Aus der Medizinischen Klinik V des Universitätsklinikums Heidelberg
(Ärztlicher Direktor: Prof. Dr. Carsten Müller-Tidow)
Labor für Myelomforschung
Leitung: Priv.-Doz. Dr. med. Dr. biol. hom. Dirk Hose

Transcriptome Profiling Assessing Pathogenesis and Prognosis of Plasma Cell Dyscrasias
–
Bioinformatic Basis for Clinical Application

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)
an der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
Martina Emde-Rajaratnam
aus
Brilon

2020

Dekan: Prof. Dr. med. Hans-Georg Kräusslich

Doktorvater: Priv.-Doz. Dr. med. Dr. biol. hom. Dirk Hose

Contents

List of Figures

List of Tables

List of Codes

Nomenclature

A	absent expression
AMM	asymptomatic multiple myeloma
BH	Benjamini-Hochberg method
BMME	bone marrow microenvironment
BMPC	bone marrow plasma cell
CAR	chimeric antigen receptor
CO	consistency
CoMMpass	relating clinical outcomes in multiple myeloma to personal assessment of genetic profile
CP	CoMMpass cohort
CRAB	hyper**c**alcaemia, **r**enal impairment, **a**neamia and **b**one disease
CTA	cancer testis antigen
docval	documentation by value
edgeR	empirical analysis of digital gene expression in R
EFS	event free survival
ENSG	Ensembl gene identifiers
FDR	false discovery rate
GEP-R	gene expression profiling report
GMMG	German speaking myeloma multicenter group
GPI	gene expression based proliferation index
GRCh	genome reference consortium human build
GTF	gene transfer format
HD	Heidelberg cohort
HDHRS	Heidelberg high risk score
HMCL	human myeloma cell line
HRD	hyperdiploidy
iFISH	interphase fluorescence *in-situ* hybridisation
Ig	immunoglobulin
IMiD	immune modulatory imide drug
IMWG	international myeloma working group
ISS	international staging system
JHT	Jonckheere-Terpstra test
LfM	multiple myeloma research laboratory (Labor für Myelomforschung
M	marginal present expression
M-protein	monoclonal protein
MBC	memory B cell

MC	molecular classification
MGUS	monoclonal gammopathy of undetermined significance
MM	multiple myeloma
MMR	relapsed multiple myeloma
MMRF	multiple myeloma research foundation
MP	marginal and present expression (taken together)
NA	not available
NR	not reached
OS	overall survival
P	present expression
PA call	present/absent call
PAD	bortezomib (formerly called PS-341), adriamycin, dexamethasone
PANP	presence-absence calls from negative strand matching probesets
PCA	principle component analysis
PI	proteasome inhibitor
PPC	polyclonal plasmablastic cell
R-ISS	revised international staging system
RNA-seq	RNA-sequencing
RPI	RNA-sequencing based proliferation index
siRNA	small interfering RNA
SJ	splice junction
SMM	smouldering multiple myeloma
STAR	spliced transcripts alignment to a reference
TAD	thalidomide, adriamycin, dexamethasone
TCB	T cell bispecific
TeG	testing group
TG	training group
TMM	weighted trimmed mean of M-values method
VAD	vincristine, adriamycin, dexamethasone
VCD	bortezomib (velcade), cyclophosphamide, dexamethasone
VG	validation group
WES	whole exome sequencing
WGS	whole genome sequencing

1 Introduction

Primary aim of this thesis is to form the bioinformatic basis for an implementation of RNA-sequencing (RNA-seq) based transcriptome profiling in clinical routine application, assessing risk stratification and potential targets in the malignant plasma cell disease multiple myeloma.
This chapter is divided in six parts. First, the development of normal bone marrow plasma cells is described. Second, a general introduction in multiple myeloma is given comprising a clinical part, including signs and symptoms, treatment, and pathogenesis. Third, molecular profiling and diagnostics in malignant plasma cell diseases are depicted. Fourth, risk assessment and classifications using molecular profiling and fifth, the assessment of potential targets are described. At the end of this chapter, the aims of this thesis are presented.

1.1 Plasma cell development

Bone marrow plasma cells (BMPC) are part of the adaptive immune system [170]. They develop from lymphoid progenitor cells via different intermediate states to B cells, which leave the bone marrow and circulate between blood and peripheral lymphoid tissues [170]. If a so called "naïve B cell" encounters the presentation of its specific antigen in lymphoid organs, e.g. a surface structure of a pathogen, the B cell starts to proliferate and matures to a so called "polyclonal plasmablastic cell" (PPC) [170]. Subsequently, the PPC differentiate into memory B cells (MBC) and plasma cells [170]. This process involves complex molecular changes [121, 170], including repeated DNA rearrangements, e.g. somatic hypermutation and class switch recombination, necessary for diversity and high affinity of immunoglobulins. Plasma cells home in the bone marrow and interact with the local, adjacent microenvironment [170, 189]. BMPCs are long living [121] and thus enable long term immunological memory, continuously synthesising and secreting immunoglobulins (Ig, antibodies) [170]. The latter recognises, binds and opsonises foreign antigens [170]. Thereby, the complement system is activated, phagocytosis initiated, and the pathogen eliminated [170]. The basic structure of such an Ig consists of two identical parts building an Y-shaped protein. Each part is composed of a heavy and light polypeptide chain, forming a constant region and the variable antigen binding-site [170]. There are five types of Ig: A, D, E, G, M, each determined by their heavy chain type: α, δ, ϵ, γ, or μ [170]. The light chain type is either κ or λ [170].
In physiological condition, plasma cells comprise is 0.1%-1% of nucleated cells in the bone marrow [121, 189, 227]. This figure does not change significantly over time, as

first, plasma cells do not proliferate and second, resident plasma cells in niches are in constant competition with juvenile plasma cells [189]. Each new wave of plasma cells dislocates a part of the resident plasma cell population from their niches, to an extent keeping the total number of niched plasma cells identical [189].

1.2 Multiple myeloma

In this section, first multiple myeloma is introduced generally regarding epidemiology, signs and symptoms, followed by a brief description of the treatment paradigms and of the pathogenesis of the disease.

1.2.1 General introduction

Multiple myeloma is a haematological disease characterised by the accumulation of malignant plasma cells in the bone marrow [214]. It represents approximately 1-2% of all newly diagnosed cancer diseases and is the third most common haematological cancer, after leukaemia and non-Hodgkin-lymphoma [106, 116]. The five-year survival probability of multiple myeloma patients is about 50% [106, 116] and the ten-year survival rate about 30% [116]. The lifetime risk of developing myeloma is 0.82% [106].

Most myeloma patients initially visit a physician due to bone pain (67%) [118]. Further unspecific clinical signs and symptoms are, for example, fatigue, weakness and weight loss. For current standard of care diagnosis, blood, urine and bone marrow samples are examined. Using electrophoresis, the monoclonal Ig (M-protein) in serum and urine is determined and subsequently quantified. In an individual patient, the amount of M-protein is a surrogate for the total number of malignant plasma cells, which produce and secrete this Ig [121, 129]. The total number of malignant plasma cells is also called tumour mass. The bone marrow samples after aspiration are used for enumeration of malignant plasma cells and molecular profiling (see section 1.3). Further routine diagnosis procedures include imaging techniques as whole-body computer aided tomography, and whole-body magnetic resonance imaging for determining lytic bone and focal lesions, respectively [93].

Regarding malignant plasma cell diseases, three consecutive entities or stages are delineated: monoclonal gammopathy of undetermined significance (MGUS), asymptomatic multiple myeloma (AMM), and multiple myeloma (MM) [111]. MGUS and AMM are differentiated by the amount of M-protein and the percentage of plasma cell infiltration [111]. Both entities are asymptomatic. MGUS is most often incidentally detected [130] and the prevalence of being diagnosed with MGUS increases

during live [130]. MGUS progresses with about 1% probability per year to symptomatic myeloma [131]. AMM is characterised by a M-protein of $\geq$30 g/L and/or a plasma cell infiltration in the bone marrow of $\geq$10% [214]. AMM progresses with about 10% probability per year in the first 5 years to symptomatic myeloma [129]. In MM, a single aberrant plasma cell clone proliferates unregulated. This accumulation of malignant plasma cells transforms the bone marrow microenvironment (BMME) and causes clinical signs, symptoms and end organ damage [121]. The displacement of normal haematopoiesis by accumulating plasma cells leads to **a**naemia [103, 111] and generation of focal and osteolytic lesions in the bone (**b**one disease) [111]. During bone degradation, calcium can be released and hyper**c**alcemia and **r**enal impairment occur [100, 111, 128]. The kidney is further affected by the high amount of M-protein, or, especially, light chains, produced by the accumulated myeloma cells [111, 170]. These clinical signs and symptoms regarding end organ damages caused by myeloma are summarised as CRAB criteria (hyper**C**alcaemia, **R**enal impairment, **A**naemia and **B**one disease) by the International Myeloma working group (IMWG) in 2003 [111]. Presence of a CRAB-criterion is seen as indication for systemic treatment [111].
In 2014, the IMWG modified and clarified the criteria, regarding modern imaging methods as computer tomography [190]. The IMWG included biomarkers, which are intended to determine asymptomatic patients with a high probability of progressing to symptomatic myeloma. For these patients, likewise an indication to initiate treatment can be seen. The main underlying concept is that an early effective therapy might prevent end organ damage and may improve the overall survival of the patients [190]. Hence, the asymptomatic patients fulfilling the so called "SLiM" criteria are now considered as "symptomatic" by the IMWG [190]. Patients neither fulfilling the "SLiM-CRAB" criteria nor being MGUS patients are termed smouldering myeloma patients (SMM). This term had previously been used synonymously with the term AMM [190]. If not otherwise specified, the above described definition of AMM is used within this thesis.

1.2.2 Treatment paradigms

Major aims of a treatment in myeloma are to revert or prevent end organ damage [190] and to enhance the survival of the patients. The long-term aim is the cure of the disease by functional eradication of malignant plasma cells, currently only reached in a small subfraction of patients [17]. As in all malignant diseases, treatment needs balancing between the benefit (efficacy) and toxicity: side effects need to be bearable for the patient and treatment related mortality has to be considered [97, 212]. Thus, not all myeloma patients are treated. The indication to initiate treatment is currently seen

if signs or symptoms of multiple myeloma are present, e.g. if the CRAB criteria are fulfilled (see section 1.2.1) [111]. Hence, the current recommendation is not to treat patients with MGUS or SMM outside clinical trials [190]. The treatment of AMM patients fulfilling the "SLiM" criteria is recommended by the IMWG, but internationally debated [75, 153, 192], e.g. in the MM3 trial of the "Arbeitsgemeinschaft medikamentöser Tumortherapie" (EudraCT No: 2018-000924-32). Only to treat symptomatic patients had been the recommendation due to a lack of early treatment clinical trials showing prolonging of overall survival of either MGUS or AMM patients and due to the inability of delineating patients of especially high risk of progression. This general paradigm has changed, at least if patients are treated within clinical trials: biomarkers indicating "imminent" risk of progression are available [97, 190]. Usually, imminent risk is defined as 80% progression probability within two years, e.g. by the "SLiM" criteria [190]. Clinical trials showed a benefit for early treatment in terms of response rate [122, 132], but, most importantly, of overall survival [152]. Hence, if the "SLiM" criteria (see section 1.2.1), are present, an indication for treatment can be seen [190]. However, MGUS and especially AMM patients are in any case closely observed [111].
In symptomatic myeloma patients, two groups of patients are distinguished, patients who can and those who can not be treated intensively [128]. This distinction depends on physical condition, comorbidities and the biological age of the patient.
Patients up to an age of 70 years without significant comorbidities are routinely treated intensively [204]. This treatment can safely be extended to fit to older patients [204]. A typical treatment strategy comprises first, 3 to 6 cycles of a so called "induction" treatment [10, 41, 77, 183], followed by autologous steam cell collection. Patients are treated with high-dose melphalan and subsequently autologous stem cells are reinfused [10, 41, 77, 183]. In Germany, the current standard is a combination of the proteasome inhibitor bortezomib, the corticosteroid dexamethasone and a cytostatic drug, e.g. cyclophosphamide (VCD) [74]. Within clinical trials, the immunomodulatory drug (IMiDs) lenalidomide in combination with bortezomib and dexamethasone are tested, with or without a CD38-antibody (e.g. daratumumab or isatuximab, see also section 1.5), as exemplified by the German speaking myeloma multicenter group (GMMG) HD7 trial (ClinicalTrials.gov identifier: NCT03617731). Aim of the intensive treatment is achieving an as deep as possible remission of the disease and, in a subfraction of patients, functional eradication of myeloma cells [17].
Not intensively treatable patients receive comparable regimen, but in reduced intensity and, especially, without high-dose melphalan treatment and stem cell transplantation [128]. These are currently e.g. combinations of lenalidomide, dexamethasone, with or without daratumumab [147].

1.2.3 Pathogenesis

Multiple myeloma is characterised by a broad inter-patient heterogeneity in terms of clinical signs and symptoms, patient survival, and underlying pathogenetic mechanisms [121, 214]. The mechanisms are mainly based on chromosomal aberrations and nucleotide changes, which alter gene expression [100, 102, 121, 217, 223] and thus affect the cell proliferation [101], the response to treatment and ultimately the survival of a patient [12, 103, 174, 175].
In the following, the underlying chromosomal aberrations, changes in ploidy and dysregulation of D-type cyclin expression are described. Subsequently, recent findings about the nucleotide changes are introduced and the prognostic impact of the chromosomal aberrations is depicted. At the end of this section, the current pathogenetic model of the Multiple Myeloma Research Laboratory (LfM) is presented.

1.2.3.1 Chromosomal aberrations and D-type cyclin expression as unifying events

Multiple myeloma pathogenesis is generally thought to follow two pathways: IgH-translocations and hyperdiploidy (HRD) [214]. IgH-translocations affect the IgH locus on chromosome 14, e.g. t(11;14), t(4;14), t(6;14), t(14;16) and t(14;20) [214]. HRD is characterised by a gain of odd numbered chromosomes (3, 5, 7, 9, 11, 15, 19 and 21). Both dysregulates the expression of cyclin D family members, *CCND1*[1], *CCND2* and *CCND3*, which is one of the hallmarks and the "unifying" event of malignant plasma cell diseases [22, 214]. In normal plasma cells, *CCND2* and *CCND3* are expressed at low levels, whereas *CCND1* is not [214]. In malignant plasma cells either *CCND2* or *CCND3* are overexpressed, or *CCND1* is aberrantly expressed [22, 103, 214]. The expression of one of the D-type cyclins is exclusive in most patients [214]. In HRD patients, the aberrant expression pattern can be explained by numerical aberrations (e.g. a gain of 11q13, see below) or by indirect overexpression of *CCND2* in HRD patients [214]. Likewise, alterations in the IgH-locus can impact directly or indirectly the cyclin D expressions [214]. The translocation t(11;14) or t(6;14) juxtaposes *CCND1* (located at chromosomal locus 11q13) or *CCND3* (located at 6p21) in the proximity of the IgH-enhancer on chromosome 14, respectively [214]. This enhancer is highly active in normal and malignant plasma cells due to the Ig-production [160] and leads to high expression levels of the translocated cyclin [214]. The expressions of *CCND1* or *CCND3* are so typical for the respective patient group, that they can be used to predict

[1] A gene is depicted by its commonly known name in this thesis. The gene symbol is presented in paranthesis, if it differs. Both are written in italic font, except in tables and figures, in which standard font is used for clarity purposes. For gene symbol description see "HUGO Gene Nomenclature Commitee; The resource for approved human gene nomenclature": `https://www.genenames.org/`.

the underlying translocations t(11;14) and t(6;14), respectively [169]. In contrast, high expression of *CCND2* (located at chromosomal locus 12p13) is not directly explainable by translocations, as t(12;14) translocations involving *CCND2* are extremely rare [214]. However, *CCND2* overexpression is indirectly associated with the translocation t(4;14) [214].
Further frequent chromosomal changes are numerical aberrations, which are present in a broad number of patients, e.g. Walker *et al.* [243] showing 21 chromosomes are affected among 114 myeloma samples. The most common are deletions of 17p13 or 13q14 and gains of 1q21 [182, 243].

1.2.3.2 Nucleotide changes

The term "nucleotide changes" denotes single nucleotide variants or insertions and deletions. Although there is a variety of mutations, e.g. in median 43 mutations per patient (range 1 to 1939) in the CoMMpass cohort (see section 1.4.2.3, [169]), no "unifying" myeloma associated nucleotide change is present [214]. Relatively frequent changes affect e.g. the genes for KRAS or BRAF (mutation V600E/K) [37, 146], involved in the signal transduction of the mitogen-activated protein kinase/ERK signalling pathway [37, 214, 252]. *KRAS* and *NRAS* (or members of the respective pathways) are the most frequently mutated genes in MM, followed by *FAM46C* (*TENT5C*), *DIS3* and *TP53* [37, 146]. Another relatively frequently affected pathway is NF-κB with nucleotide variants in genes like *TNFRSF1A* or *TRAF3* [37].

1.2.3.3 Prognostic impact

Besides their role in pathogenesis of myeloma, molecular alterations convey clinical interest in impacting on the prognosis of a patient. The chromosomal aberrations 13q14, 17p13, 1q21, t(4;14), t(14;16) are all associated with adverse progression free and overall survival in MM [25]. Additionally, 1q21, 17p13 and t(4;14) are associated with a shorter time to progression from AMM to MM [175, 191]. HRD is the only aberration pattern which is associated with survival in contrary manner in AMM and MM: it has a favourable (or neutral) prognosis in MM [125, 176], but adverse prognosis in AMM [175]. Chromosomal aberrations are very rarely present as single alteration in both in AMM [97] and MM [25, 52, 176, 214], and the number of chromosomal aberrations is associated with adverse progression free and overall survival [12, 25, 176, 214]. Whereas the association of survival with chromosomal aberrations is well-known, this is much less the case for the association with nucleotide changes. For instance, mutations in the genes *TP53*, *ZFHX4*, *CCND1*, *ATM* and *ATR* have been reported to be associated with poor prognosis by Walker *et al.* [242] in 2015, while

mutations in *IRF4* and *EGR1* seem to be associated with favourable survival [242]. Mutations in the most frequent mutated genes *KRAS* and *NRAS* are not associated with survival [242].

1.2.3.4 Pathogenetic model

In the following two prominent aspects of myeloma pathogenesis are assessed: first, the cellular origin of myeloma cells, in terms of the B cell differentiation state they develop from, and secondly, when during progression of pre-MGUS stage to MGUS, AMM and MM aberrations arise.

Regarding the first aspect, properties of myeloma cells can be compared to different stages of B cell development and their respective properties [214]. On the one hand, myeloma cells express the surface protein CD138 (SDC1), a hallmark of plasma cells [121], and are able to produce monoclonal Ig [214]. Hence, genetically, they have undergone the same series of DNA rearrangements as plasma cells, e.g. somatic hypermutation and class switch recombination, necessary for the generation of Ig (see section 1.1) [121, 214]. On the other hand, myeloma cells proliferate, which is typical for activated B cells and plasmablasts. Hence, they either evolve from proliferating cells, keeping their proliferative ability, or from non-proliferating cells, which regained their ability to proliferate [214].

Regarding the second aspect, i.e. the point in time, when during progression through different stages myeloma typical progression inducing aberrations originate, two main hypotheses are discussed. The most obvious idea is a "multistep model" [214] during disease progression [86, 167], as MM samples harbour a higher median number of aberrations compared to AMM or MGUS samples [244]. However, recent longitudinal studies have shown that aberrations are already present in previous myeloma stages and only few *de novo* or secondary alterations occur [97, 215]. This is one of the main reasons subsequently leading to the hypothesis, that disease progression is triggered by oncogenic aberrations *ab initio* or at least at pre-MGUS stage [22, 97]. Based on this pathogenetic model of the LfM, disease progression and evolvement of clinical signs and symptoms occur due to an increasing number of plasma cells and their interaction with the BMME, triggered by the initial set of alterations driving progression to the precursor stages [97].

1.3 Molecular profiling

In Europe and the US, molecular profiling is implemented at different levels in large myeloma centres. At the LfM at the University Hospital Heidelberg, molecular profiling is integrated in extended clinical routine diagnostics [99]. An overview of the

processing is depicted in figure 1.1, for details see also section 2.1.1. The analysed bone marrow aspirates are necessary for the diagnosis and staging of the disease e.g. in quantifying plasma cell infiltration. In brief, 60-80 ml of bone marrow is aspirated in local anaesthesia from the iliac crest [97]. Subsequently, bone marrow smears are generated for the determination of bone marrow plasma cell infiltration. The bulk of the aspirate is used for plasma cell purification by density gradient centrifugation and subsequent selection via anti-CD138 immuno-microbeads [97, 99, 101, 212]. The purity of the purified plasma cells is controlled with flow cytometry [97]. Analogously, normal bone marrow plasma cells from healthy donors are purified as comparator population.

In extended clinical routine, the CD138-purified plasma cells are used at the LfM to assess pathogenetically or prognostically relevant genetic alterations, using interphase fluorescence *in-situ* hybridisation (iFISH), and to determine gene expression, using DNA-microarrays [99, 100, 175, 217] as well as RNA-seq.

Detection of chromosomal aberrations by iFISH. iFISH is a conventional method for the detection of recurrent chromosomal alterations in CD138-purified malignant plasma cells [99]. For this, fluorescence-labelled DNA probes are predefined for chromosomal regions of interest and hybridised with the chromosomes of CD138-purified malignant plasma cells (see also section 2.1.1) [175, 176]. Two principal types of chromosomal alterations are assessable by this technique: numerical aberrations (gains or losses) and translocations (see above, section 1.2.3) [99]. iFISH is a DNA-based method and therefore relatively simple regarding sample processing and consignment (e.g. in clinical multi-centre trials), in comparison to DNA-microarray or RNA-seq. The latter two are RNA-based and require a higher level of sampling quality, especially in terms of timing, as delay in sample processing alters gene expression [155]. iFISH is validated in multiple trials and analyses [11, 12, 40, 174]. However, it requires a pre-selection of specific genes or regions and the number of probes is limited by the number of cells per iFISH spot and the median number of purifiable malignant plasma cells.

Transcriptome profiling by DNA-microarrays. A DNA-microarray is a small glass slide ("chip") divided in quadratic areas, each consisting of approximately one million fixed copies of one "probe" [97]. Each probe is a 25-mer oligonucleotide, designed to hybridise with a specific mRNA in human cells. For this type of expression analysis, the extracted RNA is amplified and labelled, using biotinylated nucleotides and subsequent fluorescence staining [97]. The expression of a mRNA corresponding to a specific probe can be recognised by the fluorescence intensity, which correlates with the number of transcripts bound to this specific probe (see also section 2.1.1) [97]. In this

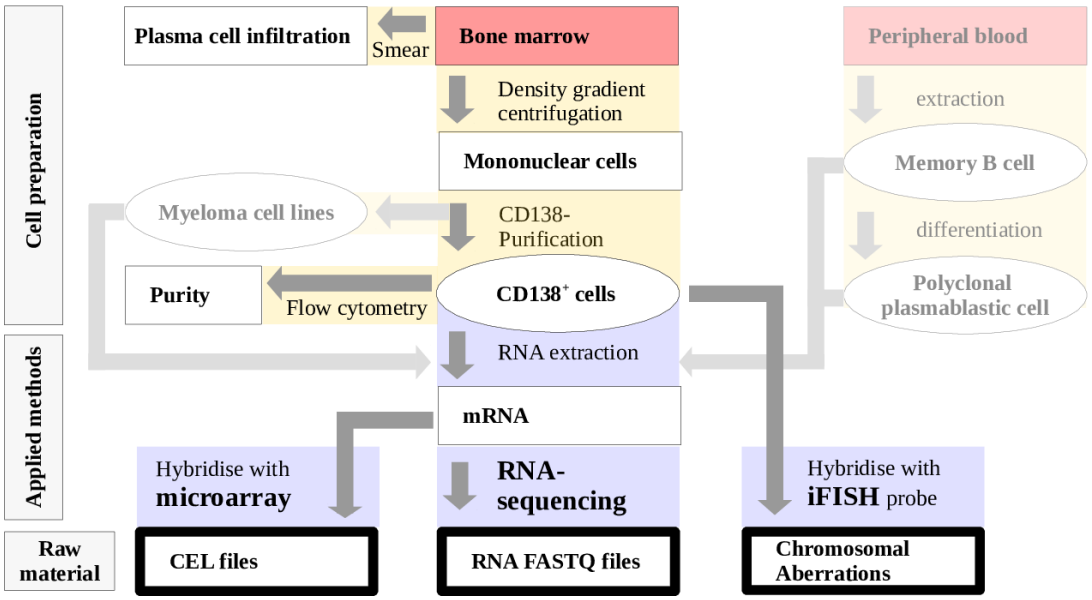

Figure 1.1: Flowchart of the molecular profiling pipeline in the multiple myeloma research laboratory Heidelberg. Bone marrow aspirate, extracted from the iliac crest, is purified by density gradient centrifugation and selection of plasma cells by anti-CD138 immuno-microbeads. The purified plasma cells are spinned on glass slides and used for interphase fluorescence *in-situ* hybridisation (iFISH). The RNA is extracted from normal and malignant plasma cells and used to determine gene expression, using DNA-microarrays and RNA-seq. In pale colours, the generation of the comparator populations, e.g. memory B cells, polyclonal plasmablastic cells, and human myeloma cell lines is depicted.

thesis Affymetrix U133 2.0 plus microarrays were used, termed "DNA-microarray" or "microarray". As iFISH, microarray processing is very standardised and can in principle be used in clinical routine [16, 27, 99, 156, 219], as exemplified by the LfM in the GMMG-MM5 trial [99]. The main advantage besides applicability is, that per microarray about fifty thousand probes can be analysed simultaneously.

Transcriptome profiling by RNA-seq. Next generation sequencing allows the analysis of the whole genome (WGS), exome (WES) and transcriptome (RNA-seq), using high-throughput massive parallel sequencing by synthesis (see also section 2.1.1). Not in the focus of this thesis, WGS and WES allow the detection of structural and numeric DNA alterations. By sequencing tumour and normal samples, variants not present in the germline (somatic) can be determined. As the exome is smaller than the whole genome (2%-3% of the whole genome [85]), WES is cheaper. However, the necessary exome capturing may lead to a non-uniform coverage. Both DNA-microarray and RNA-seq can be used for gene expression analysis. Their advantage over iFISH is the possibility of assessing more variables, including prognostically relevant biological variables (as proliferation see section 1.4.2.1) and potential targets (see section 1.5). RNA-seq can, in comparison to microarrays, be performed from low input as 10 pg of RNA (compared to about 100 ng for microarrays with the Affymetrix small sample labelling protocol) [99, 211]. This allows to investigate also plasma cell dyscrasias

with low tumour mass (i.e. MGUS, AMM). Further, no *a priori* sequence definition is necessary, hence, the detection of mutated sequences (see section 1.5) and splice variants becomes possible. Splice variants are alternative transcripts of a gene, formed during the splicing process of the pre-mRNA to mRNA.
The methods are also described in section 2.1.1 and the protocols in Hose [97] (microarray and iFISH) and Seckinger *et al.* [211] (RNA-seq).

1.4 Classification and risk assessment

Multiple myeloma is heterogeneous in terms of clinical factors, molecular alterations (see section 1.2.3) and individual prognosis of a patient [127, 128]. Almost all myeloma patients relapse and the overall survival ranges from months to more than 20 years [139, 156]. It is thus desirable to stratify patients regarding risk and molecular subentities. In theory, a stratification approach prior to the start of the therapy could be used for a treatment with individual intensity, duration and type. A respective strategy would of course have to be tested in randomised clinical trials. One concept is to treat patients with specifically adverse prognosis more aggressively. An example for such a strategy is the GMMG CONCEPT trial (ClinicalTrials.gov identifier: NCT03104842) for previously untreated symptomatic myeloma patients of high risk as defined by presence of at least one of the chromosomal aberration (deletion 17p, t(4;14), or more than three copies of 1q21) and ISS-stage ("international staging system") II or III.
As stratification systems, likewise individual genes associated with prognosis can be used. An example is the expression of *AURKA* [102], associated with adverse survival. In principle, patients expressing *AURKA* could be treated with a specific inhibitor [102]. The long way of such an approach to clinical realisation is exemplified by the AURKA inhibitor VX-680, which failed due to intolerable cardiac side effects, i.e. induction of a long QT-syndrome [71].
In the following, the risk assessment used in extended clinical routine at the LfM at the University Hospital Heidelberg will be described.

1.4.1 Using clinical parameters

The risk of MM patients can be assessed by clinical parameters as tumour mass surrogates and bone damage. In the Durie-Salmon staging, first presented in 1975, the M-protein level and the grade of bone damage (detected by X-ray) were used to surrogate tumour mass and to delineate three risk groups [61]. Thirty years later, the ISS suggests the amount of albumin in serum and the amount of β_2-microglobulin as tu-

mour mass surrogate for defining three risk groups [84].
Clinical staging systems can be enhanced by inclusion of chromosomal alterations. In 2014, the IMWG recommended, besides ISS, to test for presence of deletion 17p13, translocation t(4;14) and gain 1q21 due to association with poor survival [42]. One year later, Palumbo *et al.* [182] developed a new classification, the revised international staging system (R-ISS), combining original ISS and presence of high-risk cytogenetic abnormalities, as deletion 17p13 and translocation t(4;14) [182].

1.4.2 Using gene expression profiling

Beyond clinical parameters and iFISH, classification and risk stratification of myeloma can be based on gene expression profiling, traditionally using DNA-microarrays. Patients can be classified according to surrogates of biological variables, e.g. proliferation (GPI [101]), survival (UAMS70 [219], RS [197], EMC92 [124], and IFM15 [56]), and molecular subtypes (TC [22, 43] and molecular classification (MC) [254]).
The clinical usage of microarrays has been limited to few centres worldwide, including the "University of Arkansas School of Medical Sciences", the "centre hospitalier universitaire" Montpellier and the LfM at the University Hospital Heidelberg. Although the use is in principle possible in extended clinical routine as described in 2011 by the LfM by Meissner *et al.* [156] (see also section 1.4.3), it is not recommended for clinical routine [42, 66, 159].

1.4.2.1 Assessing proliferation as example for biological surrogates

Proliferation of malignant plasma cells is one of the most prominent adverse prognostic factors in multiple myeloma [83, 101, 205, 225]. It can, at least theoretically, be targeted by defined treatment options (e.g. tubulin polymerase inhibitors [101] or AURKA inhibitors [102]). Hose *et al.* [101] developed the gene expression-based proliferation index (GPI) at the LfM at the University Hospital Heidelberg (for a detailed description see section 2.5). It includes 50 proliferation or cell cycle associated genes and allows the determination of proliferation in a clinical setting [101]. The GPI was developed as surrogate of a biological variable and it was (intendedly) not fitted to survival, to independently assess the impact of malignant plasma cell proliferation on survival [101].

1.4.2.2 Assessing survival

The expression of a group of genes can be used directly to stratify the survival of patients. Four main risk stratifications are described, differing in the applied methods (for a detailed description see section 2.5):

The University of Arkansas School of Medical Sciences in Little Rock developed the UAMS70 score [219]. It uses log-rank tests to identify 70 genes associated with short survival in MM. Most of the genes are located on chromosome 1. Numerical alterations there (see section 1.2.3) e.g. 1q21 gains (detected by iFISH), are likewise prognostic [219].
Rème *et al.* [197] presented together with the LfM the gene expression-based risk score RS. Patients were divided into three groups, using 19 prognostic genes, selected with overall survival analysis [197]. For every new sample, the same normalisation parameters were used, which ensures the comparability of new samples [197].
Using the GMMG-HD4/Hovon-65 sample data (EudraCT 2004-000944-26 [222]), Kuiper *et al.* [124] from the Erasmus medical centre identified 92 genes associated with survival in MM, comprised in the EMC92, later commercialised as SKY92 score by Skyline diagnostics [238].
The Intergroupe Francophone du Myélome developed the IFM15, a risk-based score, which is based on 15 genes associated with poor prognosis in MM [56].

1.4.2.3 Assessing molecular entities

By assessing molecular entities, patients are divided in biological or pathophysiological subgroups. In contrast to risk stratification, these groups need not, but can, have a prognostic impact. In the simplest case, alterations, conventionally detected by other methods such as iFISH, can be predicted using altered gene expression. Examples include the gene expression profiling report (GEP-R) [156] predicting the translocation t(4;14). In the CoMMpass study ("relating Clinical outcomes in Multiple Myeloma to Personal Assessment of Genetic Profile") by the multiple myeloma research foundation (MMRF), the translocations t(11;14), t(6;14), t(4;14), t(14;16) are predicted each by a single gene expression value [51, 169].
A more integrative approach is to develop a molecular classification, to group patients according to common underlying pathogenetic mechanisms. Examples for this are the TC [22] and the MC [254] classification.
The TC classification by Bergsagel *et al.* [22] (2005) is based on the expression of eight genes (*CCND1*, *CCND2*, *CCND3*, *FGFR3*, *MMSET* (*WHSC1*), *MAF*, *ITGB7*, *CX3CR1*), dividing the patients into 8 groups (4p16, MAF, 6p21, 11q13, D1, D1+D2, D2, none), which are associated with IgH-translocations and cyclin D expressions. The modified TC classification by Chng *et al.* [43] (2007) adds *MAFB* to specify the "MAF" group.
The MC classification of the University of Arkansas for Medical Sciences by Zhan *et al.* [254] distinguishes MM in seven transcriptional signatures (MS, MF, CD1, CD2,

HY, PR, LB) based on unsupervised hierarchical clustering of MGUS, MM and human myeloma cell line (HMCL) samples. Each subgroup has a unique expression pattern, mainly based on nine genes (*MAF*, *MAFB*, *FGFR3*, *MMSET*, *CCND1*, *CCND2*, *CCND3*, *FRZB* and *DKK1*). MF, MS and PR ("proliferation") are associated with survival and are high risk groups, while LB (less focal lesions, "low bone disease"), CD1 ("CCND1"), CD2, ("CCND3") and HY ("Hyperdiploidy") are defined as low risk groups [254].
Groups delineated in the TC and the MC classification show an incomplete overlap, as they are based in part on expression values of the same gene. For instance, the MS group of the MC classification is mainly associated with *FGFR3* expression and hence overlaps with the FGFR3 group of the TC classification. Likewise, the MF group has either a high expression in *MAF* or *MAFB* and overlaps with the MAF group. Additionally, both classifications include classes based on the expressions of *CCND1*, *CCND2* and *CCND3* (MC: CD1 and CD2; TC: D1, D1+2, D2).

1.4.3 Reporting of gene expression profiling (GEP-report)

To be applicable in extended clinical routine, risk stratifications and molecular classifications need to be reported to physicians and patients in a validated, reproducible and understandable manner. For gene expression data and their combination with clinical parameters the GEP-R was developed by Meissner *et al.* [156] at the LfM. This is an academic reporting tool for using gene expression data of DNA-microarrays in clinical practice. It is based on non-commercial software frameworks, including the open-source software R. Validated conventional clinical scores and clinical parameters can be entered (e.g. ISS). The calculation of proliferation-based scores (GPI), risk-based scores (UAMS70, IFM15) and molecular classifications (EC, TC, MC) is implemented. Additionally, the expression of targets (e.g. *AURKA*, *IGF1R*) is assessed for an exemplification of potential individualisation of treatment. The GEP-R performs a quality and identity control, determining sex, light chain type and heavy chain type of a sample. Different risk stratification methods do not necessarily lead to the same result, which complicates and confuses clinical and therapeutic interpretation [156]. To avoid this, GEP-R includes a strategy by combining all prognostic factors in the HM-metascore. This generates a "summarising" risk assessment. The use of a cohort-based normalisation and cohort-based thresholds implies the need of normalising new samples by applying the normalisation parameters of the training group. This allows evaluating every new patient individually [156].
The GEP-R is used routinely at the LfM at the University Hospital Heidelberg, and recently the application has been evaluated in the GMMG-MM5 phase III clinical trial

[99]. The study demonstrates that DNA-microarray analysis can be performed for >80% of the patients and the results of the GEP-R are available within 4-6 weeks [99].

1.5 Potential targets

The term "target" defines a biological structure (e.g. a protein or mRNA) affectable by a pharmacological inhibitor or immunotherapeutical agent, which exists or can be developed (see section 1.2.2) [97]. Currently "actionable" targets include targets for which compounds are approved[2] (e.g. CD38 by daratumumab [55] or isatuximab [57]) or in later stage development for multiple myeloma (e.g. BCMA, ClinicalTrials.gov identifiers: NCT03486067 and NCT03836053) or have been described but are currently not in clinical testing for MM (e.g. AURKA [102] or IGF1R [68]). Currently not actionable targets are for instance the potential vaccination targets HM1.24 and MAGEA1 [100, 156, 209].

Targets can be further differentiated whether they are expressed in normal and malignant plasma cells (e.g. *CD38* [213] and *BCMA* [212]), in malignant plasma cells only (termed aberrant expression, e.g. *FGFR3* [156]) or frequently show a significant higher expression in malignant plasma cells (termed overexpression, e.g. *CCND1*, *CCND2* and *CCND3* [156]).

A third possible categorization relates to the mutually non-exclusive therapeutic treatment strategies of the targets: They are assessable by i) monoclonal antibodies or by chimeric antigen receptor (CAR) T cells, ii) small molecule inhibitors, iii) mutation specific agents, iv) vaccination, and v) theoretical (not currently addressed) but potentially assessable by therapeutic small interfering RNA (siRNA) molecules.

i) Cell surface antigens which are broadly expressed in normal and malignant plasma cells are for instance CD38 and BCMA (TNFRSF17) [212, 213]. They can be targeted exemplarily with three main immunotherapeutic approaches: monoclonal antibodies, bispecific monoclonal antibodies or with modified T cells. By opsonisation of the target surface proteins with monoclonal antibodies, the innate immune system and the complement cascade is activated, and the cell lysed (cellular cytotoxicity) or cellular phagocytosis is induced. A good target is for instance CD38, shown at the LfM by Seckinger *et al.* [213] and others [57, 147]. It is broadly expressed in multiple myeloma [213] and CD38-antibodies are available and/or approved[2] Bispecific monoclonal antibodies effectively link tumour cells via a surface protein (e.g. BCMA) and effector cells, e.g. T cells (via CD3) [212]. This coupling leads to the induction of

[2]National Cancer Institute: Drugs Approved for Multiple Myeloma and Other Plasma Cell Neoplasms; Online resource: `https://www.cancer.gov/about-cancer/treatment/drugs/multiple-myeloma`; Status: 27.02.2020, 12:55

myeloma cell death (in case of T cells). For instance, BCMA has been shown at the LfM by Seckinger *et al.* [212] and others to be an ideal target due to expression height and pattern. BCMA is the receptor for the myeloma cell growth factors BAFF (TNFSF13B) and APRIL (TNFSF13) [166, 212] and is mandatory for the survival of long-living BMPCs [180]. The assessment of *BCMA* expression was part of this thesis (see sections 2.8, 3.5, 4.4). The project [212] led to the development of the compound CC-93269 [212] currently in clinical trials (ClinicalTrials.gov identifier: NCT03486067). Alternatively, T cells can be triggered to attack myeloma cells by *ex vivo* engineering of the CAR or the T cell receptor [31, 69, 120, 194, 206, 212].
ii) Especially aberrantly but also frequently over expressed genes in comparison to normal plasma cells are ideal targets for small molecule inhibitors. Aberrantly expressed genes targeted by small molecule inhibitors are for instance HGF and FGFR3 [100]. A further ideal target for inhibitors is AURKA. The gene is associated with plasma cell proliferation and a clinical inhibitor was developed (VX-680) [102] although it failed in clinical application [71] (see section 1.4). iii) An example for a mutation creating a target in cancer cells is the BRAF mutation V600E [8]. The kinase BRAF is involved in cell division and mutations in BRAF increase the proliferation of the cancer cells [24]. It is affectable e.g. by the clinical applicable inhibitors vemurafenib [24, 109] and dabrafenib [91].
iv) Therapeutic or preventive cancer vaccines induce the generation of monoclonal antibodies against antigens expressed on target structures. In multiple myeloma, they are designed to specifically bind myeloma cells, without affecting normal tissue [209]. Possible targets for this strategy are frequently overexpressed, either constitutively or aberrantly [209]. Examples for the former are *HM1.24* [108, 113, 198, 209], or *RHAMM* (*HMMR*) [76, 81, 82, 209]. Examples for the latter comprises cancer testis antigens (CTAs) as *MAGEA3* [48, 156, 209], *NYESO1* (*CTAG1B*) [18, 126, 178, 209, 237], or *WT1* [209], which have been assessed as part of this thesis. While HM1.24, RHAMM, WT1 and NYESO1 are currently not actionable by vaccination, a vaccine for MAGEA3 has completed a clinical trial in multiple myeloma (ClinicalTrials.gov identifier: NCT01380145) [47].
v) If a gene expressed in myeloma cells is associated with a cancer relevant pathway (e.g. CCND1 [242]), a knock down of it may, theoretically, improve the survival of patients. RNA-seq enables the expression analysis of mRNA target sequences for therapeutic siRNA molecules. siRNAs can knock down a target, by binding specifically to its target sequence in the cell, which is then degraded [112]. For the transport into the cell vectors, e.g. "nanocarriers", which could enter the cell and release the siRNA *in vivo* could be used [112]. Up to now highly speculative in myeloma, several "nanocar-

riers" containing siRNAs are in clinical trials (phase I) in other cancer types [112].

1.6 Aim of the thesis

The primary aims of this thesis are to form the bioinformatic basis for the implementation of RNA-seq-based transcriptome profiling in translational myeloma research and extended clinical routine application, to use it for novel risk stratification, and assessment of targets.
Gene expression profiling using DNA-microarrays has provided detailed information on the pathophysiology of malignant plasma cell diseases [124], risk stratification, target assessment, and the report of these data in clinical routine has been implemented (e.g. GEP-R) [99, 156]. Expression profiling by RNA-seq has several advantages over DNA-microarrays and is therefore considered the new gold standard: it allows the use of lower amounts of sample material, which is especially useful for the assessment of early disease entities, a more precise quantification of gene expression, and the assessment of mutated sequences as well as splicing variants. To implement RNA-seq, it is further necessary to connect new to previously obtained microarray results. Thus, the aims of the thesis are: i) to establish a practicable pipeline to analyse and to lay a basis to report RNA-seq data, ii) to transfer and connect current risk stratifications and molecular classifications based on DNA-microarray to RNA-seq technology, iii) to discover novel prognostic genes and to develop a novel RNA-seq-based risk stratification, iv) to analyse potential and especially actionable therapeutic targets regarding expression, alternative splicing as potential mechanism of resistance, and mutations, and v) to prospectively test these theoretical target analysis strategies in a consecutive large patient cohort in three clinical relevant examples: first, in terms of BCMA, in the basis for the development of the T cell bispecific (TCB) antibody CC-93269 (formerly known as EM-801) [98, 212]. Second, in terms of the seemingly ideal target CD38 to assess the question, why two thirds of the patients lack of activity of anti-CD38 treatment despite seemingly ubiquitous expression on myeloma cells [213]. And third, in terms of CTAs as vaccination targets, analysing their expression pattern for clinical applicability. Figure 1.2 and the following text give an overview regarding the strategy to complete these tasks. DNA-microarray and RNA-seq expression data of in total 842 CD138-purified plasma cell samples from 798 patients (MGUS, AMM, MM and relapsed MM (MMR)), 18 samples of the B cell lineage (MBC, PPC and BMPC) and 26 samples of HMCLs shall be analysed. They are depicted in the thick, black coloured frames at the left side of figure 1.2. The previously established microarray analysis will be translated and compared to the RNA-seq-based assessment (light grey coloured arrows). The available data of myeloma patients will be splitted in a training

(TG, 194 MM samples), validation (VG, 108 MM samples) and testing group (TeG, 233 MM samples). The risk stratifications and classifications based on microarray expression data will be transferred and adjusted to RNA-seq expression data within the TG (blue coloured boxes). The impact on event-free and overall survival in MM will be assessed for all stratifications and validated (brown coloured box) on four independent data sets: TeG, AMM, MMR data and the external CoMMpass cohort (n=767) [51, 169]. RNA-seq expression will be used in combination with the overall survival data to discover novel prognostic genes. These genes will be used to generate a novel RNA-seq risk stratification, using the score developing pipeline of Rème *et al.* [197] (yellow coloured boxes).

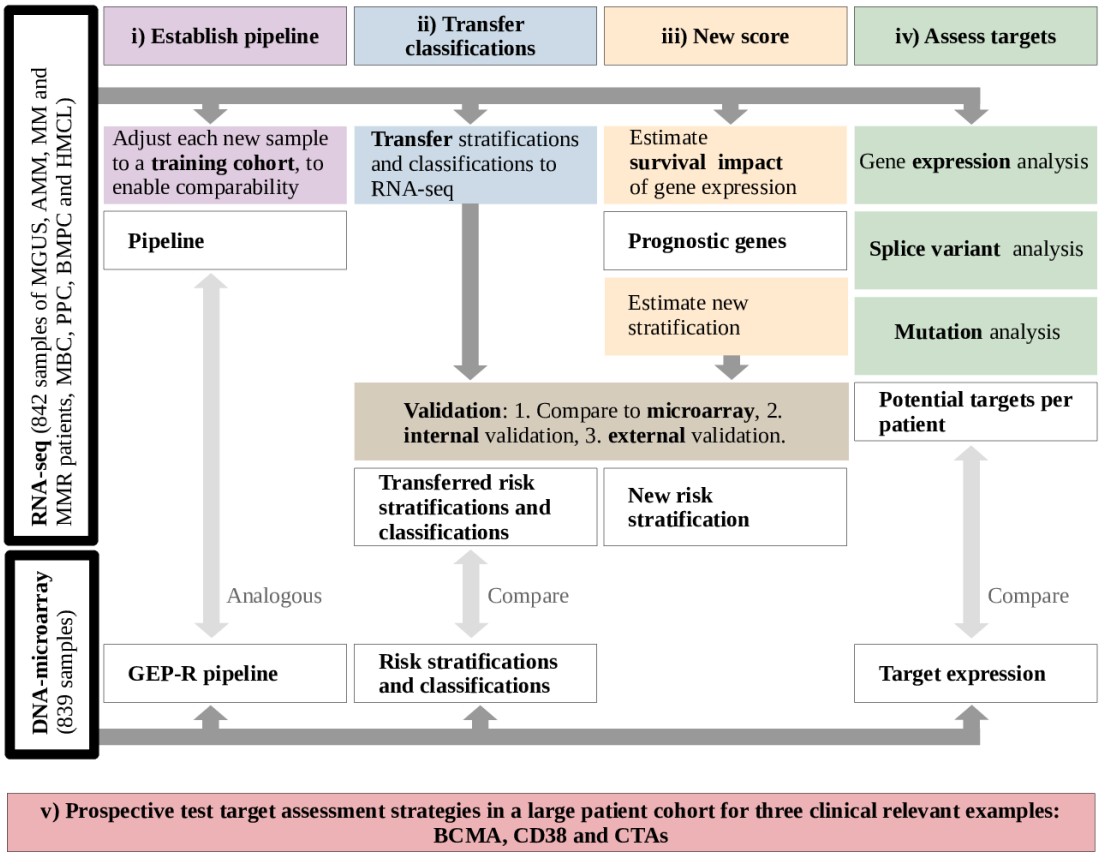

Figure 1.2: Aims of the thesis. This thesis has five aims, which are depicted with small Roman numbering at the top (and bottom) of the figure. The five aims are further depicted with coloured background. For detailed description, see text. Violet: creation of a pipeline for new samples, analogous to the GEP-R. Blue: transfer of DNA-microarray risk stratifications and classifications. Yellow: creation of a new score, based on detected novel prognostic genes. Green: analyse of targets regarding expression, splice junctions and mutations. Red: assessment of the strategies in practical examples. RNA-seq: RNA-sequencing; GEP-R: GEP-report; MGUS: monoclonal gammopathy of undetermined significance; AMM: asymptomatic multiple myeloma; MM: multiple myeloma; MMR: relapsed MM; MBC: memory B cell; PPC: polyclonal plasmablasts; BMPC: bone marrow plasma cell; HMCL: human myeloma cell line; CTA: cancer testis antigen.

The assessment of targets includes the setup of expression, splice junction and mutation analysis (depicted in green coloured boxes). The target expression determined by RNA-seq will be compared to their expression on microarrays. Splice junction analysis shall consecutively be used for the assessment of alternative splicing of e.g. CD38, which is a potential explanation for lack of activity of anti-CD38-based treatment in a large subgroup of myeloma patients.
Mutation analysis will be performed for the targetable BRAF mutation V600E [8]. Patients with present mutation are treatable by the inhibitors vemurafenib [24, 109] and dabrafenib [91].

2 Materials and methods

In the following chapter, available data and applied statistical methods are depicted. Subsequently, the steps of the thesis pipeline are described in detail (see figure 2.1). The pipeline comprises the analysis of microarray and RNA-seq data for **A** the pre-processing (see section 2.3), **B** the classification creation (see section 2.5 and 2.6), **C** the classification calculation (see section 2.4 and 2.7), and **D** the target analysis (see section 2.8).
Computations were performed on Ubuntu 14.04 and Windows Server 2012 using open source software. The used tools and their versions are listed in supplementary table B.5. Command line tools were used on Ubuntu, while the analyses with the programming language and statistical software framework R [188] were performed on Windows. R-Studio [203] was used as integrated development environment for R. R provides a collection of basic pre-built functions for data processing, extensible via "packages", which contain libraries of R functions. Additional R packages were installed using the Bioconductor software project [72]. In the following, used functions are denoted by corresponding packages, unless the function belongs to a base package (e.g. stats or graphics) [20, 188]. All package versions are listed in supplementary table B.6.

2.1 Materials

In this section the molecular profiling procedures applied at the LfM at the University Hospital Heidelberg are described in detail. For further information regarding sample processing see Hose [97] and Seckinger *et al.* [211]. An overview is depicted in figure 1.1. Results of the molecular profiling analyses, i.e. raw data, including DNA-microarray CEL files, RNA-seq FASTQ files and the chromosomal aberrations, are used as starting point for this thesis. In the second part of this section, the available data and the number of samples are summarised and applied annotations are depicted.

2.1.1 Molecular diagnostics

Cell purification and preparation. Bone marrow was aspirated from the iliac crest of healthy donors and patients (see also figure 1.1). Both normal and malignant bone marrow plasma cells were purified in two steps. First the mononuclear cell fraction was separated by standard protocol of density gradient centrifugation over Ficoll Hypaque (Biochrom) and cells were counted. Then, the cells were purified with anti-CD138 immuno-microbeads using an automated magnetic-activated cell sorter (autoMACS Pro; Miltenyi Biotec) [97, 99, 101, 212]. The negative fraction of the CD138-

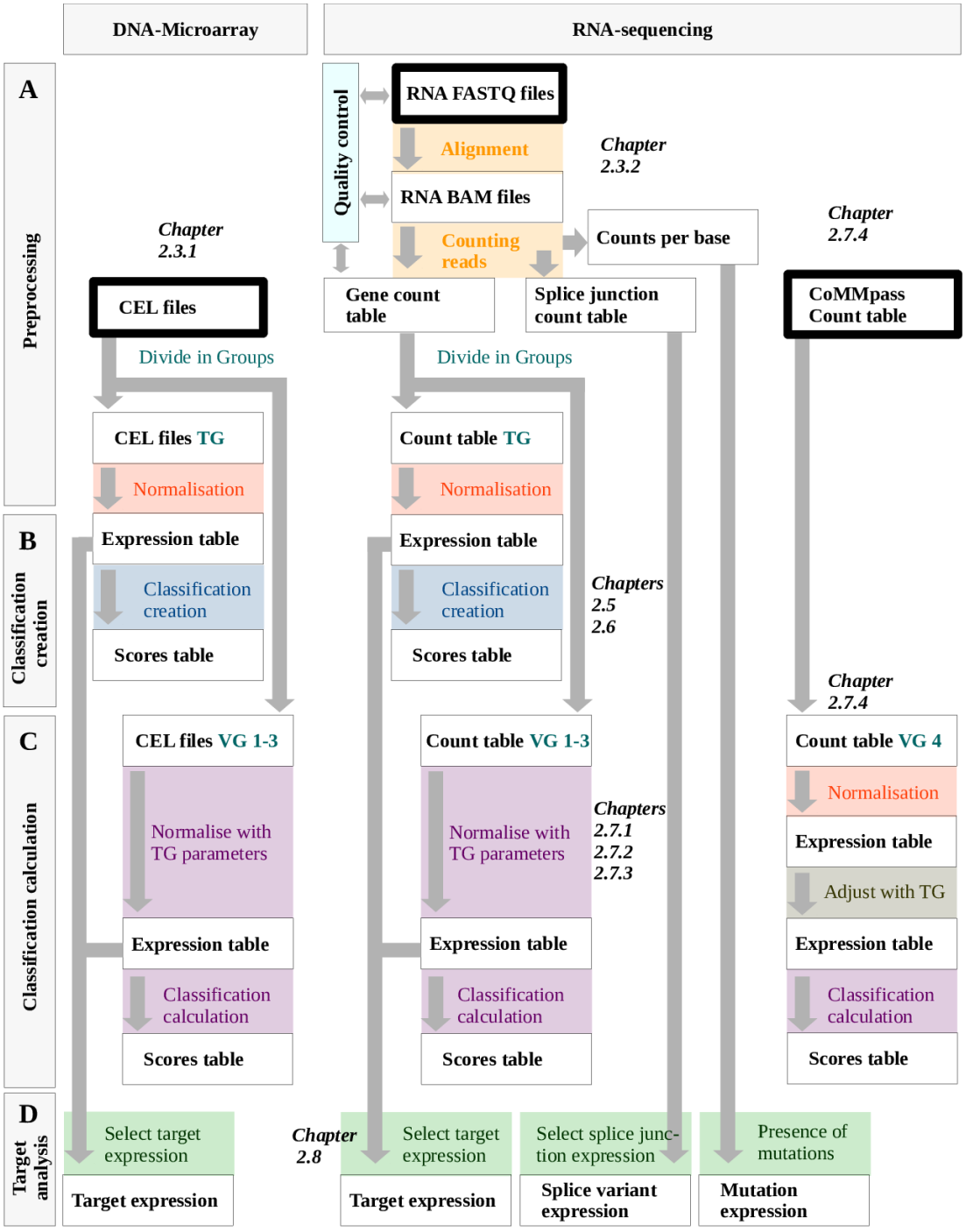

Figure 2.1: Flowchart of the bioinformatic pipeline in this thesis. Development of an RNA-sequencing stratification and transfer of microarray stratifications, classifications and target assessment to future RNA-sequencing (RNA-seq) technology. The raw starting data are depicted in thick, black frames. The samples are divided in training (TG) and validation groups (VG), depicted in cyan coloured text. The processing steps include the preprocessing (**A**), the stratification and classification creation (**B**) and calculation (**C**), and the target analysis (**D**). Subsections are depicted with coloured background. Yellow colour: Preprocessing of raw RNA-seq FASTQ files to read count table. Red colour: Normalisation of microarray and RNA-seq data. Blue colour: Stratification and classification creation on the TG, for microarray and RNA-seq. Violet colour: Normalisation and stratification and classification estimation of the VGs, for microarray and RNA-seq. Four validation groups were consecutively processed: an internal VG, an internal test group, early stages, relapsed myeloma samples and the external group CoMMpass. The latter is separately adjusted, depicted in brown colour. Green colour: Target analysis regarding expression, splice variants and mutations. Light blue colour: Quality control is performed at three processing steps.

purification was used in selected cases to generate HMCLs either at the LfM of the University hospital Heidelberg (HG1, HG13, and HG19 [97, 213]) or at the Centre Hospitalier Universitaire Montpellier (XG1, XG2, XG3, XG4, XG6, XG7, XG11, and XG13 [165, 255]). Other HMCLs were obtained commercially (L363, SK-MM-2, LP-1, OPM-2, U266, RPMI-8226, AMO-1, JJN3, KARPAS-620, KMS-12-BM, KMS-11, NCIH-929, MOLP-8, KMM-1, and EJM [97, 164, 213]).

