Distance-based methods for the analysis of Next-Generation sequencing data

DISSERTATION

zur Erlangung des akademischen Grades

Doctor rerum naturalium
im Fach der Informatik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

M.Sc. Bioinformatik Raik Otto

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:

1. Prf. Dr. Ulf Leser
2. Prf. Dr. Christine Sers
3. Prf. Dr. Stefan Keller

Datum der Einreichung: 02.03.2021
Datum der Disputation: 25.08.2021

Acknowledgement

A plethora of sophisticated persons has invaluably contributed to the creation of this thesis by providing valuable advise, insights and orientation. Some people contributed during the creation of the thesis, some even years before.

Explicit gratitude is devoted to the thesis' supervisor, Ulf Leser. His significant scientific support and clear guidance through scientific landscapes of increased complexity proved valuable and supportive. Moreover, Christine Sers' initiation and management of the MAPTor-NET project which provided the scientific and funding framework for this thesis deserves explicit thankfulness. I thank Katharina Detjen and Pamela Riemer for their close cooperation and the conduction of laboratory validation experiments. Moreover I'd like to thank my wife who constantly supported the creation of the thesis. Bertram Weiß inspired the conception of the presented methods via valuable scientific exchange while supervising the M.Sc. thesis. Eventually, I would like to express sincere gratefulness to Christian Maasz and Guido Walter for their guidance during earlier times by not only teaching knowledge as such but life itself what directed me towards this thesis.

Zusammenfassung

Die Analyse von Next-Generation Sequencing (NGS) Daten ist ein zentraler Aspekt der modernen genomischen Forschung. Bei der Extraktion von Daten aus den beiden am häufigsten verwendeten Quellorganismen bestehen jedoch vielfältige Problemstellungen. Wir untersuchen in dieser Arbeit die Problemstellungen der Fehlidentifizierung von Krebszelllinienkulturen und den Mangel an geeigneten Patienten-abgeleiteten Datensätzen für das Trainieren von maschinellen Lernmodellen. Die Fehlidentifizierung von Krebszelllinienkulturen stellt eine bedeutende Fehlerquelle dar und wird durch die Abwesenheit geeigneter Computer-gestützter Kultur-Identifizierungsalgorithmen zusätzlich erschwert. Im Kontrast hierzu sind Patienten-abgeleitete Biopsien selten von Fehlidentifizierungen betroffen jedoch für seltene Krebsarten, insbesondere solche mit hoher Subtyp-Diversität, wenig bis überhaupt nicht verfügbar was z.B. die Anwendung von artifizieller Intelligenz stark einschränkt. Diese Thesis präsentiert Lösungsansätze für diese Problemstellungen wobei übergreifend das Konzept der Abstandsquantifizierung zwischen sequenzierten Entitäten verwendet wird.

Im ersten Kapitel wird ein neuartiger Ansatz vorgestellt welcher einen Abstand zwischen Krebszellinienkulturen auf Grundlage ihrer kleinen genomischen Varianten bestimmt um die Kulturen zu identifizieren. Eine Voll-Exom sequenzierte Kultur wird durch paarweise Vergleiche zu Referenzdatensätzen identifiziert so ein gemessener Abstand geringer ist als dies bei nicht verwandten Kulturen zu erwarten wäre. Die Wirksamkeit der Methode wurde verifiziert, jedoch verbleiben Einschränkung da nur das Sequenzierformat des Voll-Exoms unterstützt wird.

Daher wird im zweiten Kapitel eine publizierte Modifikation des Ansatzes vorgestellt welcher die Unterstützung der weitläufig genutzten Bulk Ribonucleic acid (RNA) sowie der Panel-Sequenzierung ermöglicht. Die Ausweitung der Technologiebasis führt jedoch zu einer Verstärkung von Störeffekten welche zu Verletzungen der mathematischen Konditionen einer Abstandsmetrik führen. Daher werden die entstandenen Verletzungen durch statistische Verfahren zuerst quantifiziert und danach durch dynamische Schwellwertanpassungen erfolgreich kompensiert.

Das dritte Kapitel stellt eine neuartige Daten-Aufwertungsmethode (Data-Augmentation) vor welche das Trainieren von maschinellen Lernmodellen in Abwesenheit von neoplastischen Trainingsdaten ermöglicht. Ein abstraktes Abstandsmaß wird zwischen neoplastischen Entitäten sowie Entitäten gesundem Ursprungs mittels einer transkriptomischen Dekonvolution hergestellt. Die Ausgabe der Dekonvolution erlaubt dann das effektive Vorhersagen von klinischen Eigenschaften von seltenen jedoch biologisch vielfältigen Krebsarten wobei die prädiktive Kraft des Verfahrens der des etablierten Goldstandard ebenbürtig ist.

Abstract

The analysis of Next-Generation Sequencing (NGS) data is a central aspect of modern Molecular Genetics and Oncology. However, the analysis of sequencing data derived from the frequently sequenced source organisms, Cancer Cell Lines (CCLs) and patient-derived neoplasms, remains susceptible to errors and subject to constraints. This thesis addresses the erroneous misidentification of CCLs and constraining training data scarcity of rare and diverse cancer types. The shared element of the contributions is the quantification of an abstract distance between sequenced entities.

The first scientific contribution is the development of a method which identifies Whole-exome-sequenced CCLs via the quantification of a distance between their sets of small genomic variants. A distinguishing aspect of the method is that it was designed for the computer-based identification of NGS-sequenced CCLs. An identification of an unknown CCL occurs when its abstract distance to a known CCL is smaller than is expected due to chance. The method performed favorably during benchmarks but only supported the Whole-exome-sequencing technology.

The second contribution therefore extended the identification method by additionally supporting the Bulk mRNA-sequencing technology and Panel-sequencing format. However, the technological extension incurred predictive biases which detrimentally affected the quantification of abstract distances. Hence, statistical methods were introduced to quantify and compensate for confounding factors. The method revealed a heterogeneity-robust benchmark performance at the trade-off of a slightly reduced sensitivity compared to the Whole-exome-sequencing method.

The third contribution is a method which trains Machine-Learning models for rare and diverse cancer types which present with little or no training data. A distance is quantified between neoplastic entities and cells of healthy origin via transcriptomic deconvolution. Machine-Learning models are subsequently trained on these distances to predict clinically relevant characteristics. The performance of such-trained models was comparable to that of models trained on both the substituted neoplastic data and the gold-standard biomarker Ki-67. No proliferation rate-indicative features were utilized to predict clinical characteristics which is why the method can complement the proliferation rate-oriented pathological assessment of biopsies.

The thesis revealed that the quantification of an abstract distance can address sources of erroneous NGS data analysis, but as well found that the distance quantification-concept is susceptible to confounding factors and is therefore most effectively applied to the analysis and comparison of homogeneously sequenced entities.

Contents

Chapter 1

Introduction

The decryption of the human genome was a pivotal event because the creation of a human reference genome rendered multiple advances in the domain of Life-Sciences possible [1, 2]. The decryption, id est (i.e.) comprehensive genotyping of the genome's nucleotides sequence, was facilitated by the introduction of the *shotgun*-sequencing technology. The novel *shotgun*-sequencing technology represented a technological leap relative to the established chain termination method (Sanger-Sequencing) which is why it is referred to as Next-Generation Sequencing (NGS) technology [3]. The instrumental advantages of the NGS technology were low per-base sequencing costs and a high throughput rate whose combination stimulated the widespread utilization of NGS technology and public availability of NGS datasets [4].

A plethora of Life-Science domains profited from the corresponding increase in availability of genomic data since complex empirical studies which relied on increased sample-sizes became feasible [5]. Cross-disciplinary scientific domains did, however, gain particular scientific momentum because they combined the diverse types of knowledge required for the correct extraction of information from high-throughput NGS datasets. The required knowledge encompasses scientific fields such Molecular Genetics, Biostatistics and Computer Sciences whose intersection is found in the data-analysis domains of Computational Biology and Bioinformatics [2].

The importance of data-analysis for NGS-based genomic research is highlighted by the degree to which data-driven research has influenced the research paradigm of Molecular Genetics [5]. The research paradigm in the low-throughput era was hypothesis-driven in that the sequencing data had the purpose of supporting or refuting scientific hypotheses that were formulated before the sequencing itself took place [3]. The modern research paradigm is, in contrast, frequently data-driven in that data is generated first and explanatory hypothesis established after a thorough analysis of the data has been conducted [6].

The demands posed on the data-integrity and the data quality-control have increased correspondingly with the data volume, technological diversity and reliance on the NGS data-analysis. A primary reason for this development lies within the *up-stream* positioning of the data-analysis process within the workflow structure of a research project. An error in the data-analysis process may reduce the scientific value of the corresponding study significantly because the data-analysis output is the input for domain experts whose *down-stream*-located task is to interpret and contextualize the data-analysis results, potentially without means to verify the correctness of the data [7]. Furthermore, currently utilized quality-assurance solutions were at least partially developed before the advent of the NGS technology and can require additional overhead quality-assurance experiments in addition to the actual sequencing experiments if the established methods are applicable at all [8, 9]. Therefore, addressing the sources of erroneous NGS data-analysis based on the NGS data itself via application of computer-based *in-silico* methods is of great importance for multiple Life-Science domains which is why it is the scientific subject of this thesis.

Sequencing data-based Cancer Cell Line identification

All sequencing data within the domain of Oncology is derived from the sequencing of biological entities which, in a non-exhaustive listing, constitute of Cancer Cell Line (CCL), patient-derived biopsies, organoids and xenografts [9]. CCLs are high throughput two dimensional petri-dish model cultures which are most widely utilized in the Life-Sciences. However, since the beginning of research on CCLs have CCL-misidentifications been a risk factor associated with their utilization which is why the ability to identify CCLs is crucial for genomic research [9]. Gold-standard CCL identification methods, such as Short Tandem Repeat (STR) exist, but their conception predates the introduction of the NGS-technology which is why data-driven research is conceptually disadvantaged with respect to CCL identification. The disadvantage materializes in that pre-NGS technology methods require additional experiments conducted on the physical CCL culture for identification. The physical availability of the sequenced culture is, however, frequently not given in the era of the data-driven research paradigm for instance because CCL NGS data is virtually exchanged via the internet. Furthermore, the STR gold-standard method cannot be applied to NGS data because long tandem repeats are conscientiously not counted by NGS-analysis software due to the difficulty associated with correctly resolving tandem-repeat structures [10]. In summary, the great risk of CCL-misidentification in conjunction with a lack of support of the gold-standard method for NGS data-analysis motivates the development of a generic CCL NGS data-based identification method [11].

Sequencing data-augmentation for Machine-Learning of rare and diverse cancer types

CCLs excel because their sample-sizes can be comparatively easily scaled. However, they are limited in their capacity to reflect the biological reality of a patient's *in-situ* neoplasm [12]. Patient-derived data is therefore superior in its scientific value but commonly limited with respect to its sample-sizes [13]. The limited sample-sizes of patient-derived material exacerbates the acquisition of suitable biopsy material that comprehensively covers the neoplastic diversity [14]. The availability of sufficient amounts of training data is, however, instrumental for an effective training of Machine-Learning (ML) models to avoid, including to but not limited to, overfitting, a class-balanced classification or regression performance and the reduction of the model-complexity [15]. This lack of NGS training data therefore has the ramification of precluding the full exploitation of the scientific potential of NGS data with respect to the current endeavors of personalizing the patient-treatment and drug-regime.

An approach utilized in the domain of Machine-Learning (ML) to augment training data, i.e. to increase the amount of suitable training data, is to, for instance, perturb the available data in order to generated altered data that can be added to the original training data [16]. In the domain of Oncology, training data sizes can be increased by the inclusion of data of healthy origin should the addition prove informative with respect to a neoplasm [17]. Properties of a neoplasm can exempli gratia (e.g.) be identified via the quantification of the similarity of a neoplasm to a healthy cell since a malignant neoplasm is generally less similar to a healthy cell than a benign neoplasm. Chapter 5 presents a method that augments the data of a cancer type with low incidence rate to address the limited data sample-sizes. Such augmented data is subsequently shown to allow for the training of ML models whose predictive power is comparable to that of a model trained on substituted neoplastic data.

1.1 Aim

The aim of the thesis is to develop methods which address the problems of CCL misidentification and lack of training data. This thesis utilizes the quantification of abstract distances as conceptual framework for the development of the methods. In the context of the thesis, 'distance-quantification' signifies that entities whose distance cannot be qualified with a physical distance are given an abstract distance in order to predict their properties such as their identity or clinical grading.

The reason why we apply abstract distance quantification is that distances are generally geometrically interpretable and applicable to pairs of entities. The pair-wise nature allows for empirical sampling of otherwise latent parameters via an analysis of the distribution of all pair-wise distances. Statistical tests can thereafter decide on the properties of the entities conditioned on the sampled parameters. An example is the identification of a significantly small distance between CCLs. If the distribution of all quantified pair-wise distances is known, one can resolve for the distance-value at 95% of all distance-values are greater and thereby determine an empirical threshold for distance-based identification. Secondary considerations are the run-time which is critical given the high volume of the analyzed data. The training of, for instance, Deep-Learning Networks can take a considerable amount of time and resources to train what would impede the development of the methods.

Consequently, the distance-quantification over NGS data is the central element of all methods: CCLs are identified via pair-wise comparisons of their NGS-derived distances and rare cancer types are classified via a distance-quantification of their NGS data to NGS data derived from healthy donors.

1.2 Contributions

The scientific contributions of the thesis are three novel Bioinformatic approaches which address CCL misidentification and training data-augmentation. The contributions' methodology, theoretical background, benchmark performance and a critical discussion of their advantages and disadvantages are presented. Abstract distance-quantification is utilized by all contributions but differs with respect to the type of utilized NGS data (Chapter 3 and 4: small variants, Chapter 5: transcriptomic) and data format (Chapter 3: Whole Exome-Sequencing (WES), Chapter 4: Bulk Ribonucleic acid (RNA)-seq and Panel-seq) and type of problem that is solved (identification versus augmentation).

1.2.1 Whole-Exome sequencing technology-based CCL identification

We present a NGS data-based CCL identification method in Chapter 3 which identifies CCLs based on a distance metric over small variants that are rare what represents a stark contrast to established methods which are inflexibly dependent on predetermined genomic entities such as Single-Nucleotide Polymorphisms (SNPs). The approach, called UNIQUe variant identification Of canceR cell liNes (Uniquorn), is structured such that the identification of CCLs data with limited Data-Heterogeneity is first established in Chapter 3 and its generalization for diversely sequenced CCLs in Chapter 4. The Uniquorn method was benchmarked favorably on more than 700 CCLs. The method was, however, only benchmarked on WES sequenced CCLs to limit the technological diversity and ensure the accurate empirical approximation of intractable distributions.

1.2.2 Generalized NGS technology-based CCL identification

Chapter 4 reports on the extension of Uniquorn WES to identify CCLs that were either Bulk RNA, WES or Panel-sequenced. The Chapter addresses the problem that Bulk RNA-sequencing and Panel-sequencing technologies were not supported by the Uniquorn WES approach but are frequently utilized to sequence CCLs. The Uniquorn extension is therefore designed to cope with a great amount of Data-Heterogeneity in order to support the majority of the currently utilized NGS technologies.

The distance metric applied in Chapter 3 proved ineffective due to significant differences with respect to the amount of variants called by diverse sequencing technologies for identical CCLs. The Uniquorn methodology was thus modified via integration of empirical resampling techniques to quantify the strength of the technological heterogeneity which represented a confounding factor. The technological factor could thereafter be compensated for by a dynamic adjustment of the identification thresholds according to the strength of the technological bias, rendering the method applicable for significantly more use-cases.

1.2.3 Prediction of Clinical Characteristics of rare and diverse Neoplasms

We report on a novel method to predict clinically relevant characteristics of rare yet biologically diverse Neuroendocrine Neoplasms (NENs) via transcriptomic deconvolution in Chapter 5. The Data-Augmentation method first analyzes whether a distance-quantification via deconvolution is possible and secondly whether such derived distances are informative with respect to clinical characteristics. Generally, clinical characteristics can be predicted by Machine-Learning (ML) models trained on neoplastic data. The required amounts of neoplastic training data are, however, frequently not available for rare and diverse cancer types. Therefore, a Data-Augmentation of the training data is conducted via a substitution of the neoplastic training data with ubiquitously available data of healthy origin. The substitution of the training data is based on the distance between a neoplasm to data of healthy origin i.e. the deconvolution results are utilized as base for the distance quantification and subsequent model training.

1.3 Thesis outline

The thesis is composed of a General Introduction, a Scientific Background and three contribution Chapters in addition to a Conclusion Chapter and Appendix, see Figure 1.1.

The background Chapter 1 introduces the most important deoxyribonucleic acid (DNA) and RNA

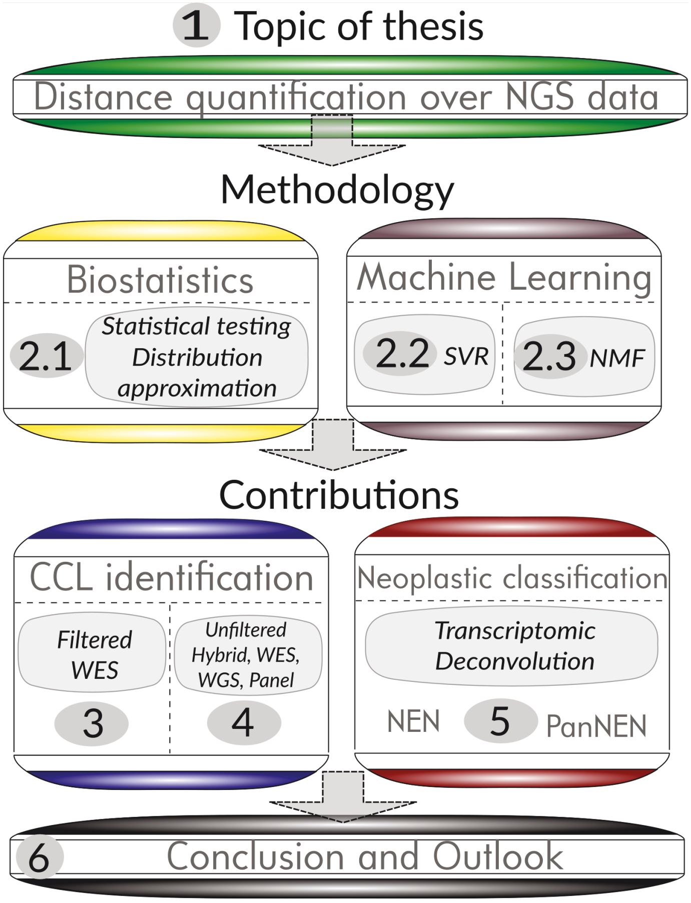
1 Topic of thesis
Distance quantification over NGS data
Methodology
Biostatistics
2.1
Statistical testing Distribution approximation
Machine Learning
2.2 SVR
2.3 NMF
Contributions
CCL identification
Filtered WES
Unfiltered Hybrid, WES, WGS, Panel
3
4
Neoplastic classification
Transcriptomic Deconvolution
NEN
5
PanNEN
6
Conclusion and Outlook

Figure 1.1: Overview of the thesis. The thesis comprises of six Chapters and multiple (Sub)-Sections. Sections 2.1, 2.2 and 2.3 provide contextualizing background information regarding the scientific concepts of Biostatistics and ML algorithms which were applied as part of the scientific contributions. Contribution Chapters for CCL identification and classification-by-deconvolution each consist of an introduction, methods, results and discussion Section. The thesis finishes with a Conclusion Chapter that integrates the contributions' findings with respect to abstract distance-quantification, followed by the Supplementary Material.

sequencing technologies relevant to the Chapters 3 and 4. Thereafter, the concept of a distance metric over a space is defined and the general context of distance-quantification over NGS data presented. Identification by distance-quantification requires the determination of a threshold where sufficient similarity is achieved. The Uniquorn methods applies statistical hypothesis tests to determine that threshold and hypothesis tests are therefore explained in the succeeding Section 2.3. The following Section 2.3.2 introduces empirical sampling and empirical testing methods and Section 2.3.2 displays how the identification threshold determination could be conducted when standard hypothesis tests were not applicable. We present the methodological background of transcriptomic deconvolution in Section 2.4 and continue with an introduction of and explanation to why two ML methods, Support Vector Machine Regression (SVR) and Non-negative Matrix Factorization (NMF), were utilized in Chapter 5.

In the main part of the thesis, the contributions are presented. Chapter 3 commences with an introduction to CCL cultures to render the biological background palpable. We outline the historic development of CCL identification methods to motivate why *in-silico* identification methods are required. Thereafter, the specific concept of distance-quantification via the matching of small variants is explained in Section 3.1.2. A benchmark based on WES data from three major CCL screening studies serves to estimate the performance of the distance-quantification identification method.

Chapter 4 presents the motivation to additionally support the Bulk RNA-sequencing and Panel-sequencing technologies. Section 4.1 outlines the differences and similarities between the WES-only and the universal identification method. The following Section benchmarks training data from a WES scenario but in addition contains hundreds of Bulk RNA and panel-sequenced CCLs. Chapter 4 concludes with a discussion of the suitability of the generalized identification method as universal identification algorithm, the method's advantages and disadvantages and a listing of identified confounding factors that deny the correct identification of CCLs.

Chapter 5 begins by presenting the biological background of NENs which is distinct in the sense that NENs are simultaneously rare and biologically diverse. Furthermore, we outline why neoplastic NGS data, required for ML model training, is not available and thus a training on data derived from healthy donors motivated. The novelty of the neoplastic deconvolution approach is rendered comprehensible and its potential within the field of NEN research is explained. The explanation focuses in particular on the approach that NENs are classified based on their abstract distance to a healthy training sample as quantified by a reconstruction error and relative cell-type proportion predictions. Multiple NEN datasets were benchmark and the method's performance with respect to the prediction of clinical characteristics determined as comparable to the current gold-standard biomarker. Chapter 5 concludes with a discussion of the classification-by-deconvolution aspects that proved successful while simultaneously elaborating on currently present limitations of the method.

Chapter 6 aggregates the distance-quantification-related findings of Chapters 3 to 5 and discusses potential future research. The Appendix contains the Supplementary Material Sections for Figures 7.1 & tables 7.2 and an abbreviation register 8 which conclude the thesis.

1.4 Own prior work

The Uniquorn WES method was published in Otto et al. 2017 [18] and its generalization in Otto et al. 2019 [19]. Following contributions were made by the authors of these publications: For [18], Raik Otto, Ulf Leser and Christine Sers wrote the manuscript while Raik Otto developed the method. Ulf Leser and Raik Otto wrote the second publication [19] and Raik Otto conceived the method. Jan-Niklas Rössler contributed to the visualizations presented in Chapter 3. The research presented in Chapter 5 is scheduled for publication in 2021. In case of an acceptance, the contributions of the authors to Chapter 5 will be structured as follows:

Conceptualization	Raik Otto, Christine Sers, Ulf Leser
Data curation	Carsten Grötzinger, Katharina Detjen, Pamela Riemer, Bertram Wiedenmann, Guido Rindi
Formal analysis	Raik Otto
Funding acquisition	Ulf Leser, Christine Sers, Katharina Detjen
Investigation	Ulf Leser, Pamela Riemer, Bertram Wiedenmann, Guido Rindi
Methodology	Raik Otto, Ulf Leser
Administration	Ulf Leser, Christine Sers
Resources	Raik Otto, Pamela Riemer, Katharina Detjen, Carsten Grötzinger, Christine Sers, Ulf Leser
Software	Raik Otto
Supervision	Ulf Leser, Christine Sers
Validation	Raik Otto, Katharina Detjen, Pamela Riemer, Bertram Wiedenmann, Guido Rindi
Visualization	Raik Otto
Writing	Raik Otto, Katharina Detjen, Ulf Leser, Christine Sers

Chapter 2

Scientific Background

Chapter 2 provides the theoretical background of the thesis. The Chapter begins with subsection 2.1, a short description of the commonalities and differences of those NGS technologies that are relevant within the context of the thesis. Thereafter, a mathematical definition of distance metrics is given which is required to quantify distances in Section 2.2. Section 2.3 defines and explains the statistical tests relevant for an understanding of Chapters 3 and 4. Subsequently, Section 2.4 elaborates on the deconvolution-based distance-quantification as applied in Chapter 5. The introduction finishes with Section 2.5, a description of ML algorithms required for the transcriptomic deconvolution.

2.1 Next-Generation sequencing

NGS technologies can be classified according to the type of molecular entity which they genotype [2, 5]. In the context of this thesis, DNA and RNA-based technologies are relevant due to their wide-spread utilization and the fact that small genomic variants and messenger Ribonucleic Acid (mRNA) expression levels can be obtained from their analysis, respectively.

A sub-classification of NGS technologies, applied within the framework of this thesis, is the sequencing format which will refer to the discrimination of NGS technology by the amount of targeted loci. The sequencing format therefore relates to the volume of data that is generated during a sequencing run. For example, Whole Genome-Sequencing (WGS) and Panel-sequencing both sequence DNA and therefore belong to the same technology, but differ with respect to their format because the amount of covered DNA basepairs differs by a factor of $\sim 10^6$ what has major ramifications for the CCL identification methods. See Figure 2.1 for a depiction of the technologies and formats relevant to this thesis.

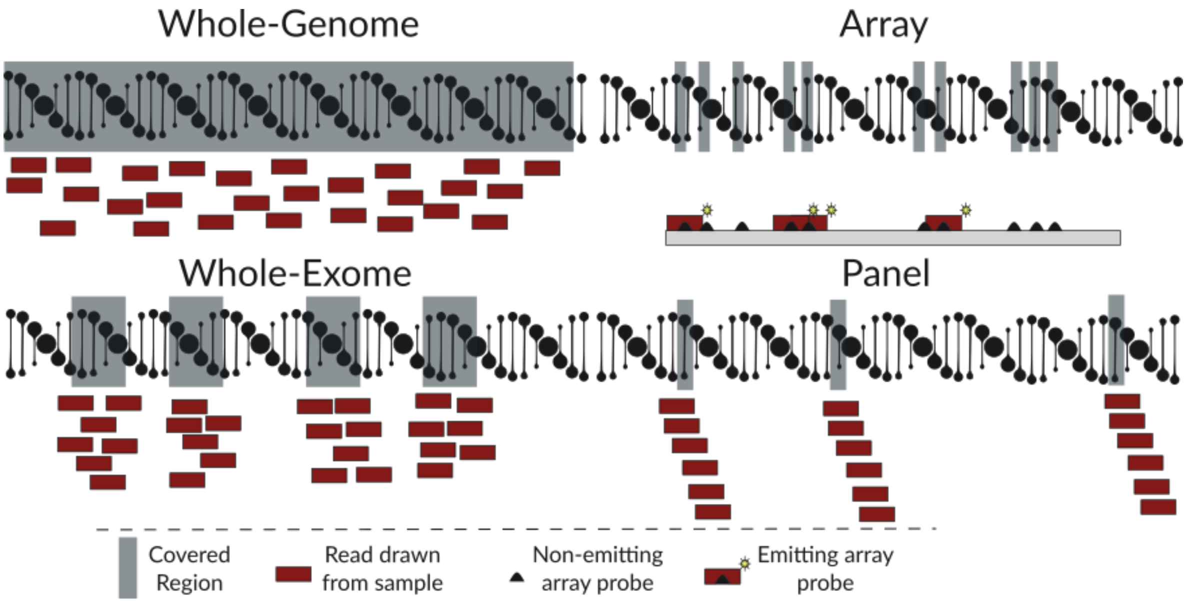

Figure 2.1: Overview of sequencing technologies and formats relevant to this thesis. A primary difference between sequencing technologies and formats are the genomic loci that are genotyped, highlighted by gray frames located on the double helix that represents a genome. Whole Genome-Sequencing (WGS) genotypes all regions of the genome with exclusion of the telomeres, the centromere and highly repetitive regions where repeat-resolution is generally not sufficiently reliable for short read-based high-throughput technologies [2]. Whole Exome-Sequencing (WES), arrays and Panel-sequencing share the property of genotyping only selected regions. The difference between WES and Panel-sequencing is that WES targets the whole exome while Panel-sequencing only targets a few hundred genes. Arrays differ from WGS, WES and Panel-sequencing technologies by utilization of probes located on the surface of a chip which emit light due to laser-excitement when a molecule binds what indicates that the corresponding mRNA was expressed in the sample. mRNA arrays cover comparatively short parts of the genome, limited by the length of the sequence with which the probe with which the molecule hybridizes.

Whole-Genome sequencing

In the context of this thesis, WGS is defined as the *shot-gun* sequencing of a whole human eukaryotic genome and informs about a mutational or wild-type status of DNA located within the nucleus and the mitochondrium [20]. A wild-type status is defined as a basepair call at a given genomic locus that is identical to the reference genome's basepair call and a mutation indicates a divergence from the reference. When a mutation substitutes a single genomic basepair (length of one basepair) and has a population prevalence of greater or equal to 5%, it is considered a Single-Nucleotide Polymorphism (SNP) and else an either somatic or private Single-Nucleotide Variant (SNV). Insertions and Deletions (InDels) are defined as genomic insertions or deletions of lengths of one up to ten basepairs. Substitutions of lengths of two up to ten are not qualified by a population prevalence and will therefore be referred to as substitutions. Large-scale structural somatic Copy Number Aberrations (CNAs) or germline Copy Number Variations (CNVs) are not included in the mutation term and are not subjected to analyses within the thesis.

A distinctive advantage of the WGS format compared to other approaches is that up to 84% of the (human) genome can be confidently sequenced [21]. This extensive coverage enables algorithms to call large-scale CNVs and determine the ploidy of a genome what is significantly more challenging for other sequencing technologies although constraints with respect to the calling of tandem-repeats remain [22].

WGS typically analyzes the DNA of a mixture of cells. In cases of healthy cells from a single donor, this circumstance is not considered a confounding factor since all cells from the same human being are assumed to possess the same genomic sequence with the exception of gametocytic cells and non-nucleic cells such as erythrocytes [23]. The identity assumption does not hold for neoplastic cells whose genome is subject to somatic i.e. non-germline mutations and structural genomic aberrations, respectively. Small variants such as SNVs up to large-scale CNVs can be called from WGS data. WGS data is therefore technologically suitable for NGS data-based CCL identification. Note, however, that the WGS technology is not intensively benchmarked by the contributions because no large-scale publicly available studies of WGS sequenced CCLs exist.

A drawback of the WGS technology is that it suffers from greater sequencing costs, higher storage space requirements, Random Access Memory (RAM) footprint and in general lower sequencing coverage compared to WES and Panel-sequencing, which is why the relative utilization rate is generally low compared to other sequencing formats [23].

Whole-Exome sequencing

Whole Exome-Sequencing (WES) is defined as the sequencing of the ~2% of the genome or ~22,000 genes in a human being that are currently assumed to be proteinogenic [24]. The primary rational is to obtain information about small DNA variants and InDels from the genomic regions that have a direct impact on the primary up to quaternary protein structure level [25]. WES always targets the expressed part of the genome which is comparatively small what is important for the thesis, since the locus-restriction significantly reduces the heterogeneity of WES data which is why the proof-of-concept in Chapter 3 is based on WES data.

WES DNA and Panel DNA sequencing both utilize primers that bind to characteristic basepair sequences usually located at the 3' end of a gene's transcriptional start-site or shortly before its exons'

3' prime end (for eukaryotes) [23]. The strand direction is of importance because the human DNA-dependent RNA-polymerase matches ('reads') DNA basepairs from their 3' to their 5' end and polymerizes ('writes') mRNA basepairs from their 5' to their 3' end [26].

The 3' upstream sequence-dependency incurs the problem that the primer sequences have to be chosen such that they are located 3' upstream and are either unique or at least shared by as few genomic loci as possible. The sequence of the gene which is not part of the primer itself does not have to known beforehand [25]. The polymerized sequences of the gene therefore may differ from the reference genome up to a limited extent what allows for the discovery of novel small variants such as SNVs or small InDels and substitutions. Novel variants do not have to be transcribed to be picked up by WES DNA sequencing, which is an advantage over the RNA-sequencing technology and motivates the support of the WES technology for CCL identification purposes.

Panel-sequencing

Panel-sequencing, also known as *targeted-sequencing*, is the sequencing technology with smallest amount of covered genes. Panel-sequencing pursues two objectives: flexibility with respect to targeted genes and a reduction of sequencing costs [27]. A legion of commercially available DNA and mRNA-sequencing panels can target $\sim$50 up to few a hundred genes what further reduces costs by standardization. These comparatively affordable pre-designed panels are applied by a wide range of researchers since the panel-design suffices to answer various clinical and scientific questions [28]. Genomic loci that do not harbor translated genes, such as pseudogenes, can as well be targeted, given the existence of a unique primer-capture sequence, but the commercially standardized kits generally target genes with translated exons [29].

A disadvantage is that the low volume of data generated can lead to strand-bias problems even for high coverage panels, in particular in case of hyperploid neoplastic genomes [30]. A strand-bias is defined as the almost exclusive sequencing of only one or a few strands when multiple strands exist. The ramification is that mutations on one strand are either not picked up or hemizygous mutations (assuming di-ploidy) are mistakenly called as being homozygous. CCLs are frequently panel-sequenced after e.g. a drug treatment since only a limited amount of genomic loci is considered to be of particular interest and thus, Panel-sequencing data is analyzed in Chapter 4 of this thesis [27]. In analogy to the WES technology, 3' primer sequences have to be known and 5' sequences that are down-stream and not part of the primer may vary.

In spite of their scientific usefulness do panel-sequenced datasets present with challenges with respect to the CCL identification since only few genes are genotyped. Therefore, Panel-sequencing was not integrated into the proof-of-concept method shown in Chapter 3 but supported by the generalized extension presented in Chapter 4.

RNA-sequencing

Sequencing of Ribonucleic acid (RNA) refers to the quantification and basepair sequence-detection of RNA molecules. Different types of RNA-molecules exist, such as the long non-coding RNA, but this thesis will in the following only refer to translated mRNA molecules. Various sub-variants of RNA-sequencing exist between which the major differentiation is whether a mixture of cells is sequenced (Bulk-sequencing) or single-cell sequenced (Single-cell sequencing). Further subtypes differ in their inclusion of an intermediate complementary deoxyribonucleic acid (cDNA) step and how they isolate the mRNA molecules (poly-A-tailed mRNA selection) versus depletion of ribosomal Ribonucleic acid (rRNA) (ribo-depletion). We will present the subtypes most relevant to this thesis.

Bulk RNA-sequencing

RNA Bulk-sequencing is in the following defined as the sequencing of translated single-stranded RNA molecules that are not part of the ribosome and possess a poly-A tail with free-floating and unspecific primers [31]. Both the sequence and the expression levels of mRNA can be determined by Bulk-sequencing. Primers are required to bind to the ends of mRNA molecules after the molecules were isolated and fragmented in order to reverse transcribe the mRNA into cDNA, see Figure 2.2 [28]. No specific primer sequences are required because the poly-a tail identifies mRNA molecules after they have either been extracted from a cell convolute or the ribosomal RNA been depleted. Therefore, all mRNA molecules in a mixture can, theoretically, be detected which is why RNA-sequencing is as well called *Full transcriptome sequencing*. The sequence-independence is important because the sequence of the mRNA fragments does not have to be known before the sequencing and thus, novel and simultaneously transcribed mutations such as somatic SNVs, can be called [31].

The term *Bulk*-sequencing serves to differentiate this technology from the Single-cell sequencing (scRNA) technology by indicating that in Bulk RNA-sequencing, a mixture of cells with possibly diverging mRNA expression levels and sequences is analyzed. The Bulk aspect represents a significant limitation for cancer-related analyses since cancer is commonly sub-clonal i.e. different sub-populations with distinct expression pattern, drug-responses and survival curves exist of which the most resilient sub-population may ultimately dominate the cancer once a relapse occurs [32].

Bulk-sequencing as well implies the undifferentiated simultaneous sequencing of cells of different types. Algorithms which predict which types of cells where present in the sequenced convolute, and in what relative proportions, exist and are generally based on the deconvolution of the transcriptome [34, 35]. Deconvolution into neoplastic sub-clonal populations with individual and general unknown characteristics is currently only possible with strong limitations since the cell-type or sub-clonal population has to be known for the model training step [32]. Deconvolution algorithms are applied in Chapter 5 to predict the clinically relevant characteristics of neoplasms.

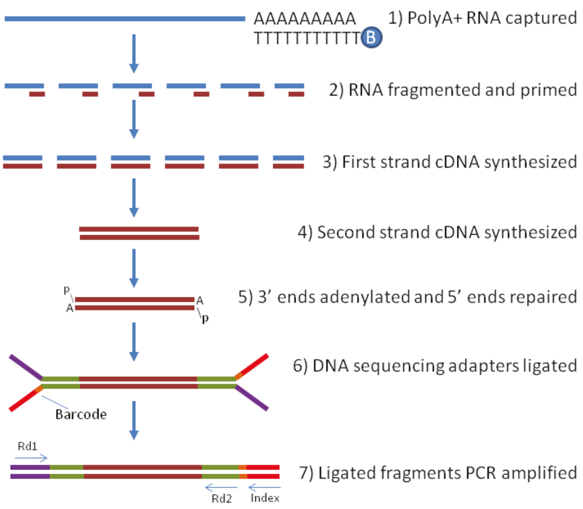
AAAAAAAAA
TTTTTTTTTTT
B
1) PolyA+ RNA captured
2) RNA fragmented and primed
3) First strand cDNA synthesized
4) Second strand cDNA synthesized
P
A
A
P
5) 3′ ends adenylated and 5′ ends repaired
6) DNA sequencing adapters ligated
Barcode
Rd1
7) Ligated fragments PCR amplified
Rd2
Index

Figure 2.2: RNA Bulk-sequencing. The process of mRNA-sequencing as conducted by the Ion Torrent technology platform is depicted. The single-stranded mRNA is first captured whereafter primer sequences are attached and two cDNA reverse-transcriptions conducted. The reverse-transcriptions serve to stabilize the molecule since double-stranded DNA is more stable than single-stranded RNA. Attached barcodes allow to identify molecules and enable the binding of the molecules to surface-probes on a sequencing lane during an eventual polymerase chain reaction (PCR) step (not depicted). Source Figure [33]

mRNA microarray

Messenger RNA microarrays are an established technology platform whose origins trace back to the 1980s [36]. We will discuss the type of array that binds mRNA and follows the Affymetrix^{TM} technology. Sequences of nucleotides, called probes, are located on a chip and hybridize to mRNA fragments. The probe sequences are fixed and preferably specific for single genes or even their exons what allows the quantification of their expression levels or detection of differential splicing in eukaryotes. After mRNA molecules are hybridized to the probes, they are illuminated with laser-light and the expression levels thus quantified via an optical i.e. electro-magnetic signal [37].

Like Bulk RNA-sequencing can mRNA-arrays only quantify convoluted mixtures of cells, assuming that no sorting of cell-types takes place beforehand. Arrays can, however in contrast to Bulk RNA-seq, only quantify gene and exon expression levels, respectively, and not the basepair sequence what is a major limitation and excludes them from the CCL identification approach. Arrays that pickup structural DNA variants, such as CNVs, exist but as of 2020, arrays are mainly utilized to quantify mRNA expression levels.

mRNA arrays represented a significant fraction of all high throughput datasets produced throughout the 1990s and 2000s but are gradually replaced by more powerful technologies such as Bulk RNA or Single-cell sequencing. Nevertheless, arrays will most likely remain in utilization for an extended period of time due to their technological maturation and inexpensiveness [38]. Neoplastic expression data created with mRNA technology is utilized in Chapter 5 of this thesis for deconvolution purposes.

An array-based subtype of WES DNA sequencing is the Hybrid capture sequencing technology whose sequencing format is typically located between the full exome and Panel-sequencing formats with about 500 to 2k targeted genes [5]. Hybrid capture sequencing is reported to possess a superior capacity to discover novel variants compared to full exome-sequencing while simultaneously achieving an advantageous mutational resolution[5]. Primary disadvantages are the fixed probe sequences on the array and a reduced exomic locus coverage. The sequencing format finds application in Chapters 3 and 4 of the thesis.

Single Cell-RNA sequencing

Single-cell sequencing (scRNA) quantifies single mRNA molecules within single-cells [39] and is the most recently introduced technology relevant to this thesis. As of 2021, scRNA sequencing platforms are evolving quickly with regard to multiple aspects such as the mode of cell-isolation, procedure of reverse transcription, second strand synthesis and sequencing library generation [40].

The scRNA technology is the greatest contributor in terms of samples in Chapter 5 and thus essential to this thesis. Single-cell sequencing can cover the full transcriptome or genome provided that a high amount of identical cells are sequenced to aggregate their signal to balance the high drop-out rate [41]. This technology particularly excels with respect to the sequencing of neoplastic genomes due to their single-cell resolution that allows to identify sub-clonal populations of a neoplasm. Furthermore, cell-development stages can be traced what allows for an improved understanding of cell-type differentiation trajectories. Nonetheless, a high degree of technological diversity exists (see Figure 2.3) what exacerbate the replicability of experiments.

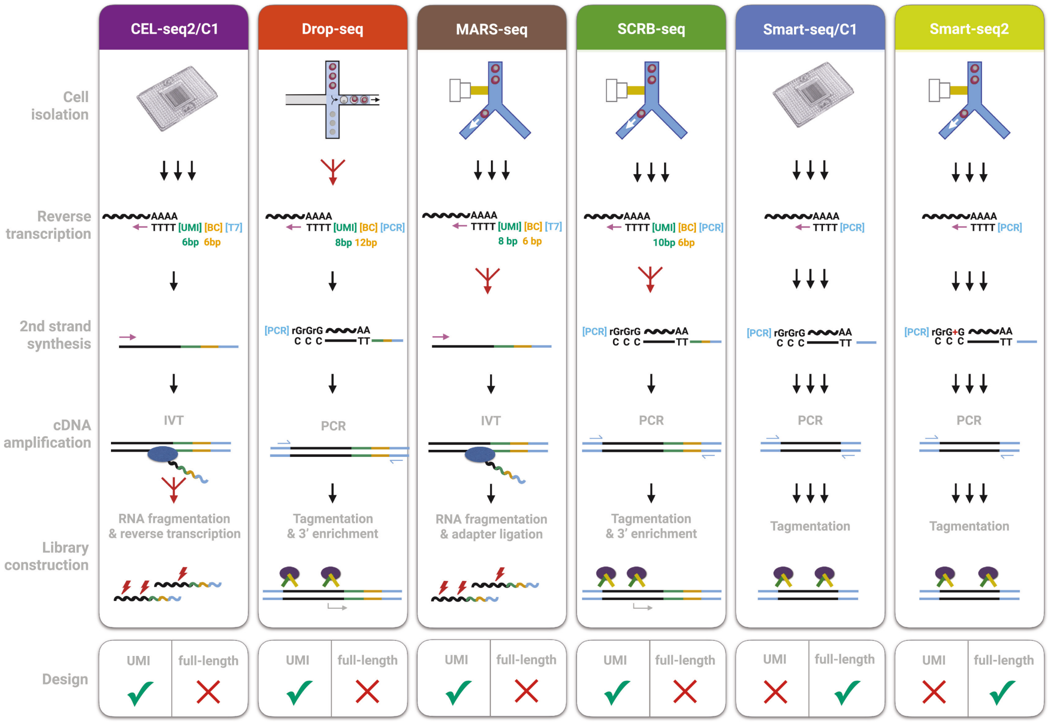

Figure 2.3: Overview of scRNA technologies relevant to this thesis. Main sources of heterogeneity and idiosyncrasies between different scRNA platforms are the cell-isolation, mRNA-isolation, strand synthesis, cDNA amplification and library construction along with Unique Molecular Identifier (UMI) utilization. This thesis analyses data from the Cell Expression by Linear amplification and Sequencing (CEL-seq), Drop-seq, Smart-seq C1 and Smart-seq2 platforms and evaluates which technology proved most suited for the purpose of neoplastic transcriptomic deconvolution by training on healthy scRNA data, see 5.2.1. It is illustrated why multi-technology benchmarks have to be an integral part of Bioinformatics analysis since the diversity and comparatively recent introduction of the scRNA technology can cause a significant degree of result volatility for different technologies [35]. Source Figure: [39]

A major advantage of the scRNA technology for the purpose of transcriptomic deconvolution is, that no cell-type specific cell-surface marker are required for the identification of a since-cell sequenced cell [42]. Therefore, a cell-sorting during the pre-processing is facultative what increases the turn-over rate and decreases the procedural complexity and run-time [43]. A corresponding disadvantage is, that unsorted-cells have to be assigned a cell-type after the sequencing since knowledge of the cell-type is critical for *down-stream* analyses. Various procedures for the cell-type assignment exist and no widely-accepted gold-standard method has been developed what increases the volatility of scRNA data-derived analyses when the data was generated with different technologies and algorithms [42]. The scRNA technology is liberally utilized in Section 5 where it provides information on individual cell-types what represents a fundamental requirement for transcriptomic deconvolution.

In summary, scRNA provides a multitude of valuable insights in to the genome and transcriptomes but remains a still maturing technology which, due to the scarcity of the sequenced material which ranges in the nanograms, requires a high amount of replicates.

2.2 Abstract distance-quantification

The quantification of a distance based on NGS data is the shared element of all contributions presented in the thesis. Here, we will detail on the two types of distance-quantification utilized by the scientific contributions. To that end, we will first contextualize the distance-quantification concept for the domain of Bioinformatics. Thereafter, we introduce the distance-quantification concept for sequenced entities based on their assigned positions in a metric space spanned over small variants as applied in Chapters 3 and 4. Lastly, we will present the quantification of distance between a neoplasm and healthy cells based on a vector norm as applied in Chapter 5.

Contextualization of the distance-quantification concept

At the core of the distance-quantification lies the comparison of features of entities [44]. Within this thesis, a distance-quantification is modeled geometrically via the assignment of a numeric degree of similarity between two entities by a function based on the amount of matching features of two entities. Entities' positions are geometrically modeled such that their features assign the entities a position in a metric space. A function maps a degree of similarity i.e. a distance to two entities based on their positions what requires the space to be a metric space that adheres to mathematical conditions displayed in detail in Subsection 2.2. These mathematical conditions are required because the intuitive notion that the knowledge of the positions of two entities in a given space automatically renders the quantification of a reasonable distance possible cannot generally be assumed and has to be mathematically defined [45].

Distance-quantification is an abstract concept commonly applied in various domains connected to Bioinformatics such as Phylogenetics and Oncology [46, 47]. In NGS-based Phylogenetics, genomic features are investigated based on NGS data in order to create a tree that reflects the evolutionary lineage of e.g. species. A representative procedure to construct a lineage tree based on the genetic similarity of species can be implemented as follows [48]: First, all pair-wise distances between the samples are calculated followed by an update step that merges the two samples with least distance into a new, united sample while deleting the samples' entries from the table of pair-wise similarities. Thereafter, all pair-wise distances between the newly created sample to all remaining samples are inserted into the table of distances. This update step is recursively applied until only one sample remains that comes to serve as the root of the lineage tree. This sequence of mergers then represents a possible lineage history of the compared species based on their pair-wise genetic similarities.

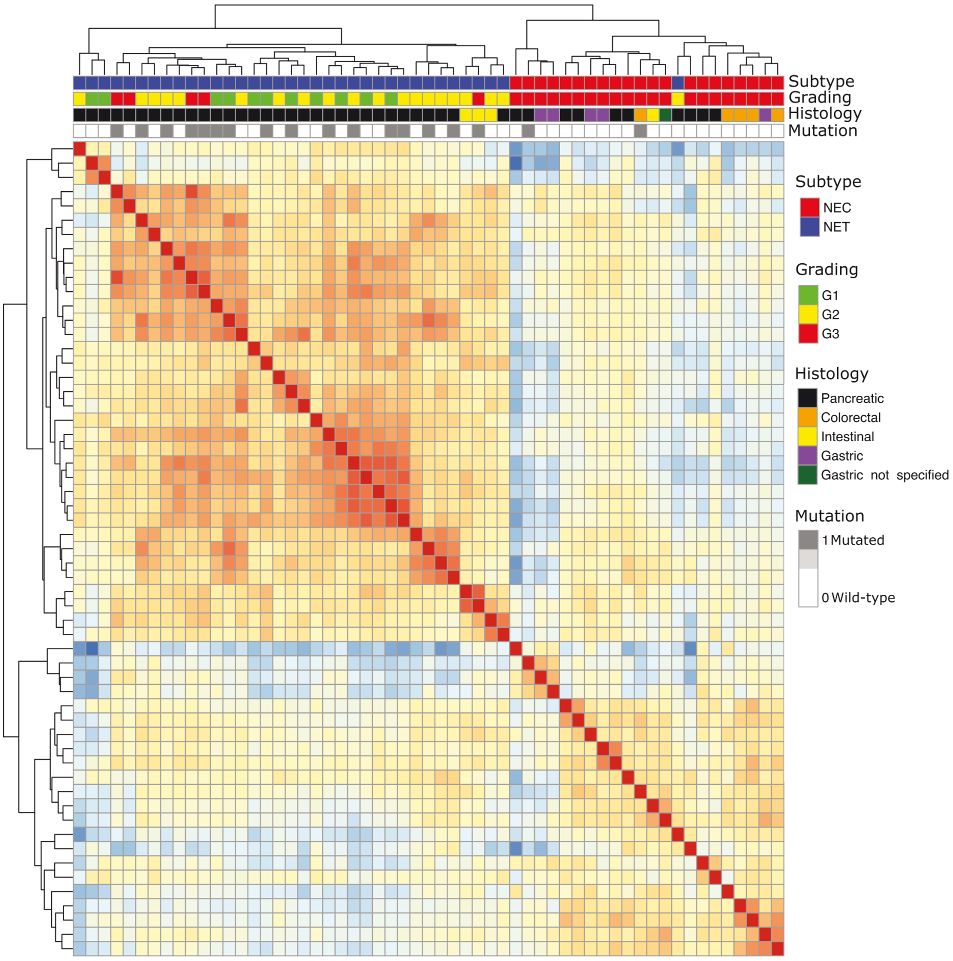

Figure 2.4: Heatmap of clustered neoplasms. The heatmap illustrates how an abstract distance-quantification between transcriptome-sequenced neoplasms renders the extraction of contextualized information via a clustering possible. 69 neoplasms are shown whose pair-wise transcriptomic-distance has been calculated and their positions been arranged such that pair-wisely most similar neoplasms are juxtaposed. Brighter colors indicate a higher degree of similarity and darker colors less similarity. The dendrograms on either rows or columns show the pair-wise clustering according to pair-wise distance. The upper rows show metadata with respect to the neoplastic subtype (carcinoma/ NEC versus tumor/ NET), neoplastic grading (from G1 to G3), Histology and mutational status. All 69 samples were analyzed in Chapter 5 where their clustering pattern was instrumental with respect to the association of transcriptomic activity and clinical phenotype. The heatmap provides the information that samples cluster according to their neoplastic subtype (one outlier) and only to a lesser extent according to their grading. Histology is unrelated to the samples' cluster pattern and mutational status hints at an underlying stratification process that distinguishes mutated from wild-type sample but as well underlines that the mutation stratification is heterogeneous within the tumor/ NET subtype field. In summary, the biological aspect such as subtype and grading can be discerned from the biological processes that are not suited to explain commonalities between neoplasms what is a critical finding from an oncological perspective.

The distance-quantification concept is as well widely established for stratification and clustering of neoplastic entities [49]: Neoplasms with comparable phenotype are grouped based on their genomic properties to identify disease-causing genotype-to-phenotype relationships. Given the hypothetical scenario that clustering was applied to a population only consisting of two phenotypes and that a single mutation was responsible for the phenotype, a clustering based on the mutation would stratify the population identically to the phenotype. However, a dichotomic clustering is generally not possible and after grouping a population based on a set of genomic properties, various hypothesis have to be tested in order to identify the genotype-to-phenotype relationship. An illustration of how clustering extracts contextualized information by analyzing single data points as an ensemble is depicted in Figure 2.4.

Implementation of the distance-quantification

The quantification of a distance between two sequenced entities, x and y requires knowledge of their position in a metric space [50]. The positions are in turn being determined by the entities' respective set of features [45]. For illustration purposes, let x and y consist of one feature which can take the dichotomic values 0 or 1. The quantification of a distance $d \in \mathbb{R}_+$ is then defined as the mapping of a scalar to the space location-defining feature sets of x and y, $D(x, y) = d$ where D is the metric or distance function. Multiple calculation rules for D exist and one of the simplest is to apply the absolute value distance:

$$D(x, y) = |x - y| \tag{2.2.1}$$

Formula 2.2.1 illustrates that in case x and y are either both 0 or both 1 their distance is 0 and 1 in any other case. The dichotomic feature domain of 0 and 1 is, however, only applicable in few scenarios and a suitable and commonly applied calculation rule for features with a domain in $\mathbb{R}$ is the Euclidean distance rule. The calculation rule of the Euclidean distance for two sequenced entities is then defined as follows [51]:

$$D_{point}(x, y) := \sqrt[2]{(x - y)^2} \tag{2.2.2}$$

Equation 2.2.2 highlights a mathematical condition that has to hold for metrics: the limitation of distances to positive domains, here, achieved by squaring the difference $x - y$. The squaring induces, however, a super-linearity for the input domain differences $(x-y)^2$ ($D^2(1, 3) = 4$, $D^2(1, 4) = 9$) which is why application of the square root recovers the linearity (example: $D(1, 3) = 2$, $D(1, 4) = 3$). In the more realistic case of multiple features per entity, where $n \in \mathbb{Z}_+$ indicates the amount of features. The Euclidean calculation rule is defined as follows:

$$D(x, y) := \sqrt[2]{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \tag{2.2.3}$$

Note, that x and y are generally defined as vectors $\vec{x}, \vec{y}$ when more than one feature is present. The same index i in the vector refers to the same feature in different entities. For brevity, x and y will in the following refer to their vector-formulation $\vec{x}$ and $\vec{y}$ and any exceptions will be mentioned.

More formally, a non-negative distance d can be quantified between all points in the set X of a space if the space possess a distance metric D which induces a topology [50]:

$$D : X \times X \to d \tag{2.2.4}$$

$$d \in \mathbb{R}_0^+ \tag{2.2.5}$$

For any points x and y in X, the following conditions have hold for the distance mapping for D to be considered a (full) metric:

$$\text{Semi-positivity} := D(x, y) \geq 0 \tag{2.2.6}$$

$$\text{Identity} D(x, y) := 0 \leftrightarrow x = y \tag{2.2.7}$$

$$\text{Symmetry} := D(x, y) = D(y, x) \tag{2.2.8}$$

$$\text{Triangle inequality} := D(x, z) \leq D(x, y) + D(y, z) \tag{2.2.9}$$

Equation 2.2.6 requires a distance to be greater than or equal to zero. Equation 2.2.7 enforces that the distance is equal to zero if and only if (iff) x and y are identical. Equation 2.2.8 requires that the distance of x to y is identical to the distance from y to x. Equation 2.2.9 stipulates that the underlying space is flat in that the direct line connecting two points (the geodesic) has a quantified distance that is equal to or less than the length of any other (indirect) line connecting the same two points.

Multiple Bioinformatics algorithms, such as phylogenetic tree constructors utilize ultra metrics which require the additional condition $D(x, z) \leq max\{D(x, y), D(y, z)\}$. The ultra-metric condition is, however, not applied in this thesis and therefore not considered in the following. If metric conditions specified in equations 2.2.6 to 2.2.9 are not fulfilled, D can still qualify as (partial) metric, depending on the condition that is not fulfilled. Relevant to this thesis, the symmetry condition 2.2.8 is dropped in Chapters 3 and 4 where a quasi-metric is applied in equation 3.2.3.

