* (2019) **42**:128–35. doi:10.1159/000495366
55. Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* (2014) **26**:319–30. doi:10.1016/j.ccr.2014.07.014
56. Kovacs A, Storkel S, Thoenes W, Kovacs G. Mitochondrial and chromosomal DNA alterations in human chromophobe renal cell carcinomas. *J. Pathol.* (1992) **167**:273–7. doi:10.1002/path.1711670303
57. Nagy A, Wilhelm M, Sükösd F, Ljungberg B, Kovacs G. Somatic mitochondrial DNA mutations in human chromophobe renal cell carcinomas. *Genes Chromosom. Cancer* (2002) **35**:256–60. doi:10.1002/gcc.10118
58. Moch H, Cubilla AL, Humphrey PA, Reuter VE, Ulbright TM. The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs-Part A: Renal, Penile, and

Testicular Tumours. *European Urology* (2016) **70**:93–105. doi:10.1016/j.eururo.2016.02.029

59. Zhou H, Zheng S, Truong LD, Ro JY, Ayala AG, Shen SS. Clear cell papillary renal cell carcinoma is the fourth most common histologic type of renal cell carcinoma in 290 consecutive nephrectomies for renal cell carcinoma. *Human Pathology* (2014) **45**:59–64. doi:10.1016/j.humpath.2013.08.004
60. Chan S, Siegel EL. Will machine learning end the viability of radiology as a thriving medical specialty? *Br J Radiol* (2019) **92**:20180416. doi:10.1259/bjr.20180416
61. Chockley K, Emanuel E. The End of Radiology? Three Threats to the Future Practice of Radiology. *J Am Coll Radiol* (2016) **13**:1415–20. doi:10.1016/j.jacr.2016.07.010
62. Moran S, Martínez-Cardús A, Sayols S, Musulén E, Balañá C, Estival-Gonzalez A, et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *The Lancet Oncology* (2016) **17**:1386–95. doi:10.1016/S1470-2045(16)30297-2
63. Jurmeister P, Schöler A, Arnold A, Klauschen F, Lenze D, Hummel M, et al. DNA methylation profiling reliably distinguishes pulmonary enteric adenocarcinoma from metastatic colorectal cancer. *Mod Pathol* (2019) **32**:855–65. doi:10.1038/s41379-019-0207-y
64. Luo R, Song J, Xiao X, Xie Z, Zhao Z, Zhang W, et al. Identifying CpG methylation signature as a promising biomarker for recurrence and immunotherapy in non-small-cell lung carcinoma. *Aging (Albany NY)* (2020) **12**:14649–76. doi:10.18632/aging.103517
65. Zhao Y, Pan Z, Namburi S, Pattison A, Posner A, Balachander S, et al. CUP-AI-Dx: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. *EBioMedicine* (2020) **61**:103030. doi:10.1016/j.ebiom.2020.103030
66. Søndergaard D, Nielsen S, Pedersen CN, Besenbacher S. Prediction of Primary Tumors in Cancers of Unknown Primary. *Journal of Integrative Bioinformatics* (2017) **14**. doi:10.1515/jib-2017-0013
67. Bhowmick SS, Bhattacharjee D, Rato L. Identification of tissue-specific tumor biomarker using different optimization algorithms. *Genes Genom* (2019) **41**:431–43. doi:10.1007/s13258-018-0773-2
68. Daoud M, Mayo M. A survey of neural network-based cancer prediction models from microarray data. *Artificial Intelligence in Medicine* (2019) **97**:204–14. doi:10.1016/j.artmed.2019.01.006
69. Dwivedi AK. Artificial neural network model for effective cancer classification using microarray gene expression data. *Neural Comput & Applic* (2018) **29**:1545–54. doi:10.1007/s00521-016-2701-1

70. Long NP, Park S, Anh NH, Nghi TD, Yoon SJ, Park JH, et al. High-Throughput Omics and Statistical Learning Integration for the Discovery and Validation of Novel Diagnostic Signatures in Colorectal Cancer. *IJMS* (2019) **20**:296. doi:10.3390/ijms20020296
71. Wei IH, Shi Y, Jiang H, Kumar-Sinha C, Chinnaiyan AM. RNA-sequencing accurately identifies cancer biomarker signatures to distinguish tissue of origin. *Neoplasia* (2014) **16**:918–27. doi:10.1016/j.neo.2014.09.007
72. Xiao Y, Wu J, Lin Z, Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine* (2018) **153**:1–9. doi:10.1016/j.cmpb.2017.09.005
73. Xiao Y, Wu J, Lin Z, Zhao X. A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-sequencing data. *Computer Methods and Programs in Biomedicine* (2018) **166**:99–105. doi:10.1016/j.cmpb.2018.10.004
74. Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, et al. The International Cancer Genome Consortium Data Portal. *Nat Biotechnol* (2019) **37**:367–9. doi:10.1038/s41587-019-0055-9
75. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Research* (2013) **41**:D991-5. doi:10.1093/nar/gks1193
76. Tekpli X, Lien T, Røssevold AH, Nebdal D, Borgen E, Ohnstad HO, et al. An independent poor-prognosis subtype of breast cancer defined by a distinct tumor immune microenvironment. *Nat Commun* (2019) **10**:5499. doi:10.1038/s41467-019-13329-5
77. Jiang Y, Sun A, Zhao Y, Ying W, Sun H, Yang X, et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* (2019) **567**:257–61. doi:10.1038/s41586-019-0987-8
78. Zeng Z, Cao Z, Tang Y. Identification of diagnostic and prognostic biomarkers, and candidate targeted agents for hepatitis B virus-associated early stage hepatocellular carcinoma based on RNA-sequencing data. *Oncol Lett* (2020) **20**:231. doi:10.3892/ol.2020.12094
79. Li S, Garrett-Bakelman FE, Chung SS, Sanders MA, Hricik T, Rapaport F, et al. Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat Med* (2016) **22**:792–9. doi:10.1038/nm.4125
80. Mann JE, Kulkarni A, Birkeland AC, Kafelghazal J, Eisenberg J, Jewell BM, et al. The molecular landscape of the University of Michigan laryngeal squamous cell carcinoma cell line panel. *Head & Neck* (2019) **41**:3114–24. doi:10.1002/hed.25803

81. Ma Y-S, Huang T, Zhong X-M, Zhang H-W, Cong X-L, Xu H, et al. Proteogenomic characterization and comprehensive integrative genomic analysis of human colorectal cancer liver metastasis. *Mol Cancer* (2018) **17**:139. doi:10.1186/s12943-018-0890-1
82. Scarlett CJ, Salisbury EL, Biankin AV, Kench J. Precursor lesions in pancreatic cancer: morphological and molecular pathology. *Pathology* (2011) **43**:183–200. doi:10.1097/PAT.0b013e3283445e3a
83. Rehan Akbani, Kadir C. Akdemir, B. Arman Aksoy, Monique Albert, Adrian Ally, Samirkumar B. Amin, et al. Genomic Classification of Cutaneous Melanoma. *Cell* (2015) **161**:1681–96. doi:10.1016/j.cell.2015.05.044
84. Abida W, Cyrta J, Heller G, Prandi D, Armenia J, Coleman I, et al. Genomic correlates of clinical outcome in advanced prostate cancer. *Proc Natl Acad Sci USA* (2019) **116**:11428–36. doi:10.1073/pnas.1902651116
85. Beltran H, Prandi D, Mosquera JM, Benelli M, Puca L, Cyrta J, et al. Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat Med* (2016) **22**:298–305. doi:10.1038/nm.4045
86. Crooks DR, Maio N, Lang M, Ricketts CJ, Vocke CD, Gurram S, et al. Mitochondrial DNA alterations underlie an irreversible shift to aerobic glycolysis in fumarate hydratase-deficient renal cancer. *Sci. Signal.* (2021) **14**. doi:10.1126/scisignal.abc4436
87. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* (2020) **17**:261–72. doi:10.1038/s41592-019-0686-2
88. *Scikit-learn: Machine learning in Python* (2011).
89. Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association* (1952) **47**:583–621. doi:10.1080/01621459.1952.10483441
90. Cameron Davidson-Pilon, Jonas Kalderstam, Paul Zivich, Ben Kuhn, Andrew Fiore-Gartland, AbdealiJK, et al. *CamDavidsonPilon/lifelines: v0.23.1*. Zenodo (2019).
91. Kotsiantis SB. Decision trees: a recent overview. *Artif Intell Rev* (2013) **39**:261–83. doi:10.1007/s10462-011-9272-4
92. Breiman L. *Machine Learning* (2001) **45**:5–32. doi:10.1023/A:1010933404324
93. *Visualizing data using t-SNE* (2008).
94. Marquardt A, Solimando AG, Kerscher A, Bittrich M, Kalogirou C, Kübler H, et al. Subgroup-Independent Mapping of Renal Cell Carcinoma-Machine Learning Reveals Prognostic Mitochondrial Gene Signature Beyond Histopathologic Boundaries. *Front. Oncol.* (2021) **11**:621278. doi:10.3389/fonc.2021.621278
95. Kullback S, Leibler RA. On Information and Sufficiency. *Ann. Math. Statist.* (1951) **22**:79–86. doi:10.1214/aoms/1177729694