The purity of the positive fraction, including the percentage of plasma cells, was controlled with flow cytometry using a fluorescence-activated cell cytometric sorter (FACSCalibur; Becton Dickinson). For this monoclonal antibodies against CD38 (clone HB-7, FITC labelled; Becton Dickinson) and CD138 (clone B-B4, PE-labelled; Milteny Biotech) were utilised. The CD138-purified plasma cells are used at the LfM to assess genetic alterations, using iFISH (in collaboration with the department of human genetics, University Hospital Heidelberg), and to determine gene expression, using DNA-microarrays [100, 175, 217] and RNA-seq. These methods are described in the following sections.

In addition to bone marrow analysis, peripheral blood is examined. MBCs were extracted from peripheral blood and differentiated *in vitro* to PPCs [97, 164, 213].

iFISH. Chromosomal alterations in the CD138-purified plasma cells can be assessed by iFISH. Specific fluorescence DNA probes are generated and hybridised with chromosomes in malignant plasma cells spinned on glass slides [175, 176]. In this thesis, iFISH probes for the assessment of numerical alterations of the chromosomal regions 1q21, 5p15, 5q31 or 5q35, 8p21, 9q34, 11q13, 11q22.3 or 11q23, 13q14.3, 15q22, 17p13 and 19q13 were used. For the detection of IgH translocations an IgH-break-apart probe, t(4;14)(p16.3;q32.3), t(11;14)(q13;q32.3), and t(14;16)(q32.3;q32) was used (Poseidon Probes, Kreatech) [97]. The hybridisation of CD138-purified plasma cells was performed according to the manufacturer's instructions (Kreatech and MetaSystems) [99]. Per probe, at least 10% of all used cells have to be altered to count a gain, deletion or translocation. A threshold of 60% was used to distinguish subclonal (<60%) from clonal ($\geq$ 60%) aberrations [97]. The ploidy status was assessed, using the score of Wuilleme *et al.* [251]. This defines patients as HRD, in case of gains of two or more of the three chromosomes 5, 9, 15 [251].

iFISH can also be used to assess the minimum percentage of malignant plasma cells within a given sample by dividing the number of cells carrying a myeloma specific aberration (as depicted above) by the total number of counted cells [175].

The information about genetic alterations is available as CSV table, containing the aberration status for each patient.

DNA-microarrays. A DNA-microarray consists of quadratic areas, each containing

a probeset, which consists of about one million fixed copies of one 25-mer DNA-strand, called "probe" [97]. The probe is designed to hybridise with a specific mRNA in human cells [97]. Several genes (>55% of all genes present on Affymetrix U133 2.0 plus GeneChips) are represented by more than one probeset. For this analysis, RNA is extracted from cells with the AllPrep DNA/RNA Mini kit (Qiagen) [99, 213]. Quality and quantity are measured using the Agilent 2100 bioanalyzer (Agilent). Subsequently, in two amplification cycles, RNA is first reverse transcribed to cDNA, amplified and afterwards transcribed to labelled cRNA using biotinylated nucleotides based on the small sample labelling protocol vII (Affymetrix). The cRNA is fragmented and hybridised to U133 2.0 plus GeneChip microarrays (Affymetrix) [99–102, 212, 216, 217]. Fluorescence intensity of the DNA-microarrays is scanned. For each spot the fluorescence intensity correlates with the number of transcripts bound to the specific probe. Resulting images are analysed by microarray image analysis software and saved as a so called "CEL" file. This text-based file contains position, calculated intensity values, standard deviation and number of pixels for each probe.

RNA-seq. RNA-seq allows the analysis of the transcriptome using high-throughput massive parallel sequencing by synthesis. RNA-seq is performed using the protocol optimised for low input analysis of CD138-purified plasma cells [211]. The extracted total RNA (5 ng, minimum 10 pg) is used to generate the full-length double-stranded cDNA [209, 211, 212]. For this, first-strand cDNA is synthesised, applying the SMARTer Ultra Low RNA Kit (Clontech laboratories, Illumina), and purified using SPRI AMPure XP Beads (Beckman Coulter). Subsequently, the double-stranded cDNA is amplified by long-distance PCR and again purified using SPRI AMPure XP Beads (Beckman Coulter). The Agilent 2100 BioAnalyzer and the Agilent High Sensitivity DNA Kit (Agilent Technologies) are used to quantify and validate the purification. Full-length cDNA is randomly sheared into smaller fragments using the Covaris system, and 10 ng are used for library preparation, according to the Illumina Sequencing protocol (New England BioLabs) and using the NEBNext ChIP-Seq Library Prep Master Mix Set. [211]

Sequencing is performed with 2x50-bp or 2x75-bp paired-end unstranded RNA-seq on an Illumina HiSeq2000 [211]. For this, a so called "flow cell" is used, which is a specific glass slide with covalently attached high-density forward and reverse primers [158]. It contains eight lanes of two columns each built of 96 (2x50-bp RNA-seq) or 432 tiles (2x75-bp RNA-seq). Every patient is sequenced on one lane, which results in 2x96 or 2x432 images per patient, each showing the randomly distributed sequence fragment clusters. The output images of the sequencer are converted to two FASTQ files per patient, using the Illumina tool bcl2fastq [211]. FASTQ files are text-based

files and contain the nucleotide sequences of the fragments (so called "reads"), as well as a quality string consisting of one quality score per base of every sequence [46]. This base quality score indicates the probability that the corresponding base call is wrong, and it is encoded in an Illumina format, ranging from 0 to 93 [46].

2.1.2 Patients and samples

All patients have been diagnosed according to standard criteria [60, 62, 111]. MM is classified according to IMWG [111, 190]. The designation AMM includes patients with an amount of M-protein of $\geq$30 g/L and/or a plasma cell infiltration in the bone marrow of $\geq$10%, as used and defined by Seckinger and Hose [214]. In total, 729 CD138-purified plasma cell samples from 52 MGUS, 142 AMM, 535 MM patients, and 10 healthy donors from the university hospitals of Heidelberg and Montpellier have been analysed. All patients were previously untreated. Additionally, 69 samples of MMR were used. This cohort is called "HD cohort" in this thesis. All patients have given their written informed consent in accordance with the Declaration of Helsinki between January 2002 and February 2015. The ethics committee of the Medical Faculty of the Ruprecht-Karls-University Heidelberg and the Centre Hospitalier Universitaire Montpellier have approved the studies (ethic vote no. 229/2003 and S152/2010). All available samples are summarised in table 2.1.

Table 2.1: Patients, samples and investigations. Depicted is the cohort size per analysis. Event free survival (EFS) data for relapsed patients were not used due to heterogeneity of second line treatment protocols. BMPC: Bone marrow plasma cell; MBC: Memory B cell; PPC: polyclonal plasmablasts; MGUS: monoclonal gammopathy of undetermined significance; AMM: asymptomatic multiple myeloma; MM: multiple myeloma; MMR: relapsed multiple myeloma; HMCL: Human myeloma cell line; iFISH: interphase fluorescence *in situ* hybridisation; RNA-seq: RNA-sequencing; OS: overall survival; SJ: splice junction; n: number of samples; NA: not available.

	n	**iFISH**	**Microarray**	**RNA-seq**	**EFS**	**OS**	**SJ**
all	842	798	839	842	673	602	512
MBC	4	NA	4	4	NA	NA	4
PPC	4	NA	4	4	NA	NA	4
BMPC	10	NA	9	10	NA	NA	9
MGUS	52	52	52	52	NA	NA	52
AMM	142	142	142	142	142	NA	29
MM	535	535	534	535	531	534	388
MMR	69	69	69	69	NA	68	0
HMCL	26	NA	25	26	NA	NA	26

Bone marrow aspirates have been purified from 798 patients (52 MGUS, 142 AMM, 535 MM, 69 MMR). For all patients iFISH data were generated and are available.
CEL files from U133 2.0 plus GeneChip microarrays are available for 797 patients (52

MGUS, 142 AMM, 534 MM, 69 MMR) and for 9 healthy donors. RNA-seq FASTQ files from paired-end and unstranded Illumina HiSeq2000 analyses were generated and are available for 798 patients (52 MGUS, 142 AMM, 535 MM, 69 MMR) and 10 healthy donors.

Overall- and/or event free survival (OS and EFS) data are available for 744 patients (142 AMM, 534 MM, 68 MMR). An event for EFS analysis is defined as disease progression or death of the patient. For OS analysis, all cases of death are considered as event. Survival data include the status and the survival time of a given patient. The status is 0 if no event occurred and 1 if an event occurs. The survival time is defined as the period from the date of bone marrow extraction to the date of the last visit or, if an event occurs, the date of the event. The survival time is depicted in months. Patients undergoing an allogenic stem cell transplantation are censored four weeks after previous autologous stem cell transplantation. For these patients the last available status and survival time, before the censoring event occurred, is used.

Additional data are available for 4 MBC, 4 PPC and 26 HMCL samples. Unlike myeloma and MGUS patient data, MBC and PPC data for RNA-seq and DNA-microarray are not based on the same donor samples due to sample availability.

All available clinical parameters, as age, ISS stage or light chain type, are listed in table 2.2 and the number of patients for which each parameter is available is given. If a parameter was not determined or applicable, it is depicted as "NA" (not available). In supplementary table B.7, the characteristics are depicted for the TG, VG and TeG. No significant differences in distribution of the variables could be found (Pearson's chi-squared test, see section 2.2.2).

For risk stratification and molecular classification training, validation and testing the available data were divided in three groups. This is necessary to prevent overfitting, which occurs, if a stratification or classification is perfectly adjusted to the training set and thus well applicable in this set only, i.e. not to a new set. Hence, the data are fitted on the training group (TG), validated on the validation group (VG) and tested on an internal "hold out" set, which is not included in stratification or classification training and creation (TeG). An example for the impact of overfitting can be seen in the first calculations of the RS-seq, see section 4.2.3. The three groups were divided, using the `sample` function in R [20, 199]. The number of samples in each group was specified. The initial proportions have been 48% of the samples in the TG and 26% in the VG and the TeG, respectively. Due to the increase of the number of available samples over time, the current size of the TeG is comparable to the TG (see table 2.3). Two different TG were used. TG 1 includes 4 MBC, 4 PPC, 9 BMPC, 26 MGUS, 26 HMCL, 194 symptomatic and 19 asymptomatic myeloma samples. Nine BMPCs were used, as the

Table 2.2: Patient characteristics. MGUS: monoclonal gammopathy of undetermined significance; AMM: asymptomatic multiple myeloma; MM: multiple myeloma; MMR: relapsed multiple myeloma; ISS: International staging system; R-ISS: revised ISS; NA: not available; NR: not reached; n: number of patients. For MM the median overall survival is depicted in months.

		MGUS		**AMM**		**MM**			**MMR**	
Variable	**Level**	**n**	**%**	**n**	**%**	**n**	**%**	**Median survival [months]**	**n**	**%**
Sex	female	28	53.8	69	48.6	218	40.7	90.9	27	39.1
	male	24	46.2	73	51.4	317	59.3	88.8	42	60.9
Age [years]	≤60	21	40.4	65	45.8	287	53.6	103.5	24	34.8
	>60	31	59.6	77	54.2	248	46.4	78.5	45	65.2
Monoclonal protein [g/l]	<20	48	92.3	59	41.5	102	19.1	90.2	24	34.8
	≥20	4	7.7	34	23.9	67	12.5	95.1	13	18.8
	≥30	0	0.0	41	28.9	280	52.3	90.	11	15.9
	NA	0	0.0	8	5.6	86	16.1		21	30.4
ISS stage	1	44	84.6	110	77.5	225	42.1	128.4	23	33.3
	2	3	5.8	15	10.6	161	30.1	80.7	14	20.3
	3	4	7.7	7	4.9	134	25.0	55.8	4	5.8
	NA	1	1.9	10	7.0	15	2.8		28	40.6
R-ISS stage	I	24	46.2	56	39.4	115	21.5	NR	11	15.9
	II	14	26.9	38	26.8	226	42.2	86.7	17	24.6
	III	1	1.9	4	2.8	70	13.1	41.5	2	2.9
	NA	13	25.0	44	31.0	124	23.2		39	56.5
Meta score risk	low	16	30.8	24	16.9	45	8.4	147.6	1	1.4
	medium	35	67.3	108	76.1	415	77.6	99.6	37	53.6
	high	0	0.0	0	0.0	59	11.0	37.2	3	4.3
	NA	1	1.9	10	7.0	16	3.0		28	40.6

last BMPC-sample was included at a later time. The TG 1 is used for the GPI, as its creation requires additional cell types. This cohort was also used for the t(4;14) classification, as the translocation is not biologically limited to MGUS or myeloma (e.g. presence in cell lines). TG 2 is used for all other stratifications and classifications and includes the 194 symptomatic myeloma samples. The validation group VG consists of 108 and the test group TeG of 233 symptomatic myeloma patient samples. TG 2 does not comprise other cell types, as processing these together with malignant plasma cell samples can skew the distribution (data not shown).

As external validation, the CoMMpass cohort data (version IA13) were used [51, 169]. This publicly available database from the MMRF comprises data for 767 newly diagnosed MM patients, including pre-analysed RNA-seq data, WGS data, survival data and predictions for translocations. The pre-analysis is depicted in figure 2.3. The MMRF have used the "Genome Reference Consortium Human Build 37" (GRCh37) as

reference genome. The following files were downloaded: clinical information, counts determined with HTseq [5] and aligned with STAR [58] against the human genome and OS and EFS survival data per patient.

Table 2.3: Patients, samples and investigations of TG, VG and TeG. Overview of samples per cohort in **a** training group (TG) 1 and **b** TG 2 and **c** the validation (VG) and test group (TeG). MBC: memory B cell; PPC: polyclonal plasmablasts; BMPC: bone marrow plasma cell; MGUS: monoclonal gammopathy of undetermined significance; AMM: asymptomatic multiple myeloma; MM: multiple myeloma; MMR: relapsed MM; HMCL: human myeloma cell line; iFISH: interphase fluorescence *in situ* hybridisation; RNA-seq: RNA-sequencing; DNA-seq: DNA sequencing; EFS: Event free survival; OS: Overall survival; n: number of samples; NA: not available.

	Set	Entity	n	iFISH	Microarray	RNA-seq	EFS	OS
a	TG 1	MBC	4	NA	4	4	NA	NA
	TG 1	PPC	4	NA	4	4	NA	NA
	TG 1	BMPC	9	NA	8	9	NA	NA
	TG 1	MGUS	26	26	26	26	NA	NA
	TG 1	AMM	19	19	19	19	19	NA
	TG 1	MM	194	194	193	194	193	194
	TG 1	HMCL	26	NA	25	26	NA	NA
b	TG 2	MM	194	194	193	194	193	194
c	VG	MM	108	108	108	108	105	107
	TeG	MM	233	233	233	233	233	233

2.1.3 Annotations

The "Genome Reference Consortium Human Build 38" (GRCh38) was used as human genome reference. Corresponding data were downloaded from the genome database and browser Ensembl [253]. The detailed download links are listed in supplementary table B.4. The primary assembly of the human genome sequence GRCh38 (release 77), not containing alternative sequences, has been downloaded in FASTA format. FASTA is a text-based format for storing sequences with their description, e.g. the chromosomal location. For annotating the human genome, a "Gene Transfer Format" (GTF) file (release 82) was used. It includes feature descriptions, e.g. information for each gene, transcript or exon.

2.2 General statistical methods

In this section, the general statistical methods used in this thesis are explained. It includes a description of comparison methods, significance tests and graphical representation of the data. Each method is selected based on the data type: survival data, categorical data or continuous data.

In general, p-values were used as statistical confidence measure. In case of several p-value estimations, defined in this thesis as at least 20 comparisons in the same cohort

(see section 4.1.5), there is an increasing probability of observing significant p-values by chance. Hence, the p-values were adjusted for multiple testing. Two methods are commonly used. Applying the Bonferroni adjustment, a significant p-value has to be smaller than the significance threshold α divided by the number n of performed tests [179]. Using the false discovery rate (FDR), all p-values are ranked and divided by the respective percentile rank [179]. The FDR is the proportion of false positive amongst all predictions. The latter approach was used in this thesis. For this, the function `p.adjust` was used with the Benjamini-Hochberg method (BH), which controls the FDR [21]. An effect was considered as statistically significant if the p-value or the adjusted p-value (where applicable) of its corresponding statistical test was smaller than 5%, unless it is stated otherwise.

2.2.1 Survival data

Survival analysis is the examination of events in a population over time. Events are the death of a patient (for OS and EFS) or disease progression (for EFS only, see section 2.1.2). Survival analysis was performed using the R package survival [228, 229]. A survival-object was generated with the function `Surv`, using the survival time and the status of the patients. The survival-object was used as *response* variable in the model formulas, used in the following functions. In univariate analysis, where only one variable is considered, only this covariant was entered in the formula (*response* $\sim$ *covariant*). In multivariate analysis the effect of several covariants is investigated. Hence, the extended formula *response* $\sim$ *covariants* was used, while several covariants were concatenated with a $+$ sign.

Differences between population subgroups were tested using the log-rank tests of the G-rho family of tests [88] by the R function `survdiff`. The resulting chi-squared statistic (*chisq*) was used to calculate the log-rank p-value with `pchisq` [20] by `1-pchisq(`*chisq*`,` df`)`. The degree of freedom df corresponds to the number of groups in the survival analysis minus one.

The Kaplan-Meier method [117] in the R function `npsurv` in the package rms [87] was applied to calculate the survival curves by calculating nonparametric survival estimates for censored data [65]. The median survival time was calculated using additional function `quantile` [20, 110] with a probability of 0.5.

Survival was plotted using the function `survplot` of the rms package Harrell [87]. Time is depicted in months and the p-value is based on the log-rank test of the G-rho family of tests.

Hazard ratios were determined for the univariate and the multivariate OS and EFS analysis, using Cox's proportional hazard model [6, 229]. A hazard rate is the frequency

of an event (e.g. death) during a time period. Therefore, the hazard ratio depicts the proportion of two hazard rates. A hazard ratio less than 1 indicates a decrease of the hazard and a ratio larger than 1 indicates an increase. The hazard model was estimated with the R function `coxph` [6, 229] in the package survival with default options. This function results in a coxph-object, including hazard ratios with a 95% confidence interval. Regarding categorical data, the ratio is always calculated comparing the first group against the other levels, with the first group always being the low risk group. Regarding continuous data, the model was estimated excluding the first 24 months for OS and the first 18 months for EFS, as with them the hazard functions were dependent on time. Fulfilling this "proportional hazard assumption" is a prerequisite for model validity. The assumption was verified with the function `cox.zph`, which performs a Schoenfeld residuals chi-squared statistic for each covariate [80, 229]. The assumption of proportional hazards has to be discarded, if the p-value is significant ($p \leq 0.05$).
The coxph-object also includes the results of two p-value tests for the hazard ratios: the Wald test calculates a p-value per level [241], and the log-rank test calculates a p-value over all values for the hazard ratios [88].
The univariate and multivariate hazard ratios are illustrated as forestplot, using the function and package `forestplot` [78]. This depicts the hazard ratios with their 95% confidence interval and the p-values of the Wald test per level. A hazard ratio is independently predictive, if its p-value is significant in multivariate analysis.
Three measures were used in this thesis to evaluate the performance ("goodness of fit") of a risk prediction model: the Brier score, R^2 and concordance statistic.
Brier score. The function and package `pec` in R [161] was applied to calculate the Brier score. The lower the score, the better is the prediction model. The survival object (*surv.obj*) was used as input. The option `cens.model` was used to determine the survival data with the Kaplan-Meier method and `multiSplitTest=T` for the Van de Wiel tests [235]. Bootstrap re-sampling was performed for cross-validation of the result (`splitMethod="bootcv"`). The number of bootstrap samples `B` was defined as one third of the number of samples and the number of bootstrap samples for resampling `M` as two thirds of the number of samples. As time interval for validation and testing years with data of at least 10% of patients were considered for the Brier scores depicted in this thesis. This is a tested time interval of 0 to 72 months for EFS and of 0 to 108 months for OS (option `testIBS`). For risk stratification creation, using the TG, a maximum endpoint of 108 months for EFS and of 120 months for OS was used.
R^2. The function `R2` in the R package `pec` was used to calculate the R^2 according to Graf *et al.* [79] and Gerds *et al.* [73]. The option *what="BootCvErr"* and a time point of 108 months for OS and 72 months for EFS was used. According to Hielscher *et al.*

[92], this R^2 measure performs better than the R^2 measure of Nagelkerke [173], which is the default of the `coxph` function.

Concordance. The concordance is included in the coxph-object, calculated internally with `survConcordance`, and extracted, using the function `summary(coxph-object)` [228, 229]. The concordance ranges from 0.5 to 1, whereby a value of 1 indicates an excellent predictability of the model and a value of 0.5 indicates no predictability [196, 228, 229]. A typical concordance for survival data is 0.6 to 0.79 [196, 208, 228, 229].

2.2.2 Categorical data

The relationship between categorical variables (e.g. presence or absence of expression of a specific gene) was depicted in a confusion matrix (see table 2.4). In the top left of the matrix, the percentage of consistency (*CO*) is depicted, which was calculated by dividing the number of overlapping values by the number of all values.

Significant differences between categorical variables were determined using a Pearson's chi-squared test with Yates' continuity correction, applying the R function `chi.square` [2, 95, 184].

Table 2.4: Model of confusion matrix. Depicted are the consistencies (CO_1 and CO_2) and the non-overlap (nCO_1 and nCO_2) of microarray and RNA-sequencing (RNA-seq). The overall consistency (*CO*) is calculated by dividing the number of overlapping values by the number of all values.

$CO\% = \frac{CO_1+CO_2}{CO_1+nCO_1+CO_2+nCO_2} * 100$	**microarray**	
RNA-seq	CO_1	nCO_1
	nCO_2	CO_2

2.2.3 Continuous data

Differences of continuous variables (e.g. gene expression values) between groups (e.g. disease entity) were assessed by exact Wilcoxon rank-sum test [19], using the `wilcox.test` function in R.

A Jonckheere-Terpstra (JHT) test [114] was applied to test for ordered differences among groups using the R package clinfun [218]. If the number of analysed patients is larger than 100, the number of permutations for the reference distribution (`nperm`) was set to 1000 to obtain a permutated p-value.

The correlation of two paired variables was calculated using `cor.test` with method `m=pearson`, performing a Pearson's product moment correlation [23, 94]. The resulting correlation coefficient *r* is presented with its confidence interval. The correlation coefficient ranges from 0 to 1, where values below 0.1 indicate negligible correlation

and values above 0.9 very strong correlation [9, 195, 210].
Continuous data were graphically depicted in point plots, boxplots and a principal component analysis (PCA). Boxplots were generated with the `boxplot` function in R [20, 35, 171], in combination with the `jitter` function [35, 36]. Numbers at the bottom of the plot depict the number of points for each box.
The PCA [185] was performed to visualise a multivariate data analysis in a scatter plot. The PCA performs a dimensionality reduction. For this, a data matrix is approximated by describing it as a product of small matrices, the principal components. Using the first two principal components, which explain most of the variance between the data, reduces the matrix to its essential patterns and transforms it to a new coordinate system. In this thesis, PCA analysis was performed via the command `prcomp` [20, 151, 239]. The default options were used, with exception of the option `scale=TRUE` to receive uniform variances prior to the analysis.

2.3 Preprocessing

2.3.1 DNA-microarray preprocessing

For DNA-microarray analysis, the GEP-R pipeline developed by Tobias Meissner *et al.* at the LfM [156] was adapted. The pipeline is shown on the left side of figure 2.2.

2.3.1.1 Normalisation

The DNA-microarray gene expression data of the TG data were normalised with two methods, which both adjust for background intensities, e.g. optical noise and non-specific binding. First, GC-robust multi-array average [250] in the function `just.gcrma` (package gcrma [249]) was used, an R function considering GC-content and using robust multi-array average as normalisation method across chips. The parameters of the normalisation were saved, using the `preproc` and the `wrap.val` function in the docval package [123], which enables the normalisation of new samples with the same parameters. To normalise the VG and the TeG samples, the `wrap.val.add` function of the docval package [123] was used, applying the parameters from the TG. Second, the whole cohort was normalised with the `mas5` algorithm from the affy package in R [70], which normalises by scaling all samples to the same mean, which is 500 by default.

2.3.1.2 Presence of expression

Presence or absence of gene expression was determined for gcrma normalised data, using "Presence-Absence Calls from Negative Strand Matching Probesets" (PANP) al-

gorithm with `pa.calls` function in the package panp in R [247]. "Negative strand matching probesets" are probes without known hybridisation partner, supposed to represent the background noise. The PANP algorithm uses them to estimate for each probe a probability of present expression and the resulting p-values are used to stratify the probes in three classes: present (P), marginal present (M) and absent (A) expression. For the VG and the TeG samples, presence or absence of gene expression was determined with the modified PANP function of the GEP-R [156]. Present and absent expression determination using microarrays is referred to as "PA call" in this thesis.

2.3.1.3 Quality control

The GEP-R was used to control the quality of the microarray CEL files, comparing the quality parameters of a new file to a reference cohort [156]. Parameters include e.g. the number of expressed genes or the background noise, which should be similar in all samples. Furthermore, reproducibility, hybridisation performance, RNA degradation and the detection of artefacts is controlled [156].

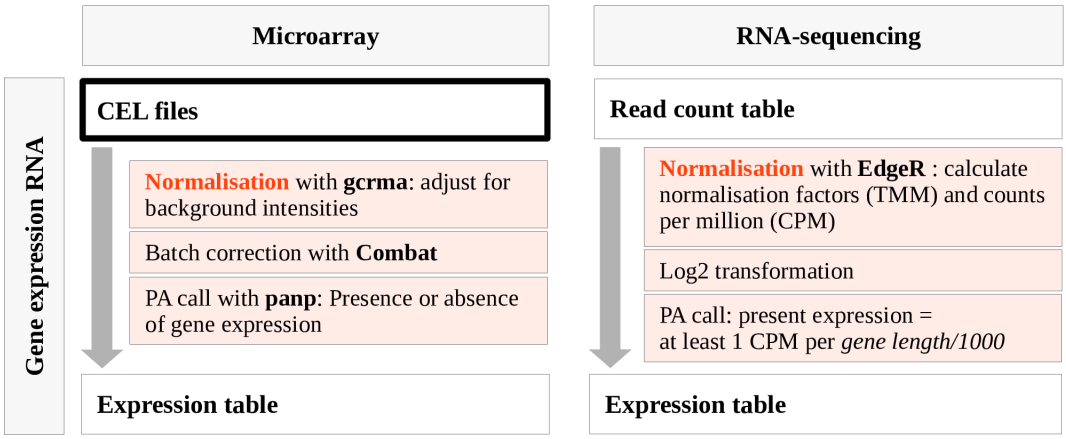

Figure 2.2: Flowchart of normalisation of RNA expression normalisation assessed by microarray and RNA-seq. Depicted are methods for: microarray and RNA-sequencing (RNA-seq) gene expression analysis used in this thesis. Left side: Microarray CEL files were normalised and batch corrected. Calls of presence and absence of expression were assessed with PANP function in R (PA call). Right side: RNA-sequencing reads were counted per gene and normalised by determining counts per million (CPM). Presence and absence of expression was assessed on RNA-seq with a threshold of one $CPM * 1000/gene\ length$. Both pipelines result in a gene expression table.

2.3.2 RNA-sequencing preprocessing

The development and application of the RNA-seq pipeline was part of this thesis. Three main steps have been performed for the general RNA-seq analysis pipeline: Alignment of the FASTQ files to a reference genome, counting reads, and normalisation. In figure 2.3 the preprocessing is depicted, and, on the right side of figure 2.2, the

normalisation strategy is shown.

The first step to process FASTQ files is to align the reads to a database or a reference genome in order to reconstruct the full-length transcripts and assign the sequencing reads to the annotated genes [133]. The second step is to count the reads aligning to a feature, e.g. a gene. A read may align to several genes and is then called "multiple mapping read". It can then either be counted for each gene, or not counted at all [5]. As genes can overlap, mapping reads may be ambiguous and not assignable to one gene [5]. The third step is to normalise the read counts. The comparison between different samples can be enabled, minimising biases like influence of sequencing depth or differences on the RNA composition among samples [28, 168]. The comparisons within a sample becomes possible by correcting the counts by e.g. gene length.

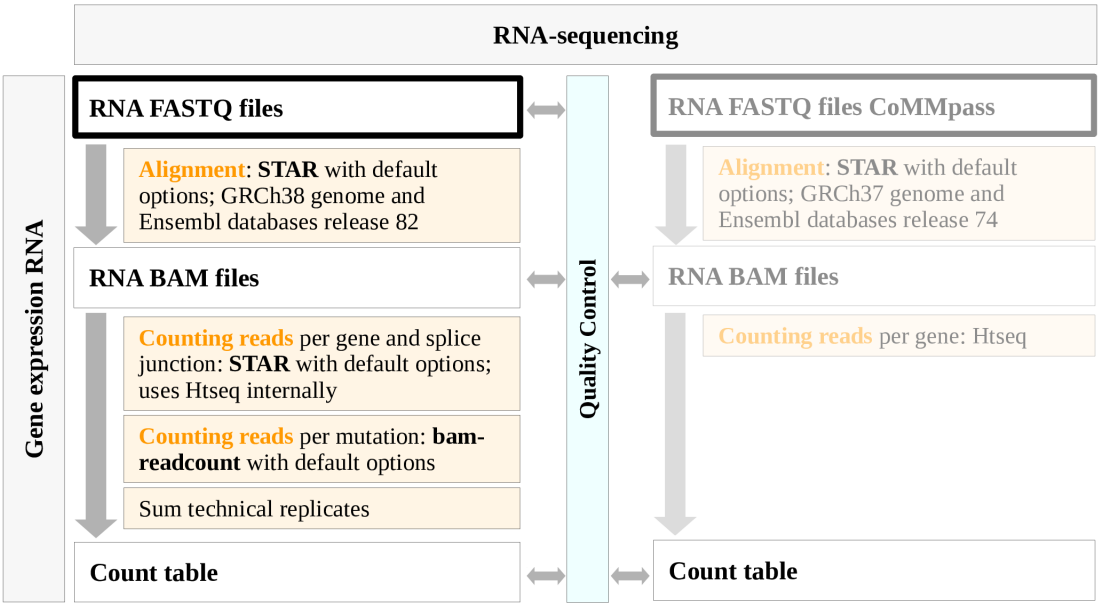

Figure 2.3: Flowchart of RNA-seq preprocessing. Left part: RNA FASTQ files from the Heidelberg cohort were aligned with STAR against the human genome GRCh38 and reads were counted. Quality control was performed before alignment, after alignment and after read count. Right part: The analysis of the CoMMpass cohort performed by the multiple myeloma research foundation (MMRF) is depicted in pale colours.

2.3.2.1 Alignment and read count per gene, per splice junction and per base

"Spliced Transcripts Alignment to a Reference" (STAR, [58]) is a fast, universal RNA-seq aligner developed by Dobin *et al.* [58]. It uses a two-step algorithm, first searching for the longest perfect matches (seeds), and then clustering the seeds and stitching the seed combinations with the best alignment score for a read. STAR includes the detection of chimeric and fusion alignments and the identification of splice junctions [58]. It internally uses `htseq-count` [5] for counting reads, either per gene or per transcript. By default, `htseq-count` is used in the mode `union`, counting only reads which can

be assigned to exactly one gene, i.e. multiple mapping reads are discarded. The results of STAR are aligned files in BAM or SAM format, read count files and various log files, usable for troubleshooting and quality control. The code belonging to this section can be found in supplementary code C.1. Reads were counted three times, per gene, per splice junction and at a specific position per base.
The reference FASTA file with the human genome sequence (GRCh38; release 77), recommended by STAR, and the most recent annotation GTF file (release 82) were used to generate genome indices for STAR with default options, while a parallelisation was performed (`-runThreadN 15`) (line 16). The genome indices are necessary to enable a fast mapping and have to be created only once for all alignments.
Reads in both paired-end FASTQ files of the samples were aligned against the genome indices. FASTQ files were uncompressed during the analysis (`-readFilesCommand gzip -cd`). To speed up computations, the alignment was parallelised (`-runThreadN 15`), the genome has been loaded and kept in memory between samples (`-genomeLoad LoadAndKeep`) and the RAM limit was expanded (`-limitBAMsortRAM 31532137230`). Reads were counted per gene (`-quantMode GeneCounts`) and splice junction and not mapping reads were saved in an extra FASTQ file (`-outReadsUnmapped Fastx`). Default options for counting reads with `htseq-count` exclude multiple mapping and ambiguous reads. The resulting binary BAM files are sorted by coordinate (–outSAMtype BAM SortedByCoordinate) (line 19). Subsequently, `samtools` [141] was used for generating an index (`samtools index`) (see line 22).
To quantify reads at a specific position, the tool `bam-readcount` [135] was used (see supplementary code C.1 line 25). It counts the aligned reads at a single nucleotide position and returns the number of reads of the mapped base. The mapping quality of the read, the base quality and the number of mapping reads on the plus and on the minus strand is returned. Many reads mapping to only one strand may indicate PCR duplicates. Likewise, the average position of a nucleotide within a read is included. A position at the edges of a read indicates lower quality. This function was exemplary used to count the reads on chromosome 7 at position 140753336, using the STAR alignments per gene. The position in the example is known as protein mutation V600E or V600K (see "Short Genetic Variations database" [220], dbSNP identifier: rs113488022) (see also section 2.8.3).

2.3.2.2 Alignment and read count per transcript

For transcript quantification, reads were counted using RSEM, developed by Li and Dewey [140]. The code belonging to this section is depicted in supplementary code

C.2. RSEM quantifies both uniquely and multiple reads and visualises the results. The quantification was performed in three steps. First, a RSEM specific transcript reference was generated with RSEM, using the GRCh38 genome and the GTF file (see supplementary code C.2, line 9). The option `-star` was appplied, to create STAR compatible indices. Second, STAR was used to align the reads in both paired-end FASTQ files of the samples against the genome (line 12). The same options as for the gene count were applied, with the following exceptions: Reads were counted per transcript (`-quantMode TranscriptomeSAM`), the output was not sorted (`-outSAMtype BAM Unsorted`), and not mapping reads were discarded. Third, transcript expression was counted, using the RSEM function `rsem-calculate-expression` with default options, specifying that the samples are paired end (`-paired-end`) and performing parallelisation (`-num-treads 15`) (line 15).
Plots presenting the read depth for each transcript were generated with the RSEM function `rsem-plot-transcript-wiggles`, using the option `-show-unique` to differentiate between uniquely and multiply mapping reads (line 19). The transcripts for each gene (e.g. BCMA and CD38) are provided in a text file (line 18).

2.3.2.3 Normalisation

The read counts per gene were normalised in order to enable the comparison between different samples [28, 168], accounting for two main technical influences: RNA composition and library size. The former is the amount of reads per gene (see section 4.1) and the latter is the total number of mapped reads, which reflects the sequencing depths. For this, the Bioconductor package "empirical analysis of digital gene expression in R" (edgeR) [38, 154, 200] was used, a tool for differential expression analysis of count-based expression data from RNA-seq or similar technologies. The code belonging to this section can be found in supplementary code C.5.
To assess a read count table, the read counts for unstranded RNA-seq were extracted per sample from STAR output files ReadsPerGene.out.tab (column 2) and the columns were merged in a table of counts in R (rows=genes, columns=samples). For several patients, more than one RNA-seq analysis has been performed. These technical replicates were summed with `aggregate` in R (see supplementary code C.5, line 2) [4].
The table of counts was filtered by excluding genes with no counts in all samples (lines 5-6). Counts were normalised by calculating the normalisation factors with the weighted trimmed mean of M-values (TMM) (line 10), in order to account for the RNA composition. This method calculates the relative quantitative changes, also called log-fold changes, per gene between the samples. It calculates a scaling factor per sample which minimises the log-fold changes [201]. The counts per million (CPM) were com-

puted in order to remove sequencing depths bias with accounting for the library size and the scaling factors (line 11).
As the microarray gcrma normalised data are log2 transformed, log2 transformation was performed with a prior count of 1 (line 14) for RNA-seq data for comparability. The prior count was used to omit the usage of the undefined logarithm of 0 and a value of 1 was chosen for a better visualisation, as 0 stays 0 after performing the log2 transformation ($log2(0+1) = 0$).

2.3.2.4 Presence and absence of gene expression

On DNA-microarrays, presence or absence of gene expression was determined with the PANP algorithm in R [247] (see section 2.3.1.2).
On RNA-seq, one CPM normalised count adjusted by gene length was used as threshold for presence. For this, the gene length was determined three times, using the human genome GTF file (release 82). First, the stop position of each gene was subtracted from its start position. Second, the length of all exons per gene was calculated, subtracting the overlapping regions. Third, the length of all exons per transcript was calculated, subtracting the overlapping regions, and the length of all transcripts belonging to one gene was averaged (median length). The latter median gene length was divided by 1000 per gene and used as thresholds for CPM normalised expression values to distinguish between presence and absence per gene (see supplementary code C.17). See section 4.1.3 for the discussion of the gene length.
On microarrays, three "levels" of expression were distinguished: present (P), marginal present (M) and absent (A) expression (see figure 2.4). Marginal thereby defines a group with presence at a lower certainty. Using RNA-seq, two groups were delineated,

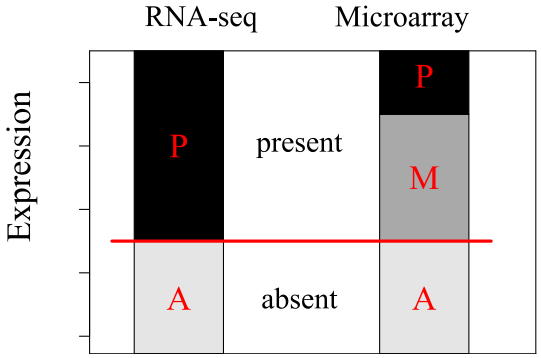

Figure 2.4: Model of presence and absence determination of gene expression assessed on microarray and RNA-sequencing (RNA-seq). On microarray three groups are distinguished: absent (A), marginal present (M) and present (P) expression. M depicts lower certainty of expression. On RNA-sequencing two groups are distinguished, A and P. For comparison between the two methods, the P and M group of microarray analysis are combined (depicted by the red coloured line).

A and P. The presence of expression was compared between microarrays and RNA-seq data. The microarray groups M and P were combined (MP) and compared to P group on RNA-seq. The calculation of the percentage of consistency is described in section 2.2.2. Present and absent expression determination using RNA-seq is referred to as PA-seq call in this thesis.

2.3.2.5 Quality control

Quality control was performed at three steps in the pipeline: before and after alignment and after summing up the technical replicates. First, the raw RNA FASTQ files were controlled with FastQC [49], designed for quality control of raw sequencing data from high throughput experiments. It provides an overview of unusual and possibly problematic areas in eleven categories by returning a "pass", "warn" or "fail" for each category. The categories are: "per base sequence quality", which provides an overview of the distribution of the quality scores among all nucleotides per position in the FASTQ file; "per tile sequence quality", which provides the average quality of each flowcell tile per read position; "per sequence quality scores", which provides the distribution of the quality scores among all sequences; "per base sequence content", which provides the average frequency of every nucleotide per read position; "per sequence GC content", which compares the density of the GC contents per read to a theoretical normal distribution; "per base n content", which provides the average number of "N" at each position of the read; "sequence length distribution", which provides the read length distribution; "sequence duplication level", which provides the density of the number of duplicates per sequence in comparison to the relative number of sequences; "over-represented sequences", which provides the most frequent (>0.1%) sequences; "k-mer content", which provides imbalanced sequences of 7 bases length (called 7-mer); and "adapter content", which compares the 7-mers to an adapter database. Detected over-represented sequences were aligned to the human genome, using the sequence similarity search tool BLAT [119] from Ensembl [253] with default options (see also discussion in section 4.1.2). For a quick overview, the results of FastQC are given in graphs and tables. FastQC was used in the non-interactive mode using the command line (see supplementary code C.3).

Second, the alignment was controlled with the STAR final log files. Files with less than 60 % mapped reads were discarded (discussed in section 4.1.2, see supplementary code C.4 lines 4 to 13).

Third, the read count table was controlled. For library size a threshold of at least 10 million reads [144] after summing up technical replicates was defined (see supplementary code C.4 lines 15 to 20).

2.4 Present stratifications and classifications

Patients were classified and stratified by existing methods using three types of input factors: conventional clinical factors, molecular alterations (iFISH), and gene expression derived factors (DNA-microarray). Clinical risk assessment was performed regarding the ISS as defined by Greipp *et al.* [84], and using tumour mass surrogates as M-Protein according to Kyle *et al.* [129]. Molecular alterations were used in combination with clinical factors in the R-ISS score [182]. Gene expression determined with DNA-microarrays was applied to assess biological variables as proliferation (GPI) [101], risk stratifications based on the expression of specific gene sets (UAMS70 [219], RS [197], EMC92 [124], and IFM15 [56]), and to classify multiple myeloma patients in different molecular disease subentities (TC [22, 43] and MC [254]).
Two different training groups were utilised (see section 2.1.2): TG 1 is used in stratifications and classifications, for which other cell types are necessary (GPI) or which are not limited to myeloma patients (t(4;14)), while in all other cases TG 2 is used.

2.5 Transfer of microarray-based stratifications and classifications to RNA-sequencing

All RNA-seq stratification and classification calculations are implemented as similar as possible to the original DNA-microarray-based assessment of proliferation (GPI [101]), of risk (UAMS70 [219], RS [197], EMC92 [124], and IFM15 [56]), and of molecular classification of myeloma (TC [22, 43] and MC [254]) (see section 4.2 for the discussion of the classification selection). For this, all genes of the underlying publications were used, and only cutoffs for the stratifications in the different groups were adjusted. The latter was necessary as evidently absolute values of scores do not overlap due to different underlying laboratory methods. The pipeline comparing the implementation of microarray stratifications and classifications with their implementation on RNA-seq is depicted in figure 2.5.
For the stratifications and classifications available in the GEP-R, the approach of Meissner *et al.* [156] for their calculation was used.
As described for microarray-based stratifications and classifications and in section 2.1.2, normalisations with TG 1 were used for GPI and t(4;14), while normalisations with TG 2 were used for all other stratifications and classifications. The validation group VG was applied to compare and validate the survival performances of risk stratifications and classifications. The test group TeG was used to confirm and validate their performance. The normalisation of the training groups was performed with edgeR as described in section 2.3.2.3.

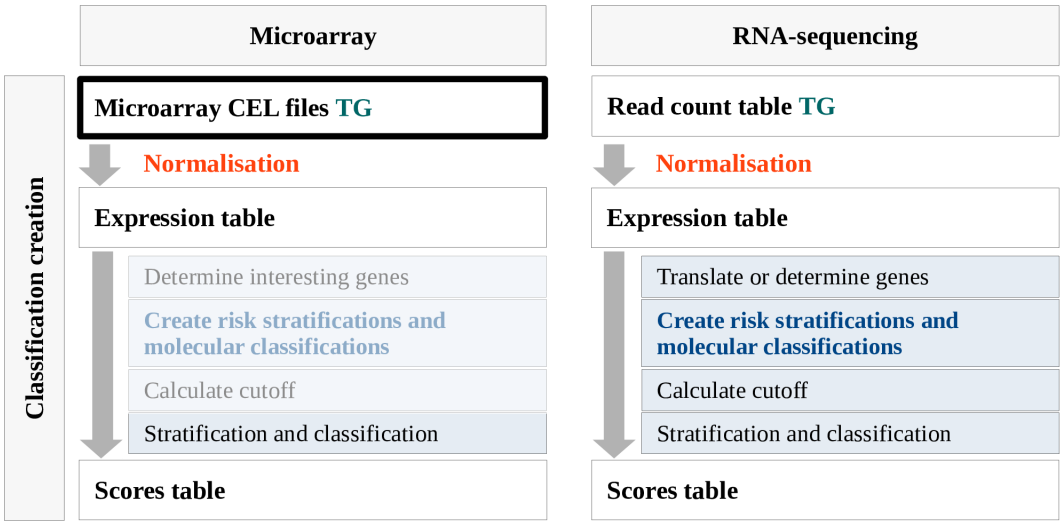

Figure 2.5: Flowchart of stratification and classification training. Right side: Development of the stratifications and classifications with the training group (TG) based on RNA-sequencing (RNA-seq). Left side: Development of the stratifications and classifications on microarray data, depicted in pale colours as they were present prior to this thesis. Both methods are as similar as possible. The normalisation is described in detail in figure 2.2.

The first step of all stratifications and classifications is the gene translation in order to use the initially described genes. The translation of the gene names was performed in R, using the `select` function in the package hgu133plus2.db [29]. Translations not present in the normalisation (e.g. due to lack of expression) were excluded. If one microarray probeset matched several "Ensembl gene identifiers" (ENSG) used in RNA-seq analysis, their expression values were added. If several probesets matched the same ENSG, it was only used one time, while associated values were averaged. For each translated gene of the stratifications, the RNA-seq expression was correlated with the microarray expression of the genes. The genes with a correlation $r \leq 0.15$ were excluded and genes with a correlation $r \leq 0.6$ were further controlled in two steps. First, the translation was controlled using the online GeneAnnot search tool [33, 34, 64], which is based on the GeneCards database [226]. In the database, specificity, sensitivity and the gene number per probeset is deposited. The "specificity" is defined as the number of probes in a probeset, matching to the target gene, divided by the number of probes, matching to any gene. The "sensitivity" is the number of probes in a probeset, matching to the target gene divided by all probes of the probeset. The "gene number" is the number of genes matching to the probeset. Genes with inconsistent translation were excluded (see section 4.1.4 for discussion of gene translation). Second, per gene, the percentage of samples with present expression in microarray and absent expression in RNA-seq has been determined (nCO_1, see section 2.2.2). The genes with $nCO_1 > 30\%$ and a correlation $r < 0.4$ were excluded.

2.5.1 Assessing proliferation (RPI)

Hose *et al.* [101] developed the gene expression-based proliferation index (GPI) at the LfM at the University Hospital Heidelberg. It includes 50 genes, which were selected by the gene ontology terms "cell proliferation" or "cell cycle". The genes were overexpressed in normal and malignant proliferating cells (HMCL and PPC) in comparison to non-proliferating cells (BMPC and MBC) [101]. The cutoff for the three risk groups was chosen in dependency of BMPC expression and the highest expression of the MM group [101]. The GPI score is determined by summing up the expression of the 50 genes, considering whether a gene is generally expressed above the background noise [101]. The GPI was developed as surrogate of a biological variable and it was not fitted to survival [101].
The translation to RNA-seq data, the RNA-seq-based proliferation index (RPI), was created in four steps. First, the 50 "GPI-genes" were translated into ENSGs. Second, the PANP algorithm, implemented for microarrays in the R package panp [246] was simulated on RNA-seq by using a threshold of one CPM normalised count, adjusted by gene length, as described in section 2.3.2.4. Third, analogously to microarrays, the sum of the expressions of the translated genes was calculated. Fourth, the two GPI cutoffs (*GPIcut*) were transferred from microarray to RNA-seq. For this, the GPI and the RPI of the TG 1, using MBC, HMCL, MGUS, AMM and MM samples, were correlated, and a linear regression line was fitted using $lm(RPI \sim GPI)$. The slope m and the y-interception value b of the regression line were used to determine the new cutoffs (*RPIcut*) with $RPIcut = m * GPIcut + b$. The samples were stratified in three groups.

2.5.2 Assessing survival

2.5.2.1 UAMS70-seq

The UAMS70 is composed of 70 genes associated with short survival in MM, which were identified using log-rank tests [219]. Of these genes, 51 were described "highly" expressed and 19 genes were "lowly" expressed in the underlying publication. Based on these genes, Shaughnessy *et al.* [219] defined a score delineating high risk and low risk myeloma. The high-risk expression pattern was similar to the one of HMCL and the low-risk expression pattern was similar to the one of MGUS and normal plasma cells [219].
Five steps were necessary to transfer the UAMS70 to the UAMS70-seq score on RNA-seq data. First, the 70 probesets were translated to 71 ENSGs. Second, the expression of the genes was centred, by subtracting the given microarray centre values from the expression values, for each gene. Third, the centred expressions were weighted and

the score was calculated in two steps by calculating the matrix-vector product of the expression values and a weighting matrix, and by multiplying the resulting values with a microarray weighting vector. Fourth, the risk scores were divided in three groups using a k-means clustering, which minimises intra-cluster variance (R function `kmeans` [67, 89, 145, 149]). The centre of each group was calculated. Fifth, the samples were stratified determining the shortest squared distance to the group centres. The low risk and the medium risk group were merged as in the original publication for microarray data [219].

2.5.2.2 RS-seq

Rème *et al.* [197] together with the LfM presented a gene expression-based risk score (RS) with 19 genes dividing patients into three groups. This score was fitted to an overall survival analysis. Prognostic genes were selected using a running log-rank test [197]. This algorithm performed survival analyses for every gene and selected the most predictive genes. In dependence of the expected number of deaths, a gene was associated with good or poor prognosis [197]. The score was calculated by subtracting the expression of "good" prognosis genes from the expression of "bad" prognosis genes. Two optimal cutoffs were selected, dividing patients into three groups [197].
Four steps were necessary to determine the RS-seq on RNA-seq. First, the 19 probesets of the original score were translated to ENSGs. Second, the expression of each gene was multiplied with a positive or negative "prognosis" factor which is -1 if the gene is associated with good prognosis according to the original RS and 1 if a gene is associated with poor prognosis. Third, the values were summed up. Fourth, new cutoffs were determined with the multi-cutoff running log-rank algorithm written by Rème *et al.* [197], using a set of parameters for a minimal size of a risk group (w), FDR (fdr), a chi-squared statistic threshold ($x2$) between two survival curves and a minimal number of events (cx) of 2. The function used is explained in detail in section 2.6. The samples were stratified in three groups, according to the best cutoff set. For estimating the cutoffs, the parameters w, $x2$, and fdr were varied: w from 18 to 37 in steps of 1 (representing 9 to 19% of patients, discussed in section 4.2.2), $x2$ from 0.01 to 0.05 in steps of 0.01, and fdr from 0.01 to 0.05 in steps of 0.01. In sum, the running log-rank algorithm ran 500 times.

2.5.2.3 EMC92-seq

Kuiper *et al.* [124] identified 92 genes associated with survival in MM. For this, they shrank the gene set by univariate Cox regression analyses and then used supervised PCA, resulting in a 92 gene survival signature (EMC92). Two groups were defined,

"standard risk" and "high risk". The cutoff was defined by using the proportion of patients with overall survival less than two years as cutoff for high risk [124].
Four steps were necessary to transfer the EMC92 to the EMC92-seq on RNA-seq data. First, the 92 genes were translated into ENSGs. Second, the normalised and log transformed gene expression per patient ($nc.log$) was standardised by mean variance, analogous to microarray standardisation. For this, the mean ($m.TG$) and the standard deviation ($sd.TG$) per gene over all samples of the TG were calculated and standardisation was calculated using the formula $\frac{nc.log - m.TG}{sd.TG}$. Third, the standardised values were multiplied with the original weighting scores and summed. Fourth, the cutoff was determined using the proportion of patients with OS of less than two years as group size of the high risk group. Samples were stratified in a standard risk and a high risk group. For simplification of inter-score comparison, the former group is called "low risk" group in this thesis.