In summary, an abstract semi-positive distance can be quantified between entities based on their $n-$dimensional respective set of features provided that the metric D is semi-positive, symmetric, identical sample of zero distance and the underlying space flat.

Vector norms

Chapter 5 follows the same geometric interpretation in that features (here mRNA expression levels) of sequenced entities are mapped onto a position in a space but differs in that a vector norm and not a Euclidean calculation rule is utilized for obtaining the scalar d. The first vector contains the gene expression features of a neoplasm and the second vector the reconstruction of the first vector conducted by a constraint ML algorithm. The difference vector obtained from subtraction of the two vectors possesses a geometric length. This length (but not orientation) is a quantification of the distance between the neoplasms and healthy cells. Note, that the orientation of the vector provides knowledge about the cell-types which make up the neoplasm and are informative when predicting clinical characteristics because the expression of marker genes, which is contained in the difference vector, is by definition characteristic for cell-types. A norm on the difference vector then serves to obtain a discrete and semi-positive distance-quantification. Importantly, note that the vector norm distance, as applied in this thesis, is not quantified between two entities existing in reality for example between CCLs but between one existing entity (the cancer) and a prototypic representative of e.g. an adult stem-cell. The hypothesis underlying this type of distance-quantification is that the adult stem-cell represent a cell-type with high proliferation potential. The smaller the distance of neoplasm is to the stem-cell, the greater the proliferation potential of the neoplasm is assumed to be which is why an informative correlation with the grading of the neoplasm could exist.

Metrics on vector spaces are a special case of general distance metrics presented in equation 2.2.5 [52]. A vector norm is defined as a mapping $||.||$ of a set of vectors V onto the set of semi-positive, real-valued numbers $\mathbb{R}_0^+$ for every vector $v \in V$ [45].:

$$||.|| : V \to \mathbb{R}_0^+ \tag{2.2.10}$$

$$v \to ||v|| \tag{2.2.11}$$

The norm $||.||$ has to fulfill the following conditions in order to qualify as a distance metric [53] [52]:

$$\text{Definite quadratic form } ||v|| = 0 \longrightarrow v = 0 \tag{2.2.12}$$

$$\text{Absolute homogeneity } ||\alpha \cdot v|| = |\alpha| \cdot ||v|| \tag{2.2.13}$$

$$\text{Subadditiveness } ||v + v'|| \leq ||v|| + ||v'|| \tag{2.2.14}$$

Equation 2.2.12 requires that a vector norm only maps a vector to zero if v itself is zero. 2.2.13 requires that the norm of v multiplied with vector α is identical to the vector norm of v multiplied by the absolute value of α, i.e. the $L1$ norm of α. Equation 2.2.14 requires that the vector norms of the addition of v' to v is less than or equal to the added vector norms of both v and v'.

Application of NGS data for distance-quantification

The application of distance metrics and vector norms from Sections 2.2 and 2.2 assumes that NGS data can either directly or after a transformation be plugged-in to the formulas as features while preserving the required conditions. The thesis posits that NGS data can indeed be plugged-in into the distance-quantification equation and information be derived from the metric due to the assumptions stated in table 2.1.

Assumption	Definition
Exclusivity	The determined numerical distance exclusively quantifies the biological distance between two entities
Zero identity	A distance $D(x, y) = 0$ is zero when the same organism was sequenced in absence of technologically confounding factors
Biased identity	A distance $D(x, y) \geq 0$ can be greater than zero and still x and y be identical when technical bias distorts the distance-quantification
Testability	A statistical test can determine whether a distance is sufficiently small to assume that x and y are biologically identical when a technological bias is present

Table 2.1: Assumptions of distance-based identification. The table specified which assumptions are made when assuming that sequenced entities can be identified via the quantification of an abstract distance between them.

See the contributions Chapters for a detailed modeling of the NGS-to-distance-quantification.

Note that the thesis utilizes the terms *distance* and *similarity* synonymously since they can be transformed into each other:

$$S(x, y) = \frac{1}{1 + D(x, y)} \tag{2.2.15}$$

$$D(x, y) = \frac{1}{S(x, y)} - 1 \tag{2.2.16}$$

Equation 2.2.15 transforms a distance metric into a similarity metric and equation 2.2.16 transforms a similarity metric into a distance metric provided that both D and S are euclidean in nature [54].

2.3 Statistical tests

Statistical tests render unto a scientist the possibility of taking informed decisions with respect to the inequality of properties of probabilistic distributions [55]. Statistical tests were applied by all scientific contributions and in the following, the hypothesis tests with greatest relevance to this thesis will be explained: the z, t, binomial and Logrank test.

2.3.1 Hypothesis testing

The intuition regarding a statistical hypothesis tests is that they allow to quantify the probability of committing a mistake when assuming the inequality of scalars or vectors. The tests are called *hypothesis* tests because they assume a so called null-hypothesis H_0 which is subsequently rejected or retained based on the probability to commit a mistake when rejecting H_0 [55]. If H_0 is rejected, the alternative hypothesis H_1, which states the inequality of parameters, is assumed. The value that the tested parameter is compared to is either a fixed scalar, in case of one sample location-tests, or alternatively, the estimated parameter of another probability distribution.

We commence by introducing the two-tailed one sample location z-test due to its simplicity and in order to suitably illustrate the general aspects of statistical testing. A one-sided test is limited to testing in either greater or smaller direction, a two-sided test tests simultaneously on both greater or smaller conditions. Given a set of measured realizations $x_1, x_2, \dots x_n$ of a distribution $\hat{X}$ approximated by sampling, the sample location z-test tests on inequality of the empirically approximated mean μ_x of $\hat{X}$ of another scalar M given the standard error SE of the samples drawn from the distribution. Note that SE specifically refers to the variance observed for the samples from X when approximating the true value of a parameter of interest, i.e. SE is not the standard deviation of the usually latent and intractable distribution. H_0 and H_1 are defined as follows for a two-sided one sample z-test:

$$H_0 := \mu_x = M \tag{2.3.1}$$

$$H_1 := \mu_x \neq M \tag{2.3.2}$$

H_0 is rejected if the probability of incorrect rejection is lower than a user-defined threshold, commonly referred to as α-level. The probability of an incorrect rejection is calculated based on the assumed distribution, a test-statistic which quantifies the statistical power of the difference between between μ_x and M given the standard error SE. The z-value is a quantification of the corresponding statistical strength of the relationship of these variables:

$$z = \frac{M - \mu_x}{SE} \tag{2.3.3}$$

SE is then weighted by the amount of observed point-wise data n:

$$SE := {}^{\sigma}\!/\!_{\sqrt{n}} \tag{2.3.4}$$

$$\sigma := \sqrt{\sum_{i=1}^{n} (\mu_x - x_i)^2} \tag{2.3.5}$$

Note that the z-test assumes that the real σ is known what, however, frequently is not the case and only the sample's variance can be quantified. The probability to reject H_0 although H_0 has, by ground-truth, to be retained is called the p-value. Differently formulated, the p-value provides the information how often a sample from a population would look like as was observed due to chance if H_0 was true. To calculate the p-value, the cumulative distribution function (CDF) of the probability density function (PDF) of z-scores has to be known. In a majority of cases and due to the central limit theorem of statistics, shown in equation 2.3.7, the PDF of the distribution of z-values is a Gaussian distribution with expected value $\mu = 0$ and standard deviation $\sigma = 1$ [55]:

$$N(\mu = 0, \sigma = 1, x) = \frac{1}{1 \cdot \sqrt{2\pi}} e^{\frac{1}{-2}(\frac{x-0}{1})^2} \tag{2.3.6}$$

The difference between μ_x and M is called *significant* when z-score $\geq$ z-critical holds, while the distance between the z-critical and the z-score is known as the statistical *power*. Parameter α is the upper bound of the probability to incorrectly reject H_0. α takes on values in the interval $[0, .., 1]$ over the body $\mathbb{R}^{+}_{\nvdash}$. The *critical value* is the smallest z-score that results in a p-value that is equal or smaller than α. The *critical value* is the CDF of the Gaussian distribution, labeled $\Phi(z)$ (Phi), resolved for the z-score at which the integral function equals α:

$$\int_{-z}^{z} N(\mu = 0, \sigma = 1, x)dx = \frac{1}{2}\left[1 + erf\left(\frac{x-0}{1 \cdot \sqrt{2}}\right)\right] = \Phi(z) = \alpha \tag{2.3.7}$$

The function *erf* refers to the Gaussian error function which quantifies the probability that the z-value falls within the range of $[-x, x]$. A convenient aspect of the z-test definition is that the *critical value* always equals 1.96 for a two-sided test with α-level of 5% due to the assumption A) that the standard deviation of X's population is known and B) that the CDF of the z-scores follows a Gaussian distribution [56].

Since α is an upper bound and zero the lower limit, a region of acceptance R_0 is directly created depending on the value that was chosen for α:

$$z_{crit} = \int_z^z N(\mu = 0, \sigma = 1, x)dx = \alpha \tag{2.3.8}$$

$$R_0 := (z_{crit}, \cdots, 0] \tag{2.3.9}$$

If the p-value falls within the interval of R_0, the test on $\mu_x = M$ will lead to a rejection of H_0 and $\mu_x \neq M$ will be accepted. In the case that H_0 was incorrectly rejected, a statistical error of Type-I, was committed while incorrectly retaining H_0 is a Type-II error.

2.3.1.1 t-test

The cumulative distribution function (CDF) of the error-distribution has to be known in order to resolve for the critical value of the test-statistic. An important reason why hypothesis testing is applied in a plethora of scientific domains is that the CDF that controls the Type-I error can be obtained via the central limit theorem [56].

The central limit theorem states that the errors that occur when estimating the potentially unknown expected value μ of a probabilistic distribution via empirical estimation of μ follow a Gaussian CDF $\Phi(z)$ [56]. Let the empirically estimated mean S_n be defined as $\sum_{i=1}^{n} x_i/n$. The properly normalized error of μ-estimation, Δ, then depends on the sample-size n and the difference between S_n and μ [57]:

$$\Delta := \sqrt{n}(S_n - \mu) \tag{2.3.10}$$

Application of the Lindeberg-Lévy theorem then quantifies Δ by utilization of $\Phi(z)$ from equation 2.3.7 assuming an asymptotic sample-size growth of $n \to \infty$ for any $z \in \mathbb{R}$ resolves to [58]:

$$\lim_{n \to \infty} CDF_n([\Delta \leq z]) = \Phi(z) \tag{2.3.11}$$

The CDF of the Δ of independently and identically distributed (i.i.d.) random variables can thus be expressed as follows for sufficiently large n. Note that μ is fixed but that σ^2 is a variable:

$$CDF_n(\Delta, n) \xrightarrow{n=\infty} N(\mu = 0, \sigma^2) \tag{2.3.12}$$

We summarize that the distribution of the errors required for determination of the critical z-value is generally a Gaussian CDF $\Phi(z)$ due to the central limit theorem. However, any test that assumes an approximate Gaussian distribution has to ensure that a sufficiently large sample-size is provided for equation 2.3.11 to be applicable and, in addition, that the specific value of σ^2 is known or can be approximated.

Here, we present the t-test which is frequently applied in the scientific domain of Bioinformatics. A commonality between the earlier introduced z-test and the t-test is that both assume the expected value of the tested probability distribution to follow a Gaussian distribution. An important difference is that the t-test empirically estimates σ^2 whereas the z-test assumes the population's σ^2 to be known. However, a z-test may still be applicable provided that the sample-size is sufficiently large to reliably estimate σ^2, what generally is assumed to be the case for a $n \geq {\sim}1000$ [57].

The t-test is of great importance for the domain of Bioinformatics and, in particular, for the analysis of differential gene expression [59]. In Section 5, differentially expressed genes between pancreatic neuroendocrine and exocrine cell-types are identified via a t-test.

The t-test tests on the possible inequality of parameters of two distributions X and Y. H_0 states that $\mu_X = \mu_Y$ and H_1 that $\mu_X \neq \mu_Y$. Here, we present a two-sided t-test for the basic case of equal variance between X and Y as well as equal sample sizes n_X and n_Y. In case of differing sample sizes and variances, a Welch test can be applied at the cost of test-sensitivity due to increased type-II errors (H_0 should be rejected but is retained) [60]. In the following, s^2 is utilized en lieu of σ^2 to distinguish an empirical variances from the population's variances that is generally unknown.

The t-test applies a test-statistic that is based on a t-distribution with single values of the distribution being referred to as eponymous t-values. These are calculated based on the difference of the tested parameters μ_X and μ_Y and their pooled standard deviation s_{pooled} [59]:

$$t := \frac{\mu_X - \mu_Y}{s_{pooled} \cdot \sqrt{\frac{2}{n}}} \tag{2.3.13}$$

The unbiased estimator for the pooled variance is then defined as follows:

$$s_{pooled} := \sqrt{\frac{s_X^2 + s_Y^2}{2}} \tag{2.3.14}$$

Identification of the critical-value t_{crit} requires resolving $\Phi(t)$ for the desired α-level. H_0 is rejected if $t \geq t_{crit}$ holds after correction of t_{crit} due to approximation of the unknown population's variance by subtracting two degrees of freedom from $n_X + n_Y$.

2.3.1.2 Binomial test

The Binomial test tests on the significant departure of an empirical distribution of realizations $x_1, x_2, \ldots x_n$ from their expected Binomial distribution. Chapters 3 and 4 utilize a Binomial test to determine whether an observed amount of matches between a query and a reference CCL can still be explained with the amount of matching variants expected due to chance based on an empirically approximated sample distribution.

The Binomial test requires that a tested probability distribution X takes on discrete and dichotomic values e.g. $X = 1$ or $X = 0$ and follows a Binomial distribution $B(p, n, k)$. Given $B(p, n, k)$, n indicates the amount of trials for success, k the amount of observed successes and p the probability of success with $q = 1 - p$ [61]:

$$B(p, n, x = k) = \binom{n}{x} p^x (q)^{n-x} \tag{2.3.15}$$

$$\binom{n}{x} := \frac{n!}{x!(n-x)!} \tag{2.3.16}$$

$$\{p | p \in \mathbb{R},\ [0 \geq p \geq 1]\}$$

Where ! indicates the factorial sign. The nomenclature of $\binom{n}{x}$ in place of the common convention of $\binom{n}{k}$ is chosen to establish clarity in following equations 2.3.17 and 2.3.18 where x is the probability of a partial sum to observe exactly x successes as opposed to observing 0 up to k successes for the test as a whole.

The Binomial-test thus quantifies the probability that an observed outcome of a Binomial distribution 2.3.15 was created with a chosen chance-of-success parameters p_0. The aim of the test is to reject or retain H_0 which states that the observed outcome was created with the chosen p_0 given the probability for an upper bound on the error of committing a type-I of α. The p-value of a Binomial test are the summed up point-wise probabilities of observing $0, \cdots, k$ or alternatively $k + 1, \cdots n$ successes. The here shown one-sided test either tests if the observed amount of successes k was too low or too high assuming a specific $p = p_0$:

$$\sum_{i=0}^{k} B(p, n, x = i) \leq \alpha, \quad p < p_0 \tag{2.3.17}$$

$$\sum_{i=k+1}^{n} B(p, n, x = i) \leq \alpha, \quad p > p_0 \tag{2.3.18}$$

For H_0 to be rejected we thus have to demonstrate that α-level is greater than the summed point-wise probabilities, however, depending on whether we test on p_0, either the lower tail $(0, \cdots, k)$ or higher tail $(k+1, \cdots n)$ of the Binomial distribution is summed over. There are multiple approaches to conduct a two-sided Binomial test on $p \neq p_0$ and here, we describe the simplest approach: the p-values for rejection of $p = p_0$ according to 2.3.17 and 2.3.18 are both calculated but followed by a halving of α to compensate for double-testing. A two-sided test thus defaults to plugin-in values for k and $p = p_o$ and testing the CDF of the Binomial distribution $F(p = p_0, n, x)$ on less than or equal to $^{\alpha}/_{2}$ (two-sided case):

$$Z(p_0, n, x) \sum_{i=0}^{k} F(p_0, n, x = i) \leq \frac{\alpha}{2} \vee \sum_{i=k+1}^{n} F(p_0, n, x = i) \leq \frac{\alpha}{2} \tag{2.3.19}$$

Critical-value determination is conducted in analogy to the z and t-test by resolving $F(p, n, x)$ for the desired α-level. A key difference to the t-test is that large sample-sizes are generally adversarial to the effective application of a Binomial test. For large sample-sizes, the underlying Binomial distribution approximates a Gaussian distribution due to the central limit theorem what would motivate the application of a z-test that is superior in sensitivity and possesses computational advantages [61].

2.3.1.3 Logrank-test

The non-parametric Logrank test is instrumental for clinical trials since the test allows to identify significant differences with respect to the efficacy of patient treatments. For instance, one can compare the effectiveness of an established drug-treatment with a novel but not sufficiently tested drug. The intention of the test would be to determine whether the novel drug should be adopted due to a significant increase in the overall patient survival time.

Advantages of the test are the absence of user-defined parameters and that the frequently occurring discarding of patients from the survival statistic can be adequately modeled via sample-censorship. The Logrank-test is central to Chapter 5 where it is demonstrated that classification-by-deconvolution can subtype cancer patients into cohorts that differ with respect to their overall patient survival time.

The test rejects or retains the null-hypothesis H_0 that the hazard functions $h_i(t)$ of c-many compared cohorts differ [62]:

$$H_0 := h_i(t) == h_i'(t) \tag{2.3.20}$$

$$i, i' \in \{i | i \in \mathbb{N}, 1 > i \leq \infty, i \neq i'\} \tag{2.3.21}$$

A hazard function is the frequency at which an event, e.g. the passing of a cancer patient, is expected to occur by chance within a given interval [63]. Importantly, the expected value of a hazard function $h_i(t)$ follows, by assumption, a hypergeometric distribution with a fixed set of parameters that is directly derived from the input data [62]. Since H_0 states equality of the hazard functions, the Logrank-test is designed to show that at least one hazard function does not utilize the same set of parameters given a risk that the assumption of diverging parameters is incorrect. Let N_j be the amount of possible events over all cohorts (e.g. surviving patients) and O_j be the marginal amounts of observed events (e.g. deceased patients), here shown for $c = 2$:

$$O_j := O_{i,j} + O_{i',j} \tag{2.3.22}$$

$$N_j := N_{i,j} + N_{i',j} \tag{2.3.23}$$

The variance-weighted difference between expected and observed events is quantified for all c cohorts. Importantly, the Logrank-test probes a time-series that consists of d equidistant intervals. The d weighted divergences are summed up as scalar which can then be attributed with a type-I error risk for H_0 rejection since its CDF is known.

Let there be d many intervals and c many cohorts with indicator variables j and i with j indicating an interval and i a cohort. Assume N_j to be the amount of patients that can still decease in interval j. O_j is, in contrast, the observed amount of events within interval j. The expected amount of events for each cohort i in interval j is then calculated as follows:

$$E_{i,j} = O_j \frac{N_{i,j}}{N_j} \tag{2.3.24}$$

The variance for interval j and cohort i is analogously approximated as:

$$V_{i,j} = E_{i,j}(\frac{N_j - N_{i,j}}{N_j})(\frac{N_j - O_j}{N_j - 1}) \tag{2.3.25}$$

The distribution of the differences between E_j and O_j asymptotically follows a Gaussian distribution due to the central limit theorem. The divergence of at least one cohort i is thus quantified by a z-value over all j as shown here [62]:

$$Z(O_{i,j}, E_{i,j}, V_{i,j}) = \frac{\sum_{j=1}^{J}(O_{i,j} - E_{i,j})}{\sqrt{\sum_{j=1}^{J} V_{i,j}}} \xrightarrow{d} \mathcal{N}(0,1) \tag{2.3.26}$$

Equation 2.3.26 follows from the Lindeberg-Lévy theorem stated in equation 2.3.11, for a sufficiently large J. Solving for the critical value, as demonstrated in equation 2.3.8, determines the threshold at which H_0 can be rejected with upper confidence bound α what can for instance indicate a survival advantages of the patients treated with the novel drug.

2.3.2 Empirical sampling and testing

Hypothesis testing is not possible when the error CDF is unknown, test-statistics for H_0 rejection unreliable or the underlying data *spurious* what leads to a loss of goodness-of-fit of the CDF to the error function [64]. *Spurious* data, as defined by Fricker et al. [65], refers to a mathematical relationship in which two or more events or variables are associated but not causally related, due to either coincidence or the presence of a certain third, latent factor. To overcome these challenges, empirical sampling and testing can be applied to obtain empirical p-values. We will present the basic theoretical principles of empirical sampling to demonstrate how the sampling enables empirical testing with methods that were applied in Chapters 4 and 5.

Empirical sampling is defined as the sampling of the distribution of type-I errors via empirical CDF approximation with the aim of obtaining p-values. The empirically sampled distribution can, by assumption, be utilized to inform about the probability of type-I errors or, alternatively, the confidence interval in which an unknown parameter is located with a given likelihood [66]. The key aspect of empirical distribution sampling is the precision with which a given CDF can be approximated by sampling. This sampling precision is described by the concept of confidence interval CDF bands by the Dvoretzky–Kiefer–Wolfowitz inequality (DKW), illustrated in Figure 2.5.

Given an intractable CDF $F(x)$ whose PDF was evaluated at n x-coordinates, the empirically approximating CDF function $F_n(x)$ is defined as follows[67]:

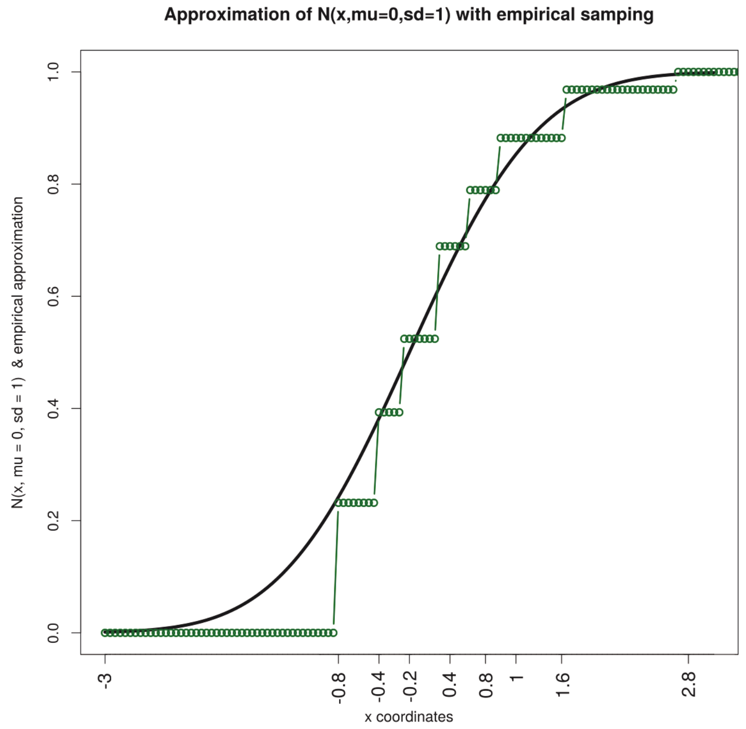

Figure 2.5: Example of an empirical distribution sampling with confidence intervals. The plot shows the empirical approximation of the CDF of a standard normal distribution via an empirical sampling from the PDF. The true CDF is the continuous orange line and the empirical sample the cyan stepped function. The flanking purple upper and lower stepped lines are example confidence intervals which bound the location of the true CDF based on the DKW what quantifies the precision with which a function can be empirically approximated. Source Figure [67]

$$F_n(x) :)\frac{1}{n}\sum_{i=1}^{n} 1_{\{X\leq x\}},\ x \in \mathbb{R} \tag{2.3.27}$$

Where 1 is the indicator function that equates to 1 when the condition is fulfilled and 0 otherwise. The precision of the approximation can then be quantified by the DKW depending the sample-size parameter n and the error ϵ :

$$pr\left(\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| > \epsilon\right) \leq 2^{-2n\epsilon^2}, \forall\epsilon > 0 \tag{2.3.28}$$

The CDF bands constrain the approximated function $F(x)$ into a tube which is composed of the empirical function $F_n(x)$ and an error term ϵ:

$$F_n(x) - \epsilon \leq F(x) \leq F_n(x) + \epsilon \text{ where } \epsilon := \sqrt{\frac{ln\frac{2}{\alpha}}{2n}} \tag{2.3.29}$$

Based on the tube-approximations via DKW, H_0 may be rejected or has to be retain if the tested parameter is located within an interval given the supremum on the upper bound of the acceptable type-I error limit α. In the following we show how three different empirical tests, applied in Chapters 4 and 5, reject H_0 via empirical p-value estimation.

Empirical parameter estimation

The Jackknife resampling technique for parameter estimation enumerates all possible leave-one-out subset input constellations. The aim is to evaluate these enumerated input-subsets in order to approximately learn the distribution of a parameter given the enumerations. The *resampling* term indicates that smaller subsets of an original samples are evaluated and *enumeration* indicates that all possible leave-one-out input subsample combinations are evaluated [68].

The Jackknife technique will be applied in Chapters 3 and 4 to determine the likelihood $\hat{x}$ to observe matching small variants between CCLs. The reason is that it can be shown that the expected value of $\hat{x}$, as obtained by resampling, is identical to the standard arithmetic mean [68]. The variance estimation is the key aspect of Jack-Knife resampling and calculated as follows:

$$Var(\hat{x}) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{x}_i - \hat{x})^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \hat{x})^2 \qquad (2.3.30)$$

The confidence interval of a parameter $\hat{x}$ can thus be determined as shown in equation 2.3.30 and application of DKW allows us to resolve for a critical value at which an observed CCL matching parameter x_j significantly diverges from the background distribution [69]:

$$Pr(x_j \geq \hat{x}) \leq \alpha \qquad (2.3.31)$$

The primary advantage of the Jack-Knife approach is its simplicity. A disadvantage is that Jack-Knife tests are not exact tests in that α-levels do not necessarily approach the true value of the tested parameter since only a resampling occurs. This implies that the true population can differ when the sample possess a different variance than the distribution population.

P-value correction via permutation sampling

In the following, the Westfall & Young permutation test for p-value correction is presented which finds application in Chapters 4 and 5. Permutation techniques differ from enumeration approaches in that distributions and not parameters are estimated. Furthermore, sampling occurs without replacement and permutation tests are generally *exact tests* which asymptotically equal the true value [70].

The Westfall & Young test infers the distribution of the type-I errors by perturbing labels of a given cohort (in contrast to Jack-Knife leave-on-out sampling that excludes samples). The perturbation therefore refers to the shuffling of the cohort-membership of the samples. The aim is to approximate the real distribution of p-values via the perturbation. After approximating the p-value distribution, the original p-value that relates to the tested cohort-definition is adjusted. The Young & Westfall test seeks to control the Family-wise error rate (FWER) while remain sensitive to H_0 rejection compared to approaches such as the Benjamini-Hochberg multiple-testing correction [71, 72].

Assuming that there are n phenotypes, there are n^m possible shuffles for m samples. Out of these shuffles, the ones with only a single phenotype are excluded, since the calculation of a difference requires at least two cohorts. We therefore obtain $n^m - 2n$ many possible case for the Westfall & Young test.

Let $b = n^m - 2n$ be the number of all allowed enumerations of the cohort-phenotypes of datasets X and p_0 be the initial p-value derived from the unperturbed cohort-labeling. The Westfall & Young in its *step-down max-T* version conducts the following procedures for the b^{th} permutation, $g = 1, ..., b$: [71]:

1. Shuffle the cohort-labels of the original matrix X b times to obtain perturbations
2. Compute the test-statistic $t_1, t_g, ..., t_b$ for each perturbation (g is an index variable)
3. Order t_g decreasingly to obtain monotonicity

The H_0-distribution corrected p-value $\hat{p}_{H0}$ is calculated as follows:

$$\hat{p}_{H0} := 1/b \sum_{g=1}^{b} I_{Indicator}(t_g | t_g \leq p_{H0}) \qquad (2.3.32)$$

A key disadvantage of the exact permutation tests are they can be computation-intensive. Despite the comparatively low size of class labels in common NGS data sets (compared to the amount of genes), a permutation-based approach can fail to terminate in high sample-size experiments, in which case one can limit the sample-size and, as a trade-off, decrease the test's accuracy. An additional disadvantage is that any set of samples drawn from the approximated distribution suffers from the *Behrens–Fisher* problem that states that the same variance has to be assumed for every class-label permutation for the test results to be exact what is frequently not possible [73].

P-value calculation via probabilistic empirical Sampling

Monte Carlo procedures refer to a group of methods which have in common that they sample repeatedly from a probability PDF to infer information about a parameter of the CDF when the CDF itself is unknown or intractable [68]. An advantage of Monte Carlo methods is their versatility in that they can be applied in various scientific domains since they are based on the evaluation of probability functions which are ubiquitously found in multiple scientific fields [74]. In particular, the empirical p-value estimation via Monte Carlo sampling has become an established procedure in the field of Bioinformatics [75]. Reasons for the utilization of Monte Carlo methods for p-value estimation are that Monte Carlo methods neither require exhaustive enumeration nor an asymptotic sampling assumptions [75]. Chapter 5 applies empirical Monte Carlo p-value estimation to quantify how much more likely a given transcriptomic deconvolution of matrix X is compared to the deconvolution of a randomly perturbed matrix X'.

Monte Carlo methods estimate the p-value via probing the domain of the p-value distribution CDF to learn the co-domain what informs about the location of the true p-value within a confidence interval [64]. Monte Carlo sampling is consequently defined as drawing without replacement until a desired width of the confidence interval is achieved based on a sample-size power-calculation [76].

Let M be the amount of Monte Carlo samples of size n $x_1, x_2, \cdots, x_n$ from a given distribution. Monte Carlo p-value estimation then quantifies the amount of times that samples resulted in a significant p-value k given a specified critical value threshold. The estimate of the $p-$value $\hat{p}$ is defined as follows [70]:

$$\hat{p} = \frac{1}{M}\sum_{j=1}^{M} k_j = \frac{k+1}{M+1} \tag{2.3.33}$$

The true p-value is then located in an encapsulating interval with width $\hat{\sigma}$:

$$\hat{\sigma} = \left[\frac{1}{M-1}\sum_{j=1}^{M}(k_j - \hat{p})^2\right]^{\frac{1}{2}} \tag{2.3.34}$$

With the probability for the interval defined by $\hat{p}$ following the Chebyshev inequality, e.g. 68% for 1 $\hat{\sigma}$[77]:

$$\text{Pr}(\hat{p} \pm \hat{\sigma}) = \hat{p} \pm \hat{\sigma} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{M}} \tag{2.3.35}$$

Disadvantages of the empirical Monte Carlo p-value estimation are that the p-value are assumed to follow a continuous, uni-modal distribution and that an uncorrected $\hat{p}$ can underestimate the true p-value. The p-value underestimation is a result of its Binomial expected value i.e. because it is not a distribution that is being sampled but the variable $\hat{p}$ [76]. The probability to underestimate $\hat{p}$ is calculated as follows [76]:

$$P(\hat{p} \leq \alpha) = \frac{\lfloor M\alpha \rfloor + 1}{M + 1} \tag{2.3.36}$$

To limit the underestimation as presented in equation 2.3.36 for the purpose of transcriptomic deconvolution, perturbation sample-sizes are generally set to 1000 by default what is as well the case for Chapter 5.

2.4 Deconvolution

Transcriptomic Deconvolution, also known as compressed sensing [78], is subject to ongoing intensive research and applied in Chapter 5 of the thesis where it contributes to the augmentation of neoplastic data for ML model training purposes. The intuition of a deconvolution is that one matrix can be factorized into two matrices whose multiplication recreates i.e. *reconstructs* the original matrix, see Figure 2.6. Deconvolution commences at the state where two matrices are given and the transcriptome, i.e. the original matrix, has to be reconstructed via a multiplication of the matrices. Importantly, one matrix has a width of one dimension, i.e. can be modeled as a vector, and its entries are not known at the onset of the reconstruction. A deconvolution algorithm chooses the entries such that the reconstruction is optimal as quantified by a reconstruction error. The such determined vector reprents the abundance of cell-types in the reconstructed transcriptome [34]. The abundance of the specific cell-types are considered clinically relevant since the relative cell-type proportions can potentially provide information regarding a neoplasm's clinical characteristics.

Definition

We will refer to the query, a neoplastic transcriptome, as vector C, to the cell-type expression signatures as matrix B and to the cell-type proportions as vector F. Matrix B is of dimensions # (amount of) genes $\times$ # cell-types and provides information on what marker-genes are characteristic for which cell-types. Vector F is of length # genes and indicates which cell-types are present according to the reconstruction of C of length # genes [80]:

$$C = B \times F + \epsilon,\ \epsilon \in \mathbb{R} \tag{2.4.1}$$
$$\text{with} \sum F = 1, F \in x | x \in [0, \ldots, 1], x \in \mathbb{R}0$$

The reconstruction illustrated in equation 2.4.1 assumes that C can be reconstructed by a linear combination of a semi-positive vector F with the marker gene signature matrix B while accepting an in-

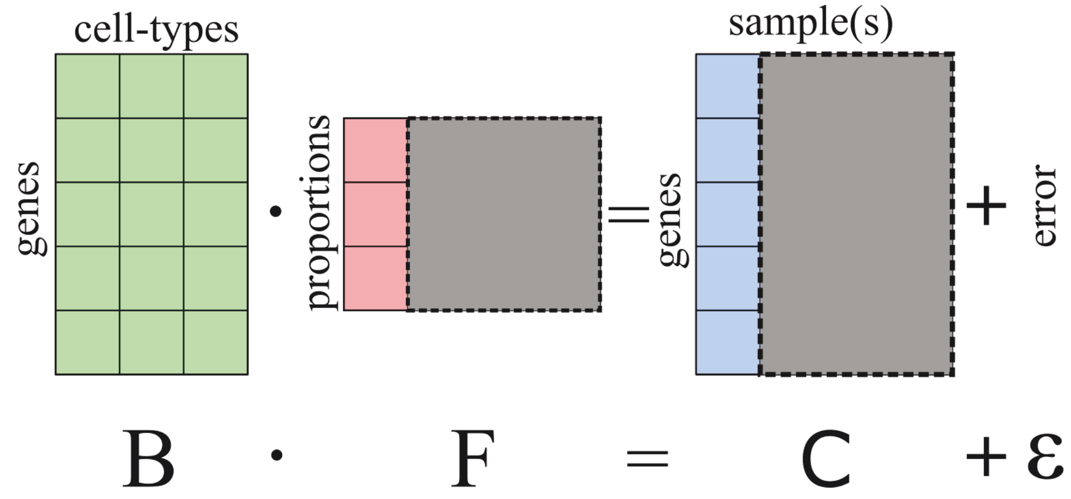

Figure 2.6: Deconvolution matrix factorization. The figure illustrates that transcriptomic deconvolution is based on a matrix factorization with error term. Matrix B consists of marker genes which distinguish cell-types. B has a width of the amount of cell-types and a height of the amount of genes. The initially unknown matrix (vector) F is determined by a deconvolution algorithm such that the (normed) distance to the original matrix C becomes minimal depending on the target function which quantifies the error. Gray areas serve to distinguish the commonly applied single-sample deconvolution from the matrix formulation with a width greater than one, i.e. the deconvolution is applied to single samples while the defactorization is defined for a multi-sample contexts. The names of the matrices are chosen such that they match equation 2.4.1. The figure was procured from [79] and modified.

curred error-term ϵ (Epsilon), $\epsilon \in \mathbb{R}$ [81, 82, 83]. A deconvolution consists of a training and a prediction phase: During the training phase C and F, whose cell-type proportions make up a convolute, are known. Solving equation 2.4.1 for B finalizes the conceptually simple training phase. Challenges associated with deconvolution generally center in the process of how B is obtained and a plethora of approaches that include, but are not limited to, unsupervised Monte Carlo maximum-likelihood [84] and supervised pseudo-matrix calculation have been applied to solve for B [85].

During the prediction phase, C and B are known, but F is unknown and equation 2.4.1 is solved for F. This is defined as actual *deconvolution step* because vector F is calculated during the reconstruction of C. The intricate element about a deconvolution is that a reconstruction is almost always possible, if ϵ is not constraint. Therefore, an empirical p-value is utilized to evaluate how non-random and statistically likely a result is. Commonly calculated by empirical resampling, the p-value quantifies how likely a randomly composed C would have resulted in the same reconstruction error, described in detail in Section 2.3.2.

Based on Equation 2.4.1, the reconstruction of C is generally formulated as an optimization problem by introducing a loss function L that quantifies the reconstructive error ϵ. During training, the optimization problem is to minimize ϵ via optimal choice of B with fixed F. For the prediction, B is fixed and F is iterated over to reconstruct C with minimal error [86]:

$$\min L(|C - (B \times F)|) \tag{2.4.2}$$

This optimization problem can be solved by various algorithms. The choice of the deconvolution algorithm and scRNA training dataset is critical due to the significant differences between the algorithms that may be suitable for different combinations of training and query data [35]. Algorithms differ primarily with respect to their predictive performance for different NGS technologies because they were developed for and benchmarked on a limited selection of technologies, most often Bulk RNA-sequencing and mRNA-arrays. Moreover, run time and RAM footprint differ greatly as well [87].

Properties and Constraints

To reduce both run time and avoid numerical problems while reduction the RAM footprint, feature-reduction algorithms are commonly utilized. These feature reduction algorithms are subdivided into algorithms which require the manual pre-selection of features and semi-supervised algorithms that determine a subset based on an optimization criterion [88, 82]. The manually selected subset features of the supervised approach are features that are generally statistically associated with only a single cell-type and can thus be identified via a one-versus-all differential expression analysis. Due to the prefiltering, B from equation 2.4.1 generally contains less genes than the query C because it only composes of genes that are characteristic for cell-types. The deconvolution is therefore restricted to the marker genes what may cause thousands of gene in C to be ignored for deconvolution purposes. Nonetheless, effective deconvolution in even as few as ten marker genes has been demonstrated for healthy pancreatic tissue and the determination of tumor immune-cell infiltration [35].

An important aspect that explains why deconvolution gained scientific momentum from 2015 onwards are improvements with respect to the marker gene selection [34]. Before 2015, marker genes had to be selected such that they avoid co-linearity constraints what reduced the amount of suitable genes [89]. Co-linearity (sometimes also called multi-linearity, [90]) is defined as a situation where two or more coefficients in a statistical model are linearly related [89]. Due to co-linearity, parameter estimates may be unstable, standard errors of estimates inflated and consequently inference statistics biased [91]. The co-linearity constraint was overcome by algorithms published from 2015 onword, such as Bulk Sequence single-cell deconvolution analysis pipeline (BSeq-sc) and MUlti-Subject SIngle Cell deconvolution (MuSiC) [34, 28] which are utilized in Chapter 5. The ramification of the dropped constraint was that marker genes may be expressed in more than one cell-type (what may imply their correlation) what dramatically increased the statistical power of the deconvolution [84].

p $>>$ n Problems and Over-Determination of B

Algorithms for transcriptomic deconvolutions are designed to address '$p >> n$' problems (p significantly greater than n) where p is the amount of features (here genes) and n the amount of samples. '$p >> n$' problems entail that the amount of features p is significantly greater than the amount of samples n [92]. '$p >> n$' problems therefore imply that the mathematical solution is over-determined and a feature reduction in the form of a Linear Dimensionality Reduction (LDR) technique may be required.

Consequently, the cell-type signature matrix B has to be over-determined, i.e. comprise of more genes than samples, to predict cell-type proportions [83]. In high-throughput sequencing settings with several thousands of genes, over-determination is generally assured, but mRNA panel sequencing approaches that cover a small amount of genes can restrict the applicability which is why the thesis only focuses on the deconvolution of over-determination-assuring Bulk-RNA sequencing technologies [34].

2.5 Algorithms for Transcriptomic Deconvolution

Various types of ML algorithms are commonly applied for deconvolution, including supervised kernel-based methods and linear factorization as well as unsupervised likelihood-maximization approaches [28, 34]. The thesis will, however, be limited to the supervised algorithms relevant to Chapter 5, namely the ν-SVR, derivative of the C-Support Vector Machine (SVM) algorithm and the NMF algorithm. We will introduce the C-SVM algorithm first due to its historical precedence and more intuitive formulation followed by a highlighting of commonalities and differences to the applied ν-SVR. Secondly, the NMF algorithm will be introduced due to its widespread utilization for deconvolution in addition to its utilization by the scientific contribution of this thesis.

Support Vector Machines

C-SVM algorithms are a class of widely utilized ML algorithms whose purpose it is to decide on the binary classification problem of whether a given sample is part of one class or another [15]. Given a d-dimensional space over the set of $\mathbb{R}^d$, the C-SVM, in its original formulation, classifies samples via the introduction of a $d-1$ dimensional hyperplane that partitions the space and the contained samples [93]. Intuitively, the coordinate equation of a hyperplane H in a three-dimensional euclidean space is spanned by two vectors $\vec{\mu}, \vec{u} \in \mathbb{R}^3$ that define the edges of the infinite hyperplane (third vector feature set to 0 for dimension $d-1$). H is then defined as set of locations whose positional vector $\vec{x}$ fulfills the following equation for an arbitrarily chosen but fixed $z \in \mathbb{R}^1$:

$$H := \{(\mu, u, z) \in \mathbb{R}^3 | x_1 \cdot \mu + x_2 \cdot u = z\} \tag{2.5.1}$$

Where $\cdot$ denotes the component-wise multiplication. For the generic, less intuitive d-dimensional case, the Hesse normal form based on a normal vector $\vec{w}$ is commonly chosen. Vector $\vec{w}$ is determined by solving the scalar product $\langle w, \vec{\mu}, \vec{u} \rangle$ equation for a vectors that is a member of the set of vectors whose scalar product with H is zero, i.e. which is pointing orthogonally to H. Next, an intercept scalar, as well called 'bias' term, b is determined to connect the normal vector to the origin of the coordinate system. Given both $\vec{w}$ and b, we can define the hyperplane H coherently to the commonly utilized SVM literature [94]:

$$H := \{\vec{x} \in \mathbb{R}^d | \langle \vec{x}, w \rangle + b = 0\} \tag{2.5.2}$$

For brevity and to obtain coherence with the literature reporting on SVM theory, the normal vector $\vec{w}$ will in the following be referred to as w without vector indicator marks. Importantly, vectors $\vec{x}$ which do not fulfill equation 2.5.2 have a scalar product $\langle \vec{x}, w \rangle$ that is either positive (geometrically located *above* H) or negative (geometrically located *below* H). A C-SVM can therefore, in theory, partition the space and separate all samples of class one in one subspace and the other samples in a different subspace of $\mathbb{R}^d$, by choosing H, such that the scalar product indicates a class membership via its sign. The samples that are closest to H are the eponymous Support Vectors a_i which enclose H. The space between H and a_i is called a margin. The width of the margin between between H minus the bias and any a_i is 1 since H is chosen such that the scalar product $\langle a_i, w \rangle$ minus the bias amounts to either 1 or -1 with no other training sample located within the margin by definition of a_i, see Figure 2.7.

The standardization $\frac{2}{||w||}$ of the margin's width allows to easily determine whether a given samples is located within the margin between H and a_i. The choice of H, during the training phase, is such that no other samples has a corresponding scalar product of equal or less than one (hard-margin definition) except for the Support Vectors which will be utilized as constraints in the following optimization equation for w. Given that a prediction y_i of class membership ($y = 0 \vee y = 1$) is to be made for a new sample x_i, the prediction is formulated as follow:

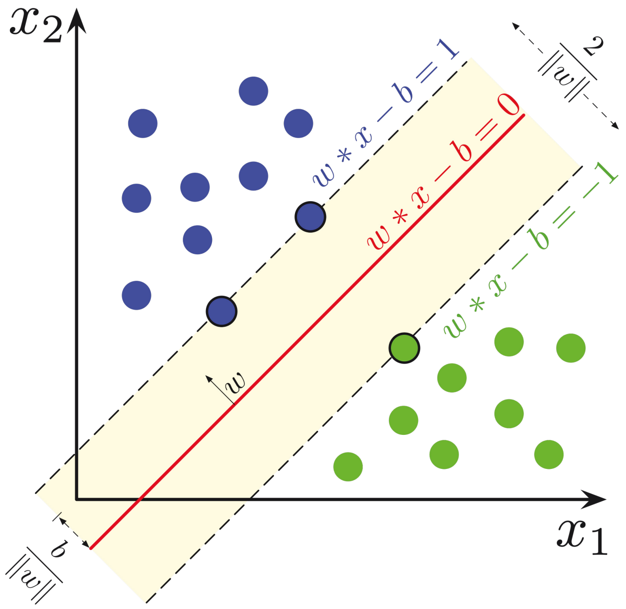

Figure 2.7: Classification Concept of a Support Vector Machine. The partitioning hyperplane H is defined as the set vectors $\vec{x}$ whose scalar product (scalar product operator denoted by $*$) with the normal vector w and a bias vector b is zero (scalar product brackets and vector arrows omitted for clarity in the Figure). Any samples' scalar product with H is either positive, negative or zero what indicates the class membership due to a location within one of the partitions if the product is not zero. During the training phase, H is chosen such that the samples closest to H have a scalar product of one and these samples are the eponymous Support Vectors a_i. The concept of a SVM-based binary classification is therefore to determine H such that all training data samples of opposite classes are optimally separated via a partition of the space (hard-margin definition). New, to-be-classified samples can be classified based on the sign of their scalar product with H. Source Figure [95]

$$y_i = sign(w \cdot \vec{x}_i - b) \tag{2.5.3}$$

In the hypothetical case that a linear separation is possible and every training sample can be correctly classified via equation 2.5.3, a C-SVM with hard margin can be applied, i.e. no incorrect classification occurs and H perfectly partitions the space. A hard margin C-SVM, however, is generally not possible and violations of the margin have to be accepted for practical purposes what leads to the definition of a soft margin C-SVM. A soft margin C-SVM first introduces an error function $loss_{err}$ to quantify the degree of incorrectly located samples. One of the most frequently chosen error functions is the hinge-loss function [96]:

$$loss_{err} := \frac{1}{n} \sum_{i=1}^{n} max(0, 1 - y_i(w \cdot \vec{x}_i - b)) \tag{2.5.4}$$

Theoretically, w could be chosen such that equation 2.5.4 becomes minimal for a given set of training samples i.e. that no sample available during the training phase violates the margin with a scalar product with H of less than one. An illustrative yet highly hypothetical approach that ensures such absence of errors is to first define H according to a loss function of choice to subsequently define all violating datapoints as support vector to achieve margin-clearance. The trade-off is, however, that the complexity of w as commonly measured by the square norm of $||w||^2$ increases what has the negative effect that the C-SVM model overfits on a given training dataset i.e. will not perform well on different datasets. Therefore, a trade-off between sparsity of the C-SVM model i.e. a reduction of $||w||$ and the minimization of the loss function 2.5.4 is required. The trade-off is parameterized by the C variable that penalizes the classification error during training time:

$$\text{minimize } C \cdot loss_{err} + ||w||^2 \tag{2.5.5}$$

$$\text{subject to } C \in [0, \ldots, \infty], C \in \mathbb{R}$$

Importantly, increases in the C parameter lead to more complex models that utilize more training samples as Support Vectors. Therefore, the standard C-SVM differs from the ν-SVM via the choice of w to minimize equation 2.5.6 what will be further elucidated in Subsection 2.5.0.1.

Given the minimization function 2.5.6, w can be determined via a numerical optimization algorithm which commonly implies the formulation of a Lagrange optimization function. A Lagrangean formulation of the optimization function requires a reformulation of equation 2.5.6 in that 2.5.4, $loss_{err}$, is reformulated as ξ_i (Xi) which quantifies the margin-violation of a given $\vec{x}_i$ [94]. The lagrangean primal problem formulation is then defined as follows for n training samples [15]:

$$\text{minimize } \frac{1}{n}\sum_{i=1}^{n} \xi_i + \frac{1}{2}||w||^2 \tag{2.5.6}$$

$$\text{subject to } y_i(w \cdot \vec{x}_i - b) \geq 1 - \xi_i \geq 0, \forall i \tag{2.5.7}$$

$$\xi_i \geq 0, \xi_i \in \mathbb{R} \tag{2.5.8}$$

The corresponding dual formulation is omitted for brevity. The linearity of the hyperplane H is, however, the key constraint of an SVM in that non-linearly separable datapoints can only be partially or not at all be linearly separated. Crucially, the SVM in its original formulation [97] introduces the

Kernel trick, a projection of the input space onto a generally higher dimensional space in which a lower-dimensionally linear hyperplane can effectively separate samples non-linearly in the higher-dimensional space. The corresponding kernel function $k(\vec{x}_i, \vec{x}_j)$, however, cannot be arbitrarily chosen and is limited to the constraint of being positive semi-definite [98]. The reason for this constraint is primarily motivated by computational complexity aspects since a non positive semi-definite kernel function would generally require an update of the higher-order projections of all samples x if a new samples was to be added. A semi-positive kernel function allows to assume linearity of the kernel projection via addition of a linear mapping operator ϕ (phi) which preserves the scalar product result of the mapping without re-calculation of all higher-order embeddings $k(x, x_i) = \langle \phi(x), \phi(x_i) \rangle$. Consequently, constraint 2.5.8 is redefined as follows for the kernel-trick:

$$y_i(w^T \phi(x_i) + b) \geq \xi_i \tag{2.5.9}$$

Such trained SVM models can effectively address a plethora of binary classification problems [99].

2.5.0.1 ν-Support Vector Machine Regression

In Chapter 5, a ν Support Vector Machine Regression was applied which is a derivative of the original C-SVM formulation and differs in the type of problem that is solved (regression versus classification) and the prioritization of the training error (ξ) over the model complexity (C). First, the regression versus classification aspect will be highlighted and secondly the ν parameter. Note, however, that both aspect are independent from each other, i.e. ν-SVM formulations are possible as well as C-SVR models.

The standard C-SVM classifies samples as member of two classes via the introduction of a hyperplane H by establishing a margin between H and Support Vectors a_i where the margin, in an ideal scenario, does not contain any other datapoints. A SVR shares the property of establishing a hyperplane H over a set of training samples but in contrast to a SVM attempts to contain as many training samples as possible within its margins with a minimally normed $||H||$ [100]. H in its formulation as orthogonal normal vector w is utilized to predict the value of a generally unknown function $f(x)$ given a specific input x:

$$f(x) = \sum_{i=1}^{n} w^T \cdot \vec{x} + b \tag{2.5.10}$$

$$f(x) \in \mathbb{R}, w^T, \vec{x}, b \in \mathbb{R}^d$$

The advantage of a regression-formulation as presented in equation 2.5.10 is that a SVR-algorithm is outlier-robust and computationally superior to comparable ML models [99] since the kernel-trick can be seamlessly reapplied to the regression:

$$f(x) = \sum_{i=1}^{n} w^T \cdot \phi(\vec{x}) + b \tag{2.5.11}$$

Identical to the default SVM hard versus soft-margin definition, an error term ξ_i is required to compensate for violations of the margins. Note, that violations occur not between the Support Vectors and H as is the case for a C-SVM but outside of the infinite tube spanned by the vectors. Since the violation term ξ differs depending on what side of H the violation occurred, an asterik (*) will in the following serve to distinguish the sides of the hyperplanes where the violation occurred. Note that the violation term of the C-SVM was symmetric since the distance to H remained invariant to the direction of the violation while the distance of an incorrectly regressed sample to the margins on either side will generally differ. The minimization problem of a C-SVR is therefore formulated as follows [98]:

$$\text{minimize } \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i + \xi_i^* \tag{2.5.12}$$

$$\text{subject to} \begin{cases} f(x_i)\langle w^T, x_i\rangle - b \leq \xi_i \\ \langle w^T, x_i\rangle + b - f(x_i) \leq \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

Based on the primal optimization formulation 2.5.12, the corresponding dual formulation can be calculated analogously to the classification formulation, however, for brevity we will only present the primal problem formulation. In summary, the objective functions of a SVM and a SVR differ but the underlying hyperplane concepts remains comparable.

ν versus C-SVM

The ν parameter differentiates the applied SVM model from the default model presented earlier and primarily modifies how the optimization presented in equation 2.5.7 is solved, i.e. the choice of H differs although the underlying optimization formulation remains the same [94]. The primary motivation to utilize ν is to obtain an optimization formulation where as few as possible features have to be utilized for the problem formulation while limiting the error during training time [93]. Given the context of Bioinformatics with data types such as e.g. gene expression data that includes ten-thousands of genes or more, the ν formulation is frequently prioritized over C formulation to reduce the calculation complexity.

As shown in equation 2.5.12 for the SVR and equation 2.5.6 for the SVM, parameter C acts as a penalty term for incorrectly classified or regressed samples. An increase of C will generally reduce the training error but render the model more complex as measured by $||w||^2$ since more samples are utilized as Support Vectors. C, however, is not directly interpretable since it is generally unclear to what extent the model complexity and the training error will increase when C is changed [99]. In contrast, the ν formulation is interpretable because ν serves as an upper bound on the fraction of margin violations and as a lower bound on the amount of samples utilized as Support Vectors [99, 94].

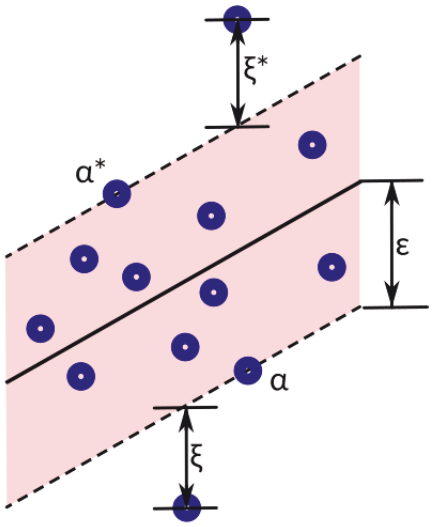

Figure 2.8: ϵ-independent SVR loss function. As proven by the Karush-Kuhn-Tucker condition (KKT) equation, the area bounded by - and +ϵ and the samples contained within the tube be can ignored, merely the slack variables ξ have impact on the optimization equation, what reduces computational complexity and helps to identify the optimal set of support-vectors [101].

Importantly, the ν formulation applies the concept of an ϵ (epsilon)-insensitivity which is defined as the omission of training samples from the primal optimization equation iff the samples are located within a distance ϵ around H [102]. The key concept of the ν parameter is to choose ϵ such, that aforementioned constraints on the amount of Support Vectors and the error during training time are preserved [99], see Figure 2.8.