96. McInnes L, Healy J, Melville J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (2018).
97. *Performance Comparison of Dimension Reduction Implementations — umap 0.5 documentation* (2022) [cited 2022 Jun 01]. Available from: https://umap-learn.readthedocs.io/en/latest/benchmarking.html
98. Marquardt A, Landwehr L-S, Ronchi CL, Di Dalmazi G, Riester A, Kollmannsberger P, et al. Identifying New Potential Biomarkers in Adrenocortical Tumors Based on mRNA Expression Data Using Machine Learning. *Cancers* (2021) **13**:4671. doi:10.3390/cancers13184671
99. Liu H, Shi J, Wilkerson ML, Lin F. Immunohistochemical evaluation of GATA3 expression in tumors and normal tissues: a useful immunomarker for breast and urothelial carcinomas. *Am J Clin Pathol* (2012) **138**:57–64. doi:10.1309/AJCP5UAFMSA9ZQBZ
100. Peng Y, Butt YM, Chen B, Zhang X, Tang P. Update on Immunohistochemical Analysis in Breast Lesions. *Arch Pathol Lab Med* (2017) **141**:1033–51. doi:10.5858/arpa.2016-0482-RA
101. Ritchie ME, Phipson B, Di Wu, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* (2015) **43**:e47. doi:10.1093/nar/gkv007
102. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* (2007) **8**:118–27. doi:10.1093/biostatistics/kxj037
103. Haghverdi L, Lun AT, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* (2018) **36**:421–7. doi:10.1038/nbt.4091
104. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-sequencing data. *Nat Commun* (2018) **9**:284. doi:10.1038/s41467-017-02554-5
105. Marquardt A, Kollmannsberger P, Krebs M, Argentiero A, Knott M, Solimando AG, et al. Visual Clustering of Transcriptomic Data from Primary and Metastatic Tumors-Dependencies and Novel Pitfalls. *Genes (Basel)* (2022) **13**. doi:10.3390/genes13081335
106. Fuhrmann DC, Brüne B. Mitochondrial composition and function under the control of hypoxia. *Redox Biol* (2017) **12**:208–15. doi:10.1016/j.redox.2017.02.012
107. Al Tameemi W, Dale TP, Al-Jumaily RM, Forsyth NR. Hypoxia-Modified Cancer Cell Metabolism. *Front Cell Dev Biol* (2019) **7**:4. doi:10.3389/fcell.2019.00004
108. Simonnet H, Alazard N, Pfeiffer K, Gallou C, Béroud C, Demont J, et al. Low mitochondrial respiratory chain content correlates with tumor aggressiveness in renal cell carcinoma. *Carcinogenesis* (2002) **23**:759–68. doi:10.1093/carcin/23.5.759

109. Mahin KF, Robiuddin M, Islam M, Ashraf S, Yeasmin F, Shatabda S. PanClassif: Improving pan cancer classification of single cell RNA-sequencing gene expression data using machine learning. *Genomics* (2022) **114**:110264. doi:10.1016/j.ygeno.2022.01.001

110. Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, et al. Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. *Cancer Cell* (2016) **29**:723–36. doi:10.1016/j.ccell.2016.04.002

111. Wu F, Fan J, He Y, Xiong A, Yu J, Li Y, et al. Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat Commun* (2021) **12**:2540. doi:10.1038/s41467-021-22801-0

112. González-Silva L, Quevedo L, Varela I. Tumor Functional Heterogeneity Unraveled by scRNA-sequencing Technologies. *Trends in Cancer* (2020) **6**:13–9. doi:10.1016/j.trecan.2019.11.010

113. Yamawaki TM, Lu DR, Ellwanger DC, Bhatt D, Manzanillo P, Arias V, et al. Systematic comparison of high-throughput single-cell RNA-sequencing methods for immune cell profiling. *BMC Genomics* (2021) **22**:66. doi:10.1186/s12864-020-07358-4

114. Yesudhas D, Dharshini SA, Taguchi Y-H, Gromiha MM. Tumor Heterogeneity and Molecular Characteristics of Glioblastoma Revealed by Single-Cell RNA-sequencing Data Analysis. *Genes (Basel)* (2022) **13**. doi:10.3390/genes13030428

115. Bischoff P, Trinks A, Obermayer B, Pett JP, Wiederspahn J, Uhlitz F, et al. Single-cell RNA-sequencing reveals distinct tumor microenvironmental patterns in lung adenocarcinoma. *Oncogene* (2021) **40**:6748–58. doi:10.1038/s41388-021-02054-3

116. Liu Y, Feng W, Dai Y, Bao M, Yuan Z, He M, et al. Single-Cell Transcriptomics Reveals the Complexity of the Tumor Microenvironment of Treatment-Naive Osteosarcoma. *Front Oncol* (2021) **11**:709210. doi:10.3389/fonc.2021.709210

117. Fustero-Torre C, Jiménez-Santos MJ, García-Martín S, Carretero-Puche C, García-Jimeno L, Ivanchuk V, et al. Beyondcell: targeting cancer therapeutic heterogeneity in single-cell RNA-sequencing data. *Genome Med* (2021) **13**:187. doi:10.1186/s13073-021-01001-x

118. Goldman SL, MacKay M, Afshinnekoo E, Melnick AM, Wu S, Mason CE. The Impact of Heterogeneity on Single-Cell Sequencing. *Front Genet* (2019) **10**:8. doi:10.3389/fgene.2019.00008

119. Mustachio LM, Roszik J. Single-Cell Sequencing: Current Applications in Precision Onco-Genomics and Cancer Therapeutics. *Cancers* (2022) **14**. doi:10.3390/cancers14030657

120. Dohmen J, Baranovskii A, Ronen J, Uyar B, Franke V, Akalin A. Identifying tumor cells at the single-cell level using machine learning. *Genome Biol* (2022) **23**:123. doi:10.1186/s13059-022-02683-1

121. Lyu B, Haque A. “Deep Learning Based Tumor Type Classification Using Gene Expression Data,”. In: Shehu A, editor. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. New York, NY: ACM (2018). p. 89–96.