2.5.2.4 IFM15-seq

The IFM15 of Decaux *et al.* [56] is based on 15 genes predicting poor prognosis. These have been identified using iterative univariate Cox analyses in combination with re-sampling and survival prediction of MM patients. An equation to calculate the score has been generated, based on PCA. Patients were divided in two groups, using the 75% quartile as cutoff [56].
On RNA-seq, the IFM15-seq was implemented in three steps. First, the 15 genes were translated to ENSGs. Second, the score was calculated from the normalised RNA-seq expression values using the original equation, with original weighting scores [56]. Third, scores were ranked and divided into quartiles, using the `quantile` [20, 110] function in R. The cutoff was calculated as 75% quartile and patients were stratified in two groups.

2.5.3 Assessing molecular entities

2.5.3.1 TC-seq

The TC classification [22, 43] is based on the expression of a set of nine genes (set *a*: *CCND1*, *CCND2*, *CCND3*, *FGFR3*, *MMSET*, *ITGB7*, *CX3CR1*, *MAF*, *MAFB*) and on a set of ten genes (set *b*: *TGFBI*, *CD14*, *CD163* (represented by two probesets), *FCGR3B*, *FCGR3A*, *CD5L*, *NDUFA2*, *CCL18*, *IK*, *TMCO6*) [156]. Patients are classified into 8 groups (4p16, maf, 6p21, 11q13, D1, D1+D2, D2, none) (see also section 1.4.2.3) [22, 43].
The TC classification was transferred in five steps. In contrast to all other stratifications and classifications, the TC-seq according to Chng *et al.* [43] was performed without

log2 transformation. First, the probesets of the two gene sets, set *a* and set *b*, were translated into ENSGs. Second, the geometric mean of the expression values of set *b* was calculated, by $exp(mean(log2(a+1)))$. Third, the median expression was determined for each gene of set *a*, which is called "control value" in the original paper. Fourth, new raw cutoffs were calculated by multiplying the median expression values with the given norm cutoffs. Fifth, the TC classification was calculated using the given equations in the underlying publication and in the GEP-R [43, 156].

2.5.3.2 MC-seq

The MC classification by Zhan *et al.* [254] distinguishes seven transcriptional signatures (MS, MF, CD1, CD2, HY, PR, LB) based on unsupervised hierarchical clustering. One hundred over- or under-expressed genes, identified by the nearest shrunken centroid, were used to generate a class predictor, which classified the samples with 98% accuracy [254]. Each subgroup has a characteristic expression pattern, mainly based on nine genes ((*MAF*, *MAFB*, *FGFR3*, *MMSET*, *CCND1*, *CCND2*, *CCND3*, *FRZB* and *DKK1*)) [254].

Two steps were necessary to create the MC-seq classification on RNA-seq data. First, the probesets were translated to ENSGs. Second, these genes were used in the R package pamr [90] to create a predictor.

This software "Prediction analysis of microarrays" of Tibshirani *et al.* [230] is a method for class prediction, based on gene expression data [230]. The objective of the tool is to find a subset of genes, whose expression values can predict the classes. The subset with the smallest prediction error and at the same time the smallest number of genes is selected. The package iteratively calculates standardised centroids for each class [230] for several sets of genes. A new sample is predicted by determining the closest (squared) distance to the centroids (also called nearest shrunken centroid). As input, a vector with the given classes and a dataset of gene expression values are necessary. Therefore, pamr can also be used on RNA-seq data. First, the predictor was trained using `pamr.train`. With `pamr.adaptthresh`, scaling thresholds for each group were estimated, to further minimise the number of genes. The training was repeated with the new scales. Afterwards, the classifier was cross-validated with `pamr.cv`. The results of cross-validation were plotted with `pamr.plotcv`. A cutoff for the best predictor was chosen by selecting the minimal misclassification error rate. If more than one minimal error occurred, the cutoff with the fewest number of genes was used. The genes are listed with `pamr.listgenes`. For a new sample, `pamr.predict` was used.

2.5.3.3 Translocation prediction

The prediction of translocation t(4;14) was generated using all genes of the RNA-seq normalisation. The ENSGs were filtered by variance ($variance \geq 0.5$). Then, the pamr package [90] (see above, section 2.5.3.2) a was used to create a predictor.

2.6 Novel risk assessment: HDHRS

One of the main aims of this thesis was the generation of a novel risk stratification using RNA-seq data. The resulting Heidelberg high risk score (HDHRS) for survival in symptomatic patients was generated on RNA-seq data according to the method published by Rème *et al.* [197] for microarray data, including the four steps explained by Rème *et al.* [197]: 1.) Normalisation , 2.) gene selection, 3.) score calculation and 4.) cutoff estimation.

1.) On RNA-seq data, gene filtering was performed before normalisation, as recommended in the edgeR user's guide [39]. One CPM in at least n patients is advised, with n between 2 and the smallest group size, which is 18 in the TG (see section 4.2.2). For clinical applicability, it was decided to use $n = 9$, which is half of the minimum group size. Hence, first, genes were filtered retaining only genes which had at least one CPM in nine patients. Genes were also excluded if the variance of CPM of all patients was less than 0.15 [197]. Then, the TG was normalised as described in section 2.3.2.3.

2.) Prognostic genes were selected using a running log-rank test. For this, a survival-object was generated with the function `Surv`. Expression values were sorted in increasing order for each ENSG, resulting in the vector *indx*. Each expression value in *indx* was used as threshold, dividing the patients into two groups. For each threshold, the differences between the groups were tested with the R function `survdiff`. The best expression threshold, resulting in at least two events per group and the maximal value of the chi-squared statistic was selected. The log-rank p-value was calculated with the function `pchisq` [20], multiplied by the number of samples, and adjusted with the function `p.adjust` and the method BH. The threshold for the corrected p-values was varied from 0.1 to 0.01 in steps of 0.0025 (see also 4.1.5). For each ENSG a prognosis factor was specified, which was -1 if the number of deaths in the high expression group was less than expected (good prognosis), and 1 in the opposite case (poor prognosis).

3.) The expressions of the ENSGs were multiplied with their prognosis factor. The values were summed, resulting in the HDHRS score.

4.) The risk groups were determined using the algorithm for risk group optimisation (multi-cutoff running log-rank algorithm) and the samples were classified in three groups. Patients were ordered according to their HDHRS score and divided into three groups multiple times. For this, the cutoff points s_i and s_j with $i \in [1, n-w]$ and

$j \in [i+w, n]$ were used. w is the minimal size of a risk group and was varied from 18 to 37 in steps of 1 (20 approaches, discussed in section 4.2.2). The factor w reduces the number of cutoff pairs per approach to $t = \frac{(n-w)(n-(1+w))}{2}$. For each cutoff pair, a survival-object was estimated with the function `Surv`. Differences between the groups were assessed using the R function `survdiff`. Three differences were calculated, resulting in three chi-squared statistics: between group 1 and 2 (cs_{12}), between group 2 and 3 (cs_{23}) and a global one (cs). The threshold ($x2$) in `qchisq((1-x2),1)` of cs_{12} and cs_{23} was varied from 0.01 to 0.05 in steps of 0.01 (5 approaches). The global chi-squared statistic (cs) was used to determine the log-rank p-values with function `pchisq`. The p-values were adjusted with the function `p.adjust` and the BH method. The threshold for the adjusted p-values (fdr) was varied from 0.01 to 0.05 in steps of 0.01 (5 approaches). As recommended by Rème *et al.* [197], the minimal number of events (cx) in each group was two. Hence, for each gene set the algorithm was running 500 times. For all cutoff pairs which fulfil the above thresholds, cs_{12} and cs_{23} were centred by subtracting their column means and scaled by dividing them by their standard deviations, using the `scale` [20] function in R. The minimum was determined from both scaled values, which was used to select the cutoff pair with the absolute minimal scaled value of the chi-squared statistics.
To describe the retained prognostic genes in the score, a reactome pathway analysis was performed. For this, the `select` function in the package hgu133plus2.db [29] was used to translate the ENSG to Entrez gene IDs and these were used in the package reactome.db [143] to determine the associated pathways. The number of genes per pathway was counted and divided by the number of genes with at least one database entry. The main pathways were selected and the pathways of the HDHRS were compared to the ones of the other stratifications.

2.7 Stratification and classification validation and testing

In the underlying GEP-R the "documentation by value" (docval) package [123] has been used to normalise new samples, using the parameters from the training cohort. This enables the analysis of new samples in clinical practice by ensuring the comparability to former samples (see also section 4.1.1). For RNA-seq data, a new strategy needed to be implemented normalising a new sample with the TG, enabling the use of the previously determined cutoffs. For this, new raw counts were attached to the raw counts of TG with the `cbind` function in R (see supplement C, code C.5, lines 3 to 16) and normalisation of the samples was performed with edgeR as described in section 2.3.2.3. Then, normalised counts of the new sample were extracted. This normalisation was performed for each sample of the VG and TeG twice, once with TG 1 and

once with TG 2. The complete normalisation function is shown in supplementary code C.6. The pipeline, comparing the calculation of the stratifications and classifications on microarray and on RNA-seq, is depicted in figure 2.6.
For each stratification and classification, an R function was developed, which was used for validation and testing. These functions are listed in supplement C (code C.7, C.8, C.10, C.9, C.13, C.15).

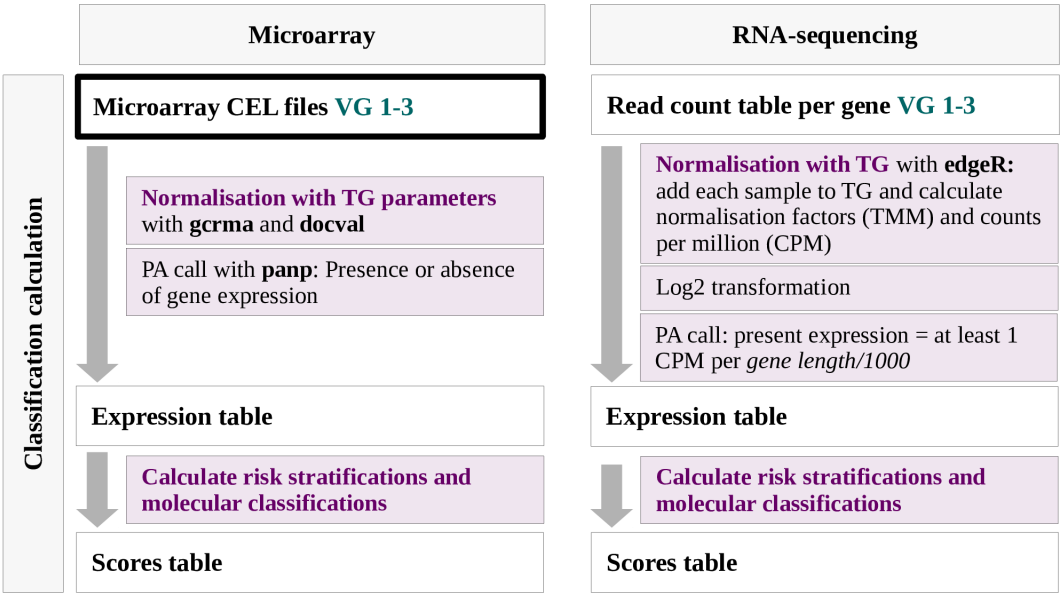

Figure 2.6: Flowchart of stratification and classification calculation. Usage of stratifications and classifications with the validation group (VG), testing group (TeG), early stages (AMM) and relapsed myeloma (MMR). Left side: Usage on microarray data. Calls of presence and absence of expression were assessed with PANP function in R (PA call). Right side: RNA-sequencing reads were counted per gene and normalised with the TG by determining counts per million (CPM). Presence and absence of expression was assessed on RNA-seq with a threshold of one $CPM * 1000/gene\ length$. Both methods are as similar as possible.

2.7.1 Internal validation

For the stratifications with multiple results from the TG (RS and HDHRS), the TG was used to reduce the initial sets and the VG was used to determine one final result. Validation of the risk stratifications was performed in two steps 1.) examining the proportion of the classes and 2.) comparing performance of survival analyses. Validation and testing of the molecular classifications were only performed in step one. The steps and strategy are discussed in section 4.2.2.

1.) examining the proportion of the classes

a In each class should be at least 9% of the patients, which are at least 18 patients in the TG and at least 10 patients in the VG (exceptions: MC, TC and t(4;14) classification) (see section 4.2.2).

b The proportions should be similar to microarray proportions: Each group should differ in less than 20 percentage points and the proportions of the stratifications should stay in the same order per score (visual inspection).

2.) comparing performance of survival analyses

a Survival curves were analysed graphically, excluding results with "wrongly" ordered curves and intersecting curves in an interval from 24 months to the last but one event per curve.

b The log-rank test of the survival analysis should be significant ($p < 0.05$).

c A Brier score analysis was performed for TG with a time interval of 0 to 120 months for OS and of 0 to 108 months for EFS, and for VG and testing groups of 0 to 108 months for OS and of 0 to 72 months for EFS. Stratifications with not significant Brier score for EFS and OS were excluded. For HDHRS, the Brier scores of EFS and OS had to be significant.

d Stratifications with a concordance below 0.6 for EFS and OS were excluded

RPI and risk-based stratifications (UAMS70-seq, RS-seq, EMC92-seq, IFM15-seq, HDHRS) were further validated in early stage patients (AMM), relapsed myeloma patients (MMR) and an external cohort (see section 2.7.3 and 2.7.4 below).

2.7.2 Independent testing on the TeG

The independent TeG was used for testing translated stratifications and classifications. Expression data of each patient in the cohort were normalised with TG 1 and with TG 2, and the stratifications and classifications were calculated. The TeG was used to determine the "success" of the translation from DNA-microarray to RNA-seq. "Success" was defined as fulfilling the proportion criteria (1a and 1b) and the first two of the survival criteria (2a and 2b), described above for the VG (see also section 4.2.2).

2.7.3 External testing in early stage and relapsed myeloma patients

Like for VG and TeG, each AMM and each MMR sample was normalised twice, once with TG 1 and once with TG 2. The resulting expression values were used to calculate the stratifications. The same cutoffs were applied as for symptomatic MM. In AMM the high risk group and the medium risk group were merged, due to (expectedly) few samples in the high risk group. In MMR, the low risk group and the medium risk group were merged, due to (expectedly) few samples in the low risk group. In AMM the progression rate to MM and in MMR the OS were used for survival analysis. Regarding RPI stratification, which was trained on the TG 1, the 19 AMM included in the TG 1 were not used for validation.

2.7.4 External testing on CoMMpass cohort

The CoMMpass cohort was used as external testing cohort. Samples before inclusion in treatment were selected. As the downloaded alignment has been performed against the human genome GRCh37, and not GRCh38 (unavailable), normalisation with TG 1 and TG 2 could not be performed. Hence, the CoMMpass cohort was normalised separately with edgeR as described in section 2.3.2.3. Subsequently, the normalised data were standardised with a modified Z-score normalisation, in order to adjust differences in the gene expression pattern of the CoMMpass cohort ($GeneExpr_{CP}$) *versus* the HD cohort, using the means μ_{CP} and μ_{HD} and the standard deviations σ_{CP} and σ_{HD}.

$$StandardisedGeneExpr_{CP} = \frac{GeneExpr_{CP} - \mu_{CP}}{\sigma_{CP}} * \sigma_{HD} + \mu_{HD} \quad (1)$$

The age of the patients of the HD (median 59 years) and CoMMpass cohort (median 64 years) is significantly different (see section 3.4.2 and 4.3). To exclude the influence of this difference, a subgroup of the HD cohort with comparable age distribution was selected. For this, of every ten years the same proportion of patients as in CoMMpass cohort was selected in HD cohort, by chance. This subgroup was used for standardisation. The flowchart of stratification calculation on CoMMpass cohort is depicted in figure 2.7.

One gene (ENSG00000276234) of the HDHRS stratification was missing in the CoMMpass cohort. Hence, the expression of this gene was replaced by the median expression of that gene in the HD subgroup.

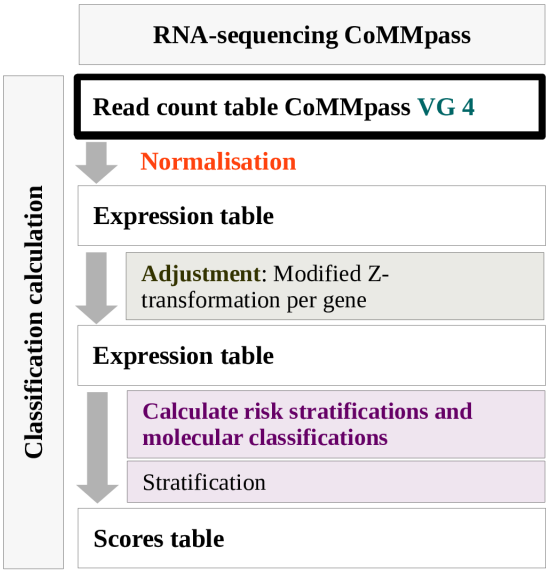

Figure 2.7: Flowchart of stratification calculation for the CoMMpass cohort. The CoMMpass cohort was normalised with edgeR. The expression values per gene were adjusted with Z-score normalisation (see also equation 1) to a subgroup of the Heidelberg cohort. The subgroup was selected according to the age distribution of the CoMMpass cohort.

For the translocations in the CoMMpass cohort [169], available RNA-seq and WGS data were applied. The CCND1-call was used for t(11;14), CCND3-call for t(6;14), MMSET-call for t(14;16) and MAF-call for t(4;14). For the aberrations gain 1q21, deletion 13q14, deletion 17p13 and hyperdiploidy, given WGS data were used.

2.7.5 HDHRS validation on microarray

To further validate the novel HDHRS, the score was transferred to DNA-microarrays, i.e. the inverse strategy compared to the implementation of microarray-based scores. The translation of the ENSGs to gene symbols was performed in R, using the `select` function in the package hgu133plus2.db [29]. The package jetset [142] was used to find the best matching probeset for a gene symbol. Jetset uses a scoring method, which quantifies specificity, coverage and robustness. The "specificity" is defined as the fraction of probes of a probeset, which specifically match to the target gene [142]. The "coverage" is the fraction of transcripts of a target gene, which specifically matches to the probeset [142]. The "robustness" is defined as the probability, that the target sequence is synthesised, considering transcript degradation and enzyme processivity [142]. Jetset returns exactly one match between one probeset and the sought gene symbol (see section 4.1.4 for the discussion of the gene translation).
As for RS-seq, the expression of each gene is multiplied with the prognosis factor and the values were summed up for HDHRS-GEP. New cutoffs were calculated with the running log-rank S3 algorithm published by Rème *et al.* [197], varying the parameters w, $x2$, and fdr: w from 18 to 37 in steps of 1 (representing 9 to 19% of patients, discussed in section 4.2.2), $x2$ from 0.01 to 0.05 in steps of 0.01, fdr from 0.01 to 0.05 in steps of 0.01.

2.8 Evaluation of potential targets

Twenty-five exemplary targets affectable by different treatment strategies (see section 1.5 and 3.5) were assessed in this thesis. Targets were evaluated with three methods, assessing the expression, determining splice variants, and detecting mutated targets. As the target evaluation in this thesis did not prerequisite training and fitting, all patients were used. The normalisations with TG 1 was applied.

2.8.1 Target expression

The R package panp [246] was used to classify an expression as "absent" or "present". On RNA-seq the PA-seq call, described in section 2.3.2.4 was used.
Overexpression in RNA-seq is defined as a present expression value higher than the

median expression in the BMPC samples plus 3 times the standard deviation. This is the same definition as previously used in the GEP-R [156] (see supplementary code C.18). A distinction was made between overexpression and aberrant expression. The latter was defined as both overexpression in MM (see above) and absent expression in BMPC samples. In the GEP-R, absent expression in BMPC is defined as absent (A) or marginal (M) expression of all BMPCs in the PA call. Due to the different method used for assessment of gene expression, absence and presence of expression needed to be defined differently for RNA-seq data. Absence of BMPC expression was defined as 90% BMPC samples with absent expression (n=9), in the PA-seq call (due to higher sensitivity of the RNA-seq-based threshold, see section 4.1.3).
For survival analysis using the expression of the targets, the patients were stratified in two groups, either by applying the PA/PA-seq call, or by applying a maximally selected log-rank statistics. For the latter, the function `maxstat.test` of the package maxstat [104, 105, 136, 137] was used, applying `smethod="LogRank"`. The function determines an optimal cutoff for distinguishing a low and a high risk group. The cutoffs for EFS and OS were averaged, to generate a unique threshold per target.

2.8.2 Splice variants

Alternative splicing events can potentially eliminate the target sequence of especially cell surface proteins targeting immunotherapeutic approaches, e.g. CD38 or BCMA. The splice variant analysis has been performed for a subset of the currently available samples (see table 2.1) within the framework of two co-authored articles of Seckinger *et al.* [212, 213]. The counted splice junctions were filtered for annotated junctions of the genes *CD38* and *BCMA*, using the human genome GTF file (release 82). The splice junctions of each sample were discarded if less than 10 reads were detected (see supplementary code C.16, line 12). For comparison: The raw number of reads in MM samples over all splice junctions in *CD38* ranges from 179 to 23996, while the gene length is 1560. The percentage of each splice junction in comparison to all other splice junction was determined per patient sample. A specific splice junction is defined as belonging to only one transcript.
A transcript was counted as present if each splice junction of the full-length transcript was detected (see supplementary code C.16, line 20 to 22). If no alternating splice junction is spanned by at least 10 reads, only this transcript is called present. The number of reads of alternating, annotated splice junctions was calculated per patient for 1%, 5% and 10%.
No normalisation was performed.

2.8.3 Mutation detection

As clinically relevant example for mutation detection, the BRAF mutations (V600E and V600K, dbSNP identifier: rs113488022) were used. Therefore, reads at the corresponding position 140753336 on chromosome 7 in the human genome GRCh38 were counted per base using `bam-readcount` [135] (see section 2.3.2.1). The reference base at this position is "A". Mutations to be counted as present required: i) at least two reads covering the mutation, ii) the highest mean mapping quality of 255, iii) a base quality of at least 30, iv) at least one read in each strand direction, v) an average base position in the intermediate 85% of the nucleotides and vi) a variant allele frequency of at least 10%. The latter is calculated by dividing the number of reads spanning the mutation by the number of all reads spanning the position. The code for this section is depicted in supplementary code C.19. Within the same argumentation as for iFISH, a threshold of 60% was assessed to distinguish clonal and subclonal mutations (see iFISH in section 2.1.1).

3 Results

This chapter is divided in five parts. First, the developed RNA-seq reporting pipeline, including quality control and determination of presence of expression, is depicted. Second, the performance of the transferred risk stratifications and classifications is described. Third, the performance of the novel risk stratification is shown. Fourth, the outcome of the validation of the stratifications is depicted. At the end of this chapter target assessment for clinical application is presented, including expression, splice variant and mutation analyses.

3.1 RNA-sequencing analysis pipeline

The pipeline is delineated in figure 3.1. It contains three main parts: First, the alignment and read count (depicted in yellow colour), second the normalisation, assessment of presence or absence of gene expression, calculation of risk stratifications and molecular classifications (depicted in violet colour), and third, the target analyses (depicted in green colour). The code for an exemplary sample is depicted in supplementary code C.20. Comparability of stratifications, classifications and target expressions to former analysed samples was ensured in the second step, by normalising each sample with the TG (see also section 4.1.1). Likewise, quality control was established in this thesis and included in the pipeline (depicted in blue colour). RNA-seq can be performed in 90% of all patients [99], using a low input amount of RNA of 0.01 to 1 ng. The quality of the RNA-seq files was considered as sufficient in 97% of all RNA-seq files.

3.1.1 RNA-sequencing data quality

Quality control was performed for 853 samples based on 983 RNA-seq files (in case of FASTQ files 983 RNA-seq file-pairs), including technical replicates, e.g. repetitions due to quality issues.

3.1.1.1 Quality control

Quality control was performed before alignment, using FastQC and after alignment using the number of mapping reads and the library size.

FastQC. Prior to alignment raw RNA FASTQ files were controlled with FastQC. This tool basically assumes a random and diverse library[3] containing only sparsely du-

[3]FastQC: Evaluating Results; Online resource: `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/2%20Basic%20Operations/2.2%20Evaluating%20Results.html`; Status: 30.04.2020, 12:05

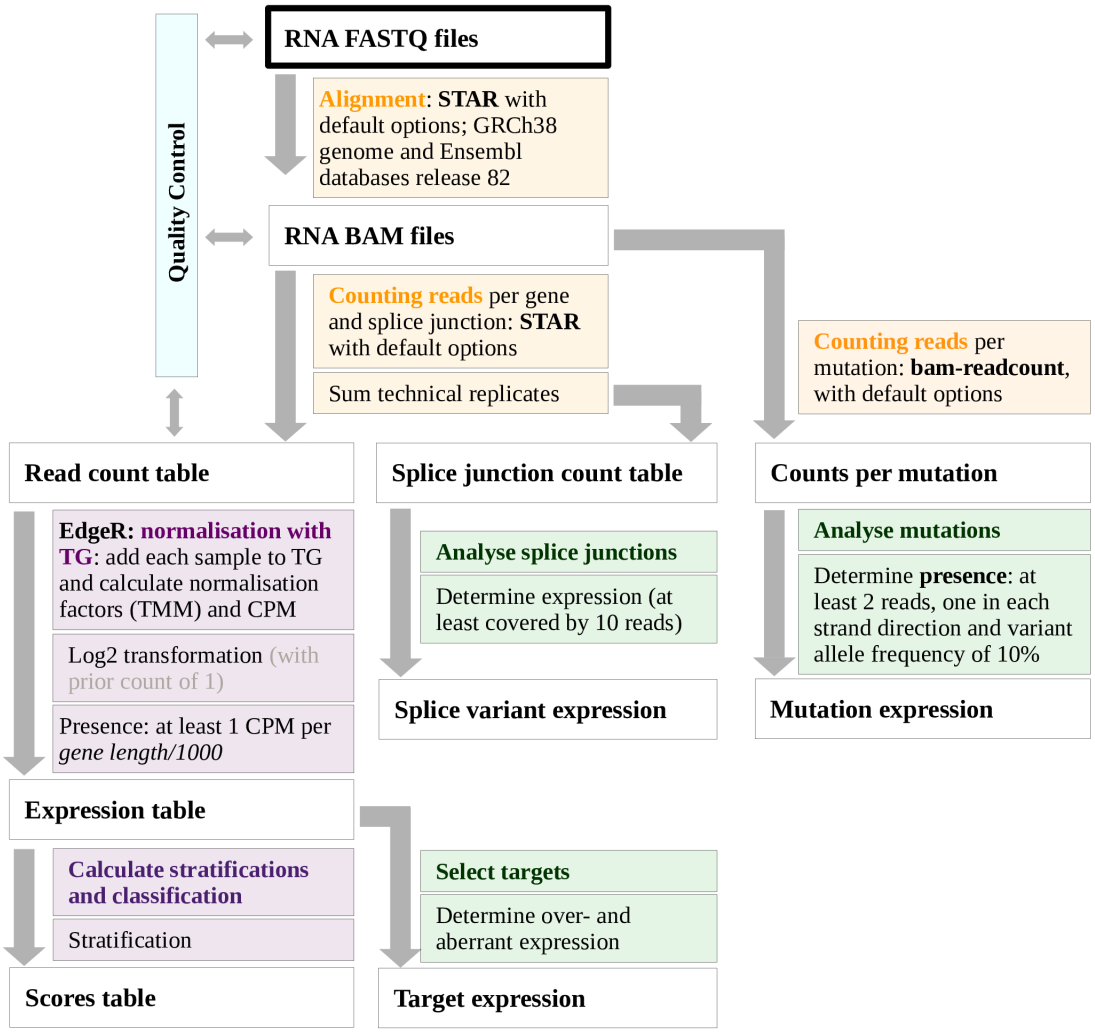

Figure 3.1: Flowchart of RNA-sequencing stratifications, classifications and target assessment analysis pipeline. Raw starting data are raw RNA-sequencing (RNA-seq) FASTQ files, depicted in a thick, black frame. The pipeline contains three main parts: First, preprocessing of the FASTQ files to a read count table (yellow coloured background), second, normalisation with the training group (TG) and risk stratification and classification calculation of a new sample (violet coloured background), third, target analysis regarding expression, splice variants and mutations (green coloured background). Quality control is performed before and after alignment (blue colour background). In this thesis, the quality of 97% of all RNA-seq files was considered as sufficient. CPM: counts per million.

plicated sequences[4] [7]. The tool further suggests to interpret the results only as indications if the library composition differs[3]. Libraries in RNA-seq have a wide dynamic range and contain highly duplicated sequences, representing reads in highly expressed genes (e.g. Ig genes, see section 3.1.1.3 and 4.1.1). Therefore, the "sequence duplication level" is interpreted as just giving information about individual sample expression patterns, and is not used as exclusion criterion (see section 3.1.1.3 and 4.1.1). FastQC

[4]FastQC: Duplicate Sequences; Online resource: `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/8%20Duplicate%20Sequences.html`; Status: 30.04.2020, 12:05

detected 47 different overrepresented sequences, representing more than 1% of the total number of reads in each assessed patient file. They are 50, 51 or 75 nucleotides long. For validation, the sequences were mapped to the human genome GRCh38, using the ENSEMBL search tool BLAT. Twenty of the sequences map several times to chromosome 2 or chromosome 22 in IGKV or IGLV regions and one poly-G sequence maps to the long intergenic non-protein coding RNA 486 (*LINC00486*). This high level of Ig expression is expected, as the former Ig-regions are typical for plasma or MM cells (see below). The latter poly-G sequence marginally exceeds the threshold (percentage of 1.003%). Two of the 26 non-mapping sequences are poly-A sequences of differing length. The first nucleotides of the 24 remaining sequences map to the PCR primer sequences "AAGCAGTGGTATC" and "AACGCAGAGT" or to the Illumina multiplexing index read sequences "GATCGGAAGAGCAC" and "ACGTCTGAACTCCAGTCAC". By default, STAR performs soft-clipping of the ends of the reads, which discards the poly-A, the primer and the index sequences, and maps the remaining sequence of the read. Hence, no file was excluded due to the category "overrepresented sequences". Furthermore, overrepresented sequences contain 7-mers, and increase the "k-mer content", thus this category is no exclusion criterion also. Likewise, overrepresented sequences are known to potentially affect the overall composition and the "per base sequence content"[5] [7], hence the 19 samples assessed with a "fail" in this category were not excluded. Samples failing the category "per tile sequence quality" were only excluded if a large region was affected. A large region is defined in this thesis as more than 10% of all tiles or 10% of tiles at one read position and failed in three samples, which thus were excluded. The "per sequence GC content" of all files is expectedly different from theoretical distribution, and homogeneous within all files, hence it is no exclusion criterion. All of the files passed the remaining five categories ("per base sequence quality", "per sequence quality scores", "per base n content", "sequence length distribution" and "adapter content"). Each file had a uniform sequence length of 100, 102, 150, 152, 154 or 160 bp.

Number of mapped reads. After alignment, 16 files with less than 60 % mapped reads were excluded. Main reason for not mapping in the remaining samples are too short reads (88.9% to 100% of all not mapping reads per file). These "too short reads", more precisely too short aligned reads, include reads, which are either *a priori* too short or map only partial and are trimmed by soft-clipping. For instance, this is defined in the default options of STAR by at least 66% matched bases per read (option `-outFilterMatchnminOverLread`) and by an alignment score of at least 0.66 (op-

[5] FastQC: Per Base Sequence Content; Online resource: `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html`; Status: 08.10.2019, 19:05

tion `outFilterScoreMinOverLread`). The minimum number of mapped reads for an individual sample in the analysed cohort, which was not excluded, was 2579784.

Library size. Technical replicates were summed up and subsequently the library size was controlled. Eight files with a library size less than 10 million reads were excluded. In total, 27 files were discarded, whereof 11 were not repeated due to material constraints, hence, 842 samples were analysed in this thesis.

For comparison, in the CoMMpass cohort the library size was controlled. Only files with a number of reads considered as sufficient were published, which corresponds to a minimum library size of 19840000 reads. As the percentage of unmapped reads is not accessible, it could not be compared. With the published requirement of at least 60 million reads (i.e. read-pairs) for each library [51, 169], the minimum number of mapping reads can be calculated as at least 33.1%.

3.1.1.2 Exclusion of potential batch affects

Read length and the sequencing run convey possible batch effects. Hence, their impact was analysed, performing a PCA. The PCA shows no clustering (data not shown). Additionally, most of the sequencing runs mainly contained one entity for clinical application reasons, which does not allow batch correction as the necessary assumption of equal constitution of the batches (e.g. BMPC, HMCL and MM patient samples in the same run) is not fulfilled. Hence a batch correction was not implemented.

3.1.1.3 Analysis of most highly expressed genes

Genes with the highest expression were determined: 44 different genes were detected, related to more than 10% of all mapping raw reads per sample. In 534 of the 535 MM samples, at least one of these genes was highly expressed. All of these genes are Ig genes, except *B2M* (β_2-microglobulin, in 3 patient samples $\sim$10.1%, respectively). β_2-microglobulin is known to be highly expressed in MM patients with risk stratifications in part being based on this gene, see section 1.4.1. The maximum percentage of raw reads mapping to one gene (*IGHG1*, ENSG00000211896) is 56%. This is explainable considering the investigation of plasma cells (see section 3.1.1.3 and 4.1.1) and is intrinsically determined by their function and thus expected. For comparison: Regarding normalised counts, the expression of *IGHG1* in symptomatic MM ranges from 3.5 to 19.7.

3.1.2 Presence of expression

For RNA-seq, one CPM normalised count adjusted by gene length was used as threshold for presence of expression, as utilised for RPI and for target prediction. This

Table 3.1: Comparison of gene length estimations and assessed presence of target expression (see section 3.5). **a** Gene length was estimated three times for the targets: First, the stop position of each gene was subtracted from its start position (maximum gene length). Second, the length of all exons per gene was calculated, subtracting the overlapping regions (maximum exon length). Third, the length of all exons per transcript was calculated, subtracting the overlapping regions, and the length of all transcripts belonging to one gene was averaged (median transcript length). **b** Percentage of patient samples with present expression of at least 1 CPM per *gene length*/1000.

a Name	maximum gene length	maximum exon length	median transcript length	number of transcripts	**b** maximum gene length	maximum exon length	median transcript length	1 raw read
BCMA	2962	1118	668	3	100%	100%	100%	100%
CD38	74956	6855	1560	5	98.69%	100%	100%	100%
HM1.24	2710	1119	459	3	100%	100%	100%	100%
CD74	11293	3184	791.5	12	100%	100%	100%	100%
NYESO1/2	1679	993	870.5	2/3	0%	0%	0%	46.92%
HGF	71433	8897	1147.5	10	40.93%	82.43%	94.21%	99.44%
FGFR3	15566	4834	4041.5	10	7.66%	7.85%	7.85%	85.42%
MAGEA1	4596	1710	1710	1	18.5%	26.17%	26.17%	64.3%
MAGEA3	3596	1788	1724	3	26.54%	31.96%	32.52%	68.04%
MMSET	110784	19776	1212	27	10.28%	11.78%	96.26%	100%
IGF1R	315560	13509	572	17	0%	18.88%	77.01%	98.69%
TP53	25772	3936	2331	27	51.4%	97.01%	98.13%	100%
AURKA	22949	2928	2112	13	2.8%	60%	70.84%	99.07%
CCND1	13388	4830	560.5	6	64.67%	74.02%	84.11%	99.25%
CCND2	31579	7157	550.5	4	39.63%	50.09%	73.46%	98.5%
CCND3	115425	5802	592	23	21.68%	97.01%	99.81%	100%
RHAMM	31743	3936	1402	8	1.68%	66.17%	88.41%	99.25%
CD20	15009	4872	767.5	12	39.63%	57.76%	82.8%	98.88%
GPRC5D	11373	1124	903	3	96.45%	99.81%	99.81%	100%
MUC1	7093	4717	927	29	0.37%	0.56%	12.71%	73.83%
CSF1	20751	5418	1009	9	0%	1.5%	24.67%	86.36%
WT1	47856	4113	2421	9	0.19%	0.56%	0.56%	61.31%
SSX2	62623	1950	1028	4	0%	0%	0%	3.36%
NKG2D	19522	3030	1553	5	0%	0%	0%	30.09%

is based on the assumption that one normalised count corresponds to at least 5 raw counts, which can be defined as (arbitrary) minimum threshold of expression, see e.g. the edgeR user's guide [39]. In the HD cohort, of all genes one normalised count corresponds to at least 7 raw counts in TG, 9 in VG and at least 7 raw counts in TeG. Thus, the probability of false positive "present" determination is very low (see section 4.1.3).

The length of the genes or the genetic sequences, assigned to an ENSG was approximated with the median length of all transcripts belonging to a gene. It ranged from 8

to 205000 bp. Genetic sequences shorter than 20 bp are IG heavy diversity sequences or T cell receptor diversity sequences. In table 3.1 the gene length of the potential targets estimated with three methods is depicted: first, the whole gene length, subtracting the start position from the stop position of the gene, second, the maximum length of all exons per gene and third the median length of all transcripts. The PA call on DNA-microarrays was compared to the PA-seq determination on RNA-seq. For this, all ENSGs were translated to probesets. Of 57566 ENSGs in RNA-seq data 20162 ENSGs can be translated in probesets, using the R package hgu133plus2.db. Due to the multiple matching probesets, the jetset package in R was used to determine the best fitting probeset for every ENSG. After this step, 18804 "translatable" genes remained. The ENSGs, matching to more than one gene symbol, were removed, resulting in 18771 remaining genes. The mean calculated consistency between present and absent expression assessment per sample is 84%, with a range from 71% to 87% for all MM samples. In table 3.2 the patients with the minimum and maximum consistency are depicted (TG 1 normalisation). The implemented PA-seq call function for RNA-seq is depicted in supplementary code C.17.

Table 3.2: Exemplary confusion matrices of present and absent expression determination per sample on RNA-sequencing (RNA-seq) *versus* microarrays. Depicted is the number (and percentage) of the 18771 translatable genes with present (P) and absent (A) expression prediction on RNA-seq in rows and on microarray in columns. **a** Exemplary patient with maximum consistency (*CO*). **b** Exemplary patient with minimum consistency.

a

RNA-seq	**Microarray**		
CO = 87%	**A**	**P**
A	*8485 (45%)*	850 (5%)
P	1473 (8%)	*7963 (42%)*

b

RNA-seq	**Microarray**	
CO = 71%	**A**	**P**
A	*7184 (38%)*	1182 (6%)
P	4292 (23%)	*6113 (33%)*

3.2 Transferred risk stratifications and classifications

One main aim of this thesis is the translation of stratifications and classifications from DNA-microarrays to RNA-seq, either by directly translating (GPI, RS, UAMS, EMC92, IFM15, TC), or newly setting-up using the pamr-predictor (MC and t(4;14)). Cutoffs were adjusted in three ways (see section 2.5). First, the GPI cutoffs were transferred by correlating the scores. Second, multiple comparisons of survival plots were performed to obtain the best cutoff for RS-seq (500 executions), for HDHRS (2500 executions) and HDHRS-GEP (500 executions). Third, the cutoffs for the UAMS70-, EMC92-, IFM15- and TC-seq were calculated according to the original methods.
The resulting RNA-seq stratifications and classifications are applied to MM patient samples. The proportions of the independent TeG (see supplementary figure A.1 and table B.8) are for consistency compared to those obtained on microarray TeG cohort.

Likewise, the survival analyses for RNA-seq stratifications are compared to the microarray analyses. For this, calculated Brier scores and the univariate Cox regression analysis are used (including concordance and hazard ratios). Brier scores, R^2 and concordance are listed in table 3.3.

As quality criteria determining "success" of the translation from DNA-microarray to RNA-seq, both proportion criteria and two of the four survival criteria demanded to be fulfilled (see section 2.7.4). This means, in each class should be at least 9% of the samples and they should differ in less than 20 percentage points in comparison to the DNA-microarray scores (see also 4.2.2). Further, the log-rank test of the survival analysis should be significant ($p < 0.05$), the survival curves should be in logical order and not intersecting in a time interval from 12 months to the last but one event per curve.

Performances and comparisons are described for each risk stratification and classification in the following section depicting the evaluation of the implementation, the continuous scores, the categorical stratifications and classifications, and the comparison to DNA-microarray.

Table 3.3: Evaluation of the performance of risk prediction models in the test group (TeG). The table shows the values of three different survival comparison methods for all risk stratifications: Brier scores (Brier), R^2 and concordance (C) with standard error (SE). A "-" indicates a p-value $p > 0.1$, * indicates $p < 0.05$ and ** indicates $p < 0.01$. Depicted are event free survival (EFS) and overall survival (OS) for **a** RNA-sequencing-based stratifications **b** microarray-based stratifications **c** international staging system (ISS) and revised ISS (R-ISS)

		EFS				OS			
		Brier	R^2	**C**	**C SE**	**Brier**	R^2	**C**	**C SE**
a	**HDHRS**	0.1534*	0.13	0.62	0.02	0.1601	0.17	0.66	0.03
	RPI	-	0.04	0.57	0.02	-	0.01	0.60	0.03
	UAMS70-seq	-	0.02	0.57	0.02	-	0.07	0.59	0.03
	RS-seq	-	0.04	0.56	0.01	-	0.03	0.58	0.02
	IFM15-seq	-	0.02	0.57	0.02	-	0.07	0.58	0.02
	EMC92-seq	0.1554*	0.03	0.56	0.01	-	0.00	0.58	0.01
b	**HDHRS-GEP**	0.1559	0.08	0.61	0.02	0.1576*	0.25	0.66	0.03
	GPI	-	0.00	0.55	0.02	-	0.00	0.61	0.03
	UAMS70	-	0.03	0.57	0.02	0.1704	0.14	0.60	0.02
	RS	-	0.02	0.58	0.02	-	-0.01	0.62	0.03
	IFM15	-	0.01	0.54	0.01	-	0.11	0.57	0.02
	EMC92	-	0.02	0.55	0.01	-	0.07	0.57	0.01
c	**ISS**	0.1577	0.08	0.58	0.02	0.1625**	0.15	0.63	0.03
	R-ISS	-	0.08	0.60	0.02	0.155**	0.28	0.67	0.03

3.2.1 RNA-seq-based proliferation index

Implementation. The microarray GPI developed by Hose *et al.* [101] includes 50 probesets, associated with cell proliferation. They were translated into 50 ENSGs and their present expression values were summed. The gene list is in supplementary table

B.9 (see also section 2.5.1). Risk stratification cutoffs were transferred from microarray to RNA-seq by correlation of the GPI and RPI for the TG and linear regression, resulting in a low cutoff of $lcut = 121.9601$ and a high cutoff of $hcut = 202.7359$ (see figure 3.2). The implemented function for calculating the RPI on RNA-seq is depicted in supplementary code C.7.

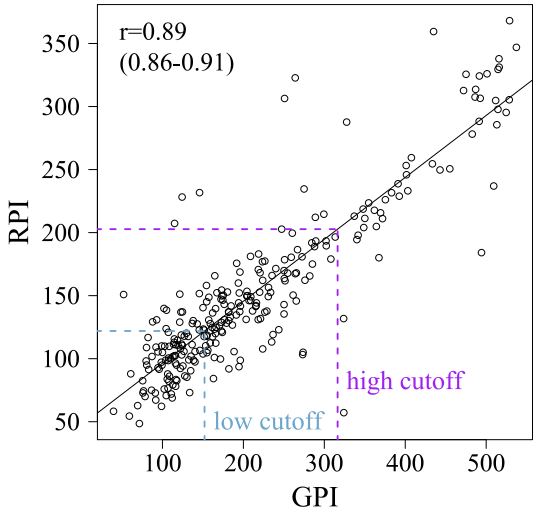

Figure 3.2: Cutoff estimation of RPI on the training group (TG). Shown is the correlation of RPI and GPI on TG 1 with the regression line. The two GPI cutoffs *GPIcut* were transferred to the new cutoffs *RPIcut*, using the slope m and the y-interception value b of the regression line by $RPIcut = m * GPIcut + b$. The correlation coefficient (r) is displayed with its confidence interval.

Evaluation of the continuous score. The continuous RPI score in MM is highly prognostic in association with survival, its hazard significantly increases over time. This holds true for both EFS and OS (log-rank test $p < 0.001$, both). As depicted in figure 3.3a, a significant, stage dependent increase from MGUS over AMM and symptomatic MM to MMR is observed (JHT test: $p = 0.001$). The RPI of MBC and BMPC is significantly lower than the RPI of MM, and the RPI of PPC and HMCL is significantly higher than the RPI of MM. RPI score and GPI score show a correlation coefficient of $r = 0.85$, depicted in figure 3.3b.

Evaluation of the categorical stratification. RPI significantly delineates three groups for OS, in TG and VG, shown in supplementary figure A.2 and in TeG shown in figure 3.4a and 3.4b. The median survival times for low, medium and high risk in TeG are 38, 33 and 25 months for EFS and 103, 92 and 55 months for OS. The concordance of the RPI is 0.57 for EFS and 0.6 for OS (see table 3.3). The univariate Cox regressions show significant hazard ratios for low *versus* high risk group for OS (2.8) and for EFS (2.25) (see figure 3.4).

Comparison of both platforms. Pairwise comparison of RPI and GPI per patient in TeG shows no changes (0%) between low and high risk group, while the consistency

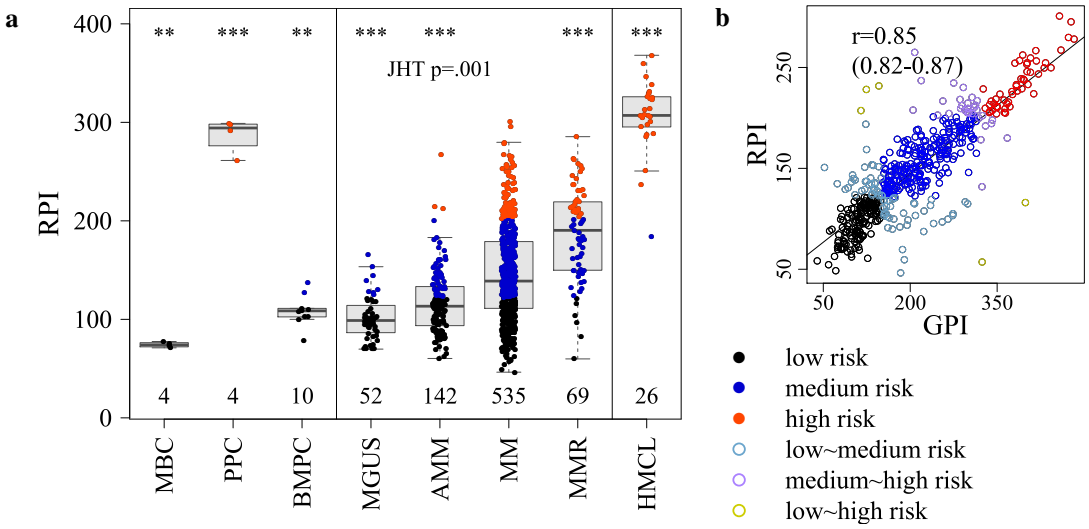

Figure 3.3: Continuous RNA-sequencing-based proliferation index (RPI) investigated on the whole cohort. **a** RPI grouped by disease entity, compared to non-malignant plasma cells and precursors (BMPC, MBC, PPC) and human myeloma cell lines (HMCL). Significant differences are depicted by 1, 2 or 3 asterisks, indicating significant p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. A Jonckheere-Terpstra test (JHT) was performed to test the significance of the score increase from MGUS over AMM and symptomatic MM to MMR. **b** Correlation of RPI and GPI. Samples stratified in the low risk group of the RPI and in the medium risk group of the GPI or *vice versa* (low∼medium risk) are depicted in light blue, medium∼high risk samples are depicted in violet and low∼high risk samples are depicted in yellow. The correlation coefficient (r) is displayed with its confidence interval. MBC: memory B cell; PPC: polyclonal plasmablasts; BMPC: bone marrow plasma cell; MGUS: monoclonal gammopathy of undetermined significance; AMM: asymptomatic multiple myeloma; MM: multiple myeloma; MMR: relapsed MM.

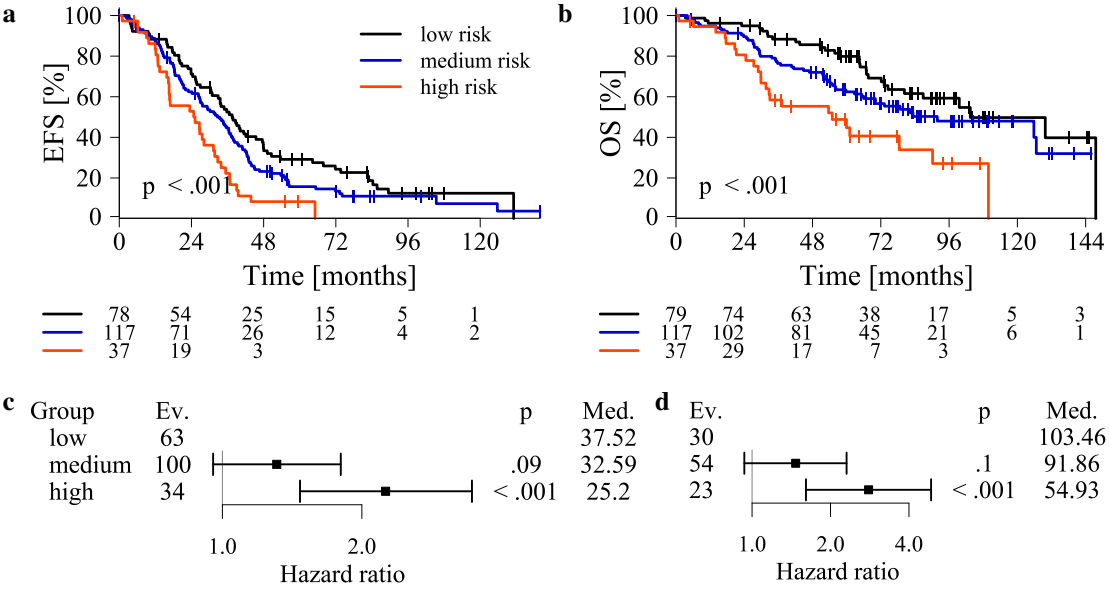

Figure 3.4: Survival analysis regarding RNA-sequencing-based proliferation index (RPI) for the test group (TeG). Performance of the RPI of symptomatic multiple myeloma patient samples for the TeG in **a** event free survival (EFS) and **b** overall survival (OS). Univariate Cox regression was performed for **c** EFS and **d** OS. Shown is the hazard ratio on a logarithmic scale with a 95% confidence interval. Ev.: number of events; Med: median survival time in months; p: p-value.

is 85% (see table 3.4). The proportions of the groups in RPI and GPI are comparable for the low risk group (33.91% and 32.19%), while the high risk group is larger in RPI (15.88%) than in GPI (9.87%).

Table 3.4: Confusion matrix of RPI and GPI stratification on the TeG. Depicted is the number (and percentage) of patients per RPI group in rows and per GPI group in columns. In the top left of the table the consistency (*CO*) is depicted.

		GPI		
RPI	$CO = 84.6\%$	**low risk**	**medium risk**	**high risk**
	low risk	*69 (30%)*	10 (4%)	0
	medium risk	6 (3%)	*108 (46%)*	3 (1%)
	high risk	0	17 (7%)	*20 (9%)*

3.2.2 Risk stratifications

3.2.2.1 UAMS70-seq

Implementation. The UAMS70 score [219] is based on 70 probesets, which were translated to ENSGs. Two probesets (227547_at and 237964_at) have no translation to ENSGs, whereas 225834_at (*MGC57827*) can be translated into four ENSGs (ENSG00000263513, ENSG00000196550, ENSG00000188610, ENSG00000215784). Of the 71 resulting genes, three were excluded due to very low correlation with microarray expression ($r \leq 0.15$) and one due to low correlation $r < 0.4$ and high PA *versus* PA-seq call difference ($nCO1 > 30\%$, see section 2.2.2 and 2.5). Four additional genes were excluded due to ambiguous translation. The translated gene list is shown in supplementary table B.10. The UAMS70-seq score was calculated following the original publication [219] and in analogy to the score implementation in the GEP-R [156] (see also section 2.5.2.1). New cutoffs were determined with k-means clustering for three groups, resulting in the three centres of -6.083014, -5.190644 and -4.096769. As in the original publication, low and medium risk group were merged. The implemented function for estimating the UAMS70-seq on RNA-seq is depicted in supplementary code C.8.

Evaluation of the continuous score. In MM, the continuous UAMS70-seq score is highly prognostic, i.e. associated with survival. Its hazard significantly increases over time for both EFS and OS (log-rank test $p < 0.001$ and $p = 0.02$). A significant (JHT test: $p = 0.001$) stage dependent increase from MGUS over AMM and symptomatic MM to MMR is found (see figure 3.5a). The UAMS70-seq score of MBC and BMPC is significantly lower than the UAMS70-seq of MM, and the UAMS70-seq of PPC and HMCL is significantly higher. Both results are expected and identical for DNA-microarrays (data not shown). The UAMS70-seq and the UAMS70 show a correlation coefficient of $r = 0.77$ (see figure 3.5b).