The corresponding ν-optimization problem is formulated as follows for slack variable ξ_i, ξ_i^* and the ϵ width:

$$\text{Minimize } \frac{1}{2}||w||^2 + C \cdot \frac{1}{n}\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$\text{subject to} \begin{cases} (w^T \cdot x_i + b) - f(x_i) \leq \epsilon + \xi_i \\ f(x_i) - (w^T \cdot x_i + b) \leq \epsilon + \xi_i^* \\ C, \xi_i, \xi_i^* \geq 0 \end{cases}$$

A well defined loss function L_ϵ states that, based on the KKT proof, only residuals greater than ϵ affect the choice of H and in turn the definition of the normal vector w [103]:

$$L_\epsilon = \begin{cases} 0, \text{if } |f(x) - w \times F| \leq \epsilon \\ |f(x) - (w \times F)| - \epsilon, \text{else} \end{cases}$$

The ϵ-constraint on w, shown in equation 2.5.13, is labeled $\epsilon - insensitivity$ because the ϵ parameter introduces an $\epsilon - tube$ around the hyperplane within which residuals do not affect the loss function, illustrated in Figure 2.8. As outlined by the KKT-proof the ϵ-independence reduces the computational complexity since the samples within the ϵ-tube can be omitted from the optimization equation [101]. The ν parameter is then solved for C and the ν-SVR redefines the optimization problem as follows:

$$\text{Minimize } \frac{1}{2}||w||^2 + C(\nu\epsilon + \frac{1}{n}\sum_{i=1}^{n}(\xi_i + \xi_o^*))$$

$$\text{subject to} \begin{cases} (w^T \cdot x_i + b) - f(x_i) \leq \epsilon + \xi_i \\ f(x_i) - (w^T \cdot x_i + b) \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^*, \epsilon \geq 0 \end{cases}$$

The optimal solution for the loss function in equation 2.5.13 is, analogously to the standard C-SVM, found via Lagrange optimization with Lagrange variables $\alpha, \eta, \beta \geq 0$. Equation 2.5.13 utilizes β is a weight on the classification error ϵ, η as weight on the slack variables ξ and α a margin-size weight on w. The ν-SVR optimization problem has to be solved for w, b, $\xi_i^{(*)}$ [102]:

$$\begin{aligned} L(w,b,\alpha_i^{(*)},\beta,\xi^*,\epsilon,\eta_i^{(*)}) = &\frac{1}{2}||w||^2 + C\nu\epsilon + \frac{C}{n}\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &- \sum_{i=1}^{n}\alpha_i(\xi_i + y_i - (w^T \cdot x_i) - b + \epsilon) \\ &- \sum_{i=1}^{n}\alpha_i^*(\xi_i^* + (w^T \cdot x_i) + b - y_i + \epsilon) \\ &- \beta\epsilon - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*) \end{aligned} \tag{2.5.13}$$

The ν-SVR kernel-projection into a higher-dimensional space is formulated analogously to the C-SVR, the kernel being indicated by the variable κ (Kappa). κ induces the transformation-invariant dot product $\langle x_i, x_j \rangle$ of the feature space [97]:

$$\kappa(x,y) = \phi(x) \cdot \phi(y) \tag{2.5.14}$$

In summary, given the complex regression problems within the domain of Bioinformatics, the ν-SVR has found ample attention due to its dimensionality-reduction property and limitation of the error during training time, thereby facilitating the training of overfit-robust models for e.g. the transcriptomic deconvolution [34, 86]. Distance-quantification concepts with respect to a transcriptomic deconvolution conducted by a ν-SVR algorithm are involved in both the methodological and clinical aspects. The methodological aspect refers to the distance of samples on the hyperplane H and whether they are located within the margin or ϵ-tube.

2.5.0.2 Non-negative matrix factorization

Here, we present the Non-negative Matrix Factorization (NMF) algorithm due to its utilization as transcriptomic deconvolution ML algorithm in Chapter 5 of this thesis and its widespread application for deconvolution in general. The NMF algorithm, in its original design, will be presented, followed by an outline of selected aspects which are particularly relevant to this thesis.

2.5.0.3 Definition NMF

A Non-negative Matrix Factorization (NMF) algorithm solves a linear multivariate problem that addresses the factorization of a matrix into semi-positively defined matrices [104]. Key aspects of a NMF algorithm, that will be discussed in this section 2.5.0.2, are the non-negativeness of the factorization products, its target function, matrix sparseness, eventual convergence and its initialization problem.

The basic principle of a NMF shows resemblance to the deconvolution definition from Equation 2.4.1 and is defined as reconstruction of a given input by numerically estimating two entities whose product approximates i.e. reconstructs the input. More precisely, given three matrices, C,B and F, let the following matrix definitions apply:

$$C \in \mathbb{R}_{+}^{m \times n}, \quad B \in \mathbb{R}_{+}^{m \times k}, \quad F \in \mathbb{R}_{+}^{k \times n} \tag{2.5.15}$$

A NMF is then defined as the factorization of C into B and F [105]:

$$\begin{aligned} C &= B \times F + \epsilon, \epsilon \in \mathbb{R} \\ B &\in \mathbb{R}^{m \times k},\ F \in \mathbb{R}^{k \times n} \\ B,F &\geq 0, k \leq min(m, n) \end{aligned} \tag{2.5.16}$$

The term of $B \times F + \epsilon$ is called the Non-negative Matrix Factorization (NMF) of C [106]. The matrices are labeled C, B and F to indicate continuity with equation 2.4.1 in Section 2.4 where C is the to-be-deconvolved transcriptome, B the cell-type signature matrix, ϵ the error term and F the vector of cell-type fractions.

The optimal choice of k, is an assumption regarding the amount of latent factors that generate C. k is a central aspect of the NMF and its value can generally not be efficiently determined apriori [107]. Publications frequently utilize a mixture of domain specific knowledge to render the choice of k biologically interpretable. A numerical optimization of k can, for instance, take place via an Evidence lower bound (ELBO) statistic target function which is an approximation of an observed distribution with a predicted, learned distribution [108]. Alternatively, the cophonetic correlation can be used which quantifies how well observed data can be reproduced. The key element of cophonetic correlations is, that a dendrogram is utilized which quantifies how faithfully pair-wise distances, observed in the input data, can be reproduced with the learned NMF model [109].

An important aspect of the NMF algorithm is its choice of goodness-of-reconstruction target function D which governs the results for matrices B and F in equation 2.5.17 [104]. The reconstruction-error function D has to be (1) continuously differentiable, (2) convex in B and F and (3) zero iff $C = B \times F$ [106]. The aim of a NMF is then to minimize D:

$$\text{minimize } D(C, B, F) \quad B \geq 0, \quad F \geq 0 \tag{2.5.17}$$

Note, that B and F serve as input variables for D in contrast to C which parameterizes $D(C, B, F)$. Nonetheless, D will in the following refer to the distance of a norm on C and on $B \times F$ for simplicity.

From the perspective of Bioinformatics, an interpretable factorization of C is of great importance and generally achieved if the matrix B can compress and group latent information from C with F providing information on the relative importance of the information contained in B [110]. For example: In the case of a transcriptomic deconvolution, B would contain the cell-type expression signatures that together constitute the transcriptome while F would harbor the relative proportions of the cell-types within the convolute. The achieve the feature-compression in B, a sparse solution is generally required when equation 2.5.17 is solve. The sparseness is integrated into the solution via the introduction of norms on B and F as follows [105]:

$$\begin{aligned} &\text{minimize } ||B|| \times ||F|| + \epsilon \\ &\text{such that } = ||B|| \leq \lambda, ||F|| \leq \lambda, B, F \geq 0, \lambda > 0 \end{aligned} \tag{2.5.18}$$

Where $\lambda \in \mathbb{R}+$ is a user-defined sparseness-inducing variable that acts as a direct trade-off between the training error and model sparseness. The numerical decomposition of C requires solving a bi-convex problem since equation 2.5.21 is either only convex in B or only in F what denies the identification of a global minimum by naive optimization [110]. However, there are multiple ways to solve the NMF and the most commonly utilized are the multiplicative update rule, the gradient descent approach and the alternative least-squares method [111]. Here, we present the multiplicative update rule since it is generally considered to be the first published NMF decomposition method [112]:

$$F \leftarrow F \frac{(B^T \times C)}{(B^T B \times F)} \tag{2.5.19}$$

$$B \leftarrow B \frac{(C \times F^T)}{(B \times F F^T)} \tag{2.5.20}$$

Equations 2.5.19 and 2.5.20 illustrate that the bi-convex problem can be solved despite the absence of a closed single-update step formula by iteratively fixing either F or B and a subsequent update of the other matrix. It is noteworthy that the numerically approximated matrix B is chosen such that B consists of pair-wisely independent vectors what is commonly referred to as *clustering* property of NMFs since the information of different latent structures in C is clustered in the orthogonal vectors of B [113]. The decomposition is finished once the updates of the matrices converge or numerical changes of the matrices' entries are smaller than a specified threshold.

An important drawback of the NMF decomposition is that the factorization, as shown in Equation 2.5.17, might not be unique [114] and depends on the initialization of B and F what is known as *initialization problem* [107]. Publications have reported that equations 2.5.19 and 2.5.20 converge against a local saddle point, however, not necessarily against the global minimum which is why multiple factorizations are computed, each with a different initialization to increase the chances of identifying the global minimum [115]. The default procedure with respect to the NMF initialization is to randomly initialize B and F with a limited set of positive non-zero entries such that a rapid convergence is ensured [115].

2.5.0.4 NMF for Transcriptomic Decomposition

Due to its sparseness, guaranteed convergence and clustering-properties, the NMF is a widely applied algorithm for LDR and $'p >> n'$ problems in general and transcriptomic deconvolution in particular as reported by Moffitt et al. whose algorithm is utilized in Chapter 5 [116].

The NMF utilized in Chapter 5 differs primarily in the formulation of the target function while the remaining properties and definitions remain comparable to a standard NMF model. The target function D was based on the squared L2-norm for i genes and j samples and defined as follows:

$$D^{L2}(C, B, F) := ||C - B \times F||^2 = \sum_m \sum_n (C_{mn} - [B \times F]_{mn})^2 \tag{2.5.21}$$

Where $[B \times F]_{mn}$ denotes the mnth entry of the matrix product of $B \times F$ [117]. The squared error function ensures a balanced convergence of both B and F while improving the grouping and sparsity of the solution [117].

A deconvolution model that utilizes a NMF algorithm commences by training on the scRNA input data to determine matrix B. Matrix C, consisting of the purified cell-type data, is then reconstructed until convergence of $B \times F$ is achieved as defined in the multiplicative update equations 2.5.19 and 2.5.20. During run time with a new C', the equation is generally solved for F via creation of the pseudoinverse $pinv()$ of B via transformation of equations 2.5.21 to $C' \times pinv(B) = F$. F then represents the cell-type proportion predictions.

Chapter 3

Identification of Exome-Sequenced Cancer Cell Lines via Distance-Quantification

This Chapter illustrates how the distance-quantification concept can address the challenging problem of CCL-misidentification via the conception of a NGS data-optimized CCL-identification method. The Chapter serves as introduction and contextualization to the following Chapter 4 and is based on the publication of Otto et al., 2017 [18]. The Chapter first contextualizes the CCL-identification problem, explains how CCLs are assigned a position in a space and how CCLs are identified based on their respectively assigned positions. Thereafter, the approach is benchmarked, the advantages and disadvantages discussed and the Chapter finalized by a conclusion that evaluates the efficiency of the approach.

3.1 Introduction

CCLs are crucial to modern Life-sciences and of particular importance for the domain of Oncology since CCLs further experiments on neoplastic cell-cultures, help to investigate the cancer etiology and aid in the validation of driver mutation candidates [118, 119]. Additionally, usage of CCLs avoids ethical and legal issues when compared to patient-based studies [120].

CCLs are, however, susceptible to misidentification i.e. the incorrect assumption that CCL A is being analyzed when in truth CCL B was sequenced [118, 10, 121, 122, 123]. An example of a well-known case of misidentification, that negatively affected a wide range of researchers, was the confusion of the widely used *MDA-MB-435* mammary CCL with the *M14* melanoma CCL which caused massive losses of research time and resources [124]. According to studies, 5-10% of all CCLs are misidentified [125, 8]. Accordingly, many journals currently require authors to ensure identity of the CCLs they employed in experiments upon publication. Therefore, identification methods which are able to detect the misidentification of CCLs are of instrumental value to the community of scientists concerned with Oncology.

CCL are increasingly Next-Generation sequenced what has the unintended side-effect that misidentification of CCLs is furthered by the fact that the established laboratory-centric identification methods can not be applied. Computer-based identification methods for homogeneously sequenced CCL exist, but are not suited for real-world scenarios where highly heterogeneous CCLs have to be identified. We therefore present the novel *in-silico* identification method Uniquorn for Whole-Exome sequenced CCLs.

3.1.1 CCL Identification - Overview

Traditionally, CCL identification is carried out *in-vitro* in a laboratory using specific assays such as STR genotyping [126], SNP panel identification assay (SPIA) [10], MinION [11] or Multiplex Cell Authentication (MCA) [9]. These assays are costly to perform, time consuming and require physical availability of CCLs. Additionally, a reference database that defines what genomic features are characteristic and sufficient for identification of a given CCL has to be available [18]. Such databases are, however, only generated by major institutions such as the American Type Culture Collection (ATCC) and do not provide information on custom-made CCLs of which there can be many in individual labs [9]. Moreover, database information can be misleading if a given CCL culture loses genomic entities that were defined as characteristic by the database due to natural evolution or drug-exposition pressure, thereby motivating the development of new identification methods.

Given the nowadays common scenario of having to identify sequenced CCLs based on their computer-based NGS data, STR and SPIA identification is not applicable due to absence of the required information in the NGS data [118]. The mRNA-array focused SPIA method requires highly reliable ploidy-calls of SNPs at specific genomic loci what, given the heterogeneous nature of NGS data and cancer in general, cannot be assumed. STR is even less suited to identify CCLs based on their NGS data since the calling of the required repeat-counts at specific genomic loci is both technically challenging and repeat-count information absent in the majority of NGS data sets. Even if the required information was available, the effectiveness of STR and SPIA on lab- and project-specific NGS data sets were unclear. Both methods were evaluated only with homogeneous NGS profiles, i.e., references and query samples were sequenced using the same technologies, algorithms, and filtering methods; on top, these procedures require that the ploidy of the reference samples R matches the ploidy of the query sample q. Such a scenario of homogeneous, easily comparable NGS data sets is quite different from that typically found today, where different labs use different technologies, leading to heterogeneous NGS profiles. Hudson et al., for instance compared the small missense variant calls accompanying identical CCLs (as defined by the creators of the reference libraries) between Cancer Cell Line Encyclopedia (CCLE) and catalogue of somatic mutations in cancer Cancer Cell Line Project (COSMIC CLP) and found them coinciding at only 43% [127]. A prominent case highlighting the extend of data-heterogeneity is the *Ishikawa-Heraklion-02er* CCL which was DNA-genotyped by the Broad institute, finding 213 missense mutations, and the Sanger institute, which reported 52 pair-wisely different missense mutations [127]. Causes for the data heterogeneity between large-scale sequencing projects are complex and include technical and design aspects. For example, sequencing of sub-clonal and aneuploid CCL cultures may cause heterogeneous sequencing results [36]. Furthermore, studies differ in their aims and priorities, leading to different choices of algorithmic parameters and workflow designs which in turn can cause differing genotyping results even for the same CCLs [128].

Another aspect that exacerbates the identification of CCLs based on their NGS data is the label and header of a NGS data-containing Variant Calling Format (VCF) file usually indicates which CCL was sequenced to generate the data. However, no nomenclature system that could help avoid idiosyncratic and misleading CCL-names has been universally adopted thus far, leading to highly bewildering naming ambiguities such as *TT* (derived from thyroidal tissue) and *T.T* (derived from esophageal tissue), which are different CCLs with almost identical names[129]. Another example that underlines that CCL names cannot be reliably utilized to infer their relationship are the *NCI/ADR-RES* derived from the *OVCAR-8*; two CCLs with a common origin but significantly different names, obscuring their close relationship [118, 123, 130].

The pressing need for a *in-silico* identification method is highlighted by the fact that already today, most experiments on CCLs involve extensive sequencing [11]. Computer-based CCL *in-silico* identification approaches are an increasingly attractive complement to laboratory-based identification methods [10]. In the *in-silico* scenario, only the NGS-genotyped information of the to-be-identified CCL (termed query) and CCLs of a reference-collection (termed reference library) are required for identification. Identification occurs by matching of small variants contained in the NGS data to identify the reference CCL that shows a significant amount of matching variants i.e. high genomic similarity.

The *in-silico* approach has several advantages from a Bioinformatics vantage point: sequence features of the CCL in the reference library can be obtained once and distributed electronically (no physical access required). Additionally, sequence features of the query CCL are often by-products of the original experimentation (no additional cost). The comparison of the features can be performed rapidly and without additional experimental efforts. Figure 3.1 compares the *in-silico* with the *in-vitro* approach which quantifies the distance between a query and reference CCLs to determine whether the query's distance is significantly small to a reference CCL.

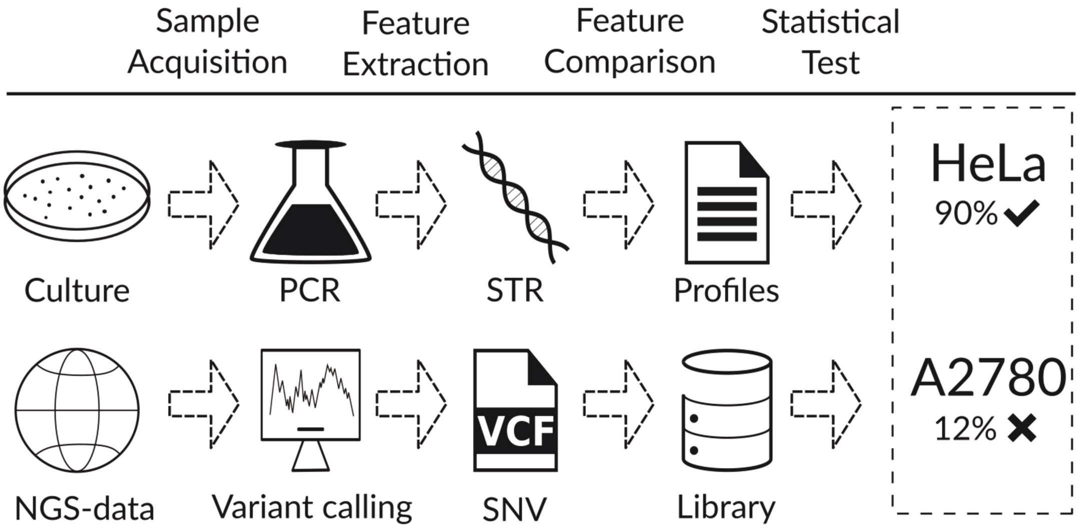

Figure 3.1: Novelty of in-silico identification concept. The gold-standard STR method (top) compares tandem counts at specific genomic loci. STR-counts are generally unavailable in NGS-data and therefore, a CCL whose NGS data is available has to be additionally STR-genotyped which requires the physical availability of the to-be-identified CCL sample to conduct a PCR. Even *in-silico* identification methods that can utilize NGS-derived SNPs are dependent on the genotyping of the loci that harbor the SNPs. SNP-calls of specific loci however, may not be available due to panel sequencing of the to-be-identified CCL or are incomparable due to utilization of divergent sequencing platforms and filtering of SNP during driver-mutation identification. The Uniquorn *in-silico* workflow (bottom) requires neither physical availability nor genotyping of specific loci but in contrast works with every NGS-technology that genotypes small variants. Uniquorn does require sets of reference CCLs, called reference libraries, to match the variants of the to-be-identified CCL and the reference CCLs. After calculating the variant overlap, a statistical test determines whether a variant overlap is sufficiently unlikely to occur by chance in which case the unknown CCL is predicted to be identical to the reference CCL i.e. is identified.

The feasibility of CCL *in-silico* identification based on their NGS data has been demonstrated by multiple publications, however, with the severe limitation that all CCLs NGS data used therein was generated homogeneously by one laboratory with one sequencing technology and identical variant-calling software pipeline [10, 11, 9]. Real-world scenarios, however, show a strong data heterogeneity where *in-silico* identification has to take into account different sequencing technologies, software, natural genomic evolution and mutational-pressure of anti-cancer drugs which can alter a genome dramatically and induce subclonality.

Overview of the Uniquorn WES method

We present Uniquorn WES, a novel *in-silico* CCL identification method for the reliable and fast identification of Whole-Exome sequenced CCLs. Uniquorn WES requires no additional *in-vitro* experiments and solely relies on small variant calls for identification. Uniquorn WES is designed to compare small variant fingerprints derived from a wide range of WES technologies, with differing quality, depth, and sequencing-scope what renders Uniquorn WES useful for large and internet-exchange-based research projects.

Uniquorn WES was developed as complement to the established *in-vitro* methods since it can identify CCLs when neither STR nor SPIA can be applied. Uniquorn supports WES technology and turns the small variants pairwise identity of the query sample to any sample from a reference library R, taking into account the prevalence of each variant in the library and a statistical assessment of the observed number of common variants. We evaluated our algorithm on three high-profile CCL datasets with altogether 1988 reference samples, namely COSMIC CLP (1024), CCLE (904) and National Cancer Institute 60 (NCI-60) CellMiner (60). WES profiles between these libraries are highly heterogeneous, because different laboratories created the data using different WES technologies and software, what even includes covering partly different genomic regions [127]. SNP-based identification using the available data is not generally possible, as in two out of these three sets all SNPs were filtered to facilitate identification of driver mutations and ploidy calls for same SNPs diverge between the repositories for same CCLs. Furthermore, neither of these data sets contains information on STRs. In such a rather difficult setting, Uniquorn WES achieves a sensitivity of 97% at a specificity of 99%. We also show that several CCLs found to be identical, were differently labeled by different CCL repositories. Finally, we confirm a very low probability of random false positive hits by comparing all reference libraries' CCLs with 1024 genomes of the 1000 genomes projects [131, 132].

3.1.2 Identification via distance-quantification based on NGS data

The concept of distance-quantification between CCLs was introduced by Demichelis et al. in 2008 who demonstrated that the matching of SNPs can quantify a distance between CCLs as part of the SNP panel identification assay (SPIA) algorithm [10]. Key questions of distance-based identification is A) whether positions indicate distances and B) at what threshold a distance is sufficiently small to predict that two entities are identical. Demichelis et al. applied a binomial test based on shared SNPs and set the identification p-value threshold of the binomial test to 0.05 which resulted in a advantageous classification performance. Up to this point however, no method has been published that extends the distance-based identification to heterogeneously genotyped i.e. sequenced CCLs.

The identification-by-Distance-quantification concept of Uniquorn builds on the Demichelis et al. approach and assigns every CCL a position in a space with a distance-quantifying metric based on the CCLs' NGS data, see Figure 3.2. The main difference to Demichelis are significant differences with respect to the analyzed genomic entity (SNP versus SNVs without population prevalence) and the omission of any assumption regarding the analyzed NGS-data other than the limitation to WES-technology.

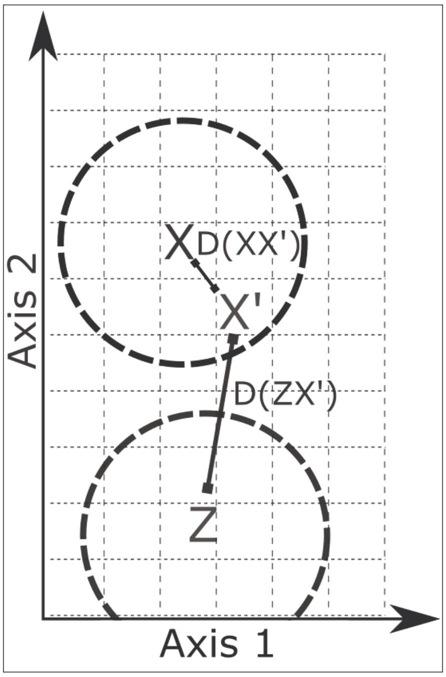

Figure 3.2: Identification by distance-quantification. The identification-by-distance-quantification concept is illustrated at hand of two reference CCLs X and Z and a query CCL X' which is identical to X. It is shown that all CCLs will be assigned a position in a metric space based on their small weighted variants. The reference CCL is surrounded by a spherical area within which a distance $D(.)$ is sufficiently small to assume an identity of two CCLs as is the case for X and X'. The size of the sphere is defined by the test on sufficient similarity: A position which results in a p-value smaller than 0.05 lies within the sphere, everything else outside.

The purpose of the position attribution in a metric space, as explained in the introduction Section 2.2, is to quantify the distance between the data-points. The underlying rational is that an unknown query CCL can be identified if the query's distance to a known reference CCLs is sufficiently small. The small distance is, by assumption, observed because the same CCL were sequenced (see Exclusivity assumption), see Figure 3.3.

Importantly, the Uniquorn approach assumes a comparatively homogeneous data landscape with respect to the utilized technology since it is limited to the WES technology. The homogeneity assumption is critical because only iff the conditions of a distance metric are fulfilled, can a meaningful distance be derived from the positions of the CCLs. The degree of Data-Heterogeneity that the Uniquorn WES approach is subjected to is illustrated in Table 3.1.

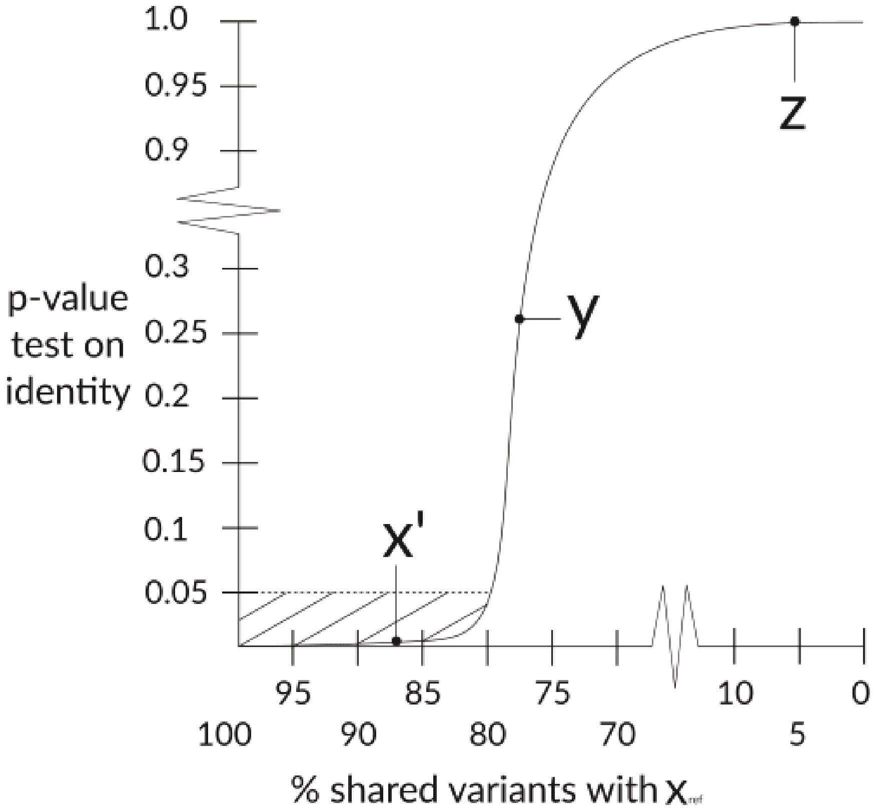

Figure 3.3: Decision on CCL similarity via distance-quantification. Determination of CCL identity requires the definition of sufficient similarity of two CCL with respect to their sets of small variants. Commonly, a p-value is calculated that quantifies the risk of committing a mistake when assuming that two CCLs are identical. The y-axis shows this type-I risk while the x-axis shows the amount of matching small variants between a known reference CCL and query CCL in percent. The figure illustrates that the p-value between the unrelated reference CCL x and query CCLs y and z is not significant ($\alpha = 0.95$) although y shares 80% of variants with x. The (ground-truth identical) query CCL x' however, is assigned a significant p-value smaller than 0.05. The similarity between two CCLs based on their small variants can thus be translated into a p-value. The resulting p-value consequently allows to decide on sufficient similarity in the variant-space. Note that the plot is an example and the underlying p-value function not necessarily representative for real-world scenarios.

Reference Library	Total amount of variants	Cancer Cell Lines	∅ Variants Per CCL	Variant calling software	SNP-MAF Filter*
COSMIC CLP	760E5	1024	7,4E5	Caveman[133]	> 0.0
CCLE	140E5	904	1,5E5	Pindel [134] MuTect [135]	≥ 0.05
Cellminer	0,68E5	60	0,01E5	GATK[136]	> 1.0 (none)

Table 3.1: Degree of data heterogeneity for Uniquorn WES. The three by amount of sequenced samples largest CCL sequencing studies are shown. The studies diverge significantly in multiple key aspect, most importantly for identification however, with respect to the absolute and the average number of variants, which differ by orders of magnitude, choice of algorithms, technologies and mount of sequenced CCLs. STRs, utilized by the gold-standard technologies are not covered and SNPs - required for SNP-based identification - are incoherently fully included, filtered according to Minor Allele Frequency (MAF) threshold or fully excluded. Filter*: Exclusion of SNPs with a MAF of greater than x. Nonetheless, the WES or its derivative, the hybrid-capture sequencing technology was utilized by all studies.

3.2 Method

Let q be an unknown to-be-identified query CCL whose NGS dataset is available in a VCF-file. Uniquorn WES then determines the distance $D(q, r)$ between q and all reference data sets r in reference library R to decide whether the distance is sufficiently small to assume an identity of q and r. The distance between q and r and is designed to range from 0 to positive infinity:

$$D : q \times r \to d \in [0, \infty), d \in \mathbb{R}_{+,0} \tag{3.2.1}$$

Uniquorn quantifies this distance based on the amount of matching variants between q and r where the sets of variants are modeled as genomic position-ordered vector of substitutions and InDels. Each variant is defined by its start position and its length.

3.2.1 Pre-processing

Uniquorn identifies samples only based on their informative variants, the informativeness being modeled as high weight (higher informativeness) or low weight (less informativeness). To this end, each variant v found in any sample of the given reference library R is weighted during pre-processing according to the inverse of its frequency f_v in R using:

$$w(v) = 2^{-(f_v - 1)} \tag{3.2.2}$$

Variants with weight lower than a threshold are discarded during pre-processing, the default threshold being 0.5. Other thresholds can be chosen, depending on the desired trade-off between sensitivity and false positive rate (see Table 3.2 and discussion section 3.4). Variant weights are library-dependent i.e. the same variants will receive different weights in different libraries to reflect the inherent divergence of sequencing technologies and algorithms.

3.2.2 Variant matching

Next, the absolute amount of overlapping variants $O(q,r)$ between q and r is calculated based on the coordinates of the variants $var_q \in q$ and $var_r \in r$:

$$O(q,r) := \sum_{i=1}^{|r|} var_i^r \in var_q \tag{3.2.3}$$

Note that the overlap depends on $|r|$ and not $|q|$ i.e. the overlap of r's variants with q is quantified and not q's overlap with r. This distinction will become important in the next confidence score calculation step 3.2.3 that quantifies the likelihood to observe a specific $O(q,r)$ given all $O(q, r \in R)$.

3.2.3 Confidence score calculation

Uniquorn WES calculates a confidence score CS which empirically quantifies the certainty that q and a r are identical i.e. their distance sufficiently small given all overlaps between q and $\forall r$. The score is based on the probability $\hat{P_r^q}$ to observe a given amount of overlapping variants $O(q,r)$ by chance conditioned on the expected rate of matching variants p between unrelated CCLs. The true value of p is generally not available because a ground-truth knowledge of p would require meta-information regarding the way how the profiles of q and r were obtained which is commonly not available. Therefore, we developed a simple yet effective empirical heuristic to approximate p based on all $O(q,r)$ where $|R|$ represents the amount of reference CCLs in R:

$$\hat{p} := \frac{1}{|R|} \cdot \sum_{i=1}^{|R|} \frac{(q, r_i)}{|r_i|} \tag{3.2.4}$$

By modeling assumption, the maximal posterior likelihood of $\hat{p}$ equates the empirical mean of all $O(q,r)$ in R. $\hat{p}$ is subsequently utilized as parameter of a binomial function that quantifies $\hat{P_r^q}$. Let T be the overall amount of variants in R, N be the number of variants in r, n the subset of these also found in q, and $k = N - n$ the amount of variants in r not found in q. Then, the approximated lower-tail probability $\hat{P_k}$ to miss exactly k variants from r in q given p due to chance is:

$$\hat{P}_k := \binom{N}{k} q_r^k p_r^{N-k} = \frac{N!}{k!(N-k)!}(1-p_r)^k p_r^{N-k} \tag{3.2.5}$$

Following Mi et al., we next compute a p-value by summing up the probabilities to miss 0 up to k variants [137].

$$\hat{P_r^q} = \sum_{N-n=k}^{N} \hat{P_k} \tag{3.2.6}$$

$\hat{P_r^q}$ is corrected for multiple testing by the Benjamini-Hochberg method to obtain q-values. The corrected $\hat{P_r^q}$ is utilized as quantification of the error type one risk when rejecting the null hypothesis H_0 [72]. Uniquorn identifies CCLs by testing on the rejection of the null hypothesis H_0 that states q and r share a given amount of variants due to chance. The alternative hypothesis H_1 formulates that q and r show a similarity that is not due to chance. We then transform the $\hat{P_r^q}$ q-value to obtain distance-suitable confidence score values $D(q, r)$ with a domain between 0 and ∞ and decide whether to reject H_0 given a defined threshold:

$$D(q, r) = -1 \cdot log_e \hat{P_r^q} \tag{3.2.7}$$

If the distance i.e. confidence score threshold and the second threshold regarding the minimal amount of matching variants (default five) are met, the variant profile of a reference CCLs r is predicted to stem from the same CCL as the profile of q. Note that this implies that multiple CCLs from the same reference library might be predicted to be identical to q. We find this strategy to have advantages over the option to simply return the best matching reference sample.

The score is library-specific i.e. the score obtained from the comparison of query q with a sample r from reference library R assesses the probability that q is identical to r independently of all other libraries. Importantly, the likelihood to observe a given overlap $O(q, r)$ as shown in equation 3.2.3 is approximated via an empirical sampling over all $O(q, r)$ in R in equation 3.2.4 and thus the confidence scores for identity not statistically independent. The default threshold for $D(q, r)$ was chosen such that it balances sensitivity and specificity in during the benchmark, see Receiver on Operator Characteristic (ROC)-curve in Supplementary Figure 7.1.

3.2.4 Evaluation

We benchmarked Uniquorn WES using 1988 CCLs from the three datasets described above (see Table 3.2) as query sample against each of the three reference libraries; thus, we performed $1988 \cdot 1988$

~4E6 comparisons in total. A True Positive (TP) identification was counted when Uniquorn predicted that a query was identical to a reference CCL in accordance with the gold-standard (see gold-standard creation); analogously for True Negative (TN) predictions. A False Positive (FP) prediction was counted when Uniquorn predicted query and reference CCL to be identical but not the gold-standard. False Negative (FN) predictions were cases were query and reference CCL were assessed as not being identical by our algorithm but identified as such in the gold-standard.

Note that the maximal number of TP predictions per query in this evaluation scheme depends on whether this CCL was present in only one or in more than one datasets (many such cases exist; see Figure 3.4). If a CCL existed only in a single reference library, only one TP prediction can occur. If it is part of two libraries or has related identified CCLs within the same library, four TP predictions are possible, since each will be used as query and should identify both itself and the related sample; for CCLs in all three libraries, maximally nine TP predictions can be found. Using our gold-standard, a maximum of 3573 TP predictions was possible.

3.2.4.1 Gold-Standard creation

The gold-standard defines which pairs of CCLs are considered identical within our evaluation. To create a gold-standard we first defined all CCLs with the same regularized name as identical. CCL names were regularized by removing any non-alpha-decimal and capitalization of all remaining characters. In a second step, we manually confirmed or rejected the identity of all CCLs whose names only differed by a small prefix or suffix, such as *MDA-MB-435* and *MDA-MB-435s*. In a third step, we screened the literature for cases were CCLs with same regularized name were reported as being different, e.g. *TT* and *T.T*, and adapted the gold-standard accordingly for these cases. Note that pairs of identical CCLs may be part of the same or of different reference libraries (See Figure 3.4).

After the evaluation, we furthermore checked all FP predictions to determine whether these are indeed FPs predictions or errors in the gold-standard (see Discussion); one such example is the pair *SNB19* and *U-251*, which have completely different names but denote the biologically identical CCLs[10]. The entire gold-standard is available in supplementary material File 7.1.

3.2.5 Utilized Datasets

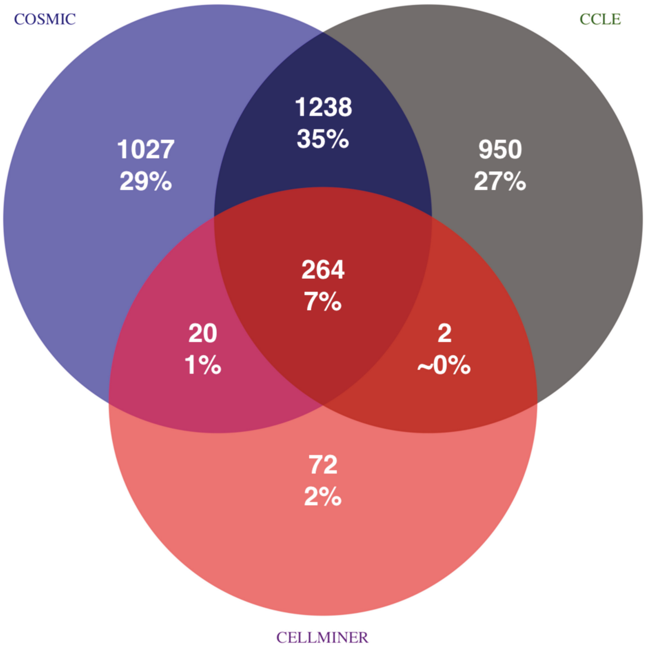

Figure 3.4: Source of true positive CCL identifications based on the gold-standard. Total amount, percentage, and source of all 3573 TPs identifications for each of the 1988 CCL samples are shown. For instance, 1238 TPs are identified because copies of the same or highly similar CCL are contained in COSMIC CLP and CCLE. Positive identification within a single circle are due to relatedness of CCLs within the same library and self-identifications. 43% of all possible TP cross-identifications are due to CCL copies in different reference libraries. Percentages do not sum up to 100% due to rounding errors.

3.2.5.1 Reference Libraries

Uniquorn WES compares NGS data of a given query sample q with that of samples r from a given CCL library R. Currently, three large libraries are integrated into the package: (1) COSMIC CLP, obtained January 13th 2016 from `http://sftp-cancer.sanger.ac.uk` (2) CCLE, obtained January 13th 2016 from `http://www.broadinstitute.org/ccle` and (3) CellMiner, obtained January 13th 2016 from `http://discover.nci.nih.gov/cellminer`. All data sets are based on the same reference genome HG19/ GrCH37. SNV profiles and CCL-names were directly parsed from the files provided. Note that the Uniquorn package also features an Application Programming Interface (API) for adding novel, possibly in-house-created, reference libraries.

Table 3.1 shows most important characteristics of the three libraries. COSMIC CLP is the largest dataset with 1024 WES genotyped CCLs from 30 tissues. CCLE contains 904 hybrid-capture genotyped CCLs from more than 36 tissues. The CellMiner project comprises WES genotype data of the NCI-60 panel from nine tissues.

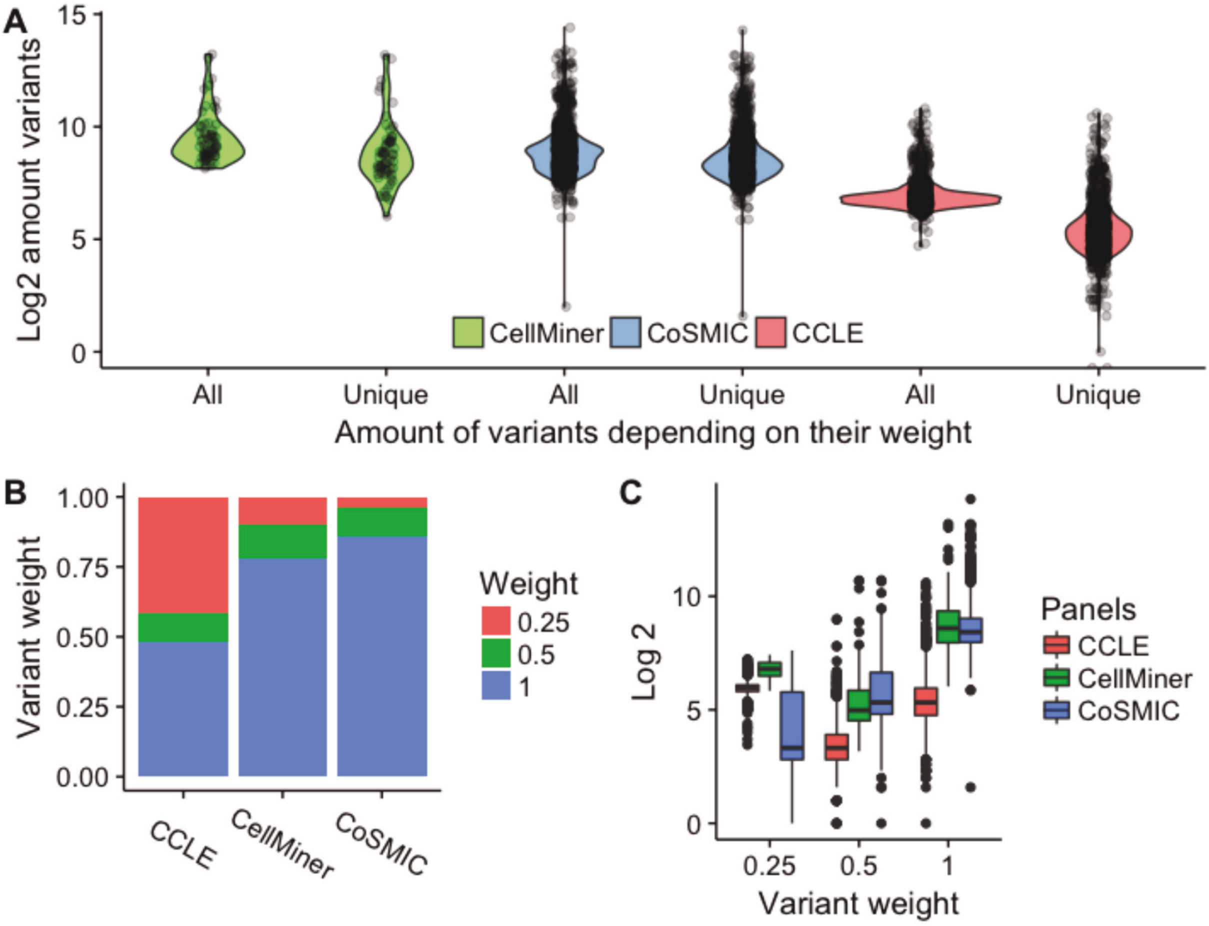

Figure 3.5: Detailed analysis of the variant class heterogeneity in the three largest CCL studies. The distribution of CCLs variant frequencies and weights across libraries. A: Number of *rare* variants in CCLs according to Uniquorn's weighting scheme. 'All' shows the log-amount of variants per CCLswithout any filtering (weight 0.0) and *Unique* indicates the amount of variants that remain after all variants were filtered that were present in more than a single CCLs (weight 1.0). Differences between software, technologies and filters (non-exhaustive) i.e. heterogeneous data-processing lead to different amounts of filtered, non-unique mutations as shown by the significantly different reduction of variants between the CellMiner (medium), COSMIC CLP (low) and CCLE panel (strong), see Table 3.1 for the sources of heterogeneity. It is shown, that all panels possess unique, i.e. *rare* variants on which the Uniquorn identification method is based. B: Distribution of weights per library. At least 50% of variants are high-weight (rare) variants. CCLE shows significantly less unique variants than COSMIC CLP and CellMiner, which explains the strong difference between raw and filtered variants in Figure A. C: Number of variants per reference sample for different weight thresholds in the different reference libraries. CCLs from COSMIC CLP show a high amount of unique variants on average, especially when compared to those from CCLE.

3.3 Results

The Uniquorn WES method identifies a Whole-exome sequenced query CCLs by comparing its variant profile to that of all CCLs in a given set of Whole-exome sequenced reference libraries, see Figure 3.6. To this end, each variant in a reference library is weighted according to its inverse frequency. Only rare variants are used further. To assess the impact of different thresholds for this weight, we studied the distribution of variant counts in each of the three libraries (Figure 3.5A). As can be seen in Figure 3.5B, more than 50% of variants are unique within their library (weight 2 or higher), which means that even a very stringent threshold of 1.0 would filter out less than half of all variants. In Figure 3.5C, we show the distribution of the number of variants per CCL using different weight thresholds. When using only unique variants, CCL from CCLE library have on average 153 variants in their profile (COSMIC CLP: 744; CellMiner: 1139).

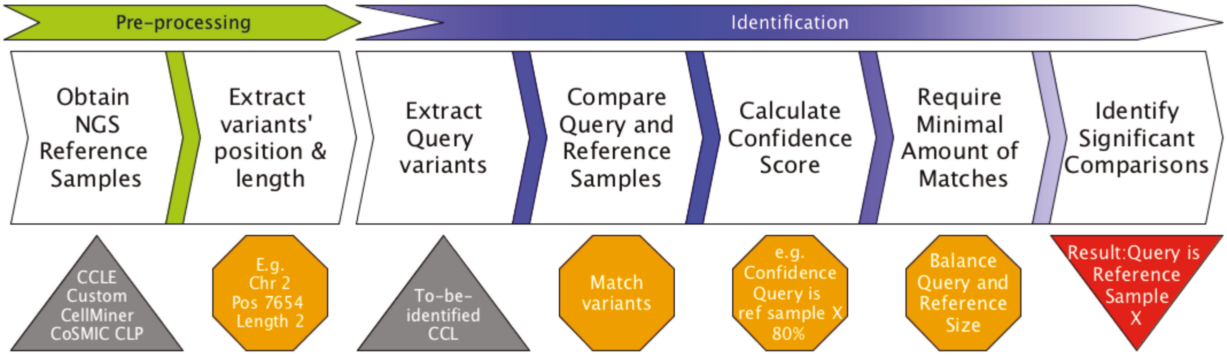
Pre-processing
Identification
Obtain NGS Reference Samples
Extract variants' position & length
Extract Query variants
Compare Query and Reference Samples
Calculate Confidence Score
Require Minimal Amount of Matches
Identify Significant Comparisons
CCLE Custom CellMiner CoSMIC CLP
E.g. Chr 2 Pos 7654 Length 2
To-be-identified CCL
Match variants
e.g. Confidence Query is ref sample X 80%
Balance Query and Reference Size
Result:Query is Reference Sample X

Figure 3.6: Uniquorn WES workflow. CCLs from a reference library are compared to a given query sample q based on their set of small variants (variant profile) obtain by Whole-Exome DNA sequencing. Variants are weighted according to their prevalence within the library (e.g. CCLE) and frequent variants are subsequently excluded. Thereafter, Uniquorn computes a confidence score quantifying the likelihood for each reference sample r being identical to q. The significance of differences in the number of variants in q versus r (for instance due to different sequencing scopes) are assessed in a second test.

Distance-quantification

In related publications, distances are generally modeled as n-dimensional space over the domain of real-valued numbers where n is the amount of features e.g. a set SNVs. The distances can thereafter stratify and cluster samples based on their genotype what yields experimentally testable hypotheses to explain the phenotype to genotype relationship. Uniquorn follows that approach but limits the compared entities to small variants such as SNVs and InDels. All small variants are sorted according to their genomic locus, starting with the autosomal chromosome 1 up to the autosomal chromosomes X and Y, if present. The sorted variants are then formatted as vector where very position in the vector corresponds to the sorted genomic locus.

The matching of small variant can intuitively be modeled as a logical *and* ($\wedge$) operation between the vectors of compared CCLs. The comparison is not symmetric because once CCL is defined as reference and all variants contained in the querry's vector that do not correspond to the genomic loci of the reference are discarded. In contrast, should a variant be contained in the reference but not the query, a zero entry is made in the query's vector while a matching variant is assigned a one. A Minkowski-distance is then quantified between these binary vectors via a logical $\wedge$ operation [130]. In summary, a metric space is created that has the same dimension as the amount of variant that the reference CCL has because all small variants serve as features for the position. A position is obtained by setting all variant that are present in either query or reference to one.

As illustration which is not applied: One could sum over the query vector and divide by the dimensionality of the reference to obtain the percentage of matching variants to obtain a naive distance-quantification, between the CCLs. The main problem in that hypothetical example is that the linear distance obtained from a pair-wise matching precludes an effective testing for sufficient similarity of the CCLs since the dimensionality is highly volatile and depends on the amount of variants in the reference. Therefore, a subsequent statistical tests overcomes that factors on the amount of variance in the reference is require for an efficient identification i.e. test on sufficient similarity of the CCLs.

3.3.1 Cross-validation benchmark

We benchmarked the accuracy of Uniquorn WES using three high impact and diverse CCL libraries, namely COSMIC CLP, CCLE and CellMiner, which as ensemble entailed 1988 CCLs. We manually identified duplicates in this set and tested how reliably Uniquorn WES would detect them. To this end, each of the 1988 CCLs samples was once utilized as query-sample and all three libraries as references. Since Uniquorn compares a single query-sample to all reference-samples, $1988 \cdot 1988{\sim}4 \cdot 10^6$ comparisons occurred during the cross-validation benchmark, underlining the size of the benchmark. Uniquorn predicted for each of the query-reference-pairs whether they were derived from the same CCL.

As only 3573 of these $\sim 4 \cdot 10^6$ pairs are duplicates according to our gold-standard, the positive predictive value (PPV) is a particularly important evaluation measure. The benchmark results shown in Table 3.2 and Figure 3.7 indicate a very high specificity (at least 99%) across a range of weight thresholds, which can be explained by the extremely large number of true negatives. The more important metric is sensitivity, which is also very high for thresholds 0.5 and 0.25, correctly identifying 3474 and 3461 of the 3573 identical or related CCLs, respectively. Limiting the comparison to unique variants (weight threshold 1.0) yields the best PPV and lowest False Positive Rate (FPR), but lower weights of 0.5 and 0.25 result in higher sensitivity. Quantitative regularization slightly reduces identification efficiency, but suppresses as significant fraction of FP predictions, see Figure 3.7 for an analysis of the impact of the weight parameter on the identification performance.

Weight Threshold	1.0	0.5	0.25	0.0
Maximally Possible TPs	3573			
True positives	3027 (3372)	3474 (3521)	3461 (3528)	3111 (3485)
False negatives	546 (201)	99 (52)	112 (45)	462 (88)
False positives	22 (18)	37 (94)	59 (155)	4631 (7689)
Sensitivity %	85 (94)	97 (99)	97 (99)	87 (98)
Specificity	99			
F1 %	91 (97)	98 (98)	98 (97)	55 (47)
Positive predictive value	99 (99)	99 (97)	98 (96)	40 (31)

Table 3.2: Uniquorn WES benchmark. A higher threshold enforces utilization of more specific variants but reduces the amount of considered variants. Depending on the threshold (0.0, 0.25, 0.5, 1.0) between 3027 and 3474 of the 3573 true relationships between CCLs are successfully recovered. Numbers in brackets show results when the to-be-expected amount of matching variants is set manually to ten variants; numbers without brackets show statistically estimated background-noise strength (regularized, see methods Section 3.2).

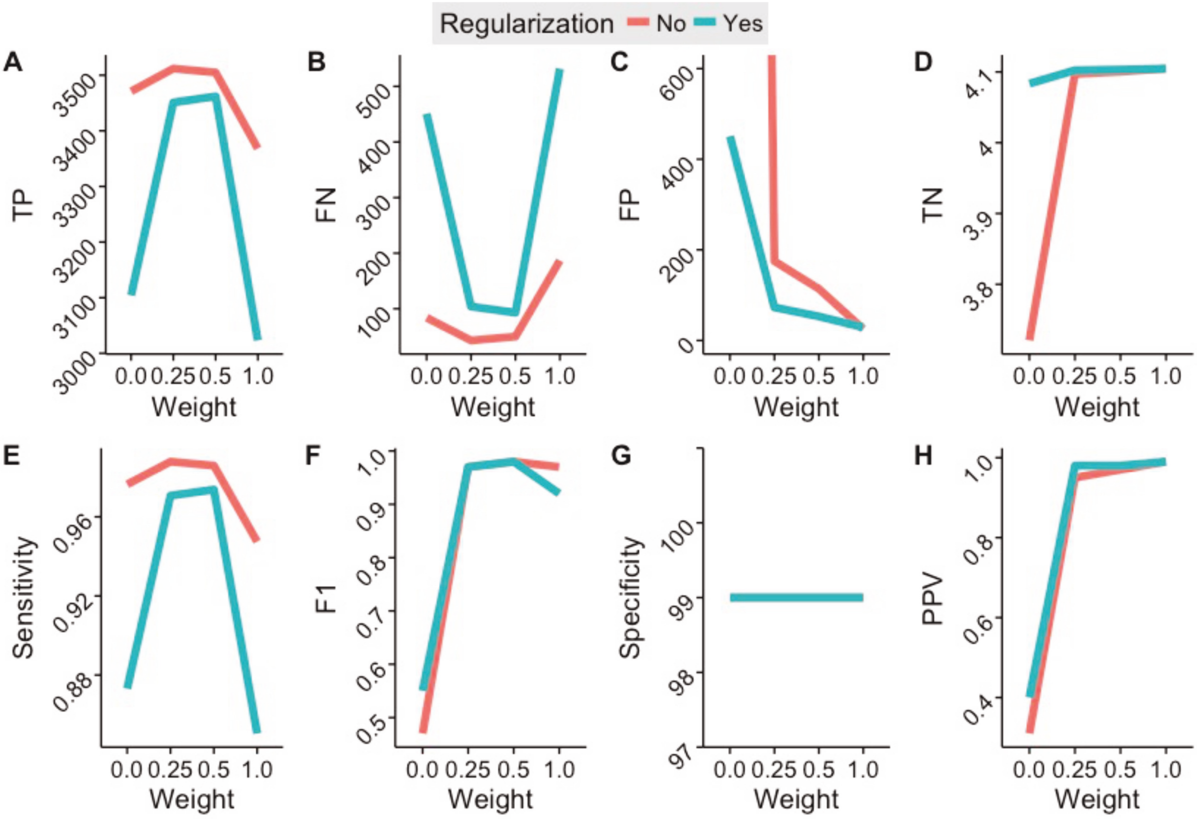

Figure 3.7: Results of the cross-identification benchmark depending on regularization and variant inclusion weight. (A) Amounts of TP, (B) FN, (C) FP and (D) TN predictions. (E) Sensitivity. (F) F1-Score (harmonic mean of specificity and sensitivity). (G) Specificity. (H) PPV. Best specificity and sensitivity values are achieved using a weight threshold of 0.5. A threshold of 1.0 achieves the least false positives, most true negatives, and the highest positive.

3.3.2 Out-group benchmark

The previous evaluation measured the performance of Uniquorn WES when searching a CCL of a reference library within the set of reference libraries. We also tested how the method performs when it has to deal with profiles that are not derived from CCLs. Specifically, we used 1092 WGS profiles from the 1000 genomes dataset as query samples and tested whether Uniquorn would assign them to a reference CCLs, any such assignment certainly would be an error [131]. Note that these comparisons work on very heterogeneous sequencing technologies, namely WGS-sequenced profiles (1000 genomes) with much smaller hybrid and WES profiles (reference libraries). This implies large differences in terms of common SNPs (contained in 1000 genomes profiles, filtered in the references) and in the sheer number of variations (on average, a 1000 genomes profile consists of ~5E7 variations per sample compared to ~5E2 variations in the reference profiles). Using a weight threshold of 1.0 and regularization to cater for this difference, Uniquorn did not produce a single FP prediction.

These comparisons highlight the importance of our regularization step; omitting this filter, the comparison would produce 167 FP predictions for the ~2E6 comparisons. Based on this and the previous experiments, Uniquorn's default confidence-score threshold is set to 10 ($\sim -log_2(0.001)$). By default, the regularization filter automatically measures the strength of the background-noise and adjusts the required amount of matching mutations accordingly. However, users can set both thresholds manually to adapt to different reference libraries or to change the balance between false prediction rates and sensitivity, see Figure 7.1 for a ROC-curve analysis.