122. Grewal JK, Tessier-Cloutier B, Jones M, Gakkhar S, Ma Y, Moore R, et al. Application of a Neural Network Whole Transcriptome-Based Pan-Cancer Method for Diagnosis of Primary and Metastatic Cancers. *JAMA Netw Open* (2019) **2**:e192597. doi:10.1001/jamanetworkopen.2019.2597

123. Cascianelli S, Molineris I, Isella C, Masseroli M, Medico E. Machine learning for RNA-sequencing-based intrinsic subtyping of breast cancer. *Sci Rep* (2020) **10**:14071. doi:10.1038/s41598-020-70832-2

124. Kuksin M, Morel D, Aglave M, Danlos F-X, Marabelle A, Zinovyev A, et al. Applications of single-cell and bulk RNA-sequencing in onco-immunology. *Eur J Cancer* (2021) **149**:193–210. doi:10.1016/j.ejca.2021.03.005

125. Li Y, Kang K, Krahn JM, Croutwater N, Lee K, Umbach DM, et al. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC Genomics* (2017) **18**:508. doi:10.1186/s12864-017-3906-0

126. Laird PW. Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet* (2010) **11**:191–203. doi:10.1038/nrg2732

127. Kim K-Y, Kim SH, Ki DH, Jeong J, Jeong HJ, Jeung H-C, et al. An attempt for combining microarray data sets by adjusting gene expressions. *Cancer Res Treat* (2007) **39**:74–81. doi:10.4143/crt.2007.39.2.74

128. Groskopf J, Aubin SM, Deras IL, Blase A, Bodrug S, Clark C, et al. APTIMA PCA3 molecular urine test: development of a method to aid in the diagnosis of prostate cancer. *Clin Chem* (2006) **52**:1089–95. doi:10.1373/clinchem.2005.063289

129. d'Ydewalle C, Ramos DM, Pyles NJ, Ng S-Y, Gorz M, Pilato CM, et al. The Antisense Transcript SMN-AS1 Regulates SMN Expression and Is a Novel Therapeutic Target for Spinal Muscular Atrophy. *Neuron* (2017) **93**:66–79. doi:10.1016/j.neuron.2016.11.033

130. Woo CJ, Maier VK, Davey R, Brennan J, Li G, Brothers J, et al. Gene activation of SMN by selective disruption of lncRNA-mediated recruitment of PRC2 for the treatment of spinal muscular atrophy. *Proc Natl Acad Sci USA* (2017) **114**:E1509-E1518. doi:10.1073/pnas.1616521114

131. Meng L, Ward AJ, Chun S, Bennett CF, Beaudet AL, Rigo F. Towards a therapy for Angelman syndrome by targeting a long non-coding RNA. *Nature* (2015) **518**:409–12. doi:10.1038/nature13975

132. Hsiao J, Yuan TY, Tsai MS, Lu CY, Lin YC, Lee ML, et al. Upregulation of Haploinsufficient Gene Expression in the Brain by Targeting a Long Non-coding RNA

Improves Seizure Phenotype in a Model of Dravet Syndrome. *EBioMedicine* (2016) **9**:257–77. doi:10.1016/j.ebiom.2016.05.011

133. Carlevaro-Fita J, Lanzós A, Feuerbach L, Hong C, Mas-Ponte D, Pedersen JS, et al. Cancer LncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. *Commun Biol* (2020) **3**:56. doi:10.1038/s42003-019-0741-7

134. Endo H, Shiroki T, Nakagawa T, Yokoyama M, Tamai K, Yamanami H, et al. Enhanced expression of long non-coding RNA HOTAIR is associated with the development of gastric cancer. *PLoS ONE* (2013) **8**:e77070. doi:10.1371/journal.pone.0077070

135. Okugawa Y, Toiyama Y, Hur K, Toden S, Saigusa S, Tanaka K, et al. Metastasis-associated long non-coding RNA drives gastric cancer development and promotes peritoneal metastasis. *Carcinogenesis* (2014) **35**:2731–9. doi:10.1093/carcin/bgu200

136. Cairns J. Mutation selection and the natural history of cancer. *Nature* (1975) **255**:197–200. doi:10.1038/255197a0

137. Klein A, Olendrowitz C, Schmutzler R, Hampl J, Schlag PM, Maass N, et al. Identification of brain- and bone-specific breast cancer metastasis genes. *Cancer Letters* (2009) **276**:212–20. doi:10.1016/j.canlet.2008.11.017

138. Lin W, Noel P, Borazanci EH, Lee J, Amini A, Han IW, et al. Single-cell transcriptome analysis of tumor and stromal compartments of pancreatic ductal adenocarcinoma primary tumors and metastatic lesions. *Genome Med* (2020) **12**:80. doi:10.1186/s13073-020-00776-9

139. Pan H, Diao H, Zhong W, Wang T, Wen P, Wu C. A Cancer Cell Cluster Marked by LincRNA MEG3 Leads Pancreatic Ductal Adenocarcinoma Metastasis. *Front. Oncol.* (2021) **11**:656564. doi:10.3389/fonc.2021.656564

140. Li Y, Hu X, Lin R, Zhou G, Zhao L, Zhao D, et al. Single-cell landscape reveals active cell subtypes and their interaction in the tumor microenvironment of gastric cancer. *Theranostics* (2022) **12**:3818–33. doi:10.7150/thno.71833

141. Gan Y, Li N, Zou G, Xin Y, Guan J. Identification of cancer subtypes from single-cell RNA-sequencing data using a consensus clustering method. *BMC Med Genomics* (2018) **11**:117. doi:10.1186/s12920-018-0433-z

142. Xu K, Wang R, Xie H, Hu L, Wang C, Xu J, et al. Single-cell RNA-sequencing reveals cell heterogeneity and transcriptome profile of breast cancer lymph node metastasis. *Oncogenesis* (2021) **10**:66. doi:10.1038/s41389-021-00355-6

143. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA-sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* (2020) **11**:2285. doi:10.1038/s41467-020-16164-1

144. Argentiero A, Solimando AG, Krebs M, Leone P, Susca N, Brunetti O, et al. Anti-angiogenesis and Immunotherapy: Novel Paradigms to Envision Tailored Approaches in Renal Cell-Carcinoma. *J Clin Med* (2020) **9**. doi:10.3390/jcm9051594

145. Bedke J, Albiges L, Capitanio U, Giles RH, Hora M, Lam TB, et al. The 2021 Updated European Association of Urology Guidelines on Renal Cell Carcinoma: Immune Checkpoint Inhibitor-based Combination Therapies for Treatment-naive Metastatic Clear-cell Renal Cell Carcinoma Are Standard of Care. *European Urology* (2021) **80**:393–7. doi:10.1016/j.eururo.2021.04.042

146. Bedke J, Albiges L, Capitanio U, Giles RH, Hora M, Lam TB, et al. Updated European Association of Urology Guidelines on Renal Cell Carcinoma: Nivolumab plus Cabozantinib Joins Immune Checkpoint Inhibition Combination Therapies for Treatment-naïve Metastatic Clear-Cell Renal Cell Carcinoma. *European Urology* (2021) **79**:339–42. doi:10.1016/j.eururo.2020.12.005

147. Fiala O, Ostašov P, Rozsypalová A, Hora M, Šorejs O, Šustr J, et al. Metformin Use and the Outcome of Metastatic Renal Cell Carcinoma Treated with Sunitinib or Pazopanib. *Cancer Manag Res* (2021) **13**:4077–86. doi:10.2147/CMAR.S305321

148. Amin S, Lux A, O'Callaghan F. The journey of metformin from glycaemic control to mTOR inhibition and the suppression of tumour growth. *Br J Clin Pharmacol* (2019) **85**:37–46. doi:10.1111/bcp.13780

149. Liu Y, Li J, Song M, Qi G, Meng L. High-Concentration Metformin Reduces Oxidative Stress Injury and Inhibits the Growth and Migration of Clear Cell Renal Cell Carcinoma. *Comput Math Methods Med* (2022) **2022**:1466991. doi:10.1155/2022/1466991

150. Song A, Zhang C, Meng X. Mechanism and application of metformin in kidney diseases: An update. *Biomed Pharmacother* (2021) **138**:111454. doi:10.1016/j.biopha.2021.111454

151. Schmid P, Adams S, Rugo HS, Schneeweiss A, Barrios CH, Iwata H, et al. Atezolizumab and Nab-Paclitaxel in Advanced Triple-Negative Breast Cancer. *N Engl J Med* (2018) **379**:2108–21. doi:10.1056/NEJMoa1809615

152. Nishino M, Ramaiya NH, Hatabu H, Hodi FS. Monitoring immune-checkpoint blockade: response evaluation and biomarker development. *Nat Rev Clin Oncol* (2017) **14**:655–68. doi:10.1038/nrclinonc.2017.88

153. Lei Y, Li X, Huang Q, Zheng X, Liu M. Progress and Challenges of Predictive Biomarkers for Immune Checkpoint Blockade. *Front Oncol* (2021) **11**:617335. doi:10.3389/fonc.2021.617335