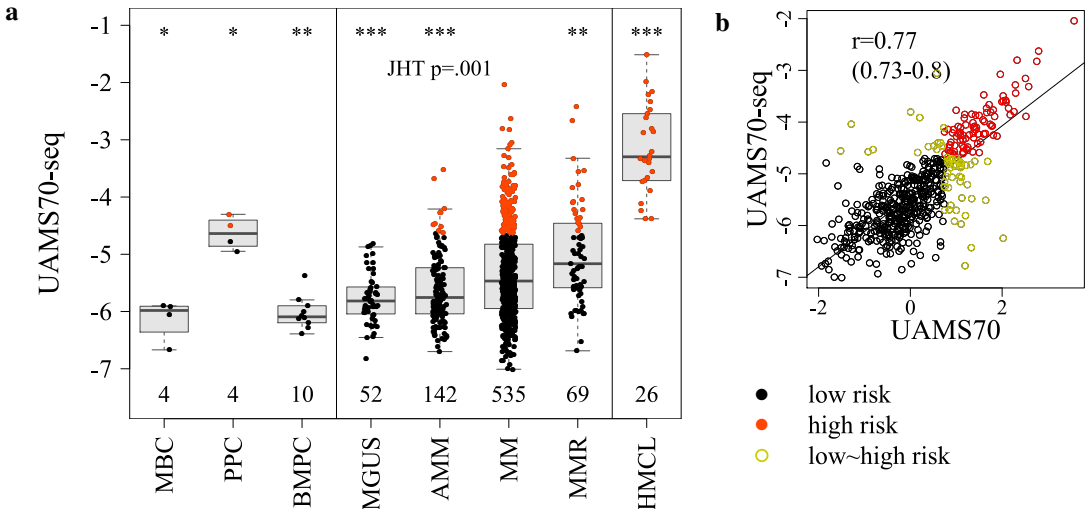

Figure 3.5: Continuous UAMS70-seq score investigated on the whole cohort. **a** UAMS70-seq grouped by disease entity, compared to non-malignant plasma cells and precursors (BMPC, MBC, PPC) and human myeloma cell lines (HMCL). Significant differences are depicted by 1, 2 or 3 asterisks, indicating significant p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. A Jonckheere-Terpstra test (JHT) was performed to test the significance of the score increase from MGUS over AMM and symptomatic MM to MMR. **b** Correlation of UAMS70-seq and UAMS70. Samples stratified in the low risk group of the UAMS70-seq and high risk group of the UAMS70 or *vice versa* (low~high risk) are depicted in yellow. The correlation coefficient (r) is displayed with its confidence interval. MBC: memory B cell; PPC: polyclonal plasmablasts; BMPC: bone marrow plasma cell; MGUS: monoclonal gammopathy of undetermined significance; AMM: asymptomatic multiple myeloma; MM: multiple myeloma; MMR: relapsed MM.

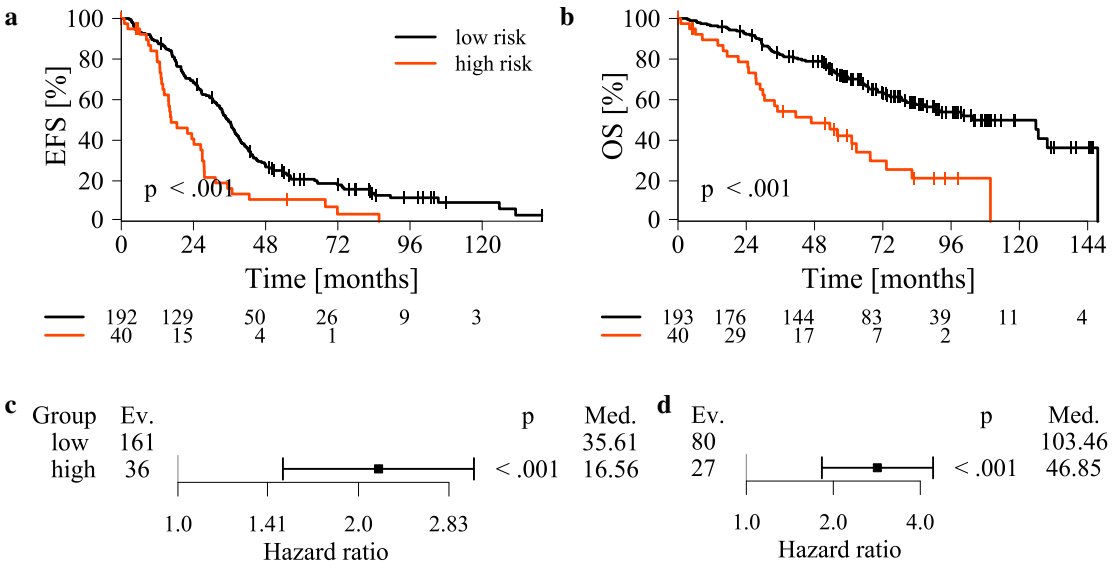

Figure 3.6: Survival analysis regarding UAMS70-seq on the test group (TeG). Performance of the UAMS70-seq of symptomatic multiple myeloma patient samples for the TeG in **a** event free survival (EFS) and **b** overall survival (OS). Univariate Cox regression was performed for **c** EFS and **d** OS. Shown is the hazard ratio on a logarithmic scale with a 95% confidence interval. Ev.: number of events; Med: median survival time in months; p: p-value.

Evaluation of the categorical stratification. The UAMS70-seq significantly delineates two groups of patients for OS and EFS, for both, TG and VG (supplementary figure A.3) as well as TeG (figure 3.6a and 3.6b). The two UAMS70-seq groups show a median survival time for EFS of 36 and 17 and for OS of 103 and 47 months, respectively. The concordance is 0.57 for EFS and 0.59 for OS (see figure 3.3). The Brier score is not significant for EFS and OS. Univariate Cox regression shows significant hazard ratios of low *versus* high risk group for EFS (2.16) and for OS (2.84).

Comparison of both platforms. Pairwise comparison of UAMS70-seq and UAMS70 per patient in TeG shows 10% changes between low and high risk group (see table 3.5). The proportions of the groups are similar on RNA-seq and microarray data, with a high risk group in UAMS70-seq of 17.17% and in UAMS70 of 23.61%.

Table 3.5: Confusion matrix of UAMS70-seq and UAMS70 stratification on the TeG. Depicted is the number (and percentage) of patients per UAMS70-seq group in rows and per UAMS70 group in columns. In the top left of the table the consistency (*CO*) is depicted.

		UAMS70	
UAMS70-seq	*CO = 90.2%*	**low risk**	**high risk**
	low risk	*174 (75%)*	19 (8%)
	high risk	4 (2%)	*36 (16%)*

3.2.2.2 RS-seq

Implementation. The RS score presented by Rème *et al.* [197] comprises 19 prognostic probesets, which were translated to ENSGs (see also section 2.5.2.2). Two of the probesets (233660_at and 235353_at) were excluded before the RS-seq score was implemented due to inconsistent translations. The gene list is depicted in supplementary table B.11. To determine the cutoff sets, the running log-rank S3 algorithm was executed 500 times, resulting in 23 different sets. Per definition, in TG ($w \geq 19$), all sets fulfil the proportion criteria of at least 9% patients per group. Fifteen of the 23 sets were excluded due to crossing curves in survival analyses and one due to a non-significant Brier score for EFS and OS. The remaining 7 sets were validated on the VG. Three of these have a low risk group size smaller than 9% of all patients and the other 4 show survival curves crossing between high risk and medium risk. The last cutoff set, which has no significant Brier score in TG, was tested on VG, but had crossing survival curves. Hence, the groups were adjusted to get as near as possible the original RS proportions, by combining the calculated cutoffs by taking the highest high cutoff (*hcut=29.37*) to obtain the smallest high risk group and the highest low risk cutoff (*lcut=9.01*) to get the largest low risk group. This cutoff set passed the visual inspection of the survival plots in TG and VG, although it has neither the best Brier score nor concordance. The cutoff set, with similar proportions to the original RS proportions,

was compared to an exact copy of the RS proportions to RS-seq, but this cutoff set is similar in survival analysis and has a larger Brier score. The implemented function for estimating the RS-seq on RNA-seq is depicted in supplementary code C.10.

Evaluation of the continuous score. The continuous RS-seq score is highly prognostic in association with survival in MM, as its hazard significantly increases over time, for EFS and OS (log-rank test $p < 0.001$, both). A significant (JHT test: $p = 0.001$) stage dependent increase from MGUS over AMM and MM to MMR is observed, shown in figure 3.7a. The RS-seq of BMPCs is significantly lower than the one of MM and the RS-seq of PPC and HMCL is significantly higher than the one of MM, as expected. RS-seq and RS show a correlation coefficient of $r = 0.84$ (see figure 3.7b).

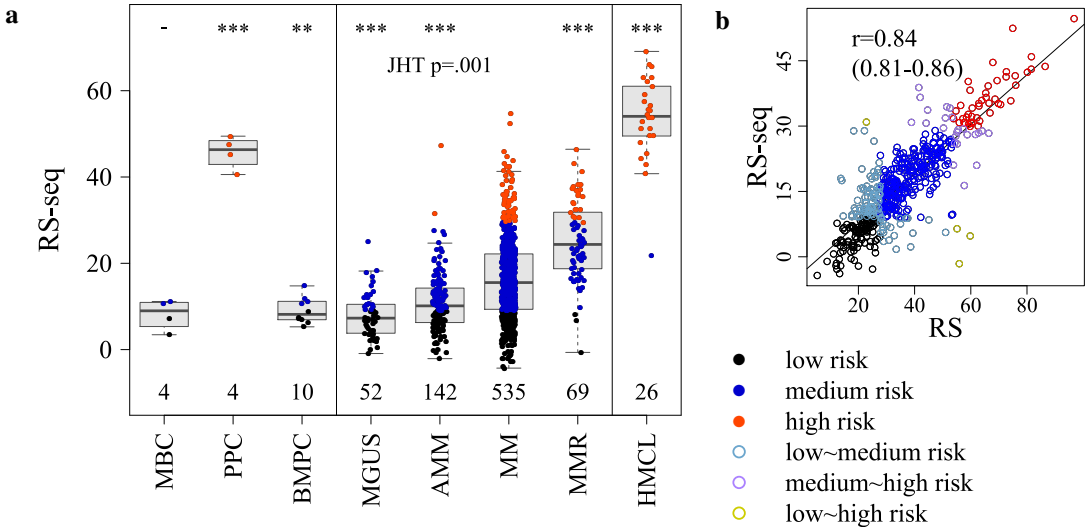

Figure 3.7: Continuous RS-seq score investigated on the whole cohort. **a** RS-seq grouped by disease entity, compared to non-malignant plasma cells and precursors (BMPC, MBC, PPC) and human myeloma cell lines (HMCL). Significant differences are depicted by 1, 2 or 3 asterisks, indicating significant p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. A Jonckheere-Terpstra test (JHT) was performed to test the significance of the score increase from MGUS over AMM and symptomatic MM to MMR. **b** Correlation of RS-seq and RS. Samples stratified in the low risk group of the RS-seq and medium risk group of the RS or *vice versa* (low∼medium risk) are depicted in light blue, medium∼high risk samples are depicted in violet and low∼high risk samples are depicted in yellow. The correlation coefficient (r) is displayed with its confidence interval. MBC: memory B cell; PPC: polyclonal plasmablasts; BMPC: bone marrow plasma cell; MGUS: monoclonal gammopathy of undetermined significance; AMM: asymptomatic multiple myeloma; MM: multiple myeloma; MMR: relapsed MM; HMCL.

Evaluation of the categorical stratification. RS-seq delineates three groups with significantly different survival for EFS and OS, in TG and VG, shown in supplementary figure A.4 and TeG shown in figure 3.8a and 3.8b. The Brier score is not significant for EFS and OS in TeG. The median survival time for EFS is 40 months for the low risk group, 32 months for the median risk group, and 17 months for the high risk group and 130, 83 and 37 months for OS, respectively. The concordance is 0.56 for EFS and 0.58

for OS (see figure 3.3). Univariate Cox regression (see figure 3.8c and 3.8d) shows significant hazard ratios for the comparison between low *versus* medium risk and low *versus* high risk group in both OS (1.81 and 3.3) and EFS (1.62 and 3.21).

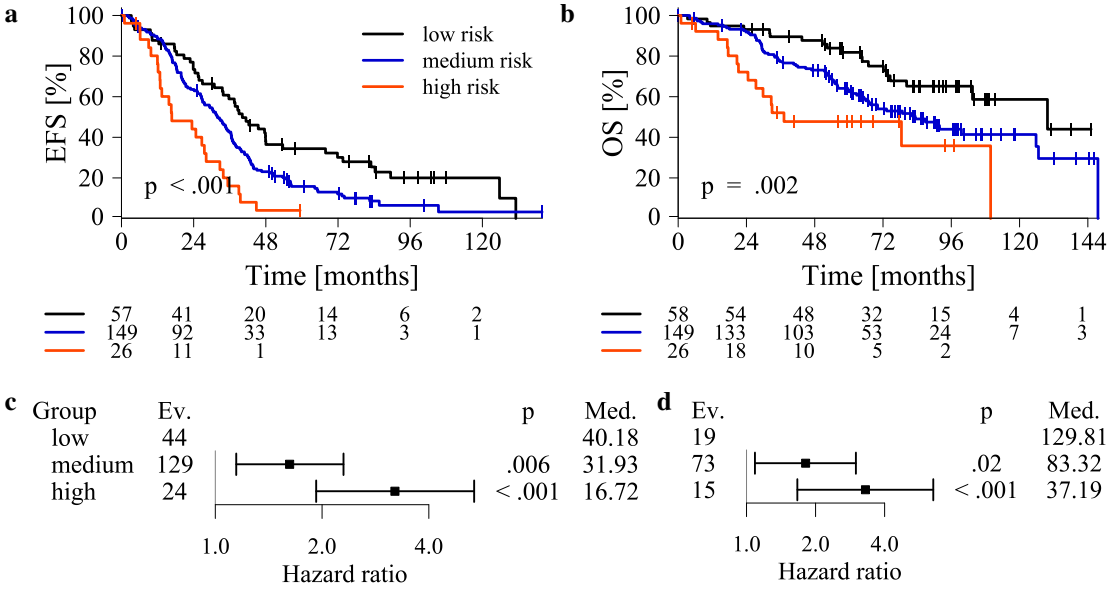

Figure 3.8: Survival analysis regarding risk score based on RNA-seq (RS-seq) for the test group (TeG). Performance of the RS-seq of symptomatic multiple myeloma patient samples for the TeG in **a** event free survival (EFS) and **b** overall survival (OS). Univariate Cox regression was performed for **c** EFS and **d** OS. Shown is the hazard ratio on a logarithmic scale with a 95% confidence interval. Ev.: number of events; Med: median survival time in months; p: p-value.

Comparison of both platforms. Pairwise comparison of the RS-seq and the RS per patient in TeG shows no switch (0%) between the low and the high risk group. The consistency is 82% (see table 3.6). The low risk group in RS-seq is half the size of the low risk group in RS (24.89% and 43.78%), but in both cases, the medium risk group is the largest group. High risk groups have comparable proportions.

Table 3.6: Confusion matrix of RS-seq and RS stratification on the TeG. Depicted is the number (and percentage) of patients per RS-seq group in rows and per RS group in columns. In the top left of the table the consistency (*CO*) is depicted.

		RS		
RS-seq	*CO = 81.6%*	**low risk**	**medium risk**	**high risk**
	low risk	*55 (24%)*	3 (1%)	0
	medium risk	31 (13%)	*116 (50%)*	2 (<1%)
	high risk	0	7 (3%)	*19 (8%)*

3.2.2.3 EMC92-seq

Implementation. One hundred ENSGs matched to the 92 probesets of the EMC92-stratification [124]. Three probesets matched several ENSGs and probeset 243018_at

(*BBOX1-AS1*) was missing. Fourteen genes were excluded due to very low correlation ($r < 0.15$) with microarray expression or due to low correlation ($r < 0.4$) and high number of samples with present expression prediction on microarray and absent expression prediction on RNA-seq ($nCO_1 > 30\%$, see section 2.2.2 and 2.5). Six additional genes were excluded due to ambiguous translation. Two probesets (211714_x_at and 209026_x_at) had the same ENSG (ENSG00000196230), which was only used once and the original weighting scores were averaged. In total 79 genes were used. The gene list is shown in supplementary table B.12. The EMC92-seq score was calculated using the same processing as in the underlying publication [124] (see also section 2.5.2.3). To calculate an RNA-seq-based cutoff, the proportion of patients with OS of less than two years *versus* equal or more than two years was used. In the HD cohort 21 of 194 patients (10.8%) progressed within the first two years, resulting in a cutoff of 1.01. The plots for TG and VG for this cutoff are shown in supplementary figure A.5. The performance of the TeG cohort is shown in figure 3.10. The implemented function for estimating the EMC92-seq on RNA-seq is depicted in supplementary code C.9.

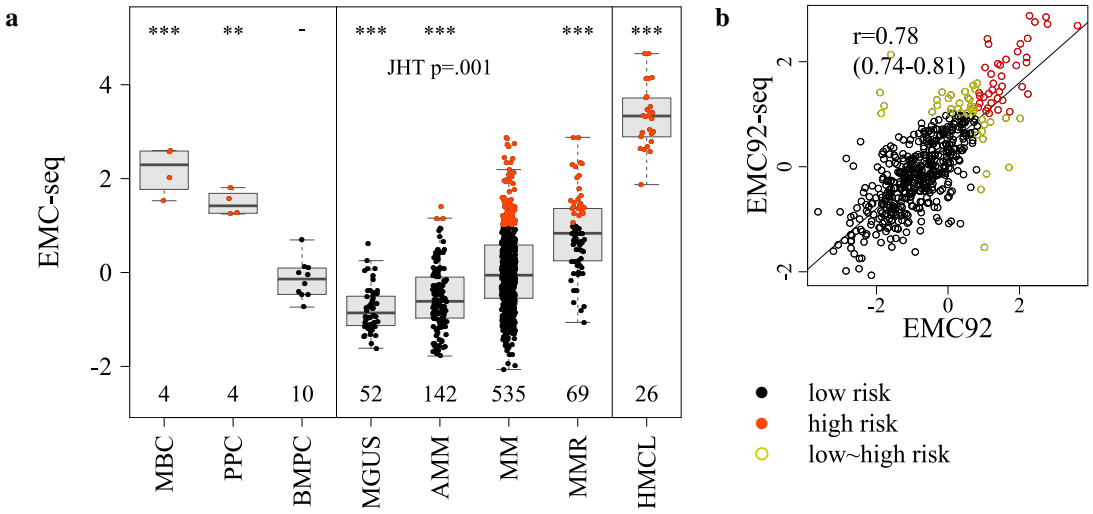

Figure 3.9: Continuous EMC92-seq score investigated on the whole cohort. **a** EMC92-seq grouped by disease entity, compared to non-malignant plasma cells and precursors (BMPC, MBC, PPC) and human myeloma cell lines (HMCL). Significant differences are depicted by 1, 2 or 3 asterisks, indicating significant p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. A Jonckheere-Terpstra test (JHT) was performed to test the significance of the score increase from MGUS over AMM and symptomatic MM to MMR. **b** Correlation of EMC92-seq and EMC92. Samples stratified in the low risk group of the EMC92-seq and high risk group of the EMC92 or *vice versa* (low∼high risk) are depicted in yellow. The correlation coefficient (r) is displayed with its confidence interval. MBC: memory B cell; PPC: polyclonal plasmablasts; BMPC: bone marrow plasma cell; MGUS: monoclonal gammopathy of undetermined significance; AMM: asymptomatic multiple myeloma; MM: multiple myeloma; MMR: relapsed MM.

Evaluation of the continuous score. The continuous EMC92-seq score in MM is highly prognostic, i.e. associated with survival. The hazard significantly increases over time for both EFS and OS (log-rank test $p < 0.001$, both). There is a significant

(JHT test: $p = 0.001$) stage dependent increase from MGUS over AMM and symptomatic MM to MMR, shown in figure 3.9a. EMC92-seq of MBC, PPC and HMCL is significantly higher than EMC92-seq of MM. EMC92-seq and EMC92 show a correlation coefficient of $r = 0.78$.

Evaluation of the categorical stratification. The EMC92-seq significantly delineates two groups in TeG for OS and EFS (see figure 3.10a and 3.10b). The median survival time for the low and the high risk group for EFS is 35 and 14 months. For OS, it is 103 and 30 months. The concordance for EMC92-seq is 0.56 months for EFS and 0.58 months for OS (see figure 3.3). The Brier score is significant for EFS. The hazard ratios of univariate Cox regression of low *versus* high risk is significant for EFS (3.86) and OS (3.93).

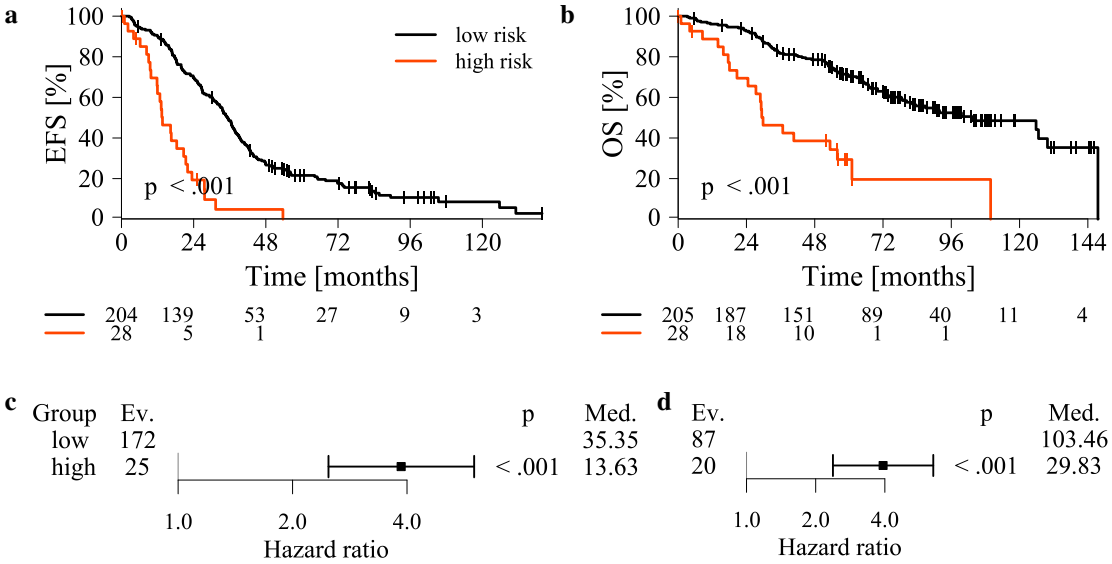

Figure 3.10: Survival analysis regarding EMC92-seq for the test group (TeG). Performance of the EMC92-seq of symptomatic multiple myeloma patient samples for the TeG in **a** event free survival (EFS) and **b** overall survival (OS). Univariate Cox regression was performed for **c** EFS and **d** OS. Shown is the hazard ratio on a logarithmic scale with a 95% confidence interval. Ev.: number of events; Med: median survival time in months; p: p-value.

Comparison of both platforms. Pairwise comparison of EMC92-seq and EMC92 per patient in TeG shows consistency between low and high risk group in 92% of the patients (see table 3.7).

Table 3.7: Confusion matrix of EMC92-seq and EMC92 stratification on the TeG. Depicted is the number (and percentage) of patients per EMC92-seq group in rows and per EMC92 group in columns. In the top left of the table the consistency (*CO*) is depicted.

		EMC92	
EMC92-seq	*CO = 92.2%*	**low risk**	**high risk**
	low risk	*200 (86%)*	5 (2%)
	high risk	13 (6%)	*15 (6%)*

3.2.2.4 IFM15-seq

Implementation. The IFM15 is based on 15 genes associated with poor prognosis in MM. It was calculated on RNA-seq as described in the underlying publication [56] and the score implementation in the GEP-R [156]. All 15 translated genes are listed in supplementary table B.13. The cutoff was determined as 75% quartile (absolute value *7.672949*). Samples were classified in two groups. Results for TG and VG for this cutoff are shown in supplementary figure A.6. In figure 3.12 the performance of the TeG cohort is shown. The implemented function for estimating the IFM15-seq on RNA-seq is depicted in supplementary code C.13.

Evaluation of the continuous score. The continuous IFM15-seq score in MM is highly prognostic in association with survival. Its hazard significantly increases over time for both, EFS and OS (log-rank test $p < 0.001$ and $p = 0.001$). A stage dependent increase from MGUS over AMM and symptomatic MM to MMR is observed with significant JHT test ($p = 0.001$, see figure 3.11a). The IFM15-seq of MBC, PPC and HMCL is significantly higher than IFM15-seq of MM. The IFM15-seq and the IFM15 show a correlation coefficient of $r = 0.81$, depicted in figure 3.11b.

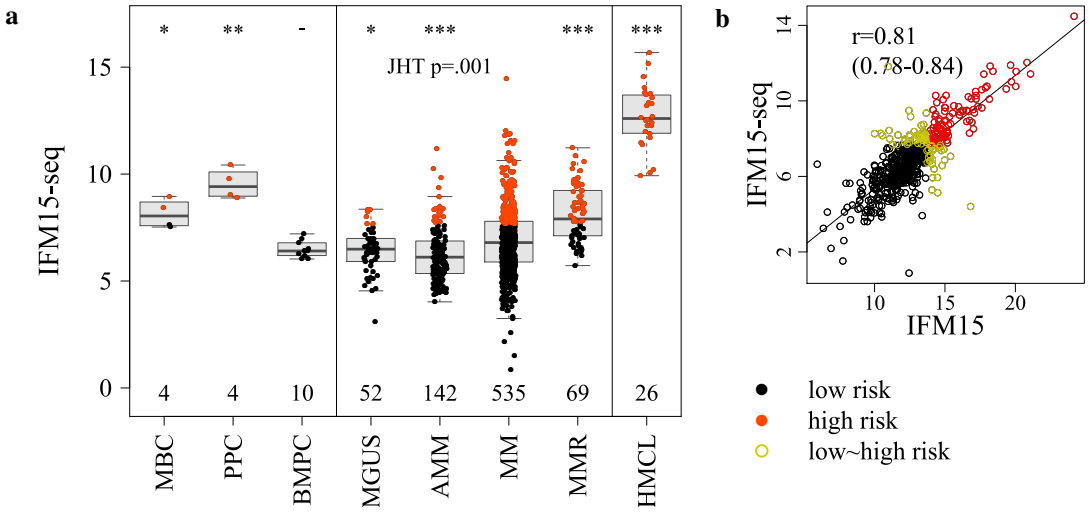

Figure 3.11: Continuous IFM15-seq score investigated on the whole cohort. **a** IFM15-seq grouped by disease entity, compared to non-malignant plasma cells and precursors (BMPC, MBC, PPC) and human myeloma cell lines (HMCL). Significant differences are depicted by 1, 2 or 3 asterisks, indicating significant p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. A Jonckheere-Terpstra test (JHT) was performed to test the significance of the score increase from MGUS over AMM and symptomatic MM to MMR. **b** Correlation of IFM15-seq and IFM15. Samples stratified in the low risk group of the IFM15-seq and high risk group of the IFM15 or *vice versa* (low∼high risk) are depicted in yellow. The correlation coefficient (r) is displayed with its confidence interval. MBC: memory B cell; PPC: polyclonal plasmablasts; BMPC: bone marrow plasma cell; MGUS: monoclonal gammopathy of undetermined significance; AMM: asymptomatic multiple myeloma; MM: multiple myeloma; MMR: relapsed MM.

Evaluation of the categorical stratification. IFM15-seq significantly delineates two groups for OS and EFS (see figure 3.12a and 3.12b), with a median EFS of 36 and 21

and OS of 103 and 66 months. The concordance is 0.56 for EFS and 0.58 for OS (see figure 3.3). The Brier score is neither significant for OS nor EFS. Univariate Cox regression showed significantly different hazard ratios of low *versus* high risk for EFS (1.86) and for OS (1.98).

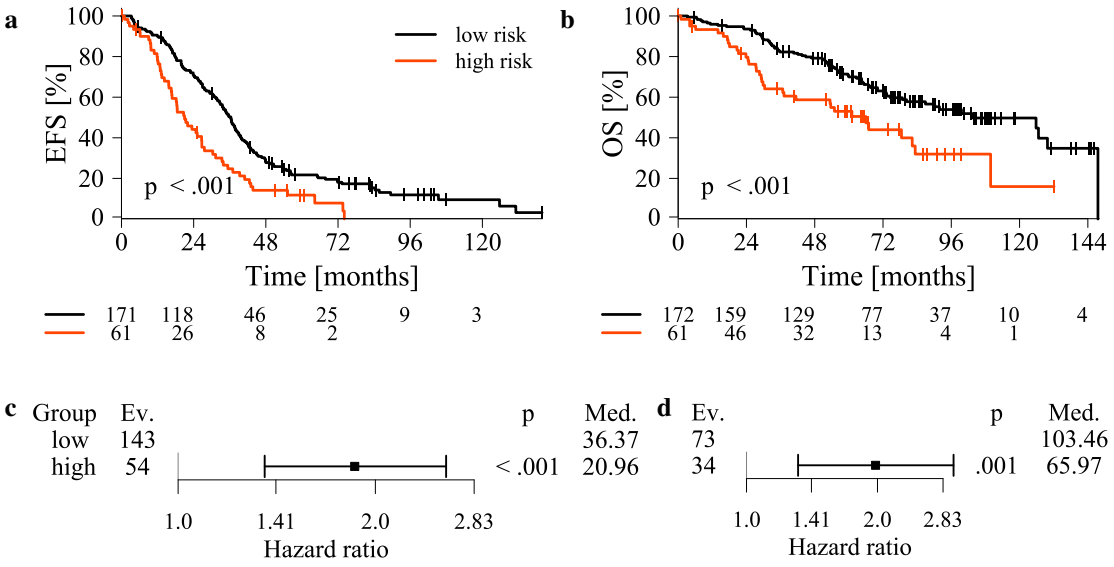

Figure 3.12: Survival analysis regarding IFM15-seq for the test group (TeG). Performance of the IFM15-seq of symptomatic multiple myeloma patient samples for the TeG in **a** event free survival (EFS) and **b** overall survival (OS). Univariate Cox regression was performed for **c** EFS and **d** OS. Shown is the hazard ratio on a logarithmic scale with a 95% confidence interval. Ev.: number of events; Med: median survival time in months; p: p-value.

Comparison of both platforms. Pairwise comparison of IFM15-seq and IFM15 per patient in TeG shows 15% changes between the low and high risk group (see table 3.8). The proportions of the groups are comparable in RNA-seq and DNA-microarray data (26.18% and 20.6% of high risk, respectively).

Table 3.8: Confusion matrix of IFM15-seq and IFM15 stratification on the TeG. Depicted is the number (and percentage) of patients per IFM15-seq group in rows and per IFM15 group in columns. In the top left of the table the consistency (*CO*) is depicted.

		IFM15	
IFM15-seq	*CO* = 85%	**low risk**	**high risk**
	low risk	*161 (69%)*	11 (5%)
	high risk	24 (10%)	*37 (16%)*

3.2.3 Molecular classifications

3.2.3.1 TC-seq

The TC classification according to Bergsagel *et al.* [22] uses two gene sets (set *a* and set *b*) to classify patients in 8 groups. The 10 probesets (representing 9 genes) of set *a* were

translated into 9 ENSGs. *CCND1* is represented by two probesets (208711_s_at and 208712_at), hence the new norm cutoffs for *CCND1* were estimated by dividing the two raw cutoffs by the sum of both given medians. The 11 probesets (representing 10 genes) of set *b* were translated into 14 unique ENSGs, present in normalised RNA-seq data. Four ENSGs were duplicates. Ten unique ENSGs were used and the expression values for 204006_s_at, matching *FCGR3B* and *FCGR3A*, and 213550_s_at, matching *IK* and *TMCO6*, were added, respectively. The TC classification on RNA-seq was calculated with the instructions and equation in the original publication [22] and in the GEP-R [156] resulting in 8 groups (see section 2.5.3.1). The implemented function for estimating the TC-seq on RNA-seq is depicted in supplementary code C.12.
The "11q13" and "D1" group are the largest groups with 24% and 34% of all MM patients. The "6p21" group is the smallest group with 1% of patients. The comparison of TC-seq and TC for TG and VG is depicted in in supplementary table B.15 and the TeG comparison in table 3.9. 81% of all MM patients are are classified as belonging to the same groups on RNA-seq and on DNA-microarrays. There is no single diverging group, differences in classification are found in all groups.

Table 3.9: Confusion matrix of TC-seq and TC classification on the TeG. Depicted is the number (and percentage) of patients per TC-seq group in rows and per TC group in columns. In the top left of the table the consistency (*CO*) is depicted.

		TC 2007							
TC2007-seq	*CO = 81%*	**11q13**	**6p21**	**D1**	**D1+D2**	**D2**	**FGFR3**	**MAF**	**none**
	11q13	*49 (21%)*	0	8 (3%)	0	0	0	1 (<1%)	0
	6p21	0	*3 (1%)*	1 (<1%)	0	0	0	0	0
	D1	6 (3%)	1 (<1%)	*68 (30%)*	0	1 (<1%)	1 (<1%)	2 (1%)	1 (<1%)
	D1+D2	0	0	2 (1%)	*14 (6%)*	4 (2%)	0	2 (1%)	0
	D2	0	0	0	0	*15 (7%)*	0	2 (1%)	1 (<1%)
	FGFR3	1 (<1%)	0	0	2 (1%)	1 (<1%)	*17 (7%)*	1 (<1%)	0
	MAF	0	0	5 (2%)	0	0	0	*15 (7%)*	0
	none	0	1 (<1%)	0	0	1 (<1%)	0	0	*4 (2%)*

3.2.3.2 MC-seq

The MC classification by Zhan *et al.* [254] used 688 probesets to classify MM in seven transcriptional signatures (MS, MF, CD1, CD2, HY, PR, LB). The 688 probesets were translated to 691 ENSGs, whereas 55 probesets were missing and 2 are not present in the normalised count table. The MC-seq on RNA-seq was calculated with the created

pamr-predictor (see section 2.5.3.2). The implemented function for estimating the MC-seq on RNA-seq is depicted in supplementary code C.14.
Of all MM patients, 87% are classified as belonging to the same groups in RNA-seq and DNA-microarrays on the TeG. The highest discrepancies can be found between the four groups CD2, HY, LB and PR (see table 3.10). The TG and VG is depicted in supplementary table B.16.

Table 3.10: Confusion matrix of MC-seq and MC classification on the TeG. Depicted is the number (and percentage) of patients per MC-seq group in rows and per MC group in columns. In the top left of the table the consistency (*CO*) is depicted.

		MC						
MC-seq	*CO = 86.6%*	**CD1**	**CD2**	**HY**	**LB**	**MF**	**MS**	**PR**
	CD1	*13 (6%)*	0	1 (<1%)	0	0	0	3 (1%)
	CD2	0	*46 (20%)*	2 (<1%)	2 (<1%)	1 (<1%)	0	3 (1%)
	HY	0	3 (1%)	*59 (25%)*	3 (1%)	0	1 (<1%)	3 (1%)
	LB	0	0	0	*27 (12%)*	0	0	1 (<1%)
	MF	0	0	1 (<1%)	0	*6 (3%)*	0	0
	MS	0	1 (<1%)	0	0	0	*22 (9%)*	0
	PR	0	0	0	6 (3%)	0	0	*29 (12%)*

3.2.3.3 Translocation t(4;14) prediction

The translocation t(4;14) can be predicted using gene expression profiling on DNA-microarrays [51, 156, 169]. For the implementation on RNA-seq, the pamr package (see sections 2.5.3.2 and 2.5.3.3) was used to create a predictor. The selected predictor uses seven genes (*FGFR3*, *MMSET*, *CLEC11A*, *MOB3A*, *JAM3*, *DSG2*, *SEPT9*). In contrast, the MMRF, based on biological assumptions, choose *MMSET* to predict presence and absence of the translocation t(4;14) for the CoMMpass cohort. As on microarrays, t(4;14) prediction is very precise. In the TeG, the consistency is 99% both in comparison to iFISH and in comparison to microarray prediction (see table 3.11a and 3.11b). TG and VG are depicted in supplementary table B.17.
The implemented function for t(4;14) prediction on RNA-seq is depicted in supplementary code C.15.

Table 3.11: Confusion matrices of t(4;14) prediction on the TeG on RNA-seq, microarray and iFISH. Depicted is the number (and percentage) of patients with and without predicted t(4;14) on RNA-seq in rows and **a** predicted on microarray or **b** determined with iFISH in columns. In the top left the percentage of consistency (*CO*) is depicted.

a

t(4;14)-seq	**t(4;14)-microarray**		
	CO = 99.2%	**no t(4;14)**	**t(4;14)**
	no t(4;14)	*208 (89%)*	1 (<1%)
	t(4;14)	1 (<1%)	*23 (10%)*

b

t(4;14)-seq	**t(4;14)-iFISH**		
	CO = 99.6%	**no t(4;14)**	**t(4;14)**
	no t(4;14)	*208 (89%)*	1 (<1%)
	t(4;14)	0	*24 (10%)*

3.3 Novel risk stratification

One main aim of this thesis was the *de novo* generation of a risk stratification based on the algorithm of Rème *et al.* [197]. The algorithm includes four steps: normalisation, gene selection, score calculation and cutoff determination.

Implementation. After normalisation, the number of input genes was reduced from 57449 to 17502 genes by filtering as described in section 2.6. Five different thresholds of the BH corrected p-values of 0.1, 0.075, 0.05, 0.025 and 0.01 for gene selection were used, which resulted in five different gene sets of 19, 30, 53, 74 and 90 genes. By varying the starting conditions, the algorithm for risk group optimisation ran 500 times per gene set, i.e. 2500 iterations in sum. Multiple approaches led to the same cutoff sets and subsequently to the same group proportions. In these cases, additional cutoffs leading to identical group composition were dropped. This leads to 48 cutoff sets, which were subsequently compared.

Twelve sets passed the group size criteria of at least 9% of patients (per definition this is always true in the TG), had no crossing survival curves and significant (p<0.05) Brier scores for OS and EFS. The remaining cutoff sets were validated on the VG. Four sets were excluded due to crossing survival curves. The Brier score was only significant for two cutoff sets for OS and no set for EFS. Of these two sets only one had a concordance > 0.6 for EFS and OS. This final set, with the lowest Brier score and the highest concordance, is based on the 53 gene list derived with a BH adjusted p-value threshold below 0.05. The best cutoff pair was selected with a minimal size for a risk group (w) of 20, a BH adjusted p-value (fdr) of 0.05, a chi-squared statistic threshold ($x2$) between two survival curves of `qchisq((1-0.05),1)` and a minimal number of events per group (cx) of 2. The low cutoff is $lcut = 3.87$ and the high cutoff is ($hcut = 24.10$).

Evaluation of prognostic genes. The 53 genes comprising the predictor are listed in supplementary table B.14. Thirteen of the genes are associated with good prognosis and 40 with poor prognosis. Two genes are also part of other risk stratifications: ENSG00000117650 (*NEK2*) and ENSG00000138180 (*CEP55*) are part of the proliferation assessing GPI and RPI stratification.

To obtain an overview regarding the biological and pathophysiological role of the 53 genes selected for the HDHRS, a reactome pathway analysis was performed (see figure 3.13). Of these, 14 genes have at least one database entry (26%). In comparison, 80% of the GPI genes, 47% of the RS genes, 56% of the UAMS70 genes, 66% of the EMC92 genes and 47% of the IFM15 genes have such an entry. The most frequently found pathway of the HDHRS is "metabolism", with 5 assigned genes, followed by "Cell Cycle", with 3 assigned genes, and "Signal Transduction", with 2 assigned genes.

Four of the five genes assigned to "metabolism" are associated with poor prognosis and one is associated with good prognosis. For comparison, GPI genes were selected according to the gene ontology terms "cell proliferation" and "cell cycle" and (expectedly) almost all GPI genes (35 of 40 genes) are associated with the reactome pathways "Cell Cycle", validating the principle approach.
The survival analyses for TG and VG for the chosen cutoff set are shown in supplementary figure A.7. The performance of the TeG is shown in figures 3.14 and 3.21. The implemented function for estimating the HDHRS is depicted in supplementary code C.11.

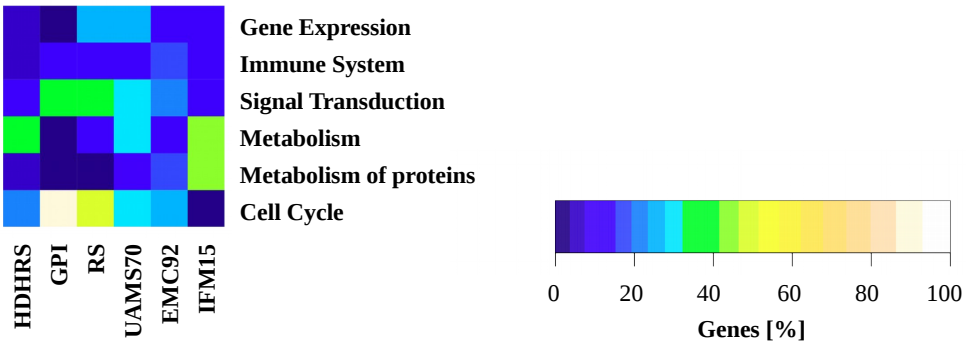

Figure 3.13: Reactome pathway analysis of risk stratification genes. Listed are the main reactome pathways occurring in at least one stratification in $\geq 1\%$ of the genes of the stratification. Most frequently found pathway of the HDHRS genes is "metabolism", with 5 assigned genes, followed by "Cell Cycle", with 3 assigned genes.

Evaluation of the continuous score. The continuous HDHRS score is highly prognostic for the TeG in MM. Its hazard significantly increases over time for both EFS and OS (log-rank test $p < 0.001$ and $p < 0.001$). A significant (JHT test: $p = 0.001$) stage dependent increase from MGUS over AMM and symptomatic MM to MMR is seen, depicted in figure 3.14a. The HDHRS of PPC and MBC is significantly higher compared with MM, whereas the HDHRS of BMPC is significantly lower.
Evaluation of the categorical stratification. The HDHRS delineates three groups of patients with significantly different EFS and OS (figure 3.14b and 3.14c), with a median EFS time for low, medium and high risk of 39, 32 and 18 months and median OS time of 148, 82, and 37 months, respectively. The proportions of low, medium and high risk group are 38.2%, 38.63% and 23.18%. The Brier score is significant for EFS (0.1534) in the TeG and has a value of (0.1601) for OS ($p \leq 0.1$). It is the smallest (best) of all RNA-seq expression-based stratifications and smaller than the one for ISS for EFS (0.1577) and OS (0.1625). The concordance is the highest concordance of all gene expression-based risk stratification regarding EFS (0.62) and OS (0.66). It is larger than the concordance for ISS (0.58 and 0.63, see table 3.3) and larger than

the concordance for R-ISS for EFS (0.6). Univariate Cox regression shows significant hazard ratios in low *versus* medium risk and low *versus* high risk group for OS (2.09 and 4.3) and for EFS (1.71 and 2.97).

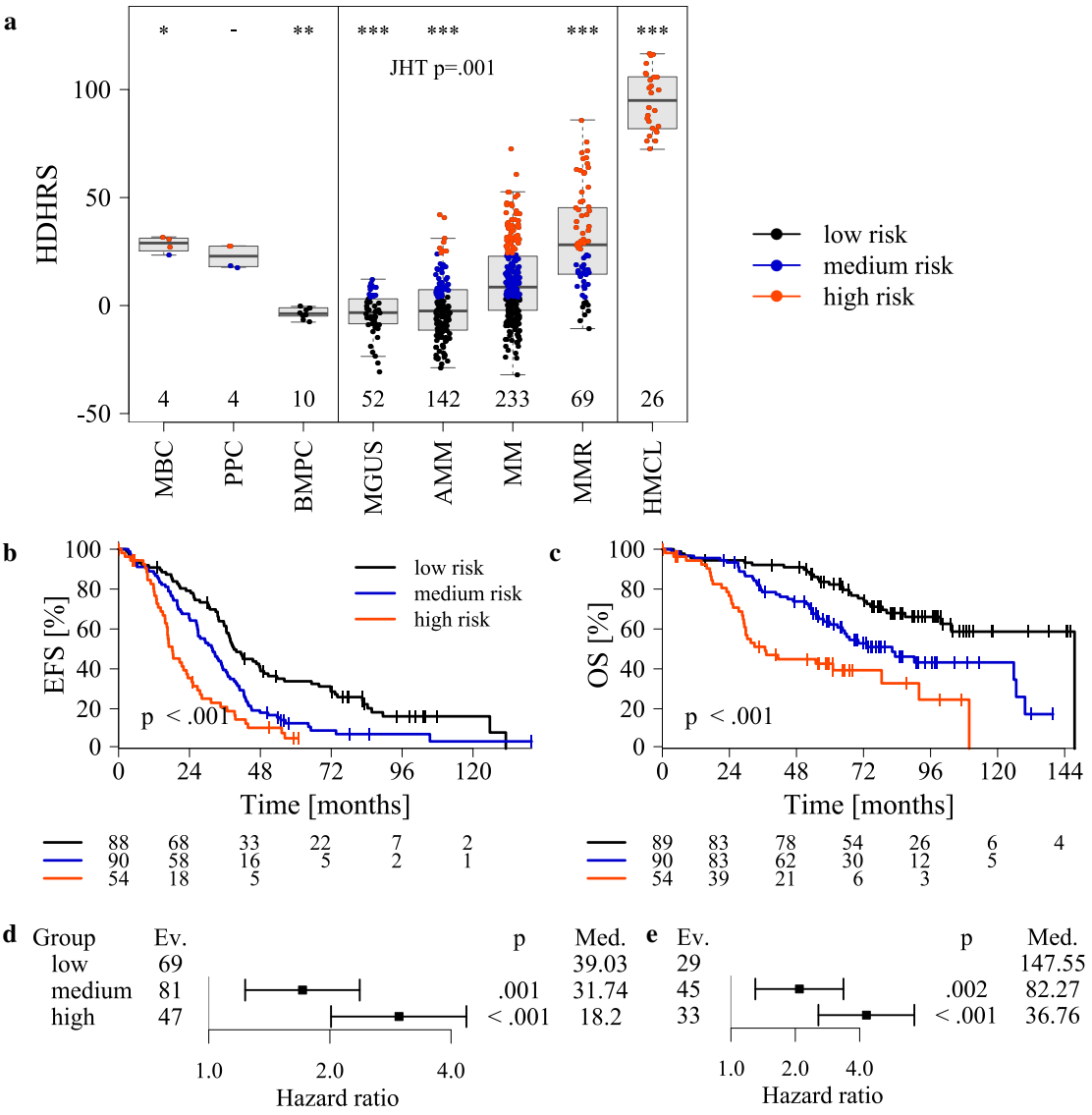

Figure 3.14: Continuous Heidelberg high risk score (HDHRS) on the TeG. **a** HDHRS grouped by disease entity, compared to non-malignant plasma cells and precursors (BMPC, MBC, PPC) and human myeloma cell lines (HMCL). Significant differences are depicted by 1, 2 or 3 asterisks, indicating significant p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. A Jonckheere-Terpstra test (JHT) was performed to test the significance of the score increase from MGUS over AMM and symptomatic MM to MMR. Performance of the HDHRS of symptomatic multiple myeloma patient samples on the TeG in **b** event free survival (EFS) and **c** overall survival (OS). Univariate Cox regression was performed for **d** EFS and **e** OS. Shown is the hazard ratio on a logarithmic scale with a 95% confidence interval. Ev.: number of events; Med: median survival time in months; p: p-value; MBC: memory B cell; PPC: polyclonal plasmablasts; BMPC: bone marrow plasma cell; MGUS: monoclonal gammopathy of undetermined significance; AMM: asymptomatic multiple myeloma; MM: multiple myeloma; MMR: relapsed MM.

Comparison to R-ISS. HDHRS and the R-ISS were compared in a multivariate Cox-regression, see figure 3.15. HDHRS and R-ISS remain highly significant for EFS and OS in medium and high risk group. The HDHRS has higher hazard ratios for EFS analysis (1.93 and 2.64) compared to the R-ISS (1.86 and 2.15), but lower hazard ratios for OS analysis (2.17 and 2.96 compared to 3.56 and 6.15).

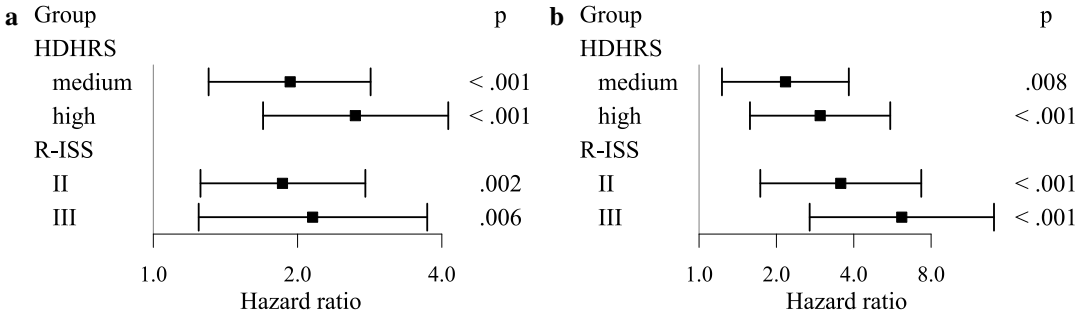

Figure 3.15: Multivariate Cox regression of HDHRS and revised ISS (R-ISS) on the test group (TeG). Multivariate Cox regression was performed for **a** event free survival (EFS) and **b** overall survival (OS). Shown is the hazard ratio on a logarithmic scale with a 95% confidence interval. p: p-value.

3.4 Stratification validation and testing

Validation of the prognostic impact of RPI and the risk-based stratifications (UAMS70-seq, RS-seq, EMC92-seq, IFM15-seq, HDHRS) was performed on independent patient cohorts of AMM and MMR (pathophysiological validation in different disease stages), as well as on external samples (CoMMpass cohort) of previously untreated myeloma patients as used for score determination in this thesis. Additionally, the HDHRS, generated on RNA-seq was transferred to and validated on DNA-microarrays. The proportions of the validation cohorts in comparison to the TeG are depicted in figure 3.16 and supplementary table B.8.

3.4.1 External testing in early stage and relapsed patients

Two independent cohorts of different patient populations in terms of stage (AMM, n=142) and disease phase (MMR, n=69) were used. For RPI stratification, trained on the TG 1, the 19 AMM included in the TG 1 are not used for validation.
In AMM due to the expected low frequency of the high risk group ($< 6\%$ in RPI, RS-seq and HDHRS, see figure 3.16 and supplementary table B.8), this group was merged with the medium risk group. RPI, UAMS-seq and RS-seq significantly delineates two groups in AMM (see supplementary figures A.2e, A.3e and A.4e, and supplementary table B.18), while the EMC92-seq, the IFM15-seq and the HDHRS are not predictive in AMM (see supplementary figures A.5e, A.6e, A.7e).

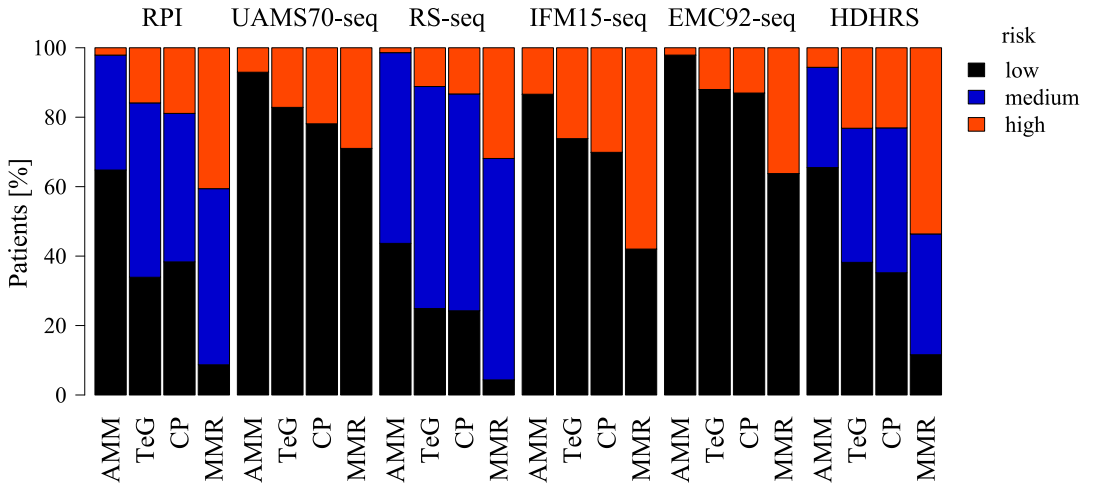

Figure 3.16: Proportions of patients regarding risk stratifications on the test and validation groups. The proportions are depicted for asymptomatic multiple myeloma patients (AMM), symptomatic multiple myeloma (MM) patients of the test group (TeG), MM patients of the CoMMpass cohort (CP) and relapsed MM patients (MMR). See also supplementary table B.8.

For MMR patients due to the low amount of low risk patients ($n \leq 8$ patients in RPI, RS-seq and HDHRS, see figure 3.16 and supplementary table B.8) low and medium risk group were merged. All stratifications, RPI, UAMS70-seq, RS-seq, EMC92-seq, IFM15-seq and HDHRS significantly delineate two groups for OS (see supplementary figures A.2f, A.3f, A.4f, A.5f, A.6f, A.7f, and supplementary table B.18).
Therefore, the stratifications can be biologically validated for early stages as well as relapsed patients.

3.4.2 External testing on CoMMpass cohort

For external validation on symptomatic MM patients, the CoMMpass cohort of the MMRF was used (see section 2.1.2).

Comparison of CoMMpass and HD cohort characteristics. Compared with HD cohort EFS and OS are to a large extend comparable (see figure 3.17a and 3.17b). The percentage of patients without progress after 1, 2 and 3 years is 80%, 60% and 45% in CoMMpass, compared to 87%, 64% and 45% in the HD cohort for EFS. For OS the survival percentages are 92%, 82% and 76% in CoMMpass and 95%, 88% and 78% in the HD cohort. The observation time of CoMMpass cohort is shorter (maximum EFS: 65 *versus* 143, maximum OS: 70 *versus* 178). The median age of the patients is higher in the CoMMpass cohort than in the HD cohort (CoMMpass: 64 years, HD: 59 years, see figure 3.17c). Both cohorts harbour similar proportions of chromosomal aberrations, despite the aberrations are determined with iFISH in the HD cohort and with RNA-seq and WGS in the CoMMpass cohort (see figure 3.17e and 3.17f).