3.3.3 Comparison to established methods

Uniquorn WES compares favorably to other methods for the identification of CCLs in terms of the amount of data and experimental work necessary (see Table 3.3). In first place, it is similar to established methods e.g. SPIA and STR-counting in that it is comparison-based. Uniquorn WES, however, is different to the aforementioned methods due to its focus on *in-silico* identification of CCLs based on variant profiles obtained from different NGS high-throughput sequencing technologies. Unlike SNP-based methods, Uniquorn does not depend on common, well characterized and publicly available genomic entities, but instead relies on rare somatic mutations, as SNP-based comparisons have severe drawbacks when applied in cancer research. First, SNPs with a MAF of $\geq 5\%$ are usually frequently filtered from datasets (to focus on driver-mutations, e.g. by CCLE) and thus cannot be assumed to be generally available for a CCL identification. Second, the loci of the most characteristic SNPs often are not genotyped during WES sequencing, and even less often so in panel sequencing. Moreover, neoplastic genomes are frequently subjected to large-scale structural variations, often removing cancer-irrelevant loci, and with polyploid chromosomes whose variant calls cannot be directly compared to diploid references. Uniquorn was designed to robustly deal with such problems.

Identification Method	Physical Sample Required	Experiment Required	Locus Coverage Required	Zygosity-pattern required	Dependent on Reference Genome
STR	X	X	-	-	-
SPIA	X	X	X	X	-
NGS-SNP	-	-	X	X	X
Uniquorn	-	-	-	-	X

Table 3.3: Properties of Uniquorn compared to established methods for CCL identification. SPIA and STR-counting require additional verification experiments to be applied to the physical CCL sample. Identification of CCLs by matching their SNP-zygosities directly from the NGS-data requires that the loci of the characteristic SNPs were sequenced and not filtered. For SPIA and NGS-SNP, zygosity calls have to be comparable (technology, ploidy, algorithms, etc.). Uniquorn WES only requires utilization of the same reference genome for variation calling. Note that CCL samples created with a specific reference genome versions can be converted into another version, e.g. by a lift-over software, thereby decreasing the gravity of this limitation.

We also compared identification results of Uniquorn WES and the SNP-based method by Demichelis et al. [10] quantitatively. 130 of the 155 CCLs used by Demichelis and colleagues are present in the Uniquorn WES benchmark set. These 130 CCLs have 265 different representations in our data set because many are present in different CCL reference libraries. Uniquorn WES identified 100% of these 265 CCLs at an inclusion weight of 0.5 (see Supplementary File 7.3). Thus, Uniquorn showed an equal performance compared to the established SNP-based identification methods.

The method was implemented in the freely available R-Bioconductor package *Uniquorn*. The software is freely available and benchmark libraries CCLE and COSMIC CLP can be freely obtained and used as Uniquorn WES reference libraries. The CellMiner Project library is included by default. Custom libraries can be created e.g. for identification of proprietary CCL samples.

3.4 Discussion

3.4.1 Analysis of false positive predictions

Analysis of the 22 FP predictions from Table 3.2 (weight 1.0) revealed that all FP-predictions were caused by a set of only 13 CCLs and have in common that their small variant profiles are very small; they have a mean size of 366 (sd=4E3) variants, while the profile sizes of CCLs that were never resulted in a FP prediction have a mean size of 3768 (sd=8E2) variants (p=0.006). 20 of these 22 FP predictions occurred with a query sample identifying a reference from a library which does not contain the query, which means that they would not occur if a lab can safely exclude a reference library from considerations. The most problematic CCLs regarding FP predictions is the *HCC-2998*, which is contained in the CellMiner and COSMIC CLP libraries. Accordingly, it was used twice as query, what resulted in five FP predictions in total (three FP predictions when used as query and two FP predictions when utilized as reference CCL). When used as query, *HCC-2998* correctly identified itself in CellMiner and COSMIC CLP with a high confidence. However, it was also predicted to be similar to three CCLs from CCLE (*JHUEM-7*, *SNU-81*, *HEC-251*). These false predictions all had very low confidence scores, sharply above the threshold, and can be explained by to the stronger influence of randomly matching variants within small profiles.

Three factors were found to be associated with FN predictions: About 100 of the 546 FN-predictions for weight 1.0 occurred between query-reference pairs that were defined as identical by the gold-standard due to either cross-contamination (e.g. ACCS and T24 [122]) or an origin within the same human being but not the same cancer-tissue (e.g. AU-565 and SKBR-3 [138]). Secondly, FN predictions are enriched in CCLs with small profiles. CCLs that failed at least once to identify a related query have on average 345 (σ=2E2) variants, while CCLs, that always identified their counterparts successfully, have on average 528 (σ=1E3) variants (p-value=1E-8). Thirdly, CCLs that are highly similar to another CCL within the same library generally perform poorly because in those cases the amount of rare variants is insufficient. For instance, *HEL* and its closely related sub-clone *HEL 92.1.7* both failed to identify themselves because they are so similar that none of their variants are unique within the library [139]. This effect can be diminished by appropriate adjustment of the weighting scheme, as can be seen by a FN-reduction of 82% from weight 1.0 to weight 0.5. However, these cases are rare within our evaluation data: As shown in Figure 3.5, unique variants are present in 1986 out of 1988 CCLs (99.9%).

The overall concept of identification by pair-wise distance-quantification in conjunction with a statistical test on sufficiently small distance proved successful in light of the benchmark given the homogeneity assumption. Importantly, an increase in heterogeneity may render the statistical tests biased and thus further research with respect to applicability of the distance concept given greater heterogeneity has to be conducted.

CCL-identification based on generic ’omics data

Every NGS technology that allows calling of small genomic variants can hypothetically be utilized to identify CCLs based on the Uniquorn WES method provided that their amount of features per sample is comparable. We therefore see an extension of the Uniquorn WES approach to RNA-sequencing data as possible without conceptual changes. Extension to panel sequencing requires the re-adjustment and optimization of thresholds to compensate for the relatively low number of variants and may require more statistical balancing such as Monte Carlo sampling. Furthermore, since fewer matching entities may already indicate that two CCLs are similar, the statistical tests for matches occurring just by chance might have to be strengthened. Usage of scRNA technologies is of high interest due to the technology’s importance but requires more adjustments to compensate for higher impact of random events (noise) given the vastly lower coverage in single cells. Less similar NGS technologies, such as methylation, Chromatin ImmunoPrecipitation-sequencing (ChIP-seq), or Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), as well require more profound changes to the Uniquorn method.

3.5 Conclusion

Uniquorn WES is a novel *in-silico* method for helping to avoid confusion of CCLs during lab processing and related *down-stream* analyses. The distance-quantification approach proved fruitful since the limited heterogeneity of the NGS data rendered the distance quantification effective in that the biological distance could be quantified as opposed to the technological distance Uniquorn works across a range of sequencing techniques and, importantly, can be applied regardless of whether SNPs were filtered or not. A favorable sensitivity of up to 97% and specificity of 99% has been achieved when benchmarking various CCL created by diverse WES technology platforms. However, Uniquorn’s limitation to sequencing data generated by the same sequencing format has to be extended to increase the amount of use-case scenarios.

Chapter 4

Distance-based identification of Bulk-RNA and Panel-sequenced Cancer Cell Lines

This Chapter demonstrates how the distance-based identification concept introduced in Chapter 3 has been significantly enhanced via a methodological extension. The purpose of the extension was to render the technology-constraint solution of Chapter 3 generic by supporting the currently most-widely utilized sequencing technologies.

The Chapter first presents difficulties encountered when applying the distance-quantification-based approach from Chapter 3 for the identification of Whole-Exome sequenced CCLs to RNA-Bulk and Panel-sequenced CCLs. We thereafter illustrate the modifications applied to the distance-based identification concept and explain how the modifications compensate for confounding factors. Next, we present a benchmark on highly heterogeneous data to evaluate whether the method is sufficiently generalized. The Chapter concludes with a discussion of the suitability of the distance-based identification concept to address the CCL identification problem on a generic base. The contributions of this Chapter are based on the publication Otto et al., 2019 [19].

4.1 Introduction

A major advantage of the inclusion of RNA and Panel-sequencing is the great increase of possible use-cases of the Uniquorn method because the reference databases can be composed of diverse types of sequencing technologies and formats. Users of the Uniquorn method can therefore compose their reference libraries much more liberally while ascertaining that a new query CCL sequenced by technology, not present in the reference database, will be efficiently identified. The major disadvantage is that technological confounding factors are incurred from comparing diversely created NGS data what severely limits the identification performance of the unmodified Uniquorn method and requires the modification of the Uniquorn 1 approach.

4.1.1 Data-Incompleteness and Data-Heterogeneity

The methodological reason that renders the distance-quantification approach susceptible towards confounding factors was the conditioning of intractable parameters on reference-databases. In order to determine the distance between a reference and a query CCL, a heuristic approximation of the baseline likelihood $\hat{p}$ to observe a match between unrelated CCLs has to be approximated. $\hat{p}$ is approximated by dividing through the amount of variants present in a reference CCL (see equation 3.2.3). One will observe diverging distance-quantifications between a query and a reference depending on the underlying sequencing technology and formats because the diverse amalgamation of different technologies and formats causes the confounding factors of Data-Incompleteness and Data-Heterogeneity to bias the distance quantifications.

Data-Incompleteness is incurred from the comparison of highly diversely generated NGS formats, when the same biological entity was sequenced but at at least partially different genomic loci. Differently formulated, Data-Incompleteness occurs when the intersect of the set of reference and query loci is different from its union, for instance when a Whole-Genome-sequenced CCL is compared to a Panel-sequenced CCL; the sequenced loci will merely show a negligible overlap. The ramification of the Data-Incompleteness is that it distorts the statistical test on sufficient similarity between two CCLs what causes false positive and negative classifications, see Figure 4.1.

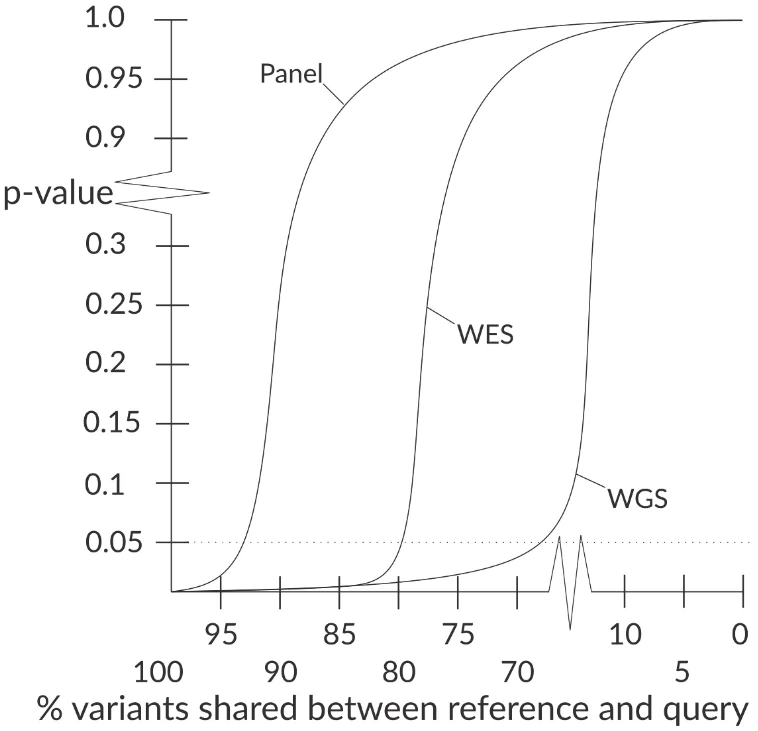

Figure 4.1: CCL test-on-identity p-value skewing caused by Data-Incompleteness and Data-Heterogeneity. The figure illustrates that the p-value of the test on identity is susceptible for confounding factors introduced by an uneven variant count between the query and the reference CCL. Given the scenario that the reference CCL was WES sequenced, about $_{\sim}80\%$ of the variants contained in the reference CCL would have to be matched by the query to result in a p-value of 0.05. The p-value of lower than 0.05 would then lead to a rejection of H_0 and the acceptance of H_1 that assumes both CCL to be of identical origin. Given a WGS or Panel sequenced reference CCL, the percentage of variants that has to match for a p-value of less than 0.05 varies compared to WES what introduces a bias due to incomplete knowledge about the correct value for p as approximated by $\hat{p}$. Data-Incompleteness in the case of Uniquorn thus causes the test-statistics of the Uniquorn WES method from Chapter 3 to fail.

Data-Incompleteness therefore introduces a bias with respect to the distance-quantification by violating the Exclusivity assumption explained in the introduction table 2.1. The consequence is that the identification effectiveness of the method is significantly reduced because the identification by distance-quantification approach assumes that the distance between CCL reflects their biological similarity i.e. remains invariant to the sequencing technology. The naive distance metric approach, however, quantifies the distance of the CCLs' technologies and not the CCLs' biological distance.

Data-Incompleteness, furthermore violates the conditions that a metric has to fulfill, such as the symmetry of the distance-quantification (equation 2.2.8) and the triangle inequality (equation 2.2.9), See Figure 4.2. The triangle inequality fails because the underlying metric space is not sufficiently 'flat' for the identification by distance-quantification approach in the sense that positions in the metric space are not anymore indicative of biological distances, see Figure 4.2.

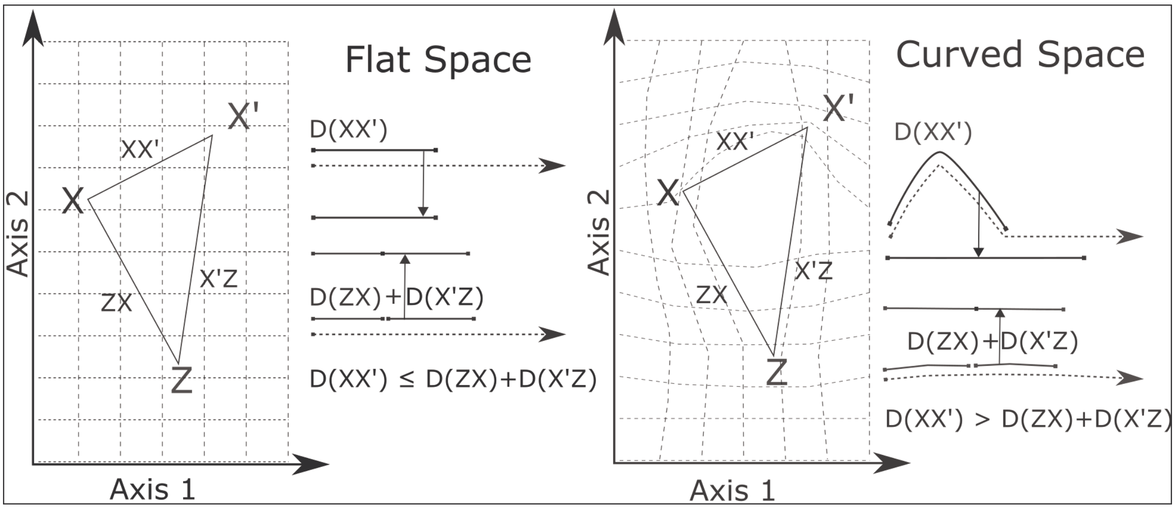

Figure 4.2: Geometric interpretation of the triangle-inequality violation. The drawing illustrates that the triangle inequality can be violated when the position-derived distance-quantification does not only quantify the biological distance but, in contrast, violates the *Biological Exclusivity* assumption from equation 2.1. The flat space grid illustrates the distance-quantification for the homogeneous WES identification method. Positions of CCLs X and X', as determined by their respective variants, are indicative of their biological distance $D(XX')$. The curved space grid illustrates the situation of the Uniquorn Bulk-RNA method. Importantly, positions can still be assigned to CCLs X and X' but the derived distance $D(XX')$ is not representative of their biological distance due to its amalgamation with the technological distance what precludes the statistical tests from effectively deciding on the sufficient similarity of CCLs.

A geometric example illustrating the Data-Incompleteness-derived curvature of the space is a scenario where a reference CCL was once sequenced with the Panel-sequencing format and once with the WGS technology i.e. a dramatic difference with respect to the amount of variants in the reference CCLs exists. Assume, that the amount of matching variants between either reference CCL and the query is identical and therefore the relative position of a query to the references is as well identical since unmatched variants in a reference do not alter the relative position of the query. The p-value of the corresponding tests on significant similarity will, however, differ because the underlying binomial test is based on the biased approximation of $\hat{p}$ i.e. the respective amount of variants on the references. The consequence is that the n-dimensional sphere that surrounds the reference, within which a test on sufficient similarity is significant, shrinks and grows based on the reference CCL's technology and format even when the location of the query remains identical.

Data-Heterogeneity occurs when the same genomic loci were sequenced but the sequencing platform or post-sequencing software differed. Even if the same biological source is sequenced, differences in the called variants will be observed [140], see Figure 4.3. This phenomenon can be caused by, but is not limited to, lane-specificity, non-identical reactant concentrations, differences between the performances of different variant calling algorithms. Data-Heterogeneity is suitably illustrated by Hudson et al. who compared the small missense variant calls of identical CCLs in the CCLE and COSMIC CLP study and found them coinciding at only 43% [127].

Consequently, the position-derived quantified distance will differ in that it becomes more volatile and random in correlation to the degree of heterogeneity. Note, that Data-Heterogeneity does not significantly affect the calculation of $\hat{p}$ but, geometrically interpreted, moves the query into or out of the n-dimensional sphere where the statistical test on sufficient similarity is significant while the sphere itself remains invariant.

In summary, Data-Incompleteness and Data-Heterogeneity have different origins (sequencing of different loci versus different processing of data from identical loci) and detrimentally affect different aspects of the identification (ineffectiveness of statistical identification tests versus volatility of distance-quantification). Importantly, both latent factors are frequently compounded when CCLs from different origins are compared what ultimately confounds the distance-quantification approach because the distances are not exclusively representative of the biological distances.

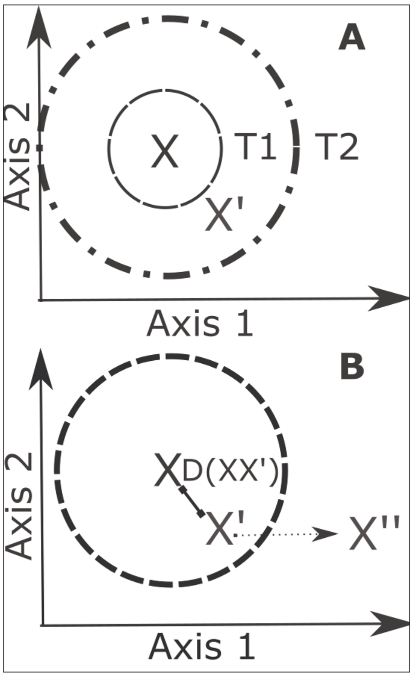

Figure 4.3: Impact of confounding factors. Data-Heterogeneity (A) has the impact that the sequencing technology of the reference decides whether a query CCL will be identified as identical or not. The reason is that different hard and software will result in differently called variants what impacts the size of the identification sphere, indicated here with technology 1 (T1) and technology 2 (T2). Data-Incompleteness (B) will cause the algorithm to incorrectly assign a position to a query CCL which is further away from the reference CCL than is correct (but never closer). The reason is, that the position of the query cannot be accurately determined because have been sequenced in the reference CCL that have not been sequenced during the generation of the query's NGS data.

4.1.2 Differences to Uniquorn WES

The rational for limiting Uniquorn WES to the Whole-Exome technology was to ensure Data-Homogeneity between the compared CCL profiles. Data-Homogeneity refers to the comparability of the sequencing technologies and formats what in particular means that identical CCLs have a comparable variant counts in their NGS data. Uniquorn universal's statistical model was, in contrast, specifically designed for the identification of CCLs sequenced with diverse sequencing technologies and formats that can cause reference CCL profiles to possess dramatically different variant counts, a challenge explained in the following Subsection 4.1.1.

Property	Uniquorn specific	Uniquorn universal
Benchmarked Technologies	DNA	DNA RNA
# Benchmarked Samples	1984	3896
# Variants Benchmarked	$0.97 \cdot 10^6$	$151 \cdot 10^6$
Benchmarked sequencing formats	Whole-Exome-seq	Panel-seq Whole-Exome-seq Bulk-RNA-seq Hybrid-capture
SNP filtering*	Yes	No

Table 4.1: The universal Uniquorn version extends the proof-of-concept Uniquorn significantly with respect to covered samples sizes, NGS-technologies and types of data processing. Furthermore, Uniquorn universal was benchmarked on a much wider and much more heterogeneous set of CCLs. *SNP-filtering refers to the post-sequencing of sequencing data regarding SNPs, such as filtering based on minor allele frequencies.

We benchmarked Uniquorn universal by identifying all identity-relationships in a set of 1612 RNA-sequenced CCLs (5309 related) and in a mixed set of 3596 RNA and DNA-sequencing CCL-profiles (11512 related). Ninety-six% of the relationships of the later RNA-sequencing CCL-profiles were correctly identified and 95% of the relationships were found in the mixed scenario i.e. when DNA-sequencing samples were used to identify RNA-sequencing samples and vice versa. A Panel-sequencing scenario was benchmarked by synthetically limiting the 3596 mixed-scenario samples to the set of genes contained in the Clearseq © /Agilent$^{\text{TM}}$, TruSight © /Illumina$^{\text{TM}}$ and Hotspot v2 © /Thermo Fisher$^{\text{TM}}$ Panel, respectively. Panel-sequencing showed sensitivities of 83% (151 genes, Clearseq), 82% (94 genes, TruSight) and 65% (49 genes, Hotspot v2). The algorithm is freely available as R package *Uniquorn* and contains the NCI-60 CCLs by default. Scientists can identify their own custom CCL-samples as well as publicly available CCL-samples.

4.2 Methods

Uniquorn universal's general identification concept and its workflow structure with respect to variant weighting and confidence score calculation remains identical to the Uniquorn proof-of-concept thus, see thesis subsection 3.2.1 for a description.

We introduced two statistical approaches which first quantify and secondly compensate Data-Heterogeneity and Incompleteness. In Subsection 4.2.1 we demonstrate how a beta-distribution based bias-factor estimation quantifies the strength of the bias on the identification p-value. Subsection 4.2.1 shows how statistical permutation resampling quantifies the volatility of the p-value calculations. Methods subsection 4.2.1 describes the thresholds that are dynamically determined based on the bias strength to compensate the effect of highly heterogeneous NGS data.

4.2.1 Quantification of spuriousness and filtering of false positive predictions

We observed that the degree of Data-Heterogeneity introduced by the diverse range of sequencing technologies caused a significant amount of matches to occur between unrelated q and r CCL samples in spite of the filtering of non-informative variants via weighting and filtering. We term these highly likely random and non-informative matches *spurious* matches because they cannot be caused by a common origin or the compared profiles but must be caused by spurious technological noise effects. The characteristic property of these spurious matches is, that are found in many q to r comparisons but by total amount seldomly exceed an absolute amount of three to ten randomly matching variants. Furthermore, false positive predictions show an amount of matching variants that is comparable to the average amount of matching variants in R.

We quantify the spuriousness between q and any r with a spuriousness variable $SP \in [0, .., 1], SP \in \mathbb{R}$. sp is estimated by the integral of the B function with parameters s_{max} and s_{mean}, where s_{max} is the maximal number of shared variants between q and any sample from r, and s_{mean} is the mean of the number of these matches. The B function has been found to suitably estimate the expected number of additional variants in that it is governed (1) by the relative number of matches and (2) by the absolute size of its input-parameters and (3) by its domain $\{x | x \in [0, .., 1], x \in \mathbb{R}\}$. Thereafter, a threshold on the acceptable amount of observed unmatched variants M_{min} is calculated as follows:

$$M_{min} := \frac{S_{mean} + S_{max} \cdot SP_R}{1 - SP_R} \tag{4.2.1}$$

Where the spuriousness of a reference database R, SP_R is computed as the integral of the beta distribution based on the ratio of the average amount m_{avg} and the maximum amount m_{max} of matching variants in R. We chose the integral of the beta distribution due to the integral's skewness, two-parameter positive integer domain for m_{avg}, m_{max} and real-valued co-domain between and including 0 and 1:

$$SP_R := \int_{p=0}^{p=1} beta_{CDF}(p; \alpha, \beta) := \int \frac{\Gamma(m_{max}) \cdot \Gamma(m_{avg})}{\Gamma(m_{max} + m_{avg})} \tag{4.2.2}$$

In a second step, we filter all overlaps with less matches than threshold T_R to exclusively retain overlaps that show a higher number of matches than expected by chance:

$$T_R := \left(\frac{m_{avg} + m_{avg} \cdot SP_L}{1 - SP_L} \right) \tag{4.2.3}$$

Thus, after quantification of the bias, a compensation is implemented.

Empirical sampling of the confidence score

In the specific case of a low-variant count e.g. Panel-sequenced q, can the exclusion or addition of a single variant render a test on q and r identity significant i.e. the test on identity becomes unstable and volatile. In particular when the variant in question is a spuriousness variant is present due to technological diversity i.e. noise can the volatility aspect be observed. The spuriousness-detection approach described in subsection 4.2.1 is generally not suited for Panel-sequenced CCL samples due to the low sample size due to insufficient variant-count sample-size. However, an exhaustive Jackknife approach which enumerates all possible leave-on-out configurations of features i.e. variants in q can be efficiently applied due to the low variant count on q, see section 2.3.2 for a description of the Jackknife method. The Jackknife-approach serves to estimate the 95% upper bound by the confidence scores CS is perturbed when single variants are excluded in order to quantify the volatility. A single leave-one-out confidence score calculation $\hat{CS_i}$ for variant var_i is defined as $\hat{CS_i} = -1 \cdot log_e D(q_i, r)$ where $q_i := var_i \notin q$. Let, v_{VS} be the increasingly ordered vector of all $\hat{CS_i}$. The upper confidence score volatility bound CS_{up} is the smallest $\hat{CS_i}$ for which the following holds:

$$CS_{up} := \hat{CS_i} \geq v_{VS}[\lceil |q| \cdot 0.95 \rceil] \tag{4.2.4}$$

Where $v_{VS}[i]$ signifies the i^{th} increasingly ordered permuted confidence score, i.e., CS_{up} is the ceiled

95% percentile bound.

Rejection of the null hypothesis

Three conditions have to be fulfilled for rejection of H_0 given to compensate for Data-Heterogeneity:

1. $CS_{q,r} \geq t$, user defined threshold (default value is t = 3)
2. $CS_{q,r} \geq CS_{up}$, the confidence scores must be greater than 95% confidence scores obtain by permutation
3. $CS_{q,r}$ must rank among the top-k positions of all r in R (default value k = 2)
4. $O(q,r) \geq T_R$, the amount of overlapping variants must be greater than can be expected due to spuriousness

Technology	Source	Genotyped Genes	Variant Calling Software	SNP Filtering*
RNA Bulk-seq	Klijn et al. GDC	Expressed alleles only	GATK RNA FreeBayes	None
Hybrid-capture	CCLE	1651	MuTect	> 0.01
Exome-seq	CGP Cellminer	20965 $> 20k$	Pindel Caveman GATK DNA	None

Table 4.2: Data differs with respect to sequencing technology, variant calling algorithms, SNP-filtering, and number of covered genes. Variants within Genomic Data Commons (GDC) and Klijn et al. repositories were manually called by first utilizing the Trimmomatic and the Spliced Transcripts Alignment to a Reference (STAR) aligner [141] and a subsequent diverging variant calling step: the Genome Analysis Toolkit (GATK)-RNA variant caller [142] was utilized for data from Klijn et al. and the FreeBayes variant-caller [143] for GDC data to increase the heterogeneity of the benchmarked data. *SNPs were pre-filtered by the creators of the data based the SNPs' minor allele frequency [144].

Evaluation

We benchmarked Uniquorn in its universal version using 3596 CCL-profiles derived from 1516 distinct CCL-samples from five libraries, each characterized by a different technology, see Table 4.2. We utilized the 3596 profiles both as reference and as queries, resulting in 3596 identification tasks and roughly 13 Million individual comparisons. Each query profile possessed between one and nine matching reference profiles (median = 3) because many CCLs are contained in more than one library. In addition to obtaining key performance indicators (Tables 4.5, 4.4 and 4.6), we also assessed whether the performance was biased related to certain properties of the profiles such as sequencing technology (Figure 4.5 and Supplementary Material Figure 7.5).

Sensitivity was defined as the fraction of all predictions which correctly predicted that two CCL profiles were similar and specificity as the fraction of all predictions which correctly stated that two CCL profiles were not similar.

4.2.2 Gold-standard creation

We created a Gold Standard based on CCL names and literature research. Firstly, names of CCLs were either parsed from the VCF-files directly (Cellminer, GDC, Klijn et al.) or extracted from the meta-file that aggregated the variant-calls of all CCL-profiles into a single document (CCLE, Cancer Genome Project (CGP)). Secondly, a pre-processing step removed all non-alpha-decimal characters and spaces from the names and capitalized the processed names. CCLs that differed only by a prefix or by a suffix, such as *MDA-MB-435* and *MDA-MB-435S*, were considered candidates for being identical and validated using literature. Also, collisions of different CCLs that had the same name after the pre-processing e.g. *TT* and *T.T* were resolved by literature research. This process resulted in 11508 identity-relationships of which 5309 are based on RNA-sequencing profiles. Supplementary Table 7.1 contains the gold-standard contains the identity-definitions based on reports and a link to the reports where needed.

4.2.3 General data procurement and creation of the Panel-sequencing data

The CCL profiles of all libraries we considered were obtained by either DNA or RNA sequencing. However, labs often only perform Panel sequencing with their samples to save on cost and labor [145, 122]. To test the capability of Uniquorn universal to identify a Panel-sequenced sample within an RNA or DNA sequenced library, we created synthetic Panel-sequencing profiles by removing all variants from a profile that fall outside the region of three predefined Panels, i.e., gene set. Firstly, we formatted all profiles into the VCF-format and secondly bedtools intersected all VCF-files with Browser Extensible Data file (BED)-files containing the genomic coordinates of the Panels [146]. The TruSight's BED-file (trusight_cancer_manifest_a.bed) was obtained from illumina.com. The websites of the Hotspot v2 thermofisher.com and the ClearSeq Panel agilent.com did not provide the Panels' genomic-coordinates in BED but comma-separated format and thus we manually converted the comma-separated files into the BED-format using BioMart[147].

We procured the data either in the VCF-format or as Binary sequence alignment file (BAM)-files, see Table 4.3. BAM-files were converted into FAST-ALL Quality file (FASTQ)-files and conscientiously processed with different variant calling algorithms to obtain VCF-files 4.2. The CCL-profiles from the CGP and CCLE repositories were extracted from the meta-files and transformed into VCF-files. R version 3.5.1 (2018-07-02) was utilized on a Linux Debian Mint operating system and benchmarks performed with the Bioconductor *Uniquorn* package 2.0.031 [148, 149].

Library	URL	Files	Date
Klijn et al.	ebi.ac.uk	BAMs	July 16th 2017
GDC	gdc.cancer.gov	BAMs	May 24th 2017
CGP	sftp-cancer.sanger.ac.uk	CosmicCLP_MutantExport.tsv	January 13th 2017
CCLE	Broadinstitute.org/ccle	CCLE_hybrid_capture1650_hg19_NoCommonSNPs_CDS_2012.05.07.maf	
Cellminer	discover.nci.nih.gov	VCFs	

Table 4.3: Origin and name of utilized files used for the benchmark are shown. Klijn et al. [150], GDC[151] CGP[152], CCLE [4] and Cellminer [153, 154] were procured.

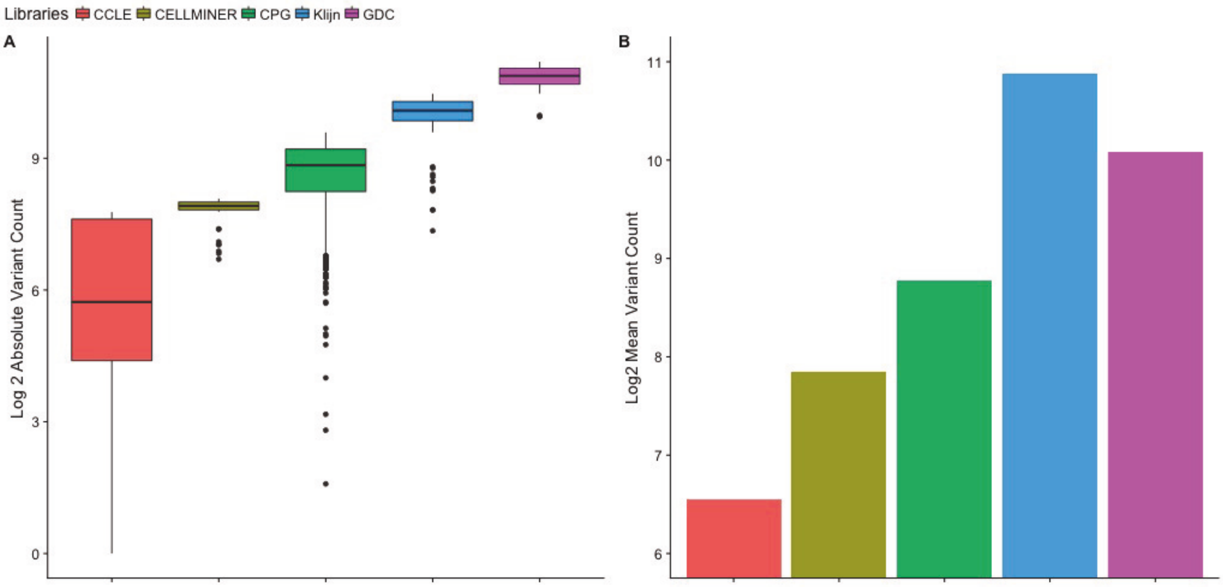

Figure 4.4: (A) Absolute amount of variants per benchmarked library. (B) Mean amount of variants per profile per benchmarked library. All repositories differed by at least one power of two with respect to the amount of variants they contain i.e. are heterogeneous. Whiskers depict the standard deviation of the mean variant-counts

4.3 Results

4.3.1 Identification of Bulk RNA and Panel-sequenced CCLs

CCLs are essential tools for cancer research but are also highly susceptible to misidentification, which makes the accurate identification of a CCL used in an experiment crucial [145, 155]. We recently published the Uniquorn proof-of-concept, a method to identify CCLs using variant profiles derived from exome DNA-sequencing or from hybrid-capture DNA-sequencing [18]. Here, we present the universal version of Uniquorn which can robustly identify RNA and Panel-sequenced CCLs derived from heterogeneous sequencing technologies while retaining Uniquorn's ability and performance to identify DNA-sequenced CCLs [19]. Furthermore, Uniquorn universal no longer relies on SNP-filtering, which brings its own problems (such as the concrete set of SNPs to filter) when using pre-computed profiles.

We benchmarked Uniquorn universal on NGS data from 1612 RNA, 1080 DNA-exome and 904 targeted hybrid-capture sequenced CCLs from five repositories, in the following called libraries, which utilized four different sequencing technologies to adequately reflect the heterogeneity of a real-world scenario (Table 4.2 and Figure 4.4). Four identification scenarios were benchmarked of which three were novel and not covered by Uniquorn proof-of-concept: RNA-sequencing identification (Table 4.4), mixed RNA-sequencing and DNA-sequencing identification (Table 4.5), Panel-sequencing identification (Table 4.6) and Uniquorn proof-of-concept's DNA-sequencing only scenario (Supplementary Material Table 7.2). It was benchmarked whether a CCL was correctly identified when comparing it to all reference CCL-profiles from all five reference libraries, leading to ~13 million CCL benchmark comparisons overall. Since a TP prediction was only possible for about 11,000 of the ~13 million comparisons, our evaluations put special emphasis on the PPV.

4.3.2 Cross-validation benchmark

The first finding was that Uniquorn universal could effectively identify full-transcriptome sequenced CCL-profiles: with default parameters (Weight Threshold 0.5), Uniquorn universal's sensitivity to identify RNA-sequenced CCLs reached 95.7% its PPV 85.5% (Table 2). The rationale for choosing 0.5 as default weight threshold is shown in Supplementary Material Figures 7.2 and 7.3.

Threshold	1.0	0.5	0.25	0.0
Possible TP	5309			
TP	5096	5082	5071	4192
FN	213	227	237	1117
FP	850	860	865	1411
Sensitivity %	96.0	95.7	95.5	79.0
Specificity %	99			
F1%	90.6	90.3	90.2	76.8
PPV%	85.7	85.5	85.4	74.8

Table 4.4: The performance of Uniquorn universal to identify full-transcriptome sequenced CCL-profiles is shown. 1612 of such profiles were identified within five reference libraries containing 3596 DNA and RNA-sequencing sequenced CCLs. Columns 2 to 5 show key measures dependent on the mutational inclusion weight (see methods). Inclusion weights 1.0, 0.5 and 0.25 showed comparable performance with sensitivities above 95%. 0.5 is the default parameter setting of the Uniquorn R-package.

The second finding was that Uniquorn universal could effectively identify CCL profiles in a real-word scenario: Heterogeneously created RNA-sequencing and DNA-sequencing CCL-profiles had to be identified by equally heterogeneously created reference CCL-profiles what resulted in an average sensitivity of 95% and average PPV of 90% (Table 4.5). Both RNA-sequencing and mixed-sequencing benchmarks showed extremely high specificity values (99.9% and higher) which were caused by the very large number of true negative predictions.

Threshold	1.0	0.5	0.25	0.0
Possible TP	11512			
TP	10951	10945	10937	9843
FN	561	567	575	1326
FP	1128	1106	1139	4626
Sensitivity %	95.1	95.1	95.0	85.5
Specificity %	99			
F1%	92.8	92.9	92.7	85.5
PPV%	90.7	90.8	90.6	85.4

Table 4.5: Uniquorn universal's ability to identify CCL-profiles created and identified by RNA-seq, DNA-exome and DNA-hybrid-capture CCL-profiles is shown to determine the expected real-word use-case performance. 3596 CCLs that were sequenced and processed with various technologies and algorithms were identified (see Tables 4.1 and 4.2 for technologies). The sensitivity was comparable to the RNA-sequencing benchmark (Table 2) with the exception of inclusion weight 0.5 which resulted in a higher F1-score and PPV than weight 1.0. A performance drop can be observed for weight threshold 0.0 where all variants, informative and non-informative, were utilized.

The 3596 available reference CCL profiles were reduced to the genomic regions covered by three of the most widely utilized ClearSight© , TruSight© and Hotspot v2© Panels to simulate Panel-sequencing benchmark profiles. Identification of the resulting $3 \cdot 3596 = 10788$ Panel-profiles revealed as third finding that Panel-sequenced profiles could be successfully identified with an average sensitivity of 82% and PPV of 68% if the Panel covered more than 100 genes (Table 4.6). Panels covering less than 100 genes were significantly less suited for CCL-identification with an average sensitivity of 60% and a PPV of 55%. Specificity always remained higher than 99%. False-negative and false-positive identifications were found to be predominantly caused by CCL-profiles that covered less than 100 genes.

Subsequently, we analyzed what factors caused Uniquorn universal to incorrectly classify i.e. identify a CCL-profile and it was determined that technological heterogeneity does not significantly impact Uniquorn universal's sensitivity and F1 score 4.5. However, although sensitivity and F1 score remained robust with respect to the utilized technology, sensitivity showed a strong positive correlation (r of 0.7) with the amount of genes covered by a profile. The uncovered a log-linear sensitivity to amount-of-covered-genes relationship is shown in Supplementary Material Figure 7.4 and the benchmark results for each library are shown in Supplementary Material Figure 7.5. In contrast, the PPV showed a limited bias with respect to utilized sequencing technology and no log-linear relationship to the amount of covered genes.

Panel	Clearsight	TruSight	Hotspot v2
Genes	151	94	49
Possible TP	11512		
TP	9505	9423	7525
FN	2007	2089	3987
FP	4591	4424	6097
Sensitivity %	82.6	81.9	65.4
Specificity %	99		
F1%	74.2	74.3	59.9
PPV%	67.4	68.1	55.2

Table 4.6: Uniquorn universal achieves sensitivities of ~83%, ~82% and ~65% while constantly showing a specificity of higher than 99% at default parameters for Panel-sequencing identification.

We focused on investigating what effect a varying choice of identification parameters had on the identification performance and found that sensitivity remained robust, in particular for high-volume full transcriptome sequenced CCLs while specificity was moderately affected and the PPV showing high sensitivity to the choice of the identification threshold, shown in Supplementary Material Figure 7.1.

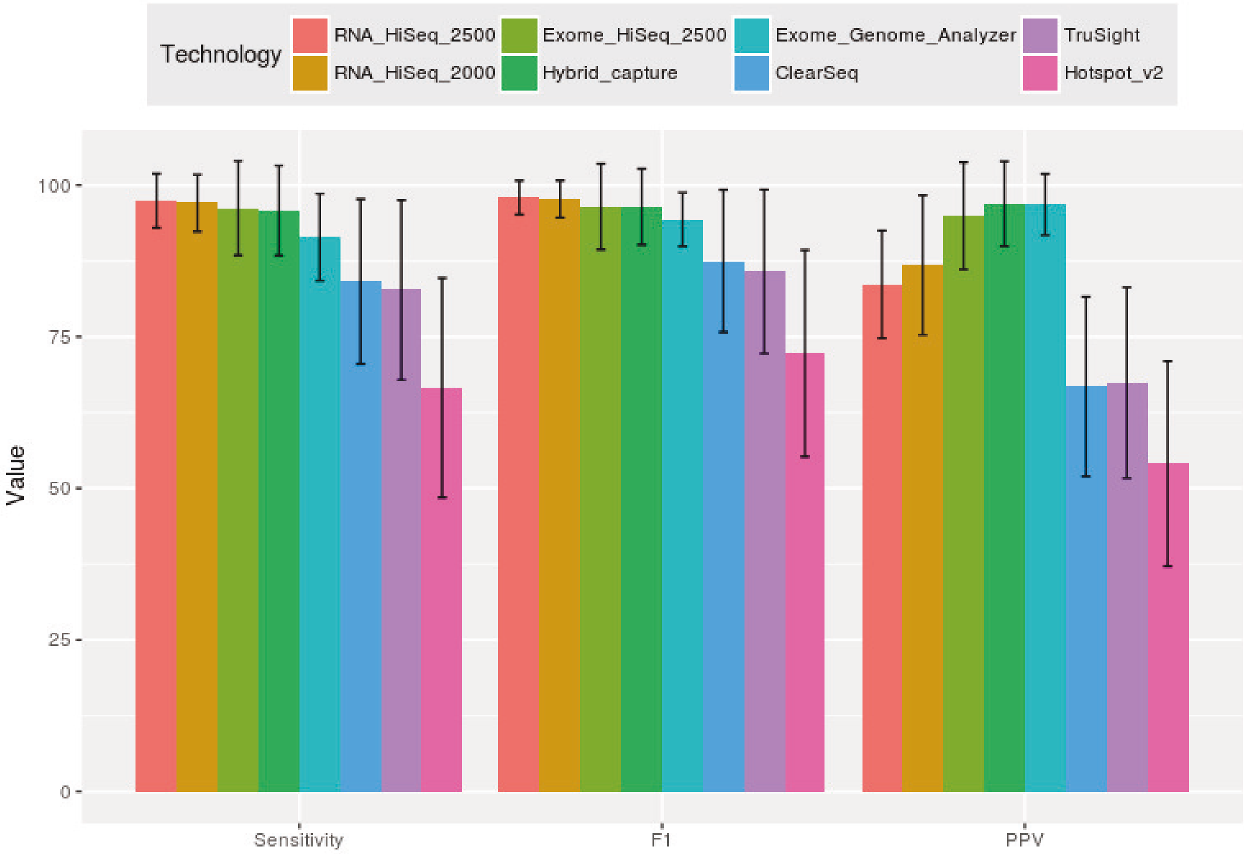

Figure 4.5: Relationship between Data-Heterogeneity and identification performance. CCL profile sequenced and processed by vastly different technologies and algorithms were identified and determined whether Uniquorn universal's identification performance remained robust in spite of the Data-Heterogeneity. Bars depict average performance, whiskers standard deviation. Profile sizes of the query CCL shrink dramatically from left (2^{10} variants) to right (50 variants). Sensitivity and F1 score are highest when full transcriptome profiles are used and lowest for small Panel-sequencing profiles but remain robust when faced with different technologies. In general, PPV decreases with the profile size with the exception of WES and hybrid-capture technologies, which show a higher sensitivity.

4.4 Discussion

4.4.1 Feasibility of the identification of Bulk and Panel-sequenced CCLs

Uniquorn universal is optimized for the identification of CCLs whose variant profiles were obtained from diverse technologies and diverging computational processing pipelines. Thus, it complements established methods by addressing some of their key limitations: 1) The physical CCL sample is not required, as it is, for instance, in the case of STR-counting-based identification) Uniquorn universal is agnostic to sequencing technology and thus able to reuse data provided by the creators of CCL libraries. The support of various RNA and DNA sequencing format as well shows the significantly advantage of relaxing the requirement on reference libraries since the do not have to be identical to the queries format i.e. one reference CCL image of one format suffices.

We benchmarked the performance of the algorithm in high-diversity scenarios, which we consider best mimic the real situation, in laboratories dealing with CCLs, confirming its ability to cope with various sequencing technologies and data-processing (Table 4.2). This considerably extends the functionality of Uniquorn proof-of-concept to also handle RNA and Panel-sequenced CCLs (Tables 4.4 and 4.6).

Panel sequenced profiles were simulated by reducing the amounts of covered genes of the 3596 available profiles from about 22,000 down to 151, 94 and 49 covered genes, respectively. Differences in the identification efficiency of the benchmarked Panels (Agilent ClearSight, Illumina TruSight, Thermo-Fisher Hotspot v2) was therefore caused by differing amounts of covered genes and not due heterogeneous technology since the variants call within the covered genes were identical for each Panel. Significant differences regarding sensitivity, F1-score and PPV were detected between the Panels, indicating that not the sequencing technology (Figure 4.5) but the number of covered genes is most influential with respect to how efficiently a CCL profile can be identified (Supplementary Material Figure 7.4). Remarkably, the identification efficiency of Panel-sequencing profiles was merely 12% to 13% lower than the efficiency measured for full transcriptome sized CCL-profiles although the Panels covered orders of magnitude less genes than the full-transcriptome profiles. An exception was the hotspot v2 Panel which showed a significantly decreased sensitivity of 65% which was 30% lower than the full-transcriptome profile identification but as well only covered 49 genes. We therefore deem the concept of identification by pair-wise distance-quantification as successfully applicable in case of Uniquorn universal due to its successful benchmark results.

4.4.2 Reasons for incorrect Identifications

By manual inspection of benchmark results (available online in [19] as 'Supplementary Material Table 1') we found that FP predictions are associated with CCLs that had diverged significantly from their origin due to long-term subclonation or exposure to drug treatment e.g. the *CEM-2*, *Jurkat* and *CCRF-CEM* CCLs. This finding is supported by reports of the same phenomenon for the same CCLs when STR-identification was applied [126]. False-negative predictions where furthermore frequently associated with CCLs whose relationship-status could not be fully resolved due to an unclear nomenclature: E.g. when it was unclear whether CCLs with a similar name were different or identical CCLs or in the case of false-positive, whether CCLs with different names were nevertheless identical but counted as FP predictions by the Gold Standard (Supplementary Material Table 7.1) which lists numerous labeling inconsistencies. Thus, low variant-counts and an unclear relationship caused by the absence of a generally applied CCL-nomenclature system are still the dominant causes of incorrect predictions.

We deem the applied empirical resampling and spuriousness quantification approaches outlined in Subsections 4.2.1 and 4.2.1 as adequate. The main reason for this conclusion are Uniquorn universal's real-world benchmark results shown table 4.5 and the comparative benchmark results between Uniquorn proof-of-concept and Uniquorn universal in supplementary table 7.2 where a minor loss of sensitivity in exchange for a significant gain of robustness was observed.

4.5 Conclusion

Uniquorn in its universal version complements established methods in particular when those cannot be applied e.g. due to absence of a physical sample. The Uniquorn universal method supports quality-assurance procedures in high-CCL-throughput laboratories since it seamlessly integrates into analysis pipelines to serve as a quick test for in-house or procured third-party CCL-profiles. The Uniquorn universal method is freely available as Bioconductor R-package *Uniquorn* and can be easily implemented.

Users of the generalized Uniquorn method can utilize their own sets of CCL-profiles as reference. However, as the run time of Uniquorn in its generalized version is very low, it is advisable to always include a wide range of reference profiles to also detect unexpected contamination. The CGP and CCLE repositories contain 1695 CCL-profiles while showing a low false-negative rate as references and are freely available. The *Uniquorn* R-package is ported with the limited NCI-60 reference Panel but a tutorial that enables researcher to easily utilize the 1695 CGP and CCLE CCLs is documented in the *Uniquorn* Bioconductor vignette. The Klijn et al. [150] and GDC CCL-repositories show suitable identification characteristics and can be obtained by application at the European Genome-phenome Archive (EGA).

The utilization of a an abstract distance concept to identify CCLs has proven susceptible for Data-Heterogeneity and Data-Incompleteness but proved successful after a modification of the concept, as judged by the benchmark results. Detailed analyses of factors influencing the identification of CCL-profiles such as SNP filtering are indicated to further improve the Uniquorn universal method. A further extension to non-neoplastic Cell Lines, single or methyl-sequenced CCLs are viable subjects for future work to further expand the range of research fields which can utilize the generalized Uniquorn method.

Chapter 5

Data-Augmentation via Distance-Quantification based on Transcriptomic Deconvolution

This Chapter introduces a novel method to predict clinically relevant properties of rare and diverse neoplasms via Data-Augmentation. The biological background will be introduced first and thereafter the distance-quantification-based data-augmentation explained. A following benchmark determines whether ML models trained on the augmented data can effectively characterize rare and diverse neoplasms based on the output of a transcriptomic deconvolution. A discussion highlights advantages and disadvantages of the approach while the conclusion Section argues whether the distance-quantification approach could effectively augment the data and render a neoplastic characterization without neoplastic training data possible.

5.1 Introduction

The personalization of the patient treatment is in the prime focus of current cancer research. It is defined as the adjustment of the treatment to individual neoplastic characteristics and promises to help identify more effective drug-regimes, to reduce side effects and ultimately to prolong the patient-survival while reducing monetary costs [156, 157]. Personalized treatment constitutes a particularly urgent need in case of rare cancers with highly variable and unpredictable clinical courses, such as NENs and, more specifically, Pancreatic Neuroendocrine Neoplasm (PanNEN). Well differentiated PanNEN are referred to as Neuroendocrine Tumors (NETs) and typically exhibit a low (G1, G2) or, in rare cases, high (G3) proliferative index, as quantified by Proliferation marker protein Ki-67 (Ki-67) gold-standard staining, with median survival of patients exceeding ten years [158]. This overall indolent course of the disease stresses the need for careful balancing of treatment benefits and side effects. In contrast, patients with poorly differentiated Neuroendocrine Carcinomas (NECs) face a dismal prognosis of a few months and hence are eligible for aggressive and ideally personalized therapies [159, 160].

Endeavours to personalize the treatment of PanNEN and the more general group of gastroenteropancreatic neuroendocrine neoplasms (GEP-NENs), arising in the gastro-enteropancreatic system, suffer from the impediments of small sample numbers and imbalanced sample characteristics. First, GEP-NENs are rare: the current age adjusted incidence rate of GEP-NENs is estimated as 7.38 cases per 100,000 persons for well differentiated GEP-NENs, currently defined as NETs [158, 161] in the United States of America where they as well only account for 1-2% of pancreatic tumors [162]. Secondly, GEP-NENs are highly heterogeneous, with low proliferative, well differentiated G1 and G2 NETs being overrepresented in all publicly available datasets, while poorly differentiated NECs are underrepresented [163]. Thirdly, ambiguity with respect to morphologic NEC-versus-NET subtype classification is a frequently encountered issue even by experienced pathologists, questioning the reliability of currently available classifications [164, 165, 166]. However, precise subtype determination constitutes a key element of personalization. It relies on a characterization of a tumor's molecular landscape [167] which can be revealed by either *in-vivo*, *in-vitro* or *in-silico* methods. *In-vivo* methods, such as medical imaging, and *in-vitro* methods, such as Ki-67 Immune Histo-Chemistry (IHS), currently constitute the gold-standard methods. However, even these approaches are limited in their ability to discern subtypes in samples with ambiguous morphologies and same proliferation rates [167]. Therefore, new approaches that complement the gold-standard approaches in case of ambiguity are highly needed.

Throughout the last decade, ML models have become the main *in-silico* approach for the classification of neoplastic samples based on the NGS of the samples' molecular landscape [168]. Machine-Learning (ML) models, however, require training on large amounts of data covering all subtypes of neoplasms. Accordingly, the beneficial usage of ML models for PanNENs is challenging since sample numbers are low while class imbalance is high: All publicly available datasets are severely skewed towards low and medium grade neoplasms.

The impact that insufficient training data has can be leveraged detrimentally if a cancer type is simultaneously rare and diverse. Two correlated aspects impair the analysis of rare yet biologically diverse cancer types: The rareness of a cancer not only diminishes the overall amount of available training data, it frequently causes a class-imbalance within the training data which in turn reduces the predictive power of the ML models [169]. Different subtypes of cancer can have different incidence-rates and therefore differ in their respective likelihood to present with biopsies required for sequencing. Provided that a rare cancer type is diverse, multiple sub-categories of the cancer have to be sufficiently populated with training data to achieve a class-balanced performance of the ML models. The most often occurring subtype of a rare cancer will therefore generally be predicted best by a ML model due to the corresponding data availability. The subtype with greatest incidence-rate might, however, might not be the most aggressive or scientifically valuable subtype of the cancer with greatest need for accurate ML predictions [166]. PanNENs are an example os such a type of cancer because its high-grade, malignant neoplasms are scarcely available compared to benign low-grade neoplasms. The need to predict the clinical characteristics of high-grade PanNENs with great statistical performance is highly pressing, yet ML model can be more effectively trained to characterize the low grade PanNENs due to the increased availability of training data [164]. The amount of samples drawn from the overall population of the cancer type therefore has to be significantly increased to cover the rarely observed subtypes what, however, is counteracted by the cancer type's rareness [169].

5.1.1 Distance-quantification via deconvolution

Distance-quantification between entities is an established concept in the field of Bioinformatical neoplastic classification [170, 171, 172]. Entities can be classified via measurement of their distance to a informative reference entity which may or may not be neoplastic itself. This entity can, for instance, be a holotype of a high-grade series carcinoma in order to quantify how alike a neoplasm is to a malignant carcinoma [17]. Importantly, the distance-quantification approach can be exploited to augment training data of neoplasms where suitable samples are scarce. The amount of available training data is limited for the types of cancer which is why the substitution of neoplastic data is tempting. Instead of quantifying the distance between neoplastic entities we will show that a distance-quantification between neoplastic and healthy entities is informative with respect to clinical meta-data such as the grading. This neoplastic-to-healthy distance-quantification takes place via deconvolving the neoplastic transcriptome into healthy cell-type fractions. The deconvolution is essential because, in analogy to the 'kernel-trick' utilized for the SVM, is the transcriptomic data of the neoplasm projected on a different space, here, the deconvolution-results.