154. Garon EB, Rizvi NA, Hui R, Leighl N, Balmanoukian AS, Eder JP, et al. Pembrolizumab for the treatment of non-small-cell lung cancer. *N Engl J Med* (2015) **372**:2018–28. doi:10.1056/NEJMoa1501824

i. SUPPLEMENTS

ENSG Identifier	Mean Position All Models	Position Best Model	Both in Top200 Genes
ENSG00000198804	1	3	Yes
ENSG00000198886	2	2	Yes
ENSG00000198712	3	5	Yes
ENSG00000212907	4	6	Yes
ENSG00000198938	5	4	Yes
ENSG00000134954	6	13	Yes
ENSG00000198899	7	7	Yes
ENSG00000248527	8	9	Yes
ENSG00000116062	9	67	Yes
ENSG00000211459	10	8	Yes
ENSG00000237973	11	69	Yes
ENSG00000158097	12	168	Yes
ENSG00000131473	13	-	No
ENSG00000186063	14	-	No
ENSG00000113580	15	31	Yes
ENSG00000198840	16	24	Yes
ENSG00000112715	17	112	Yes
ENSG00000210082	18	25	Yes
ENSG00000172845	19	15	Yes
ENSG00000166068	20	103	Yes
ENSG00000198763	21	20	Yes
ENSG00000135766	22	-	No
ENSG00000124795	23	-	No
ENSG00000265241	24	70	Yes
ENSG00000171488	25	136	Yes
ENSG00000136758	26	66	Yes
ENSG00000228253	27	96	Yes
ENSG00000138071	28	132	Yes
ENSG00000097033	29	32	Yes
ENSG00000154727	30	92	Yes
ENSG00000171388	31	37	Yes
ENSG00000009307	32	106	Yes

ENSG00000229344	33	17	Yes
ENSG00000122884	34	19	Yes
ENSG00000167772	35	138	Yes
ENSG00000172469	36	-	No
ENSG00000261371	37	-	No
ENSG00000145817	38	53	Yes
ENSG00000111252	39	26	Yes
ENSG00000267325	40	64	Yes
ENSG00000151617	41	-	No
ENSG00000275052	42	-	No
ENSG00000225630	43	47	Yes
ENSG00000120727	44	-	No
ENSG00000184014	45	176	Yes
ENSG00000143751	46	29	Yes
ENSG00000115232	47	151	Yes
ENSG00000095139	48	-	No
ENSG00000116747	49	41	Yes
ENSG00000164105	50	-	No
ENSG00000182827	51	-	No
ENSG00000166478	52	-	No
ENSG00000143756	53	-	No
ENSG00000085449	54	192	Yes
ENSG00000198744	55	72	Yes
ENSG00000130741	56	147	Yes
ENSG00000273585	57	48	Yes
ENSG00000162852	58	-	No
ENSG00000150687	59	51	Yes
ENSG00000173198	60	199	Yes
ENSG00000102755	61	38	Yes
ENSG00000143319	62	154	Yes
ENSG00000065135	63	78	Yes
ENSG00000165732	64	-	No
ENSG00000121774	65	18	Yes
ENSG00000135913	66	-	No
ENSG00000165806	67	-	No
ENSG00000134352	68	-	No

ENSG00000059804	69	-	No
ENSG00000161267	70	197	Yes
ENSG00000178175	71	-	No
ENSG00000135829	72	87	Yes
ENSG00000064989	73	177	Yes
ENSG00000198833	74	-	No
ENSG00000136643	75	-	No
ENSG00000107625	76	49	Yes
ENSG00000121481	77	-	No
ENSG00000170906	78	-	No
ENSG00000099250	79	56	Yes
ENSG00000111142	80	173	Yes
ENSG00000155380	81	-	No
ENSG00000132824	82	16	Yes
ENSG00000115504	83	-	No
ENSG00000102580	84	81	Yes
ENSG00000151151	85	-	No
ENSG00000055332	86	-	No
ENSG00000118946	87	150	Yes
ENSG00000139350	88	-	No
ENSG00000182963	89	-	No
ENSG00000088205	90	185	Yes
ENSG00000075945	91	-	No
ENSG00000099942	92	46	Yes
ENSG00000150540	93	-	No
ENSG00000181744	94	-	No
ENSG00000168675	95	-	No
ENSG00000221818	96	-	No
ENSG00000083642	97	-	No
ENSG00000165119	98	39	Yes
ENSG00000139436	99	68	Yes
ENSG00000158290	100	-	No
ENSG00000077514	101	-	No
ENSG00000120063	102	-	No
ENSG00000171862	103	109	Yes
ENSG00000152193	104	-	No

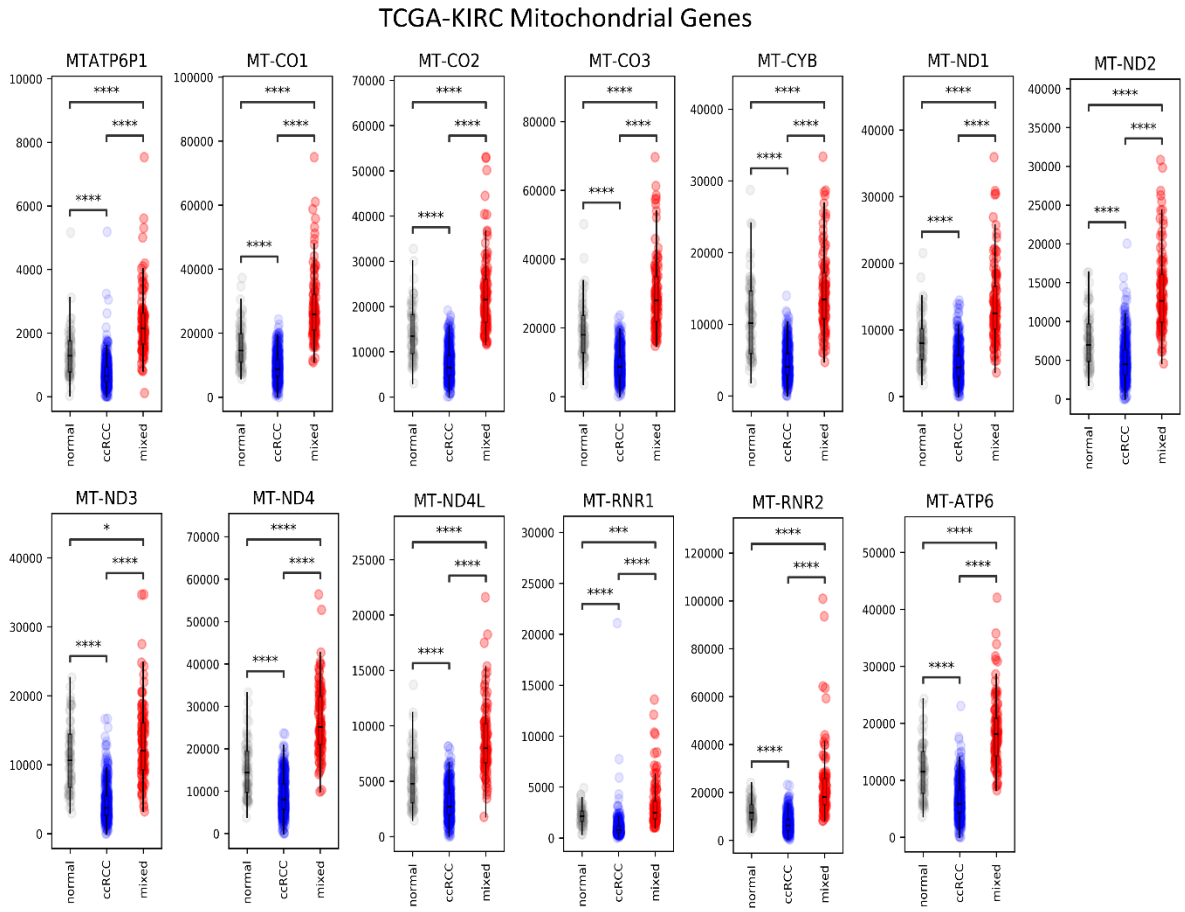
ENSG00000169860	105	40	Yes
ENSG00000198771	106	-	No
ENSG00000147113	107	-	No
ENSG00000185633	108	-	No
ENSG00000196369	109	152	Yes
ENSG00000196628	110	-	No
ENSG00000171159	111	-	No
ENSG00000135837	112	-	No
ENSG00000198700	113	-	No
ENSG00000117000	114	-	No
ENSG00000162654	115	-	No
ENSG00000116833	116	80	Yes
ENSG00000162607	117	89	Yes
ENSG00000128917	118	-	No
ENSG00000110330	119	-	No
ENSG00000133561	120	14	Yes
ENSG00000113739	121	108	Yes
ENSG00000057608	122	-	No
ENSG00000162636	123	-	No
ENSG00000174780	124	160	Yes
ENSG00000086589	125	-	No
ENSG00000076321	126	59	Yes
ENSG00000198888	127	10	Yes
ENSG00000197771	128	-	No
ENSG00000213516	129	-	No
ENSG00000134333	130	189	Yes
ENSG00000108061	131	-	No
ENSG00000112144	132	-	No
ENSG00000129521	133	-	No
ENSG00000102218	134	-	No
ENSG00000241468	135	-	No
ENSG00000082516	136	-	No
ENSG00000116473	137	-	No
ENSG00000095787	138	-	No
ENSG00000185009	139	-	No
ENSG00000185585	140	-	No