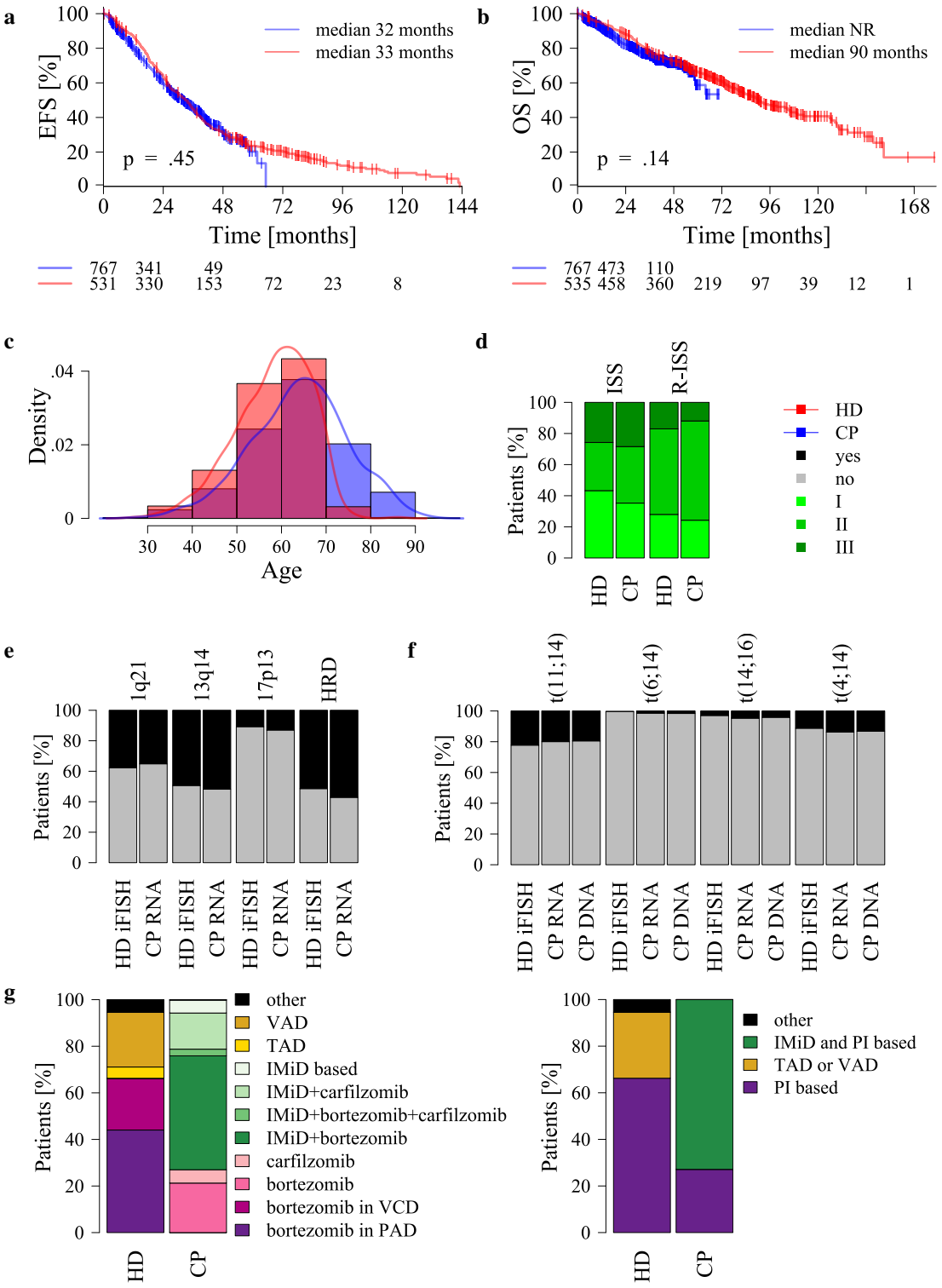

Figure 3.17: Patient characteristics of CoMMpass (CP) *versus* Heidelberg (HD) cohort. Comparison of **a** event free survival (EFS), **b** overall survival (OS), **c** density of the age distribution, **d** International staging system (ISS) stage and revised-ISS stage (R-ISS), **e** chromosomal aberrations, **f** IgH-translocations, and **g** treatment on whole symptomatic multiple myeloma cohorts. Chromosomal aberrations and IgH-translocations are determined with interphase fluorescence *in-situ* hybridisation (iFISH) in the HD cohort and with sequencing of RNA and DNA in the CP cohort. NR: not reached; HRD: hyperdiploidy; IMiD: immunomodulatory drugs (non-thalidomide, i.e. pomalidomide or lenalidomide); PI: proteasome inhibitor (bortezomib or carfilzomib); VAD: vincristine, adriamycin, dexamethasone; TAD: thalidomide, adriamycin, dexamethasone; VCD: bortezomib (velcade), cyclophosphamide, dexamethasone; PAD: bortezomib (formerly called PS-341), adriamycin, dexamethasone.

There are more patients with ISS stage I in the HD cohort (43%) than in the CoMMpass cohort (35%), while in the latter there are more patients with stage III (26% *versus* 28%, see figure 3.17d).

Significant differences are present between the two cohorts in terms of applied treatment regimen. The earliest included patients in the HD cohort were treated upfront with induction regimen of thalidomide, adriamycin and dexamethasone (TAD) or vincristine, adriamycin and dexamethasone (VAD). While 28% of the patients were administered these therapies in HD cohort, none of the CoMMpass patients received these (see figure 3.17g). The most frequent upfront therapy (66%) of the patients in HD cohort is based on the proteasome inhibitor bortezomib, either in combination with adriamycin and dexamethasone (PAD) or in combination with cyclophosphamide and dexamethasone (VCD). In the CoMMpass cohort, 27% of the patients were treated with a proteasome inhibitor (bortezomib or carfilzomib) and 52% were treated with bortezomib treatment in combination with one of the non-thalidomide immunomodulatory agents (lenalidomide or pomalidomide). IMiD-based treatment with or without combination therapy is used for 73% of the CoMMpass patients, while no HD patient received non-thalidomide IMiDs as induction regimen, although in part as maintenance treatment. In the HD cohort, all patients were intended to receive high-dose therapy and 99% of patients did receive autologous stem cell transplantation within the first line of therapy, compared with 50% of patients in CoMMpass cohort. Intrinsically different distributions of treatment regimen depict the different start to recruitment (i.e. 2002 for the HD cohort and 2012 for CoMMpass), as well as different treatment regimen in the participating countries, i.e. Germany *versus* the US and other European countries.

Evaluation of the continuous scores. The continuous RPI, UAMS70-seq, RS-seq, EMC92-seq, IFM15-seq and HDHRS scores are highly prognostic in MM patients in the CoMMpass cohort. The hazards significantly increase for OS (log-rank test $p < 0.001$, each). All scores are likewise prognostic for EFS (log-rank test $p < 0.001$, each), except the RPI, as its hazards are dependent on time i.e. the proportional hazard assumption is violated (see section 2.2.1).

Evaluation of the categorical stratification. The proportions of the risk groups are similar to the TeG proportions (see figure 3.16 and supplementary table B.8). The size of the high risk groups of TeG and CoMMpass of RPI is 15.8% *versus* 18.9%, of UAMS70-seq 17.2% *versus* 21.9%, of RS-seq 11.2% *versus* 13.3%, of EMC92-seq 12.0% *versus* 13.0%, of IFM15-seq 26.2% *versus* 30.1%, and of HDHRS 23.2% *versus* 23.1%.

In figure 3.18 EFS and OS are shown for the six risk stratifications for the CoMMpass

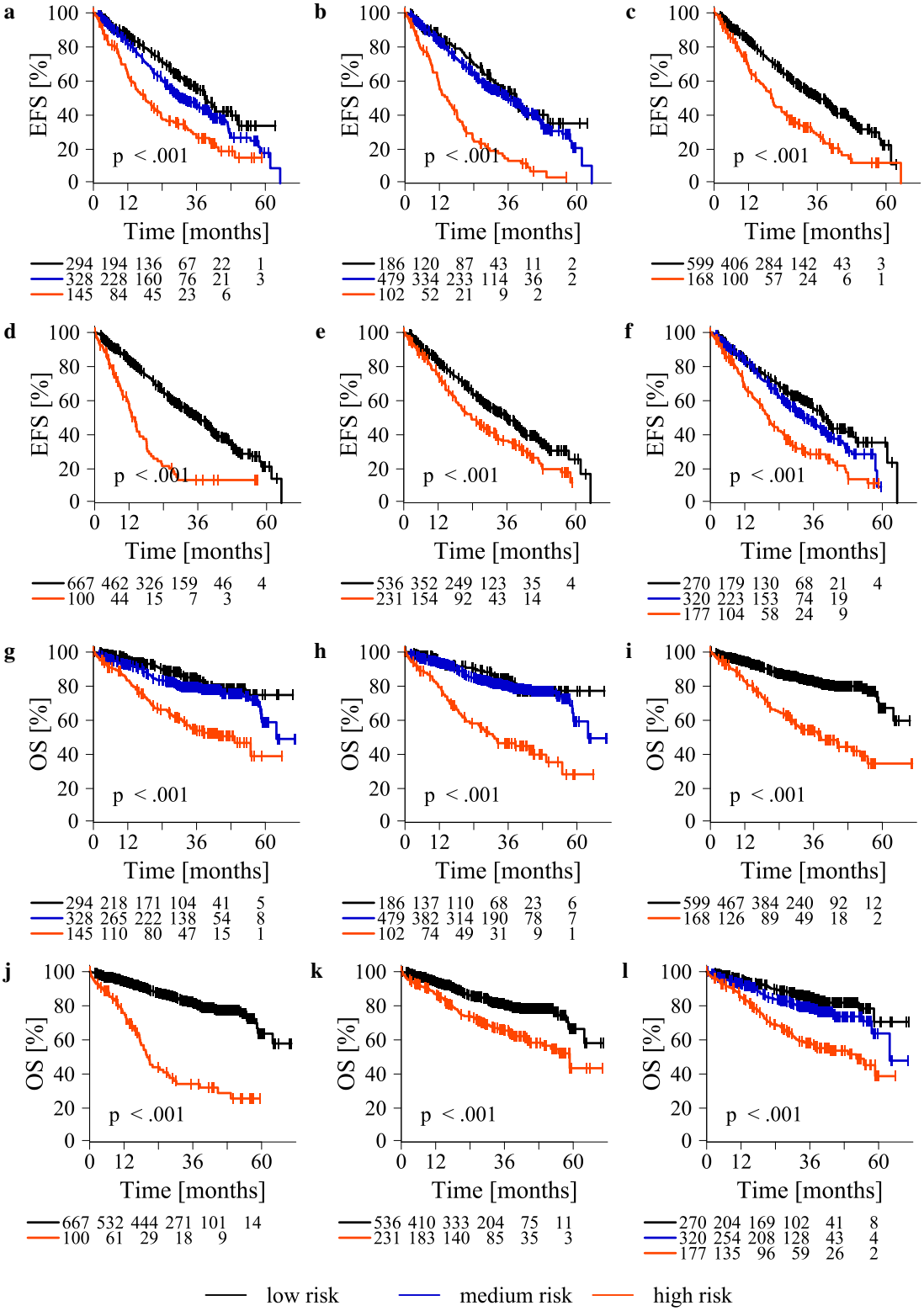

Figure 3.18: Survival analyses of risk stratifications and classifications on CoMMpass cohort. Event free survival (EFS) analyses is depicted for **a** RPI, **b** RS-seq, **c** UAMS70-seq, **d** EMC92-seq, **e** IFM15-seq and **f** HDHRS. Likewise, overall survival (OS) analyses is depicted for **g** RPI, **h** RS-seq, **i** UAMS70-seq, **j** EMC92-seq, **k** IFM15-seq, **l** HDHRS. p: p-value.

cohort. RPI (figures 3.18a and 3.18g) significantly delineates three groups, but the low and medium risk group are close. The same can be seen for the RS-seq (figures 3.18b and 3.18h) and HDHRS (figures 3.18f and 3.18l), which significant delineate three groups. EFS and OS for low and medium risk group are comparable. UAMS70-seq (figures 3.18c and 3.18i), EMC92-seq (figures 3.18d and 3.18j) and IFM15-seq (figures 3.18e and 3.18k) significantly delineate the two groups for EFS and OS.
Table 3.12 shows the results of the three different survival comparison methods (Brier score, R^2, and the concordance) for all risk stratifications for the CoMMpass cohort. The Brier score for EMC92-seq is the smallest, for EFS (0.1769, $p < 0.1$ but $p > 0.05$) and OS (0.1279). The R-ISS has the largest concordance for EFS (0.61) and the ISS for OS (0.66).

Table 3.12: Evaluation of the performance of risk prediction models on the CoMMpass cohort. The table shows the values of three different survival comparison methods for all risk stratifications: Brier scores (Brier), R^2 and concordance (C) with standard error (SE). A "-" indicates a p-value $p > 0.1$, * indicates $p < 0.05$ and ** indicates $p < 0.01$. Depicted are event free survival (EFS) and overall survival (OS) for **a** RNA-seq-based stratifications **b** international staging system (ISS) and revised ISS (R-ISS)

		EFS				OS			
		Brier	R^2	**C**	**C SE**	**Brier**	R^2	**C**	**C SE**
a	**HDHRS**	0.1804**	0.01	0.58	0.02	0.1376***	0.19	0.62	0.02
	RPI	0.1840*	-0.04	0.59	0.02	0.1361***	0.11	0.63	0.02
	UAMS70-seq	0.1867	-0.12	0.56	0.01	0.1343***	0.08	0.63	0.02
	RS-seq	0.1774*	0.05	0.59	0.01	0.1367***	0.09	0.62	0.02
	IFM15-seq	0.1865	-0.02	0.55	0.01	0.1434**	0.02	0.59	0.02
	EMC92-seq	0.1769	0.03	0.57	0.01	0.1279***	0.22	0.63	0.01
b	**ISS**	0.1794**	0.07	0.60	0.02	0.1335***	0.15	0.66	0.02
	R-ISS	0.1803**	0.02	0.61	0.02	0.1378***	0.11	0.65	0.03

3.4.3 HDHRS validation on microarray

Implementation. The RNA-seq based HDHRS was transferred to and validated on DNA-microarrays. For this, the 53 ENSGS were translated to probesets. In three cases, translation and valuation with jetset does not yield a result, although several probesets are present. Because ENSG00000188092 (*GPR89B*) matches 5 probesets, the online search tool GeneAnnot was used to find the probeset with best specificity and sensitivity (220642_x_at). The ENSG00000276234 matches two probesets and was translated in 210537_s_at on the basis of a smaller jetset score. The two ENSG (ENSG00000238269 and ENSG00000234068) are translated in the same SYMBOL (*PAGE2B*) and probeset (231307_at) and are therefore only used once (see also supplementary table B.14). Two genes, ENSG00000166415 (238253_at) and ENSG00000237424 (224456_s_at), have a very low correlation between RNA-seq and microarray expression ($r = -0.03$ and $r = 0.01$). This is due to the low and absent (A)

expression of the genes on microarray. The correlation of the calculated HDHRS-GEP with the HDHRS only change marginally if these two genes are included or not. Hence, for consistency, both genes were retained. As for RS, for HDHRS-GEP the expression of each gene was multiplied with the "prognosis factor" and the values were summed up.

New cutoffs were calculated with the running log-rank algorithm as described by Rème *et al.* [197], varying the parameters (see section 2.6). The 500 iterations resulted in 7 different cutoff sets, which were compared and evaluated visually in the same way as for HDHRS (see section 2.6). On the TG, all sets passed the filtering criteria. Hence, the four cutoff pairs with lowest Brier score and highest concordance for OS and EFS were chosen for validation on the VG. The Brier scores of two sets were significant for OS on VG. The one with the higher concordance for OS was selected and used for testing on the TeG. Initial parameters of this set are a minimal size of a risk group (w) of 18 (9%), an FDR threshold (fdr) of 0.05, a chi-squared statistic threshold ($x2$) between two survival curves of `qchisq((1-0.01),1)` and a minimal number of events per group (cx) of 2. Patients were classified in three groups, according to the low cutoff (*lcut=68.28*) and the high cutoff (*hcut=87.09*).

Survival plots for TG and VG for the HDHRS-GEP are shown in supplementary figure A.8. The results of the performance of the TeG are shown in figure 3.20.

Evaluation of the continuous score. The continuous HDHRS-GEP score is highly prognostic for the TeG in MM. Its hazard significantly increases over time, for OS (log-rank test $p < 0.001$). There is a significant (JHT test: $p = 0.001$) increase from MGUS to AMM, to MM to MMR (see figure 3.19). The HDHRS-GEP of PPC and MBC is similar to the HDHRS-GEP of MM. The HDHRS-GEP of PPC and HMCL is significantly higher than the HDHRS-GEP of MM, while the one of BMPC is significantly lower.

Evaluation of the categorical stratification. The HDHRS-GEP significantly delineate three groups of patients regarding EFS and OS (figure 3.20a and 3.20b), with a median for low, medium and high risk for EFS of 39, 29 and 21 and for OS of 148, 70 and 37 months. The proportions of low, medium and high risk group are 46.35%, 33.48% and 20.17%. The Brier score is significant regarding OS in TeG and has a lower value (0.1576) than the one of HDHRS on RNA-seq (0.1601) for OS. The concordance is the highest concordance of all microarray-based scores for EFS (0.61) and OS (0.66) (see figure 3.3). Univariate Cox regression showed significant hazard ratios in low *versus* high risk group for OS (4.61) and for EFS (2.69).

Comparison of both platforms. Pairwise comparison of HDHRS and HDHRS-GEP per patient in TeG shows few (1%) (see table 3.13 and figure 3.21) differences in terms

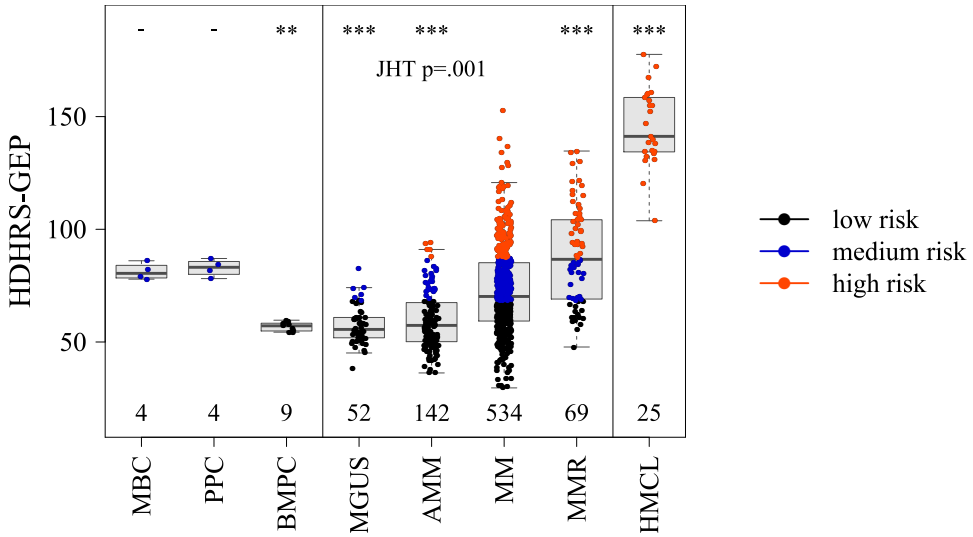

Figure 3.19: Continuous HDHRS-GEP score investigated on the whole cohort. HDHRS-GEP grouped by disease entity, compared to non-malignant plasma cells and precursors (BMPC, MBC, PPC) and human myeloma cell lines (HMCL). Significant differences are depicted by 1, 2 or 3 asterisks, indicating significant p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. A Jonckheere-Terpstra test (JHT) was performed to test the significance of the score increase from MGUS over AMM and symptomatic MM to MMR. MBC: memory B cell; PPC: polyclonal plasmablasts; BMPC: bone marrow plasma cell; MGUS: monoclonal gammopathy of undetermined significance; AMM: asymptomatic multiple myeloma; MM: multiple myeloma; MMR: relapsed MM.

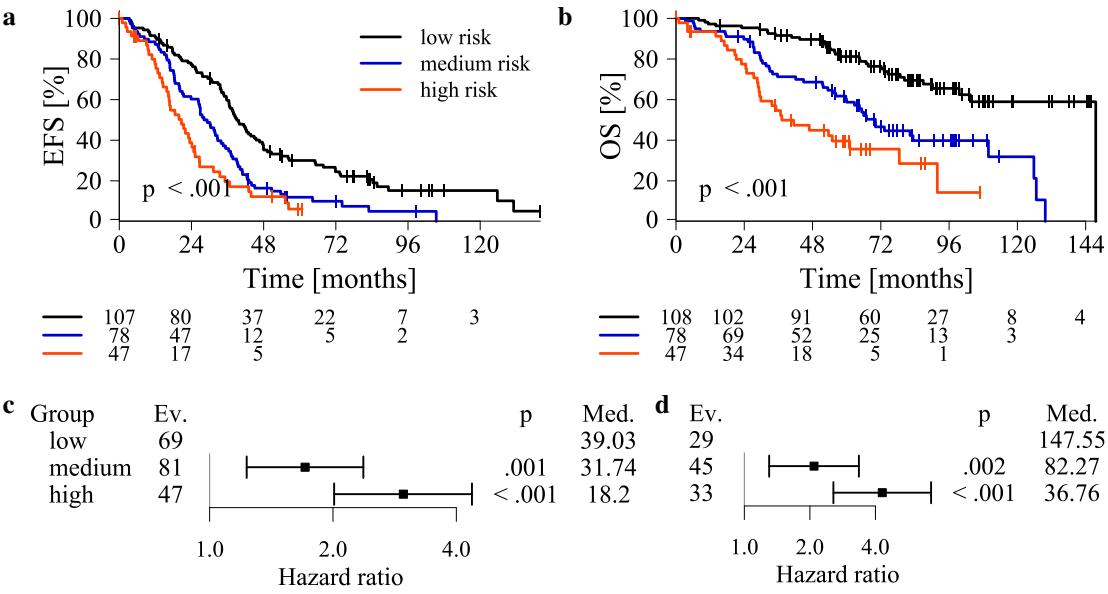

Figure 3.20: Survival analyses of HDHRS-GEP on the test group (TeG). Performance of the HDHRS-GEP of symptomatic multiple myeloma patient samples on the TeG in **a** event free survival (EFS) and **b** overall survival (OS). Univariate Cox regression was performed for **c** EFS and **d** OS. Shown is the hazard ratio on a logarithmic scale with a 95% confidence interval. Ev.: number of events; Med: median survival time in months; p: p-value.

of low to high risk group or *vice versa*. The consistency is 70%. HDHRS and HDHRS-GEP show a correlation coefficient of $r = 0.85$, with a confidence interval of 0.8-0.87.

Independent validation for asymptomatic and relapsed myeloma patients. The progression rate of the HDHRS-GEP for all AMM (see supplementary figure A.8e) is not significant, while the OS of the MMR (see supplementary figure A.8f) significantly delineates two groups, the high risk group and the combined low and medium risk group.

Table 3.13: Confusion matrix of HDHRS and HDHRS-GEP stratification on the TeG. Depicted is the number (and percentage) of patients per HDHRS group in rows and per HDHRS-GEP group in columns. In the top left of the table the consistency (*CO*) is depicted.

		HDHRS-GEP		
HDHRS	*CO = 70%*	**low risk**	**medium risk**	**high risk**
	low risk	*74 (32%)*	13 (6%)	2 (<1%)
	medium risk	34 (15%)	*50 (22%)*	6 (3%)
	high risk	0	15 (6%)	*39 (17%)*

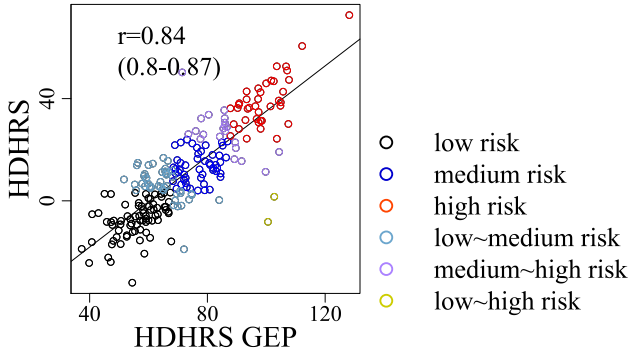

Figure 3.21: Comparison of HDHRS and HDHRS-GEP on test group (TeG). **a** correlation of HDHRS and HDHRS-GEP. Samples stratified in the low risk group of the HDHRS and medium risk group of the HDHRS-GEP or *vice versa* (low∼medium risk) are depicted in light blue, medium∼high risk samples are depicted in violet and low∼high or high∼low risk samples are depicted in yellow. The correlation coefficient (r) is displayed with its confidence interval.

3.5 Evaluation of potential targets

Microarrays and RNA-seq allow the assessment of targets of pharmacological agents. This intrinsically prerequisites that the target for a specific drug is known, which is especially the case for targets of immunotherapeutical approaches such as TCB monoclonal antibodies (see sections 3.5.2.1 and 4.4.2.1). Both, DNA-microarrays and RNA-seq allow assessing the expression of target genes. RNA-seq additionally enables the analysis of splice variants and mutated targets. In this thesis, 25 targets are exemplarily assessed, of which 15 are currently actionable (see table 3.14).

Table 3.14: Exemplary potential target list. The used translations of gene symbols (SYMBOL), microarray probeset and Ensembl gene identifier (ENSG) are depicted (all translations are listed in supplementary table B.19). In column one constitutively expressed genes are depicted in bold and an asterisk * indicates cell surface proteins. "Type" specifies the (non-exclusive) treatment strategy for a target: 1=antibody or CAR T cell therapy; 2=small molecule inhibitor; 3=mutation specific agent, 4=vaccination strategy, 5=theoretical siRNA target, frequently overexpressed in multiple myeloma and associated with a cancer relevant pathway. A + indicates actionable targets with available agent. *SSX2B* and *MAGEA6* were not selected.

Name	SYMBOL	PROBEID	ENSG	Type	Reference
BCMA*	TNFRSF17	206641_at	ENSG00000048462	1+	[212]
CD38*	CD38	205692_s_at	ENSG00000004468	1+	[55, 57, 213]
HM1.24*	BST2	201641_at	ENSG00000130303	4	[156, 209]
CD74*	CD74	209619_at	ENSG00000019582	1+	[1]
NYESO1/2*	CTAG2	207337_at	ENSG00000126890	1+ 4	[48, 156, 209]
	CTAG1A CTAG1B	210546_x_at	ENSG00000268651 ENSG00000184033		
HGF	HGF	210997_at	ENSG00000019991	2+	[193]
FGFR3*	FGFR3	204379_s_at	ENSG00000068078	2+	[232]
MAGEA1*	MAGEA1	207325_x_at	ENSG00000198681	4	[48, 156]
MAGEA3* MAGEA6*	MAGEA3 MAGEA6	209942_x_at	ENSG00000221867 ENSG00000197172	4+ -	[47, 48, 209]
MMSET*	WHSC1	209053_s_at	ENSG00000109685	5	[156]
IGF1R*	IGF1R	225330_at	ENSG00000140443	2+	[68, 223]
TP53	TP53	201746_at	ENSG00000141510	5	[156]
AURKA	AURKA	208079_s_at	ENSG00000087586	2+	[102]
CCND1	CCND1	208712_at	ENSG00000110092	5	[22, 156]
CCND2	CCND2	200953_s_at	ENSG00000118971	5	
CCND3	CCND3	201700_at	ENSG00000112576	5	
RHAMM	HMMR	207165_at	ENSG00000072571	4	[76, 81, 82, 209]
CD20*	MS4A1	228599_at	ENSG00000156738	1+	[231]
GPRC5D*	GPRC5D	221297_at	ENSG00000111291	1+	[186]
MUC1*	MUC1	213693_s_at	ENSG00000185499	4+	[30, 156]
CS1*	SLAMF7	222838_at	ENSG0000026751	1+	[45]
WT1*	WT1	206067_s_at	ENSG00000184937	4	[209, 233]
SSX2* SSX2B	SSX2 SSX2B	210497_x_at	ENSG00000241476 ENSG00000268447	4 -	[48, 156]
NKG2D*	KLRK1	1555691_a_at	ENSG00000213809	1+	[14, 15]
BRAF	BRAF	206044_s_at	ENSG00000157764	3+	[8]

The 15 "actionable" targets include targets for which compounds are approved[6] or in later stage development for multiple myeloma (e.g. BCMA, CD38, CD74, NYESO1/2, GPRC5D, NKG2D, see table 3.15) or have been described but are currently not in clinical testing for MM (e.g. AURKA [102] or IGF1R [68]). The treatment strategies for the actionable targets include 8 antibody or CAR T cell therapies, 4 clinical inhibitors and 2 vaccination strategies. For the mutated target BRAF approved inhibitors are available (e.g. vemurafenib [24, 109] and dabrafenib [91]), but not approved for MM. The ten targets until now not actionable in MM inlcude five potential antigens as targets for vaccination (HM1.24, MAGEA1, RHAMM, WT1 and SSX2 [100, 156, 209]) and five "theoretical" targets (MMSET, TP53, CCND1, CCND2 and CCND3). The latter are potentially targetable by small molecule inhibitors (e.g. CDK4/6 inhibitors [107, 172]) or siRNA-based strategies, but have not been tested clinically in myeloma up to now. All targets are listed in table 3.14.

Fourteen of the 25 targets had already been presented in the GEP-R (6 actionable, 3 until now not actionable vaccination, and 5 only theoretical targets). These were translated from probesets to SYMBOLs and ENSGs. The remaining 8 actionable and 3 until now not actionable vaccination targets were translated from SYMBOL to probesets and ENSGs. All pairs of probesets and ENSGs for the 25 targets are listed in supplementary table B.19. For three probesets more than one related gene SYMBOL and ENSG exists, see table 3.14. First, the probeset 210546_x_at of *NYESO1/2* used in the GEP-R, matches to three ENSGs and three SYMBOLs: *CTAG2*, *CTAG1B* and *CTAG1A*. For them, 2, 2 and 3 transcripts are known, respectively. Regarding GeneAnnot, the sensitivity of *CTAG2*, *CTAG1B* and *CTAG1A* is identical (0.636) and the specificity is below 0.41 in all three cases. Using the GeneCards database, the three genes are paralog to each other[7]. Hence, no gene was excluded and the maximal RNA-seq expression per sample of the three ENSGs was used as comparison to the microarray expression. Second, *MAGEA3* and *MAGEA6* fit both probeset 209942_x_at. GeneAnnot depicts the sensitivity (1 vs. 0.909) and specificity (0.462 vs. 0.439) as higher for *MAGEA3*, hence this gene was used in comparison. Third, the ENSG ENSG00000268447, matching to *SSX2B* and 210497_x_at was rejected, because it was not found on GeneAnnot, and instead the paralog[8] SSX2 (ENSG00000241476) was used.
Of the 11 targets, which were translated from SYMBOL to probesets and ENSGs, for

[6]National Cancer Institute: Drugs Approved for Multiple Myeloma and Other Plasma Cell Neoplasms; Online resource: `https://www.cancer.gov/about-cancer/treatment/drugs/multiple-myeloma`; Status: 27.02.2020, 12:55

[7]GeneCards CTAG1A; Online resource: `https://www.genecards.org/cgi-bin/carddisp.pl?gene=CTAG1A&keywords=CTAG1A`; Status: 15.10.2019, 13:33

[8]GeneCards SSX2; Online resource: `https://www.genecards.org/cgi-bin/carddisp.pl?gene=SSX2&keywords=SSX2B`; Status: 15.10.2019, 13:34

7 SYMBOLs (*CD74*, *HGF*, *RHAMM*, *CD20*, *CS1*, *WT1*, *NKG2D*) and ENSGs more than one probeset exists (3, 5, 2, 4, 3, 2, 2, respectively). For each gene the probeset with the largest jetset score were selected, see table 3.14 and supplementary table B.19.

Table 3.15: Exemplary targets in clinical trials. Not comprehensive list of targets with clinical grade agents which are currently in later stage development. The trial names are extracted from the database `clinicalTrials.gov`; Status: 21.04.2020, 11:29.

Target name	Trial name	Status	ClinicalTrials.gov identifier
BCMA	Study of CC-93269, a BCMA x CD3 T Cell Engaging Antibody, in Subjects With Relapsed and Refractory Multiple Myeloma	Ongoing	NCT03486067
	Assessment of AMG 420 in Subjects With Relapsed and/or Refractory Multiple Myeloma (AMG420)	Ongoing	NCT03836053
CD38	Study to Evaluate the Safety and Efficacy of Anti-CD38 CAR-T in Relapsed or Refractory Multiple Myeloma Patients	Ongoing	NCT03464916
CD74	Study of STRO-001, an Anti-CD74 Antibody Drug Conjugate, in Patients With Advanced B-Cell Malignancies	Ongoing	NCT03424603
HGF	A Phase 2 Trial of MP0250 Plus Bortezomib + Dexamethasone in Patients With Multiple Myeloma	Ongoing	NCT03136653
NYESO1/2	Redirected Auto T Cells for Advanced Myeloma	Completed	NCT01352286
GPRC5D	Dose Escalation Study of JNJ-64407564 in Participants With Relapsed or Refractory Myeloma	Ongoing	NCT03399799
CS1	CS1-CAR T Therapy Following Chemotherapy in Treating Patients With Relapsed or Refractory CS1 Positive Multiple Myeloma	Ongoing	NCT03710421
NKG2D	A Dose Escalation Phase I Study to Assess the Safety and Clinical Activity of Multiple Cancer Indications (THINK)	Ongoing	NCT03018405

3.5.1 Target expression

The RNA-seq expression of the exemplary 24 actionable, vaccination and theoretical targets (*BRAF* is not shown) was contrasted with their microarray expression, comparing non-malignant cells and myeloma cell lines with the myeloma stages (see figure 3.22 and supplementary figures A.9, A.10,A.11, A.12, A.13, A.14 and A.15), the PA-seq call with the PA call (see supplementary table B.20), overexpression (see supplementary table B.21), survival association and correlation of expression using the different platforms (see supplementary figure A.16).

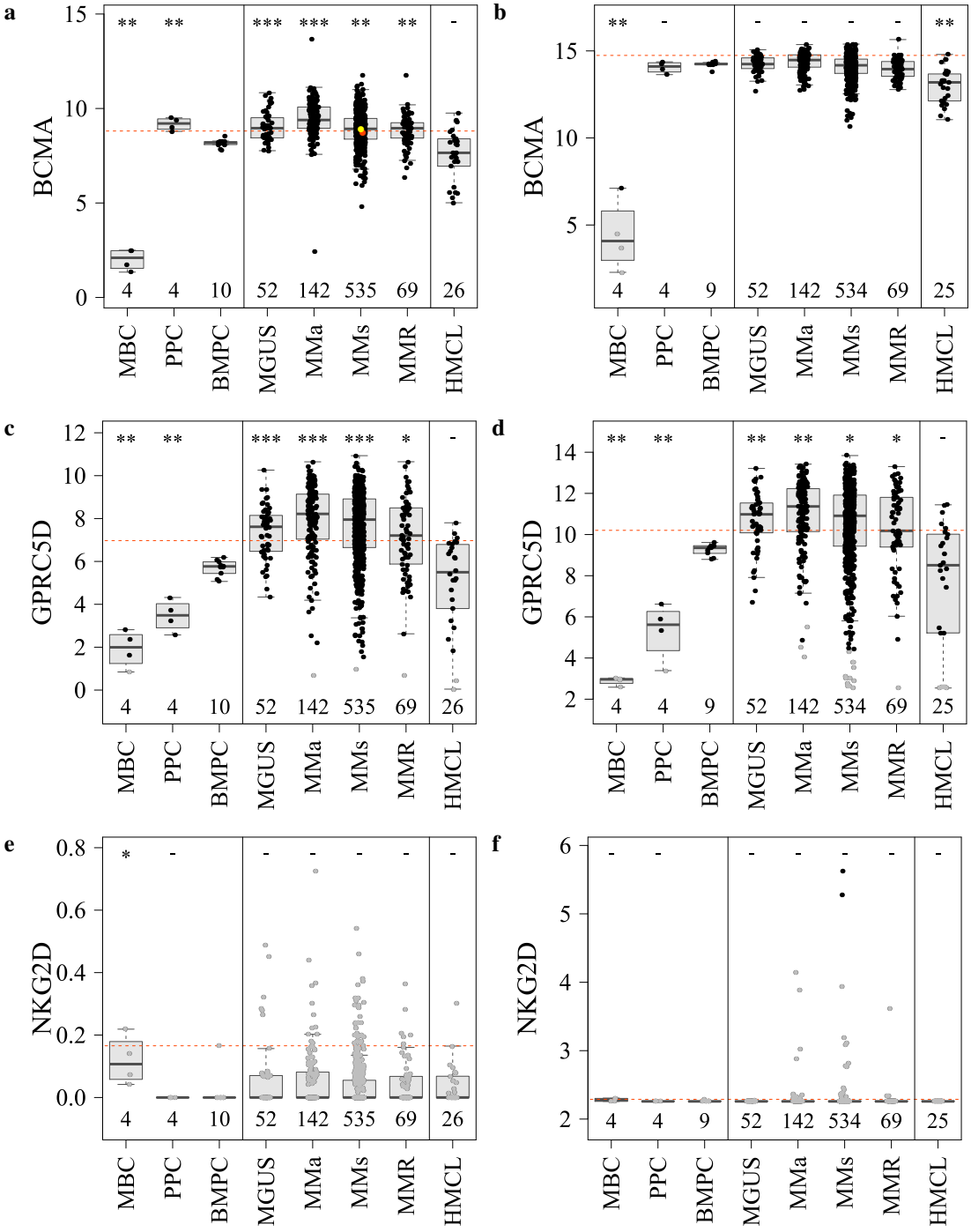

Figure 3.22: Expression of *BCMA*, *GPRC5D* and *NKG2D*. The whole cohort is assessed. Expression is grouped by disease entity, compared to non-malignant plasma cells and precursors (BMPC, MBC, PPC) and human myeloma cell lines (HMCL). Significant differences in expression compared to BMPCs are depicted by 1, 2 or 3 asterisks, for p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. RNA-seq expression is depicted for **a** *BCMA* (published with a smaller cohort in Seckinger, ..., Emde et al., Cancer Cell 2017 [212]), **c** *GPRC5D* and **e** *NKG2D*. Microarray expression is depicted for **b** *BCMA* (published with altered sample composition in Seckinger, ..., Emde et al., Cancer Cell 2017 [212]), **d** *GPRC5D* and **f** *NKG2D*. The red dashed line depicts the threshold for overexpression, defined as median expression in BMCPs plus 3 times the standard deviation. Two exemplary patients are highlighted with red (patient 1) and yellow (patient 2) dot for *BCMA*. MBC, memory B cells; PPC, polyclonal plasmablastic cells; BMPC, bone marrow plasma cells; MGUS, monoclonal gammopathy of undetermined significance; AMM, asymptomatic multiple myeloma, MM, symptomatic multiple myeloma; MMR, relapsed myeloma.

The results are published in parts by Seckinger *et al.* [212] (expression of *BCMA*), Seckinger *et al.* [213] (expression and survival analysis of *CD38*), and Schmitt *et al.* [209] (expression, correlation and survival analyses of *HM1.24*, *NYESO1/2*, *MAGEA3*, *RHAMM* and *WT1*). A summary of the results for all targets is depicted in table 3.16. The targets can be divided in four groups according to their expression: i) constitutively expressed in plasma cells, ii) aberrantly expressed, iii) overexpressed and iv) non-overlapping in RNA-seq and DNA-microarrays (NA). The expression of three exemplary targets, *BCMA*, *GPRC5D* and *NKG2D*, is depicted in figure 3.22.

i) The RNA-seq and microarray expression of *BCMA*, *CD38*, *HM1.24* and *CD74* show a correlation coefficient *r* of 0.66, 0.68, 0.60, and 0.73, respectively. All four genes are expressed in all samples in PPC, BMPCs, MGUS, AMM, MM, and MMR, except one MM sample with absent microarray expression regarding *HM1.24*. As exemplified for *CD38* (see figure 3.22), the expression of the MBCs is significantly lower than the expression of BMPC or MM, while MBC expression on microarray is absent in at least one sample. This pattern is found likewise for *BCMA* and *HM1.24*. The expression of *CD38* and *CD74* (see figure 3.22) is higher in BMPCs than in MM, whereas for *BCMA* and *HM1.24* the expression in BMPC is lower or similar to MM. BCMA is overexpressed in comparison to BMPCs (for the definition see 2.8.1) in 57 MM samples on microarray and in 299 samples on RNA-seq, *CD38* in 8 each, *HM1.24* in 175 and 336 and *CD74* in 71 and 27. Neither *BCMA* nor *HM1.24* are associated with survival. High *CD38* expression was previously published by the LfM [213] as associated with good prognosis on microarray, with a borderline significant p-value of 0.03 for EFS and of 0.02 for OS. In this thesis and a larger cohort, *CD38* marginally fails survival association with borderline (unadjusted) p-values in maxstat test of 0.08 (adjusted: 0.12) for microarray and 0.055 (adjusted: 0.08) for RNA-seq (see supplementary table B.22 for adjusted p-values).

ii) *NYESO1/2*, *HGF*, *FGFR3*, *MAGEA1* and *MAGEA3* are neither expressed in BMPCs nor MBCs, except one BMPC sample categorised as borderlinely expressing *HGF* using the RNA-seq expression threshold. Regarding unnormalised raw read counts, only *NYESO1/2* is not expressed in BMPCs. The expression in PPC is absent in all cases, except for *HGF* expression on RNA-seq, which is present at a very low level (see supplementary figure A.10). *NYESO1/2* is expressed in 67 MM samples (12.5%) on microarray and 64 samples on RNA-seq (12.0%), *HGF* in 448 (83.9%) and 504 (94.2%), *FGFR3* in 55 (10.3%) and 42 (78.5%), *MAGEA1* in 92 (17.2%) and 140 (26.2%) and *MAGEA3* in 215 (40.3%) and 174 (32.5%). The consistency (*CO*), comparing the present "P" samples in RNA-seq and "MP" samples on DNA-microarrays,

Table 3.16: Target presence, overexpression and survival association using microarray and RNA-sequencing (RNA-seq) investigated on the whole cohort. **i)** constitutively expressed in plasma cells, **ii)** aberrantly expressed, **iii)** overexpressed and **iv)** non-overlapping in RNA-seq and DNA-microarrays. Shown are the correlation coefficient *r*, present expression with consistency (*CO*) and the percentage of patients with present PA-seq call and absent PA call (nCO_1), and overexpression with *CO* between the two platforms. The survival association of high expression is summarised in the last column. Present survival association has always the same direction (i.e. adverse/adverse or good/good) on both platforms. $^-$ indicates good survival association of high expression. NA depicts non-overlapping results in RNA-seq and microarray analysis. **CD38* was previously published at the LfM [213] as associated with survival on microarray, with borderline significant p-value of 0.02 in overall survival (OS). In this thesis, using a subset of patients only, *CD38* is not associated with OS with borderline, but not significant (unadjusted) p-values of 0.08 (DNA-microarrays) and 0.055 (RNA-seq).

			Microarray			**RNA-seq**			**Comparison**			
	Name	**r**	**present BMPC (n=9)**	**present MM (n=534)**	**overexpressed**	**present BMPC (n=10)**	**present MM (n=535)**	**overexpressed**	**CO present/absent [%]**	**nCO_1 present/absent [%]**	**CO overexpressed [%]**	**survival association**
i)	BCMA	0.66	9	534	57	10	535	299	100	0	54	no
	CD38	0.68	9	534	8	10	535	8	100	0	98	yes$^-$*
	HM1.24	0.60	9	533	175	10	535	336	100	0	63	no
	CD74	0.73	9	534	71	10	535	27	100	0	11	yes$^-$
ii)	NYESO1/2	0.73	0	67	67	0	64	64	92	4	90	yes
	HGF	0.85	0	448	448	1	504	469	88	1	91	no
	FGFR3	0.90	0	55	55	0	42	42	97	3	97	yes
	MAGEA1	0.80	0	92	92	0	140	140	86	2	83	yes
	MAGEA3	0.82	0	215	215	0	174	174	85	11	84	yes
iii)	MMSET	0.78	9	245	84	10	515	59	50	0	73	yes
	IGF1R	0.68	1	166	38	9	412	11	52	1	82	yes
	TP53	0.64	9	512	308	10	525	173	95	1	44	no
	AURKA	0.76	7	390	346	6	379	64	81	11	39	yes
	CCND1	0.79	4	430	387	7	450	388	89	4	90	yes$^-$
	CCND2	0.86	9	299	206	10	393	192	78	2	87	yes
	CCND3	0.79	9	531	10	10	534	25	99	0	95	no
	RHAMM	0.74	9	462	114	10	473	139	88	5	78	yes
	CD20	0.76	4	156	76	10	443	77	44	1	90	no
	GPRC5D	0.84	9	522	350	10	534	368	98	0	16	no
iv)	MUC1	0.63	9	527	152	0	68	68	14	86	46	yes
	CS1	0.71	9	533	18	10	534	6	100	0	96	NA
	WT1	0.12	0	5	5	0	3	3	99	1	99	NA
	SSX2	0.08	0	52	52	0	0	0	90	10	90	NA
	NKG2D	0.03	0	2	2	0	0	0	100	0	0	no

is $CO \geq 85\%$ for all genes (see section 2.2.2). For all MM samples, for most genes categorised as present using RNA-seq, these are likewise overexpressed in comparison to BMPCs. This holds true for *NYESO1/2*, *FGFR3*, *MAGEA1* and *MAGEA3*. For *HGF*, of the 504 myeloma patient samples showing present expression (6.9%) 35 do not show overexpression using RNA-seq. A significant association with EFS and OS was found for *NYESO1/2*, *FGFR3*, *MAGEA1* and *MAGEA3*, while no association was found for *HGF*.

iii) RNA-seq and microarray expression of *CCND2* show a correlation coefficient of $r = 0.86$, *CCND1*, *CCND3*, *MMSET*, *AURKA*, *CD20* and *GPRC5D* show a correlation coefficient of $r \geq 0.76$, *RHAMM* and *IFG1R* of $r \geq 0.66$ and *TP53* of $r = 0.64$. *CCND2*, *CCND3*, *MMSET*, *TP53*, *RHAMM* and *GPRC5D* are found expressed in all BMPC samples on both platforms, while *CCND1*, *IGF1R*, *AURKA* and *CD20* are expressed in 4, 1, 7 and 4 BMPCs on microarray and 7, 9, 6, 10 BMPCs using RNA-seq. *CCND3* and *TP53* are expressed in all cell types in this analysis, including MBCs, with 531 (99.4%) and 512 (95.9%) MM samples with present expression on microarray and 534 (99.8%) and 525 (98.1%) on RNA-seq. *MMSET*, *IGF1R* and *CD20* are more often found expressed in RNA-seq compared with microarray data: 245 (45.9%), 166 (31.1%) and 156 (29.2%) patients show expression in DNA-microarrays *versus* 515 (96.2%), 412 (77.0%) and 443 (82.8%) in RNA-seq. The consistency is $CO \leq 52\%$ for all three gene. Frequent overexpression is detected for *CCND1*, *CCND2*, *TP53*, *RHAMM* and *GPRC5D* in 387 (72.5%), 206 (38.6%), 308 (57.7%), 114 (21.3%) and 350 (65.5%) patient samples by microarrays, and 388 (72.5%), 192 (35.9%), 173 (32.3%), 139 (26.0%) and 368 (68.8%) MM in RNA-seq. Significant association with EFS and OS is found for *CCND1*, *CCND2*, *MMSET*, *IFG1R*, *AURKA* and *RHAMM*, while no association is found for *TP53*, *CCND3*, *CD20* and *GPRC5D*.

iv) The correlation coefficient *r* for *WT1*, *SSX2* and *NKG2D* is below 0.25. SSX2 expression is not detected by RNA-seq, but frequently present in DNA-microarrays. *WT1* is present and overexpressed in a few patient samples in microarray analyses (n=5) and on RNA-seq (n=3), while *NKG2D* is only present and overexpressed in very few patient samples in microarray analyses (n=2). In contrast to the latter targets, *MUC1* and *CS1* (see figure 3.22) have a correlation coefficient *r* of 0.63 and 0.71. *MUC1* is found expressed in all BMPCs and MBCs using DNA-microarrays, but neither in BMPC nor MBC using RNA-seq. Of the 534 samples on microarray, 527 (98.7%) show present expression, while using RNA-seq only 68 (12.7%) express *MUC1*. Hence, the consistency is the lowest of all targets, with 14%. *MUC1* is significantly associated with survival for EFS and OS. *CS1* is expressed in 533 MM patient samples on DNA-microarray and 534 on RNA-seq. *CS1* expression is found to be significantly

associated with survival by the maxstat test for EFS and OS on RNA-seq, but not on DNA-microarrays.

Additionally, within the publication of Schmitt *et al.* [209] the number of expressed CTAs was analysed. In table 3.17 *NYESO1/2*, *MAGEA3*, *RHAMM* and *WT1* are depicted. Considered individually, each is expressed in 67 (13%), 215 (40%), 462 (87%), and 5 (1%) MM patient samples on microarray and in 64 (12%), 174 (33%), 473 (88%), and 3 (1%) on RNA-seq. Combining the targets, 91% of the patients show expression of at least one of the four CTAs.

Table 3.17: Presence of Cancer testis antigens (CTAs) using microarrays and RNA-seq investigated on the whole cohort. Shown are the number of present CTA (*MAGEA3*, *WT1*, *RHAMM* and *NYESO1/2*) expression on the two platforms. The number of BMPC and MM samples with at least one read is shown for RNA-seq per gene in grey colour.

	Microarray		**RNA-seq**			
Number of CTAs	**present BMPC (n=9)**	**present MM (n=534)**	**BMPC $\geq$ 1 read (n=10)**	**MM $\geq$ 1 read (n=535)**	**present BMPC (n=10)**	**present MM (n=535)**
0	0	47	0	1	0	45
1	9	271	4	57	10	310
2	0	172	4	152	0	136
3	0	42	2	187	0	44
4	0	2	0	138	0	0

3.5.2 Splice variants

Alternative splicing may eliminate the target sequence of especially immunotherapeutical agents. Using RNA-seq, the determination of splice variants is possible. Exemplary, the splice junctions of the two immunotherapeutical target genes *BCMA* and *CD38* were further analysed in this thesis.

3.5.2.1 BCMA

The cell surface antigen BCMA is mandatory for the survival of long-living BMPCs [180] and considered an ideal target as published by Seckinger *et al.* [212]. As part of the development of a BCMA-TCB antibody (CC-93269, [212], ClinicalTrials.gov identifier: NCT03486067), BCMA splice variants have been analysed and presented [98].

Three transcripts for *BCMA* are annotated, see figure 3.23. All three are protein encoding and composed of 8 exons in total. BCMA-001, BCMA-002 and BCMA-003, consist of 3, 4, and 2 exons, respectively. Four of the five splice junctions are transcript specific, while each transcript has at least one specific splice junction. Seven non-annotated splice junctions were detected with each matched by a maximum of 0.6% of all reads spanning BCMA splice junctions. These were therefore not further analysed.

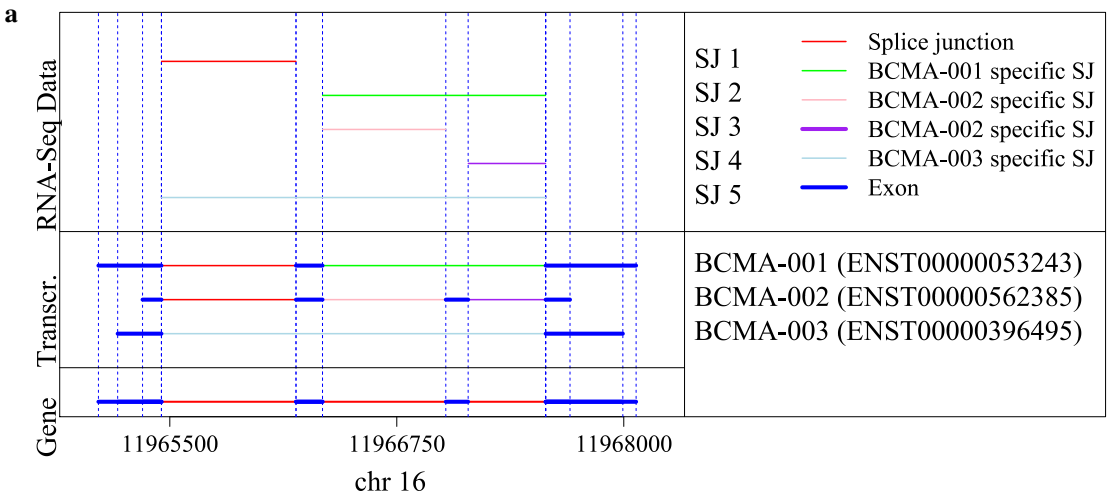

Figure 3.23: Analysis of alternative splicing of BCMA. The figure consists of three parts, containing the structure of the gene locus (bottom part), the different transcripts (centre part) and the splice junctions (top part). Exons are depicted in dark blue and splice junctions (SJ) in red, while specific splice junctions are of contrasting colour. The three transcripts for BCMA are composed by 5 splice junctions and 8 exons. The splice junction specific for BCMA-001 (SJ 2) is depicted in green, the two specific for BCMA-002 (SJ 3 and SJ 4) in light lilac and lilac, and the one specific for BCMA-003 (SJ 5) in light blue. RNA-seq: RNA-sequencing; chr: chromosome.