The main assumptions underlying classification-by-deconvolution are that:

1. Low-grade neoplasms are more similar to healthy cells than high-grade neoplasms
2. Different clinical characteristics present with different cell-type proportion and reconstruction errors
3. That deconvolution-derived results allow the quantification of meaningful distance between neoplasms and health cells

The distance-quantification-concept differs from Chapter 3 and Chapter 4 in that the distance was quantified via a norm of a vector representing the distance between a query sample and a reconstruction of that sample. Additionally, a second vector comprising of cell-type proportion predictions serves to train a ML model which predicted the grading as function of the distance to healthy cells.

5.2 Methods

5.2.1 Utilized datasets

We procured three of the GEP-NEN datasets from the publicly accessible gene omnibus database Gene Expression Omnibus (GEO) and obtained the Scarpa et al. dataset from ICGC [173, 147]. The Riemer et al. dataset was made available by C. Grötzinger, Charité Berlin. For a listing of the Riemer dataset, see Supplementary Table 7.4. Seven scRNA deconvolution training datasets were located on publicly available GEO servers with the exception of the Segerstolpe et al. dataset that was acquired from the Array Express database [147, 148]. Four additional scRNA training datasets (Grün [42], Haber [174], Stanescu [175], Yan [176]) were subjected to preliminary benchmarks but not to detailed result analyses based on ranking, which revealed inferior performance for the purpose of GEP-NEN deconvolution.

Data set	Purpose	Procurement	Date
Baron	Benchmark	GSE84133, GEO	May 5th 2018
Califano	Training	GSE98894, GEO	May 5th 2018
Fadista	Out-group	GSE50244, GEO	February 11th 2019
Grün	Training (Discarded)	GSE81076, GEO	May 5th 2018
Haber	HISC Training	GSE92332, GEO	May 5th 2018
Lawlor	Benchmark	GSE86473, GEO	May 5th 2018
Missiaglia	Benchmark	GSE73339, GEO	May 5th 2018
Sadanandam	Benchmark	GSE73338, GEO	May 5th 2018
Scarpa	Benchmark	EGAS00001001732, ICGC	June 1st 2017
Segerstolpe	Training	E-MTAB-5061, Array Express	July 15th 2018
Stanescu	Training (discared)	GSE78510, GEO	February 2nd 2018
Riemer	Benchmark	Unpublished	June 1st 2015
Yan	Training (discared)	GSE36552, GEO	February 2nd 2018

Table 5.1: Overview of the data sets obtained to train and benchmark the deconvolution framework. *Purpose* indicates for what purpose the datasets were utilized, *Source* indicates the source of the dataset and *Date* shows the date that the data was obtained.

The ranking evaluated whether the deconvolution algorithms were tested on a given datasets, the amount of sequenced cells, a stratification of the cell-types roughly correlated to the stratification in healthy tissue and the technological homogeneity of the datasets compared to each other in order to facilitate the interpretation of benchmark results.

From the sources listed in Table 5.1, we obtained 364 Bulk RNA-seq and 20,953 scRNA samples. All samples underwent a sequencing data quality assurance process based on the publication of Conesa et al., which included but was not limited to analyses of read-counts and read-qualities in addition to outlier and heteroscedasticity detection [177]. Where required, reads were clipped and adapter sequences removed. For an overview of the datasets' properties, see Table 5.2.

Dataset	Type	Grading G1 G2 G3	Patient survival information
Baron [178]	Pancreas scRNA	8569 GO	-
Califano[179]	PanNEN Bulk RNA	105 G1 & G2	x
Fadista* [180]	Pancreas Bulk RNA	89 non neoplastic	NA
Haber [174]	Pancreas scRNA	642 G0	-
Lawlor [181]	Pancreas scRNA	6102 G0	-
Missiaglia [182]	PanNEN mRNA array	46 G1 25 G2 4 G3	x
Representative Set (RepSet)	GEP-NEN Bulk RNA	14 G1 23 G2 9 G3 NET 9 G3 NEC 14 Ambiguous	✓
Riemer	GEP-NEN & Bulk RNA	0 G1 10 G2 8 G3 NET 9 G3 NEC 13 Ambiguous	✓
Sadanandam [163]	PanNEN Bulk RNA	7 12 8	x
Scarpa [183]	PanNEN Bulk RNA	14 13 2	✓
Segerstolpe [184]	Pancreas scRNA	3514 G0	-

Table 5.2: Overview of deconvolved PanNEN, GEP-NEN and out-group datasets. The distribution of samples with G3, G2 or G1 grading was unequal between the datasets: 30 of the 44 high-grade G3 samples were part of the Riemer dataset which however, lacked low-grade G1 samples. Non pancreatic colorectal and gastric samples were included in the Riemer dataset and represented exclusively G3 samples. The Scarpa dataset was limited to two G3 graded samples but contained 14 out of 78 G1 and 13 of the 69 G2 graded samples. G0 indicates that samples were non-neoplastic control samples. *The Fadista dataset was utilized for an out-group sanity test that analyzed whether proportion to proliferation rate correlation rates were cancer specific.

The Baron, Lawlor and Segerstolpe datasets were utilized to train the deconvolution algorithms on exocrine and endocrine cell-types. The Haber et al. dataset was utilized to quantify the similarity of the neoplasms to Human Intestinal Stem Cells (HISCs) by measuring their predicted relative proportion.

The Riemer and Scarpa datasets were combined to create the RepSet on the grounds that they possessed the greatest technological homogeneity of all studies while simultaneously being representative of a wide range of GEP-NEN types. The construction was necessary on the grounds that the deconvolution results vary between different grades with none of the available datasets possessing a sufficiently balanced grade distribution which would representatively demonstrate the changes of deconvolution results. The RepSet consisted of the 29 PanNEN from Scarpa et al. and the 23 PanNEN and 17 non-pancreatic GEP-NEN from Riemer et al. to which no modifications were applied. A clustering of samples according to the biological properties such as grading and not study of origin was verified as shown in Supplementary Figure 7.7. We controlled for immune-infiltration via application of the Estimation of STromal and Immune cells in MAlignant Tumours using Expression data (ESTIMATE) algorithm in the RepSet and found no immune infiltration or stromal tissue contamination [185].

Software

All read-based analyses were based on the human reference genome GRCh38 [186]. The GATK RNA-seq gold-standard pipeline as described by GATK was used for mutation calling [142]. Transcript fusion analyses were applied with the STAR fusion detection algorithm [141]. Reads were clipped and adapters removed by the trim-galore software [187]. Transcripts Per Million bases (TPM) counts were utilized for analyses and generated by the Kallisto software from the quality-assured .fastqc files [188]. We ran differential expression analyses via the 'DESeq2' R package where we formulated the design matrix based on cohort and study membership to exclude potential batch effects during differential expression analysis [189]. 'Ggplot2' and 'ggbiplot' were utilized for graphics generation. 'Survival', 'sleuth', 'biomaRt' and 'RocR' were further R packages utilized for numeric analyses and the 'stringR' R package for string related operations [190, 191, 192]. The software 'GSEA' as provided by the Broad institute, Linux version 4.0.2 was utilized for enrichment analyses [193].

5.2.2 Deconvolution algorithms

Deconvolution into relative cell-type proportions requires training on transcriptomes of cell with specified cell-type. The classification can either take place before the sequencing in form of Fluorescence Activated Cell Sorting (FACS) or via the assignment of a cell-type after a single-cell sequencing run via *in-silico* methods. We trained exclusively on scRNA data from endocrine, exocrine and stem cells that were assigned a cell-type after sequencing. The utilzied assignment algorithms differed from study to study and represented a source of volatility.
We created different deconvolution models, one endocrine-only model composed of α, β, γ and δ cell-types and an exocrine cell-type model comprising of ductal and acinar cells. The exocrine cell-type model was limited to ductal and acinar cells due to sample-size constraints. Analogously, epsilon cell-types were excluded from the endocrine-only model due to insufficient sample sizes. Endocrine cell-types were included since NENs originate from endocrine cell-types. Exocrine cell-types were included due to the demonstrated feasibility of Pancreatic Ductal Adenocarcinoma (PDAC)-deconvolution.

Additionally, the reported high carcinogenic potential of the ductal cell-type motivated the approximation of high-grade series cell-types with a model that contained a ductal cell-type signature. HISC were included in models to analyze whether a correlation between the degree of de-differentiation and the HISC cell-type proportion exists as has been reported by previous studies [17]. The HISC proportions were initially trained on both fetal and intestinal stem cells but subsequently limited to the HISC datasets due to superior statistical efficacy [28, 174].

Neither normalization nor log-transformation was applied at any point during the deconvolution and the amount of permutations per deconvolution was set to 10E3, where applicable. The three deconvolution algorithms were each trained on three pancreatic scRNA datasets and their results compared: BSeq-sc, MuSiC and Moffitt, see Supplementary Tables 7.5 (CIBERSORT), 7.6 (MuSiC) and 7.7 (Moffitt). Cell-type proportion predictions were analyzed for the optimal BSeq-sc and Baron scRNA combination and the regression coefficients of the ν-SVR as calculated by BSeq-sc utilized as cell-type proportion predictions.

The BSeq-sc 1.0 R-implementation algorithm was acquired from `cibersort.stanford.edu`. Beforehand, the most recent version 1.4 of the csSAM R-package required to run BSeq-sc had been obtained from github [81]. The MuSiC algorithm version 0.1.1 was obtained from the GitHub repository `github.com/xuranw/MuSiC`. The Moffit et al. NMF algorithm was trained according to the specifications laid out in the corresponding publication which was replicated with the 'NMF' R-package version 0.22 [194].

Before the models were trained, a differential expression analysis was performed to identify 800 marker genes whose expression was significantly higher in a given cell-type compared to all other cell-types, utilizing the limma R package [195]. Note that models were thus trained on an aggregate of about 4000-5000 genes since each cell-type contained its pair-wisely unique set of marker genes. The amount of 800 genes per cell-type was selected as trade-off between performance, computation time and memory restrictions. The MuSiC algorithm could select the most suited subset of genes independently but due to the aforementioned time and memory constraints, we limited the amount of genes deconvolved by MuSiC as well.

We ascertained that the deconvolution models were comparatively expressed between the non-pancreatic and pancreatic tissues by conducting a differential expression analyses between the pancreatic and non-pancreatic tissue followed by the determination of the intersect of the significantly differentially expressed genes with the marker gene signature. We found that only 4% of the ductal and 29% of the HISC signature genes showed a differential expression activity between pancreatic and non-pancreatic tissue and therefore concluded that the amount of marker genes with a tissue-bias was sufficiently low as not to significantly influence the Riemer and RepSet benchmark results.

We furthermore ensured that the ductal marker genes were not associated with proliferation activity by calculating their overlap with the proliferation-specific GO-annotation geneset 'CELL PROLIFERATION GO 0008283'. We found the overlap to amount to 5% and therefore not to constitute a confounding factor for the deconvolution. The machine-learning models which predict the clinical characteristic were exclusively trained on deconvolution-derived results, such as the relative cell-type proportions, which did not contain any directly proliferation-associated feature.

Machine-Learning model training and Survival tests

We applied a multiclass Random-Forest algorithm trained by the R caret package [196] to differentiate between either combined G1 and G2 and G3 GEP-NENs, consisting of NETs and NECs. We compared the predicted Ki-67, ductal and HISC proportion-based sensitivities, specificities, F1-Score and ROC curves.

The features of the ML model were the Root mean square error (RMSE) of the transcriptomic reconstruction, the reconstruction p-value and the cell-type predictions depending on the model (endocrine-only, endocrine and exocrine or endocrine and HISC). For the Fadista dataset sanity test a generalized linear model was trained identically to the deconvolution model i.e. we commenced by identifying the cell-type specific differentially expressed genes and trained a regression model on this subset of genes on the scRNA training datasets. Since no grading classification was possible in non-neoplastic tissue, we linearly predicted the Ki-67 count levels in the patient-derived data via linear regression. The Califano et al. dataset did not provide grading information, however the dataset was benchmarked as an unsupervised deconvolution cohort and found that the distribution of the resulting deconvolution models p-values were comparable to those of all other GEP-NEN cohorts.

We trained the differential expression sanity-tests on the same genes as the deconvolution algorithm i.e. the genes whose expression differentiated the ductal cell-types from the remaining cell-types in the Baron scRNA dataset. Subsequently, a generalized linear model that predicted *Ki-67* counts was trained on differentially expressed genes and the Pearson Product-Moment correlation of predicted *Ki-67* counts and Ki-67 staining levels with the ground-truth quantified.

The survival curves were trained with R-package 'Survminer', version 0.4.8 [197]. The threshold for the subgroups were determined by averaging the aggregated gradings' cell-type proportions or Ki-67 levels, e.g. aggregated G1 and G2 values were summed up and divided by two to obtain the distinguishing threshold between the 'low' and 'medium' subgroups. The grading survival statistics were utilized 'as-is' and directly tested without any alteration. Ten Riemer et al. samples either did not possess survival information or were doublets derived from the same patient with identical survival time and were thus excluded.

5.2.3 Identification of the optimal training dataset and algorithm combination

We chose BSeq-sc, MuSiC and Moffit et al. due to their proven ability to deconvolve either healthy pancreatic tissue (BSeq-sc, MuSiC) or cancerous exocrine pancreatic tissue (Moffitt et al.) [34, 28, 116]. The extension of the liberally cited Cell-type identification by estimating relative subsets of RNA transcripts (CIBERSORT), BSeq-sc utilized a ν-SVR that was optimized for parsimonious modeling due to the ν parameter which is an upper bound on the training error and lower bound on the relative fraction of support vectors thereby reducing overfitting in contrast to the default C-SVR implementation [100]. MuSiC is based on a NMF and is unique in that no specified marker genes are required due to a dynamic gene weighting that prioritizes informative genes and suppresses information from non informative genes with a reported ability to discern between closely related tissues such as exocrine and endocrine. The Moffitt et al. NMF algorithm was conceived to deconvolve pancreatic ductal carcinomas and thus benchmarked on PanNEN and NEN samples. The Moffitt et al. algorithm was implemented as specified in the related publication due to absence of a publicly available implementation.

We identified the combination of training scRNA dataset and deconvolution algorithm whose predictions where most suited by analyzing comparing the stability and significance of the resulting correlations. In particular, the correlation with *Ki-67* levels averaged over all patient-derived Bulk RNA-seq data sets together with the algorithm specific quality score was taken as measurement of effectiveness. The Pearson product moment correlations of the relative proportions and the *Ki-67* levels were subsequently calculated to compare the performance to predict sample grading and patient survival, Figure 5.5.

P-value Monte Carlo sampling

We calculated an empirical p-value in order to consider a deconvolution successful i.e. to determine whether a deconvolution was sufficiently unlikely to be caused by chance. We determined the p-value via a Monte Carlo sampling over the correlation between the neoplastic transcriptome C and its reconstruction $B \times F$. We first determined the Pearson-Product Moment correlation R between the observed C and the conducted $B \times F$ calculation and thereafter determined the percentile within which R was located in the null distribution of correlations R^*. Since R^* is generally intractable, we empirically approximated R^* via resampling over the genes g drawn from M. We thereby generated resampled random mixtures C_i^*, each with a randomly selected set of genes, such that $|C_i^*| = |C|$ (same amount of genes) held. Subsequently, a deconvolution was applied to every resampled C_i^* and every correlation R_i^* between C_i^* and its reconstitution $B_i^* \times F_i^*$ calculated. The obtained R_i^* were ordered increasingly and determined how many percent of the randomly observed R_i^* showed a correlation as great or greater than R. This percentage quantified the p-value. We ensured that the sample-size was sufficiently great by setting the amount of resampling runs to 1E3 and required the p-value to be less than 0.05.

5.3 Results

Overview

We developed a framework for predicting the clinical characteristics of rare and diverse PanNEN. The framework explicitly addresses the ubiquitous lack of publicly available PanNEN datasets via an application of a transcriptomic deconvolution combined with a specific data augmentation strategy that substitutes neoplastic training data with data of healthy origin. The underlying hypothesis of the data augmentation is that a deconvolution's output is informative with respect to the clinical characteristics of a PanNEN and therefore renders the training of comprehensive ML models on widely available data feasible. The framework consists of multiple steps, illustrated in Figure 5.1.

First, deconvolution algorithms are trained to factorize healthy transcriptomes based on scRNA data of known cells with known type. In a second step, the framework deconvolves PanNEN transcriptomes in order to predict their respective cell-type proportions and determine a sample-specific reconstruction error. Third, the deconvolution output is utilized as training data for ML algorithms for the characterization of PanNENs and non-pancreatic GEP-NENs with respect to clinical properties. The characterization comprises the prediction of the neoplastic grading, subtype (NEC or NET), and survival time. Lastly, the performance of such trained ML models is compared to the performance of baseline models directly trained on neoplastic expression and Ki-67 biomarker data.

5.3.1 Deconvolution algorithms, cell-type signatures and evaluation datasets

The efficacy of deconvolution-based approaches critically depends on the effective combination of deconvolution algorithms and scRNA training datasets [35]. We evaluated three state-of-the-art deconvo-

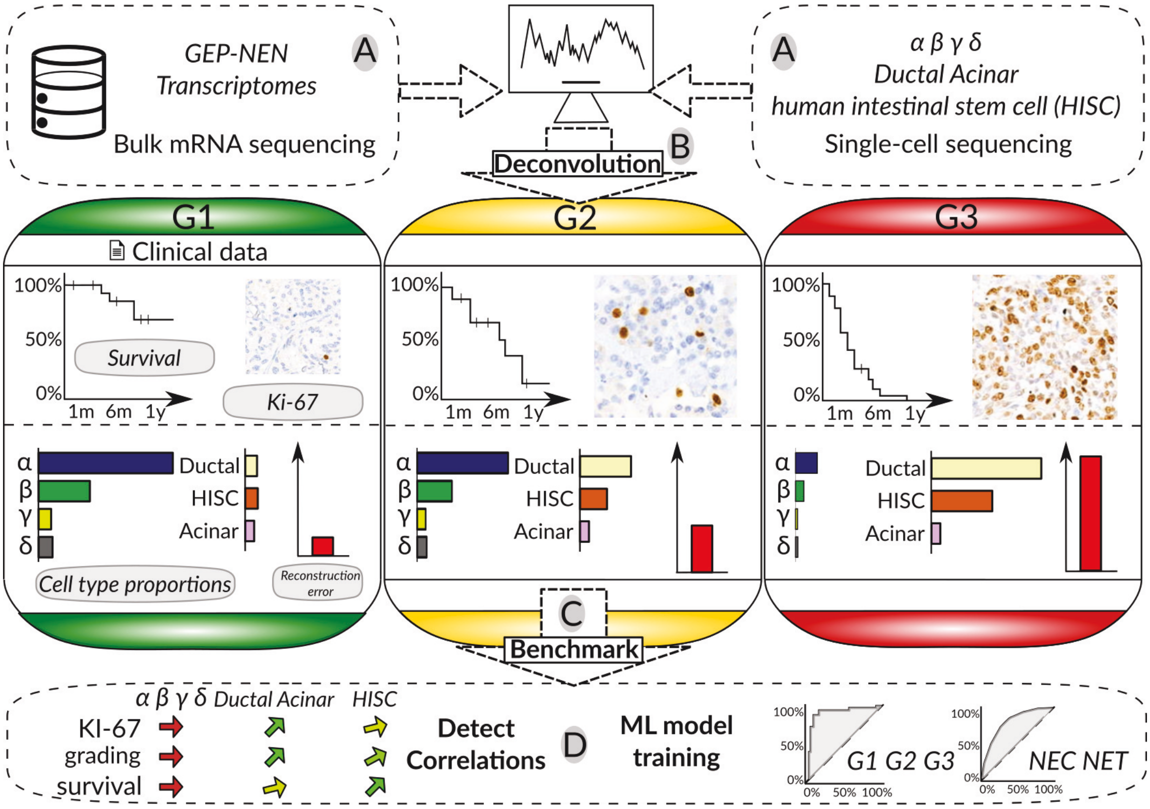

Figure 5.1: Overview of the developed framework that predicts clinical characteristics without training on neoplastic transcriptomic data. A: During a pre-processing phase, deconvolution algorithms are trained on exocrine (ductal & acinar), endocrine (α to δ) and HISC scRNA data from healthy donors. B: Deconvolution of NEN transcriptomes. Relative cell-type proportions are obtained and utilized as input of ML model that predicts grading, patient survival and carcinoma or tumor subtype. C: Direct comparison of the ML model's predictions with clinical ground truth allows to benchmark the predictive performance of the deconvolution-approach. Additionally, we calculate the Pearson Product-Moment correlations of the deconvolution-derived predictions with the establish Ki-67 gold-standard grading marker levels. Finally, the deconvolution-model's predictive performance is compared to a baseline model trained on the substituted data.

lution algorithms: CIBERSORT [34] in its BSeq-sc version, MuSiC [28] and NMF as applied by Moffitt et al. on pancreatic adenocarcinoma PDACs [116]. We furthermore identified three scRNA studies with a focus on single-cell sequencing of endocrine and exocrine cells. We refer to these datasets by the studies' first author: Baron [178], Segerstolpe [184] and Lawlor [181].

We considered three different cell-type models for the neoplastic deconvolution into cell-type proportions. The endocrine cell-type model comprises only endocrine cell-types (α to δ), since GEP-NENs are thought to originate from endocrine cell-types. The second model includes both endocrine and exocrine cell-types (ductal and acinar). The reasons for using such a model are: (i) adult pancreatic stem cells are proposed to reside in the ductal compartment, (ii) trans-differentiation of exocrine cells to endocrine cell-types occurs, at least in mouse models of pancreatic injury and regeneration, and (iii) poorly differentiated Pancreatic Neuroendocrine Carcinoma (PanNEC) frequently contain small areas with morphological differentiation of pancreatic adenocarcinoma and share a similar mutational pattern with PDACs [198]. The third model (endocrine & HISC) adds prototypic adult stem cells to the endocrine model, using transcriptome profiles of adult HISCs. The inclusion of stem cells follows from the work by Riester et al. [17] who demonstrated that patient survival time can be predicted accurately via the quantification of a distance between a sample and a holotypic stem cell. The HISC cell-type proportion was included as a measure to assess the distance of a given cell to a stem cell, but not to quantify intestinal stem cell-type proportions.

The three deconvolution algorithms were each trained on three cell-type models generated from the three different scRNA datasets. All resulting 27 combinations (3 (algorithm) x 3 (model) x 3 (scRNA dataset)) were benchmarked to identify the most effective setup, see Supplementary Table 7.5. We assessed the feasibility of a transcriptomic deconvolution of GEP-NENs by deconvolving patient-derived NENs from four PanNEN datasets and one mixed PanNEN and non-pancreatic GEP-NEN dataset (manuscript in preparation, raw data available at the EGA database, EGAD00001006657). Additionally, one non-neoplastic control-group dataset was deconvolved to obtain a non-neoplastic baseline and to quantify the impact of neoplastic transformation on the deconvolution performance. In the following, we refer to the neoplastic datasets by the name of their publication's corresponding authors, namely Califano [179], Missiaglia [182], Sadanandam [163], Scarpa [183], and Riemer (unpublished), see Table 5.1 and Supplementary Table 7.4.

Four neoplastic datasets exhibited a strong class imbalance with respect to grading and NEC versus

NET subtype, while the fifth did not provide grading information overall. The four datasets with grading information were skewed towards G1 & G2 NETs with the exception of the Riemer dataset that was skewed towards G3 NECs which has to be taken into account during the interpretation of the benchmark results, see Table 5.2. In order to derive benchmark statistics from a class-balanced dataset representative for all subtypes, we constructed an additional sixth dataset, called Representative Set (RepSet) by combining the Riemer and Scarpa datasets. We chose the Riemer and Scarpa datasets on the ground of their technological homogeneity, their complementary grading, and their favorable NEC and NET population characteristics. The RepSet composed of 69 GEP-NENs (52 PanNEN, 17 non-pancreatic GEP-NENs), sufficiently balanced with respect to grading (14 G1, 23 G2, 32 G3) and NEC versus NET subtype status (9 G3 Gastroenteropancreatic Neuroendocrine Carcinomas (GEP-NECs), 9 G3 Gastroenteropancreatic Neuroendocrine Tumors (GEP-NETs), 14 ambiguous G3 GEP-NENs). In the following, we will indicate that a cohort of samples exclusively consists of pancreatic NENs by referring to it as a PanNEN cohort, such as the G1 and G2 subcohort of the RepSet. In contrast, we will refer to a mixed cohort of pancreatic and non-pancreatic NENs as a GEP-NEN cohort, as is the case with the RepSet G3 high grade NENs.

Deconvolution of pancreatic and non-pancreatic GEP-NEN transcriptomes

We deconvolved four PanNEN, a mixed PanNEN and GEP-NEN datasets and the created RepSet for a total of 347 deconvolutions (69 G1, 59 G2, 44 G3, undetermined grading 105, 69 RepSet doublets). Nine combinations of deconvolution algorithms (three) and scRNA datasets (three) were analyzed with the aim to identify the most suited algorithm and training data combination. We also applied the method to 89 non-neoplastic samples from Fadista et al. to test the deconvolution of transcriptomes of healthy origin. We analyzed the resulting statistical significance and power of the deconvolution results to assess how well GEP-NENs can be deconvolved into healthy cell-type proportions, see Supplementary Tables 7.5, 7.6 and 7.7. Overall, we found that deconvolution of PanNEN and GEP-NEN, as measured by reconstruction p-values, differed greatly between different combinations of algorithm and training data. The greatest statistical power was obtained by the CIBERSORT algorithm trained on the Baron et al. scRNA dataset. All following results will therefore be based on this combination, if not stated otherwise.

We found that the grading had a significant impact on the reconstruction p-value and the ML model performance and therefore aggregated results by grading. We could effectively deconvolve non-neoplastic control, low-grade G1, and medium-grade G2 samples with any of the three deconvolution models. Deconvolution of G3 PanNEN was successful with a model exclusively consisting of endocrine cell-types, but inclusion of non-endocrine cell-types (models two and three, respectively) was required to deconvolve G3 GEP-NENs from non-pancreatic tissues, as illustrated in Figure 5.2 for the RepSet. This is noteworthy because GEP-NENs are supposed to originate from endocrine cells. A characteristic feature found to discern high-grade GEP-NENs from low to medium grade GEP-NENs, regardless of the tissue of origin, was that an endocrine model's p-value was consistently greater than the p-value of the related exocrine model for all G3 GEP-NENs. This finding was reproducible in all NEN datasets as shown in SM Figure 7.8 and proved important for the subsequent ML model training since it represents a characteristic classification pattern that differentiated high grade from medium and low grade samples.

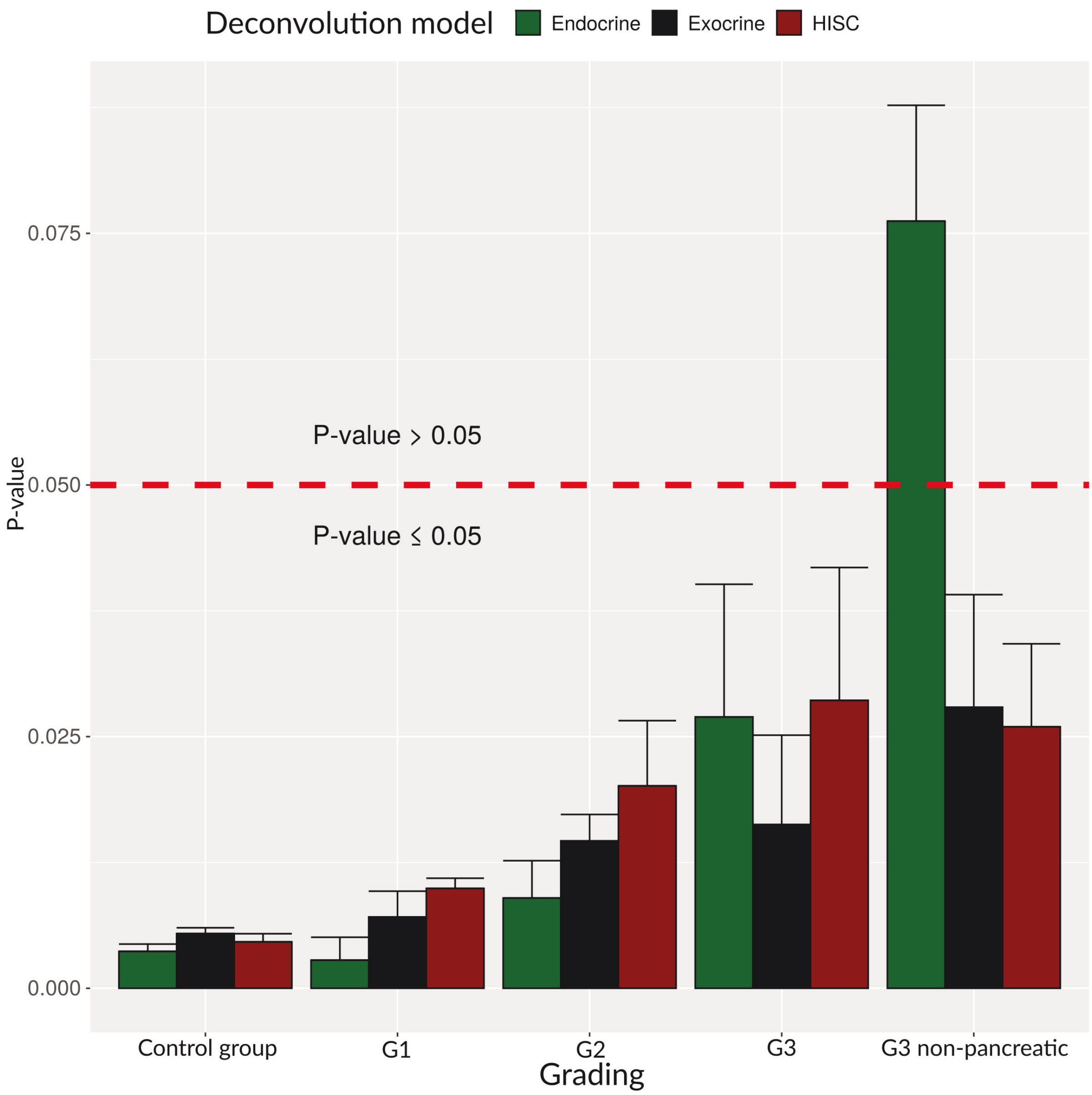

Figure 5.2: Statistical significance of the RepSet and control cohort dataset deconvolutions aggregated by grading. The figure shows empirical p-values obtained from deconvolving 89 non-neoplastic pancreatic neuroendocrine samples (Control group) and 51 RepSet PanNEN (G1-G3) and 18 high-grade non-pancreatic GEP-NEN (G3 non-pancreatic). Whiskers indicate the standard deviation of the p-values and deconvolutions with a p-value of less than 0.05 are considered successful. The transcriptomic deconvolution of PanNENs of all gradings proved possible for a model exclusively trained on healthy endocrine cells as indicated by an aggregated p-value of less than 0.05 but generally failed for high-grade non-pancreatic GEP-NENs. Importantly, a deconvolution model trained on exocrine cell-types (ductal and acinar) succeeded in deconvolving all types of GEP-NENs invariant to their grading despite showing inferior performance when deconvolving low and medium grade PanNEN.

We analyzed the predicted cell-type proportions of all datasets and found the RepSet to be representative for the general behavior of cell-type predictions, see Figure 5.3 for the RepSet predictions and Supplementary Figure 7.9 and Table 7.8 for the predictions of the remaining datasets. We found the cell-type proportion predictions of the endocrine-only model for G1 NETs to be most similar to the healthy cell-type stratification, approximately resembled that of the healthy pancreatic endocrine tissue [181]. A correlation between cell-type proportions and grading was visible in that the endocrine-only model showed an increase of the alpha cell-type proportion and a decrease in gamma cell-type proportion with increased grading. The alpha cell-type proportion was significantly positively correlated with the grading (r= 0.87) and great statistical power (p-value correlation $\leq$2E-16) even when the non-significant G3 deconvolution models with greatest alpha cell-type proportion were excluded.

Also, cell-type proportions of the endocrine & exocrine model were correlated with the grading. Similar to the endocrine model, G1 PanNENs were composed of cell-type proportions approximately equal to the healthy islet stratification, albeit that 8% of the G1 PanNEN cells were classified as ductal cells. Like the endocrine-only model, the mixed model predicted an increase in alpha cell-type proportions between G1 and G2 PanNEN. However, the increase of the predicted ductal cell-type proportion was greater than that of the alpha cell-type when comparing the G1 to G2 predictions. The ductal cell-type proportion was the highest for G3 GEP-NENs, while alpha cell-type proportions decreased compared to G2 PanNENs. Exocrine acinar cells did not contribute notably to the deconvolution of low or high-grade PanNENs. HISC model predictions were generally comparable to the combined endocrine and exocrine model. However, the increase of the HISC cell-type proportion with grading was less distinct than that of the ductal proportion of the mixed model or the alpha proportion of the endocrine-only model. In addition to analyzing the mixed RepSet, we compared the cell-type proportions for the subset of the RepSet that either exclusively consisted of PanNENs or exclusively of non-pancreatic GEP-NEN. The resulting cell-type proportions were comparable, despite a generally less well stratified cell-type proportion for medium grade non-pancreatic GEP-NEN compared to the exclusively pancreatic subset.

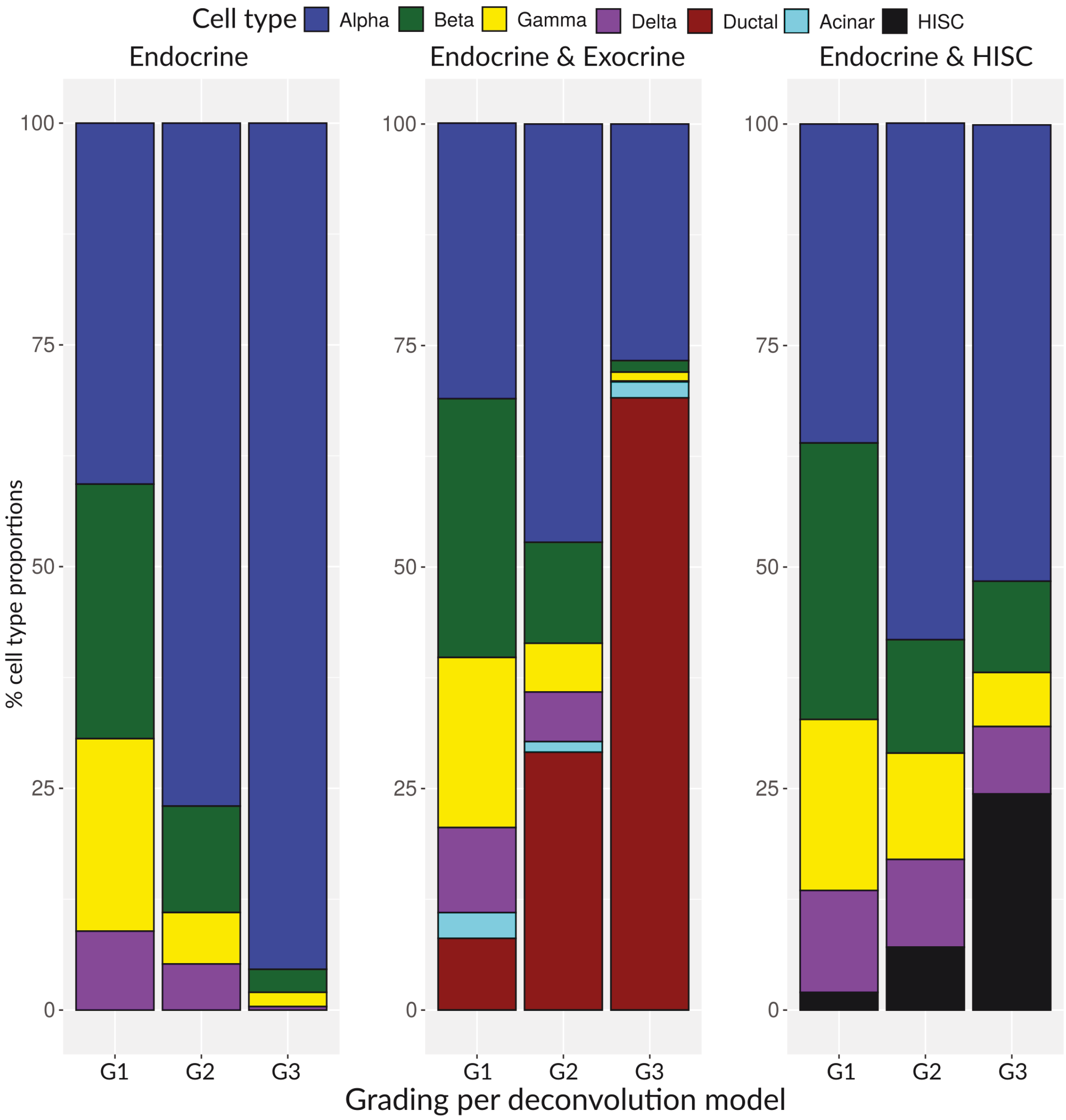

Figure 5.3: Cell-type proportions predicted by deconvolution of RepSet grouped by grading. The grading was correlated with the predicted cell-type proportions which favored a single dominant cell-type in high grade GEP-NEN while all of the other proportion predictions were reduced. Deconvolution with a model exclusively trained on endocrine cell-types (left) revealed a comparatively complex and balanced distribution of cell-type proportions which approximately resembles that of the healthy endocrine pancreas [181]. With increased grading, however, the alpha cell-type dominated the predicted cell-type proportions. Panels two (exocrine) and three (HISC) illustrate that a cell-type proportion to grading correlation remained when the deconvolution model additionally included exocrine and intestinal stem cell-types, respectively. The ductal and stem cell-types, however, replaced the alpha cell-type as the dominant cell-type for high-grade GEP-NENs, which was complemented by the observation that models which included a ductal or HISC signature generally showed superior p-values for high-grade GEP-NENs compared to exclusively endocrine models.

5.3.2 Correlation of predictedcell-type proportions with *Ki-67* count levels

We commenced the analyses by visually discovering a common clustering pattern of the cell-type proportions with Ki-67 biomarker levels in the RepSet, shown in Supplementary Figures 7.7 and 7.10, and subsequently quantified the extent to which Ki-67 levels were correlated with the ductal and HISC cell-type proportions in all datasets. In the following, italic *Ki-67* will refer to the gene's mRNA and non-italic Ki-67 to the protein). Although the alpha cell-type proportion of the endocrine-only model was correlated with the grading as shown in Figure 5.3, we discarded the alpha cell-type signature from the analysis since the p-values of the high-grade G3s in the endocrine only deconvolution model were not significant (Figure 5.2).

We calculated the Pearson Product-Moment correlations between the *Ki-67* count levels and the predicted cell-type proportions and found a significant correlation for the ductal cell-type proportion in five out of six datasets, with p-values ranging from 6E-6 (RepSet) to 0.007 (Riemer et al.), see Supplementary Tables 7.9. The insignificantly correlated dataset, Missiaglia (p-value 0.14), was found to be strongly biased for low-grade PanNENs, with low *Ki-67* levels preventing a *Ki-67*-based differentiation of G1 and G2 PanNENs. The HISC cell-type proportion was significantly correlated in four out of six datasets, p-values ranging from 2E-5 (Scarpa et al.) to 4E-2 (Riemer et al.). Overall, the HISC proportion proved to be less correlated with *Ki-67* than the ductal cell-type proportion.

5.3.3 Correlation of predicted cell-type proportions with the grading

Thereafter, we determined the correlation between the predicted cell-type proportions and the Histopathology derived grading information available for all datasets (except for the Califano dataset). The ductal cell-type proportions were significantly correlated in five out of five datasets, with p-values ranging from 1E-10 (RepSet) to 4E-2 (Riemer et al.), shown in Supplementary Table 7.9. The statistical power of the ductal cell-type proportions to the annotated grading was thus greater than the correlation with the grading-indicating *Ki-67* biomarker levels but generally coherent with the ductal cell-type proportions to *Ki-67* count level correlation. The HISC cell-type proportion proved less correlated with the grading than with Ki-67 count levels, being significantly correlated in only two out of five datasets with p-values ranging from 3E-7 (RepSet) up to 2E-5 (Scarpa et al.).

Further ANalysis Of VAriance test (ANOVA) of the distribution of cell-type proportions between the gradings revealed that the ductal proportion predictions could effectively separate G3 from G2 GEP-NENs in four out of five datasets and G1 from G2 in two out of four datasets, see Supplementary Table 7.5. The HISC proportion predictions allowed to discern G3 from G2 GEP-NENs in two out of five datasets and G1 from G2 in two out of four datasets, the HISC proportions therefore being less suited for a statistical differentiation between gradings.

5.3.4 Machine-Learning-based prediction of the grading

Due to the significant correlations between the grading and *Ki-67* levels and cell-type proportions, respectively, we decided to train a ML model on healthy pancreatic cell-type proportions to predict the grading of PanNEN and non-pancreatic GEP-NENs. The first model was called 'Deconvolution' and was trained on the deconvolution results and thus not informed about the grading-indicative *Ki-67* levels. The second model 'Expression & *Ki-67*', served as comparative baseline model and was trained on the union of all cell-type marker genes and additionally on the grading-informative *Ki-67* gene. We trained a Random Forest for either model and compared the observed predictive performances with respect to multiple key performance characteristics for all datasets, shown in Supplementary Tables 7.10 and 7.11, while key comparative predictive performances as observed for the RepSet are illustrated in Figure 5.4.

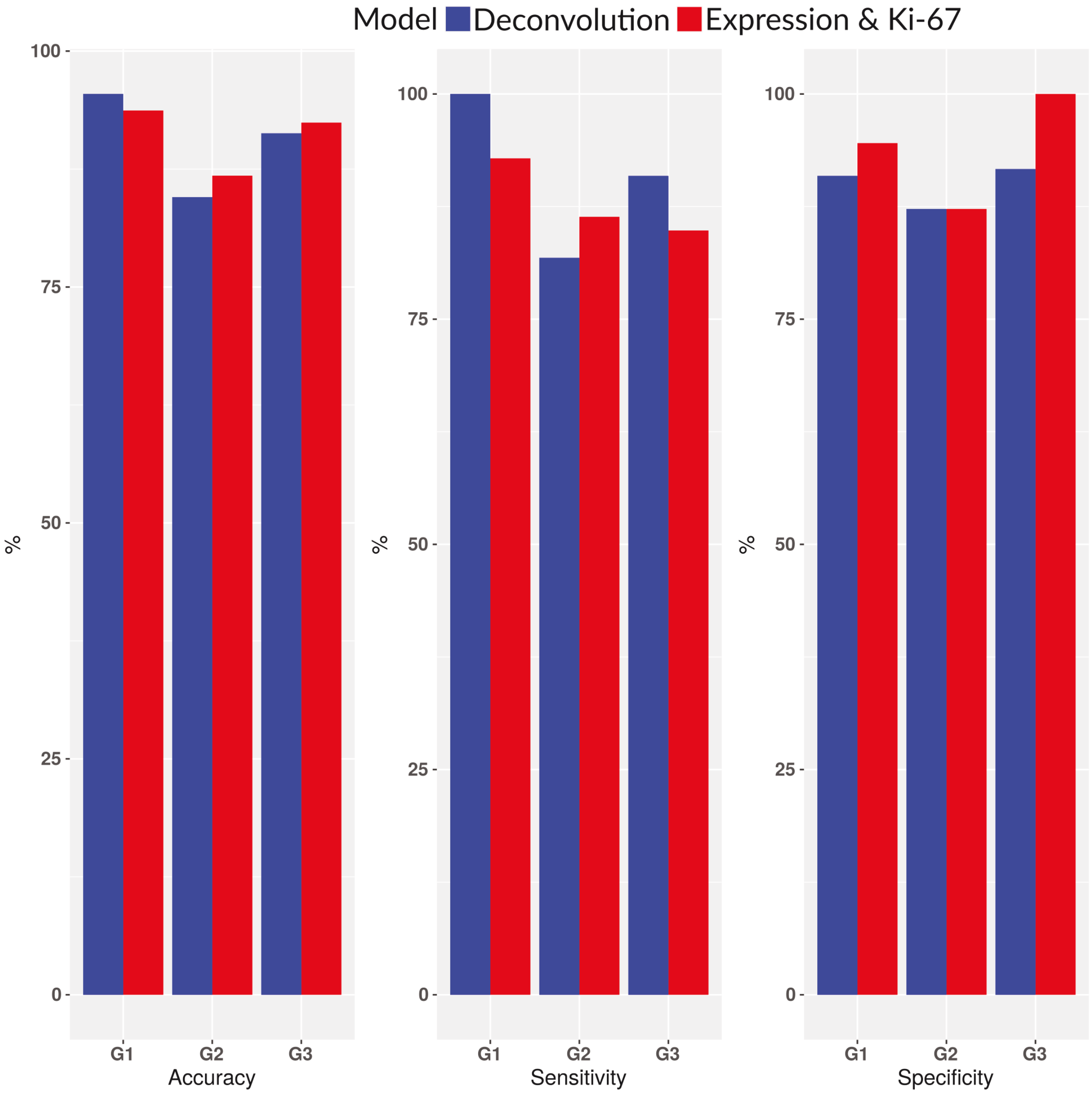

Figure 5.4: Comparative performance characteristics of grading-predicting ML models either trained on expression data or the deconvolution output for the RepSet dataset. The expression based ML model slightly outperforms the deconvolution-based model with respect to accuracy and specificity, however, the general predictive performances of either training method remain comparable. The major source of varying predictive performance was the grading of a GEP-NEN and not the training method. Percentages were averaged over 10-fold cross validations.

The predictive performance of the ML models differed notably between lowly and highly graded samples in a majority of datasets and thus we disaggregated the analyses by grading to visualize the observed differences. For the RepSet, the 'Deconvolution' model achieved a sensitivity (G1 100%, G2 82%, G3 91%) that was comparable to that of the 'Expression & Ki-67' model (G1 93%, G2 86%, G3 85%), similarly to the accuracy characteristic of the 'Deconvolution' (G1 95%, G2 85%, G3 91%)

and ’Expression & Ki-67’ (G1 94%, G2 87%, G3 92%) model. The predictive specificity showed the greatest differences between the ’Deconvolution’ (G1 91%, G2 87%, G3 92%) and ’Expression & Ki-67’ (G1 95%, G2 87%, G3 100%) model in that the expression-trained model performed 8% better for G3 GEP-NENs than the deconvolution-trained model.

Analysis of the ’Deconvolution’ model revealed that both reconstruction error of a transcriptome, as measured by the RMSE, and the ductal cell-type proportion were the most important features, thereby supporting assumptions *i*) and *ii*) specified in section 5.3.1. Important features of the ’Expression & Ki-67’ model were *Ki-67* and proliferation and cell-cycle regulating genes.

5.3.5 Classification of neoplasms as NEC or NET

We furthermore explored the relationship between deconvolution-derived ductal and HISC cell-type marker genes and the NEC versus NET subtype and found that reducing the transcriptome to either ductal or HISC marker genes rendered NETs and NECs linearly separable on a plot of a Principal Component Analysis (PCA), shown in Supplementary Figure 7.10 and 7.11. This finding was of interest because it suggested that a ML model could classify the subtype effectively based on these sets of genes.

We therefore benchmarked whether a logistic regression model trained on cell-type fractions and reconstruction RMSE could differentiate between NECs and NETs and compared its performance to that of a model trained on *Ki-67* levels. We limited the benchmark to the RepSet since it was the only dataset with a balanced NEC and NET population. Note that pathologists’ NEC and NET annotations were partially ambiguous and that samples with uncertain morphology could be assigned to either the NEC and NET cluster in the PCA on the basis of the ductal or HISC signature. A linear separation of NET and NEC subtypes, respectively, resulted from supervised clustering using the panNETassigner signature [163], see Supplementary Figures 7.10 and 7.11, which led us to manually assign a NET (one sample) or NEC (13 samples) subtype to ambiguous samples.

A sensitivity of 84% was achieved by both the combined endocrine & exocrine and HISC models in the RepSet. Specificity amounted to 91% for the combined model and 85% for the HISC model. In comparison, the model trained on *Ki-67* levels obtained a sensitivity of 84% and a specificity of 88%. The predictive performance remained comparable when the ambiguous cases were excluded, see Supplementary Table 7.12.

5.3.6 Prediction of the overall patient survival time

Information on disease-related survival was available for three datasets: Riemer et al., Scarpa et al. and their combination, the RepSet. Analyses revealed a statistically significant Pearson product-moment correlation (r=-0.45, p-value 0.017), between the cell-type proportion predictions of the 32 high-grade GEP-NENs of the Riemer and RepSet datasets and their corresponding patient survival times. We therefore applied Cox-Hazard ratio tests to quantify the extent to which the ductal and HISC cell-type proportion predictions informed about the disease-related patient survival time relative to the grading and *Ki-67* baselines.

We utilized two different cohort designs for the tests, the first design required three subgroups ('low', 'medium' and 'high' risk subgroup) while the second cohort design tested on two subgroups ('low' and 'medium' combined versus 'high' risk subgroup). The three arm design was chosen to reflect that a three-arm design is the established clinical standard and the two arm design was tested because the grading ANOVA tests indicated that G3 GEP-NENs could be well discerned from G2 GEP-NENs but not G2 GEP-NENs from G1 GEP-NENs. Three arm design testing was limited to the Scarpa et al. and RepSet because the Riemer et al. dataset only consisted of G2 and G3 GEP-NENs.

The two-arm design Cox tests revealed that the ductal signature achieved significance for all three datasets (range p-values 2.1E-3 to 4.5E-2) and the HISC signature in only two datasets (range significant p-values 2.2E-4 to 4E-2). The corresponding *Ki-67* (range p-values 3.5E-3 to 1.4E-2) and grading (range p-values 5.4E-4 to 3.6E-2) baselines were always significant, see Supplementary Table 7.5. The three-arm design resulted in Cox test p-values with slightly less statistical power for both the ductal signature (range p-values 8.6E-3 to 2.0E-2) as well as the HISC (significant p-value of 8E-3). The ductal signature tested significantly for both datasets and the HISC signature for the Scarpa et al. dataset. The *Ki-67* (range p-values 2.6E-3 to 5E-2) and grading (range p-values 2.3E-3 to 8.7E-3) baselines' statistical power remained comparable to the two-arm design and were significant for both the Scarpa et al. and RepSet.

A comparison of the baseline and cell-type proportion prediction Cox-tests results for the two-arm RepSet cohort design revealed that the ductal signature's statistical performance was superior (p-value 6.6E-3) to the *Ki-67* (p-value 1.4E-2) baseline but inferior to the grading gold-standard (p-value 2.0E-3), see Supplementary Table 7.5. Averaged results over all three datasets differed, however, in that the ductal signature (p-value 1.8E-2) was slightly less statistically powerful than both the *Ki-67* (p-value 1.1E-2) and averaged grading baseline p-values (p-value 1.3E-2). Comparison of the three-arm design provided comparable insights, including the finding that the HISC signature's p-values always remained less statistically powerful than any baseline and the ductal signature-derived Cox-test p-value. We thus concluded that the predictive performance with which ductal cell-type proportions informed about patient survival was generally comparable to that of the *Ki-67* and grading baselines based on the RepSet results despite a slightly superior distribution of the dataset-averaged p-values.

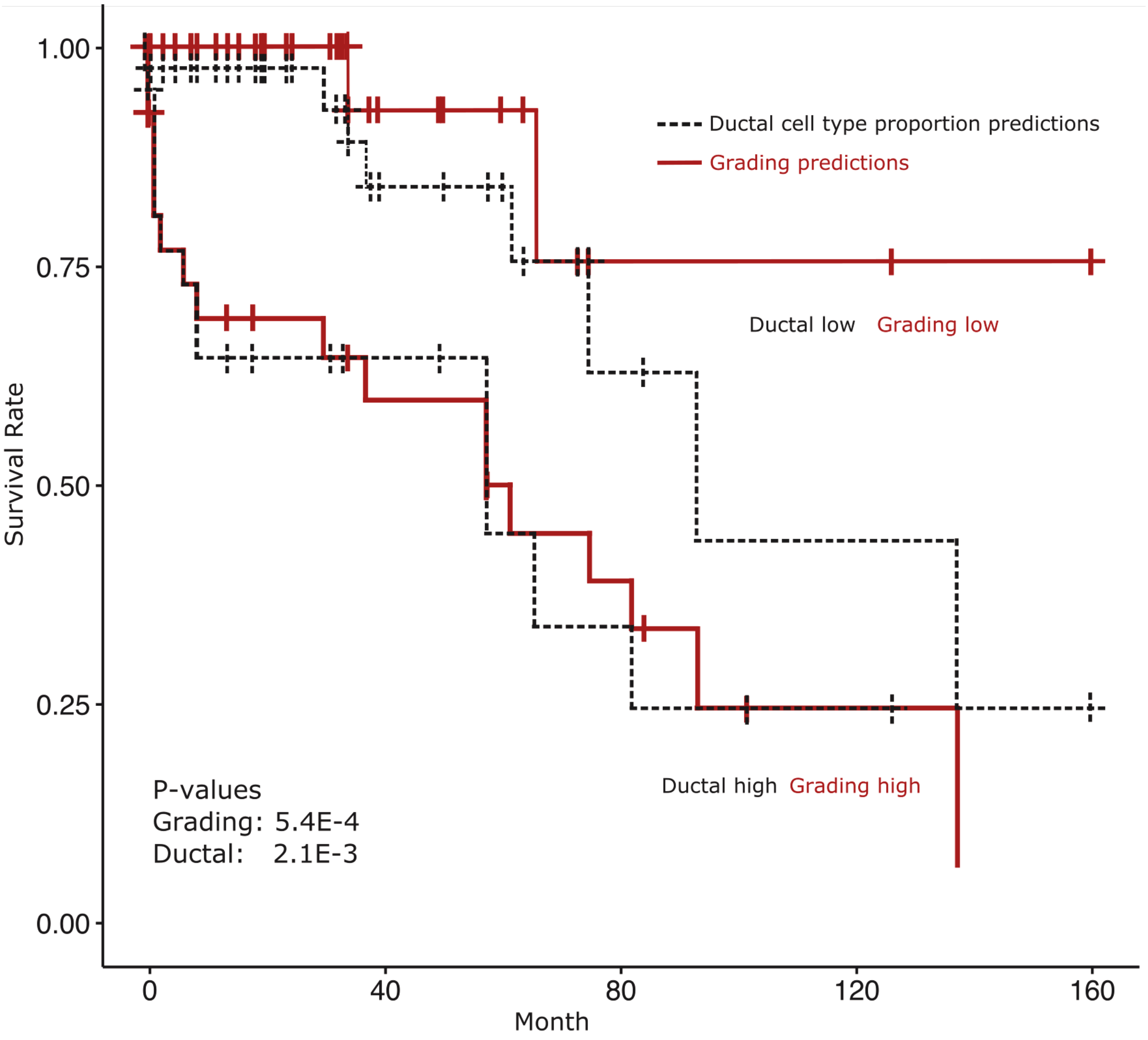

Figure 5.5: Kaplan-Meier plot comparing the predictive power of the ductal cell-type proportions and the clinical grading for the RepSet with respect to the disease-related patient survival time. The RepSet was split into two subgroups either based on the grading (G1 and G2 combined) or the ductal cell-type proportion predictions. Tests on a difference of the subgroups' disease-related patient survival time were significant in either case with the grading-based subgroup showing greater statistical power. We found that the deconvolution-derived ductal cell-type proportion predictions were informative with respect to the disease-related patient survival time and the statistical power of the corresponding Cox hazard ratio-test comparable but inferior to that of the pathologist-derived grading.