ENSG00000143344	141	-	No
ENSG00000147065	142	55	Yes
ENSG00000127463	143	-	No
ENSG00000177464	144	-	No
ENSG00000056097	145	-	No
ENSG00000004700	146	-	No
ENSG00000128708	147	-	No
ENSG00000125810	148	-	No
ENSG00000237686	149	-	No
ENSG00000134326	150	-	No
ENSG00000159167	151	-	No
ENSG00000213281	152	65	Yes
ENSG00000180776	153	163	Yes
ENSG00000131931	154	76	Yes
ENSG00000108100	155	-	No
ENSG00000182154	156	62	Yes
ENSG00000140479	157	-	No
ENSG00000081189	158	-	No
ENSG00000144118	159	-	No
ENSG00000115091	160	-	No
ENSG00000151702	161	104	Yes
ENSG00000115825	162	-	No
ENSG00000244462	163	-	No
ENSG00000145907	164	90	Yes
ENSG00000118260	165	-	No
ENSG00000076053	166	-	No
ENSG00000178057	167	-	No
ENSG00000211460	168	-	No
ENSG00000198786	169	36	Yes
ENSG00000164434	170	144	Yes
ENSG00000166265	171	-	No
ENSG00000143341	172	-	No
ENSG00000170989	173	187	Yes
ENSG00000125249	174	58	Yes
ENSG00000171208	175	-	No
ENSG00000133574	176	12	Yes

ENSG00000145860	177	-	No
ENSG00000113555	178	-	No
ENSG00000117597	179	-	No
ENSG00000164283	180	191	Yes
ENSG00000005812	181	-	No
ENSG00000117036	182	-	No
ENSG00000132024	183	-	No
ENSG00000182621	184	-	No
ENSG00000138031	185	-	No
ENSG00000155111	186	-	No
ENSG00000105289	187	50	Yes
ENSG00000164081	188	23	Yes
ENSG00000152683	189	143	Yes
ENSG00000143337	190	-	No
ENSG00000138134	191	-	No
ENSG00000160799	192	-	No
ENSG00000079308	193	-	No
ENSG00000157077	194	111	Yes
ENSG00000092964	195	100	Yes
ENSG00000013306	196	-	No
ENSG00000213250	197	-	No
ENSG00000107159	198	-	No
ENSG00000166578	199	-	No
ENSG00000204564	200	30	Yes

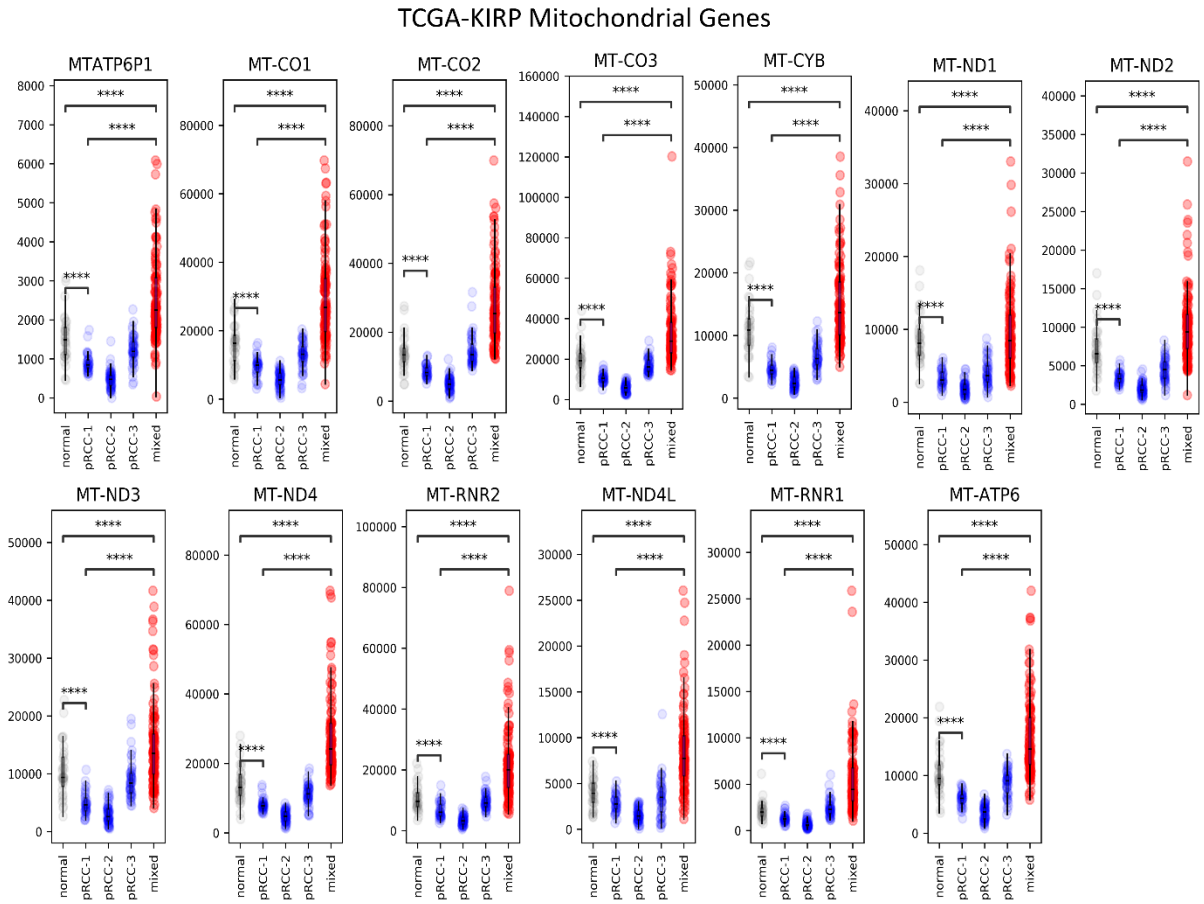
Supplementary Table 1:Random forest model comparison for subgroup prediction

Comparison for the 20 random forest models trained for the classification of the different subgroups for renal cell carcinomas, as identified by the visual clustering approach using t-SNE plot with unprocessed FPKM values. The table shows the top200 calculated genes as described for 19 of the 20 models compared to the top200 genes of the best performing model, including the ENSG Identifier, the mean position for the combined models, the position of the ENSG Identifier in the best performing model, and whether it is in the top200 for both or not.