The splice junction specific for BCMA-001, SJ 2, is expressed with at least 10 reads in all patient samples, while SJ 3 and SJ 4, specific for BCMA-002 are expressed in 40/52 and 43/52 MGUS (76.9% and 82.7%), 22/29 and 20/29 AMM (75.9% and 69.0%), and 259/388 and 263/388 MM (66.8% and 67.8%) patient samples and SJ 5, specific for BCMA-003 is expressed in 28/52 MGUS (53.8%), 18/29 AMM (62.1%), and 230/388 MM (59.3%) patient samples (see table 3.18).

The minimum percentage of reads spanning the BCMA-001 specific splice junction is 43.74% in MM, while the maximum percentages of BCMA-002 specific splice junctions, SJ 3 and SJ4 are 3.51% and 3.69%, and the maximum percentage of BCMA-003, SJ 5, is 2.55%.

Regarding the full-length transcript of BCMA-001, the percentage of patient samples containing all SJ of BCMA-001 is 100%, while the percentage of patient samples containing only SJ of BCMA-001 is 7.7% for MGUS, 10.3% for AMM and 19.8% for MM (see table 3.18). The number of patient samples having reads spanning any other

SJ than the ones for BCMA-001 above the levels of 1% is 28 MGUS, 17 AMM and 200 MM and above the levels of 5% it is 1 MGUS, 3 AMM and 3 MM. No patient sample exceeds the level of 10%.

In supplementary table B.23 the number of reads per splice junction is depicted for two exemplary patients, one with only a BCMA-001 specific splice junction and one with additional splice junctions. The results of the two exemplary patients are confirmed by RSEM transcript analysis (see supplementary figure A.20). In example one, the BCMA-002 and BCMA-003 specific SJs are present, but the maximum number of reads in BCMA-001 is 50 times higher. In example 2 the maximum number is even 100 times higher in BCMA-001. In conclusion, BCMA-001 is the main transcript of *BCMA*.

Table 3.18: Splice junctions of *BCMA* per cohort and relative abundance. **a** The 5 splice junctions (SJ) of BCMA are depicted. SJ 2 is specific for BCMA-001, SJ 3 and SJ 4 for BCMA-002 and SJ 5 for BCMA-003. The number of samples in which the respective SJ is expressed, is depicted for memory B cells (MBC), polyclonal plasmablasts (PPC), bone marrow plasma cells (BMPC), monoclonal gammopathy of undetermined significance (MGUS), asymptomatic multiple myeloma (AMM), multiple myeloma (MM) and human myeloma cell lines (HMCL). In the last two columns are the minimum (min) and maximum (max) percentages of raw reads for the specific SJs in comparison to the reads for all other annotated SJs in MM patient samples. **b** BCMA-001 transcript and the percentage of patient samples containing only this SJ. In the last three columns, the percentage of reads spanning any other SJ above the levels of 1%, 5% and 10% is delineated.

a

								MM	
	MBC	**PPC**	**BMPC**	**MGUS**	**AMM**	**MM**	**HMCL**	**Min %**	**Max %**
n	4	4	9	52	29	388	26	-	-
SJ 1	3	4	9	52	29	388	26	37.21	53.72
SJ 2	4	4	9	52	29	388	26	43.74	62.79
SJ 3	0	4	6	40	22	259	22	0.00	3.51
SJ 4	0	4	6	43	20	263	23	0.00	3.69
SJ 5	0	4	0	28	18	230	22	0.00	2.55

b

		BCMA-001		**BCMA-001 only**		**Number of reads spanning other annotated SJ**		
	n	**n**	**[%]**	**n**	**[%]**	**>1%**	**>5%**	**>10%**
MBC	4	3	75.0	3	75.0	4	0	0
PPC	4	4	100	0	0.0	4	0	0
BMPC	9	9	100	1	11.1	4	0	0
MGUS	52	52	100	4	7.7	28	1	0
AMM	29	29	100	3	10.3	17	3	0
MM	388	388	100	77	19.8	200	3	0
HMCL	26	26	100	3	11.5	19	0	0

3.5.2.2 CD38

CD38 is an immunological target for which approved and developmental drugs are available (see 1.5). Interestingly, despite malignant plasma cells from almost all myeloma patient samples express *CD38*, the compound is active as single agent in

only about one thirds of the patients [213]. Thus, alternative splicing eliminating the binding site of therapeutic antibodies was considered as a potential mechanism of up-front resistance. Results from this thesis regarding CD38 splice variants have already been published [213].

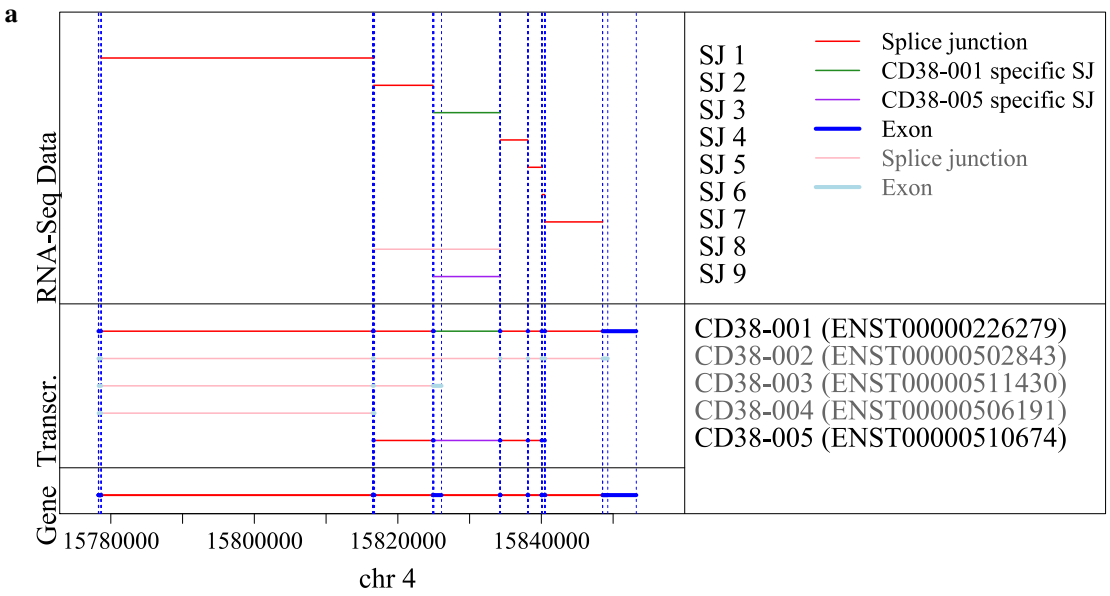

Figure 3.24: Analysis of alternative splicing of *CD38*. The figure consists of three parts, containing the structure of the gene locus (bottom part), the different transcripts (centre part) and the splice junctions (top part). Exons are depicted in dark blue and splice junctions (SJ) in red, while specific splice junctions are of contrasting colour. Five transcripts for *CD38* are annotated in GRCh38, composed by 9 splice junctions and 17 exons. The splice junction specific for CD38-001 (SJ 3) is depicted in green and the one specific for CD38-005 (SJ 9) in lilac (adapted from Seckinger, ..., Emde et al., Frontiers in Immunology 2018 [213]). RNA-seq: RNA-sequencing; chr: chromosome.

In the reference genome GRCh38 five transcripts for *CD38* are annotated, see figure 3.24. Two are protein encoding (CD38-001 and CD38-005), two are non-protein encoding due to retained intron sequences and one shows a nonsense-mediated decay (CD38-002). The five transcripts are composed of 22 exons. Of these 17 are unique exons. CD38-001, CD38-002, CD38-003, CD38-004, CD38-005 consists of 8, 7, 3, 2 and 6 exons, respectively. Nine splice junctions are annotated. Three of these are transcript specific, for the transcripts CD38-001, CD38-002 and CD38-005. Additionally, nine non-annotated splice junctions were detected. Each matched by a maximum of 2.06% of all reads spanning CD38 splice junctions. As for *BCMA*, these were not further analysed.

The splice junction specific for CD38-001 is expressed in all 469 plasma cell samples. The one specific for CD38-005 is expressed in 3/52 MGUS (5.8%), 1/29 AMM (3.4%), and 14/388 MM patient samples (3.6%), respectively (see table 3.19).

The minimum percentage of reads spanning a CD38-001 specific splice junction is 10.36 in MM, while the maximum percentage of CD38-002 or CD38-005 specific

Table 3.19: Splice junctions of CD38 per cohort and relative abundance. **a** The 9 splice junctions (SJ) of CD38 are depicted (adapted from Seckinger, ..., Emde et al., Frontiers in Immunology 2018 [213]). SJ 3 is specific for CD38-001, SJ 8 for CD38-002 and SJ 9 for CD38-005. The number of samples in which the respective SJ is expressed, is depicted for memory B cells (MBC), polyclonal plasmablasts (PPC), bone marrow plasma cells (BMPC), monoclonal gammopathy of undetermined significance (MGUS), asymptomatic multiple myeloma (AMM), multiple myeloma (MM) and human myeloma cell lines (HMCL). In the last two columns are the minimum (min) and maximum (max) percentages of raw reads for the specific SJs in comparison to the reads for all other annotated SJs in MM patient samples. **b** Depicted is the percentage of patient samples containing all splice junctions (SJ) of the CD38-001 (adapted from Seckinger, ..., Emde et al., Frontiers in Immunology 2018 [213]) transcript and the percentage of samples containing only this SJ. In the last three columns, the percentage of reads spanning any other SJ above the levels of 1%, 5% and 10% is delineated.

a

								MM	
	MBC	**PPC**	**BMPC**	**MGUS**	**AMM**	**MM**	**HMCL**	**Min %**	**Max %**
n	4	4	9	52	29	388	26	-	-
SJ 1	2	4	9	52	29	388	24	11.04	18.18
SJ 2	2	4	9	52	29	388	23	12.48	20.78
SJ 3	2	4	9	52	29	388	22	10.36	18.6
SJ 4	2	4	9	52	29	388	23	8.33	16.88
SJ 5	2	4	9	52	29	388	22	11.01	18.37
SJ 6	2	4	9	52	29	388	22	10.61	16.76
SJ 7	2	4	9	52	29	388	23	9.77	23.84
SJ 8	0	1	4	14	10	125	6	0	2.08
SJ 9	0	0	1	3	1	14	1	0	0.25

b

		CD38-001		**CD38-001 only**		**Number of reads spanning other annotated SJ**	
	n	**n**	**[%]**	**n**	**[%]**	**>1%**	**>5%**
MBC	4	1	25.0	1	25.0	0	0
PPC	4	4	100	2	50.0	0	0
BMPC	9	9	100	3	30.0	0	0
MGUS	52	52	100	36	69.2	3	0
AMM	29	29	100	17	58.6	0	0
MM	388	388	100	247	63.7	1	0
HMCL	26	22	84.6	15	57.7	0	0

splice junction is 2.08 and 0.25.

Comparing the number of raw reads of all splice junctions in the 18 patient samples with present CD38-005 specific splice junction (see supplementary figure A.17), the seven splice junctions of CD38-001 are either less present (below 100% to 81%) or more present (above 100% to 130%). The maximum observed frequency of the CD38-005 specific splice junction in the 18 patient samples is 1.8%, in comparison to CD38-001 specific splice junction.

Regarding the full-length transcript of CD38-001, the percentage of patient samples containing all SJ of CD38-001 is 100%, while the percentage of patient samples containing only SJ of CD38-001 is 69.2% for MGUS, 58.6% for AMM and 63.7% for MM (see table 3.19). The number of patients having reads spanning any other SJ above the levels of 1% is 3 MGUS, 0 AMM and 1 MM. No patient exceeds the level of 5%.

In supplementary table B.23 the number of reads per splice junction is depicted for two exemplary patients, one with present CD38-005 specific splice junction and one without. The results of the two patients are confirmed by RSEM transcript analysis (see supplementary figures A.18 and A.19). All transcripts are expressed, at least partly. In both examples, the maximum read number of CD38-001 is at least 6 times larger than the maximum read number of CD38-005. At the position of the CD38-005 specific splice junction, nucleotide position 161 and 162, CD38-005 expression shows a gap in both examples. Hence, reads overlapping this splice junction are missing.

3.5.3 Detected mutations

An example for a mutation creating a target in cancer and specifically myeloma cells is the *BRAF* (ENSG00000157764) mutation V600E (dbSNP identifier: rs113488022). Of the 535 MM patients, 11 (2.1%) show presence of this mutation, 10 carry a T instead of an A and one a G. On protein level on the minus strand, this means that valine is replaced by glutamic acid (in case of T) or by alanine (in case of G). The mean number of all reads spanning the position is 5.39 and the mean number of reads supporting the variant is 4.63. The variant allele frequency ranges between 23% and 100%. 6 of the 11 mutations are clonal, with a variant allele frequency $\geq$ 60%. In all MM patients carrying a V600 mutation, *BRAF* is expressed (PA-seq call), and the expression ranges from 2.8 to 3.9 normalised counts. Of the 524 MM patients without *BRAF* mutation at this location, 58 show *BRAF* expression (PA-seq call), and the expression ranges from 0.2 to 5.6 normalised counts. None of the MGUS or AMM patients, and one of the 69 (1.4%) MMR patients carry a *BRAF* mutation.

4 Discussion

This chapter is divided in six parts. First, the main aspects and strategies of the RNA-seq pipeline definition and assembly are discussed. Second, the risk stratifications based on RNA-seq data are critically reviewed. Third, the results of transfer of findings to early stage and relapsed multiple myeloma patients are discussed. These cohorts represent together with the CoMMpass cohort an independent validation of obtained results and strategies, as well as a biological translation to other disease entities, i.e. early and relapsed myeloma. Fourth, target assessment, including expression, splice variant and mutation analysis is critically reviewed. Fifth, the aims of the thesis are discussed. At the end of this chapter, the thesis is concluded and a suggested outlook described.

4.1 Implementation of the RNA-sequencing analysis pipeline

The underlying challenge addressed in this thesis is how to implement RNA-seq in extended clinical routine. This includes the implementation of a sequencing pipeline that delivers clinically meaningful results within an acceptable time. "Acceptable" in this setting is usually considered as about four weeks, i.e. the time for one cycle of induction treatment [99]. The pipeline should present the results comprehensively, so that they can be understood not only by bioinformaticians, but especially by clinicians and patients. For instance, for treatment choice it is necessary to determine "presence" of expression (see also section 4.4.6), e.g. if *NYESO1/2* is expressed on malignant plasma cells and a respective CAR T cell application is potentially successfull.
The RNA-seq analysis pipeline in this thesis lays the basis to enable risk stratification, classification and target assessment for individual multiple myeloma patients in extended clinical routine. It includes three main steps (see section 3.1): First, alignment and read count, second, normalisation, risk stratification and molecular classification, and third, expression, mutation and splice variant analysis of potential targets.
In the following the normalisation, necessary for the sample comparability, the minimal quality requirements, the presence of expression on RNA-seq, the gene translation and the multiple testing correction are discussed.

4.1.1 Normalisation

Sample comparability is mandatory for the application of risk stratifications and molecular classifications, i.e. the categorisation for samples may not change over time, e.g. at different timepoints within a clinical trial. Risk stratifications generally consider the quantitative expression of gene sets above an established threshold. To utilize the same

thresholds for new individual samples, it is necessary to relate the sample to a standard cohort, e.g. by a normalisation. In research practice cohort normalisation is frequently performed, as a specific cohort of patients is analysed at one time, e.g. at inclusion or conclusion of a clinical trial. In this setting, all samples are typically comparable, e.g. the same (cancer) cell type and mode of preparation. Due to the nature of the analysis, inclusion of new samples is usually not foreseen. If a new sample would be included in the respective cohort, the new normalisation can change the classification of previously analysed samples. The more new samples added, the more the initial results are altered. Nonetheless a cohort of patients is necessary to define and validate scores and thresholds. Hence, the pipeline in this thesis uses the same reference group (TG) for each new sample, so that the samples became comparable among each other and previous results are not affected by inclusion of new samples. This strategy was previously used for DNA-microarrays (GEP-R [156]) and was in this thesis for the first time implemented for RNA-seq-based approaches for use in extended clinical routine.
The main step in the pipeline for achieving comparability of samples is the normalisation. It was performed using the method TMM and CPM to account for the two most important technical influences: library size and RNA composition [39].
The library size, ranging from 10 million to 107 million reads in the HD cohort, intrinsically correlates with the number of reads per gene. This can bias results as exemplified in the following seemingly trivial example: Assuming a library size of sample A as twice the library size of sample B, the raw read count for an equally expressed gene is twice as high in sample A compared to sample B [4].
The RNA composition is the amount of reads per gene. This can be extremely different, especially between normal and malignant cell samples, mainly due to the highly varying (and generally lower) synthesis rates of Ig genes by malignant plasma cells. Clinically, this is exemplified in the amount of monoclonal protein in the HD cohort varying between 0 (asecretory myeloma patients) and 106.1 g/l. This can bias RNA-seq results: For instance, 56% of the reads of one MM patient sample map to the *IGHG1* gene (in a patient with IgG-myeloma), which can lead to an underestimation of the expression of less expressed genes [39]. For instance, comparing the first sample (library size 10415974) to a second sample of a similar library size (10421219), which only maps with 5.27% of the reads to *IGHG1*, the raw read count of the exemplary gene *TECPR2* is 91 in both cases. After normalisation, considering the different RNA compositions, the normalised counts are 13.24 for the first patient sample and 5.54 for the second patient sample.
The adjustment for library size and RNA composition is implemented in the normalisation process with edgeR, which enables the comparison between the samples within

a normalisation. Correction for gene length was intendedly not implemented already in the normalisation process. It is accounted for the assessment of presence or absence of gene expression, as discussed in section 4.1.3.

4.1.2 Minimal quality requirements

Quality control was performed in the laboratory and implemented in the bioinformatic processing for DNA-microarrays and RNA-seq samples.

4.1.2.1 Laboratory quality

For broad application of expression analyses by DNA-microarrays or RNA-seq, the limiting factor is the available amount of extracted total RNA, which in turn is depending on the number and purity of purified malignant plasma cells (see section 2.1.1). This is especially low for early stage plasma cell disease, due to the defined underlying maximum infiltration of plasma cells in the bone marrow, i.e. the tumour mass [111]. DNA-microarray analysis prerequisites 25-100 ng of total RNA and is possible in 80% of MM patients [99, 156]. As shown by several institutions including the LfM [42, 56, 99, 156, 197, 219] DNA-microarray analyses can be performed in extended clinical routine in or outside clinical trials, as e.g. the GMMG-MM5 trial (EudraCT 2010-019173-16) [99]. RNA-seq can be performed successfully from a twenty to 100 fold lower input (10 pg [99, 211] to 5 ng [213]) and especially enables the expression analyses for patients with low tumour mass (i.e. MGUS, AMM) [99, 213]. As implemented in this thesis, RNA-seq can be performed in about 90% of previously untreated myeloma patients.

4.1.2.2 Bioinformatic quality control

Using the published quality control criteria of the LfM, 93.6% of performed DNA-microarrays could be used for extended clinical routine diagnostics [156]. Surprisingly, a higher percentage of RNA-seq analyses fulfilled the RNA-seq quality criteria (97% of the 983 RNA-seq files used in this thesis). Eight ($< 1\%$) samples had to be excluded due to too small library size (<10 million [144]). The second most frequent reason for exclusion of files was the number of unmapped reads. In total, 16 files (1.6%) with less than 60% mapped reads were discarded, while the number of mapped reads vary much between different cell types. Dobin and Gingeras [59] expect $\geq 85\%$ of mapping reads. For instance, the non-malignant, polyclonal PPCs and MBCs have $> 87\%$ mapped reads, each, validating the used wet-lab procedures. PPCs have a significantly lower Ig-production rate compared to bone marrow plasma cells [115]. BMPCs, healthy plasma cells, show $\geq 75\%$ mapped reads per sample. For (malignant)

plasma cell samples the higher amount of unmapped reads can be explained by two circumstances:
First, normal and malignant plasma cells are considered to be "Ig factories" and therefore have a very high production rate of Igs, and Ig mRNA. Ig genes undergo a high number of genetic alterations during production, including DNA-rearrangement and somatic hypermutation to ensure antibody diversity (see section 1.1 and 1.2.3). These sequences thereby appear *de novo*, i.e. frequently lack overlap with germ-line DNA.
Second, no trimming or clipping of the reads was performed prior to the alignment, because STAR performs a soft-clipping during the alignment. This also trims the 24 overrepresented sequences, detected by FastQC, which do not map to the genome and contain PCR primer sequences or the Illumina multiplexing index read sequences. The clipping during the alignment increases the number of unmapped reads, because a trimmed read can be "too short" to map. This "too short" is the main reason in all samples for not mapping reads ($\geq$ 88.9% of all not mapping reads per file). Prior to the alignment, the reads have had a uniform sequence length in each sample of 100, 102, 150, 152, 154 or 160 bp. Thus, clipping during the alignment is (most probably) the main reason for "too short" reads, considering that in no case "too many mismatches" was the cause for removal of a read.

4.1.3 Credible presence of expression

The determination if a gene is "present" or "absent" (PA call) is influenced by different technically determined limitations of gene expression assessment by DNA-microarrays. For instance, DNA-microarrays have a narrow dynamic range: The upper limit of the range is given by saturation effects, as for each assessed transcript, a maximum number of mRNA-molecules can bind based on a maximum of binding sites on the chip. The minimum detectable expression is determined by background noise, e.g. non-specific binding and optical noise caused by different labelling [97]. As this background noise can "mask" the presence of few mRNAs of low expressed genes, the PA call on Affymetrix DNA-microarrays is a rather conservative estimate of presence of expression.
In contrast, RNA-seq does not show significant saturation effects *per se*. Hence, RNA-seq theoretically allows a higher sensitivity. However, RNA-seq has also some remaining technical limitations, most important the RNA composition and the library size (see section 4.1.1). Samples were adjusted for both during the normalisation with edgeR. In the following, the threshold for presence of expression for RNA-seq and the gene length estimation are discussed.

4.1.3.1 PA-seq call

In principle, one matching read could be taken as indication for expression in RNA-seq analysis. Indeed, if expression of a given target gene would need to be excluded e.g. because of high expected treatment side effects, this would be appropriate (minimising false negative discovery of expression). Such a "simple" threshold of one raw read would however comprise a variety of pitfalls. First, the threshold will not be identical between patients due to differences in the library size and RNA composition (see section 4.1.1). Likewise, the threshold would be different for different genes within one patient, as it is impacted by gene length and it is more probable to find a read in a longer than in a shorter gene. Hence, presence and absence of expression in RNA-seq data is a question of plausibility and any given threshold definition is thus inherently arbitrary.

Given the aims of the analyses performed in this thesis, a "robust" threshold definition prioritising a low rate of false positive detection was chosen. Therefore, the following strategy was applied. Based on the edgeR user's guide recommendations [39] 5 to 10 raw reads are recommended as biological plausible threshold for presence of expression. In this thesis, additionally observance of gene length was implemented follwing the argumentation depicted above. Assuming that the reads map next to each other, a "plausible" expression is suggested if the mapping reads cover at least half of the gene. This means, a gene with a length of 10000 bp will be called expressed if 100 or more raw reads map to it, in other words, 5 raw reads per 1000 bp are necessary for a "call" of expression.

To conservatively assess expression, in this thesis it was initially assumed, that one normalised count depicts at least 5 raw reads. In practice, it depicts at least 7 raw reads in the TG, 9 raw reads in VG and 7 raw reads in TeG (but none of the PA-seq call did significantly differ between the three groups for the target genes).

Hence, one normalised count per 1000 bp is used in this thesis as necessary to call a gene "expressed". For genes shorter than 1000 bp, at least 1 normalised count is defined as threshold. The definition might be willingly conservative, because first, not all reads are 100 bp, (most of the reads are in fact 150 bp), and second, one normalised count depicts more than five reads in the TG, VG and TeG. Comparing a threshold of 1 raw count to the PA-seq call for the 27 target genes, "present" expression is detected in average in 24.4% more patient samples (of all 535 patient samples) than with the PA-seq call (see table 3.1). *BCMA*, *CD38*, *HM1.24* and *CD74* are expressed in all samples in both cases. It needs to be emphasized that categorising in presence and absence of expression is a categorisation of a continuous biological variable that needs to be re-evaluated for different types of applications.

4.1.3.2 Gene length estimation

To assess counts per bp for presence of expression, the length of each gene has to be calculated. Seemingly trivial, it is not, as multiple transcripts of individual genes have to be considered. To achieve these, three approaches are suggested: First, estimating length by subtracting starting point and end point of a gene on the genome, which will however include all introns. These can widely vary in length and number. Second, merging all exons and excluding overlapping positions, which will result in the maximum length. This length is the upper limit of gene length, as not all exons need to be expressed. Therefore, this measure may frequently deliver longer gene length estimate compared with the longest transcript. Third, using the mean transcript length. This approach in turn may underestimate gene length, and overestimate expression. For instance, *BCMA* has 3 transcripts, a maximum gene length of 2962 bp, a maximum exon length of 1118 bp and a median transcript length of 668 bp. The exon between SJ 3 and SJ 4 is not present in the main transcript BCMA-001 (see figure 3.23), but in the extremely low expressed transcript BCMA-002. Including this exon (of length 124 bp) in the estimation will overestimate the gene length.

For the 24 targets, as expected, the number of patient samples indicated as expressing a gene increases with decreasing gene length (see section 3.1.2), except for the constitutively expressed genes *BCMA*, *CD38*, *HM1.24* and *CD74*. For instance, *HGF* with gene lengths as defined above of 71433 bp, 8897 bp and 1147.5 bp would be indicated as expressed in 40.93%, 82.43% and 94.21% of the patient samples, respectively.

Examples for overestimation of expression due to the PA-seq call are *IGF1R* and *MMSET*. Regarding the maximum gene length, they are expressed in 0% and 10.28% of the 535 patient samples, respectively, whereas the PA-seq call shows presence in 77.01% and 96.26% of the patient samples. Both are expressed in 166 (31.1%) and 245 (45.9%) of the 534 patients using DNA-microarray. This discrepancy results in a low consistency of 52% and 50%, respectively (see table 3.16). *IGF1R* and *MMSET* match to several probesets (5 and 6, respectively) and contain multiple transcripts (17 and 27, respectively). Only 3% of the 60619 genes contain at least this high number of transcripts. Likewise, both show a wide range of transcript lengths, *IGF1R* transcript lengths range from 303 to 11800 bp (median length 572 bp) and *MMSET* transcript lengths from 421 to 8568 bp (median length 1212). Only 2% of the genes have this broad range of transcript lengths. Most probably, the gene length is underestimated at least for *IGF1R*, resulting in a (too) conservative threshold for presence of expression, as none of the alternative probesets has a higher consistency. In case of *MMSET*, it is more likely that the chosen probeset used in the GEP-R is not the best comparison to the RNA-seq ENSG, as 209054_s_at reveals a similar correlation ($r = 0.76$ compared

to $r = 0.78$) and the consistency is extremely high with 96% (for the gene translation see the following section 4.1.4). As the probeset used in the GEP-R was translated to the ENSG and not *vice versa*, the probeset was not changed to keep a consistent approach in terms of strategy.
This high transcript number and length affects 2% to 3% of all genes. With these limitations in mind, the validity of the PA-seq call could be shown in comparison to DNA-microarrays. Comparing for each gene present "P" expression in RNA-seq to marginal "M" and present expression "P" on microarray (see figure 2.4), a very good mean consistency per sample of 84% (range: 70% to 88%) could be found. Regarding the 24 target genes (see section 3.5.1, table 3.16) 12 targets have an excellent consistency of RNA-seq PA-seq and microarray PA call ($\geq 90\%$) and eight have a good consistency ($\geq 75\%$). The four genes with a consistency $< 75\%$ are *IGF1R* and *MMSET*, described above, and *MUC1* and *CD20*, which are both discussed in the following section 4.1.4.

4.1.4 Translation of gene names between DNA-microarrays and RNA-sequencing data

The translation of the gene names between DNA-microarrays and RNA-seq data is a critical and non-trivial point for the transfer of stratifications and classifications, and the expression estimation of targets. In the following, the translation strategies, the translation difficulties and the resulting exclusion of genes part of the DNA-microarray scores are discussed.

4.1.4.1 Translation strategies

Two main strategies were performed: first, for the translation from DNA-microarray probeset to RNA-seq ENSG, and second, for the HDHRS validation on microarrays, i.e. in reversed order from ENSG to gene symbol and to probeset.
Translation from DNA-microarray probeset to RNA-seq ENSG. In this thesis, the available probesets were translated within R, using the hgu133plus2.db package [29]. Subsequently, for each gene the RNA-seq expression was correlated with the microarray expression and the PA-seq call was compared to the PA call on microarray. If the correlation coefficient was $r \leq 0.6$, the translation was controlled in two steps. First, the online GeneAnnot search tool [33, 34, 64] was used, which is based on the GeneCards database [226]. In the database specificity, sensitivity and the gene number per probeset is deposited (see section 2.5). Second, the percentage of samples with present PA-seq and absent PA call was determined (nCO_1, see section 2.2.2). This strategy was performed for the genes of the scores and for the targets used in the GEP-

R.

Translation from RNA-seq ENSG to DNA-microarray probeset. The ENSG of the HDHRS were translated to gene symbols, using the hgu133plus2.db package [29]. A "gene symbol" is frequently (55%) represented on Affymetrix U133 2.0 plus microarrays by more than one probeset, with a median number of probesets per gene being two, and the maximum number being 41. As GeneAnnot often identifies two or more "optimal" probesets it was not used for translation of gene symbols of the HDHRS or the potential targets to probesets. Instead, the package jetset [142] was used to find the best matching probeset for a gene symbol. Jetset is a theoretical tool and determines a score based on specificity, coverage and degradation resistance (see section 2.7.5). Although the probesets had been defined in the respective scores, likewise the jetset package was used to find the best matching probeset for genes part of the risk stratifications. As for translation from probeset to ENSG, the correlation was calculated and the PA-seq and PA call were compared.

4.1.4.2 Non-correlation

As stated above, in this thesis correlation was used as method to confirm the translated genes. Three main reasons for non-correlation were identified: inconclusive databases, homologous sequences and background noise.

Inconclusive databases. For two of the assessed genes, *MUC1* and *CD20*, this is the main mechanism for non-correlation. *MUC1* is represented by three probesets (207847_s_at, 211695_x_at and 213693_s_at). 21369_s_at has the best jetset score and the highest sensitivity (1) and specificity (1) on GeneAnnot, with the latter also being the case for "207847_s_at". In comparison, 211695_x_at has a sensitivity of 0.909 and a specificity of 1. The probeset 21369_s_at with the best correlation was used ($r = 0.63$, compared to $r = 0.59$ (207847_s_at) and $r = 0.20$ (211695_x_at)). However, *MUC1* is the gene with the lowest overlap in PA determination between DNA-microarrays and RNA-seq, with 14% consistency, only. 527 (98.7%) patient samples show present expression on microarray, whereas only 68 (12.7%) on RNA-seq (see also the other gene length estimations in table 3.1, which are even more conservative). 207847_s_at in contrast shows a higher consistency (89%) in PA determination and an almost identical correlation to RNA-seq (see above). To sum up, *MUC1* (ENSG00000185499) does not map to the probeset 213693_s_at, although both are recommended by jetset and GeneAnnot.

A further example is *CD20* represented by the probeset 228599_at, which shows the best jetset score. GeneAnnot in turn indicates that 210356_x_at and 228592_at are optimal choices, too (specificity and sensitivity is 1 for the three probesets).

In correlation analysis, the expressions of all three probesets correlate well with *CD20* (ENSG00000156738, $r = 0.76$ (228599_at), $r = 0.73$ (210356_x_at), $r = 0.78$ (228592_at)).

Homologous sequences. In the RNA-seq analysis pipeline, gene symbols are (mostly) unambiguously matching one ENSG. However, genes with highly homologous sequences frequently do have multiple mapping reads, which are therefore excluded during read count (see section 2.3.2.1) and consequently lead to erroneous low or not expressed genes. This matching difficulty concerns e.g. *NYESO-1/2*, which has a high sequence homology between the three genes coding for *NYESO-1/2* (*CTAG1A*, *CTAG1B*, *CTAG2*). Regarding RNA-seq analysis in this thesis, this leads to reads multiple mapping to all three genes, which are then discarded due to "low quality". On Affymetrix U133 2.0 DNA-microarrays, the two different genes *CTAG1A* and *CTAG1B* are represented by the same three probesets: 210546_x_at, 211674_x_at and 217339_x_at, with the former two also matching to *CTAG2* (GeneAnnot).

Background noise. The correlation as quality factor of the gene translation is limited for very lowly expressed genes, as the (specific) expression of the gene is "hidden" in the background noise on DNA-microarrays and therefore undetectable. Any correlation of RNA-based expression with this random noise is unspecific and the correlation coefficient low and not meaningful. This is exemplified for *WT1*, which is lowly expressed in only 5 ($< 1\%$) patients on DNA-microarrays and 3 ($< 1\%$) patients in RNA-seq. The median expression height is 0.07 for RNA-seq and 2.27 for DNA-microarrays. *WT1* lacks relevant correlation ($r = 0.12$, see table 3.16 in section 3.5.1). A different example is *MAGEA1*, which is expressed in 92 (17%) of the samples for DNA-microarray and 140 (26%) for RNA-seq. It shows a correlation coefficient of $r = 0.80$, but inspecting the correlation plot reveals, that some of the 442 patients with absent expression on DNA-microarrays show present expression on RNA-seq (n=61 (11%)) ranging from 1.45 to 6.40, which is masked in DNA-microarray by the background noise. Correlating only the expression of the patients with present *MAGEA1* expression on both platforms increases the correlation coefficient from $r = 0.80$ to $r = 0.86$.

4.1.4.3 Gene exclusion

Matching, as indicated above, represented a significant challenge. To translate the risk stratifications, it was thus necessary to balance between exclusion of genes with low correlations, and an exact translation of the scores. Gene exclusion would increase the correlation of the scores in DNA-microarrays compared to RNA-seq, as only genes with relevant contribution would be retained. At the same time, exclusion of many

genes would "destroy" the original idea that the specific gene sets of the scores surrogate prognosis. Likewise, the exclusion contradicts the idea of translating this specific on microarrays validated score, as similar as possible to RNA-seq. Hence, a low threshold of $r < 0.15$ was set, related to the first local maximum of correlation, as exclusion of more genes would reduce correlation of the scores on the TG again. (Although there are more possible thresholds, i.e. local maxima, excluding more genes by further increasing the correlation). As discussed above, preferably few genes with ambiguous translations were excluded to retain as much as possible of the initial DNA-microarray or RNA-seq-based risk stratification and classification. Genes were excluded for the UAMS70-seq, the RS-seq and the EMC92-seq. Especially the predictive power of the EMC92-seq has profited from the exclusion as the correlation in the TG increased from $r = 0.67$ to $r = 0.75$ (data not shown).
Regarding the potential targets, "exclusion" is not intended. Hence, always the probeset and ENSG recommended by jetset are used. This is not necessarily the gene with the highest correlation (see *CD20* in section 4.1.4.1) or highest consistency (see *MUC1* in section 4.1.4.1) between both platforms. This reveals that not all analyses based on DNA-microarrays are directly transferable to RNA-seq, as exemplified for *MUC1*: the used GEP-R probeset is not represented by any ENSG on RNA-seq (see section 4.1.4.1). This is a specific problem of the multiple probesets used in microarray analysis and does not affect the proof-of-principle that target assessment is possible on RNA-seq. Taken together, translation from RNA-seq ENSG to DNA-microarray probeset and using correlation as quality measure is well possible but a non-trivial task that needs to be validated for each translation.

4.1.5 Multiple testing correction

In gene expression analyses, where thousands of genes are analysed regarding typically few parameters e.g. survival, significance tests lead to a high number of p-values seemingly significant by chance. For instance, using a significance threshold of 0.05, it is expected that among 100 comparisons 5 are significant by chance, i.e. among 19 tests less than one is expected as significant by chance. Hence, in this thesis a multiple testing correction is used if 20 or more comparisons in the same cohort and analysis step are made.

4.1.5.1 Score generation.

A multiple testing correction was thus performed for the score generation and transfers according to Rème *et al.* [197]. Both steps, the gene selection (HDHRS), as well as the cutoff selection (HDHRS, HDHRS-GEP, RS-seq) were performed multiple times (see

section 3.2.2.2 and 3.3) and each time a p-value was calculated in the TG. Hence, the gene selection is corrected twice for multiple testing, first each p-value is multiplied with the number of samples (according to Rème *et al.* [197]) and second, a BH adjustment is performed. The 53 genes with a corrected p-value smaller than 0.05 were selected.
Likewise, the cutoff selection is controlled by a corrected p-value threshold of 0.05. For all other stratifications (RPI, UMAS70, EMC92, IFM15) no correction was used, following the argumentation detailed above, as 14 significance tests were performed for survival analyses per score, and 14 survival analyses were performed per cohort. To test for possible remaining bias in terms of overfitting, independent validation cohorts were used (see sections 4.2.2 and 4.3).

4.1.5.2 Target analyses

Each comparison performed for the 24 targets was adjusted, including the 24 p-values of the exact Wilcoxon rank-sum test [19] per cohort comparison and the 24 survival analyses. The adjustment was performed separately for RNA-seq and DNA-microarrays, for OS and EFS, and for division of the patients according to maxstat test and PA determination, respectively. The twenty-fifth target *BRAF* was analysed separately.

4.1.6 Success of the implemented pipeline

The RNA-seq-based risk stratifications, molecular classifications and target assessments were successfully determined for 798 consecutive patients. The implemented RNA-seq analysis pipeline was successfully applied and tested in a prospective manner in 604 consecutive patients of the validation and testing groups. As in GEP-R, all stratifications, molecular classifications and target assessments in this thesis can be calculated individually for a new sample. The RNA-seq analysis pipeline was included in extended clinical routine at the LfM and can be performed within four weeks. Prospectively, the presented RNA-seq analysis pipeline can be included in an RNA-seq report, in analogy to the GEP-R. As presented, the bioinformatical pipeline is applicable in extended clinical routine. This development was part of successful conclusion of the BMBF-funded project "CLIOMMICS" (01ZX1309).
The results of the 535 MM patients of the HD cohort are published and were presented at the 61st annual meeting of the American society of hematology 2019 in abstract 1801 by Emde *et al.* [63]. This thesis therefore gives the proof of principle in implementing and applying an RNA-seq pipeline in clinical routine.
A respective publication as full paper is currently in preparation (see also section 4.6).

4.2 Advanced risk stratification and molecular classification using RNA-sequencing

Risk stratification with conventional parameters, e.g. the amount of serum β_2-microglobulin and albumin as in the ISS-stage [84], allows a certain delineation of differences in the survival of patients. With these, risk adapted strategies are in principal already possible [103, 214]. Conventional risk stratifications as ISS-stage lack however the ability to delineate especially high and low risk patients, e.g. on the 535 MM patients in this thesis, the median OS is 128, 81 and 56 months in stage 1, 2, and 3 of the ISS. Using molecular parameters like iFISH allows better delineation, indicated by a broader range of median OS, e.g. in the R-ISS, which shows a median OS of 87 and 41 months in stage II and III, while not reached in stage I. This likewise holds true for the inclusion of gene expression, as exemplified by the RS score for the TeG (see supplementary table B.18), which shows a median OS of 127, 70 and 33 months for low, medium and high risk group. Besides implementation in risk scores, gene expression can directly surrogate biological parameters. For instance, proliferation is assessed by the GPI, based on the expression of 50 proliferation associated genes. Proliferation in turn is associated with survival and the higher the proliferation rate of malignant plasma cells is, the faster is the growth of the tumour mass and the sooner the patient will show signs and symptoms and progress [101].

In conclusion, risk stratification based on gene expression is in principle well suited for personalised decisions regarding intensity, duration and type of a treatment. An example for this has been introduced for the UAMS70 score for patients treated within the total therapy program of the university of Arkansas [236]. Here, for instance, patients with high risk could be treated more aggressively, accepting a higher rate of side effects, but most likely having a more effective treatment leading to a better survival of the patient.

Given the perceived advantages of RNA-seq over DNA-microarrays, the LfM alongside other groups decided to introduce RNA-seq for risk assessment of myeloma patients for the prospective use in clinical practice. The advantages are for instance a wider dynamic range (see section 4.1.3), lower necessary input of purified plasma cells and RNA (see section 4.1.2.1) [99, 211], and possible detection of mutated sequences and splice variants as *a priori* sequence definition is not necessary (see section 4.4).

The general process for the translation of microarray-based risk stratifications and classifications to RNA-seq includes three main steps: translation of the gene names, translation of risk stratifications and classification, and cutoff adjustment. As expected, the calculation *per se* was a straight forward process. Translation of the gene names (see section 4.1.4) and the cutoff adjustment, avoiding overfitting (see section 4.2.3)

required extensive validation. In the following the actual score selection is discussed, followed by the validation criteria and the cutoff adjustment. At the end of this section, the successfully translated RNA-seq-based stratifications and classifications are critically reviewed. The gene translation was already discussed in section 4.1.4.

4.2.1 Score selection

A variety of different microarray-based risk stratifications and molecular classifications has been published over the last decades [22, 43, 56, 101, 124, 197, 219, 254]. Whereas the choice is necessarily arbitrary, expression-based stratifications most frequently used were selected for this thesis. This includes those already implemented in extended clinical routine at the University hospitals of Heidelberg and Montpellier (e.g. in the GEP-R). Seven were successfully translated: a stratification according to the biological variable proliferation GPI [101], the most frequently used risk scores UAMS70 [219], RS [197], EMC92 [124], IFM15 [56], and the molecular classifications TC [22, 43] and MC [254].

4.2.2 Internal validation strategy

To create valid and robust stratifications and classifications on RNA-seq, these were first internally validated (for external validation see section 4.3). For this, all previously untreated MM patient samples were divided in a TG (n=194), an internal VG (n=108) and an external TeG (n=233). Patients per group were randomly selected, except for the last hundred patients analysed, which were added to expand the TeG (see section 2.1.2). This is one potential bias and had to be balanced against the benefit of obtaining a larger TeG. However, the patient groups did not differ regarding standard clinical parameters, e.g. age and monoclonal protein (see supplementary table B.8), or regarding PA-seq call of the targets (see section 4.1.3). To prevent overfitting, the stratifications were fine adjusted on the VG only regarding the proportions of the groups and regarding the survival performance. If the in the following described group size and survival criteria were not fulfilled in the VG, the score has been trained again (by varying the parameters) on the TG and then validated on the VG again, too. The TeG was not involved in stratification training, it is used as a completely independent "hold out" set (see section 4.3).

4.2.2.1 Group size proportions

Similarity is a useful advantage if DNA-microarrays based stratifications should be transfered to RNA-seq, build on existing and established algorithms and strategies. Due to the different technical background of the two platforms it can not be expected

that both analyses lead to exactly the same results. For instance, it is not possible to perfectly fit the proportions of DNA-microarrays in the RNA-seq stratifications and *vice versa* e.g. due to saturation effects (see section 4.1.3). Hence, some changes between low and medium risk as well as between medium and high risk were expected and considered acceptable. It was therefore defined that each group should differ in less than 20 percentage points and the proportions of the stratifications should stay in the same order per score, which was inspected visually. Further, in a clinical view, a group size smaller than 5% is not applicable for risk stratification, as it would affect too few patients. From a statistical point of view, for the RS score prediction a minimum of 2 events per group is recommended [134, 197], whereas e.g. Vittinghoff and McCulloch [240] recommend at least 5 events. As only 108 MM samples are in the VG, and 59 (54.6%) of these have an event for OS, a group size of 10 (9%) would reflect the 5 events. Hence, this percentage was used as minimum group size to ensure clinical and statistical relevance. It is evident that these definitions are to a certain degree arbitrary and approximative and needed to be adapted to the specific experimental requirements, as performed in one possible way in this thesis.

4.2.2.2 Survival performance

As the stratifications should delineate different risk groups, it is necessary to evaluate the survival performance. This was assessed graphically and by evaluating the performance with Brier score and concordance. Regarding the graphical inspection, intersecting curves or curves in aberrant order were excluded, e.g. if the intended "high" risk group performs better in survival compared to the "low" risk group. The analysis of the Brier score for risk stratification creation (TG) was performed using a maximum endpoint of 108 months for EFS and of 120 months for OS. In contrast, the analysis of the validation (VG) and testing groups (TeG, AMM, MMR, CoMMpass) was performed with an endpoint of 72 months for EFS and 108 months for OS, considering only the survival time until 10% of patients remaining at risk, to avoid a high impact of per chance effects due to low patient numbers. The validity of the Kaplan-Meier estimate decreases with decreasing number of patients at risk [32]. Concordance was used as rough indication with a threshold of 0.6, which is the mean concordance calculated for the five stratifications on microarray and the R-ISS in all samples for OS.

4.2.3 Cutoff adjustment

Cutoffs for delineation of groups in translated stratifications and classifications were intended to be as similar as possible to the original microarray-based cutoff calculation in terms of the used method for calculations for comparability. Whether this was pos-

sible or not depended largely on the method used in the original publication: a similar cutoff calculation was possible for UAMS70, by performing a k-means clustering, for EMC92, by fitting the cutoff to the 2 year survival rate and for IFM15, by using the 75% quartile. In contrast, for the RPI a different approach needed to be taken, as the original approach, calculating the cutoffs in dependence of the BMPC and MM expression, resulted in a group size of the high risk group <5% (data not shown). Instead, cutoff estimation was performed by correlating GPI and RPI and transferring the original cutoffs, which resulted in most similar group sizes. This was possible, as GPI and RPI have an excellent correlation coefficient of 0.89 in the TG.
Likewise, the cutoff calculation of the RS-seq needed to differ from the original publication. Using the "automatic" best-cutoff-search-method of Rème *et al.* [197], the risk of overfitting is quite high (see section 2.1.2). Overfitting lead in this case to cutoffs so highly adjusted to the TG, that they were not transferable to any new data sets, as the VG: 7 cutoffs excellently delineated three groups on the TG, respectively, but none of these delineated significant groups on the VG. Exactly for this the TG-VG-TeG strategy was used (see section 4.2.2): Although the validation has failed, it was possible to train and validate the stratification again, while there was still the independent TeG. Finally, as for GPI, the group proportions (independent of the specific expression pattern) being similar to the microarray-based proportions were introduced as selection criterion. For this, the smallest high risk group, using the highest calculated high cutoff (*hcut=29.37*) and the largest low risk group, using the highest calculated low risk cutoff (*lcut=9.01*), were used.
The implicit assumption for RPI and RS-seq cutoff selection was, that, as the identical patient cohort was used for microarray and RNA-seq analysis, likewise the risk group proportions of the stratifications should be comparable.

4.2.4 Successfully translated RNA-sequencing-based stratifications and classifications

In this subsection, the success of translation of each stratification and classification addressed in this thesis is discussed, based on the results obtained on the TeG. A successful translation was defined by four criteria: First, each class should consist of at least 9% of the samples, second each class differs in less than 20 percentage points in comparison to the DNA-microarray scores, third, the log-rank test of the survival analysis should be significant ($p < 0.05$) and fourth, the survival curves should be in the right order and not intersecting, in an time interval from 24 months to the last but one event per curve. These criteria are the same as for the validation (see section 4.2.2), without considering Brier score and concordance.

4.2.4.1 Risk stratifications

According to these criteria the stratifications based on proliferation (RPI) and risk (UAMS70-seq, RS-seq, EMC92-seq, and IFM15-seq) were successfully translated and implemented, as the criteria are achieved. The continuous scores are highly prognostic, as their hazards significantly increases over time. The translated microarray-based stratifications as well as *de novo* generated HDHRS significantly delineate groups of patients with different EFS and OS.

RPI. The proliferation assessing RPI shows the second best concordance of 0.60 for OS and of 0.57 for EFS, along with UAMS70-seq and IFM15-seq. The RPI correctly depicts the proliferation level of malignant plasma cells within the ranges of MBC- and BMPC-like (non-proliferating low risk group below the cutoff of 121.96) and PPC- and HMCL-like proliferation (proliferating high risk group above the cutoff of 202.74). The three groups are significantly associated with different survival. Patients with malignant plasma cells showing a low proliferation index have a median EFS of 38 months and a very long median OS of over 103 months. In contrast, patients with high RPI have a median EFS of 25 months and median OS of 55 months.

UAMS70-seq. The UAMS70-seq and the UAMS70 score show the lowest correlation of all transferred scores, which is nevertheless good ($r = 0.77$). Both have a concordance of 0.57 for EFS and of 0.59 and 0.60 for OS, respectively. Hence, the survival performance is comparable. Likewise, there is a high consistency of group proportions of 90.2%. The high risk group shows a median EFS of 17 months and a median OS of 47 months, compared to the low risk group of 36 and 103 months.

RS-seq. The RS-seq perfectly divides the MMs into three groups with significantly different EFS and OS. The median EFS of low, medium and high risk group is 40, 32 and 17 months and the median OS is 130, 83 and 37 months, respectively. The RS-seq is the only score with concordance lower than the one on microarray minus the standard error of the concordance, which is the case for EFS and OS. Although the correlation between RS and RS-seq is good ($r = 0.82$), the RS-seq has the smallest consistency (81.6%) between DNA-microarray and RNA-seq score estimation. It is likewise a risk score delineating patients in three groups which inherently implies a higher probability of non-concordant classifications. Other than the three group delineating GPI with proportion threshold setting, implementing the best threshold for RS was more difficult as several thresholds lead to the same (good) separation (see above, section 4.2.3). This is therefore an intrinsic property of categorisation of continuous variables and does not hamper the clinical application.

EMC92-seq. The EMC92-seq is the score with the lowest concordance for EFS (0.56, along with RS-seq) and OS (0.58, along with RS-seq and IFM15-seq). This

concordance is better than on microarray for EFS (0.55) and OS (0.57), hence the survival performance is comparable. The consistency of the group proportions is high, with 92.2%. The EMC92-seq significantly delineates two groups, showing a median EFS of 35 and 14 months and a median OS of 103 and 30 months.

IFM15-seq. As the EMC92-seq, IFM15-seq has a low concordance for EFS (0.57) and OS (0.58), but it is still better than the concordance of the IFM15 for EFS (0.54) and OS (0.57). The stratifications of both platforms show a consistency of 85%. The two groups of the IFM15-seq show a median EFS of 36 and 21 months and a median OS of 103 and 66 months.

HDHRS. The *de novo* generated risk stratification based on the algorithms of Rème *et al.* [197] significantly delineates three groups. The median EFS of the low, medium and high risk group is 40, 32 and 18 months and the median OS is 148, 82, 37 months. This is the widest range of all median OS values. The HDHRS is the stratification with the best concordance. Its concordance is similar to the one of the current standards, the R-ISS, for EFS (0.62 *versus* 0.6) and for OS (0.66 *versus* 0.67). The HDHRS-GEP performs similar well, with a concordance of 0.61 for EFS and 0.66 for OS. The Brier score of the HDHRS is better than the ISS Brier score for EFS (0.1534 *versus* 0.1577) and OS (0.1601 *versus* 0.1625). But it is higher than the Brier score of the R-ISS for OS (0.1601 *versus* 0.1550). Summarising the results of concordance and Brier score, the HDHRS performs as well as the R-ISS. The multivariate Cox regression shows, that both HDHRS and R-ISS are independent predictive and thus convey orthogonal information.

4.2.4.2 Molecular classifications

Two expression-based classifications regarding molecular subtypes (TC [22, 43] and MC [254]) and the t(4;14) prediction were successfully implemented on RNA-seq.

TC-seq. The consistency of TC-seq and TC classification is excellent with 81% in the TeG. The largest difference, with 6%, is between the group "11q13" and "D1", while all other differences in classification are $\leq 3\%$. This is expectable, as the "11q13" and "D1" group are the largest groups with 24% and 34% of all MM patients. With 1% of patients, the "6p21" group is the smallest group.

MC-seq. Likewise, the consistency of the MC-seq and the MC classification is excellent with 86.6% in the TeG. The most consistent groups are "MF" and "MS", which differ in <1%, each. The differences in the other groups are similar without a single outlying group.

t(4;14)-seq. The t(4;14) classification shows 99% consistency in comparison to iFISH (99.6%) and DNA-microarray prediction (99.2%). This was expected, based

on the good predictability with DNA-microarray expression data, shown for different datasets [156, 169].

4.2.4.3 Success of translation

In summary, risk stratification and classifications can be translated from DNA-microarray to RNA-seq and, as presented in this thesis, this translation can be validated. In the TeG, VG and TG, RNA-seq-based and the *de novo* derived score performs at least as well as the DNA-microarray-based stratifications. The concordance ranges from 0.55 to 0.62 for EFS and 0.57 to 0.66 for OS, which is typical for survival data [196, 208, 228, 229].

The similarity of risk assessment by RNA-seq and microarray is further substantiated by the high correlation (r ranges from 0.77 to 0.85), and the high consistency (from 70% to 92%) of RNA-seq- and microarray-based stratifications. Likewise, the classifications show a high consistency, ranging from 81% to 99%.

One important aspect to mention regarding the categorisation of continuous variables into groups is that they are to a certain degree inherently arbitrary. For instance, a myeloma patient sample classified as "medium" or "high" risk, based on a score close to the threshold will not have a biologically different disease or survival. Hence, slightly different thresholds will likewise be prognostic. This limitation holds true for both RNA-seq- and microarray-based methods individually, as well as for the translation from one strategy to the other. It does, however, not hamper the principal applicability of RNA-seq risk stratifications and classifications in clinical routine. Hence, the RPI, UAMS70-seq, RS-seq, EMC92-seq, IFM15-seq, HDHDRS as well as the TC, MC, and t(4;14)-seq were included in the RNA-seq analysis pipeline (see figure 3.1).