5.3.7 Out-group tests, immune cell infiltration and methodological comparison

We performed three sanity tests to verify the correlation of *Ki-67* levels with predicted cell-type fractions. Firstly, we tested whether the effect is cancer-specific i.e. absent in healthy tissue. Secondly, we compared results to a simpler baseline. Thirdly, we determined whether the correlation could be caused due to confounding contamination with exocrine tissue, immune-cells or other tissue. Regarding the first control, deconvolution of the healthy control dataset (Fadista et al.), resulted in an insignificant correlation of *Ki-67* levels, ductal and HISC fractions. Next, we determined whether a differential expression analysis in conjunction with a generalized linear model could effectively predict *Ki-67* mRNA expression levels. A logistic regression model could either predict the *Ki-67* levels with comparable (2x) or worse (3x) statistical power. Given that a logistic model was superior in only one dataset (Missiaglia), the differential expression-based model was discarded on the grounds of inferior predictive power (statistics shown in Supplementary Table 7.13). Third, ESTIMATE analysis [185] of the RepSet revealed no correlation of immune scores with annotated clinical parameters or deconvolution outcomes, data not shown.

5.4 Discussion

5.4.1 Feasibility of deconvolving GEP-NEN transcriptomes

We developed and benchmarked an *in-silico* framework that allows the application of ML models for rare cancers despite low sample numbers. The framework applies transcriptomic deconvolution trained on data of samples from healthy origin to study tumor samples. We consequently determined and compared the performance of ML models trained on the deconvolution output with respect to predicting clinically relevant neoplastic characteristics. Three different deconvolution models were generated and their ability to factorize neoplastic transcriptomes into relative cell-type proportions and a reconstruction error assessed. A model consisting exclusively of endocrine cell-types could effectively deconvolve pancreatic GEP-NENs with low and medium grading. Interestingly, this 'endocrine-only' model performed poorly for high-grade G3 PanNEN and failed to achieve significant results for G3 NENs from non-pancreatic sites, despite the fact that these G3 NENs share the defining feature of neuroendocrine marker expression. Remarkably, a deconvolution with addition of exocrine cell-type information led to improved results for G3 samples with significant deconvolution of PanNENs as well as GEP-NENs of non-pancreatic origin.

Analysis of predicted cell-type proportions G1, and to a lesser extent G2, pancreatic neoplasms showed cell-type proportion distributions that resembled the cell-type stratification in the healthy pancreatic endocrine tissue when the endocrine-only model was applied. Samples with known clinical annotation of insulinoma deconvolved primarily into beta-cells, whereas one tumor with known production of pancreatic polypeptide deconvolved into gamma-cells, which physiologically produce this hormone. Thus, in-silico lineage allocation and clinical reality converged.

All three benchmarked deconvolution models showed the same proclivity to deconvolve low grade neoplasm into multiple cell-type proportions and high grade neoplasms into a single dominant cell-type proportion whose cell-type, however, differed between the models (alpha or ductal or HISC). In the endocrine-only model, the alpha-cell type proportions increased with grading. The prediction of higher alpha cell-type proportions from G2 and G3 PanNEN transcriptomes is consistent with recent observations from studies using entirely different methodology. For instance, similarity to alpha or beta-cells, respectively, was proposed as a basis for stratification of sporadic Pancreatic Neuroendocrine Tumor (PanNET) [156, 157, 161, 159], with expression of the alpha-cell specific transcription factor Aristaless Related Homeobox (ARX) in more advanced stage PanNET. Conversely, beta-cell like tumors exhibited a more favorable clinical course [160].

The model predictions of high ductal cell proportions in high-grade PanNENs implies similarities with the expression features of fully differentiated non-transformed ductal cells. Interestingly, pancreatic ductal cells were assigned a central position in lineage trees derived *in-silico* from single-cell sequencing of adult pancreatic cells with ductal cell sub-populations giving rise to different endocrine cell-types [42]. Thus, the ductal features may reflect such an endocrine progenitor population of ductal cells in the adult pancreas. Alternatively, this may point to a 'reserve' multi-potency of adult ductal cells, capable of generating endocrine cells under specific stimuli and pressure. Indeed, exocrine pancreatic cells were found capable of reprogramming to endocrine cells in a process that reactivated embryonic multi-potency markers [199]. Conversely, an acquisition of exocrine features by islet cells occurred in a rat model of mild islet cell injury, indicating that ductal characteristics are within the range of lineage plasticity of pancreatic endocrine cells [199, 200]. Hence, the non-endocrine cell fates predicted for high-grade GEP-NENs have precedence in de-differentiation or trans-differentiation processes in rodent models. Given that admixtures of ductal adenocarcinoma are frequently present in PanNECs, ductal features could moreover reflect similarity to adeno-carcinomas. Moreover, separating RepSet NENs based on either the ductal marker genes or the previously published panNETassigner signature [163] resulted in almost identical clusters of neoplasms, further supporting that ductal cell type predictions relate to biologically relevant features of NENs. Recent findings state that NECs and NETs possess different developmental trajectories on the ground of differential hypo and hyper-methylation patterns in genes associated with cellular development [201]. NETs are reported to derive from an endocrine developmental lineage, however, the developmental lineage of NECs is reported to be closer related to the exocrine cell-type of acinar cells. Should the differential development hypothesis be further supported, then the classification of NETs as being alike to an endocrine cell-type and NECs as similar to an exocrine cell-type would further strengthen the adeno-carcinoma-similarity hypothesis which would show as prediction of NECs as an exocrine cell-type. However, the differential developmental trajectory hypothesis remains subject to further research.

The observed correlation between cell-type proportions and Ki-67 levels is a notable finding, because staining levels of Ki-67 represent the current gold-standard method for GEP-NEN grading. Moreover, the ductal and HISC cell-type proportions were either comparably well or even more strongly correlated with the ground-truth grading than Ki-67 levels. This correlation between the grading and the model-specific dominant cell-type allowed for an effective grading prediction by deconvolution-trained ML models. Their statistical power was comparable to that of a baseline model trained on expression data and Ki-67 levels. These results underline the efficacy of our deconvolution approach for characterizing PanNENs, and possibly NENs in general. The deconvolution-trained ML model that predicted clinical characteristics such as the grading did not include either Ki-67 levels or any other proliferation marker as features. Therefore, the deconvolution-trained prediction of characteristics can serve as a complementary approach to the established proliferation based methods.

While our study focused on PanNEN, the rarity of resection tissues from PanNECs led us to include high-grade GEP-NENs which originated from other locations, e.g. stomach and colorectum. Therefore, the tissue of origin could affect the deconvolution results. However, the overall performance of the *in-silico* classification model remained stable between datasets that contained GEP-NENs of purely pancreatic origin and those that included high-grade samples of diverse gastrointestinal tissue-backgrounds. Nonetheless, we ascertained that the ductal and HISC models were not biased for either pancreatic or non-pancreatic tissue by verifying that the ductal and HISC marker genes predominantly consisted of genes with similar expression in PanNENs and non-pancreatic GEP-NENs. Models trained on the ductal and HISC cell-type proportions, respectively, were able to predict the survival of GEP-NEN patients in a two-arm and in a three-arm cohort design. Partitioning the RepSet into a ductal-high and a ductal-low sub-cohort revealed a significant difference in disease-related survival time, comparable to that of partitioning the cohort by Ki-67 expression and only slightly inferior to a partitioning based on the pathologists-derived grading ground-truth. However, we observed a high degree of variation of the survival-time test's p-value between different deconvolution algorithms and different training sets, indicating a need for further fine-tuning of the method for this purpose. The prediction of clinical characteristics via a similarity measurement to stem cells previously was demonstrated [17] and could be replicated by the framework, with the difference that deconvolved HISC proportions served as quantification of similarity. However, the ductal cell-type proportions were generally found to be more suited for the prediction of the overall patient survival time than HISC cell-type proportions.

5.5 Conclusion

Our results show that the combination of transcriptomic deconvolution and ML methods for the study of PanNENs can lead to clinically meaningful results. Our proposed strategy reduces the dependency on scarcely available neoplastic training data for PanNEN and GEP-NEN research in general. Therefore, classification-by-deconvolution has the potential to support pathologists in cases of an incongruous or uncertain grading and morphological differentiation, which in turn may lead to a better personalization of the clinical management of GEP-NENs.

Future research is required to render the classification-by-deconvolution method more robust. Additionally, a gold-standard dataset is required to validate the results of a deconvolution with respect to predicted cell-type proportions. Eventually, the classification-by-deconvolution method should also be tested in other rare cancer types.

Evaluation of the distance-quantification aspect

We conclude that the distance-quantification based on transcriptomic deconvolution between neoplasms and prototypic cell-types is possible since the benchmark results indicated a successful prediction of clinical characteristics. Importantly, the distance-quantification to non-neoplastic prototypic cells allowed for the substitution of neoplastic training data. The advantages of a distance-quantification are therefore that distance-quantification can increase the amount of available training data and can allow the prediction of neoplastic properties.

The disadvantage of deconvolution-based distance quantification approach is its volatility in that different combinations of algorithms and training data sets resulted partially in highly different results. The volatility is a major obstacle because no gold-standard is currently available to identify the ground truth-prediction and the identification of the best algorithm and data combination occurred via comparison to pathologists' findings. Therefore, the distance-quantification shows promising perspectives for the field of neoplastic prediction but requires more research and a ground-truth gold-standard before a generalized and robust approach of this kind can be developed.

Chapter 6

Summary and conclusion of the thesis

Verification of the correct identity of CCLs is of great importance for multiple Life-Science domains. However, a generalized identification method that determines the identity of CCLs based on heterogeneously generated NGS data has not been established thus far. We therefore developed the Uniquorn method which serves as scientific contribution to this thesis and utilizes the concept of distance-quantification to identify CCLs. The distance-quantification concept, as shown in Chapter 3, can successfully identify CCLs based on their rare small variants but has limitations with respect to the technological diversity.

The Uniquorn method's limitation to the WES technology motivated the method's generalization in order to significantly extend the range of use-case scenarios. The second scientific contribution therefore added the support of the Bulk RNA-seq technology and Panel-sequencing format. This extension, however, incurred Data-Incompleteness and Data-Heterogeneity which caused the distance-quantification approach to fail because distance-quantifications reflected the technological similarity of the data and not the biological similarity of the sequenced CCLs. We therefore conclude that the naive 'identification-by-distance-quantification' approach presented in Chapter 3 can identify homogeneously sequenced CCLs but requires modifications to generalize to the identification of technologically diverse NGS data. We furthermore, uncovered the lower bound with respect to genes where distance-quantification fails to identify CCL profiles and approximated it as $_\sim$1E3 genes.

The modified and generalized Uniquorn method could, however, recover the identification by distance-quantification concept via the introduction of statistical resampling methods. The method first quantified the degree of Data-Heterogeneity and Data-Incompleteness to secondly adjust the identification thresholds such that the biases were compensated for. Benchmarks revealed that the such modified Uniquorn method remained robust to a high degree of technological diversity. However, identification sensitivity was lost compared to the unmodified Uniquorn WES method and therefore a trade-off between sensitivity and methodological robustness exists. We therefore conclude that the Uniquorn identification method can serve as generic CCL NGS identification method if a loss of predictive power over its WES version is accepted.

The third scientific contribution addressed the problem that rare and diverse cancer types frequently lack sufficient sample-sizes for a comprehensive ML model training. A transcriptomic deconvolution approach was therefore applied to support the ML-based classification of rare and diverse cancer types. The deconvolution aimed to augment the neoplastic training data by substituting it with data of healthy origin which is widely available. The choice of the deconvolution algorithm and training dataset strongly affected the performance of the deconvolution. A careful identification process did, however, result in the identification of a suitable algorithm (BSeq-sc) and scRNA training dataset (Baron et al.) combination. ML models were subsequently trained on deconvolution-derived relative cell-type proportion predictions and reconstruction errors to predict clinical characteristics of neoplasms. Such trained ML models could thereafter efficiently predict the grading, overall patient survival time and the carcinoma versus tumor subtype status during benchmarks which entailed five different NEN studies. The predictive performance was comparable to that of the Ki-67 biomarker and only slightly inferior to that of the pathologists'-derived grading.

We therefore consider the distance-quantification based approach of transcriptomic deconvolution to augment training data of rare and diverse cancer types possible. Transcriptomic deconvolution does, however, remain an approach in development, in particular, with respect to the deconvolution of neoplasms. Primarily, the robustness of the deconvolution still has to improve before the need for bespoke adaption of algorithms and datasets by a user to a given type of cancer can be relinquished.

Critical assessment of the contributions' development process

Insights have been gained throughout the development of the scientific contributions and consequently aspects revealed which would have improved the development of the methods.

The 'classification-by-deconvolution' approach has been benchmarked on a single type of cancer. This represents a major disadvantage since a cell-type independent generalization of any methods is highly advantageous. Due to the novel nature of the method, additional proof of concept analyses on different types of cancer would have strongly increased our understanding of the methods capabilities and limitations since cancer type-specific confounding factors would likely have been obviated when analyzing different types of cancer.

The decision not to include additional cancer types was taken on the ground of the advantages of limiting the benchmark to one type of cancer. The primary advantage was that the biological interpretation of the deconvolution results was dramatically simplified since only a single type of cancer had to be analyzed and interpreted. Additionally, data availability and biological expertise constituted limiting factors. Suitable scRNA data and Bulk RNA-sequenced biopsies would have to been identified what represents a significant increase in required time and increases the developmental complexity. Inclusion of additional cancer types would have as well required the inclusion of additional domain experts. The inclusion would have been necessary on the ground that the experts' evaluations would have been critical since deconvolution gold-standard datasets with ground-truth annotations are generally not available due to the rare and diverse nature of suitable cancer types.

A possible retrospective improvement of the CCL-identification contributions has been identified with respect to the reference dataset-generation. Analyses of the benchmark results revealed that query i.e. to-be-identified CCLs with a low amount of covered genes cannot be identified as effectively as CCLs with a large amount of covered genes. However, it has not been sufficiently analyzed to what extent the reference CCLs contribute towards an efficient identification of CCLs with a low amount of covered genes. The background is that analyses have revealed the existence of CCLs in given reference datasets whose variant profile was highly similar while simultaneously possessing likely causally connected below-average identification performances. An analysis of the reference dataset compositions would therefore have been indicated since minor adjustments to the variant selection process of similar CCL profiles could have improved the identification performance in particular with respect to panels covering a limited amount of genes. Moreover, a user is currently left to their own discretion when creating their reference datasets what is likely to decrease the identification performance when highly similar reference profiles are added.

Nonetheless, all contributions can potentially be adapted to incorporate the aforementioned improvement during the course of future research activities. An augmented automatic composition of reference database based on the reference CCLs' similarity can added and deconvolution benchmarks be conducted on different types of cancer.

Advantages of the distance-quantification concept

The thesis is based on the distance-quantification concept and application of this concept has proven feasible as illustrated by the contributions. Three main advantages of the abstract distance-quantification concept were identified. The primary advantage was the mathematically well-defined nature of the distance-metric concept. Mathematical conditions could be identified that clarified when a metric on a space is able to quantify meaningful distances. The second advantage was that the interpretability of complex scenarios was improved. Given a benign scenario where the space is 'flat' and all metric conditions hold, the assignment of positions in a metric space to CCLs allows for an intuitive visualization of relative distances of CCLs as illustrated in Figure 3.2. Given the alternative scenario of a 'curved' (positions are not indicative of distances) space caused by confounding factors such as Data-Incompleteness, illustrations of the violation of mathematical conditions such as the triangle-inequality facilitated the method development. The perspective of interpreting neoplasms as entities whose distance to a prototypic stem cell motivated the development of the classification-by-deconvolution contribution. The motivation to train ML models on cell-type proportions was based on the report by Riester et al. [17] that an abstract distance of a neoplasms to a stem-cell was clinically informative. The third advantage entails the possibility to analyze quantified distances as ensemble to approximate latent and intractable parameters. The likelihood to observe pair-wise random matches of small variants between unrelated CCLs was, for instance, approximated by empirically calculating the amount of matches within a collection of reference CCLs.

Disadvantages of the distance-quantification concept

Three major disadvantages of the distance-quantification concept were identified during the development of the methods. All disadvantages are associated with either a reduction of the usefulness of the distance-quantification or the addition of complexity to the method development due to work required to implement the distance-concept or to recover the distance-quantification's usefulness.

Lack of comparability

A quantified distance is only informative relative to other distances, i.e. comparability has to be given between distances. However, comparability is frequently not ensured. For instance, different datasets, processed with different Next-Generation Sequencing (NGS) technologies may present with different deconvolution-derived cell-type proportions for the subclass of high grade malignant carcinomas. The distance of a deconvolved entity as quantified by the cell-type proportions therefore differs depending on the utilized technology even when the predicted cell-type proportions possess the same numerical value.

Numerically identical numbers can therefore counterintuitively indicate different biological or technological distances.

The disadvantage was found to primarily occur between different datasets and requires either a normalization of the deconvolution training-data or a numerical correction of the output to balance the batch effects. A major ramification of lack of comparability is that Machine-Learning models trained the cell-type proportions from one dataset will overfit on a dataset due to the batch effect. The reason that a normalization step has not been integrated into the respective scientific contribution in Chapter 5 is, that only one dataset present with the central class of high-grade series carcinoma in sufficient sizes. Since only one dataset possesses this class, no inter-dataset normalization could be developed.

Lack of interpretability

The geometric interpretation of an abstract distance can be difficult or even misleading in a complex scenario where conditions on a metric space are at least partially violated. An example is the scenario where Data-Heterogeneity 'curves' the space within distances are quantified, as illustrated in Figure 4.1. Analogously, percentages of cell-type proportions cannot be easily interpreted both from a biological perspective which inquires what the biological interpretation of, for instance, a ductal cell-type proportions of 20% is. The mathematical aspect with regards to interpretability is connected to the lack of comparability in that a Machine-Learning model commonly utilize thresholds to separate classes. However, 20% ductal proportion can potentially be interpreted as a great proportion for one dataset but has to be interpreted as low for another, i.e. the model cannot interpret the distance correctly.

A consequence of the lack of interpretability is the exacerbation of hypothesis validation. The design a in-vitro experiment which supports or refutes a hypothesis is challenging When a biological interpretation of an observed phenomenon is not possible. The lack of interpretability does affect the training of Machine-Learning models given that it is trained and applied to the same datasets. However, the Machine-Learning model will show a reduced performance when applied to other datasets whose thresholds might not be *interpretable* i.e. comparable for the Machine-Learning model. Lack of interpretability is therefore most commonly associated with cross-datasets experiments but generally absent in scenarios where only one datasets or highly homogeneously generated datasets are analyzed.

Lack of parsimoniousness

The utilization of a distance-quantification approach adds complexity to the development of a method. The identification of CCLs can, for instance, be conducted without a geometric interpretation of their pair-wise distances. The underlying reason is, that statistical tests ultimately decide on the outcome of the identification what does not required the assignment of human being-interpretable positions in a metric space. In case of the classification-by-deconvolution approach, an abstract distance quantification was helpful in that a biological interpretation in form of the cell-types was available, but the training of Machine-Learning models was ultimately based on correlations between cell-type proportions. A geometric interpretation was not found to be identical with the most parsimonious approach to a given problem.

Modifications and future research

Suggested changes incorporate three major aspects. Firstly, future research should to ascertain that any applied distance-quantification in a metric space will remain invariant to perturbations of the data. Perturbations, such as those caused by Data-Heterogeneity, may violate the mathematical conditions required by a metric as observed in the second contribution in Chapter 4. The compensation for the increase in Data-Heterogeneity was an introduction of Biostatistics which both increased the complexity of the method and reduced the sensitivity. Secondly, a decision to integrate a geometric integration should be made before the development of the methods based on a positive trade-off gain between increased complexity and the added support of the geometric interpretation. Thirdly, the comparability of quantified distances between different datasets has to be ensured in particular with respect to the training of Machine-Learning methods. If deconvolution-by-classification is to become a generically applicable approach for the data-augmentation of rare and diverse neoplasms, the overfitting of models between different datasets has to be addressed. Therefore, the normalization of predicted cell-type proportions between different datasets would be analyzed in detail before Machine-Learning models would be trained. Ideally, a Machine-Learning model can first normalize the deconvolution-predictions of different datasets and secondly train a model on all normalized predictions with inclusion of a hold-out dataset where the generalized performance can be validated.

The thesis concludes that the concept of abstract distance-quantification is helpful during the conception stage of the contributed methods but less useful in complex scenarios with elevated Data-Heterogeneity.

6.1 Outlook

The distance-based identification of CCL, whose NGS data was generated with different NGS technologies such as methyl-sequencing, remains to be explored and can possibly present with suitable CCL identification performances. Specialized approaches which focus on the identification of Panel-sequenced datasets can as well show positive results since the universal Uniquorn version revealed an effective performance with respect to the identification of Panel-sequenced CCLs when larger amounts of loci where targeted and purpose-built methods might decrease the lower bound on the amount of loci further.

A distance-quantification based deconvolution approach represents a promising subject for future research since it can prospectively support clinicians in their challenging classification endeavors. A highly relevant future approach to deconvolution would be the combination of expression and methyl-sequencing. An advantage of the methyl-sequencing technology is that the topological domain structures are strongly associated with cell-type characteristic and remain intact even after cellular neoplastic transformation. The volatility of Single-Cell sequencing datasets motivates the development of deconvolution ensemble methods which average out the training dataset differences. Transcriptomic deconvolution can potentially be applied to determine the neoplastic cell-type-of-origin, but a gold-standard benchmark dataset will have to be generated to render that highly valuable scientific goal possible.

Chapter 7

Supplement

7.1 Supplementary figures

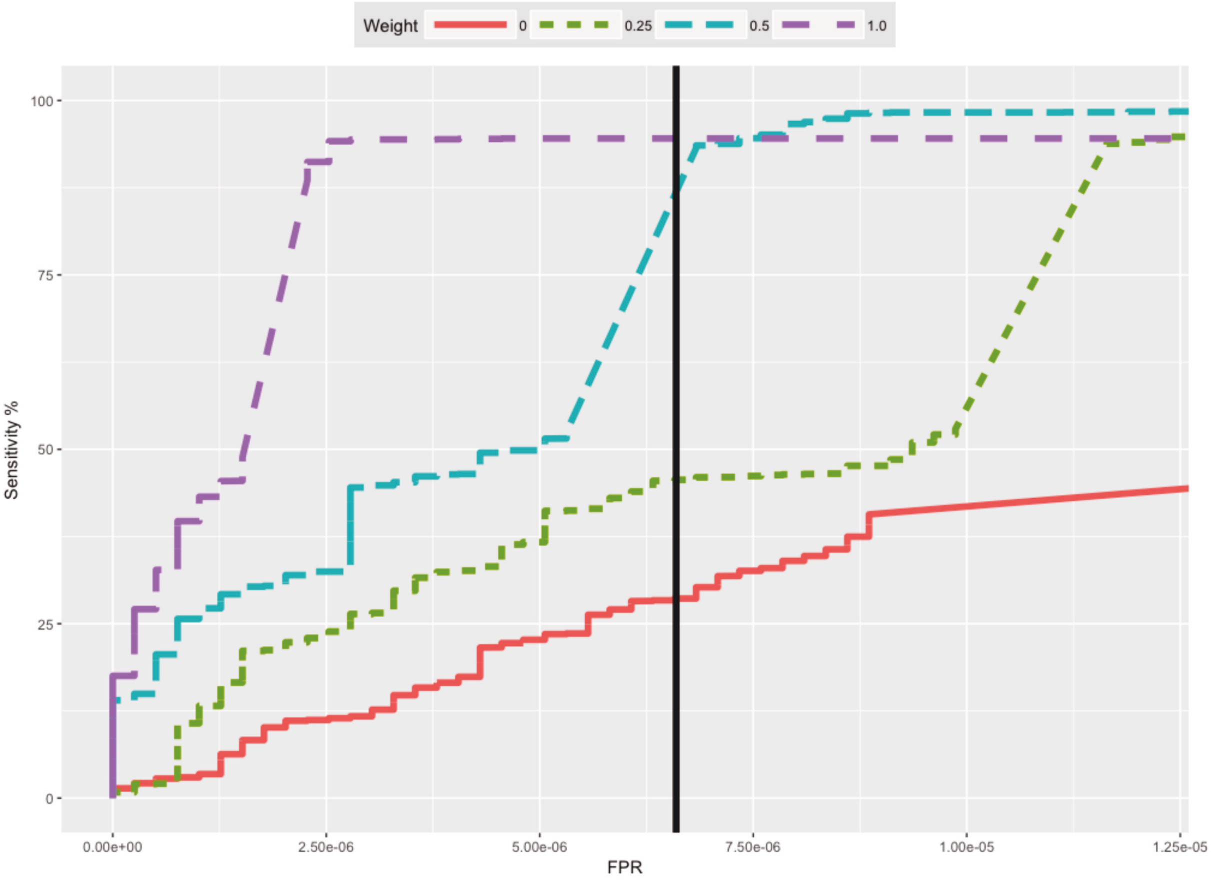

Figure 7.1: ROC-curve of the cross-identification benchmark for different weight thresholds. Thresholds 0.5 and 0.25 reached the maximal sensitivity (see also Table 3.2). The embedded plot shows the same ROC plot with an adjusted FPR-axis range to visualize the ROC-curve of inclusion weight 0.0. The vertical black line shows the Uniquorn proof-of-concept default threshold (confidence score of 10). The threshold was chosen as optimal cutoff between sensitivity and specificity.

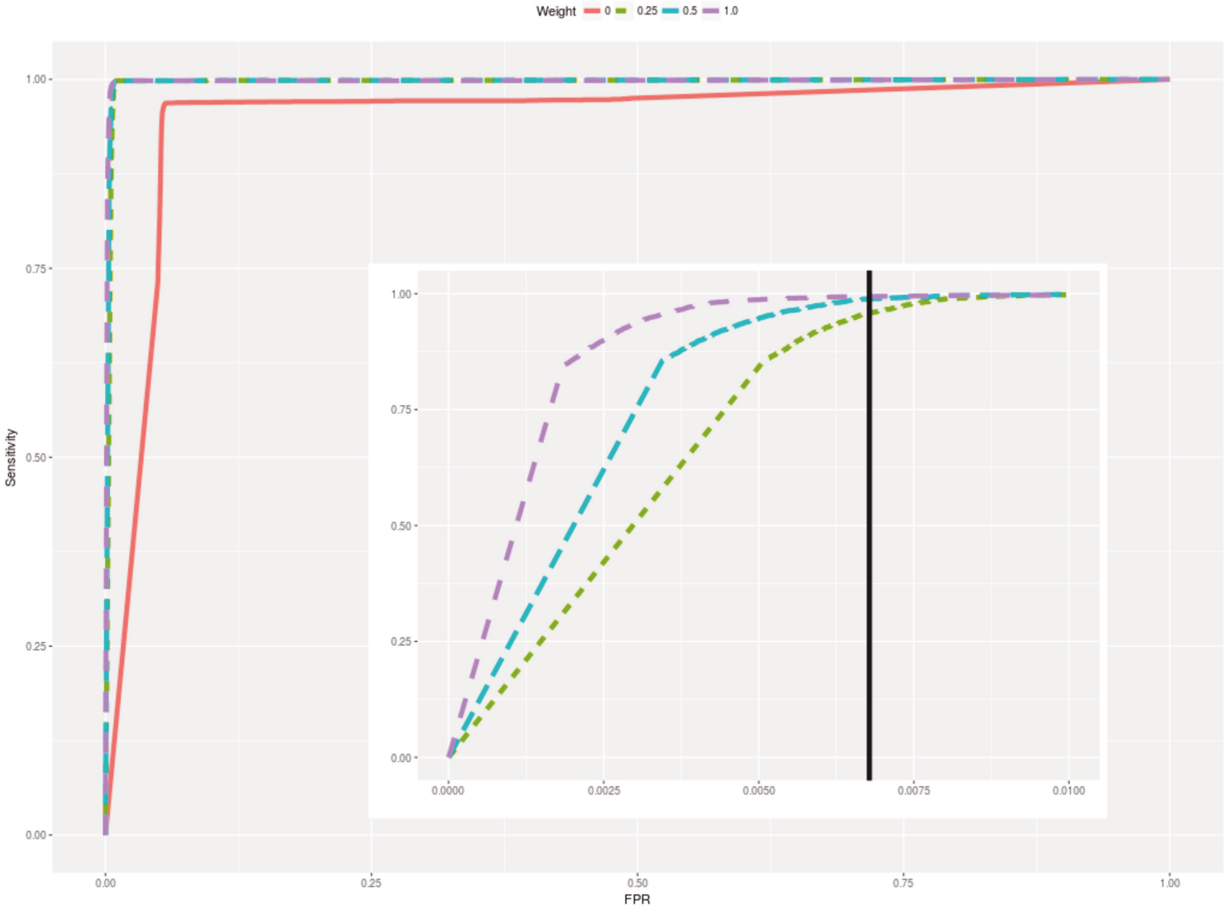

Figure 7.2: The ROC-curve iterates over the confidence score compares a score's associated sensitivity and associated specificity. It can thus be seen how the overall default weight threshold was chosen as the optimal ratio between sensitivity and specificity. The embedded plot shows the same ROC-curve plot with an adjusted FPR-axis range to visualize the ROC-curve of inclusion weight 0.0. The vertical black line shows the Uniquorn default threshold sensitivity to specificity ratio. The identified thresholds were set as default within the *Uniquorn* package.

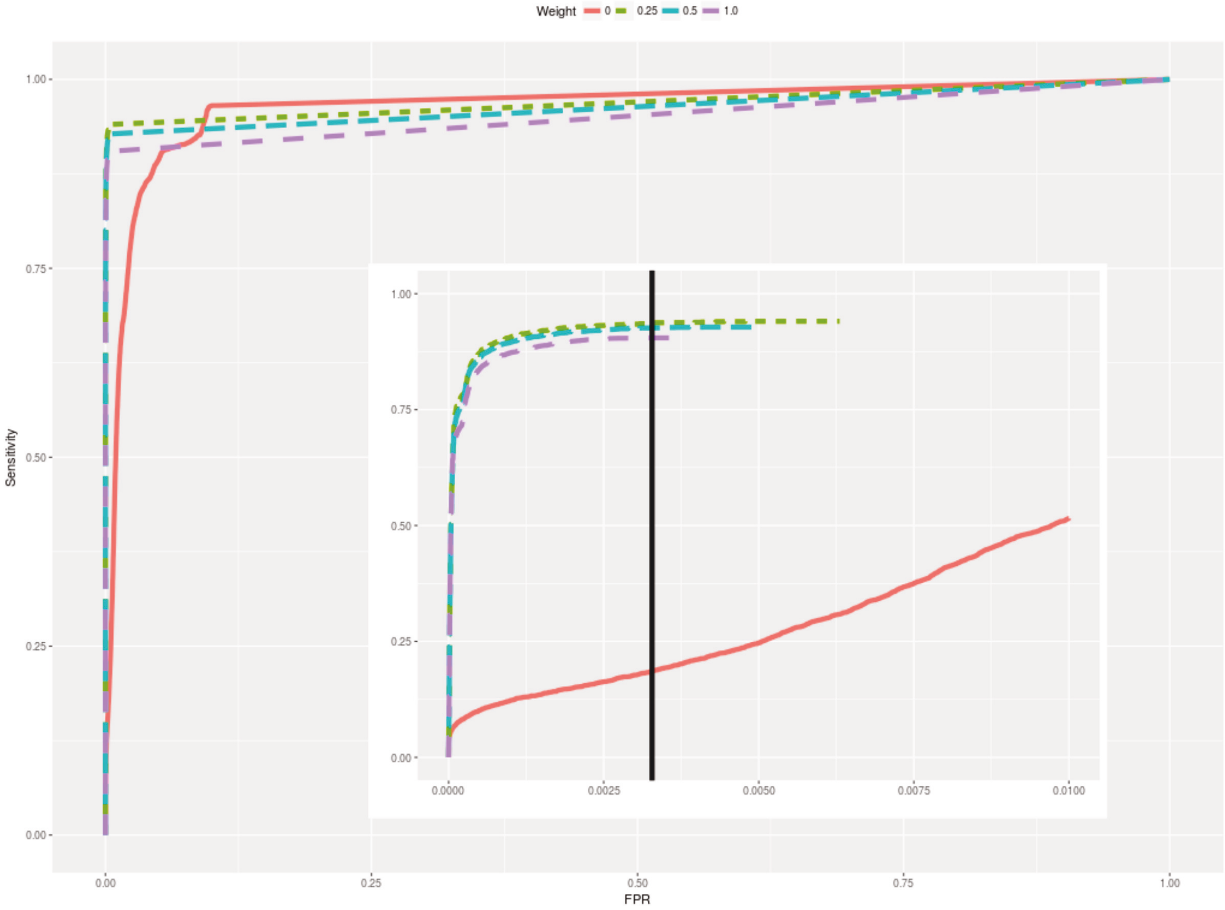

Figure 7.3: The Figure description is identical to Figure 7.2. Inclusion weights 0.5 and 0.25 show the best ratio between sensitivity and FPR. The identification of different optimal threshold for panel and non-panel-sequencing indicates, that users should actively adapt the identification threshold.

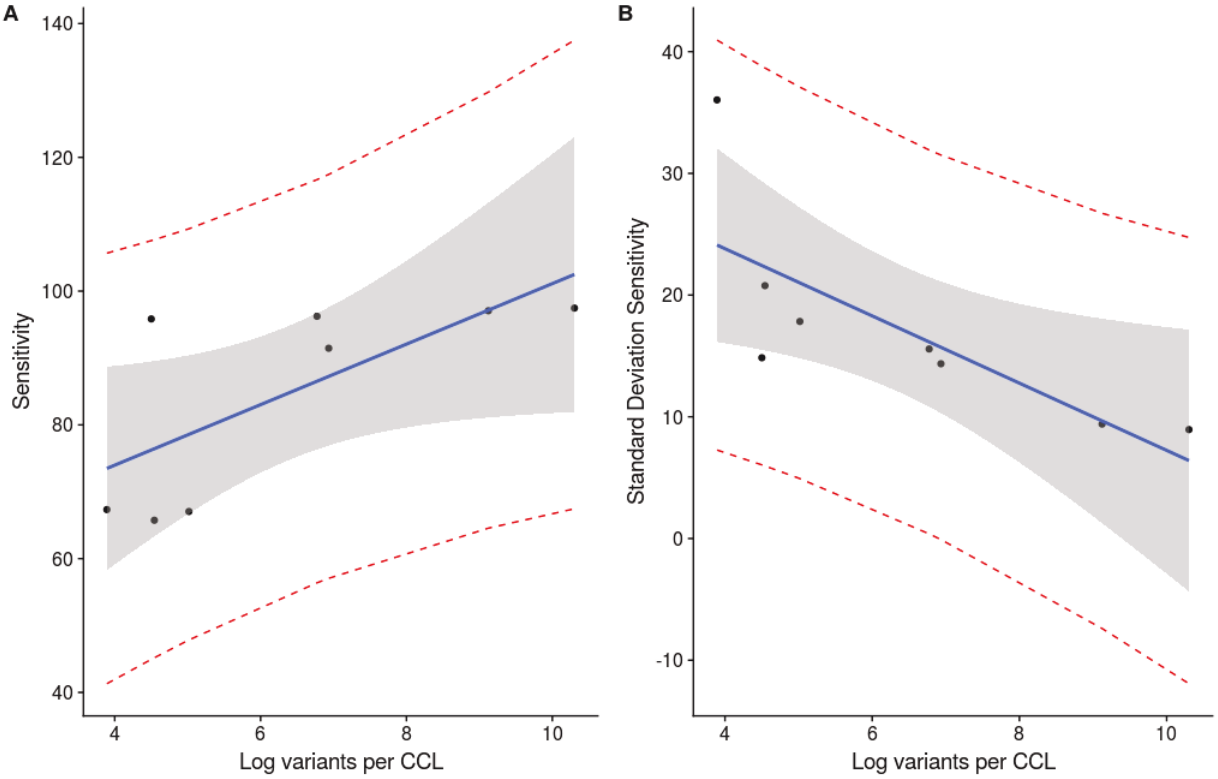

Figure 7.4: Correlation between sensitivity and average variant-count per CCL within a library. A linear regression depicts the relationship between the variant count of a CCL-profile and the sensitivity with which CCL-profiles can be identified. A log-linear correlation between a library's average number of variants per contained CCL-profile and the sensitivity with which these profiles exists. Sensitivity is correlated with an r of 0.7 and regression p-value of 0.041 and standard deviation with an r of -0.75 and p-value of 0.03. Shaded areas indicate the regression standard error and dashed lines indicate the regression's 95% confidence interval.

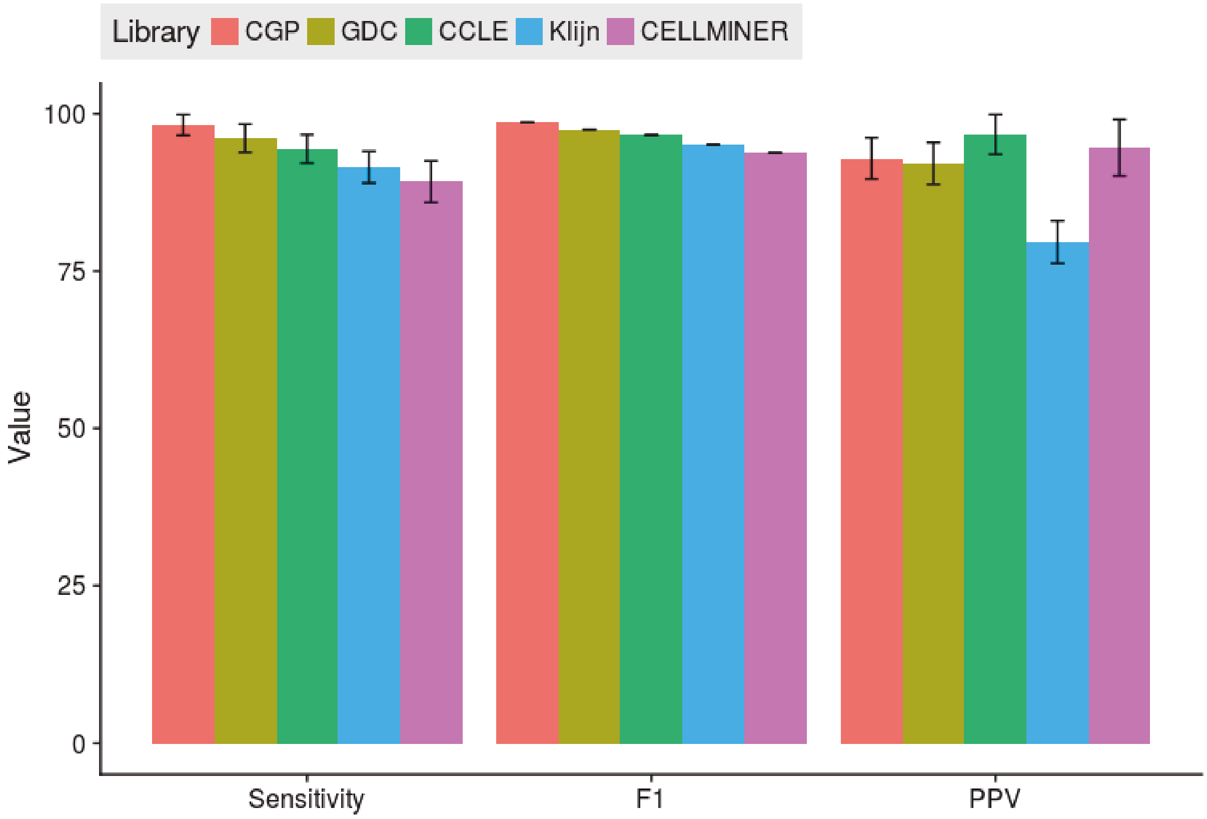

Figure 7.5: Benchmark results split by library. Sensitivity and F1 value do not show a significant change between libraries. Overall, the benchmark results remain robust with the exception of the PPV for the Klijn library which is a minor outlier due to slightly worse PPV.

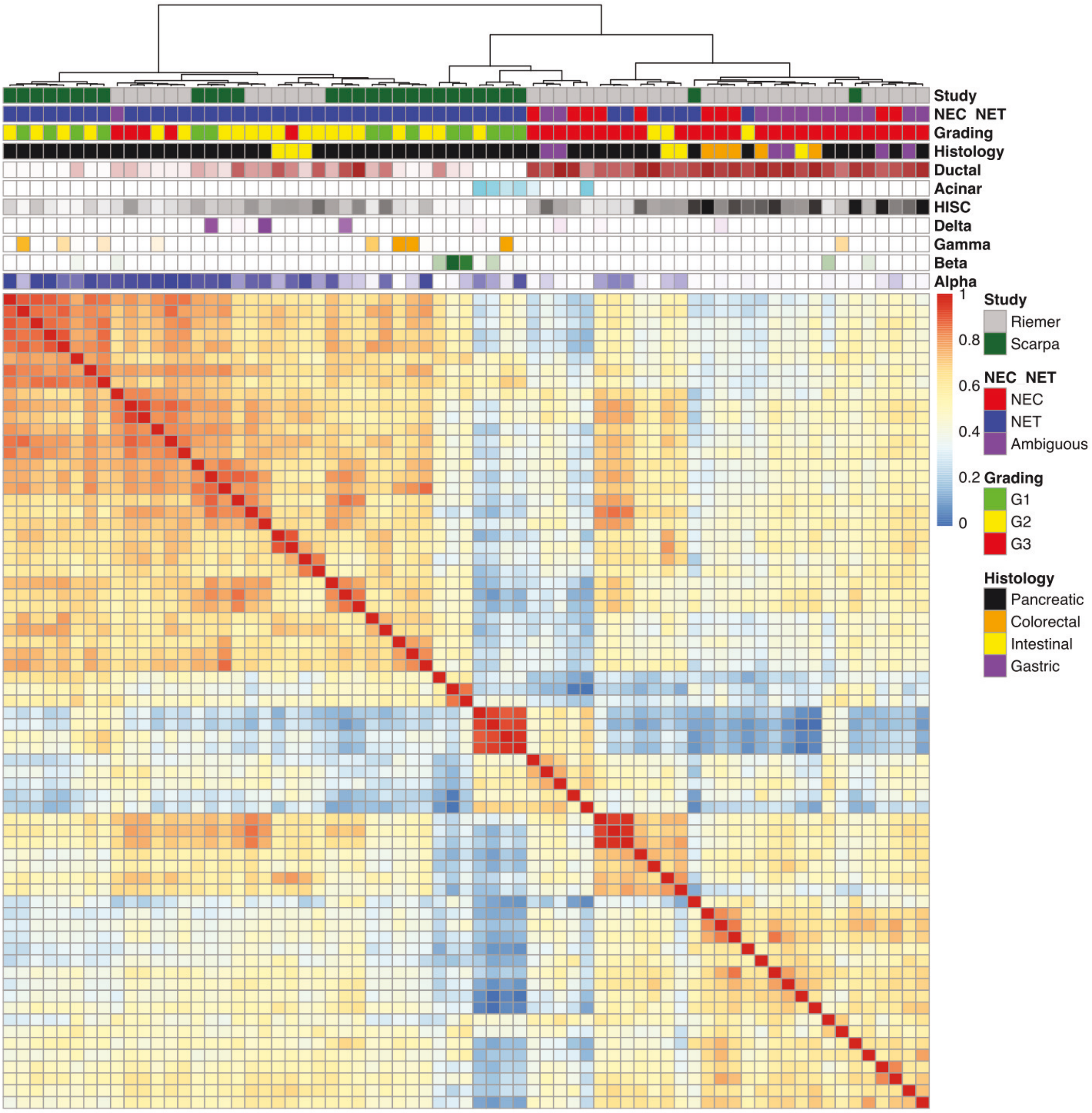

Figure 7.7: Transcriptome correlation heatmap of the deconvolution results of the RepSet. The underlying data is identical to Supplementary Figure 7.10. Deconvolution-derived relative cell type proportions along with the grading, histology and Ki-67 levels are shown in the top rows. A clustering of high ductal and HISC cell type proportions predictions is obviated for high-grade GEP-NENs. Note, that the alpha-cell type dominated samples differ depending on whether ductal or HISC cell types are included in the training data. In case that alpha and ductal cell types are incorporated into a single model (depicted), the alpha signature is highest for differentiated models and only in case of an exclusively endocrine model (not depicted) can a clustering of high alpha cell type proportion predictions be observed in high-grade series GEP-NENs.

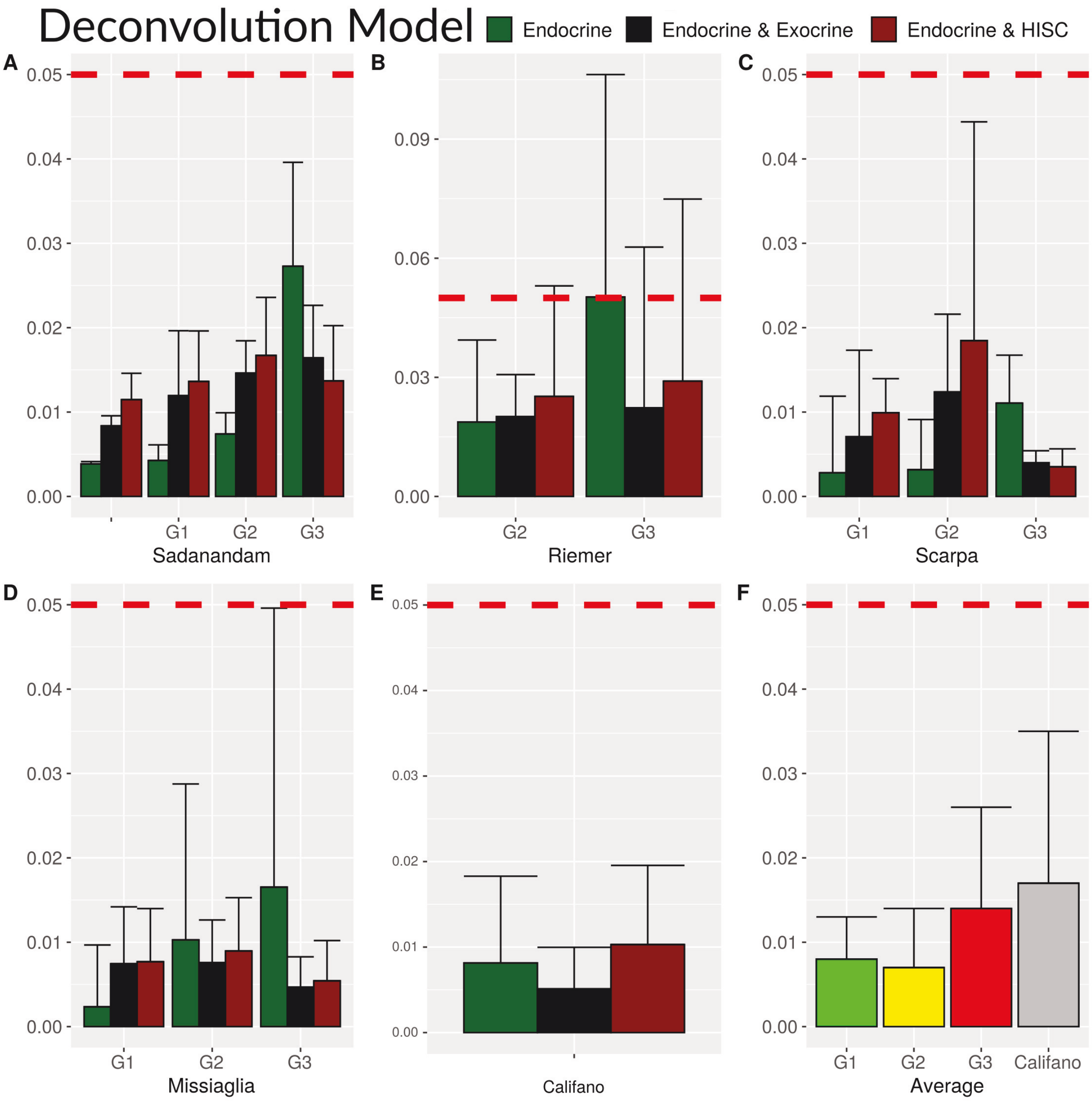

Figure 7.8: Reconstruction p-value for the Sadanandam, Riemer, Scarpa, Missiaglia and Califano datasets and the global average. The figure shows that the identified optimal combination of deconvolution algorithm and scRNA training dataset could create significant deconvolution models for all benchmarked datasets. Different datasets differed with respect to their proclivity to deconvolve, however, a deconvolution was generally possible for every dataset regardless of the specific tissue background.

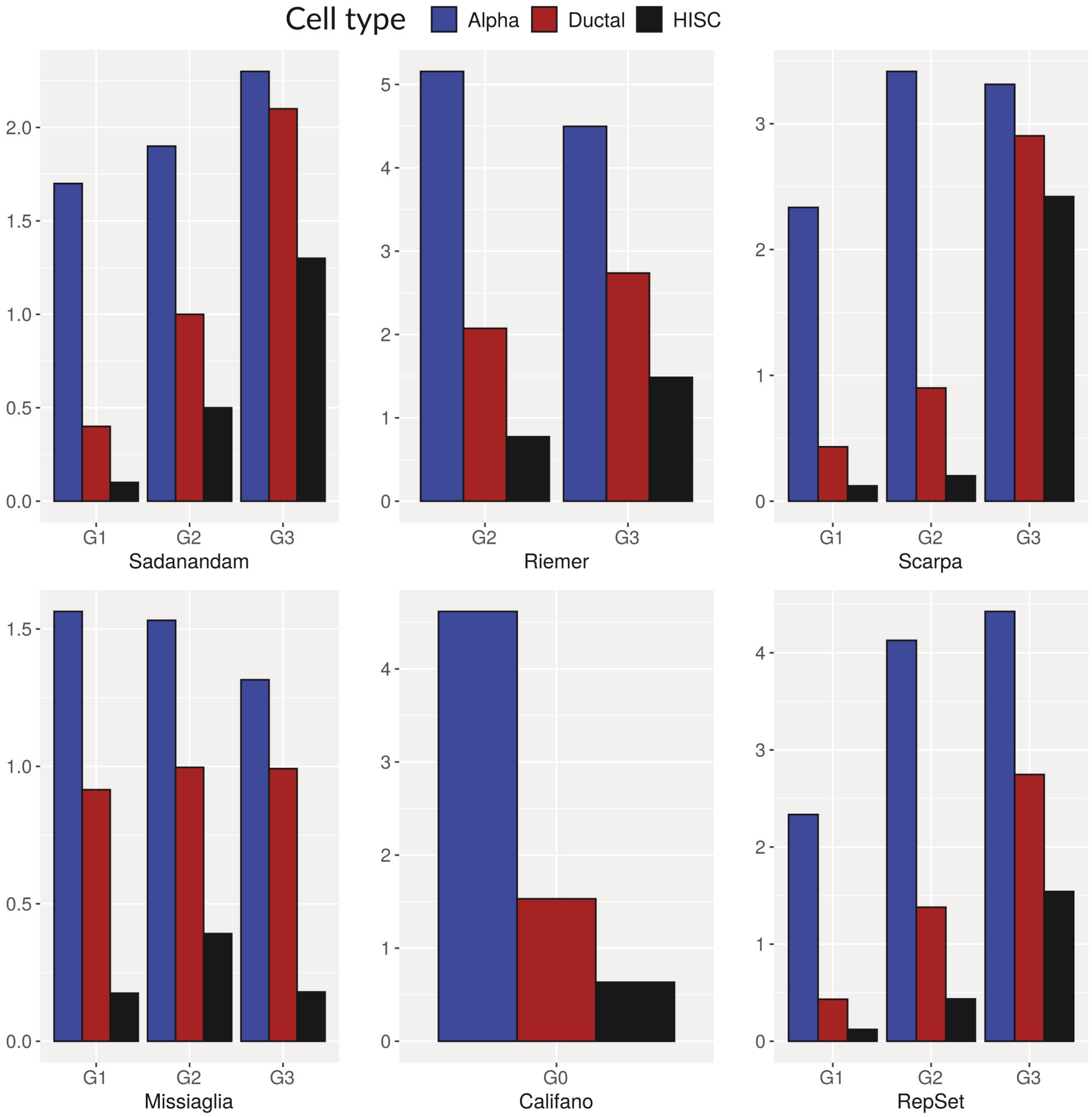

Figure 7.9: Regression coefficients for the alpha, ductal and HISC cell type proportions for all benchmarked datasets. The regression coefficients as opposed to the relative cell type proportions are shown aggregated over the gradings. It is displayed that a general trend towards increased ductal and HISC cell type proportion exists when grading increases. The alpha cell type proportion does show a general proclivity as well however with greater variance. Note that the depicted alpha cell type proportions are exclusively retrieved from the exclusive endocrine alpha to delta model and that alpha cell type proportions in the other two models are not depicted.

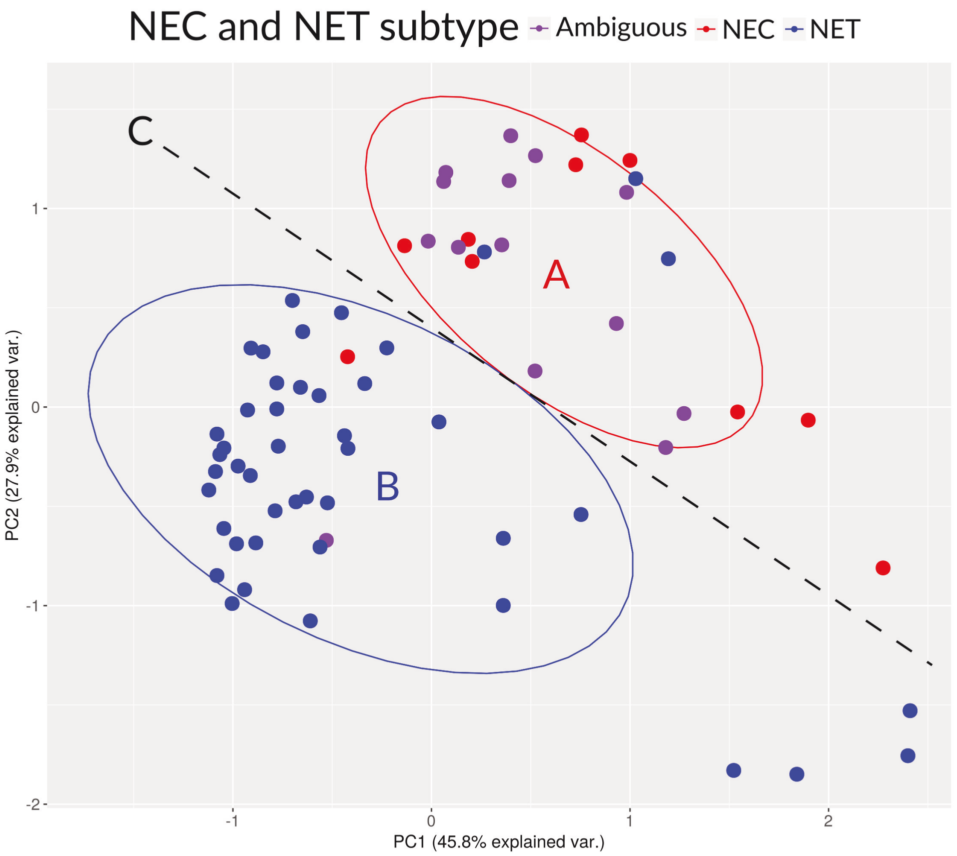

Figure 7.10: PCA of the correlation matrix of a subset of 204 genes of the RepSet. The dataset was reduced to 204 genes specified as characteristic for PanNEN molecularly defined subgroups with different metastatic potential by Sadanandam et al. [163]. A distinctive NET (A) and to a lesser extent distinctive NEC cluster (B) with inclusion of multiple ambiguous and few NET-classified samples was found when applying a PCA on the correlation matrix. Identification of a linearly separating decision boundary (C) of NECs and NETs involving both Principal-Component (PC) 1 and PC 2 was possible, despite the misclassification of NETs in the NEC cluster and a single NEC in the NET cluster. Ambiguous samples were manually assigned as NEC or NET depending on their membership in either the cluster A or cluster B for the purpose of the NEC and NET classification shown in Figure 5.3. One of the three 'NET-outlier' in the NEC-cluster was a sarcoma-like Tumor classified as NET based on NE marker expression and Ki-67 staining, which otherwise exhibited features of a small-round-blue cell tumor, recently shown to resemble small cell neuroendocrine cancer. The second outlier was obtained from a patient with an established prior history of NET and the outlier had received several courses of platinum based chemotherapy prior to tissue sampling. The third sample had been annotated as well differentiated G3 NET, but was connected to an aggressive clinical course with disease-related death occurring within seven months.