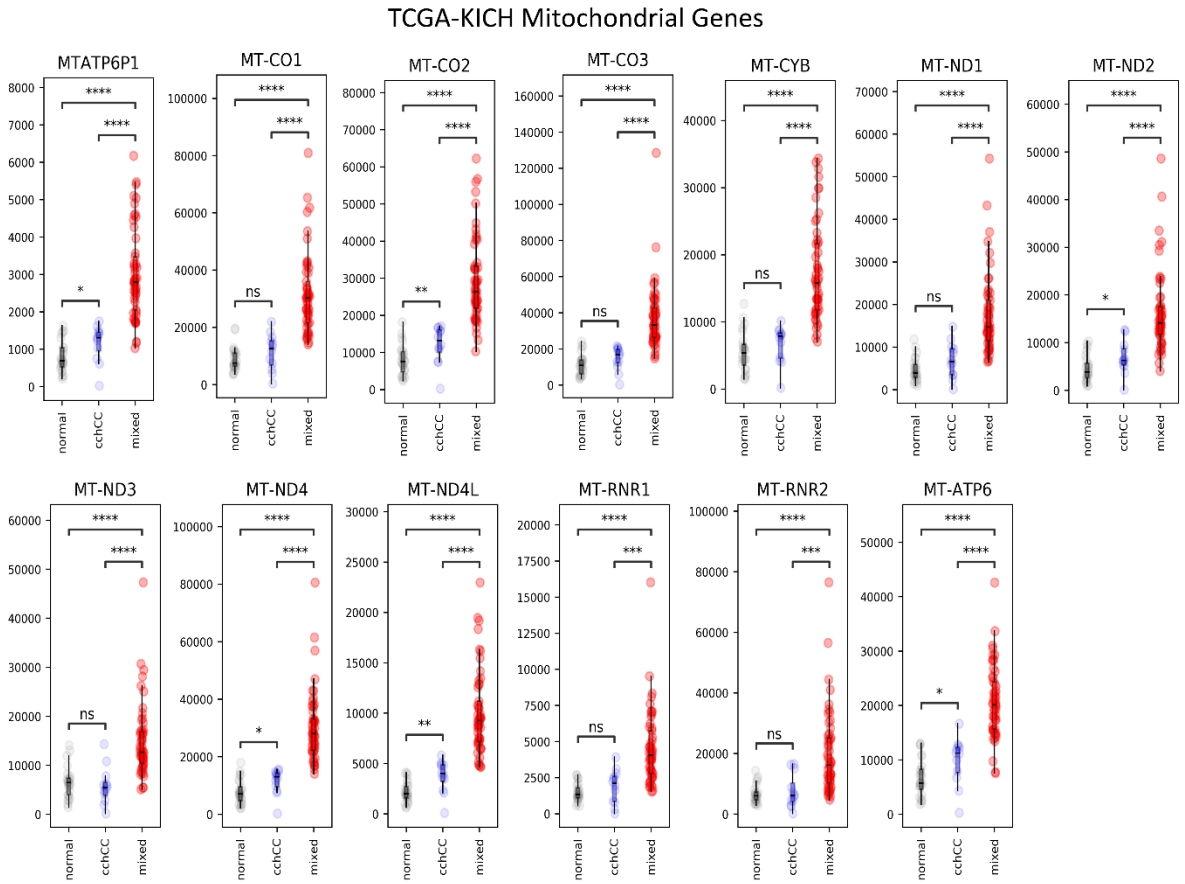
Supplementary Figure 1: Subgroup mRNA expression comparisons of mitochondrial genes for TCGA-KIRC cohort

*Expression comparison between clear cell renal cell carcinomas outside (ccRCC) and inside (mixed) the mixed subgroup and respective normal tissue samples for mitochondrial genes identified by machine learning. ns, not significant. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.*



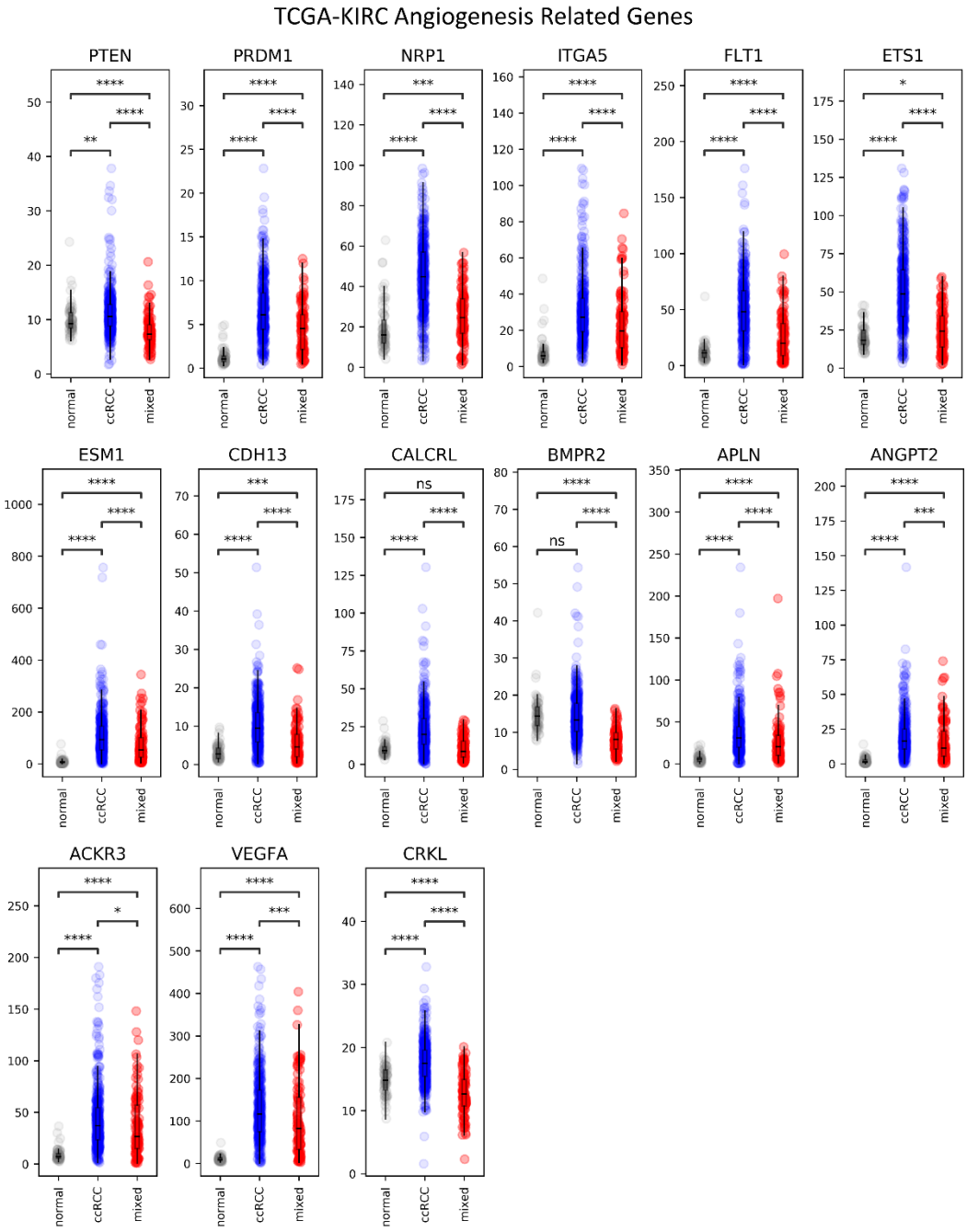
Supplementary Figure 2: Subgroup mRNA expression comparisons of mitochondrial genes for TCGA-KIRP cohort

*Expression comparison between the identified papillary renal cell carcinoma cluster outside (pRCC 1 to 3) and inside (mixed) the mixed subgroup and respective normal tissue samples for mitochondrial genes identified by machine learning. ns, not significant. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001. Figure taken from (94).*