4.3 External testing on earlier stage, relapsed and independent symptomatic myeloma patient cohorts

Two main strategies were applied to create robust stratifications and classifications on RNA-seq data, and subsequently validate them: internal and external validation. The division of the HD cohort into TG, VG and TeG (see section 4.2.2) allowed adjusting initial group definitions using the VG, but retaining the possibility to use an internal "hold out" set (TeG) for independent testing and validation of the score performance. This strategy excludes the possibility of overfitting, i.e. the over-adjustment of scores to the underlying patient cohort. As exemplified in section 4.2.3 for the RS-seq, this can indeed impose a problem. The internal validation strategy thus focuses on the statistical and technical aspects of the translation.

The external validation was performed on three independent external cohorts: first,

on early stages, comprising asymptomatic myeloma patients (AMM), second, on relapsed myeloma (MMR), both from the LfM Heidelberg, and third on an external symptomatic myeloma patient cohort (CoMMpass). The external validation focuses on additional information about the biological and pathophysiological transferability and, more statistically, to what extent scores are still prognostic in a setting of altered proportions regarding group constitution. Further, testing on the CoMMpass cohort reveals, how much the stratifications are transferable between RNA-seq data sets of different research groups.

4.3.1 Early stages

The validation on AMM assess whether the risk stratifications delineate risk only in a treatment setting, or the assessed risk is to a certain extent independent of the treatment and thus describes an intrinsic property of malignant plasma cells.
The stratifications were tested on 142 AMM patient samples. The high risk groups of the three-group stratifications (RPI, RS-seq, and HDHRS) were, as expected, small (<6%, each, see section 3.4.1). Hence, the medium and the high risk group needed to be merged. Three of the six stratifications (RPI, UAMS70-seq, and RS-seq) significantly delineated two groups, while the EMC92, the IFM15 and the HDHRS are not predictive in AMM. This shows, that expression of the selected genes for the RPI, UAMS70-seq, and RS-seq includes intrinsic factors of malignant plasma cells, which are already present at AMM stage.

4.3.2 Relapsed myeloma

Testing on MMR allows biologically to assess whether scores remain prognostic for later stage treatment, because MMR samples are, as MM samples, treated, but MMR treatment did include different treatment regimen compared to upfront induction treatment, high-dose therapy and autologous stem cell transplantation. Of course, in assessment of OS for previously untreated MM patients as in TG, VG, TeG and CoMMpass, these relapse treatment regimen are included. Testing on MMR likewise allows testing to what extent distributions of scores and proportions of high and low risk shift to higher values for MMR patients.
The stratifications were tested on 69 relapsed MMR patients. Due to the low amount of low risk patients ($n \leq 8$ patients) for the three-group stratifications (RPI, RS-seq, and HDHRS), low and medium risk group were merged for the survival analysis. All six stratifications significantly delineate two groups. The successful performance of the risk stratifications implies that they are, to a certain level, independent of the actual treatment.

4.3.3 CoMMpass

With the TG-VG-TeG three-group-validation strategy (see section 4.2.2), it was possible to show the transferability and validity of risk stratifications and classifications focusing on statistical and technical aspects of the translation. The analysis of the CoMMpass cohort of the MMRF allows to assess to what extent the stratifications are transferable between RNA-seq data sets of different research groups. More precisely, the external CoMMpass cohort comprises the validation in the setting of a different RNA-seq protocol in different institutions and different applied treatment strategies. The latter is especially relevant if a scoring system, established by one group, is intended to be used in another group, e.g. within translational clinical trials or as part of a new IMWG-recommendation. This would require highly standardised methods, or methods, which are robust enough to convey prognostic information within a range of certain experimental validation (see also section 4.5).

4.3.3.1 Risk stratifications

The stratifications for patients of the CoMMpass cohort were performed as described for the HD cohorts. From the results (see section 3.4.2) four main conclusions can be drawn:

i) The continuous scores are highly prognostic in MM, determined by a significant increase of their hazards over time for OS (log-rank test $p < 0.001$ in all cases).

ii) The risk stratifications delineate groups of patients with significantly different OS and EFS. Survival analyses for all scores are highly significant for EFS and OS ($p \leq 0.001$, each). The Brier scores for EFS of the EMC92-seq and the HDHRS are larger for the CoMMpass cohort (0.1769 and 0.1804), than for the TeG (0.1554 and 0.1534), but comparable to the ones of the ISS (0.1794) and R-ISS (0.1803) on the CoMMpass cohort. In contrast, for OS, the HDHRS Brier score of the CoMMpass cohort (0.1376) is smaller than the one of the TeG (0.1601), but as for EFS, the Brier score is comparable to the one of ISS (0.1335) and R-ISS (0.1378). Regarding the concordance, for EFS on CoMMpass cohort, the RPI (0.59), RS-seq (0.59) and EMC92-seq (0.57) perform better, compared to TeG (0.57, 0.56, 0.56, respectively). For OS, the RPI (0.63 *versus* 0.60), UAMS70-seq (0.63 *versus* 0.59), RS-seq (0.62 *versus* 0.58), EMC92-seq (0.63 *versus* 0.58), and IFM15-seq (0.59 *versus* 0.58) performed better.
Hence, regarding Brier score and concordance, the stratifications perform as well on the CoMMpass cohort, as on the TeG.

iii) The two-group risk stratifications are successfully transferred. The criteria for "success" (see section 4.2.4) were analysed for the CoMMpass cohort. The group size criterion of 9% of patients in each group was fulfilled. The comparison to DNA-

microarrays could not be performed as none were available for the CoMMpass cohort. As stated above, the log-rank test of the survival analyses was highly significant and the curves are in the right order and not intersecting.
iv) The three-group risk stratifications are statistically significant, regarding the log-rank test of the survival analyses. As the two-group risk stratifications, the group-size criterion of 9% of patients in each group was likewise fulfilled for the three-group risk stratifications. Survival curves of low and medium risk intersect for EFS and OS for the RS-seq after 36 months. Although not intersecting, for the RPI the low and medium risk curves are closely together. The HDHRS separates low and medium risk group slightly better, but not comparably well as in the TeG. Hence, the three-group risk stratifications significantly differentiate two patient groups regarding survival.
In the next section, it is discussed to what extend these discrepancies can be related to differences in the HD *versus* CoMMpass cohort

4.3.3.2 Differences between HD and CoMMpass cohort

First of all, it should be noted that the transferability of established scores on the CoMMpass cohort is quite good in general. With the depicted internal three-group validation and testing strategy overfitting as reason for the observed differences could be excluded. More probably the differences are related to three main aspects: differences between patient cohorts regarding age and treatment regimen, differences in experimental laboratory procedures and differences in bioinformatical pipelines.
Patient cohort associated differences. The CoMMpass and the HD cohort differ in composition of patients (geographically and age) and especially the applied treatment regimen. The CoMMpass cohort includes patients from Canada, Italy, Spain and the United States, whereas HD patients are mainly from Germany. Different ethnic background has previously been shown to impact on survival rates [3, 248]. The inclusion criteria for MM patients are the same. Likely most important, treatment regimen differ between the two patient cohorts. Most striking differences are the use of upfront high-dose therapy and autologous stem cell transplantation: Nearly all patients of the HD cohort (99%) received stem cell transplantation, compared to 50% of patients of the CoMMpass cohort. Further, 73% of the CoMMpass patients received (non-thalidomide) IMiD-based treatments (see section 3.4.2), which was not used in the HD cohort. The remaining patients of the CoMMpass cohort (27%) received a PI-based treatment (bortezomib or carfilzomib), which likewise received 66% of the patients in the HD cohort, but rarely in the same drug combinations. The latter is in part explainable as the first sample of the HD cohort was already included in January 2002, while CoMMpass cohort is quite recent and inclusion started in 2012. Further,

the preferred treatment regime generally differs within Europe and between Europe and the US (especially in terms of use of upfront high-dose therapy). A further significant difference is the median age in both cohorts with patients in the CoMMpass cohort being significantly older. Age likewise has an impact on the survival [3, 26, 44] and, not less importantly, impacts the given treatment regimen. Independent of these factors, the shorter follow up in the CoMMpass cohort implies fewer events and an overrepresentation of rather early events. This is likely associated with the performance of the differentiation of low and medium risk for the RPI, RS-seq and HDHRS. Here, a separation is frequently found later during the course of observation. Despite these differences, the percentage of patients without progress within the first year is 80% *versus* 87% in CoMMpass *versus* HD cohort and OS rate of 92% *versus* 95%. The rates are after three years for both cohorts are 45% for EFS, and 76% and 78% for OS, respectively.

In general, underlying patient cohorts are widely comparable, except for more patients being treated upfront with "novel agents" in the CoMMpass cohort (see figure 3.17g in section 3.4.2). From a technical point of view, this leads to a conservative estimation of score performance, as even better performance would be expected if exactly the same proportion of treatment regimen would have been used. Of course, the score performance on a cohort of patients receiving current treatment regimen is also clinically more relevant.

Experimental laboratory differences. Different experimental laboratory procedures and especially library preparation influence the sequencing results and therefore their comparability [221], also representing the experience of the LfM (data not shown). Library preparation was performed for CoMMpass with the Illumina TruSeq RNA library kit v2 [51, 169], while LfM used the SMARTer Ultra Low RNA Kit (Illumina) in combination with the NEBNext Chip-Seq Library Prep protocol (New England BioLabs) [211]. Further, in CoMMpass 150 to 500 ng RNA are used as input [51, 169], in the HD cohort 5 ng (minimum 10 pg, see also section 2.1.1) [211]. The difference in starting material itself can lead to a selection bias, as a higher initial tumour mass is necessary to reach the required amount in the CoMMpass cohort. For sequencing of the CoMMpass cohort, mainly 2x83bp reads were used, whereas the LfM used 2x50bp or 2x75bp reads. Sequencing was performed for both groups on Illumina HiSeq2000 instruments with unstranded library preparation.

Experimental bioinformatic differences. A different reference genome version was used for the alignment. CoMMpass still used the GRCh37 and HD used the (newer) GRCh38. The ENSGs are almost completely identical, although in GRCh37 3000 fewer genes are referenced compared with GRCh38. Underlying transcripts show sub-

stantial deviations: From the 57905 genes in GRCh37, 16885 have a different mean transcript length compared to GRCh38. This biases the number of matching reads and influences the risk stratification. It was not possible to align the CoMMpass cohort on GRCh38, as the FASTQ files are not accessible for download. In turn, it was not considered appropriate to align the HD cohort against the obsolete reference genome GRCh37, as all subsequent analyses and upward comparability would have been hampered. The different reference genome also necessitated the separate normalisation of the CoMMpass cohort. Hence, as the raw FASTQ files could not be obtained, the gene wise adjustment as performed was chosen as best solution (see section 2.7.4). For this, a modified Z-score standardisation was performed, following the normalisation, using a subgroup of MM patients of the HD cohort (see section 2.7.4) as reference cohort.

4.3.4 Inter-group transferability and reproducibility

The differences in the experimental design and their impact on the actual risk stratification pose a caveat for the ability to transfer risk stratification between different groups. On the one hand, the general ability to risk stratify holds true. On the other, it is likely that a patient presenting at different sites (e.g. HD and in any CoMMpass centre) would be stratified differently. To achieve inter-group transferability, a "standard" pipeline would need to be introduced to reduce technical variation in experimental laboratory and bioinformatical analysis. One necessary step for this is to provide a detailed description of the used pipeline, to make it comparable to others and reproducible. For instance, the RNA-seq experimental laboratory pipeline used in this thesis was previously published in detail by Seckinger *et al.* [211], and this thesis represents the detailed description of the bioinformatical analysis. The results of the pipeline are published and presented at the 61st annual meeting of the American society of hematology 2019 [63] (see section 4.1.6).
Summarising, the introduced stratifications are very well transferable to patient samples which are processed in the same way as the samples used for training of the stratifications. Technical variation, caused by differing laboratory pipelines, and variation in bioinformatical processing influences but does not hamper the use of risk stratification systems and classifications: these give meaningful and significant results. The degree of transferability between the groups and patient cohorts is thus the more remarkable.

4.4 RNA-seq-based target assessment

The main aim of this thesis is the implementation of RNA-seq in translational myeloma research and extended clinical routine application, in terms of risk assessment (see

above, section 4.2) and target analysis.
The latter was assessed for 25 exemplary targets, depicted in the following section 4.4.1. Subsequently, expression assessment, splice variant analysis and the detection of actionable mutations are discussed. The latter two analyses are not feasible on the routinely used Affymetrix U133 2.0 DNA-microarrays, due to the *a priori* designed sequences (probes) not including e.g. mutated transcripts.
Target analysis was exemplified for a consecutive large patient cohort for the assessment of CD38 expression [213] (during clinical trial design, see below in section 4.4.2), BCMA (during compound development of CC-93269 [212]), and CTAs as vaccination targets for assessment of a potential vaccination trial [209]. For all three publications, data obtained in this thesis have been essential and are discussed within the following sections.

4.4.1 Target selection

The 25 exemplary targets assessed within this thesis can be grouped in different not necessarily disjunct categories (see table 3.14 and 3.16). In this section, three groups are presented, actionable, vaccination, and theoretical targets.

4.4.1.1 Actionable targets

In principle, 14 addressed targets are currently actionable (see targets marked with a $^{+}$ in table 3.14). BCMA, CD38, CD74, NYESO1/2, GPRC5D, and CS1 represent surface antigens for which compounds (monoclonal antibodies, TCB) or CAR T cell treatment options are either approved[9] (CD38, daratumumab; CS1, elotuzumab) or in clinical trials (an overview of clinical trials is depicted in table 3.15). The completed trial of NYESO1/2 treatment revealed promising results [224]. *NKG2D* in principle also falls in this category, but is almost not expressed in the 534 MM patient samples on RNA-seq (0%) and DNA-microarrays (<1%). A further actionable target is the mutated gene *BRAF* (V600E mutation).
FGFR3, IGF1R, AURKA, and CD20 are special cases in this category. For these targets, compounds have been tested in clinical trials and are in principle available, but results have until now not been clinically encouraging:
FGFR3 has been suggested over 15 years ago as target for patients harbouring a t(4;14), which, in 70% of cases, also express FGFR3 [232]. Tested compounds are e.g. dovitinib (tested in multiple myeloma) [207], masitinib (ClinicalTrials.gov iden-

[9]National Cancer Institute: Drugs Approved for Multiple Myeloma and Other Plasma Cell Neoplasms; Online resource: `https://www.cancer.gov/about-cancer/treatment/drugs/multiple-myeloma`; Status: 27.02.2020, 12:55

tifier: NCT00866138) and edrafitinib (ongoing clinical trial, ClinicalTrials.gov identifier: NCT02952573). Edrafitinib has previously shown activity in solid tumours [13].
IGF1R is an important growth and survival factor in multiple myeloma, shown by the LfM in collaboration with the "centre hospitalier universitaire" Montpellier [223] and others. It has been suggested for personalised treatment approaches [96] and the monoclonal antibody AVE-1642, binding IGF1R, has been tested within clinical phase I trials in refractory multiple myeloma patients in 2007 [163] and 2011 [162]. In the latter trial, AVE-1642 was tested in combination with bortezomib, but the response rates were insufficient [162]. Moreau *et al.* [162] assumed, that the response rates could have been improved by determining systematically the expression of IGF1R prior to the treatment (see also section 4.4.6).
AURKA inhibitors have been suggested by the LfM as potential treatment option in multiple myeloma [102]. Clinically tested compounds include VX680, which failed due to induction of a long QT-syndrome [71], and Alisertib (MLN8237), not showing encouraging remission rates [202].
CD20 treatment (rituximab) has not been clinically effective (ClinicalTrials.gov identifier: NCT00003554), but CD20 CAR T cells are currently in a phase 1 clinical trial for patients with refractory B cell lymphoma or chronic lymphocytic leukaemia (ClinicalTrials.gov identifier: NCT04007029).
The group of actionable targets likewise includes multiple myeloma vaccines, e.g. for MAGEA3 and MUC1, described below.

4.4.1.2 Vaccination targets

Multiple myeloma vaccines are designed to trigger an antibody response against e.g. a surface proteins of malignant plasma cells i.e. to induce a myeloma specific immunity [209]. This is currently tested, using a recombinant MAGEA3 protein in a combination treatment, which induces a MAGEA3 specific antibody response in MM (ClinicalTrials.gov identifier: NCT01380145) [47]. The results are promising and future clinical trials are recommended [47]. Further, a MUC1 vaccine was tested in a clinical trial (ClinicalTrials.gov identifier: NCT01232712), in a small myeloma patient cohort (n=15), but with encouraging results [30].
Additional potential vaccination targets are malignant plasma cell surface proteins as NYESO1/2, MAGEA1, RHAMM, CS1, WT1, and SSX2.

4.4.1.3 Theoretical targets

MMSET, TP53, CCND1, CCND2 and CCND3 have been selected as potential targets, because they are aberrantly or differentially expressed [156] and involved in myeloma

pathogenesis [22, 37, 146, 214, 254]. They are potentially targetable by small molecule inhibitors (e.g. CDK4/6 inhibitors [107, 172]) or siRNA-based strategies. The latter strategy is currently tested in solid tumours [112], e.g. using the nanoparticle CALAA-01 (ClinicalTrials.gov identifier: NCT00689065) [54]. CALAA-01 consists of a "nanocarrier", which contains RRM2 specific siRNAs [112]. The nanoparticle binds to transferrin receptors on the cancer cell surface and delivers the siRNAs into the cell [54]. The siRNAs reduce the expression of the anti-cancer target RRM2 [54]. However, it has not been tested clinically in myeloma up to now. In the context of personalised treatment, these targets therefore remain theoretical. In the context of this thesis, they have mainly been maintained due to theoretical interest in a different category of targets and their expression. From a technical point of view, they were included to analyse the transferability of DNA-microarray-based assessment to RNA-seq on a broader set of target genes.

4.4.2 Target expression

The main benefit of analysing gene expression are the personalised treatment approaches for targets with available inhibitor or immunotherapeutical agent, to assess whether it is expressed and therefore treatable in a specific patient. This expression assessment is exemplified in the following for *CD38* [213], *BCMA* [212]), and CTAs as vaccination targets [209].

4.4.2.1 BCMA

The target expression analysis of the potential therapeutic target BCMA was assessed and published by Seckinger *et al.* [212]. In this project the IgG-based BCMA-TCB antibody was constructed, which links BCMA on myeloma cells with CD3 on T cells, leading to immune response and elimination of the myeloma cells. During the initial phase of the project, one challenge was to assess in which percentage of the myeloma patients *BCMA* is expressed to evaluate the potential patient population for a respective treatment option. Literature data at that time were only available for myeloma cell lines and small patient cohorts [212]. The analyses of the expression of *BCMA*, for 778 DNA-microarray samples and 263 RNA-seq samples, was performed within the framework of this thesis. As depicted above, *BCMA* is expressed in all malignant plasma cell samples, including those of patients with especially adverse prognosis. Based on the results of this thesis, BCMA represents an universal target in multiple myeloma treatment. For instance, the developed BCMA-TCB CC-93269 [212] is currently in clinical testing (ClinicalTrials.gov identifier: NCT03486067) [50]. The compound shows promising initial results in terms of response rates for MMR patients of

83% [50] and 89% (interview at the annual meeting of the American society of hematology 2019[10], published online 15 January 2020) at a target dose of >6 mg. However, *BCMA* expression varies significantly between myeloma patients (see figure 3.22) and could potentially interrelate with the response. The assessment of *BCMA* expression was therefore included in the RNA-seq analysis pipeline (see figure 3.1).
Furthermore, it could be shown in this thesis that *BCMA* is expressed in BMPC, but not in earlier stages such as MBCs, and in naïve B cells, pro- and pre-B cells shown by Seckinger *et al.* [212]. Hence, anti-BCMA treatment would lead to the elimination of normal plasma cells and immune suppression. But this would be reversible, which is in agreement with clinical trial results (ClinicalTrials.gov identifier: NCT03486067) reported at the 61st annual meeting of the American society of hematology 2019 [50].

4.4.2.2 CD38

As detailed above, *CD38* expression was assessed during clinical trial design for GMMG-MM5 (EudraCT 2010-019173-16) [150, 157], HD6 (ClinicalTrials.gov identifier: NCT02495922) and GMMG CONCEPT trial (ClinicalTrials.gov identifier: NCT03104842) [213]. CD38 is a normal and malignant plasma cell surface protein, which can be targeted by CD38-antibodies as daratumumab approved in Europe[11] and the USA[12] or in late stage clinical development as isatuximab [57] approved in the USA[12]. It was previously known that *CD38* is a frequently expressed target in malignant plasma cell diseases. However, only about one thirds of patients respond to anti-CD38 treatment as monotherapy [147, 148, 234]. This prompted the LfM to investigate *CD38* expression in a large cohort of patients as part of this thesis. The results were previously published by Seckinger *et al.* [213] for 62 MGUS, 259 AMM, 764 MM, 90 MMR patients with DNA-microarray data and for 52 MGUS, 29 AMM and 388 MM with RNA-seq data. The analysis revealed that *CD38* is constitutively expressed. The expression height varies in expression from 7.05 to 15.42 normalised counts with a median of 12.48 and a standard deviation of 1.04. High *CD38* expression at a level of normal bone marrow plasma cells is (surprisingly) associated with a shorter time to progression for AMM and a good prognosis for MM.
Seckinger *et al.* [213] concluded that neither lack of expression, nor alternative splicing

[10] ASH Clinical News: Early-Phase Trial Suggests Bispecific Antibody CC-93269 Has Activity in Relapsed/Refractory Multiple Myeloma; Online resource: `https://www.ashclinicalnews.org/on-location/ash-annual-meeting/early-phase-trial-suggests-cc-93269-activity-relapsed-refractory-multiple-myeloma/`; Status: 27.02.2020, 19:57

[11] European Medicines Agency, Science Medicines Health: Darzalex; Online resource: `https://www.ema.europa.eu/en/medicines/human/EPAR/darzalex`; Status: 22.04.2020, 11:54

[12] National Cancer Institute: Drugs Approved for Multiple Myeloma and Other Plasma Cell Neoplasms; Online resource: `https://www.cancer.gov/about-cancer/treatment/drugs/multiple-myeloma`; Status: 22.04.2020, 11:56

(see below section 4.4.4.2) are a reason for upfront resistance. Hence, patients should not be excluded from anti-CD38 treatment based on *CD38* expression. However, due to the variation of *CD38* expression, the analysis is included in the standard diagnostic of patients at the LfM. This diagnostic is used for prospective analysis of patients, included in clinical trials, which test anti-CD38 combination treatment. For analysis of alternative splicing as potential mechanism of resistance, see section 4.4.4.2.

4.4.2.3 Vaccination targets

The third investigated strategy regarding expression of potential targets is the assessment of CTAs. These antigens are selectively expressed on cancer cells and immune privileged regions, as e.g. testis. Therefore, it is envisioned that normal tissue would not be affected by CTA treatment. From a therapeutical point of view, CTAs are seen as ideal targets for vaccination strategies to trigger a (prophylactic) myeloma-specific immunity [209].

The expression of potential vaccination targets, e.g. *HM1.24* and CTAs (*NYESO1/2*, *MAGEA3*, *RHAMM*, *WT1*) and their association with survival have been analysed, as part of this thesis, based on the previous work of the LfM in collaboration with the "centre hospitalier universitaire" Montpellier [48]. The results were published by Schmitt *et al.* [209]. Of 458 DNA-microarray and 152 RNA-seq MM patients, published in the paper, all express *HM1.24*, 318 and 144 *RHAMM*, 209 and 77 *MAGEA3*, 40 and 20 *NYESO1/2* and 4 and 5 *WT1*. As none of the CTAs is expressed in all patients, the question raised, how many CTAs would be necessary in a vaccine, to cover at least one expressed CTA in all patients. Of the 458 DNA-microarray patient samples, 368 (80%) express *RHAMM*, *MAGEA3* or *NYESO1/2*. Hence, the clinical suggestion for therapeutic strategies is the recommendation to use "cocktails" of different vaccination peptides together covering at least one target in all myeloma patients [48, 209].

4.4.3 Survival association of target expression

Expression of targets can be associated with survival. These targets could then, in principle, be used for personalised and risk adapted treatment [96]. Association with adverse survival was found for high expression of *FGFR3*, *MAGEA1*, *MAGEA3*, *CCND2*, *MMSET*, *IGF1R*, *AURKA*, *RHAMM* and *MUC1*. In contrast, high expression of *CD38*, *CD74* and *CCND1* is associated with longer survival. The direction of survival association, adverse or not, did not differ between the platforms. The one exception is *CS1* expression, which is associated with survival on RNA-seq but not on DNA-microarrays. Survival association was always detected, if present, using the maxstat test. Survival association was not detected with the PA-seq call for the genes *CCND2*,

MMSET, *IGF1R* and *MUC10*, and for the constitutively expressed genes *CD38*, *CD74* and *CS1*. In case of *MMSET* and *MUC1*, the PA-seq call is most probably limited by the ratio of patients with absent and present expression (93.3% and 12.7% present, respectively). For *MUC1* likewise survival association was not detected using DNA-microarrays (98.7% present).
For *WT1* and *SSX2*, both weekly associated with survival, microarray- and RNA-seq-based analyses lead to contradictory results. *SSX2* is associated with survival on DNA-microarray unlike using RNA-seq analysis where the expression is extremely low, ranging from 0 to 0.25 normalised read counts (see supplementary figure A.15). *WT1* expression is associated with survival for EFS maxstat test and OS PA call on DNA-microarrays. For *CD38*, an association with survival in MM in the HD cohort had previously been published on DNA-microarrays in a large cohort of 764 patients [213] with a p-value of 0.03 for EFS and a p-value of 0.02 for OS, using the maxstat test. In the smaller cohort used in this thesis, normalised with a differing method, *CD38* marginally fails significance p-values of the maxstat test for OS of 0.08 (adjusted: 0.12) for microarrays and 0.055 (adjusted: 0.08) for RNA-seq.

4.4.4 Splice variants

RNA-seq allows the analysis of splice variants. In the context of target analysis, this especially enables assessing whether alternative splicing can lead to the elimination of therapeutic monoclonal antibody binding regions. This has been hypothesised to be a potential mechanism of resistance against CD38-antibody treatment [213] and has been screened for during the development of the BCMA-TCB antibody CC-93269 [98, 212]. In the following, the assessment of alternative splicing as part of this thesis is presented for these two exemplary targets on the HD cohort.

4.4.4.1 BCMA

During the development of the TCB antibody CC-93269 [212], the question was raised whether different *BCMA* transcripts are expressed, potentially in a way that the TCB antibody binding sequence could not be present in a subfraction of patients or myeloma cells. For *BCMA* three protein coding transcripts and five splice junctions are known. Four of the five splice junctions are transcript specific. The BCMA-001 specific splice junction (SJ 2) is spanned by at least 10 reads in each MM patient sample (see section 3.5.2.1). For all other specific splice junctions, the frequency varies (SJ 3 66.8%, SJ 4 67.8% and SJ 5 59.3%). Although all splice junctions are frequently expressed, the expression height of the two BCMA-001 splice junctions is essentially higher, as the minimum percentage of reads spanning SJ 1 is 37.21% and SJ 2 is 43.74% of all

reads spanning a BCMA splice junction. In contrast, SJ 3 is detected by maximum of 3.51%, SJ 4 by 3.69% and SJ 5 by 2.55% of the reads. Likewise, in exemplary patient 1 the expression of BCMA-001 specific splice junctions is 50 times higher compared to the BCMA-002 and BCMA-003 specific SJs, whereas in exemplary patient 2 the expression is even 100 times larger. This suggests that target sequences of BCMA directed treatment are not eliminated by alternative splicing. Hence, alternative splicing seems not to be biological relevant as mechanism of resistance against BCMA directed treatment.

4.4.4.2 CD38

During the preparation of the GMMG-MM5 (EudraCT 2010-019173-16) [150, 157], HD6 (ClinicalTrials.gov identifier: NCT02495922) and GMMG CONCEPT trial (ClinicalTrials.gov identifier: NCT03104842), the question was raised whether lack of expression or alternative splicing of *CD38* might explain the lack of efficacy of anti-CD38-antibody monotherapy in two thirds of the patients [213].
As part of this thesis, 388 MM patients were analysed. The reads spanning the splice junctions of the two protein coding transcripts (CD38-001 and CD38-005) and three further transcripts (CD38-002, CD38-003 and CD38-004) were counted. The splice junction SJ 3 is specific for CD38-001 and considerably present in 469 plasma cell samples, whereas the splice junction SJ 9, specific for CD38-005, is present in only 18 patients. And in these 18 patients, the expression is extremely low: Assuming SJ 3 is expressed 100%, the maximum observed frequency of SJ 9 is 1.8% (see section 3.5.2.2). In summary, for *CD38* one (CD38-001) of the two protein coding transcripts was identified as expressed, which excludes alternative splicing as possible upfront resistance against anti-CD38 treatment.

4.4.5 Detection of actionable mutations

Using RNA-seq, the detection of targetable mutations in myeloma cells is possible. In this thesis this is exemplified for the BRAF mutation V600E in the context of target assessment: the BRAF mutation V600E can be targeted by the clinically available inhibitors vemurafenib [24, 109] and dabrafenib [91]. *BRAF* is expressed in all 535 previously untreated myeloma patients at a low level, ranging from 2.8 to 3.9 normalised counts. The coverage of the *BRAF* mutation site is 5.39 reads spanning the position. The selection criteria are thus quite strict. The mapping quality of 255 confirms uniquely mapping reads. One mapping read in each direction confirms the reliability of the detection, by omitting the count of PCR duplicates. The nucleotides at the end of each read are not considered, because the quality in general decreases

at these positions. The V600 position of *BRAF* is detected as mutated in 2.1% of the samples. This is in the range of previously reported results for MM using genotyping (i.e. DNA-based) approaches (2.5%-4% [8, 37, 245]).
Within the setting of assessment of personalised actionable targets, mutation detection of *BRAF* is included in the RNA-seq analysis pipeline (see figure 3.1).

4.4.6 Target assessment for an individual, educated guess based treatment choice

In this section personalised treatment options for individual patients are discussed in the context of currently available approved or upcoming (currently in phase I/II clinical testing) agents. For instance, TCB or CAR T cell treatments for BCMA, CD38, GPRC5D, CD74 and NYESO1/2 are potentially highly effective. However, in a context of treatment perspective and the observed toxicities, it is unlikely that a patient could receive treatment against all the listed antigens in subsequent relapse settings. Hence, a choice is inevitable. RNA-seq could help to assist in this choice in form of an more educated than guessed approach.

4.4.6.1 Analysis of expression in individual patients

In a setting of multiple treatment options, a potential strategy for choosing a treatment is using RNA-seq expression analysis for making an educated guess. Obviously, it is useful to assess expression prior to the treatment, to exclude treatments for patients showing no expression of the respective targets. For instance, the exemplary targets *NYESO1/2*, *CD20* and *NKG2D* are expressed in subfractions of patients (12%, 83% and <1%). Due to this, NKG2D is very unlikely to be successful in clinical testing in multiple myeloma due to lack of expression. Within the setting of the multi-entity trial (ClinicalTrials.gov identifier: NCT03018405), either the very few MM patients expressing the target could be included (not likely due to the frequency), or the trial could be restricted to other disease entities. Further, it can indeed be helpful to assess the expression height of a target and first choose a treatment against the target most highly expressed (if all other known parameters would not allow a different choice). Both strategies, determining the presence or the height of expression, of course have to be tested within a clinical trial setting.
One limitation of the actual expression analysis of targets is, that no treatment targeting the respective antigen could be systematically analysed. This especially holds true for targeting BCMA, CD38, GPRC5D and CD74. It can not be excluded that the expression height might impact on the response rates of the specific agent. This further will impact the observed prognosis. Prospectively, it is useful to assess the actual expres-

sion height of the respective target and correlate it with the response to treatment. A new expression threshold for clinically present response could then potentially be implemented, which is of course possible from a bioinformatic point of view (see section 4.1.3).
Hence, to analyse this prospectively, the analysis of the expression height of the respective genes is included in the RNA-seq analysis pipeline (see figure 3.1).

4.4.6.2 Analysis of actionable mutated targets

Besides actual expression, assessment of actionable mutated targets can be used for treatment decision. For instance, *BRAF* mutation V600E is detected in 2% of the patients. This small subfraction of patients can be treated with the inhibitors vemurafenib [24, 109] and dabrafenib [91].
RNA-seq can be used to assess if the mutation is present in all cells or in a subset of cells, i.e. if the mutation is clonal or subclonal. A subclone expressing a low or non-detectable amount of target molecules will undergo a positive selection pressure and potentially be the seed of a subsequent disease progression. This clonal heterogeneity is not detectable by RNA-seq in case of unmutated targets, as a small subfraction of myeloma cells with absent expression would not be detectable. In case of mutated targets, e.g. *BRAF*, clonality is assessed in this thesis by dividing the number of reads spanning the mutation by the number of all reads spanning the position. The *BRAF* mutation is subclonal in half of the patients carrying it.
To analyse clonal heterogeneity in unmutated targets, two experimental strategies can in principal be performed: single cell sequencing or flow cytometric methods. The latter can easily be implemented in the framework of this thesis.
As for the expression height, it will be useful to assess the impact of clonal heterogeneity in the expression of mutated targets, in relation to response against respective inhibitors within a clinical trial setting.

4.4.6.3 Analysis of splice variants in individual patients

Likewise, alternative splicing can in principle be a potential strategy for choosing a treatment and can be implemented in an RNA-seq analysis pipeline (see figure 3.1). But, in case of *CD38* and *BCMA* expression alternative splicing is for neither of the targets the reason for upfront resistance against treatment in terms of eliminating the target sequences. Hence, the inclusion of these targets in the RNA-seq analysis pipeline is not necessary in clinical routine. Although this means that no actual example for this kind of setting is evident, nonetheless the principal approach is feasible.

4.5 Discussing the aims of the thesis

The primary objectives of this thesis were to form the bioinformatic basis for the implementation of RNA-seq in translational myeloma research and extended clinical routine application. The first is exemplified by the assessment of BCMA (during compound development), CD38 (during clinical trial design), and CTAs as vaccination targets for the assessment of a potential vaccination trial. For extended clinical routine application, microarray-based strategies have been translated to the current RNA-seq-based approach. This includes the assessment of risk and targets. Subsequently, an RNA-seq-based risk score has been *de novo* established and validated. Furthermore, the target analysis was extended by including the assessment of mutated transcripts and alternative splicing. In the following, the five aims and as far they are reached is discussed.

i) Establishment of a practicable pipeline to analyse and to lay a basis to perform and report RNA-seq in extended clinical routine

The developed pipeline is the proof of principle that a standardised pipeline can establish RNA-seq risk stratification and target assessment in clinical routine. The pipeline fulfils three requirements: First, it enables the independent stratification and target assessment of each new sample individually, as previously established for DNA-microarray (GEP-R [99, 156]), by ensuring the comparability of all samples. Second, RNA-seq is broadly applicable. Furthermore, it is possible in substantially more patients than microarray (about 90% *versus* about 80%), as shown in the evaluation of the LfM [99]. Likewise, less patients than in microarray analysis have to be excluded due to quality issues (3% *versus* 6.4%). Third, a model for determination of a potential threshold (and its variation) of biologically and clinically relevant expression was constructed, implemented and validated, considering the gene length (at least 1 normalised count per 1000 bp). This is necessary for the transfer of the stratifications, classifications and the target assessment.

The analysis in this thesis also covered the question in as much target assessment is comparable between DNA-microarrays and RNA-seq. Almost all targets have a consistency of $\geq 75\%$ of RNA-seq PA-seq and microarray PA call. The expression comparison shows that potential difficulties arise if genes are mostly identified as absent genes in DNA-microarrays, mainly due to background noise. To this end, RNA-seq is superior if low thresholds are envisioned.

The pipeline and its results are published and presented at the 61st annual meeting of the American society of hematology 2019 [63] (see section 4.1.6). Further, a manuscript, summarising the results is in preparation.

ii) Transfer and connect current risk stratifications and molecular classifications based on DNA-microarray to RNA-seq technology

A broad spectrum of **stratifications and classifications** have been transferred successfully to RNA-seq, including the proliferation-based RPI, the risk-based UAMS70-seq, RS-seq, EMC92-seq, and IFM15-seq as well as the molecular subtype-based TC-seq and MC-seq classification. The risk stratifications significantly delineate two or three groups of patients with different EFS and OS. The scores correlate strongly on RNA-seq and microarray (r ranges from 0.77 to 0.85) and the stratifications show few extreme differences (<1%).

One difficulty in score transfer was the **gene translation**, using primary the R database hgu133plus2.db. In case of multiple ENSG, additionally GeneAnnot was used and in case of multiple probesets jetset. The "success" was controlled by correlating the expression of both platforms and comparing the PA and PA-seq call. Three exclusion criteria were defined: First, genes with inconsistent translations between the databases were excluded, if the correlation was low ($r \leq 0.6$). Second, genes with a high number of patients showing present expression in microarray and absent expression in RNA-seq ($nCO_1 > 30\%$, see section 2.2.2), were excluded, if the correlation was very low ($r \leq 0.4$). Third, genes with no correlation at all ($r < 0.15$) were excluded. Three main reasons for non-correlation are assumed as result of the analysis in this thesis: First, mapping of probeset and ENSG is incorrect, i.e. both jetset and GeneAnnot are contradictory. This most probably is the case for *MUC1* and *CD20*. Second, on RNA-seq, genes with high homologous sequences lead to not uniquely mapping reads, which distorts the read count on RNA-seq. This is the case for *NYESO1/2*, represented by the genes *CTAG1A*, *CTAG1B* and *CTAG2*. Third, the background noise in microarrays influences the correlation, e.g. for *WT1* and *MAGEA1*.

Independent testing was performed for all stratifications on an internal test cohort (TeG) to control overfitting. Three further independent cohorts were used: an early stage patient cohort (AMM), a relapsed patient cohort (MMR) and an external cohort (CoMMpass). The stratifications were confirmed on the TeG and on the relapsed MMR group. The latter implies that the scores are independent, to a certain extent, of the actual treatment. The validation on the AMM cohort was successful for the RPI, UAMS70-seq, and RS-seq, showing that these stratifications reflect intrinsic factors of malignant plasma cells. These intrinsic factors are an underlying biological feature, i.e. they are independent of any given treatment. The testing on the CoMMpass cohort yielded four results: First, the continuous scores are highly prognostic in MM. Second, the risk stratifications delineate groups of patients with significantly different OS, as the survival analyses are highly significant for EFS and OS ($p \leq 0.001$, each),

the Brier scores are comparable to the ones of the ISS (0.1794) and R-ISS (0.1803) on the CoMMpass cohort, and the concordances for EFS and OS for RPI (0.59 and 0.63), RS-seq (0.59 and 0.62) and EMC92-seq (0.57 and 0.63) are even higher as on the TeG for RPI (0.57 and 0.60), RS-seq (0.56 and 0.58) and EMC92-seq (0.56 and 0.58). Third, according to the criteria for "success" (see section 4.2.4) the two-group risk stratifications are successfully transferred. Fourth, the three-group risk stratifications significantly differentiate only two patient groups regarding survival, as the survival curves of low and medium risk intersect (or are closely together) for EFS and OS.
Three main reasons for the differences in the performance of the stratifications have been identified: First, patient constitution in HD cohort and the CoMMpass cohort is different regarding age and applied treatment, including percentage of patients treated by high-dose therapy and autologous stem cell transplantation (50% in CoMMpass *versus* 99% in HD cohort). Second, different experimental laboratory procedures, especially library preparation, have been applied. Third, a different reference genome (GRCh37) compared to the one applied in this thesis (GRCh38) has been used.
In total, differences in technical and bioinformatic strategies are to be considered as potentially impacting on the result. However, the independent testing likewise shows that the principal approach of RNA-seq-based risk stratification is applicable in different groups. In addition, such a strategy can be implemented within a clinical study group performing multicentre clinical trials, e.g. within the two BMBF funded projects (CAMPSIMM (01ES1103), CLIOMMICS (01ZX1309)) of which this thesis is part of.
A further step for the establishment of RNA-seq-based risk stratifications and classifications in clinical routine would require the development of a "standard" processing pipeline in terms of experimental laboratory procedures as well as bioinformatic analyses (see also the presentation of the pipeline of this thesis, section 4.1.6, published and presented at the 61st annual meeting of the American society of hematology 2019 [63]).
Taking this point together, the thesis successfully allows conducting RNA-seq-based risk and target (see below) assessment as part of multicentre clinical trials.

iii) Discover novel prognostic genes and develop a *de novo* RNA-seq-based risk stratification

As intended, it was possible to *de novo* construct a respective score (HDHRS). It is based on thirteen genes associated with good prognosis and fourty genes associated with poor prognosis. The HDHRS significantly delineates three groups of patients in the TeG with different median EFS (40, 32, and 18 months) and median OS (148, 82 and 37 months). It performs as well as the R-ISS, regarding concordance for EFS (0.62

versus 0.6) and OS (0.66 *versus* 0.67), and regarding Brier scores for EFS (0.1534 *versus* 0.1577) and OS (0.1601 *versus* 0.1550).The HDHRS is independently predictive of the R-ISS in multivariate analyses and thus conveys additional prognostic information. The HDHRS delineates two groups on MMR and three group on the CoMMpass cohort (despite the methodological challenges as described above, see also section 4.3.3), implying that the HDHRS is, to a certain extent, intended of the actual treatment. In turn, the HDHRS could be successfully transferred to DNA-microarrays.

iv) Analyses of potential target structures

RNA-seq can be used to analyse targets regarding expression and, in contrast to DNA-microarrays, splice variants as well as mutations.

Of the **25 exemplary targets** assessed within this thesis, 14 are currently actionable (see table 3.14). For BCMA, CD38, CS1, GPRC5D, CD74, NYESO1/2, NKG2D and BRAF monoclonal antibodies, TCB or CAR T cell treatment options are either approved[13] (CD38, daratumumab; CS1, elotuzumab) or in clinical trials (an overview of clinical trials is depicted in table 3.15). For FGFR3, IGF1R, AURKA, and CD20 compounds have been tested in clinical trials, but results have until now not been clinically encouraging. The vaccination targets MAGEA3 [47] and MUC1 [30] are likewise actionable and in clinical trials (ClinicalTrials.gov identifiers: NCT01380145 and NCT01232712). Further potential vaccination targets are e.g. NYESO1/2, MAGEA1, RHAMM, CS1, WT1, and SSX2.

MMSET, TP53, CCND1, CCND2 and CCND3 remain of theoretical interest, regarding personalised treatment. They are potentially targetable but no actual targeting compound has been tested clinically in myeloma up to now. Nevertheless, they were included in this thesis to enlarge the categories of targets and to assess the transferability of DNA-microarray-based assessment to RNA-seq on a broader set of target genes.

In this thesis, the RNA-seq-based **assessment of target expression** was successfully implemented. For clinical application, a threshold for "present" expression was defined as at least 1 normalised count per 1000 bp (see section 4.1.3), overexpression as expression above the median BMPC expression plus three times standard deviation of BMPC expression, and aberrant expression was defined as overexpression in combination with absent expression in 90% of the BMPCs. According to these definitions, four targets are constitutively expressed (*BCMA*, *CD38*, *HM1.24* and *CD74*) and five targets are aberrantly expressed (*NYESO1/2*, *HGF*, *FGFR3*, *MAGEA1* and *MAGEA3*), detected on both platforms. Twenty targets show a very good consistency between

[13]National Cancer Institute: Drugs Approved for Multiple Myeloma and Other Plasma Cell Neoplasms; Online resource: `https://www.cancer.gov/about-cancer/treatment/drugs/multiple-myeloma`; Status: 27.02.2020, 12:55

DNA-microarray and RNA-seq PA and PA-seq call of $\geq 75\%$. RNA-seq thus can cover all target assessments previously performed by DNA-microarrays.

Expression of alternative **splice variants** can eliminate the target sequence of immunotherapeutic approaches. In this thesis, splice variant expression was exemplary assessed for the potential therapeutic targets CD38 and BCMA: For both one single transcript was identified as expressed in all samples, excluding alternative splicing as potential upfront resistance to the applied target specific treatment. These results were published (CD38 [213]) and presented (BCMA [98]).

Mutated target analysis has been implemented and is depicted for the only currently clinically targetable mutation in myeloma, the BRAF V600. This position is mutated on gene level in 2.1% of the symptomatic myeloma patients. RNA-seq can be used to assess clonal heterogeneity in terms of mutated targets, by dividing the number of reads spanning the mutation by the number of all reads spanning the position. The *BRAF* mutation is subclonal in half of the patients carrying it.

The proposed analysis strategy can easily be extended regarding future targets, mutated transcripts or, if becoming clinically relevant, alternative splicing.

v) Prospectively test the theoretical target analysis in a consecutive large patient cohort in three clinically relevant examples

The implementation of RNA-seq-based target analysis was exemplified for a consecutive large patient cohort for the assessment of *BCMA* (during compound development of CC-93269 [212]), *CD38* expression (during clinical trial design [213]), and CTAs as vaccination targets for assessment of a potential vaccination trial [209].

The analyses of the expression of ***BCMA*** in a consecutive large patient cohort was performed within the framework of this thesis. *BCMA* is expressed in all malignant plasma cell samples, i.e. *BCMA* represents an universal target in multiple myeloma treatment. The analysis of the splice junctions shows, that the expression height of the two BCMA-001 splice junctions is essentially higher than the expression height of the other splice junctions. This suggests that alternative splicing is not biological relevant as potential mechanism of resistance against BCMA directed treatment. The analyses were published by Seckinger *et al.* [212] and presented by Hose [98]. The developed BCMA TCB CC-93269 is currently in clinical testing (ClinicalTrials.gov identifier: NCT03486067) [50], showing promising initial response rates for MMR patients of 89% at a target dose of >6 mg[14].

[14] ASH Clinical News: Early-Phase Trial Suggests Bispecific Antibody CC-93269 Has Activity in Relapsed/Refractory Multiple Myeloma; Online resource: `https://www.ashclinicalnews.org/on-location/ash-annual-meeting/early-phase-trial-suggests-cc-93269-activity-relapsed-refractory-multiple-myeloma/`; Status: 27.02.2020, 19:57

CD38 monotherapy treatment is active in one thirds of the patients [147, 148, 234]. During the preparation of the GMMG-MM5 (EudraCT 2010-019173-16) [150, 157], HD6 (ClinicalTrials.gov identifier: NCT02495922) and GMMG CONCEPT trial (ClinicalTrials.gov identifier: NCT03104842), the question was raised whether lack of expression or alternative splicing of CD38 might explain this. Hence, as part of this thesis, *CD38* expression was assessed in a large cohort of patients. *CD38* is constitutively expressed. The analysis of alternative splicing variants showed that only one protein coding transcript (CD38-001) is expressed. The analyses were published by Seckinger *et al.* [213], who concluded, that neither lack of expression, nor alternative splicing are a reason for upfront resistance.

Vaccination strategies should trigger a (prophylactic) myeloma-specific immunity [209]. As part of this thesis, the expression of potential vaccination targets, e.g. *HM1.24* and CTAs (*NYESO1/2*, *MAGEA3*, *RHAMM*, *WT1*) and their association with survival have been analysed. None of the CTAs is expressed in all patients, which was published by Schmitt *et al.* [209]. The clinical suggestion of Schmitt *et al.* [209] for therapeutic strategies is to use "cocktails" of different vaccination peptides together covering at least one target in all myeloma patients [48, 209].

4.6 Conclusion and outlook

The main aim of this thesis was the implementation of RNA-seq in translational myeloma research and extended clinical routine application, in terms of risk assessment (see section 4.2) and target analysis (see section 4.4). The implemented RNA-seq analysis pipeline can be successful performed in 90% of the patients both in clinical trials (exemplified by the GMMG-MM5 phase III clinical trial [99]) and extended clinical routine. In contrast, the GEP-R, a comparable setting using DNA-microarrays, is possible in 80% of the patients [156]. This likewise includes risk stratification for patients with low tumour mass (i.e. MGUS, AMM) [99, 213]. Risk-based stratifications and molecular subtype-based classifications have been successfully transferred from Affymetrix U133 2.0 DNA-microarrays to RNA-seq-based assessment.

Target assessment has been implemented and allows the analyses of expression, mutations and splicing variants. It was exemplified regarding patient cohorts in applied translational myeloma research and extended clinical routine for the assessment of BCMA [212], CD38 [213], and CTAs as vaccination targets [209]. For all three publications, data obtained in this thesis have been essential. For *BCMA*, the proportion of patients and plasma cell precursor stages expressing the target has been assessed, to determine potential clinical side effects of the developed compound CC-93269 [212] (ClinicalTrials.gov identifier: NCT03486067). This exemplifies the usefulness of

RNA-seq in assessing potential targets for compound development. Likewise, RNA-seq can be used to assess expression prior to the treatment, to make an educated-guess choice on the selection from currently available *a priori* equivalent treatment options. In a setting where TCB or CAR T cell treatment against BCMA, CD38, GPRC5D, CD74 and NYESO1/2 is available, it can be helpful to first choose treatment against the antigen most highly expressed. BCMA and CD38 splice variant analysis has been introduced and showed that alternative splicing does not confer upfront resistance of anti-BCMA and anti-CD38 treatment. Both analysis strategies can be directly used for other target structures and will be applied to future compound developments.
Hence, this thesis is the proof of principle, that all steps necessary for targeted and risk adapted treatment strategy in myeloma can be performed using RNA-seq. The strategy has been published and presented at the 61st annual meeting of the American society of hematology 2019 [63] and a full paper publication is currently in preparation. It will include a description of the bioinformatical analysis as well as the results of stratification, classification and target analyses performed within this thesis for a large cohort of 798 consecutive patients and likewise focus on the long-term survival of the patients.
From a bioinformatic perspective, the next step for a prospective use of RNA-seq in clinical routine, is to implement a graphical user interface, e.g. as in the GEP-R [156]. This Seq-report, assessing risk and actionable targets, can be used in a tumour boards setting or as help for physicians and patients in educated-guess situation.
However, before introducing the Seq-report to clinical use in different institutions, it would be very helpful to agree on a standard for experimental laboratory procedures and bioinformatical implementation of the RNA-seq analyses. In some points, the necessity of a standard processing pipeline is already realised: For instance, the reference genomes of UCSC, termed with "hg", and the ones of NCBI, termed with "GRCh", were standardised in December 2013 in version 38. Additionally, the read file format FASTQ and the alignment format SAM/BAM are unified and broadly used. To these regards, the LfM published the experimental laboratory RNA-seq pipeline by Seckinger *et al.* [213] and the bioinformatical pipeline is depicted in detail within this thesis. The results of this thesis could be used as the basis for this standardised pipeline and lay the bioinformatic basis for the implementation of RNA-seq for risk adapted and targeted treatment strategies in clinical routine.

5 Summary

Summary

Background. Multiple myeloma is a haematological disease characterised by the accumulation of malignant plasma cells in the bone marrow. Gene expression data of 839 patients have been obtained using Affymetrix U133 2.0 microarrays in the Multiple Myeloma Research Laboratory at the University Hospital Heidelberg in extended clinical routine. RNA-sequencing was introduced as novel method thought to be superior to DNA-microarray analysis, regarding precision, lower amount of input RNA, and ability to analyse mutated transcripts and splice variants. Therefore, the primary objective of this thesis was to lay the bioinformatic basis for the implementation of RNA-sequencing in applied translational myeloma research and extended clinical routine. This includes i) to establish a practicable analysis pipeline for RNA-sequencing data, ii) to transfer and connect current stratification and classification methods based on microarrays to future RNA-sequencing technology, iii) to discover novel prognostic genes and to develop a novel RNA-sequencing-based risk stratification, iv) to analyse potential and especially actionable therapeutic targets regarding expression, alternative splicing, and mutations, and v) to prospectively test these theoretical target analysis strategies in a consecutive large patient cohort in three clinical relevant examples: BCMA, in development of the T cell bispecific antibody CC-93269 now in clinical testing, CD38, to assess potential mechanism of upfront resistance against anti-CD38 treatment, and cancer testis antigens as vaccination targets, analysing their clinical applicability.

Methods. For risk stratification, RNA-sequencing files of 535 multiple myeloma samples were divided into three groups for classification training ($\sim$40%), validation ($\sim$20%) and testing ($\sim$40%). All samples were aligned with STAR and resulting read counts were normalised using edgeR. Each microarray classification was transferred by i) translating the microarray probe sets to Ensembl gene identifiers, ii) creating RNA-sequencing classification as similar as possible to microarray classification, iii) calculating new cutoffs and iv) stratifying the patients. The classifications were validated by examining the proportions of the groups and comparing the survival performance to the microarray classifications of 534 multiple myeloma patients, respectively. Additionally, running log-rank algorithms were used to determine prognostic genes and create a novel RNA-sequencing-based classification, the HDHRS. For potential target analysis, the expression per target, e.g. *BCMA* or *CD38*, was determined, definitions for "absent" and "present" expression were established and alternative splice variants were assessed. A pipeline was developed and tested on 142 asymptomatic, 69 relapsed

and 767 symptomatic myeloma patients of the CoMMpass cohort.

Results. RNA-sequencing can be performed in 90% (in contrast to 80% by microarray) of all patients and the quality of the RNA-sequencing files is sufficient in 97% of all cases. An RNA-sequencing pipeline was successfully implemented. Proliferation-based risk assessment (GPI), risk stratifications (RS, UAMS70, EMC92 and IFM15) as well as molecular classifications (MC, TC) are as predictive on RNA-sequencing as on microarray and are significant prognostic. The novel HDHRS includes 53 discovered prognostic genes and significantly delineates three groups. All stratifications and classifications were validated on the independent test group. On the CoMMpass cohort the stratifications are significant, with laboratory and bioinformatical variations to consider.