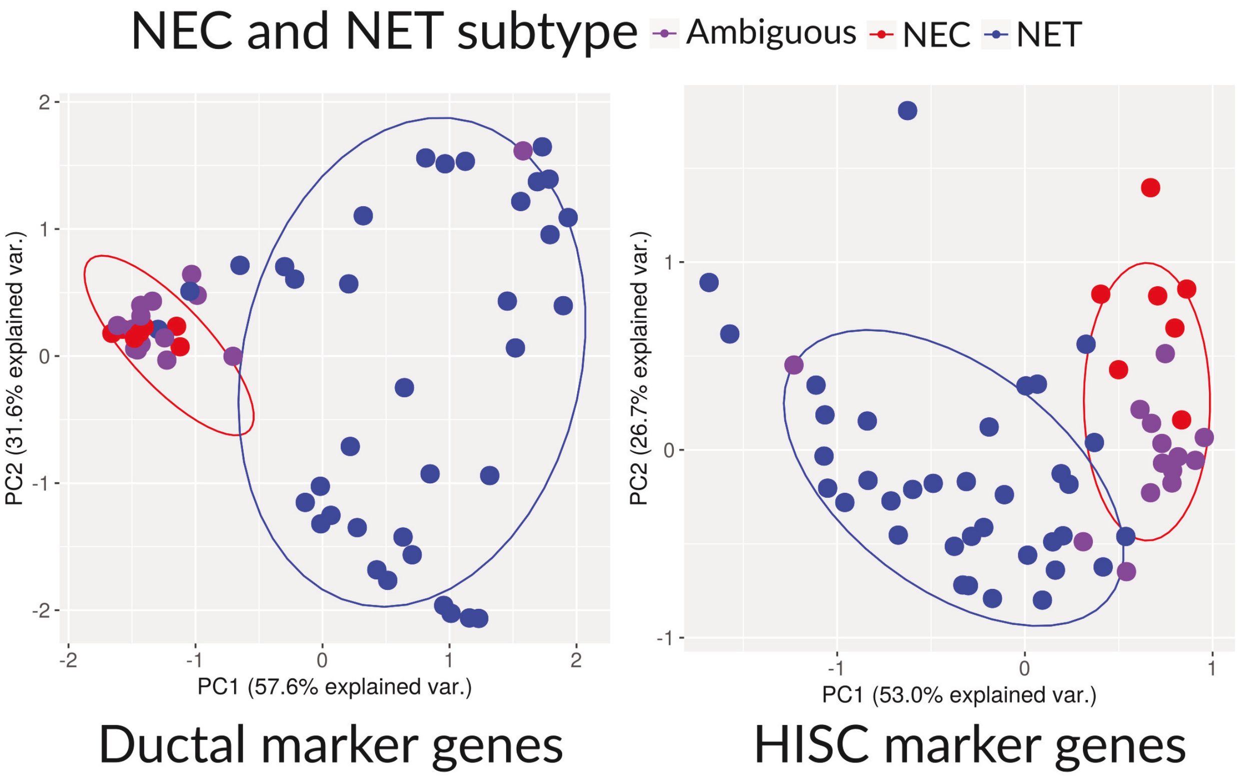

Figure 7.11: Principal component analysis of the transcriptome of the RepSet reduced to sets of cell type marker genes. Expression datasets were reduced to the marker gene sets of the ductal and HISC cell types as determined by the CIBERSORT algorithm, respectively, and their PCAs depicted after colorization for NEC and NET subtype. Purple dots represent samples whose NEC or NET classification was not unanimously possible. It is visualized that a machine learning model trained on the ductal or HISC signature-based deconvolution results can classify NECs and NETs because NECs and NET differ characteristically in the set of genes that are utilized to deconvolve the ductal or HISC cell type proportions. Note clustering by tissue of origin was identified which was supported by the finding that the underlying ductal and HISC marker genes predominantly consist of genes whose expression does not differ between the relevant tissue type of origin.

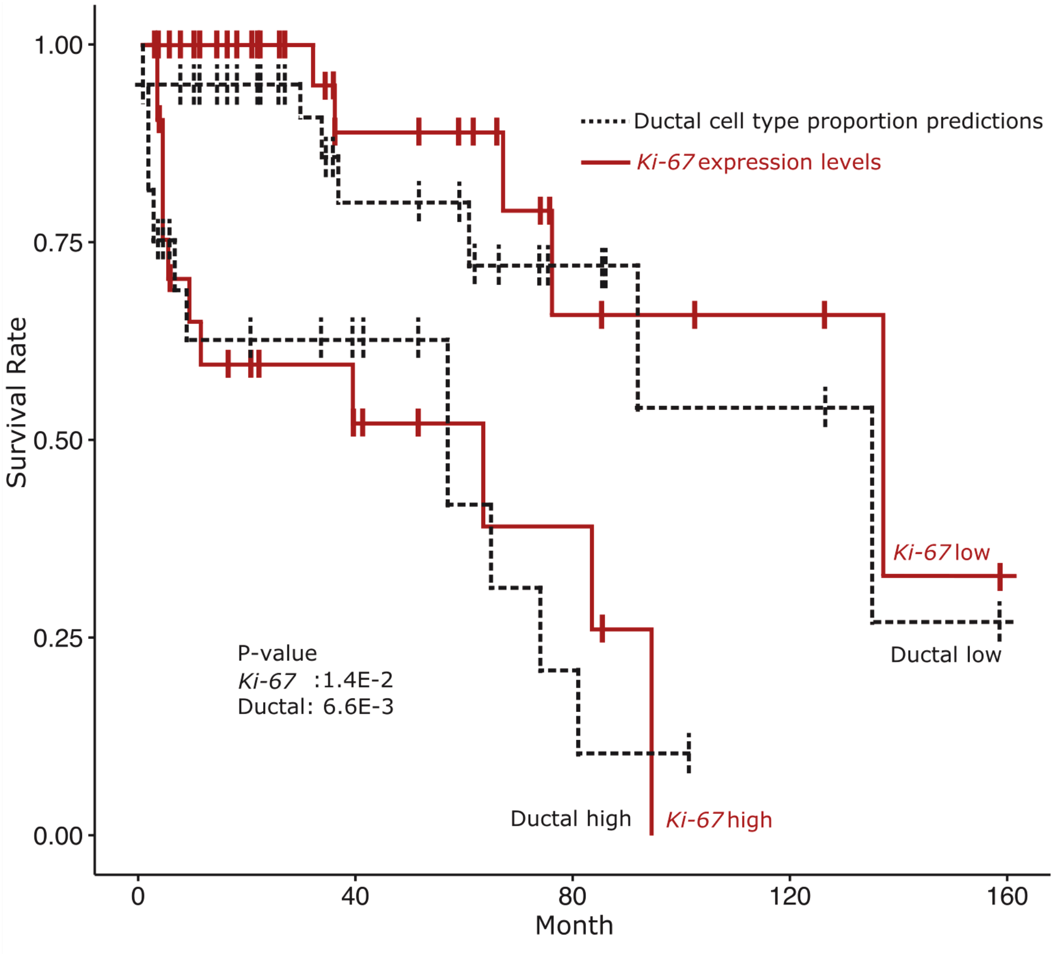

Figure 7.12: Kaplan-Meier survival plot of disease-related patient survival time comparing the predictive performance of the Ki-67 levels and ductal signature in the RepSet dataset for a two-arm design. A 'high' and 'low' risk subgroup of the RepSet were constructed based on either Ki-67 expression levels or the predicted ductal proportion predictions and Cox-hazard ratio test on difference of survival between the subgroups applied. It is illustrated that either test was significant, that the Ki-67 based test had more statistical power but that the overall trend of survival rate prediction remained comparable between ductal cell type proportion and the Ki-67 based predictions. Note that the depicted ductal proportion predictions were extracted from a CIBERSORT model trained on Lawlor et al. as opposed to ubiquitously presented Baron et al. scRNA data.

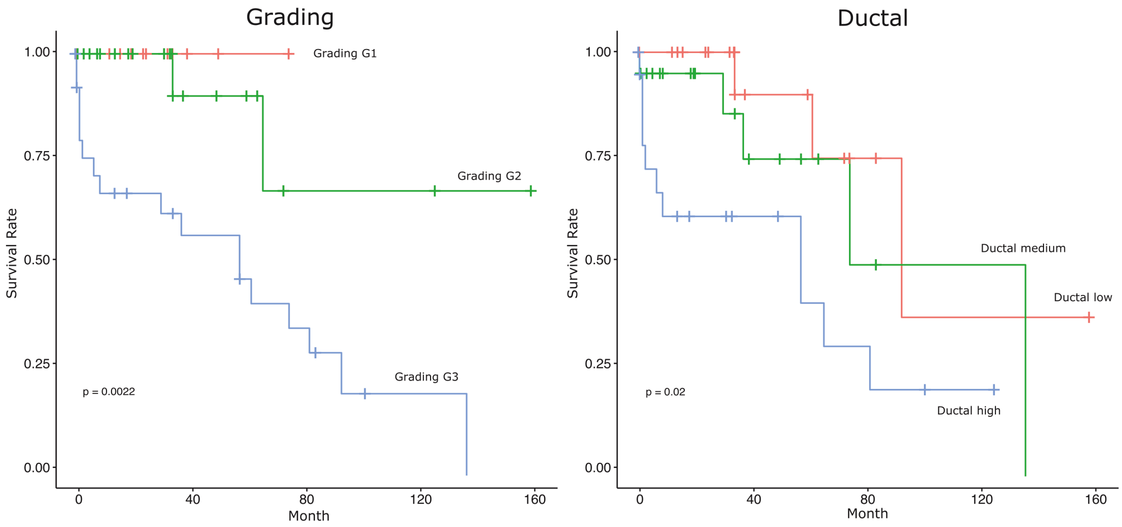

Figure 7.13: Kaplan-Meier survival plot of disease-related patient survival time in the RepSet dataset comparing the predictive performance of the grading and ductal signature for a three arm design. The grading shows a clearly superior statistical power to distinguish the three cohorts at a p-value of 0.0022, however, the ductal cell type proportion predictions remain significant at a p-value of 0.019.

7.2 Supplementary tables

CL1	CL2	Reason
7860	786O	Synonym
253J	253JBV	Synonym
BE13	PEER	Synonym
CMK115	CMK	Synonym
COLO320HSR	COLO320	Synonym
COLO704	COLO684	Synonym
EOL1	EOL1CELL	Synonym
H3255	NCIH3255	Synonym
HEC1A	HEC1	Synonym
HEC1B	HEC1	Synonym
HEL9217	HEL	Synonym
HEY	HEYA8	Synonym
HUH6CLONE5	HUH6	Synonym
KP1NL	KP1N	Synonym
LC1F	LC1SQSF	Synonym
LC1SQ	LC1F	Synonym
LC1SQSF	LC1SQ	Synonym
LU99A	LU99	Synonym
M059K	M059J	Synonym
M059KJ	M059K	Synonym
NCIADRRES	OVCAR8	Synonym
NCIH510	NCIH510A	Synonym
NCIH510A	NCIH510	Synonym
NCISNU1	SNU1	Synonym
NCISNU5	SNU5	Synonym
OC314	OC316	Synonym
OVCAR3	NIHOVCAR3	Synonym
PECAPJ15	CAPJ15	Synonym
RT112	RT11284	Synonym
SNU1	NCISNU1	Synonym
SNU5	NCISNU5	Synonym
SNUC2B	SNUC2A	Synonym
U266	U266B1	Synonym
UO31	U031	Synonym
WM793	WM793B	Synonym
WM793B	WM793	Synonym
AU565	SKBR3	Related
C3A	HEPG2	Related
COLO201	COLO205	Related
COLO775	RPMI8226	Related
COLO800	COLO818	Related
FTC133	FTC238	Related
GP2D	GP5D	Related
GT3TKB	RERFLCAI	Related
H9	HUT78	Related
HCC1588	LS513	Related
HCMB	CHL1	Related
HLE	HLF	Related
HOP92	U251	Related
HRT18	HCT15	Related
HTK	HOS	Related
IGR39	IGR37	Related
IMR5	IMR32	Related
KARPAS422	OCILY10	Related
M059J	M059K	Related
MCIXC	SKNMC	Related
MKN28	MKN74	Related
MONOMAC1	MONOMAC6	Related
NCIH1770	NCIH2106	Related
NCIH1993	NCIH2073	Related
ONCODG1	OVCAR3	Related
SHSY5Y	SKNSH	Related
SNB19	U251	Related
SNUC2A	SNUC2B	Related
SW480	SW620	Related
SW579	CGTHW1	Related
T24	ACCS	Related
TOV112D	HS571T	Related
TUR	U937	Related
U118MG	U138MG	Related
WM115	WM2664	Related
YMB1E	ZR751	Related
KPL1	MCF7	Known cross-contamination or derivation
M14	MDAMB435	Known cross-contamination or derivation
M14	MDAMB435S	Known cross-contamination or derivation
MDAMB435	M14	Known cross-contamination or derivation
MDAMB435	MDAMB435S	Known cross-contamination or derivation
MDAMB435	MDAN	Known cross-contamination or derivation
MDAN	M14	Known cross-contamination or derivation
MDAN	MDAMB435S	Known cross-contamination or derivation
OVCAR8	NCIADRRES	Known cross-contamination or derivation
SR	SR786	Known cross-contamination or derivation

Table 7.1: Known cross-contaminations and derivation Gold Standard utilized to benchark Uniquorn proof-of-concept and Uniquorn in its universal version

Weight Threshold	Maximally possible TPs	True positives	False negatives	False positives	Sensitivity %	F1 %	PPV
Uniquorn POC							
1	3573	3027	546	22	85	91	99
0.5		3474	99	37	97	98	99
0.25		3461	112	59	97	98	98
0		3111	462	4631	87	55	40
Uniquorn universal							
1	3573	3403	170	126	95	96	96
0.5		3408	165	123	95	96	97
0.25		3408	165	124	95	96	96
0		3402	171	135	95	96	96

Table 7.2: Comparison of Uniquorn performance in its proof-of-concept and universal version. The same benchmark as conducted in Uniquorn proof-of-concept was conducted and results compared. It can be see that the respective methods' performance is comparable. POF = proof-of-concept

Spia CCLs	Unified CCL Labels 1	Equivalent Uniquorn Benchmark	Identification successful at weight 0.5
184A1	184A1		
184B5	184B5		
501mel	501MEL		
786-O	786O	786O	1
A172	A172	A172	1
A375	A375	A375	1
A498	A498	A498	1
A549	A549	A549	1
ACHN	ACHN	ACHN	1
AN3-CA	AN3CA	AN3CA	1
BT-20	BT20	BT20	1
BT-474	BT474	BT474	1
BT-549	BT549	BT549	1
CAKI-1	CAKI1	CAKI1	1
Caki.2	CAKI2	CAKI2	1
CAL-51	CAL51	CAL51	1
CAMA-1	CAMA1	CAMA1	1
CAPAN-1	CAPAN1	CAPAN1	1
CCRF-CEM	CCRFCEM	CCRFCEM	1
CL-11	CL11	CL11	1
CL-14	CL14	CL14	1
COLO-205	COLO205	COLO205	1
COLO-320	COLO320	COLO320	1
COLO-824	COLO824	COLO824	1
CPC-N	CPCN	CPCN	1
DU-145	DU145	DU145	1
DU4475	DU4475	DU4475	1
EFE-184	EFE184	EFE184	1
EFM-19	EFM19	EFM19	1
EKVX	EKVX	EKVX	1
H128	NCIH128	NCIH128	1
H1339	NCIH1339	NCIH1339	1
H1395	NCIH1395	NCIH1395	1
H1437	NCIH1437	NCIH1437	1
H1450	NCIH1450		
H1770	NCIH1770	NCIH1770	1
H1819	NCIH1819		
H2009	NCIH2009	NCIH2009	1
H2141	NCIH2141	NCIH2141	1
H2171	NCIH2171	NCIH2171	1
H2195	NCIH2195		
H220	NCIH220		
H378	NCIH378	NCIH378	1
H460	NCIH460	NCIH460	1
H838	NCIH838	NCIH838	1
H889	NCIH889	NCIH889	1
HCC1143	HCC1143	HCC1143	1
HCC1187	HCC1187	HCC1187	1
HCC1395	HCC1395	HCC1395	1
HCC1419	HCC1419	HCC1419	1
HCC1500	HCC1500	HCC1500	1
HCC1569	HCC1569	HCC1569	1
HCC1599	HCC1599	HCC1599	1
HCC1806	HCC1806	HCC1806	1
HCC1937	HCC1937	HCC1937	1
HCC1954	HCC1954	HCC1954	1
HCC202	HCC202	HCC202	1
HCC2157	HCC2157	HCC2157	1
HCC2218	HCC2218	HCC2218	1
HCC2998	HCC2998	HCC2998	1
HCC33	HCC33	HCC33	1
HCC38	HCC38	HCC38	1
HCC70	HCC70	HCC70	1
HCC970	HCC970		
HCT-116	HCT116	HCT116	1
HCT-15	HCT15	HCT15	1
HL-60	HL60	HL60	1
HOP-62	HOP62	HOP62	1
HOP-92	HOP92	HOP92	1
HS 578T	HS578T	HS578T	1
HT-29	HT29	HT29	1
HUP-T3	HUPT3	HUPT3	1
HUP-T4	HUPT4	HUPT4	1
IGROV1	IGROV1	IGROV1	1

Table 7.3: Comparison of the performance to SPIA. The table lists whether regularized CCL labels were correctly found both by SPIA and Uniquorn (1 in indicator column)

K-562	K562	K562	1
KC12	KC12		
KM12	KM12	KM12	1
KO295	KO295		
KPL-1	KPL1	KPL1	1
KU-19-20	KU1920		
LOX IMVI	LOXIMVI	LOXIMVI	1
LuCaP 35	LUCAP35		
M14	M14	M14	1
Malme-3M	MALME3M	MALME3M	1
MCF-10A	MCF10A		
MCF-12A	MCF12A		
MCF7	MCF7	MCF7	1
MDA Pca 2b	MDAPCA2B	MDAPCA2B	1
MDA-MB-134-VI	MDAMB134VI	MDAMB134VI	1
MDA-MB-157	MDAMB157	MDAMB157	1
MDA-MB-175	MDAMB175		
MDA-MB-231	MDAMB231	MDAMB231	1
MDA-MB-415	MDAMB415	MDAMB415	1
MDA-MB-435	MDAMB435	MDAMB435	1
MDA-MB-453	MDAMB453	MDAMB453	1
MDA-MB-468	MDAMB468	MDAMB468	1
MDA-MB361	MDAMB361	MDAMB361	1
MDAMB-330	MDAMB330	MDAMB330	1
MEWO	MEWO	MEWO	1
MFE-280	MFE280	MFE280	1
MFE-296	MFE296	MFE296	1
MOLT-4	MOLT4	MOLT4	1
N15C6 p48	N15C6P48		
N33B2 p21	N33B2P21		
NCI-ADR-RES	NCIADRRES	NCIADRRES	1
NCI-H226	NCIH226	NCIH226	1
NCI-H23	NCIH23	NCIH23	1
NCI-H322M	NCIH322M	NCIH322M	1
NCI-H460	NCIH460	NCIH460	1
NCI-H522	NCIH522	NCIH522	1
NCI-H660	NCIH660	NCIH660	1
OVCAR-3	OVCAR3	OVCAR3	1
OVCAR-4	OVCAR4	OVCAR4	1
OVCAR-5	OVCAR5	OVCAR5	1
OVCAR-8	OVCAR8	OVCAR8	1
PC-3	PC3	PC3	1
PMC42	PMC42		
RPMI-8226	RPMI8226	RPMI8226	1
SF-268	SF268	SF268	1
SF-295	SF295	SF295	1
SF-539	SF539	SF539	1
SK-BR-3	SKBR3	SKBR3	1
SK-MEL-2	SKMEL2	SKMEL2	1
SK-MEL-28	SKMEL28	SKMEL28	1
SK-MEL-5	SKMEL5	SKMEL5	1
SK-OV-3	SKOV3	SKOV3	1
SLR 20	SLR20		
SLR 21	SLR21		
SLR 22	SLR22		
SLR 23	SLR23		
SLR 24	SLR24		
SLR 25	SLR25		
SLR 26	SLR26		
SN12C	SN12C	SN12C	1
SNB-19	SNB19	SNB19	1
SNB-75	SNB75	SNB75	1
SR	SR	SR	1
SUM190	SUM190		
SUM225	SUM225		
SUM44	SUM44		
SW156	SW156	SW156	1
SW403	SW403	SW403	1
SW620	SW620	SW620	1
SW948	SW948	SW948	1
T47D	T47D	T47D	1
T98G	T98G	T98G	1
TK-10	TK10	TK10	1
U-118 MG	U118MG	U118MG	1
U-251	U251	U251	1
U138	U138	U138MG	1
U87	U87	U87MG	1
UACC-257	UACC257	UACC257	1
UACC-62	UACC62	UACC62	1
UACC-732	UACC732		
UACC-812	UACC812	UACC812	1
UACC-893	UACC893	UACC893	1
UO-31	UO31	UO31	1
ZR-75-1	ZR751	ZR751	1
ZR-75-30	ZR7530	ZR7530	1

Table 7.3: Comparison of the performance to SPIA. The table lists whether regularized CCL labels were correctly found both by SPIA and Uniquorn (1 in indicator column)

ID	site	primary	site of met	grading	NEC/NET	sex	age	Staging	functionality	
ICGC_0425	pancreas	primary		G1	NET	Male	69	IB	No	73.78
ICGC_0427	pancreas	primary		G2	NET	Male	48	IB	Yes	59.28
ICGC_0428	pancreas	primary		G2	NET	Male	74	IB	No	3.35
ICGC_0431	pancreas	primary		G1	NET	Male	78	IIB	Yes	49.58
ICGC_0432	pancreas	primary		G1	NET	Male	76	IIB	No	19.73
ICGC_0433	pancreas	primary		G2	NET	Female	59	IIA	No	49.02
ICGC_0434	pancreas	primary		G1	NET	Male	59	IIA	No	0.82
ICGC_0435	pancreas	primary		G2	NET	Male	66	IIB	No	20.22
ICGC_0436	pancreas	primary		G1	NET	Male	61	IIB	No	38.89
ICGC_0437	pancreas	primary		G2	NET	Female	67	IB	Yes	37.51
ICGC_0438	pancreas	primary		G1	NET	Male	79	IV	No	33.9
ICGC_0440	pancreas	primary		G1	NET	Female	65	IB	Yes	32.19
ICGC_0441	pancreas	primary		G2	NET	Female	69	IB	No	33.67
ICGC_0443	pancreas	primary		G1	NET	Female	59	IB	No	24.79
ICGC_0447	pancreas	primary		G1	NET	Female	44	IA	No	26.96
ICGC_0449	pancreas	primary		G1	NET	Male	81	IA	No	22.68
ICGC_0452	pancreas	primary		G2	NET	Male	77	IIA	No	20.55
ICGC_0453	pancreas	primary		G1	NET	Male	56	IB	No	23.77
ICGC_0455	pancreas	primary		G3	NET (amb)	Male	78	IIB	No	18.18
ICGC_0456	pancreas	primary		G2	NET	Male	59	IB	No	18.67
ICGC_0457	pancreas	primary		G3	NET	Male	38	IV	No	6.87
ICGC_0459	pancreas	primary		G1	NET	Male	56	IA	Yes	12.16
ICGC_0489	pancreas	primary		G2	NET	Male	68	IB	No	14.1
ICGC_0491	pancreas	primary		G1	NET	Male	64	IA	No	15.85
ICGC_0492	pancreas	primary		G2	NET	Male	38	IIB	No	9.04
ICGC_0497	pancreas	primary		G2	NET	Male	71	IIB	No	7.96
ICGC_0498	pancreas	primary		G2	NET	Female	59	IIB	No	5.42
ICGC_0500	pancreas	primary		G1	NET	Female	75	IA	No	2.01
ICGC_0501	pancreas	primary		G2	NET	Male	64	IIA	No	1.32
1444	pancreas	primary		G3	NET	Male	72	IV	No	30
PNET06	pancreas	primary		G2	NET	Female	65	IIB	No	72
PNET37	pancreas	primary		G2	NET	Male	58	IB	No	157
1286	pancreas	primary		G3	NET (amb)*	Male	67	IV	No	34
135602	pancreas	primary		G3	NEC	Male	74	IIB	No	2
PNET17	pancreas	primary		G3	NET (amb) *	Female	73	IIB	No	98
PNET22	pancreas	primary		G3	NET (amb) *	Male	63	IIA	No	92
1401	pancreas	local recurrence		G3	NET	Female	54	IV	No	37
1418	pancreas	metastasis	lymph node	G2	NET	Female	32	IV	No	63
PNET04	pancreas	metastasis	liver	G3	NET (amb) *	Female	60	IV	Yes	61
PNET21	pancreas	metastasis	liver	G3	NET	Male	61	IV	missing	57
PNET41	pancreas	metastasis	liver	G2	NET	Male	42	IV	Yes	34
135604	pancreas	metastasis	lymph node	G3	NEC	Male	74	IV	No	2
139101	pancreas	metastasis	liver	G2	NET	Male	47	IV	No	65
PNET05	pancreas	metastasis	peritoneum	G3	NEC	Male	65	IV	No	1
PNET26	pancreas	metastasis	liver	G3	NET (amb) *	Male	64	IV	No	81
1344	pancreas	metastasis	liver	G3	NEC	Male	59	IV	No	1
105103	pancreas	metastasis	liver	G2	NET	Female	63	IV	No	31
130002	pancreas	metastasis	liver	G3	NET	Male	44	IV	No	57
130003	pancreas	metastasis	liver	G3	NET	Male	44	IV	No	57
PNET08	pancreas	metastasis	liver	G3	NEC	Female	38	IV	No	57
PNET55	pancreas	metastasis	liver	G3	NET	Male	44	IV	No	57
128802	stomach	primary		G3	NET (amb) *	Male	66	IV	No	3
140302	stomach	primary		G3	NEC	Male	71	III	missing	93
124101	stomach	metastasis	peritoneum	G3	NEC	Female	55	IV	No	74
124702	stomach	metastasis	liver	G3	NET (amb) *	Female	71	IV	No	100
132502	stomach	metastasis	liver	G3	NET (amb) *	Male	61	IV	Yes	92
IC15	small intestine	primary		G2	NET	Female	49	IV	No	124
145702	small intestine	metastasis	liver	G2	NET	Female	75	IV	No	59
110202	small intestine	metastasis	liver	G2	NET	Female	32	IV	missing	9
127402	small intestine	metastasis	liver	G3	NET	Male	59	IV	Yes	135
148402	small intestine	metastasis	other	G2	NET	Male	48	IV	No	34
IC02	small intestine	metastasis	lymph node	G3	NEC (amb)*	Female	68	IV	No	2
127403	small intestine	metastasis	liver	G3	NET	Male	59	IV	Yes	135
121103	colon	rectum	primary	G3	NEC (amb)*	Male	58	IIIB	missing	missing
136901	colon	rectum	primary	G3	NEC	Male	32	IIB	missing	missing
INET17	colon	rectum	primary	G3	NEC (amb)*	Female	61	IV	No	1
123402	colon	rectum	primary		G3	NEC	Male	57	IIIB	missing
141901	colon	rectum	metastasis	G3	NEC	Female	70	IV	No	14

Table 7.4: Cohort composition Chapter 5. The samples included in the RepSet are shown. For a full listing of all samples, please consult the corresponding publication.

Query	Training	Cor Ductal grading	Cor Ductal grading p-value	Cor HISC grading	Cor HISC grading p-value	Cor Ductal Ki-67	Cor Ductal Ki-67 p-value	Cor HISC Ki-67	Cor Ductal Ki-67 p-value	Ductal G1 versus G2**	Ductal G1 versus G3**	Ductal G2 versus G3**	HISC G1 versus G2**	HISC G1 versus G3**	HISC G2 versus G3**	Surv Ki-67 2-arm	Surv Ductal 2-arm	Surv HISC 2-arm	Surv grading 2-arm	Surv Ki-67 3-arm	Surv Ductal 3-arm	Surv HISC 3-arm	Surv grading 3-arm
Califano	Baron	NA		0.341	0.004	0.21	0.005	NA															
	Segerstolpe			0.302	0.0017	0.33	7.0E-4																
	Lawlor					0.416	1.0E-5	0.24	0.01														
Missaglia	Baron	0.329	0.004	0.122	0.297	0.163	0.14	-0.05	0.35	0.0064	0.36	0.98	0.081	0.805	0.24	NA							
	Segerstolpe	0.396	4.0E-4	-0.209	0.072	0.21	0.07	0.31	0.007	1.97E-4	0.33	0.83	0.27	0.45	0.89								
	Lawlor	0.496	6.0E-6	0.223	0.055	0.334	0.003	0.0	0.98	1.21E-5	0.03	0.99	0.056	0.82	0.86								
Sadanandam	Baron	0.667	1.0E-4	0.013	0.948	0.538	0.004	-0.15	0.05	0.97	4.0E-5	4.0E-6	0.87	0.9	0.90								
	Segerstolpe	0.038	0.852	0.115	0.568	0.093	0.646	0.4	0.04	0.25	0.99	0.18	0.75	0.85	0.38								
	Lawlor	0.562	0.001	0.17	0.388	0.57	0.406	0.45	0.02	0.32	3.0E-4	0.0019	0.55	0.68	0.12								
RepSet	Baron	0.654	1.0E-9	0.524	0.0	0.464	6.0E-5	0.48	0.02	0.033	2.0E-8	5.0E-5	0.10	3.0E-5	0.0082	0.014	0.0066	0.58	5.4E-4	0.0026	0.019	0.79	0.0022
	Segerstolpe	0.631	6.0E-9	0.488	2.15E-5	0.316	0.008	0.18	0.1	0.016	6.0E-8	6.0E-4	0.014	5.0E-5	0.15		0.063	0.11		0.17	0.29		
	Lawlor	0.655	1.0E-9	0.524	3.77E-6	0.365	0.002	0.05	0.7	0.022	2.0E-8	8.0E-5	0.10	3.0E-5	0.008		0.040	2.27E-4		0.049	0.0010		
Scarpa***	Baron	0.562	0.001	0.632	2.0E-4	0.57	0.001	0.61	2.0E-5	0.32	3.0E-4	0.0019	0.30	2.0E-7	2.0E-6	0.014	0.0020	0.0020	0.035	0.049	0.0086	0.0086	0.008
	Segerstolpe	0.578	0.001	0.693	3.03E-5	0.646	2.0E-4	0.59	7.0E-4	0.44	2.0E-5	1.0E-4	0.038	4.0E-6	1.0E-4		0.15	0.0020		0.049	0.0086		
	Lawlor	0.594	0.001	0.488	0.007	0.542	0.002	0.47	0.01	0.30	4.0E-5	3.0E-4	0.65	9.85E-4	0.0032		0.039	0.039		0.049	0.11		
Riemer	Baron	0.29	0.043	0.122	0.461	0.405	0.00749	0.32	0.04	NA		0.0022	NA		0.46	0.0035	0.045	0.38	4.9E-4	0.0035	0.045	0.38	4.9E-4
	Segerstolpe	0.151	0.359	0.028	0.864	0.132	0.42	0.33	0.04			0.0416			0.86		0.53	0.20		0.53	0.20		
	Lawlor	0.237	0.147	0.211	0.197	0.117	0.47	0.25	0.12			0.1071			0.19		0.25	0.082		0.25	0.082		
Fadista	Baron	NA				0.078	0.469	0.1	0.36	NA													
	Segerstolpe			0.053	0.621	-0.43	3.0E-5																
	Lawlor					0.126	0.099	-0.21	0.04														

Table 7.5: CIBERSORT benchmark results for varying algorithms and training datasets are shown. P-values smaller than 0.05 are significant. A p-value of less than 0.05 was considered significant. * NA no grading information is available. ** ANOVA p-value of the test on euqality of distribution means. Three and two arm design test statistics identical due to limitation to two grading classes in the Riemer dataset

Query	Training	Cor Ductal grading	Cor Ductal grading p-value	Cor HISC grading	Cor HISC grading p-value	Cor Ductal Ki-67	Cor Ductal Ki-67 p-value	Cor HISC Ki-67	Cor Ductal Ki-67 p-value	Ductal G1 versus G2**	Ductal G1 versus G3**	Ductal G2 versus G3**	HISC G1 versus G2**	HISC G1 versus G3**	HISC G2 versus G3**	Surv Ki-67 2-arm	Surv Ductal 2-arm	Surv HISC 2-arm	Surv grading 2-arm	Surv Ki-67 3-arm	Surv Ductal 3-arm	Surv HISC 3-arm	Surv grading 3-arm
Califano	Baron	NA	0.073	0.458	-0.028	0.774	NA																
Segerstolpe	0.12	0.223	0.196	0.0451																			
Lawlor	0.128	0.193	0.188	0.0542																			
Missaglia	Baron	0.261	0.0239	-0.008	0.944	0.196	0.0913	0.143	0.22	0.0126	0.833	0.709	0.912	0.866	0.773	NA							
Segerstolpe	0.19	0.102	-0.158	0.176	-0.028	0.811	-0.136	0.245	0.0462	0.998	0.542	0.466	0.657	0.951									
Lawlor	0.331	0.0037	0.047	0.687	0.147	0.208	-0.057	0.624	0.0037	0.449	0.928	0.924	0.972	0.999									
Sadanandam	Baron	0.255	0.198	0.025	0.902	0.334	0.0885	0.263	0.185	0.954	0.431	0.214	7E-4	0.0017	1								
Segerstolpe	-0.475	0.0122	0.049	0.807	-0.548	0.0031	0.248	0.213	0.651	0.0359	0.116	0.439	0.951	0.611									
Lawlor	-0.587	0.0013	0.068	0.737	-0.52	0.0055	0.108	0.593	0.148	0.0045	0.135	0.296	0.907	0.515									
RepSet	Baron	0.579	1.92E-7	0.29	0.0157	0.564	4.5E-7	-0.17	0.163	1.4E-6	0.0041	0.389	0.389	0.044	0.436	0.0143	0.33	0.854	5E-4	0.0026	0.0303	0.85	2E-3
Segerstolpe	0.44	2E-5	0.441	1E-4	0.27	0.0276	0.456	8.2E-5	0.0955	6E-5	0.118	0.373	0.001	0.0217	0.484	0.925	0.781	0.555					
Lawlor	0.7	2.25E-011	0.607	3.16E-8	0.405	6E-4	0.434	2E-4	0.0109	4.43E-10	6.8E-6	0.0323	3.23E-7	9E-4	0.0158	0.59	0.033	0.46					
Scarpa***	Baron	0.493	0.0065	0.57	0.0012	0.408	0.0278	0.115	0.552	0.14	0.0386	0.266	0.235	4E-4	0.004	0.0143	0.202	0.0736	0.035	0.049	0.444	0.202	9E-3
Segerstolpe	0.48	0.0085	0.549	0.002	0.385	0.0391	0.347	0.0648	0.14	0.0561	0.346	0.373	4E-5	0.0025	0.202	0.04	0.577	0.119					
Lawlor	0.422	0.0225	0.509	0.0048	0.54	0.0025	0.453	0.0137	0.7	0.011	0.0297	0.273	0.0071	0.0469	0.0736	0.0393	0.202	0.119					
Riemer	Baron	0.476	2E-4	0.184	0.263	0.349	0.0292	0.11	0.496	NA	2E-4	NA	0.27	0.0036	0.98	0.815	5E-4	4E-3	0.98	0.815	5E-5		
Segerstolpe	0.328	0.0416	0.197	0.228	0.073	0.659	0.421	8E-4	0.0416	0.228	0.502	0.403	0.402	0.403									
Lawlor	0.262	0.107	0.238	0.145	0.209	0.201	0.312	0.0531	0.107	0.145	0.0473	0.78	0.0473	0.78									
Fadista	Baron	NA	-0.149	0.164	-0.068	0.524	NA																
Segerstolpe	0.026	0.81	-0.024	0.821																			
Lawlor	0.078	0.469	-0.062	0.56																			

Table 7.6: MuSiC benchmark results for varying algorithms and training datasets are shown. P-values smaller than 0.05 are significant. A p-value of less than 0.05 was considered significant. * NA no grading information is available. ** ANOVA p-value of the test on euqality of distribution means. Three and two arm design test statistics identical due to limitation to two grading classes in the Riemer dataset

Query	Training	Cor Ductal grading	Cor Ductal grading p-value	Cor HISC grading	Cor HISC grading p-value	Cor Ductal Ki-67	Cor Ductal Ki-67 p-value	Cor HISC Ki-67	Cor Ductal Ki-67 p-value	Ductal G1 versus G2**	Ductal G1 versus G3**	Ductal G2 versus G3**	HISC G1 versus G2**	HISC G1 versus G3**	HISC G2 versus G3**	Surv Ki-67 2-arm	Surv Ductal 2-arm	Surv HISC 2-arm	Surv grading 2-arm	Surv Ki-67 3-arm	Surv Ductal 3-arm	Surv HISC 3-arm	Surv grading 3-arm
Califano	Baron	NA	-0.121	0.217	-0.109	0.27	NA																
Segerstolpe	-0.083	0.402	-0.366	1E-4																			
Lawlor	-0.016	0.872	0.004	0.969																			
Missaglia	Baron	-0.286	0.0128	-0.093	0.425	-0.152	0.192	-0.049	0.679	0.00249	0.9	0.47	0.857	0.77	0.906	NA							
Segerstolpe	-0.106	0.364	-0.215	0.0635	0.105	0.371	-0.232	0.0449	0.136	0.823	0.312	0.541	0.166	0.407									
Lawlor	-0.023	0.842	-0.217	0.0616	0.129	0.27	-0.322	0.00488	0.97	0.998	0.999	0.379	0.272	0.65									
Sadanandam	Baron	0.221	0.268	0.443	0.0205	0.38	0.0505	0.112	0.579	0.981	0.541	0.572	0.515	0.0588	0.272								
Segerstolpe	0.399	0.0393	0.313	0.111	0.588	0.00126	0.066	0.743	0.762	0.107	0.239	0.187	0.224	0.999									
Lawlor	0.143	0.477	-0.168	0.401	0.115	0.567	-0.107	0.596	0.596	0.736	0.984	0.243	0.625	0.786									
RepSet	Baron	-0.496	1.45E-5	0.165	0.176	-0.221	0.068	0.106	0.385	0.002	1.48E-5	0.339	0.552	0.343	0.933	0.0143	0.00563	0.794	5E-4	3E-3	0.0178	0.155	0.0022
Segerstolpe	-0.463	6.14E-5	-0.401	6E-4	0.131	0.282	-0.188	0.122	6E-4	3E-5	0.734	0.808	0.0103	2E-3	0.0562	0.0711	0.0466	0.658					
Lawlor	0.1	0.413	-0.312	0.00897	0.184	0.13	-0.085	0.486	0.871	0.688	0.936	0.848	0.0792	5E-3	0.206	0.541	0.0721	0.05					
Scarpa***	Baron	-0.405	0.0292	0.112	0.563	-0.287	0.131	0.13	0.502	0.179	0.21	0.706	0.819	0.948	1	0.0143	0.114	0.202	0.0358	0.0498	0.157	0.444	9E-2
Segerstolpe	-0.452	0.0137	-0.134	0.489	-0.261	0.171	-0.267	0.162	0.0826	0.172	0.761	0.44	0.0546	0.0136	0.114	0.0736	0.287	0.386					
Lawlor	-0.014	0.943	0.224	0.243	0.004	0.984	0.209	0.277	0.85	0.8	0.637	0.502	0.751	0.988	0.202	0.34	0.202	0.577					
Riemer	Baron	-0.004	0.981	-0.054	0.74	0.065	0.694	-0.125	0.45	NA	0.981	NA	0.743	4E-3	0.0428	0.299	5E-4	0.0036	0.198	0.299	5E-4		
Segerstolpe	0.056	0.734	-0.264	0.105	0.177	0.282	-0.371	0.0201	0.734	0.105	0.616	0.348	0.144	0.34									
Lawlor	0.189	0.248	-0.204	0.212	-0.014	0.934	-0.12	0.465	0.248	0.212	0.35	0.21	0.042	0.21									
Fadista	Baron	NA	-0.164	0.124	-0.027	0.803	NA																
Segerstolpe	-0.089	0.41	0.131	0.0896																			
Lawlor	-0.04	0.712	0.127	0.0978																			

Table 7.7: Moffitt et al. NMF benchmark results for varying algorithms and training datasets are shown. P-values smaller than 0.05 are significant. A p-value of less than 0.05 was considered significant. * NA no grading information is available. ** ANOVA p-value of the test on euqality of distribution means. Three and two arm design test statistics identical due to limitation to two grading classes in the Riemer dataset

Sample	Dataset	Model	P_value	Grading
425	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.001	G1
431	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.04	G1
432	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.014	G1
434	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.008	G1
436	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.007	G1
438	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.002	G1
440	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.0032	G1
443	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.009	G1
453	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.004	G1
459	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.0020054	G1
491	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.006	G1
427	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.007	G2
428	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.004	G2
433	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.003	G2
435	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.005	G2
437	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.019	G2
441	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G2
456	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.007	G2
489	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.001	G2
492	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.004	G2
497	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.011	G2
498	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.003	G2
501	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.006	G2
110202	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.02	G2
139101	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G2
148402	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.02	G2
IC15	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.02	G2
PNET06	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.02	G2
PNET37	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G2
PNET41	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G2
455	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.005	G3
457	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.003	G3
1286	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
1401	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
1418	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.096	G3
1444	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.096	G3
121103	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
123402	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
124101	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
124702	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
127402	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.096	G3
128802	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.096	G3
132502	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.096	G3
135602	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
135604	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
136901	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
140302	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.096	G3
141901	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
145502	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
IC02	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
INET17	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
PNET04	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0.096	G3
PNET05	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
PNET17	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
PNET21	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
PNET22	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3
PNET26	RepSet	Alpha_Beta_Gamma_Delta_Acinar_Ductal_Baron	0	G3

Table 7.8: P-values RepSet. For a complete list of p-values, please consult the publication documents.

Feature	RepSet	Riemer	Scarpa	Missiaglia	Sadanandam
Ki-67	3E-4	0.01	0.004	7E-017	0.17
Ductal	7E-8	0.04	0.001	0.004	1E-4
HISC	0.007	0.46	2E-4	0.3	0.95

Table 7.9: Correlations between cell type proportions and Ki-67.Pearson product moment correlation-derived p-values associated with the correlation of (M)Ki-67 levels and ductal and HISC fractions with the clinical grading. The (M)Ki-67 levels were significantly correlated with the grading in four out of five data sets with exception of the predominantly low to medium-grade Scarpa et al. data set. The relative ductal proportion were significantly correlated in all five data sets, always showing a significance with more statistical power to predict grading than MKi-67 mRNA counts but worse performance than Ki-67 staining level based grading as illustrated by the Missiaglia data set which only provided Ki-67 staining levels. Significant correlations are formatted in the scientific numbering scheme whereas insignificant correlations are presented as decimal numbers.

Query	Califano	Missaglia*	Sadanandam	RepSet	Scarpa	Riemer	Fadista
Dif_Exp_GLM_Cor	0.28	0.24	-0.38	0.42	0.27	0.43	0.17
Dif_Exp_GLM_P_value	0.004	0.35	0.05	0.001	0.16	0.002	0.11
Convolution_P_Value*	0.0041	0.14	0.0038	5E-4	4E-5	0.0074	0.47
Comparison	Comparable	Comparable	Supperior	Supperior	Supperior	Comparable	NA

Table 7.10: Deconvolution-trained Machine Learning performance. Positive class = NECs. * Ki-67 staining intensity and not mRNA levels utilized

Dataset	Unsupervised	Supervised			
Accuracy%	87	85			
Sensitivity%	84	81			
Specificity%	91	88			
PPV%	85	80			
Kappa	0.75	0.69			

Table 7.11: Expression and Ki-67-trained Machine Learning performance. Positive class = NECs. Logistic regression trained on deconvolution data of 57 NENs from the RepSet. ** Average obtained my averaging over the statistics, not a common model

Dataset	Unsupervised	Supervised
Accuracy%	87	85
Sensitivity%	84	81
Specificity%	91	88
PPV%	85	80
Kappa	0.75	0.69

Table 7.12: Positive class = NECs.Logistic regression trained on deconvolution data of 57 NENs from the RepSet. ** Average obtained my averaging over the statistics, not a common model

Query	Califano	Missaglia*	Sadanandam	RepSet	Scarpa	Riemer	Fadista
Dif-Exp GLM Correlation	0.28	0.24	-0.38	0.42	0.27	0.43	0.17
Dif-Exp GLM p-value	0.004	0.35	0.05	0.001	0.16	0.002	0.11
Convolution p-value	0.004	0.14	0.004	5E-4	4E-5	0.007	0.47
Comparison	Comparable	Comparable	Supperior	Supperior	Supperior	Comparable	NA

Table 7.13: Differential Expression Out-Group Benchmark. Values taken from CIBERSORT table 1, ductal correlation. * Ki-67 staining intensity and not mRNA levels utilized

Chapter 8

Bibliography

[1] Eric S.; Linton Lander et al. “Initial sequencing and analysis of the human genome”. In: *Nature* 409.February (2001), pp. 860–921. URL: www.nature.com.

[2] Leroy Hood and Lee Rowen. “The human genome project: big science transforms biology and medicine”. In: *Genome Medicine* 5.9 (2013), p. 79. ISSN: 1756-994X. DOI: 10.1186/gm483. URL: http://genomemedicine.biomedcentral.com/articles/10.1186/gm483.

[3] Daniel C. Koboldt et al. “The Next-Generation Sequencing Revolution and Its Impact on Genomics”. In: *Cell* 155.1 (Sept. 2013), pp. 27–38. ISSN: 00928674. DOI: 10.1016/j.cell.2013.09.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867413011410.

[4] Jordi Barretina et al. “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity”. In: *Nature* 483.7391 (Mar. 2012), pp. 603–607. ISSN: 0028-0836. DOI: 10.1038/nature11003. URL: http://www.ncbi.nlm.nih.gov/pubmed/22460905%7B%5C%%7D0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3320027%20http://www.nature.com/articles/nature11003.

[5] Maxime Gaudin and Christelle Desnues. “Hybrid Capture-Based Next Generation Sequencing and Its Application to Human Infectious Diseases”. In: *Frontiers in Microbiology* 9.NOV (Nov. 2018), pp. 1–9. ISSN: 1664-302X. DOI: 10.3389/fmicb.2018.02924. URL: https://www.frontiersin.org/article/10.3389/fmicb.2018.02924/full.

[6] Ryoji Amamoto et al. “Probe-Seq enables transcriptional profiling of specific cell types from heterogeneous tissue by RNA-based isolation”. In: *eLife* 8 (Dec. 2019). ISSN: 2050-084X. DOI: 10.7554/eLife.51452. URL: https://elifesciences.org/articles/51452.

[7] Chien Yueh Lee et al. “Common applications of next-generation sequencing technologies in genomic research”. In: *Translational Cancer Research* 2.1 (2013), pp. 33–45. ISSN: 22196803. DOI: 10.3978/j.issn.2218-676X.2013.02.09. URL: https://tcr.amegroups.com/article/view/962/html.

[8] J R Masters. “Human cancer cell lines: fact and fantasy.” In: *Nature reviews. Molecular cell biology* 1.3 (2000), pp. 233–236. ISSN: 1471-0072. DOI: 10.1038/35043102. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed%7B%5C&%7Did=11252900%7B%5C&%7Dretmode=ref%7B%5C&%7Dcmd=prlinks.

[9] Felipe Castro et al. “High-throughput SNP-based authentication of human cell lines”. In: *International Journal of Cancer* 132.2 (Jan. 2013), pp. 308–314. ISSN: 00207136. DOI: 10.1002/ijc.27675. arXiv: NIHMS150003. URL: http://www.ncbi.nlm.nih.gov/pubmed/21959306%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3259195%20http://doi.wiley.com/10.1002/ijc.27675.

[10] Francesca Demichelis et al. “SNP panel identification assay (SPIA): A genetic-based assay for the identification of cell lines”. In: *Nucleic Acids Research* 36.7 (Apr. 2008), pp. 2446–2456. ISSN: 03051048. DOI: 10.1093/nar/gkn089. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2367734%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[11] Sophie Zaaijer et al. “Rapid re-identification of human samples using portable DNA sequencing”. In: *eLife* 6 (2017), e27798. ISSN: 2050-084X. DOI: 10.7554/eLife.27798. URL: https://elifesciences.org/articles/27798.

[12] Stefan J. Vermeulen et al. “Did the four human cancer cell lines DLD-1, HCT-15, HCT-8, and HRT-18 originate from one and the same patient?” In: *Cancer Genetics and Cytogenetics* 107.1 (1998), pp. 76–79. ISSN: 01654608. DOI: 10.1016/S0165-4608(98)00081-8.

[13] Teresa K Attwood et al. “A global perspective on evolving bioinformatics and data science training needs”. In: *Briefings in Bioinformatics* 20.2 (Mar. 2019), pp. 398–404. ISSN: 1477-4054. DOI: 10.1093/bib/bbx100. URL: https://academic.oup.com/bib/article/20/2/398/4096809.

[14] Tri Dao et al. “A Kernel Theory of Modern Data Augmentation.” In: *Proceedings of machine learning research* 97.3 (June 2019), pp. 1528–1537. ISSN: 2640-3498. URL: http://www.ncbi.nlm.nih.gov/pubmed/31777848%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6879382.

[15] Jenni A. M. Sidey-Gibbons and Chris J. Sidey-Gibbons. “Machine learning in medicine: a practical introduction”. In: *BMC Medical Research Methodology* 19.1 (Dec. 2019), p. 64. ISSN: 1471-2288. DOI: 10.1186/s12874-019-0681-4. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0681-4.

[16] Kentaro Yamada et al. ““Dip-and-read” paper-based analytical devices using distance-based detection with color screening”. In: *Lab on a Chip* 18.10 (2018), pp. 1485–1493. ISSN: 1473-0197. DOI: 10.1039/C8LC00168E. URL: http://xlink.rsc.org/?DOI=C8LC00168E.

[17] Markus Riester et al. “PureCN: copy number calling and SNV classification using targeted short read sequencing”. In: *Source Code for Biology and Medicine* 11.1 (2016), p. 13. ISSN: 1751-0473. DOI: 10.1186/s13029-016-0060-z. URL: http://scfbm.biomedcentral.com/articles/10.1186/s13029-016-0060-z.

[18] Raik Otto et al. “Robust ¡i¿In-Silico¡/i¿ identification of cancer cell lines based on next generation sequencing”. In: *Oncotarget* (Mar. 2017), pp. 1–11. ISSN: 1949-2553. DOI: 10.18632/oncotarget.16110. URL: http://www.oncotarget.com/abstract/16110.

[19] Raik Otto et al. “Robust in-silico identification of Cancer Cell Lines based on RNA and targeted DNA sequencing data”. In: *Scientific Reports* 9.1 (Dec. 2019), p. 367. ISSN: 2045-2322. DOI: 10.1038/s41598-018-36300-8. URL: http://www.nature.com/articles/s41598-018-36300-8.

[20] Katharina Schwarze et al. “Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature”. In: *Genetics in Medicine* 20.10 (Oct. 2018), pp. 1122–1130. ISSN: 1098-3600. DOI: 10.1038/gim.2017.247. URL: http://www.nature.com/articles/gim2017247.

[21] Amalio Telenti et al. “Deep sequencing of 10,000 human genomes”. In: *Proceedings of the National Academy of Sciences* 113.42 (2016), pp. 11901–11906. ISSN: 0027-8424. DOI: 10.1073/pnas.1613365113. arXiv: 061663. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1613365113.

[22] Andreas Halman and Alicia Oshlack. “Accuracy of short tandem repeats genotyping tools in whole exome sequencing data”. In: *F1000Research* 9 (Mar. 2020), p. 200. ISSN: 2046-1402. DOI: 10.12688/f1000research.22639.1. URL: https://f1000research.com/articles/9-200/v1.

[23] Christian Gilissen et al. “Genome sequencing identifies major causes of severe intellectual disability”. In: *Nature* 511.7509 (July 2014), pp. 344–347. ISSN: 0028-0836. DOI: 10.1038/nature13394. URL: http://www.nature.com/articles/nature13394.

[24] Mihaela Pertea et al. “Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise”. In: *bioRxiv* (2018), p. 332825. DOI: 10.1101/332825. URL: https://www.biorxiv.org/content/10.1101/332825v2.

[25] Sarah B. Ng et al. “Targeted capture and massively parallel sequencing of 12 human exomes”. In: *Nature* 461.7261 (Sept. 2009), pp. 272–276. ISSN: 0028-0836. DOI: 10.1038/nature08250. URL: http://www.nature.com/articles/nature08250.

[26] “Molecular cell biology, by James Darnell, Harvey Lodish and David Baltimore; Scientific American Books, W. H. Freeman & Co., New York, NY, 1986, 1186 pages, $42.95”. In: *Gamete Research* 17.1 (May 1987), pp. 95–95. ISSN: 0148-7280. DOI: 10.1002/mrd.1120170110. URL: http://doi.wiley.com/10.1002/mrd.1120170110.

[27] Katharina Schwarze et al. “The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom”. In: *Genetics in Medicine* 22.1 (Jan. 2020), pp. 85–94. ISSN: 1098-3600. DOI: 10.1038/s41436-019-0618-7. URL: http://dx.doi.org/10.1038/s41436-019-0618-7%20http://www.nature.com/articles/s41436-019-0618-7.

[28] Xuran Wang et al. “Bulk tissue cell type deconvolution with multi-subject single-cell expression reference”. In: *Nature Communications* 10.1 (Dec. 2019), p. 380. ISSN: 2041-1723. DOI: 10.1038/s41467-018-08023-x. URL: http://www.nature.com/articles/s41467-018-08023-x.

[29] Lora J. H. Bean et al. “Diagnostic gene sequencing panels: from design to report—a technical standard of the American College of Medical Genetics and Genomics (ACMG)”. In: *Genetics in Medicine* 22.3 (Mar. 2020), pp. 453–461. ISSN: 1098-3600. DOI: 10.1038/s41436-019-0666-z. URL: http://dx.doi.org/10.1038/s41436-019-0666-z%20http://www.nature.com/articles/s41436-019-0666-z.