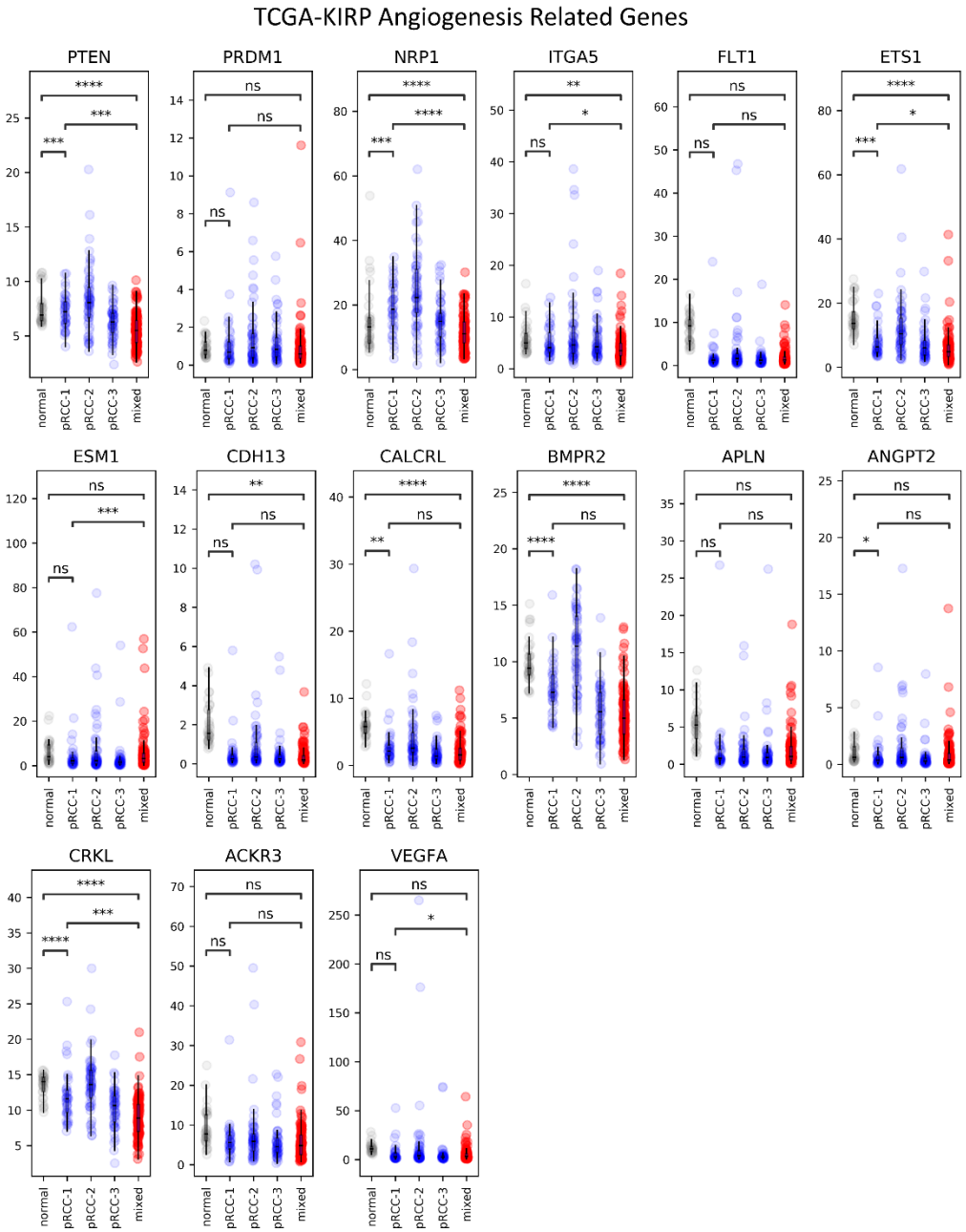
Supplementary Figure 3: Subgroup mRNA expression comparisons of mitochondrial genes for TCGA-KICH cohort

*Expression comparison between chromophobe renal cell carcinomas outside (chRCC) and inside (mixed) the mixed subgroup and respective normal tissue samples for mitochondrial genes identified by machine learning. ns, not significant. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001. Figure taken from (94).*



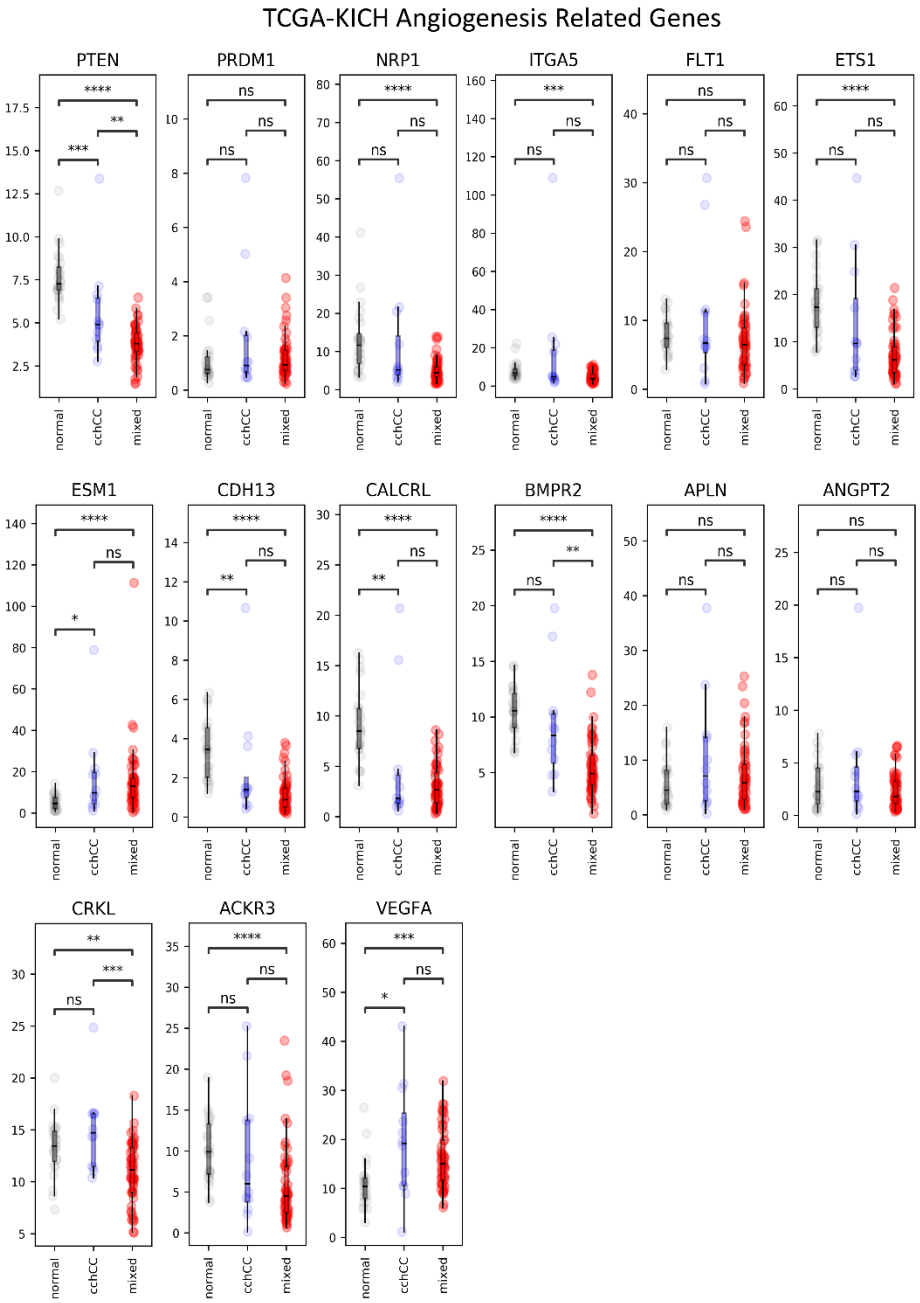
Supplementary Figure 4: Subgroup mRNA expression comparisons of angiogenesis related genes for TCGA-KIRC cohort

*Expression comparison between clear cell renal cell carcinomas outside (ccRCC) and inside (mixed) the mixed subgroup and respective normal tissue samples for angiogenesis genes identified by machine learning. ns, not significant. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001. Figure taken from (94).*



Supplementary Figure 5: Subgroup mRNA expression comparisons of angiogenesis related genes for TCGA-KIRP cohort

*Expression comparison between the identified papillary renal cell carcinoma cluster outside (pRCC 1 to 3) and inside (mixed) the mixed subgroup and respective normal tissue samples for angiogenesis genes identified by machine learning. ns, not significant. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001. Figure taken from (94).*



Supplementary Figure 6: Subgroup mRNA expression comparisons of angiogenesis related genes for TCGA-KICH cohort

*Expression comparison between chromophobe renal cell carcinomas outside (chRCC) and inside (mixed) of mixed subgroup and respective normal tissue samples for angiogenesis genes identified by machine learning. ns, not significant. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001 Figure taken from (94)..*

ii. APPENDIX

2.1 PUBLICATION LIST AND CONFERENCE CONTRIBUTIONS

2.1.1 Publications

1. **André Marquardt**, Antonio Giovanni Solimando, Alexander Kerscher, Max Bittrich, Charis Kalogirou, Hubert Kübler, Andreas Rosenwald, Ralf Bargou, Philip Kollmannsberger, Bastian Schilling, Svenja Meierjohann, Markus Krebs; Subgroup-Independent Mapping of Renal Cell Carcinoma-Machine Learning Reveals Prognostic Mitochondrial Gene Signature Beyond Histopathologic Boundaries, Front Oncol. 2021 Mar 15;11:621278, https://doi.org/10.3389/fonc.2021.621278

2. **André Marquardt**, Laura-Sophie Landwehr, Cristina L Ronchi, Guido Di Dalmazi, Anna Riester, Philip Kollmannsberger, Barbara Altieri, Martin Fassnacht, Silviu Sbiera; Identifying New Potential Biomarkers in Adrenocortical Tumors Based on mRNA Expression Data Using Machine Learning, Cancers 2021, 13, 4671, https://doi.org/10.3390/cancers13184671

3. **André Marquardt**, Philip Kollmannsberger, Markus Krebs, Antonella Argentiero, Markus Knott, Antonio Giovanni Solimando, Alexander Kerscher; Visual Clustering of Transcriptomic Data from Primary and Metastatic Tumors – Dependencies and Novel Pitfalls; Genes 13, no. 8: 1335. https://doi.org/10.3390/genes13081335

4. **André Marquardt**, Philipp Hartrampf, Philip Kollmannsberger, Antonio G Solimando, Svenja Meierjohann, Hubert Kübler, Ralf Bargou, Bastian Schilling, Sebastian E Serfling, Andreas Buck, Rudolf A Werner, Constantin Lapa, Markus Krebs; Predicting Microenvironment in CXCR4- and FAP-Positive Solid Tumors—A Pan-Cancer Machine Learning Workflow for Theranostic Target Structures. Cancers **2023**, 15, 392. https://doi.org/10.3390/cancers15020392

5. Markus Krebs, Antonio Giovanni Solimando, Charis Kalogirou, **André Marquardt**, Torsten Frank, Ioannis Sokolakis, Georgios Hatzichristodoulou, Susanne Kneitz, Ralf Bargou, Hubert Kübler, Bastian Schilling, Martin Spahn, Burkhard Kneitz; miR-221-3p Regulates VEGFR2

Expression in High-Risk Prostate Cancer and Represents an Escape Mechanism from Sunitinib In Vitro. J. Clin. Med. 2020, 9, 670. https://doi.org/10.3390/jcm9030670

6. Andreas Hofmeister, Maximilian C Thomaßen, Sabrina Markert, **André Marquardt**, Mathieu Preußner, Martin Rußwurm, Ralph T Schermuly, Ulrich Steinhoff, Hermann-Josef Gröne, Joachim Hoyer, Benjamin D Humphreys, Ivica Grgic; Development of a new macrophage-specific TRAP mouse (Mac^{TRAP}) and definition of the renal macrophage translational signature. Sci Rep 10, 7519 (2020). https://doi.org/10.1038/s41598-020-63514-6

7. Christian Michel, Elisabeth KM Mack, Christopher-Nils Mais, Lea V Fritz, Ying Wang, Lutz B Jehn, Sonja K Hühn, Clara Simon, Sabrina Inselmann, **André Marquardt**, Jennifer Kremer, Ellen Wollmer, Kristina Sohlbach, Andreas Neubauer, Cornelia A Brendel, Claudia Haferlach, Gert Bange, Andreas Burchert; Cloning and characterization of a novel druggable fusion kinase in acute myeloid leukemia. Haematologica. 2020 Aug;105(8):e395-e398. doi: 10.3324/haematol.2019.237818

8. Christina Jessen, Julia KC Kreß, Apoorva Baluapuri, Anita Hufnagel, Werner Schmitz, Susanne Kneitz, Sabine Roth, **André Marquardt**, Silke Appenzeller, Carsten P Ade, Valerie Glutsch, Marion Wobser, José Pedro Friedmann-Angeli, Laura Mosteo, Colin R Goding, Bastian Schilling, Eva Geissinger, Elmar Wolf, Svenja Meierjohann; The transcription factor NRF2 enhances melanoma malignancy by blocking differentiation and inducing COX2 expression. Oncogene 39, 6841–6855 (2020). https://doi.org/10.1038/s41388-020-01477-8

9. Julia Katharina Charlotte Kreß, Christina Jessen, **André Marquardt**, Anita Hufnagel, Svenja Meierjohann; NRF2 Enables EGFR Signaling in Melanoma Cells. Int. J. Mol. Sci. 2021, 22, 3803. https://doi.org/10.3390/ijms22083803

10. Robert Mandic, **André Marquardt**, Philip Terhorst, Uzma Ali, Annette Nowak-Rossmann, Chengzhong Cai, Fiona R Rodepeter, Thorsten Stiewe, Bernadette Wezorke, Michael Wanzel, Andreas Neff, Boris A Stuck, Michael Bette; he importin beta superfamily member RanBP17 exhibits a role in cell proliferation and is associated with improved survival of patients with HPV+ HNSCC. BMC Cancer 22, 785 (2022). https://doi.org/10.1186/s12885-022-09854-0

2.1.2 Conference Contributions

04.-06.11.2019, Heidelberg	Poster at the 4th EMBL Conference: Cancer Genomics in Heidelberg, Germany Poster title: Using Machine Learning on RNA-Sequencing Data to Identify New Signatures in Cancer for CUP Prediction
24.-26.09.2020, Online	Talk at the 72nd Jahrestagung der Deutschen Gesellschaft für Urologie e.V., Best of DGU 2020, Talk title: RNAseq-based mapping of renal cell carcinoma - Machine Learning identifies mitochondrial gene signatures beyond classical histopathology
07-08.10.2020, Würzburg	Talk at the 15th Eureka International GSLS student Symposium in Würzburg, Germany Talk title: Hidden in the Transcriptome – Machine Learning Uncovers Prognostic Mitochondrial Gene Signature in Renal Cell Carcinoma beyond Established Histopathology

iii. CURRICULUM VITAE

iii. Curriculum Vitae

iii. Curriculum Vitae

iv. ACKNOWLEDGMENTS

My deepest gratitude goes to Prof. Dr. Svenja Meierjohann, not only for placing her trust in me, but also for giving me the opportunity of a fresh start in her research group. In the last nearly four years, she has always supported me in both word and deed, even during difficult phases. Her positive nature and her talent for constructive suggestions were crucial not only for the success of this work, but also for other projects.

Another special thanks goes to my second supervisor Prof. Dr. Philip Kollmannsberger, who always supported me with professional input and helped me make this thesis what it is now.

I would also like to thank my third supervisor and collaboration partner Dr. Markus Krebs, as his willingness to listen as well as his advice - not only in the scientific field - has made him a good friend. I sincerely hope that our cooperation will continue to grow and produce successful results in the future.

Last but not least, I would like to take this opportunity to thank all those who have always stood by me during good and bad times, first and foremost my parents.

Thank you!

v. AFFIDAVIT

5.1 AFFIDAVIT

I hereby confirm that my thesis entitled „ Machine Learning Based Identification of Tumor Entities, Subgroups, and Therapy Options" is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Stuttgart, 01.03.2023 _____

Place, Date Signature

5.2 EIDESSTATTLICHE ERKLÄRUNG

Hiermit erkläre ich an Eides statt, die Dissertation „Bestimmung von Tumorentitäten, Subgruppen und Therapieoptionen basierend auf maschinellem Lernen" eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Stuttgart, 01.03.2023 _____

Ort, Datum Unterschrift