[30] Yan Guo et al. “The effect of strand bias in Illumina short-read sequencing data”. In: *BMC Genomics* 13.1 (2012), pp. 1–11. ISSN: 14712164. DOI: 10.1186/1471-2164-13-666.

[31] Kimberly R. Kukurba and Stephen B. Montgomery. “RNA Sequencing and Analysis”. In: *Cold Spring Harbor Protocols* 2015.11 (Nov. 2015), pdb.top084970. ISSN: 1940-3402. DOI: 10.1101/pdb.top084970. URL: http://www.cshprotocols.org/lookup/doi/10.1101/pdb.top084970.

[32] Li Ding et al. “Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing”. In: *Nature* 481.7382 (Jan. 2012), pp. 506–510. ISSN: 0028-0836. DOI: 10.1038/nature10738. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3267864%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract%7B%5C%%7D5Cnhttp://www.nature.com/

doifinder/10.1038/nature10738%20http://www.nature.com/doifinder/10.1038/nature10738.

[33] Travis W. Murphy et al. “Microfluidic Platform for Next-Generation Sequencing Library Preparation with Low-Input Samples.” In: *Analytical chemistry* 92.3 (Feb. 2020), pp. 2519–2526. ISSN: 1520-6882. DOI: 10.1021/acs.analchem.9b04086. URL: https://pubs.acs.org/doi/abs/10.1021/acs.analchem.9b04086%20http://www.ncbi.nlm.nih.gov/pubmed/31894965%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7002211.

[34] Aaron M Newman et al. “Robust enumeration of cell subsets from tissue expression profiles”. In: *Nature Methods* 12.5 (May 2015), pp. 453–457. ISSN: 1548-7091. DOI: 10.1038/nmeth.3337. arXiv: 15334406. URL: http://www.nature.com/articles/nmeth.3337%20http://www.ncbi.nlm.nih.gov/pubmed/25822800%7B%5C%%7D0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4739640.

[35] Meichen Dong et al. “SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references”. In: *Briefings in Bioinformatics* 00.September 2019 (Jan. 2020), pp. 1–12. ISSN: 1477-4054. DOI: 10.1093/bib/bbz166. URL: https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbz166/5699815.

[36] Can Alkan et al. “Genome structural variation discovery and genotyping.” In: *Nature reviews. Genetics* 12.5 (May 2011), pp. 363–76. ISSN: 1471-0064. DOI: 10.1038/nrg2958. URL: http://www.ncbi.nlm.nih.gov/pubmed/21358748%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4108431.

[37] Karunakaran Kaliyappan et al. “Microarray and its applications”. In: *Journal of Pharmacy and Bioallied Sciences* 4.6 (2012), p. 310. ISSN: 0975-7406. DOI: 10.4103/0975-7406.100283. URL: http://www.jpbsonline.org/text.asp?2012/4/6/310/100283.

[38] Shanrong Zhao et al. “Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells”. In: *PLoS ONE* 9.1 (Jan. 2014). Ed. by Shu-Dong Zhang, e78644. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0078644. URL: https://dx.plos.org/10.1371/journal.pone.0078644.

[39] Christoph Ziegenhain et al. “Comparative Analysis of Single-Cell RNA Sequencing Methods”. In: *Molecular Cell* 65.4 (Feb. 2017), 631–643.e4. ISSN: 10972765. DOI: 10.1016/j.molcel.2017.01.023. URL: http://biorxiv.org/lookup/doi/10.1101/035758%20http://www.ncbi.nlm.nih.gov/pubmed/28212749%20http://linkinghub.elsevier.com/retrieve/pii/S1097276517300497.

[40] Byungjin Hwang et al. “Single-cell RNA sequencing technologies and bioinformatics pipelines”. In: *Experimental & Molecular Medicine* 50.8 (Aug. 2018), p. 96. ISSN: 2092-6413. DOI: 10.1038/s12276-018-0071-8. URL: http://dx.doi.org/10.1038/s12276-018-0071-8%20http://www.nature.com/articles/s12276-018-0071-8.

[41] Yuchen Yang et al. “SAFE-clustering: Single-cell Aggregated (from Ensemble) clustering for single-cell RNA-seq data”. In: *Bioinformatics* 35.8 (Apr. 2019). Ed. by Inanc Birol, pp. 1269–1277. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty793. URL: https://academic.oup.com/bioinformatics/article/35/8/1269/5092931.

[42] Dominic Grün et al. "De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data". In: *Cell Stem Cell* 19.2 (Aug. 2016), pp. 266–277. ISSN: 19345909. DOI: 10.1016/j.stem.2016.05.010. arXiv: 1512.00567. URL: http://linkinghub.elsevier.com/retrieve/pii/S1934590916300947.

[43] Xiuli An and Lixiang Chen. "Flow Cytometry (FCM) Analysis and Fluorescence-Activated Cell Sorting (FACS) of Erythroid Cells". In: *Erythropoiesis: Methods and Protocols*. Ed. by Joyce A. Lloyd. New York, NY: Springer New York, 2018, pp. 153–174. ISBN: 978-1-4939-7428-3. DOI: 10.1007/978-1-4939-7428-3{_}9. URL: https://doi.org/10.1007/978-1-4939-7428-3_9.

[44] Hajk-Georg Drost. "Philentropy: Information Theory and Distance Quantification with R". In: *Journal of Open Source Software* 3.26 (2018), p. 765. ISSN: 2475-9066. DOI: 10.21105/joss.00765.

[45] Massoud Malek-Shahmirzadi. "A characterization of certain classes of matrix norms". In: *Linear and Multilinear Algebra* 13.2 (May 1983), pp. 97–99. ISSN: 0308-1087. DOI: 10.1080/03081088308817508. URL: http://www.tandfonline.com/doi/abs/10.1080/03081088308817508.

[46] Li Wang et al. "A reference profile-free deconvolution method to infer cancer cell-intrinsic subtypes and tumor-type-specific stromal profiles". In: *Genome Medicine* 12.1 (Dec. 2020), p. 24. ISSN: 1756-994X. DOI: 10.1186/s13073-020-0720-0. URL: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-0720-0.

[47] Chase T. Gerold et al. "Selective Distance-Based K + Quantification on Paper-Based Microfluidics". In: *Analytical Chemistry* 90.7 (Apr. 2018), pp. 4894–4900. ISSN: 0003-2700. DOI: 10.1021/acs.analchem.8b00559. URL: https://pubs.acs.org/doi/10.1021/acs.analchem.8b00559.

[48] Si Chen et al. "Phylogenetic tree construction using trinucleotide usage profile (TUP)". In: *BMC Bioinformatics* 17.Suppl 13 (2016). ISSN: 14712105. DOI: 10.1186/s12859-016-1222-3. URL: http://dx.doi.org/10.1186/s12859-016-1222-3.

[49] Enrico Tiacci et al. "BRAF Mutations in Hairy-Cell Leukemia". In: *New England Journal of Medicine* 364.24 (June 2011), pp. 2305–2315. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1014209. URL: http://www.nejm.org/doi/abs/10.1056/NEJMoa1014209.

[50] Eduard Cech and Michael Katetov. *Point Sets*. 1969, pp. 9–12.

[51] Marie Labelle and Daniel Valois. "Functional categories and the acquisition of distance quantification". In: January 2004 (2004), pp. 27–49. DOI: 10.1075/lald.32.04lab.

[52] M Lange et al. "Applications of l p -Norms and their Smooth Approximations for Gradient Based Learning Vector Quantization". In: *Esann* April (2014), pp. 23–25.

[53] G. Y. Chen et al. "A nonlinear scalarization function and generalized quasi-vector equilibrium problems". In: *Journal of Global Optimization* 32.4 (2005), pp. 451–466. ISSN: 09255001. DOI: 10.1007/s10898-003-2683-2.

[54] Ziyi Chen et al. "Inference of immune cell composition on the expression profiles of mouse tissue". In: *Scientific Reports* 7.1 (Feb. 2017), p. 40508. ISSN: 2045-2322. DOI: 10.1038/srep40508. URL: http://www.nature.com/articles/srep40508.

[55] C. Alan Boneau. "The effects of violations of assumptions underlying the t test." In: *Psychological Bulletin* 57.1 (1960), pp. 49–64. ISSN: 0033-2909. DOI: 10.1037/h0041412. URL: http://content.apa.org/journals/bul/57/1/49.

[56] D. R Bellhouse. "The Central Limit Theorem Under Simple Random Sampling". In: *The American Statistician* 55.4 (Nov. 2001), pp. 352–357. ISSN: 0003-1305. DOI: 10.1198/000313001753272330. URL: http://www.tandfonline.com/doi/abs/10.1198/000313001753272330.

[57] L. E. Clarke and Patrick Billingsley. "Probability and Measure". In: *The Mathematical Gazette* 64.430 (Dec. 1980), p. 293. ISSN: 00255572. DOI: 10.2307/3616746. URL: https://www.jstor.org/stable/3616746?origin=crossref.

[58] Georg Pólya. "Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem". In: *Mathematische Zeitschrift* 8.3-4 (Sept. 1920), pp. 171–181. ISSN: 0025-5874. DOI: 10.1007/BF01206525. URL: http://link.springer.com/10.1007/BF01206525.

[59] Thomas Lumley et al. "The Importance of the Normality Assumption in Large Public Health Data Sets". In: *Annual Review of Public Health* 23.1 (May 2002), pp. 151–169. ISSN: 0163-7525. DOI: 10.1146/annurev.publhealth.23.100901.140546. URL: http://www.annualreviews.org/doi/10.1146/annurev.publhealth.23.100901.140546.

[60] B. Derrick and P. White. "Why Welch's test is Type I error robust". In: *The Quantitative Methods for Psychology* 12.1 (Jan. 2016), pp. 30–38. ISSN: 2292-1354. DOI: 10.20982/tqmp.12.1.p030. URL: http://www.tqmp.org/RegularArticles/vol12-1/p030.

[61] Kam-Fai Wong et al. "Forward selection two sample binomial test." In: *Journal of data science : JDS* 12.4 (Oct. 2014), pp. 279–294. ISSN: 1680-743X. URL: http://www.ncbi.nlm.nih.gov/pubmed/27335577%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4914133.

[62] Richard Peto and Julian Peto. "Asymptotically Efficient Rank Invariant Test Procedures". In: *Journal of the Royal Statistical Society. Series A (General)* 135.2 (1972), p. 185. ISSN: 00359238. DOI: 10.2307/2344317. URL: https://www.jstor.org/stable/10.2307/2344317?origin=crossref.

[63] Benjamin S. Blanchard. *Logistics Engineering and Management*. 4th. New Jersey: Prentice-Hall: Englewood Cliffs, 1992, pp. 26–32.

[64] Yinglei Lai. "Conservative adjustment of permutation p-values when the number of permutations is limited". In: *International Journal of Bioinformatics Research and Applications* 3.4 (2007), p. 536. ISSN: 1744-5485. DOI: 10.1504/ijbra.2007.015420.

[65] Ronald Fricker. *The American Statistician;* vol. 50. Alexandria, 1996, pp. 278–279.

[66] A. Dvoretzky et al. "Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator". In: *The Annals of Mathematical Statistics* 27.3 (Sept. 1956), pp. 642–669. ISSN: 0003-4851. DOI: 10.1214/aoms/1177728174. URL: http://projecteuclid.org/euclid.aoms/1177728174.

[67] *Dvoretzky-Kiefer-Wolfowitz inequality*. https://upload.wikimedia.org/wikipedia/commons/9/90/DKW_bounds.svg. Accessed: 2020-02-24.

[68] Dirk P. Kroese et al. "Why the Monte Carlo method is so important today". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 6.6 (Nov. 2014), pp. 386–392. ISSN: 19395108. DOI: 10.1002/wics.1314. URL: http://doi.wiley.com/10.1002/wics.1314.

[69] Rick Routledge. "Fisher's Exact Test". In: *Encyclopedia of Biostatistics*. Chichester, UK: John Wiley & Sons, Ltd, July 2005. DOI: 10.1002/0470011815.b2a10020. URL: http://doi.wiley.com/10.1002/0470011815.b2a10020.

[70] IBM. “Exact Tests”. In: *IBM Knowledge Center* (2018). URL: https://www.ibm.com/support/knowledgecenter/en/SSLVMB%7B%5C_%7D23.0.0/spss/base/idh%7B%5C_%7Dexact.html.

[71] P. H. Westfall et al. “On Adjusting P-Values for Multiplicity”. In: *Biometrics* 49.3 (Sept. 1993), p. 941. ISSN: 0006341X. DOI: 10.2307/2532216. URL: https://www.jstor.org/stable/2532216?origin=crossref.

[72] Yoav Benajmini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing Author (s): Yoav Benjamini and Yosef Hochberg Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 57 , No . 1 Published by :” in: *Journal of the Royal Statistical Society. Series B* 57.1 (1995), pp. 289–300.

[73] Sudhir Paul and You-gan Wang. “A Review of the Behrens-Fisher Problem and Some of Its Analogs: Does the same size fit all?” In: (2000).

[74] Resknik. “Gibbs Sampler Tutorial”. In: June (2010).

[75] B V North et al. “A note on the calculation of empirical P values from Monte Carlo procedures.” In: *American journal of human genetics* 71.2 (Aug. 2002), pp. 439–41. ISSN: 0002-9297. DOI: 10.1086/341527. URL: http://www.ncbi.nlm.nih.gov/pubmed/12111669%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC379178.

[76] Belinda Phipson and Gordon K. Smyth. “Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn”. In: *Statistical Applications in Genetics and Molecular Biology* 9.1 (2010), pp. 1–12. ISSN: 15446115. DOI: 10.2202/1544-6115.1585. arXiv: arXiv:1603.05766v1.

[77] Kai Yu et al. “Efficient p-value evaluation for resampling-based tests”. In: *Biostatistics* 12.3 (2011), pp. 582–593. ISSN: 14654644. DOI: 10.1093/biostatistics/kxq078.

[78] De-jian LI et al. “Compressed sensing based deconvolution algorithm for time-domain UWB channel modeling”. In: *The Journal of China Universities of Posts and Telecommunications* 19.1 (Feb. 2012), pp. 62–68. ISSN: 10058885. DOI: 10.1016/S1005-8885(11)60229-X. URL: https://linkinghub.elsevier.com/retrieve/pii/S100588851160229X.

[79] Quartl. *Matrix multiplication qtl1.svg*. Online; accessed February 12th, 2021. 2021. URL: https://commons.wikimedia.org/wiki/File:Matrix_multiplication_qtl1.svg.

[80] Emmanuel Vincent et al. “From Blind to Guided Audio Source Separation: How models and side information can improve the separation of sound”. In: *IEEE Signal Processing Magazine* 31.3 (May 2014), pp. 107–115. ISSN: 1053-5888. DOI: 10.1109/MSP.2013.2297440. URL: http://ieeexplore.ieee.org/document/6784053/.

[81] Shai S. Shen-Orr et al. “Cell type-specific gene expression differences in complex tissues”. In: *Nature Methods* 7.4 (Apr. 2010), pp. 287–289. ISSN: 1548-7091. DOI: 10.1038/nmeth.1439. URL: http://dx.doi.org/10.1038/nmeth.1439%20http://www.nature.com/articles/nmeth.1439.

[82] Francisco Avila Cobos et al. “Computational deconvolution of transcriptomics data from mixed cell populations”. In: *Bioinformatics* 34.January (Jan. 2018). Ed. by Jonathan Wren, pp. 1–11. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty019. URL: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty019/4813737.

[83] Alexander R. Abbas et al. "Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus". In: *PLoS ONE* 4.7 (July 2009). Ed. by Patrick Tan, e6098. DOI: 10.1371/journal.pone.0006098. URL: http://dx.plos.org/10.1371/journal.pone.0006098.

[84] Konstantina Dimitrakopoulou et al. "Deblender: a semi?/unsupervised multi-operational computational method for complete deconvolution of expression data from heterogeneous samples". In: *BMC Bioinformatics* 5 (2018), pp. 1–17. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2442-5. URL: https://link.springer.com/article/10.1186/s12859-018-2442-5?utm%7B%5C_%7Dsource=researcher%7B%5C_%7Dapp%7B%5C&%7Dutm%7B%5C_%7Dmedium=referral%7B%5C&%7Dutm%7B%5C_%7Dcampaign=MKEF%7B%5C_%7DUSG%7B%5C_%7DResearcher%7B%5C_%7Dinbound.

[85] Joseph D Szustakowski. "Package 'DeconRNASeq' Title Deconvolution of Heterogeneous Tissue Samples for mRNA-Seq data". In: (2018), pp. 1–6. URL: https://git.bioconductor.org/packages/DeconRNASeq.

[86] Shahin Mohammadi et al. "A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues". In: *Proceedings of the IEEE* 105.2 (Feb. 2017), pp. 340–366. ISSN: 0018-9219. DOI: 10.1109/JPROC.2016.2607121. arXiv: 1510.04583. URL: http://arxiv.org/abs/1510.04583%20http://dx.doi.org/10.1109/JPROC.2016.2607121%20http://ieeexplore.ieee.org/document/7676285/.

[87] Zuoli Dong et al. "Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection." In: *BMC cancer* 15.1 (2015), p. 489. ISSN: 1471-2407. DOI: 10.1186/s12885-015-1492-6. URL: http://www.biomedcentral.com/1471-2407/15/489.

[88] Niya Wang et al. "UNDO: A Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples". In: *Bioinformatics* 31.1 (2015), pp. 137–139. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu607.

[89] Carsten F. Dormann et al. "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance". In: *Ecography* 36.1 (Jan. 2013), pp. 27–46. ISSN: 09067590. DOI: 10.1111/j.1600-0587.2012.07348.x. URL: http://doi.wiley.com/10.1111/j.1600-0587.2012.07348.x.

[90] Aylin Alin. "Multicollinearity". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.3 (May 2010), pp. 370–374. ISSN: 19395108. DOI: 10.1002/wics.84. URL: http://doi.wiley.com/10.1002/wics.84.

[91] Chris Chatfield. "Model Uncertainty, Data Mining and Statistical Inference". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 158.3 (1995), p. 419. ISSN: 09641998. DOI: 10.2307/2983440. URL: https://www.jstor.org/stable/2983440.

[92] Martin Slawski and Matthias Hein. "Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization". In: *Electronic Journal of Statistics* 7.1 (2013), pp. 3004–3056. ISSN: 1935-7524. DOI: 10.1214/13-EJS868. arXiv: 1205.0953. URL: http://projecteuclid.org/euclid.ejs/1386943911.

[93] N.H. Farhat. "Photonic neural networks and learning machines". In: *IEEE Expert* 7.5 (Oct. 1992), pp. 63–72. ISSN: 0885-9000. DOI: 10.1109/64.163674. URL: http://ieeexplore.ieee.org/document/163674/.

[94] Bernhard Schölkopf et al. "New Support Vector Algorithms". In: *Neural Computation* 12.5 (May 2000), pp. 1207–1245. ISSN: 0899-7667. DOI: 10.1162/089976600300015565. URL: http://www.mitpressjournals.org/doi/10.1162/089976600300015565.

[95] Lahrmam. *SVM Margin*. Online; accessed November 13, 2020. 2018. URL: https://commons.wikimedia.org/wiki/File:SVM_margin.png.

[96] Chi Jin and Liwei Wang. “Dimensionality dependent PAC-Bayes margin bound”. In: *Advances in Neural Information Processing Systems* 2 (2012), pp. 1034–1042. ISSN: 10495258.

[97] Vapnik VN Boser B, Guyon I. “A Training Algorithm for Optimal Margin Classifiers”. In: *Lecture Notes in Computer Science* (1992).

[98] Alex J. Smola and Bernhard Scholkopf. “A tutorial on support vector regression”. In: *Statistics and Computing* 14 (2004), pp. 199–222. ISSN: 0960-3174. arXiv: arXiv:1011.1669v3. URL: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=1CAD92EF8CCE726A305D8A41F873EEFC?doi=10.1.1.114.4288%7B%5C&%7Drep=rep1%7B%5C&%7Dtype=pdf%7B%5C%%7D0Ahttp://download.springer.com/static/pdf/493/art%7B%5C%%7D3A10.1023%7B%5C%%7D2FB%7B%5C%%7D3ASTCO.0000035301.49549.88.pdf?auth66=1408162706%7B%5C_%7D8a28764ed0fae9.

[99] Chih-Chung Chang and Chih-Jen Lin. “Training v -Support Vector Classifiers: Theory and Algorithms”. In: *Neural Computation* 13.9 (2001), pp. 2119–2147. ISSN: 0899-7667. DOI: 10.1162/089976601750399335. URL: http://www.mitpressjournals.org/doi/abs/10.1162/089976601750399335.

[100] Harris Drucker et al. “Support vector regression machines”. In: *Advances in Neural Information Processing Dystems* 1 (1997), pp. 155–161. ISSN: 10495258. DOI: 10.1.1.10.4845. URL: http://papers.nips.cc/paper/1238-support-vector-regression-machines.pdf.

[101] Morgan A. Hanson. “Invexity and the Kuhn-Tucker theorem”. In: *Journal of Mathematical Analysis and Applications* 236.2 (Aug. 1999), pp. 594–604. ISSN: 0022247X. DOI: 10.1006/jmaa.1999.6484. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022247X99964843.

[102] Debasish Basak et al. “Support vector regression”. In: *Neural Information Processing Letters and Reviews*. 2007, pp. 203–224.

[103] Emma Villeneuve. “Hyperspectral data deconvolution for galaxy kinematics with MCMC”. In: Eusipco (2012), pp. 2477–2481.

[104] Inderjit S. Dhillon and Suvrit Sra. “Generalized nonnegative matrix approximations with Bregman divergences”. In: *Advances in Neural Information Processing Systems* (2005), pp. 283–290. ISSN: 10495258.

[105] Rashish Tandon and Suvrit Sra. “Sparse nonnegative matrix approximation: new formulations and algorithms”. In: *Tech Report No. 193, Max-Planck* 193 (2010), p. 19. URL: http://www.kyb.tuebingen.mpg.de/techreports.html%7B%5C%%7D0Ahttp://www.kyb.mpg.de/publications/attachments/MPIK-TR-193%7B%5C%%7D5C%7B%5C%%7D5C%7B%5C_%7D[0].pdf.

[106] Zhao Li et al. “Nonnegative Matrix Factorization on Orthogonal Subspace”. In: *Pattern Recognition Letters* 31.9 (July 2010), pp. 905–911. ISSN: 01678655. DOI: 10.1016/j.patrec.2009.12.023. URL: http://dx.doi.org/10.1016/j.patrec.2009.12.023%20http://linkinghub.elsevier.com/retrieve/pii/S0167865509003651.

[107] Ludmil B. Alexandrov et al. “Signatures of mutational processes in human cancer”. In: *Nature* 500.7463 (Aug. 2013), pp. 415–421. ISSN: 0028-0836. DOI: 10.1038/nature12477. URL: http://dx.doi.org/10.1038/nature12477%2010.1038/nature12477%20http://www.nature.com/nature/journal/v500/n7463/abs/nature12477.html%7B%5C#%7Dsupplementary-information%20http://www.nature.com/doifinder/10.1038/nature12477.

[108] Xitong Yang. “Understanding the Variational Lower Bound”. In: *variational lower bound, ELBO, hard attention* (2017), pp. 1–4. ISSN: 0928-4931.

[109] Robert R. Sokal and F. James Rohlf. “the Comparison of Dendrograms By Objective Methods”. In: *Taxon* 11.2 (1962), pp. 33–40. ISSN: 0040-0262. DOI: 10.2307/1217208.

[110] Daniel D. Lee and H. Sebastian Seung. “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401.6755 (Oct. 1999), pp. 788–791. ISSN: 0028-0836. DOI: 10.1038/44565. URL: http://www.nature.com/articles/44565.

[111] Michael W. Berry et al. “Algorithms and applications for approximate nonnegative matrix factorization”. In: *Computational Statistics & Data Analysis* 52.1 (Sept. 2007), pp. 155–173. ISSN: 01679473. DOI: 10.1016/j.csda.2006.11.006. URL: http://linkinghub.elsevier.com/retrieve/pii/S0167947306004191.

[112] D. Daniel Lee and H. Sebastian Seung. “Algorithms for Non-negative Matrix”. In: *Nips* 13 (2001), pp. 556–562. ISSN: 10987576. DOI: 10.1109/IJCNN.2008.4634046. arXiv: 0408058 [cs].

[113] Ludmil B. Alexandrov et al. “A mutational signature in gastric cancer suggests therapeutic strategies”. In: *Nature Communications* 6.OCTOBER (2015), p. 8683. ISSN: 2041-1723. DOI: 10.1038/ncomms9683. URL: http://www.nature.com/doifinder/10.1038/ncomms9683.

[114] Friedman J. Hastie T. Tibshirani R. *The Elements of Statistical Learning*. 12th. New york: Springer, 2017, pp. 553–554.

[115] Daniel D. Lee and H. Sebastian Seung. “Algorithms for Non-negative Matrix Factorization”. In: *Advances in Neural Information Processing Systems 13*. Ed. by T. K. Leen et al. MIT Press, 2001, pp. 556–562. URL: http://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf.

[116] Richard A Moffitt et al. “Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma”. In: *Nature Genetics* 47.10 (Oct. 2015), pp. 1168–1178. ISSN: 1061-4036. DOI: 10.1038/ng.3398. URL: http://www.nature.com/articles/ng.3398.

[117] Juan José Burred. “Detailed derivation of multiplicative update rules for NMF”. In: March (2014), pp. 1–8.

[118] Amanda Capes-Davis et al. “Check your cultures! A list of cross-contaminated or misidentified cell lines”. In: *International Journal of Cancer* 127.1 (2010), pp. 1–8. ISSN: 00207136. DOI: 10.1002/ijc.25242.

[119] Philip L Lorenzi et al. “DNA fingerprinting of the NCI-60 cell line panel.” In: *Molecular cancer therapeutics* 8.4 (Apr. 2009), pp. 713–24. ISSN: 1535-7163. DOI: 10.1158/1535-7163.MCT-08-0921. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4020356%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[120] Simon A. Forbes et al. “COSMIC: Somatic cancer genetics at high-resolution”. In: *Nucleic Acids Research* 45.D1 (2017), pp. D777–D783. ISSN: 13624962. DOI: 10.1093/nar/gkw1121.

[121] Hao Hu et al. “Evaluating information content of SNPs for sample-tagging in re-sequencing projects”. In: *Scientific Reports* 5 (2015), p. 10247. ISSN: 2045-2322. DOI: 10.1038/srep10247. URL: http://www.nature.com/doifinder/10.1038/srep10247.

[122] Janyaporn Phuchareon et al. “Genetic profiling reveals cross-contamination and misidentification of 6 adenoid cystic carcinoma cell lines: ACC2, ACC3, ACCM, ACCNS, ACCS and CAC2”. In: *PLoS ONE* 4.6 (2009), pp. 6–13. ISSN: 19326203. DOI: 10.1371/journal.pone.0006040.

[123] Mordechai Liscovitch and Dana Ravid. "A case study in misidentification of cancer cell lines: MCF-7/AdrR cells (re-designated NCI/ADR-RES) are derived from OVCAR-8 human ovarian carcinoma cells." In: *Cancer letters* 245.1-2 (Jan. 2007), pp. 350–2. ISSN: 0304-3835. DOI: 10.1016/j.canlet.2006.01.013. URL: http://www.ncbi.nlm.nih.gov/pubmed/16504380.

[124] James M Rae et al. "MDA-MB-435 cells are derived from M14 melanoma cells–a loss for breast cancer, but a boon for melanoma research." In: *Breast cancer research and treatment* 104.1 (July 2007), pp. 13–9. ISSN: 0167-6806. DOI: 10.1007/s10549-006-9392-8. URL: http://www.ncbi.nlm.nih.gov/pubmed/17004106.

[125] Roderick A.F. MacLeod et al. "Widespread intraspecies cross-contamination of human tumor cell lines arising at source". In: *International Journal of Cancer* 83.4 (Nov. 1999), pp. 555–563. ISSN: 0020-7136. DOI: 10.1002/(SICI)1097-0215(19991112)83:4<555::AID-IJC19>3.0.CO;2-2. URL: http://doi.wiley.com/10.1002/%7B%5C%%7D28SICI%7B%5C%%7D291097-0215%7B%5C%%7D2819991112%7B%5C%%7D2983%7B%5C%%7D3A4%7B%5C%%7D3C555%7B%5C%%7D3A%7B%5C%%7D3AAID-IJC19%7B%5C%%7D3E3.0.CO%7B%5C%%7D3B2-2.

[126] Walther Parson et al. "Cancer cell line identification by short tandem repeat profiling: power and limitations". In: *The FASEB Journal* 19.3 (Dec. 2004), pp. 434–436. ISSN: 0892-6638. DOI: 10.1096/fj.04-3062fje. URL: http://www.fasebj.org/cgi/doi/10.1096/fj.04-3062fje%20http://www.ncbi.nlm.nih.gov/pubmed/15637111.

[127] Andrew M Hudson et al. "Discrepancies in Cancer Genomic Sequencing Highlight Opportunities for Driver Mutation Discovery". In: *Cancer Research* 74.22 (2014), pp. 6390–6396. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-14-1020. URL: http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-14-1020.

[128] Jian Li et al. "An NGS workflow blueprint for DNA sequencing data and its application in individualized molecular oncology". In: *Cancer Informatics* 15 (2016), pp. 87–107. ISSN: 11769351. DOI: 10.4137/CIN.S30793.

[129] Bairoch A. *The Cellosaurus: a cell line knowledge resource*. URL: https://web.expasy.org/cellosaurus/.

[130] Levi A. Garraway et al. "Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma". In: *Nature* 436.7047 (2005), pp. 117–122. ISSN: 00280836. DOI: 10.1038/nature03664.

[131] David M. Altshuler et al. "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491.7422 (2012), pp. 56–65. ISSN: 14764687. DOI: 10.1038/nature11632.

[132] Jacob A Tennessen et al. "Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes." In: *Science* 64.May (2012), pp. 1–8. ISSN: 10959203. DOI: 10.1126/science.1219240. URL: http://www.ncbi.nlm.nih.gov/pubmed/22604720.

[133] David Jones et al. "cgpCaVEManWrapper: Simple execution of caveman in order to detect somatic single nucleotide variants in NGS data". In: *Current Protocols in Bioinformatics* 2016.December (2016), pp. 15.10.1–15.10.18. ISSN: 1934340X. DOI: 10.1002/cpbi.20.

[134] Kai Ye et al. "Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads". In: *Bioinformatics* 25.21 (2009), pp. 2865–2871. ISSN: 13674803. DOI: 10.1093/bioinformatics/btp394. arXiv: NIHMS150003.

[135] Kristian Cibulskis et al. "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples". In: *Nature Biotechnology* 31.3 (2013), pp. 213–219. ISSN: 1087-0156. DOI: 10.1038/nbt.2514. arXiv: NIHMS150003. URL: http://www.nature.com/nbt/journal/v31/n3/abs/nbt.2514.html%7B%5C%%7D5Cnhttp://www.nature.com/nbt/journal/v31/n3/pdf/nbt.2514.pdf.

[136] A. McKenna et al. "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data". In: *Genome Research* 20.9 (Sept. 2010), pp. 1297–1303. ISSN: 1088-9051. DOI: 10.1101/gr.107524.110. arXiv: arXiv:1011.1669v3. URL: http://genome.cshlp.org/cgi/doi/10.1101/gr.107524.110.

[137] Huaiyu Mi et al. "Large-scale gene function analysis with the panther classification system". In: *Nature Protocols* 8.8 (2013), pp. 1551–1566. ISSN: 17502799. DOI: 10.1038/nprot.2013.092.

[138] P.-H. Tseng et al. "Overcoming Trastuzumab Resistance in HER2-Overexpressing Breast Cancer Cells by Using a Novel Celecoxib-Derived Phosphoinositide-Dependent Kinase-1 Inhibitor". In: *Molecular Pharmacology* 70.5 (Aug. 2006), pp. 1534–1541. ISSN: 0026-895X. DOI: 10.1124/mol.106.023911. URL: http://molpharm.aspetjournals.org/cgi/doi/10.1124/mol.106.023911.

[139] P Martin and T Papayannopoulou. "HEL cells: a new human erythroleukemia cell line with spontaneous and induced globin expression". In: *Science* 216.4551 (June 1982), pp. 1233–1235. ISSN: 0036-8075. DOI: 10.1126/science.6177045. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.6177045.

[140] Tobias Marschall et al. *Computational Pan-Genomics: Status, Promises and Challenges*. Tech. rep. Mar. 2016. DOI: 10.1101/043430. URL: http://biorxiv.org/lookup/doi/10.1101/043430.

[141] Alexander Dobin et al. "STAR: Ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts635. arXiv: 1201.0052.

[142] Geraldine A Van Der Auwera et al. *From FastQ data to high confidence varant calls: the Genonme Analysis Toolkit best practices pipeline*. Vol. 11. 1110. 2014. DOI: 10.1002/0471250953.bi1110s43.From.

[143] Anthony M. Bolger et al. "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (Aug. 2014), pp. 2114–2120. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btu170. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu170.

[144] Erik Garrison and Gabor Marth. "Haplotype-based variant detection from short-read sequencing". In: (July 2012), pp. 1–9. arXiv: 1207.3907. URL: http://arxiv.org/abs/1207.3907.

[145] Amanda Capes-Davis et al. "Match criteria for human cell line authentication: Where do we draw the line?" In: *International Journal of Cancer* 132.11 (2013), pp. 2510–2519. ISSN: 00207136. DOI: 10.1002/ijc.27931.

[146] Aaron R. Quinlan and Ira M. Hall. "BEDTools: A flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (2010), pp. 841–842. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq033.

[147] Junjun Zhang et al. "International cancer genome consortium data portal-a one-stop shop for cancer genomics data". In: *Database* 2011 (2011), pp. 1–10. ISSN: 17580463. DOI: 10.1093/database/bar026.

[148] Damian Smedley et al. "The BioMart community portal: An innovative alternative to large, centralized data repositories". In: *Nucleic Acids Research* 43.W1 (2015), W589–W598. DOI: 10.1093/nar/gkv350.

[149] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019. URL: https://www.R-project.org/.

[150] Christiaan Klijn et al. "A comprehensive transcriptional portrait of human cancer cell lines." In: *Nature biotechnology* 33.3 (Mar. 2015), pp. 306–12. ISSN: 1546-1696. DOI: 10.1038/nbt.3080. URL: http://www.ncbi.nlm.nih.gov/pubmed/25485619.

[151] Grossman et al. "Toward a Shared Vision for Cancer Genomic Data". In: *The New England journal of medicine* 375.12 (2016), pp. 1109–1112. ISSN: 15334406. DOI: 10.1056/NEJMp1002530. arXiv: arXiv:1011.1669v3. URL: http://scholar.google.com/scholar?hl=en%7B%5C&%7DbtnG=Search%7B%5C&%7Dq=intitle:New+engla+nd+journal%7B%5C#%7D0.

[152] John G. Tate et al. "COSMIC: The Catalogue Of Somatic Mutations In Cancer". In: *Nucleic Acids Research* 47.D1 (2019), pp. D941–D947. ISSN: 13624962. DOI: 10.1093/nar/gky1015.

[153] Uma T. Shankavaram et al. "CellMiner: A relational database and query tool for the NCI-60 cancer cell lines". In: *BMC Genomics* 10 (2009), pp. 1–10. ISSN: 14712164. DOI: 10.1186/1471-2164-10-277.

[154] William C. Reinhold et al. "CellMiner: A web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set". In: *Cancer Research* 72.14 (2012), pp. 3499–3511. ISSN: 00085472. DOI: 10.1158/0008-5472.CAN-12-1370.

[155] Shahram Azari et al. "Profiling and authentication of human cell lines using short tandem repeat (STR) loci: Report from the National Cell Bank of Iran." In: *Biologicals : journal of the International Association of Biological Standardization* 35.3 (June 2007), pp. 195–202. ISSN: 1045-1056. DOI: 10.1016/j.biologicals.2006.10.001. URL: http://www.ncbi.nlm.nih.gov/pubmed/17254797.

[156] Ignasi Morán et al. "Human β Cell Transcriptome Analysis Uncovers lncRNAs That Are Tissue-Specific, Dynamically Regulated, and Abnormally Expressed in Type 2 Diabetes". In: *Cell Metabolism* 16.4 (Oct. 2012), pp. 435–448. ISSN: 15504131. DOI: 10.1016/j.cmet.2012.08.010. URL: http://linkinghub.elsevier.com/retrieve/pii/S1550413112003610.

[157] Mariantonia Di Sanzo et al. "Clinical Applications of Personalized Medicine: A New Paradigm and Challenge". In: *Current Pharmaceutical Biotechnology* 18.3 (Apr. 2017), pp. 194–203. ISSN: 13892010. DOI: 10.2174/1389201018666170224105600. URL: http://www.eurekaselect.com/openurl/content.php?genre=article%7B%5C&%7Dissn=1389-2010%7B%5C&%7Dvolume=18%7B%5C&%7Dissue=3%7B%5C&%7Dspage=194.

[158] Arvind Dasari et al. "Trends in the Incidence, Prevalence, and Survival Outcomes in Patients With Neuroendocrine Tumors in the United States". In: *JAMA Oncology* 3.10 (Oct. 2017), p. 1335. ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2017.0589. URL: http://oncology.jamanetwork.com/article.aspx?doi=10.1001/jamaoncol.2017.0589.

[159] H. Sorbye et al. "Predictive and prognostic factors for treatment and survival in 305 patients with advanced gastrointestinal neuroendocrine carcinoma (WHO G3): The NORDIC NEC study". In: *Annals of Oncology* 24.1 (2013), pp. 152–160. ISSN: 09237534. DOI: 10.1093/annonc/mds276.

[160] Massimo Milione et al. "The Clinicopathologic Heterogeneity of Grade 3 Gastroenteropancreatic Neuroendocrine Neoplasms: Morphological Differentiation and Proliferation Identify Different Prognostic Categories". In: *Neuroendocrinology* 104.1 (2017), pp. 85–93. ISSN: 0028-3835. DOI: 10.1159/000445165. URL: https://www.karger.com/Article/FullText/445165.

[161] Arvind Dasari et al. “Comparative study of lung and extrapulmonary poorly differentiated neuroendocrine carcinomas: A SEER database analysis of 162,983 cases”. In: *Cancer* 124.4 (2018), pp. 807–815. ISSN: 10970142. DOI: 10.1002/cncr.31124.

[162] C. Lepage et al. “Endocrine Tumours: Epidemiology of malignant digestive neuroendocrine tumours”. In: *European Journal of Endocrinology* 168.4 (Apr. 2013), R77–R83. ISSN: 0804-4643. DOI: 10.1530/EJE-12-0418. URL: https://eje.bioscientifica.com/view/journals/eje/168/4/R77.xml.

[163] A. Sadanandam et al. “A Cross-Species Analysis in Pancreatic Neuroendocrine Tumors Reveals Molecular Subtypes with Distinctive Clinical, Metastatic, Developmental, and Metabolic Characteristics”. In: *Cancer Discovery* 5.12 (Dec. 2015), pp. 1296–1313. ISSN: 2159-8274. DOI: 10.1158/2159-8290.CD-15-0068. URL: http://cancerdiscovery.aacrjournals.org/cgi/doi/10.1158/2159-8290.CD-15-0068.

[164] AO Oladejo. “Gastroenteropancreatic Neuroendocrine Tumors (GEP-NETs) - approach to diagnosis and managment”. In: *Annals of Ibadan Postgraduate Medicine* 7.2 (2011), pp. 29–33. ISSN: 1597-1627. DOI: 10.4314/aipm.v7i2.64085.

[165] Olca Basturk et al. “The high-grade (WHO G3) pancreatic neuroendocrine tumor category is morphologically and biologically heterogenous and includes both well differentiated and poorly differentiated neoplasms”. In: *American Journal of Surgical Pathology* 39.5 (2015), pp. 683–690. ISSN: 15320979. DOI: 10.1097/PAS.0000000000000408.

[166] L. H. Tang et al. “Well-Differentiated Neuroendocrine Tumors with a Morphologically Apparent High-Grade Component: A Pathway Distinct from Poorly Differentiated Neuroendocrine Carcinomas”. In: *Clinical Cancer Research* 22.4 (Feb. 2016), pp. 1011–1017. ISSN: 1078-0432. DOI: 10.1158/1078-0432.CCR-15-0548. URL: http://clincancerres.aacrjournals.org/cgi/doi/10.1158/1078-0432.CCR-15-0548.

[167] Fleur Broekman. “Tyrosine kinase inhibitors: Multi-targeted or single-targeted?” In: *World Journal of Clinical Oncology* 2.2 (2011), p. 80. ISSN: 2218-4333. DOI: 10.5306/wjco.v2.i2.80. URL: http://www.wjgnet.com/2218-4333/full/v2/i2/80.htm.

[168] Gökcen Eraslan et al. “Deep learning: new computational modelling techniques for genomics”. In: *Nature Reviews Genetics* (Apr. 2019). ISSN: 1471-0056. DOI: 10.1038/s41576-019-0122-6. URL: http://www.nature.com/articles/s41576-019-0122-6.

[169] Gwang Ha Kim et al. “Learning models for endoscopic ultrasonography in gastrointestinal endoscopy”. In: *World Journal of Gastroenterology* 21.17 (2015), pp. 5176–5182. ISSN: 22192840. DOI: 10.3748/wjg.v21.i17.5176.

[170] Michael J. Nasse and Jörg C. Woehl. “Realistic modeling of the illumination point spread function in confocal scanning optical microscopy”. In: *Journal of the Optical Society of America A* 27.2 (Feb. 2010), p. 295. ISSN: 1084-7529. DOI: 10.1364/JOSAA.27.000295. URL: https://www.osapublishing.org/abstract.cfm?URI=josaa-27-2-295.

[171] Kiarash Ahi and Mehdi Anwar. “Developing terahertz imaging equation and enhancement of the resolution of terahertz images using deconvolution”. In: ed. by Mehdi F. Anwar et al. May 2016, 98560N. DOI: 10.1117/12.2228680. URL: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2228680.

[172] Chengqi Xu et al. “Algebraic analysis of the Van Cittert iterative method of deconvolution with a general relaxation factor”. In: *Journal of the Optical Society of America A* 11.11 (Nov. 1994), p. 2804. ISSN: 1084-7529. DOI: 10.1364/JOSAA.11.002804. URL: https://www.osapublishing.org/abstract.cfm?URI=josaa-11-11-2804.

[173] Tanya Barrett et al. “NCBI GEO: Archive for functional genomics data sets - Update”. In: *Nucleic Acids Research* 41.D1 (2013), pp. 991–995. ISSN: 03051048. DOI: 10.1093/nar/gks1193.

[174] Adam L. Haber et al. “A single-cell survey of the small intestinal epithelium”. In: *Nature* 551.7680 (2017), pp. 333–339. ISSN: 14764687. DOI: 10.1038/nature24489. URL: http://dx.doi.org/10.1038/nature24489.

[175] Diana E. Stanescu et al. “Single cell transcriptomic profiling of mouse pancreatic progenitors.” In: *Physiological genomics* 49.2 (2017), pp. 105–114. ISSN: 1531-2267. DOI: 10.1152/physiolgenomics.00114.2016. URL: http://physiolgenomics.physiology.org/lookup/doi/10.1152/physiolgenomics.00114.2016%20http://www.ncbi.nlm.nih.gov/pubmed/28011883%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5336600.

[176] Liying Yan et al. “Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells”. In: *Nature Structural & Molecular Biology* 20.9 (Sept. 2013), pp. 1131–1139. ISSN: 1545-9993. DOI: 10.1038/nsmb.2660. URL: http://dx.doi.org/10.1038/nsmb.2660%20http://www.nature.com/articles/nsmb.2660.

[177] Ana Conesa et al. “A survey of best practices for RNA-seq data analysis”. In: *Genome Biology* 17.1 (2016), p. 13. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0881-8. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8.

[178] Maayan Baron et al. “A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure”. In: *Cell Systems* 3.4 (Oct. 2016), 346–360.e4. ISSN: 24054712. DOI: 10.1016/j.cels.2016.08.011. arXiv: 15334406. URL: http://linkinghub.elsevier.com/retrieve/pii/S2405471216302666%20http://dx.doi.org/10.1016/j.cels.2016.08.011.

[179] Mariano J Alvarez et al. “Functional characterization of somatic mutations in cancer using network-based inference of protein activity.” In: *Nature genetics* 48.8 (2016), pp. 838–47. ISSN: 1546-1718. DOI: 10.1038/ng.3593. URL: http://www.nature.com/doifinder/10.1038/ng.3593%7B%5C%%7D5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/27322546.

[180] J. Fadista et al. “Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism”. In: *Proceedings of the National Academy of Sciences* 111.38 (2014), pp. 13924–13929. ISSN: 0027-8424. DOI: 10.1073/pnas.1402665111.

[181] Nathan Lawlor et al. “Single-cell transcriptomes identify human islet cell signatures and reveal cell-type?specific expression changes in type 2 diabetes”. In: *Genome Research* 27.2 (Feb. 2017), pp. 208–222. ISSN: 1088-9051. DOI: 10.1101/gr.212720.116. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.212720.116.

[182] Edoardo Missiaglia et al. “Pancreatic Endocrine Tumors: Expression Profiling Evidences a Role for AKT-mTOR Pathway”. In: *Journal of Clinical Oncology* 28.2 (Jan. 2010), pp. 245–255. ISSN: 0732-183X. DOI: 10.1200/JCO.2008.21.5988. URL: http://ascopubs.org/doi/10.1200/JCO.2008.21.5988.

[183] Aldo Scarpa et al. "Whole-genome landscape of pancreatic neuroendocrine tumours". In: *Nature* 543.7643 (Feb. 2017), pp. 65–71. ISSN: 0028-0836. DOI: 10.1038/nature21063. URL: http://www.nature.com/doifinder/10.1038/nature21063.

[184] Åsa Segerstolpe et al. "Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes". In: *Cell Metabolism* 24.4 (Oct. 2016), pp. 593–607. ISSN: 15504131. DOI: 10.1016/j.cmet.2016.08.020. URL: http://linkinghub.elsevier.com/retrieve/pii/S1550413116304363.

[185] Kosuke Yoshihara et al. "Inferring tumour purity and stromal and immune cell admixture from expression data". In: *Nature Communications* 4.1 (Dec. 2013), p. 2612. ISSN: 2041-1723. DOI: 10.1038/ncomms3612. URL: http://www.nature.com/articles/ncomms3612.

[186] Valerie A. Schneider et al. "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly". In: *Genome Research* 27.5 (May 2017), pp. 849–864. ISSN: 1088-9051. DOI: 10.1101/gr.213611.116. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.213611.116.

[187] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17.1 (May 2011), p. 10. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200. URL: http://journal.embnet.org/index.php/embnetjournal/article/view/200.

[188] Nicolas L Bray et al. "Near-optimal probabilistic RNA-seq quantification". In: *Nature Biotechnology* 34.5 (May 2016), pp. 525–527. ISSN: 1087-0156. DOI: 10.1038/nbt.3519. URL: http://www.nature.com/articles/nbt.3519.

[189] Michael I Love et al. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (Dec. 2014), p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

[190] Steffen Durinck et al. "Mapping identfiers for integration of genomic datasets with the R/Bioconductor package biomaRt". In: 4.8 (2009), pp. 1184–1191. DOI: 10.1038/nprot.2009.97.Mapping.

[191] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0. 2019. URL: https://CRAN.R-project.org/package=stringr.

[192] TM. Therneau and P. Grambsch. *Modeling Survival Data: Extending the Cox Model*. 2000, pp. 1–25. ISBN: 0-387-98784-3.

[193] A. Subramanian et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (Oct. 2005), pp. 15545–15550. ISSN: 0027-8424. DOI: 10.1073/pnas.0506580102. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0506580102.

[194] Renaud Gaujoux and Cathal Seoighe. "Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study". In: *Infection, Genetics and Evolution* 12.5 (2012), pp. 913–921. ISSN: 15671348. DOI: 10.1016/j.meegid.2011.08.014. URL: http://dx.doi.org/10.1016/j.meegid.2011.08.014.

[195] Marylyn D. Ritchie et al. "Methods of integrating data to uncover genotype?phenotype interactions". In: *Nature Reviews Genetics* 16.2 (2015), pp. 85–97. ISSN: 1471-0056. DOI: 10.1038/nrg3868. URL: http://dx.doi.org/10.1038/nrg3868%7B%5C%%7D5Cnhttp://www.nature.com/doifinder/10.1038/nrg3868.

[196] Max Kuhn. "Building Predictive Models in R Using the caret Package". In: *Journal of Statistical Software, Articles* 28.5 (2008), pp. 1–26. ISSN: 1548-7660. DOI: 10.18637/jss.v028.i05. URL: https://www.jstatsoft.org/v028/i05.

[197] Alboukadel Kassambara et al. *survminer: Drawing Survival Curves using 'ggplot2'*. R package version 0.4.9. 2021. URL: https://CRAN.R-project.org/package=survminer.

[198] Björn Konukiewitz et al. "Pancreatic neuroendocrine carcinomas reveal a closer relationship to ductal adenocarcinomas than to neuroendocrine tumors G3." In: *Human pathology* 0.0 (2018), #pagerange#. ISSN: 1532-8392. DOI: 10.1016/j.humpath.2018.03.018. URL: http://www.ncbi.nlm.nih.gov/pubmed/29596894.

[199] Aref G. Ebrahimi et al. "Beta cell identity changes with mild hyperglycemia: Implications for function, growth, and vulnerability". In: *Molecular Metabolism* 35.February (May 2020), p. 100959. ISSN: 22128778. DOI: 10.1016/j.molmet.2020.02.002. URL: https://doi.org/10.1016/j.molmet.2020.02.002%20https://linkinghub.elsevier.com/retrieve/pii/S2212877820300314.

[200] Fong Cheng Pan et al. "Spatiotemporal patterns of multipotentiality in Ptf1a-expressing cells during pancreas organogenesis and injury-induced facultative restoration". In: *Development* 140.4 (Feb. 2013), pp. 751–764. ISSN: 0950-1991. DOI: 10.1242/dev.090159. URL: http://dev.biologists.org/cgi/doi/10.1242/dev.090159.

[201] Tincy Simon et al. "An Integrative Genetic, Epigenetic and Proteomic Characterization of Pancreatic Neuroendocrine Neoplasms (PanNENs) defines Distinct Molecular Features of alpha- and beta-cell like Subgroups". In: *bioRxiv* (2020). DOI: 10.1101/2020.06.12.146811. eprint: https://www.biorxiv.org/content/early/2020/06/12/2020.06.12.146811.full.pdf. URL: https://www.biorxiv.org/content/early/2020/06/12/2020.06.12.146811.

Chapter 9

List of Abbreviations

ANOVA ANalysis Of VAriance test

ARX Aristaless Related Homeobox

API Application Programming Interface

ATAC-seq Assay for Transposase-Accessible Chromatin using sequencing

ATCC American Type Culture Collection

AUC Area Under the Curve

BAM Binary sequence alignment file

BED Browser Extensible Data file

BRAF B-Raf Proto-Oncogene

BSeq-sc Bulk Sequence single-cell deconvolution analysis pipeline

CCL Cancer Cell Line

CCLE Cancer Cell Line Encyclopedia

CDF cumulative distribution function

cDNA complementary deoxyribonucleic acid

CEL-seq Cell Expression by Linear amplification and Sequencing

CGP Cancer Genome Project

ChIP-seq Chromatin ImmunoPrecipitation-sequencing

CIBERSORT Cell-type identification by estimating relative subsets of RNA transcripts

COSMIC CLP catalogue of somatic mutations in cancer Cancer Cell Line Project

CNA Copy Number Aberration

CNV Copy Number Variation

csSAM computes cell-specific differential expression from measured cell proportions using SAM

DACO Data Access Compliance Office

DNA deoxyribonucleic acid

DKW Dvoretzky–Kiefer–Wolfowitz inequality

e.g. exempli gratia

EGA European Genome-phenome Archive

ELBO Evidence lower bound

ESTIMATE Estimation of STromal and Immune cells in MAlignant Tumours using Expression data

FACS Fluorescence Activated Cell Sorting

FASTQ FAST-ALL Quality file

FISH Fluorescence in situ hybridization

FN False Negative

FP False Positive

FPR False Positive Rate

FWER Family-wise error rate

GATK Genome Analysis Toolkit

GEO Gene Expression Omnibus

GEP-NEC Gastroenteropancreatic Neuroendocrine Carcinomas

GEP-NEN Gastroenteropancreatic Neuroendocrine Neoplasm

GEP-NET Gastroenteropancreatic Neuroendocrine Tumor

GDC Genomic Data Commons

GEO Gene Expression Omnibus

GEP-NEN gastroenteropancreatic neuroendocrine neoplasm

GRC Genome Reference Consortium

GSEA Gene set enrichment analysis

GSOA Gene set over-representation analysis

HISC Human Intestinal Stem Cell

HGP Human Genome Project

IHS Immune Histo-Chemistry

i.e. id est

iff if and only if

KKT Karush-Kuhn-Tucker condition

KS Kolmogorov-Smirnoff

Ki-67 Proliferation marker protein Ki-67

LDR Linear Dimensionality Reduction

ICGC International Cancer Genome Consortium

i.i.d. independently and identically distributed

InDel Insertion and Deletion

MA mean-average

ML Machine-Learning

MAF Minor Allele Frequency

MCA Multiplex Cell Authentication

MKi-67 Marker Of Proliferation Ki-67

mRNA messenger Ribonucleic Acid

MuSiC MUlti-Subject SIngle Cell deconvolution

NCI-60 National Cancer Institute 60

NGS Next-Generation Sequencing

NEC Neuroendocrine Carcinoma

NEN Neuroendocrine Neoplasm

NET Neuroendocrine Tumor

NMF Non-negative Matrix Factorization

NNLS Non-negative Least Squares Factorization

PanNEC Pancreatic Neuroendocrine Carcinoma

PanNEN Pancreatic Neuroendocrine Neoplasm

PanNET Pancreatic Neuroendocrine Tumor

PC Principal-Component

PCA Principal Component Analysis

PDAC Pancreatic Ductal Adenocarcinoma

PCR polymerase chain reaction

PDF probability density function

PPV positive predictive value

RAM Random Access Memory

ROC Receiver on Operator Characteristic

RepSet Representative Set

RMSE Root mean square error

RNA Ribonucleic acid

rRNA ribosomal Ribonucleic acid

scRNA Single-cell sequencing

SNP Single-Nucleotide Polymorphism

SNV Single-Nucleotide Variant

SPIA SNP panel identification assay

STAR Spliced Transcripts Alignment to a Reference

STR Short Tandem Repeat

SVM Support Vector Machine

SVR Support Vector Machine Regression

TCGA The Cancer Genome Atlas Program

TKI Multi-targeted tyrosine kinase inhibitor

TIMER Tumor IMmune Estimation Resource

TN True Negative

TP True Positive

TP53 Tumor Protein 53

TPM Transcripts Per Million bases

UMI Unique Molecular Identifier

Uniquorn UNIQUe variant identification Of canceR cell liNes

VC Vapnik-Chervonenkis

VCF Variant Calling Format

WES Whole Exome-Sequencing

WGS Whole Genome-Sequencing

List of Figures

List of Tables

Chapter 10

Declaration of authorship and independence

Ich erkläre, dass ich die Dissertation selbständig und nur unter Verwendung der von mir gemäß § 7 Abs. 3 der Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät, veröffentlicht im Amtlichen Mitteilungsblatt der Humboldt-Universität zu Berlin Nr. 42/2018 am 11.07.2018 angegebenen Hilfsmittel angefertigt habe.

Ort, Datum, Unterschrift ....