Nineteen potential targets, five theoretical targets and the mutated target *BRAF* were assessed. The actionable targets *BCMA*, *CD38*, and *CD74* are expressed in all, the cancer testis antigens *MAGEA3* (33%), *RHAMM* (88%) and *NYESO1/2* (12%) in subfractions of myeloma patients, and 2% show expression of mutated (V600E) *BRAF*.

To assess a potential mechanism of upfront resistance, splice variants of BCMA and CD38 have been analysed. For both targets only one main transcript is expressed.

Discussion. RNA-sequencing has been implemented successfully in applied translational myeloma research and extended clinical routine application: All intended risk stratifications and classifications could be successfully transferred and a novel RNA-seq-based risk score (HDHRS) be developed and validated.

The target assessment was exemplified regarding patient cohorts for assessment of *BCMA*, *CD38*, and cancer testis antigen as vaccination targets. For *BCMA*, the proportion of patients and plasma cell precursor stages expressing the target has been assessed, to determine potential clinical side effects of the developed compound CC-93269. This exemplifies the usefulness of RNA-sequencing in assessing potential targets for compound development. Likewise, RNA-sequencing is useful to assess expression prior to therapy, to exclude treatments for patients lacking expression of the respective targets. BCMA and CD38 splice variant analyses showed that alternative splicing does not confer up-front resistance of anti-BCMA and anti-CD38 treatment. Both analysis strategies can and will be directly applied to future compound developments.

Hence, this thesis is the proof of principle, that all steps necessary for targeted and risk adapted treatment strategy in myeloma can be performed using RNA-sequencing. The results of this thesis, including the developed pipeline, are the fundamental step for the implementation of RNA-sequencing for risk adapted and targeted treatment strategies in clinical routine.

Zusammenfassung

Hintergrund. Das multiple Myelom ist eine hämatologische Erkrankung, die durch die Ansammlung maligner Plasmazellen im Knochenmark gekennzeichnet ist. Im Labor für Myelomforschung des Universitätsklinikums Heidelberg wurden Affymetrix U133 2.0-Microarray Genexpressionsdaten von 839 Patienten in erweiterter klinischer Routine untersucht. Die RNA-Sequenzierung wurde als neuartige Methode vorgestellt, die der Microarray Analyse überlegen sein soll, hinsichtlich Präzision, geringerer Menge an Input-RNA und der Möglichkeit Transkripte sowie Spleißvarianten zu analysieren. Daher bestand das Hauptziel dieser Arbeit darin, die bioinformatische Grundlage für die Implementierung der RNA-Sequenzierung in der angewandten translationalen Myelomforschung und erweiterten klinischen Routine zu schaffen. Das beinhaltet i) die Erstellung einer praktikablen Analysepipeline für RNA-Sequenzierungsdaten, ii) die Übertragung und die Verbindung von aktuellen Einteilungs- und Klassifizierungsmethoden basierend auf Microarrays zur zukünftigen RNA-Sequenzierungstechnologie, iii) die Entdeckung neuer prognostischer Gene und die Entwicklung einer auf RNA-Sequenzierung basierenden Risikoeinteilung, iv) die Analyse potentieller und insbesondere anwendbarer therapeutische Zielgene in Bezug auf Expression, alternatives Spleißen und Mutationen und v) die prospektive Testung dieser theoretischen Zielgen-Analysestrategien in einer großen Patientenkohorte in drei klinisch relevanten Beispielen: BCMA, bei der Entwicklung des gerade in klinischen Tests befindlichen, bispezifischen T-Zell-Antikörpers CC-93269, CD38, zur Analyse der klinischen Anwendbarkeit von Anti-CD38-Behandlung, und Krebs-Hoden-Antigenen als Impfziel unter Bewertung des möglichen Mechanismus der Vorabresistenz.

Methoden. Für die Risikostratifizierung wurden RNA-Sequenzierungs Proben von 535 Myelom Patienten in drei Gruppen eingeteilt, eine Trainings- ($\sim$40%), eine Validierungs- ($\sim$20%) und eine Testgruppe ($\sim$40%). Alle Proben wurden mit STAR aligniert und die resultierenden Read Anzahlen mit edgeR normalisert. Jede Microarray-Klassifizierung wurde übertragen durch i) Translation der Microarray-Sondennamen in Ensembl-Gennamen, ii) Erzeugung einer RNA-Sequenzierungs-Klassifizierung, die so ähnlich wie möglich ist zur Microarray-Klassifizierung, iii) Berechnung neuer Grenzwerte und iv) die Einteilung der Patienten. Die Validierung der Klassifikationen erfolgte durch die Untersuchung der Proportionen der Gruppen und dem Vergleich der Überlebenszeitanalyse zu den Microarray-Klassifikationen von 534 Patienten mit multiplem Myelom. Zusätzlich wurden laufende Log-Rank-Algorithmen verwendet, um prognostische Gene zu bestimmen und eine neue auf RNA-Sequenzierung basierende Klassifizierung zu erstellen, den HDHRS. Für die

Analyse potentieller Zielgene wurde die Expression pro Zielgen, z.B. *BCMA* oder *CD38*, bestimmt, eine Definition für "fehlende" und "vorhandene" Expression etabliert und alternative Spleißvarianten bewertet. Eine Pipeline wurde entwickelt und an 142 asymptomatischen, 69 rezidivierten und 767 symptomatische Myelompatienten der CoMMpass-Kohorte getestet.

Ergebnisse. Die RNA-Sequenzierung kann bei 90% aller Patienten durchgeführt werden (im Gegensatz zu 80% bei Mikroarray) und die Qualität der RNA-Sequenzierungsdaten ist in 97% der Fälle ausreichend. Eine RNA-Sequenzierungspipeline wurde erfolgreich implementiert. Proliferation basierte Risikobewertung (GPI), Risikoklassifizierungen (RS, UAMS70, EMC92 und IFM15), sowie molekulare Klassifikationen (MC, TC) sind signifikant prognostisch und ebenso prädiktiv für die RNA-Sequenzierung wie auf Microarrays. Der neue HDHRS umfasst 53 prognostische Gene und unterscheidet drei Gruppen signifikant. Alle Klassifikationen wurden in der unabhängigen Testgruppe validiert. Die Klassifikationen sind auf der CoMMpass Kohorte signifikant, unter Berücksichtigung Labor- und bioinformatischen Variationen. Neunzehn potenzielle, fünf theoretische und das mutierte Zielgen *BRAF* wurden analysiert. Die angreifbaren Zielgene *BCMA*, *CD38* und *CD74* sind in allen Patienten exprimiert, die Krebs-Hoden-Antigene *MAGEA3* (33%), *RHAMM* (88%) und *NYESO1/2* (12%) in Subfraktionen von Myelompatienten und 2% zeigen die Expression des mutierten (V600E) *BRAF* Genes. Um einen möglichen Mechanismus der Vorabresistenz auszuschließen, wurden die Spleiß-Varianten von BCMA und CD38 untersucht. Für beide Zielgene wird nur ein Haupttranskript exprimiert.

Diskussion. Die RNA-Sequenzierung wurde erfolgreich implementiert und in der Myelomforschung und der erweiterten klinischen Routine angewandt. Alle untersuchten Risiko-Einteilungen und Klassifikationen konnten erfolgreich übertragen werden und eine neue auf RNA-Sequenzierung basierende Klassifizierung (HDHRS) konnte entwickelt und validiert werden. Die Zielgen Analyse wurde am Beispiel von *BCMA*, *CD38* und Krebs-Hoden-Antigenen als Impfzielgene in Bezug auf Patientenkohorten beispielhaft dargestellt. Für *BCMA* wurde der Anteil der Patienten und Plasmazellvorläufer-Stadien bestimmt, die das Zielgen exprimieren, um mögliche klinische Nebenwirkungen des entwickelten Medikaments CC-63269 zu bestimmen. Dies veranschaulicht die Nützlichkeit von RNA-Sequenzierung bei der Bewertung potenzieller Zielgene für die Wirkstoffforschung. Ebenso ist RNA-Sequenzierung nützlich, um die Expression vor Beginn der Therapie zu bewerten, um Patienten von der Behandlung auszuschließen, denen die Expression des jeweiligen Zielgens fehlt. BCMA und CD38 Spleißvariantenanalysen zeigten, dass alternatives Spleißen nicht die Ursache für Resistenzen gegen Anti-BCMA- und Anti-CD38-Behandlung sind. Beide

Analysestrategien können und werden direkt auf zukünftige Wirkstoff Entwicklungen angewendet werden.
Somit ist diese Dissertation der Grundsatzbeweis, dass alle Schritte, die notwendig sind für eine zielgerichtete und risikoadaptierte Behandlungsstrategie beim Myelom, mit RNA-Sequenzierung durchführbar sind. Die Ergebnisse dieser Arbeit, einschließlich der entwickelten Pipeline, sind der grundlegende Schritt für die Implementierung von RNA-Sequenzierung für risikoadaptierte und zielgerichtete Behandlungsstrategien in der klinischen Routine.

6 References

[1] Abrahams, C. L., Li, X., Embry, M., Yu, A., Krimm, S., Krueger, S., Greenland, N. Y., Wen, K. W., Jones, C., DeAlmeida, V., Solis, W. A., Matheny, S., Kline, T., Yam, A. Y., Stafford, R., Wiita, A. P., Hallam, T., Lupher, M., and Molina, A. (2018). **Targeting CD74 in multiple myeloma with the novel, site-specific antibody-drug conjugate STRO-001**. *Oncotarget*, 9:37700–37714. doi:10.18632/oncotarget.26491.

[2] Agresti, A. (2007). **An introduction to categorical data analysis**. Wiley series in probability and statistics. Wiley-Interscience, 2. edition. doi:10.1002/0470114754. URL: `http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10278250`.

[3] Ailawadhi, S., Aldoss, I. T., Yang, D., Razavi, P., Cozen, W., Sher, T., and Chanan-Khan, A. (2012). **Outcome disparities in multiple myeloma: a SEER-based comparative analysis of ethnic subgroups**. *British Journal of Haematology*, 158:91–98. doi:10.1111/j.1365-2141.2012.09124.x.

[4] Anders, S. and Huber, W. (2019). **Differential expression of RNA-Seq data at the gene level the DESeq package**, version 1.36.0. URL: `https://bioconductor.org/packages/release/bioc/vignettes/DESeq/inst/doc/DESeq.pdf`. [Last visited: 16.10.2019].

[5] Anders, S., Pyl, P. T., and Huber, W. (2015). **HTSeq–a Python framework to work with high-throughput sequencing data**. *Bioinformatics*, 31:166–169. doi:10.1093/bioinformatics/btu638.

[6] Andersen, P. K. and Gill, R. D. (1982). **Cox's Regression Model for Counting Processes A Large Sample Study**. *The Annals of Statistics*, 10:1100–1120. doi:10.1214/aos/1176345976.

[7] Andrews, S. (2010). **FastQC: a quality control tool for high throughput sequence data**. URL: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc`. [Last visited: 18.10.18].

[8] Andrulis, M., Lehners, N., Capper, D., Penzel, R., Heining, C., Huellein, J., Zenz, T., von Deimling, A., Schirmacher, P., Ho, A. D., Goldschmidt, H., Neben, K., and Raab, M. S. (2013). **Targeting the BRAF V600E mutation in multiple myeloma**. *Cancer discovery*, 3:862–869. doi:10.1158/2159-8290.CD-13-0014.

[9] Anonymous (2008). **Pearson's Correlation Coefficient**. In: *Encyclopedia of Public Health*, editor Kirch, W., pages 1090–1091, Dordrecht. Springer Netherlands. doi:10.1007/978-1-4020-5614-7_2569.

[10] Attal, M., Harousseau, J. L., Stoppa, A. M., Sotto, J. J., Fuzibet, J. G., Rossi, J. F., Casassus, P., Maisonneuve, H., Facon, T., Ifrah, N., Payen, C., and Bataille, R. (1996). **A prospective, randomized trial of autologous bone marrow transplantation and chemotherapy in multiple myeloma. Intergroupe Français du Myélome**. *The New England journal of medicine*, 335:91–97. doi:10.1056/NEJM199607113350204.

[11] Avet-Loiseau, H., Attal, M., Campion, L., Caillot, D., Hulin, C., Marit, G., Stoppa, A.-M., Voillat, L., Wetterwald, M., Pegourie, B., Voog, E., Tiab, M., Banos, A., Jaubert, J., Bouscary, D., Macro, M., Kolb, B., Traulle, C., Mathiot, C., Magrangeas, F., Minvielle, S., Facon, T., and Moreau, P. (2012). **Long-term analysis of the IFM 99 trials for myeloma Cytogenetic abnormalities t(4;14), del(17p), 1q gains play a major role in defining long-term survival**. *Journal of clinical oncology*, 30:1949–1952. doi:10.1200/JCO.2011.36.5726.

[12] Avet-Loiseau, H., Attal, M., Moreau, P., Charbonnel, C., Garban, F., Hulin, C., Leyvraz, S., Michallet, M., Yakoub-Agha, I., Garderet, L., Marit, G., Michaux, L., Voillat, L., Renaud, M., Grosbois, B., Guillerm, G., Benboubker, L., Monconduit, M., Thieblemont, C., Casassus, P., Caillot, D., Stoppa, A.-M., Sotto, J.-J., Wetterwald, M., Dumontet, C., Fuzibet, J.-G., Azais, I., Dorvaux, V., Zandecki, M., Bataille, R., Minvielle, S., Harousseau, J.-L., Facon, T., and Mathiot, C. (2007). **Genetic abnormalities and survival in multiple myeloma The experience of the Intergroupe Francophone du Myélome**. *Blood*, 109:3489–3495. doi:10.1182/blood-2006-08-040410.

[13] Bahleda, R., Italiano, A., Hierro, C., Mita, A., Cervantes, A., Chan, N., Awad, M., Calvo, E., Moreno, V., Govindan, R., Spira, A., Gonzalez, M., Zhong, B., Santiago-Walker, A., Poggesi, I., Parekh, T., Xie, H., Infante, J., and Tabernero, J. (2019). **Multicenter Phase I Study of Erdafitinib (JNJ-42756493), Oral Pan-Fibroblast Growth Factor Receptor Inhibitor, in Patients with Advanced or Refractory Solid Tumors**. *Clinical Cancer Research*, 25:4888–4897. doi:10.1158/1078-0432.CCR-18-3334.

[14] Barber, A., Meehan, K. R., and Sentman, C. L. (2011). **Treatment of multiple myeloma with adoptively transferred chimeric NKG2D receptor-expressing T cells**. *Gene Therapy*, 18:509–516. doi:10.1038/gt.2010.174.

[15] Barber, A., Zhang, T., Megli, C. J., Wu, J., Meehan, K. R., and Sentman, C. L. (2008). **Chimeric NKG2D receptor-expressing T cells as an immunotherapy for multiple myeloma**. *Experimental hematology*, 36:1318–1328. doi:10.1016/j.exphem.2008.04.010.

[16] Barlogie, B., Anaissie, E., van Rhee, F., Haessler, J., Hollmig, K., Pineda-Roman, M., Cottler-Fox, M., Mohiuddin, A., Alsayed, Y., Tricot, G., Bolejack, V., Zangari, M., Epstein, J., Petty, N., Steward, D., Jenkins, B., Gurley, J., Sullivan, E., Crowley, J., and Shaughnessy, J. D. (2007). **Incorporating bortezomib into upfront treatment for multiple myeloma Early results of total therapy 3**. *British journal of haematology*, 138:176–185. doi:10.1111/j.1365-2141.2007.06639.x.

[17] Barlogie, B., Anaissie, E. J., van Rhee, F., Shaughnessy, J. D., Haessler, J., Pineda-Roman, M., Hollmig, K., Epstein, J., and Crowley, J. J. (2008). **Total therapy (TT) for myeloma (MM)—10% cure rate with TT1 suggested by >10yr continuous complete remission (CCR) Bortezomib in TT3 overcomes poor-risk associated with T(4;14) and DelTP53 in TT2**. *Journal of Clinical Oncology*, 26:8516. doi:10.1200/jco.2008.26.15_suppl.8516.

[18] Batchu, R. B., Moreno, A. M., Szmania, S. M., Bennett, G., Spagnoli, G. C., Ponnazhagan, S., Barlogie, B., Tricot, G., and van Rhee, F. (2005). **Protein transduction of dendritic cells for NY-ESO-1-based immunotherapy of myeloma**. *Cancer Research*, 65:10041–10049. doi:10.1158/0008-5472.CAN-05-1383.

[19] Bauer, D. F. (1972). **Constructing Confidence Sets Using Rank Statistics**. *Journal of the American Statistical Association*, 67:687–690. doi:10.1080/01621459.1972.10481279.

[20] Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). **The new S language**. Wadsworth and Brooks/Cole.

[21] Benjamini, Y. and Hochberg, Y. (1995). **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300.

[22] Bergsagel, P. L., Kuehl, W. M., Zhan, F., Sawyer, J., Barlogie, B., and Shaughnessy, Jr, John (2005). **Cyclin D dysregulation: an early and unifying pathogenic event in multiple myeloma**. *Blood*, 106:296–303. doi:10.1182/blood-2005-01-0034.

[23] Best, D. J. and Roberts, D. E. (1975). **Algorithm AS 89 The Upper Tail Probabilities of Spearman's Rho**. *Applied Statistics*, 24:377. doi:10.2307/2347111.

[24] Bollag, G., Tsai, J., Zhang, J., Zhang, C., Ibrahim, P., Nolop, K., and Hirth, P. (2012). **Vemurafenib The first drug approved for BRAF-mutant cancer**. *Nature reviews drug discovery*, 11:873–886. doi:10.1038/nrd3847.

[25] Boyd, K. D., Ross, F. M., Chiecchio, L., Dagrada, G. P., Konn, Z. J., Tapper, W. J., Walker, B. A., Wardell, C. P., Gregory, W. M., Szubert, A. J., Bell, S. E., Child, J. A., Jackson, G. H., Davies, F. E., and Morgan, G. J. (2012). **A novel prognostic model in myeloma based on co-segregating adverse FISH lesions and the ISS Analysis of patients treated in the MRC Myeloma IX trial**. *Leukemia*, 26:349–355. doi:10.1038/leu.2011.204.

[26] Bringhen, S., Mateos, M. V., Zweegman, S., Larocca, A., Falcone, A. P., Oriol, A., Rossi, D., Cavalli, M., Wijermans, P., Ria, R., Offidani, M., Lahuerta, J. J., Liberati, A. M., Mina, R., Callea, V., Schaafsma, M., Cerrato, C., Marasca, R., Franceschini, L., Evangelista, A., Teruel, A.-I., van der Holt, B., Montefusco, V., Ciccone, G., Boccadoro, M., Miguel, J. S., Sonneveld, P., and Palumbo, A. (2013). **Age and organ damage correlate with poor survival in myeloma patients: meta-analysis of 1435 individual patient data from 4 randomized trials**. *Haematologica*, 98:980–987. doi:10.3324/haematol.2012.075051.

[27] Broyl, A., Hose, D., Lokhorst, H., de Knegt, Y., Peeters, J., Jauch, A., Bertsch, U., Buijs, A., Stevens-Kroef, M., Beverloo, H. B., Vellenga, E., Zweegman, S., Kersten, M.-J., van der Holt, B., el Jarari, L., Mulligan, G., Goldschmidt, H., van Duin, M., and Sonneveld, P. (2010). **Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients**. *Blood*, 116:2543–2553. doi:10.1182/blood-2009-12-261032.

[28] Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments**. *BMC Bioinformatics*, 11:94. doi:10.1186/1471-2105-11-94.

[29] Carlson, M. (2016). **hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2)**, version 3.2.3. URL: `https://bioconductor.org/packages/3.4/data/annotation/html/hgu133plus2.db.html`. [Last visited: 17.10.2019].

[30] Carmon, L., Avivi, I., Kovjazin, R., Zuckerman, T., Dray, L., Gatt, M. E., Or, R., and Shapira, M. Y. (2015). **Phase I/II study exploring ImMucin, a pan-major histocompatibility complex, anti-MUC1 signal peptide vaccine, in multiple myeloma patients**. *British Journal of Haematology*, 169:44–56. doi:10.1111/bjh.13245.

[31] Carpenter, R. O., Evbuomwan, M. O., Pittaluga, S., Rose, J. J., Raffeld, M., Yang, S., Gress, R. E., Hakim, F. T., and Kochenderfer, J. N. (2013). **B-cell maturation antigen is a promising target for adoptive T-cell therapy of multiple myeloma**. *Clinical cancer research*, 19:2048–2060. doi:10.1158/1078-0432.CCR-12-2422.

[32] Carter, R. E. and Huang, P. (2009). **Cautionary Note Regarding the Use of CIs Obtained From Kaplan-Meier Survival Curves**. *Journal of Clinical Oncology*, 27:174–175. doi:10.1200/JCO.2008.18.8011.

[33] Chalifa-Caspi, V., Shmueli, O., Benjamin-Rodrig, H., Rosen, N., Shmoish, M., Yanai, I., Ophir, R., Kats, P., Safran, M., and Lancet, D. (2003). **GeneAnnot Interfacing GeneCards with high-throughput gene expression compendia**. *Briefings in bioinformatics*, 4:349–360. doi:10.1093/bib/4.4.349.

[34] Chalifa-Caspi, V., Yanai, I., Ophir, R., Rosen, N., Shmoish, M., Benjamin-Rodrig, H., Shklar, M., Stein, T. I., Shmueli, O., Safran, M., and Lancet, D. (2004). **GeneAnnot Comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes**. *Bioinformatics*, 20:1457–1458. doi:10.1093/bioinformatics/bth081.

[35] Chambers, J. M., Cleveland, W. S., Kleiner, B., and Turkey, P. A. (1983). **Graphical methods for data analysis**. Wadsworth and Brooks/Cole.

[36] Chambers, J. M. and Hastie, T. J. (1992). **Statistical models in S**. Chapman & Hall.

[37] Chapman, M. A., Lawrence, M. S., Keats, J. J., Cibulskis, K., Sougnez, C., Schinzel, A. C., Harview, C. L., Brunet, J.-P., Ahmann, G. J., Adli, M., Anderson, K. C., Ardlie, K. G., Auclair, D., Baker, A., Bergsagel, P. L., Bernstein, B. E., Drier, Y., Fonseca, R., Gabriel, S. B., Hofmeister, C. C., Jagannath, S., Jakubowiak, A. J., Krishnan, A., Levy, J., Liefeld, T., Lonial, S., Mahan, S., Mfuko, B., Monti, S., Perkins, L. M., Onofrio, R., Pugh, T. J., Rajkumar, S. V., Ramos, A. H., Siegel, D. S., Sivachenko, A., Stewart, A. K., Trudel, S., Vij, R., Voet, D., Winckler, W., Zimmerman, T., Carpten, J., Trent, J., Hahn, W. C., Garraway, L. A., Meyerson, M., Lander, E. S., Getz, G., and Golub, T. R. (2011). **Initial genome sequencing and analysis of multiple myeloma**. *Nature*, 471:467–472. doi:10.1038/nature09837.

[38] Chen, Y., Lun, A., McCarthy, D., Zhou, X., Robinson, M., and Smyth, G. (2016). **edgeR: Empirical Analysis of Digital Gene Expression Data in R**, version 3.16.5. URL: `https://bioconductor.org/packages/3.4/bioc/html/edgeR.html`. [Last visited: 17.10.2019].

[39] Chen, Y., McCarthy, D., Ritchie, M., Robinson, M., and Smyth, G. (2016). **edgeR: differential expression analysis of digital gene expression data User's Guide**, version 30 June 2016. URL: `https://bioconductor.org/packages/3.4/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf`. [Last visited: 31.10.2019].

[40] Chiecchio, L., Protheroe, R. K. M., Ibrahim, A. H., Cheung, K. L., Rudduck, C., Dagrada, G. P., Cabanas, E. D., Parker, T., Nightingale, M., Wechalekar, A., Orchard, K. H., Harrison, C. J., Cross, N. C. P., Morgan, G. J., and Ross, F. M. (2006). **Deletion of chromosome 13 detected by conventional cytogenetics is a critical prognostic factor in myeloma**. *Leukemia*, 20:1610–1617. doi:10.1038/sj.leu.2404304.

[41] Child, J. A., Morgan, G. J., Davies, F. E., Owen, R. G., Bell, S. E., Hawkins, K., Brown, J., Drayson, M. T., and Selby, P. J. (2003). **High-dose chemotherapy with hematopoietic stem-cell rescue for multiple myeloma**. *The New England journal of medicine*, 348:1875–1883. doi:10.1056/NEJMoa022340.

[42] Chng, W. J., Dispenzieri, A., Chim, C.-S., Fonseca, R., Goldschmidt, H., Lentzsch, S., Munshi, N., Palumbo, A., Miguel, J. S., Sonneveld, P., Cavo, M., Usmani, S., Durie, B. G. M., Avet-Loiseau, H., and International Myeloma Working Group (2014). **IMWG consensus on risk stratification in multiple myeloma**. *Leukemia*, 28:269–277. doi:10.1038/leu.2013.247.

[43] Chng, W. J., Glebov, O., Bergsagel, P. L., and Kuehl, W. M. (2007). **Genetic events in the pathogenesis of multiple myeloma**. *Best practice & research. Clinical haematology*, 20:571–596. doi:10.1016/j.beha.2007.08.004.

[44] Chretien, M.-L., Hebraud, B., Cances-Lauwers, V., Hulin, C., Marit, G., Leleu, X., Karlin, L., Roussel, M., Stoppa, A.-M., Guilhot, F., Lamy, T., Garderet, L., Pegourie, B., Dib, M., Sebban, C., Lenain, P., Brechignac, S., Royer, B., Wetterwald, M., Legros, L., Orsini-Piocelle, F., Voillat, L., Delbrel, X., Caillot, D., Macro, M., Facon, T., Attal, M., Moreau, P., Avet-Loiseau, H., and Corre, J. (2014). **Age is a prognostic factor even among patients with multiple myeloma younger than 66 years treated with high-dose melphalan: the IFM experience on 2316 patients**. *Haematologica*, 99:1236–1238. doi:10.3324/haematol.2013.098608.

[45] Chu, J., Deng, Y., Benson, D. M., He, S., Hughes, T., Zhang, J., Peng, Y., Mao, H., Yi, L., Ghoshal, K., He, X., Devine, S. M., Zhang, X., Caligiuri, M. A., Hofmeister, C. C., and Yu, J. (2014). **CS1-specific chimeric antigen receptor (CAR)-engineered natural killer cells enhance in vitro and in vivo antitumor activity against human multiple myeloma**. *Leukemia*, 28:917–927. doi:10.1038/leu.2013.279.

[46] Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants**. *Nucleic acids research*, 38:1767–1771. doi:10.1093/nar/gkp1137.

[47] Cohen, A. D., Lendvai, N., Nataraj, S., Imai, N., Jungbluth, A. A., Tsakos, I., Rahman, A., Mei, A. H.-C., Singh, H., Zarychta, K., Kim-Schulze, S., Park, A., Venhaus, R., Alpaugh, K., Gnjatic, S., and Cho, H. J. (2019). **Autologous Lymphocyte Infusion Supports Tumor Antigen Vaccine-Induced Immunity in Autologous Stem Cell Transplant for Multiple Myeloma**. *Cancer immunology research*, 7:658–669. doi:10.1158/2326-6066.CIR-18-0198.

[48] Condomines, M., Hose, D., Raynaud, P., Hundemer, M., de Vos, J., Baudard, M., Moehler, T., Pantesco, V., Moos, M., Schved, J.-F., Rossi, J.-F., Rème, T., Goldschmidt, H., and Klein, B. (2007). **Cancer/Testis Genes in Multiple Myeloma: Expression Patterns and Prognosis Value Determined by Microarray Analysis**. *The Journal of Immunology*, 178:3307–3315. doi:10.4049/jimmunol.178.5.3307.

[49] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). **A survey of best practices for RNA-seq data analysis**. *Genome biology*, 17:13. doi:10.1186/s13059-016-0881-8.

[50] Costa, L. J., Wong, S. W., Bermúdez, A., de la Rubia, J., Mateos, M.-V., Ocio, E. M., Rodríguez-Otero, P., San-Miguel, J., Li, S., Sarmiento, R., Lardelli, P., Gaudy, A., Boss, I., Kelly, L. M., Burgess, M. R., Hege, K., and Bensinger, W. I. (2019). **First Clinical Study of the B-Cell Maturation Antigen (BCMA) 2+1 T Cell Engager (TCE) CC-93269 in Patients (Pts) with Relapsed/Refractory Multiple Myeloma (RRMM): Interim Results of a Phase 1 Multicenter Trial**. *Blood*, 134:143. doi:10.1182/blood-2019-122895, ASH Annual Meeting Abstract.

[51] Craig, D. W., Liang, W., Venkata, Y., Kurdoglu, A., Aldrich, J., Auclair, D., Allen, K., Harrison, B., Jewell, S., Kidd, P. G., Correll, M., Jagannath, S., Siegel, D. S., Vij, R., Orloff, G., Zimmerman, T. M., Network, M. C., Capone, W., Carpten, J., and Lonial, S. (2013). **Interim Analysis Of The Mmrf Commpass Trial, a Longitudinal Study In Multiple Myeloma Relating Clinical Outcomes To Genomic and Immunophenotypic Profiles**. *Blood*, 122:532, ASH Annual Meeting Abstract.

[52] Cremer, F. W., Bila, J., Buck, I., Kartal, M., Hose, D., Ittrich, C., Benner, A., Raab, M. S., Theil, A.-C., Moos, M., Goldschmidt, H., Bartram, C. R., and Jauch, A. (2005). **Delineation of distinct subgroups of multiple myeloma and a model for clonal evolution based on interphase cytogenetics**. *Genes, chromosomes & cancer*, 44:194–203. doi:10.1002/gcc.20231.

[53] Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J. (2018). **xtable: Export Tables to LaTeX or HTML**, version 1.8-2. URL: `https://cran.r-project.org/src/contrib/Archive/xtable/xtable_1.8-2.tar.gz`. [Last visited: 17.10.2019].

[54] Davis, M. E., Zuckerman, J. E., Choi, C. H. J., Seligson, D., Tolcher, A., Alabi, C. A., Yen, Y., Heidel, J. D., and Ribas, A. (2010). **Evidence of RNAi in humans from systemically administered siRNA via targeted nanoparticles**. *Nature*, 464:1067–1070. doi:10.1038/nature08956.

[55] de Weers, M., Tai, Y.-T., van der Veer, M. S., Bakker, J. M., Vink, T., Jacobs, D. C. H., Oomen, L. A., Peipp, M., Valerius, T., Slootstra, J. W., Mutis, T., Bleeker, W. K., Anderson, K. C., Lokhorst, H. M., van de Winkel, J. G. J., and Parren, P. W. H. I. (2011). **Daratumumab, a novel therapeutic human CD38 monoclonal antibody, induces killing of multiple myeloma and other hematological tumors**. *The Journal of Immunology*, 186:1840–1848. doi:10.4049/jimmunol.1003032.

[56] Decaux, O., Lodé, L., Magrangeas, F., Charbonnel, C., Gouraud, W., Jézéquel, P., Attal, M., Harousseau, J.-L., Moreau, P., Bataille, R., Campion, L., Minvielle, S., and Intergroupe Francophone du Myélome (2008). **Prediction of survival in multiple myeloma based on gene expression profiles reveals cell cycle and chromosomal instability signatures in high-risk patients and hyperdiploid signatures in low-risk patients: a study of the Intergroupe Francophone du Myélome**. *Journal of clinical oncology*, 26:4798–4805.

[57] Deckert, J., Wetzel, M.-C., Bartle, L. M., Skaletskaya, A., Goldmacher, V. S., Vallée, F., Zhou-Liu, Q., Ferrari, P., Pouzieux, S., Lahoute, C., Dumontet, C., Plesa, A., Chiron, M., Lejeune, P., Chittenden, T., Park, P. U., and Blanc, V. (2014). **SAR650984, a novel humanized CD38-targeting antibody, demonstrates potent antitumor activity in models of multiple myeloma and other CD38+ hematologic malignancies**. *Clinical cancer research*, 20:4574–4583. doi:10.1158/1078-0432.CCR-14-0695.

[58] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics*, 29:15–21. doi:10.1093/bioinformatics/bts635.

[59] Dobin, A. and Gingeras, T. R. (2015). **Mapping RNA-seq Reads with STAR**. *Current protocols in bioinformatics*, 51:11.14.1–19. doi:10.1002/0471250953.bi1114s51.

[60] Durie, B. G. (1986). **Staging and kinetics of multiple myeloma**. *Seminars in oncology*, 13:300–309.

[61] Durie, B. G. and Salmon, S. E. (1975). **A clinical staging system for multiple myeloma. Correlation of measured myeloma cell mass with presenting clinical features, response to treatment, and survival**. *Cancer*, 36:842–854.

[62] Durie, B. G. M., Harousseau, J.-L., Miguel, J. S., Bladé, J., Barlogie, B., Anderson, K., Gertz, M., Dimopoulos, M., Westin, J., Sonneveld, P., Ludwig, H., Gahrton, G., Beksac, M., Crowley, J., Belch, A., Boccadaro, M., Cavo, M., Turesson, I., Joshua, D., Vesole, D., Kyle, R., Alexanian, R., Tricot, G., Attal, M., Merlini, G., Powles, R., Richardson, P., Shimizu, K., Tosi, P., Morgan, G., and Rajkumar, S. V. (2006). **International uniform response criteria for multiple myeloma**. *Leukemia*, 20:1467–1473. doi:10.1038/sj.leu.2404284.

[63] Emde, M., Seckinger, A., Benes, V., Moreaux, J., Beck, S., and Hose, D. (2019). **RNA-Sequencing Based Assessment of Targets, Risk and Long Term Survival for Personalized Treatment of Multiple Myeloma**. *Blood*, 134:1801. doi:10.1182/blood-2019-131159, ASH Annual Meeting Abstract and Poster.

[64] Ferrari, F., Bortoluzzi, S., Coppe, A., Sirota, A., Safran, M., Shmoish, M., Ferrari, S., Lancet, D., Danieli, G. A., and Bicciato, S. (2007). **Novel definition files for human GeneChips based on GeneAnnot**. *BMC Bioinformatics*, 8:446. doi:10.1186/1471-2105-8-446.

[65] Fleming, T. H. and Harrington, D. P. (1984). **Nonparametric estimation of the survival distribution in censored data.** *Communications in Statistics*, 13:2469–2486.

[66] Fonseca, R., Bergsagel, P. L., Drach, J., Shaughnessy, J., Gutierrez, N., Stewart, A. K., Morgan, G., van Ness, B., Chesi, M., Minvielle, S., Neri, A., Barlogie, B., Kuehl, W. M., Liebisch, P., Davies, F., Chen-Kiang, S., Durie, B. G. M., Carrasco, R., Sezer, O., Reiman, T., Pilarski, L., and Avet-Loiseau, H. (2009). **International Myeloma Working Group molecular classification of multiple myeloma Spotlight review**. *Leukemia*, 23:2210–2221. doi:10.1038/leu.2009.174.

[67] Forgy, E. (1965). **Cluster Analysis of Multivariate Data Efficiency versus Interpretability of Classification**. *Biometrics*, 21:768–769.

[68] García-Echeverría, C., Pearson, M. A., Marti, A., Meyer, T., Mestan, J., Zimmermann, J., Gao, J., Brueggen, J., Capraro, H.-G., Cozens, R., Evans, D. B., Fabbro, D., Furet, P., Porta, D. G., Liebetanz, J., Martiny-Baron, G., Ruetz, S., and Hofmann, F. (2004). **In vivo antitumor activity of NVP-AEW541-A novel, potent, and selective inhibitor of the IGF-IR kinase**. *Cancer Cell*, 5:231–239.

[69] Garfall, A. L., Maus, M. V., Hwang, W.-T., Lacey, S. F., Mahnke, Y. D., Melenhorst, J. J., Zheng, Z., Vogl, D. T., Cohen, A. D., Weiss, B. M., Dengel, K., Kerr, N. D. S., Bagg, A., Levine, B. L., June, C. H., and Stadtmauer, E. A. (2015). **Chimeric Antigen Receptor T Cells against CD19 for Multiple Myeloma**. *The New England journal of medicine*, 373:1040–1047. doi:10.1056/NEJMoa1504542.

[70] Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). **affy–analysis of Affymetrix GeneChip data at the probe level**. *Bioinformatics*, 20:307–315. doi:10.1093/bioinformatics/btg405.

[71] Gavriilidis, P., Giakoustidis, A., and Giakoustidis, D. (2015). **Aurora Kinases and Potential Medical Applications of Aurora Kinase Inhibitors: A Review**. *Journal of clinical medicine research*, 7:742–51. doi:10.14740/jocmr2295w.

[72] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). **Bioconductor: open software development for computational biology and bioinformatics**. *Genome biology*, 5:R80. doi:10.1186/gb-2004-5-10-r80.

[73] Gerds, T. A., Cai, T., and Schumacher, M. (2008). **The performance of risk prediction models**. *Biometrical journal*, 50:457–479. doi:10.1002/bimj.200810443.

[74] Gerecke, C., Fuhrmann, S., Strifler, S., Schmidt-Hieber, M., Einsele, H., and Knop, S. (2016). **The Diagnosis and Treatment of Multiple Myeloma**. *Deutsches Arzteblatt international*, 113:470–476. doi:10.3238/arztebl.2016.0470.

[75] Ghobrial, I. M., Badros, A. Z., Vredenburgh, J. J., Matous, J., Caola, A. M., Savell, A., Henrick, P., Paba-Prada, C. E., Schlossman, R. L., Laubach, J. P., Rosenblatt, J., Yee, A., Wisch, J. S., Farber, C. M., Maegawa, R. O., Usmani, S. Z., Cappuccio, J., Rivotto, B., Noonan, K., Reyes, K., Munshi, N. C., Anderson, K. C., and Richardson, P. (2016). **Phase II Trial of Combination of Elotuzumab, Lenalidomide, and Dexamethasone in High-Risk Smoldering Multiple Myeloma**. *Blood*, 128:976. doi:10.1182/blood.V128.22.976.976, ASH Annual meeting Abstract.

[76] Giannopoulos, K., Li, L., Bojarska-Junak, A., Rolinski, J., Dmoszynska, A., Hus, I., Greiner, J., Renner, C., Döhner, H., and Schmitt, M. (2006). **Expression of RHAMM/CD168 and other tumor-associated antigens in patients with B-cell chronic lymphocytic leukemia**. *International journal of oncology*, 29:95–103. doi:10.3892/ijo.29.1.95.

[77] Goldschmidt, H., Sonneveld, P., Cremer, F. W., van der Holt, B., Westveer, P., Breitkreutz, I., Benner, A., Glasmacher, A., Schmidt-Wolf, I. G. D., Martin, H., Hoelzer, D., Ho, A. D., and Lokhorst, H. M. (2003). **Joint HOVON-50/GMMG-HD3 randomized trial on the effect of thalidomide as part of a high-dose therapy regimen and as maintenance treatment for newly diagnosed myeloma patients**. *Annals of hematology*, 82:654–659. doi:10.1007/s00277-003-0685-2.

[78] Gordon, M. and Lumley, T. (2017). **forestplot Advanced Forest Plot Using 'grid' Graphics**, version 1.7.2. URL: `https://cran.r-project.org/src/contrib/Archive/forestplot/forestplot_1.7.2.tar.gz`. [Last visited: 17.10.2019].

[79] Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). **Assessment and comparison of prognostic classification schemes for survival data**. *Statistics in medicine*, 18:2529–2545.

[80] Grambsch, P. M. and Therneau, T. M. (1994). **Proportional Hazards Tests and Diagnostics Based on Weighted Residuals**. *Biometrika*, 81:515. doi:10.2307/2337123.

[81] Greiner, J., Li, L., Ringhoffer, M., Barth, T. F. E., Giannopoulos, K., Guillaume, P., Ritter, G., Wiesneth, M., Döhner, H., and Schmitt, M. (2005). **Identification and characterization of epitopes of the receptor for hyaluronic acid-mediated motility (RHAMM/CD168) recognized by**

CD8+ T cells of HLA-A2-positive patients with acute myeloid leukemia. *Blood*, 106:938–945. doi:10.1182/blood-2004-12-4787.

[82] Greiner, J., Schmitt, M., Li, L., Giannopoulos, K., Bosch, K., Schmitt, A., Dohner, K., Schlenk, R. F., Pollack, J. R., Dohner, H., and Bullinger, L. (2006). **Expression of tumor-associated antigens in acute myeloid leukemia Implications for specific immunotherapeutic approaches**. *Blood*, 108:4109–4117. doi:10.1182/blood-2006-01-023127.

[83] Greipp, P. R., Lust, J. A., O'Fallon, W. M., Katzmann, J. A., Witzig, T. E., and Kyle, R. A. (1993). **Plasma cell labeling index and beta 2-microglobulin predict survival independent of thymidine kinase and C-reactive protein in multiple myeloma**. *Blood*, 81:3382–3387.

[84] Greipp, P. R., San Miguel, J., Durie, B. G. M., Crowley, J. J., Barlogie, B., Bladé, J., Boccadoro, M., Child, J. A., Avet-Loiseau, H., Harousseau, J.-L., Kyle, R. A., Lahuerta, J. J., Ludwig, H., Morgan, G., Powles, R., Shimizu, K., Shustik, C., Sonneveld, P., Tosi, P., Turesson, I., and Westin, J. (2005). **International staging system for multiple myeloma**. *Journal of Clinical Oncology*, 23:3412–3420. doi:10.1200/JCO.2005.04.242.

[85] Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., and Shyr, Y. (2017). **Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis**. *Genomics*, 109:83–90. doi:10.1016/j.ygeno.2017.01.005.

[86] Hallek, M., Bergsagel, P. L., and Anderson, K. C. (1998). **Multiple myeloma Increasing evidence for a multistep transformation process**. *Blood*, 91:3–21.

[87] Harrell, JR, F. E. (2018). **rms Regression Modeling Strategies**, version 5.1-2. URL: `https://cran.r-project.org/src/contrib/Archive/rms/rms_5.1-2.tar.gz`. [Last visited: 17.10.2019].

[88] Harrington, D. P. and Fleming, T. R. (1982). **A Class of Rank Test Procedures for Censored Survival Data**. *Biometrika*, 69:553. doi:10.2307/2335991.

[89] Hartigan, J. A. and Wong, M. A. (1979). **Algorithm AS 136: A K-Means Clustering Algorithm**. *Applied Statistics*, 28:100–108.

[90] Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G. (2014). **pamr: Pam: prediction analysis for microarrays**, version 1.55. URL: `https://cran.r-project.org/src/contrib/Archive/pamr/pamr_1.55.tar.gz`. [Last visited: 17.10.2019].

[91] Hauschild, A., Grob, J.-J., Demidov, L. V., Jouary, T., Gutzmer, R., Millward, M., Rutkowski, P., Blank, C. U., Miller, W. H., Kaempgen, E., Martín-Algarra, S., Karaszewska, B., Mauch, C., Chiarion-Sileni, V., Martin, A.-M., Swann, S., Haney, P., Mirakhur, B., Guckert, M. E., Goodman, V., and Chapman, P. B. (2012). **Dabrafenib in BRAF-mutated metastatic melanoma A multicentre, open-label, phase 3 randomised controlled trial**. *The Lancet*, 380:358–365. doi:10.1016/S0140-6736(12)60868-X.

[92] Hielscher, T., Zucknick, M., Werft, W., and Benner, A. (2010). **On the prognostic value of survival models with application to gene expression signatures**. *Statistics in medicine*, 29:818–829. doi:10.1002/sim.3768.

[93] Hillengass, J., Hose, D., Raab, M. S., Bertsch, U., Hegenbart, U., Bärtsch, M. A., Mai, E. K., and Goldschmidt, H. (2017). **Patienten-Handbuch Multiples Myelom 2017**. *Sektion Multiples Myleom*. URL: `https://www.klinikum.uni-heidelberg.de/fileadmin/medizinische_klinik/Abteilung_5/docs/Veranstaltungen/MM_Tage_Patientenhandbuch/Finales_Patientenhandbuch_mit_Cover.pdf`. [Last visited: 28.10.2019].

[94] Hollander, M., D. A. Wolfe (1973). **Nonparametric statistical methods**. *New York: John Wiley & Sons*, pages 185–194.

[95] Hope, A. C. A. (1968). **A Simplified Monte Carlo Significance Test Procedure**. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30:582–598.

[96] Hose, D. (2014). **Experimental and conceptual basis for personalized and risk-adapted treatment of multiple myeloma**. Habilitation treatise, Ruprecht-Karls-Universität, Heidelberg.

[97] Hose, D. (2015). **Asymptomatic multiple myeloma - molecular background of progression, evolution, and prognosis**. Inaugural Dissertation, Justus-Liebig-Universität, Gießen. URL: `http://geb.uni-giessen.de/geb/volltexte/2015/11674/pdf/HoseDirk_2015_08_18.pdf`. [Last visited: 17.10.2019].

[98] Hose, D. (2019). **Symposium 10: Targeting B-cell maturation antigen with T-cell bispecific antibodies for the treatment of malignant plasma cell dyscrasias - preclinical and clinical development of EM801**. 81th Annual Meeting of the Japanese Society of Hematology, Tokyo, Invited lecture.

[99] Hose, D., Beck, S., Salwender, H., Emde, M., Bertsch, U., Kunz, C., Scheid, C., Hänel, M., Weisel, K., Hielscher, T., Raab, M. S., Goldschmidt, H., Jauch, A., Moreaux, J., and Seckinger, A. (2019). **Prospective target assessment and multimodal prediction of survival for personalized and risk-adapted treatment strategies in multiple myeloma in the GMMG-MM5 multicenter trial**. *Journal of hematology & oncology*, 12:65. doi:10.1186/s13045-019-0750-5.

[100] Hose, D., Moreaux, J., Meissner, T., Seckinger, A., Goldschmidt, H., Benner, A., Mahtouk, K., Hillengass, J., Rème, T., de Vos, J., Hundemer, M., Condomines, M., Bertsch, U., Rossi, J.-F., Jauch, A., Klein, B., and Möhler, T. (2009). **Induction of angiogenesis by normal and malignant plasma cells**. *Blood*, 114:128–143. doi:10.1182/blood-2008-10-184226.

[101] Hose, D., Rème, T., Hielscher, T., Moreaux, J., Messner, T., Seckinger, A., Benner, A., Shaughnessy, Jr, John D, Barlogie, B., Zhou, Y., Hillengass, J., Bertsch, U., Neben, K., Möhler, T., Rossi, J. F., Jauch, A., Klein, B., and Goldschmidt, H. (2011). **Proliferation is a central independent prognostic factor and target for personalized and risk-adapted treatment in multiple myeloma**. *Haematologica*, 96:87–95. doi:10.3324/haematol.2010.030296.

[102] Hose, D., Rème, T., Meissner, T., Moreaux, J., Seckinger, A., Lewis, J., Benes, V., Benner, A., Hundemer, M., Hielscher, T., Shaughnessy, Jr, John D, Barlogie, B., Neben, K., Krämer, A., Hillengass, J., Bertsch, U., Jauch, A., de Vos, J., Rossi, J.-F., Möhler, T., Blake, J., Zimmermann, J., Klein, B., and Goldschmidt, H. (2009). **Inhibition of aurora kinases for tailored risk-adapted treatment of multiple myeloma**. *Blood*, 113:4331–4340. doi:10.1182/blood-2008-09-178350.

[103] Hose, D. and Seckinger, A. (2014). **Biologie des multiplen Myeloms**. *Der Onkologe*, 20:208–216. doi:10.1007/s00761-013-2568-z.

[104] Hothorn, T. and Lausen, B. (2003). **On the Exact Distribution of Maximally Selected Rank Statistics**. *Computational Statistics & Data Analysis*, 43:121–137.

[105] Hothorn, T. and Lausen, B. (2017). **maxstat: Maximally Selected Rank Statistics**. R, version 0.7-25. URL: `https://cran.r-project.org/web/packages/maxstat/index.html`. [Last visited: 17.10.2019].

[106] Howlader, N., Noone, A. M., Krapcho, M., Miller, D., Brest, A., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, Chen, H. S., Feuer, E. J., and Cronin, K. A., editors (2019). **SEER Cancer Statistics Review, 1975-2016, National Cancer Institute**. National Cancer Institute. URL: `https://seer.cancer.gov/csr/1975_2016/`. [Last visited: 10.09.2019, based on November 2018 SEER data submission, posted to the SEER web site, April 2019.].

[107] Huang, X., Di Liberto, M., Jayabalan, D., Liang, J., Ely, S., Bretz, J., Shaffer, Arthur L., I., Louie, T., Chen, I., Randolph, S., Hahn, W. C., Staudt, L. M., Niesvizky, R., Moore, M. A. S., and Chen-Kiang, S. (2012). **Prolonged early G1 arrest by selective CDK4/CDK6 inhibition sensitizes myeloma cells to cytotoxic killing through cell cycleâĂŞcoupled loss of IRF4**. *Blood*, 120:1095–1106. doi:10.1182/blood-2012-03-415984.

[108] Hundemer, M., Schmidt, S., Condomines, M., Lupu, A., Hose, D., Moos, M., Cremer, F., Kleist, C., Terness, P., Belle, S., Ho, A. D., Goldschmidt, H., Klein, B., and Christensen, O. (2006). **Identification of a new HLA-A2-restricted T-cell epitope within HM1.24 as immunotherapy target for multiple myeloma**. *Experimental hematology*, 34:486–496. doi:10.1016/j.exphem.2006.01.008.

[109] Hyman, D. M., Puzanov, I., Subbiah, V., Faris, J. E., Chau, I., Blay, J.-Y., Wolf, J., Raje, N. S., Diamond, E. L., Hollebecque, A., Gervais, R., Elez-Fernandez, M. E., Italiano, A., Hofheinz, R.-D., Hidalgo, M., Chan, E., Schuler, M., Lasserre, S. F., Makrutzki, M., Sirzen, F., Veronese, M. L., Tabernero, J., and Baselga, J. (2015). **Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations**. *New England Journal of Medicine*, 373:726–736. doi:10.1056/NEJMoa1502309.

[110] Hyndman, R. J. and Fan, Y. (1996). **Sample Quantiles in Statistical Packages**. *The American Statistician*, 50:361–365. doi:10.2307/2684934.

[111] International Myeloma Working Group (2003). **Criteria for the classification of monoclonal gammopathies, multiple myeloma and related disorders: a report of the International Myeloma Working Group**. *British journal of haematology*, 121:749–757. doi:10.1046/j.1365-2141.2003.04355.x.

[112] Jain, S., Pathak, K., and Vaidya, A. (2018). **Molecular therapy using siRNA Recent trends and advances of multi target inhibition of cancer growth**. *International journal of biological macromolecules*, 116:880–892. doi:10.1016/j.ijbiomac.2018.05.077.

[113] Jalili, A., Ozaki, S., Hara, T., Shibata, H., Hashimoto, T., Abe, M., Nishioka, Y., and Matsumoto, T. (2005). **Induction of HM1.24 peptide-specific cytotoxic T lymphocytes by using**

peripheral-blood stem-cell harvests in patients with multiple myeloma. *Blood*, 106:3538–3545. doi:10.1182/blood-2005-04-1438.

[114] Jonckheere, A. R. (1954). **A Distribution-Free k-Sample Test Against Ordered Alternatives**. *Biometrika*, 41:133. doi:10.2307/2333011.

[115] Jourdan, M., Caraux, A., De Vos, J., Fiol, G., Larroque, M., Cognot, C., Bret, C., Duperray, C., Hose, D., and Klein, B. (2009). **An in vitro model of differentiation of memory B cells into plasmablasts and plasma cells including detailed phenotypic and molecular characterization**. *Blood*, 114:5173–5181. doi:10.1182/blood-2009-07-235960.

[116] Kaatsch, P., Spix, C., Katalinic, A., Hentschel, S., Luttmann, S., Stegmaier, C., Waldeyer-Sauerland, M., Waldmann, A., Caspritz, S., Christ, M., Ernst, A., Folkerts, J., Hansmann, J., Klein, S., Kranzhöfer, K., Kunz, B., Manegold, K., Penzkofer, A., Treml, K., Weg-Remers, S., Wittenberg, K., Barnes, B., Bertz, J., Buttmann-Schweiger, N., Dahm, S., Fiebig, J., Haberland, J., Kraywinkel, K., Wienecke, A., Wolf, U., Meisegeier, S., Franke, M., and Werth, K. (2017). **Krebs in Deutschland für 2013/2014**. Robert-Koch-Institut and Gesellschaft der epidemiologischen Krebsregister in Deutschland, 11. edition.

[117] Kaplan, E. L. and Meier, P. (1958). **Nonparametric Estimation from Incomplete Observations**. *Journal of the American Statistical Association*, 53:457. doi:10.2307/2281868.

[118] Kariyawasan, C. C., Hughes, D. A., Jayatillake, M. M., and Mehta, A. B. (2007). **Multiple myeloma Causes and consequences of delay in diagnosis**. *QJM : monthly journal of the Association of Physicians*, 100:635–640. doi:10.1093/qjmed/hcm077.

[119] Kent, W. J. (2002). **BLAT–the BLAST-like alignment tool**. *Genome research*, 12:656–664. doi:10.1101/gr.229202.

[120] Klebanoff, C. A., Rosenberg, S. A., and Restifo, N. P. (2016). **Prospects for gene-engineered T cell immunotherapy for solid cancers**. *Nature medicine*, 22:26–36. doi:10.1038/nm.4015.

[121] Klein, B., Seckinger, A., Moehler, T., and Hose, D. (2011). **Molecular pathogenesis of multiple myeloma: chromosomal aberrations, changes in gene expression, cytokine networks, and the bone marrow microenvironment**. In: *Multiple Myeloma, Recent Results in Cancer Research*, editors Moehler, T. and Goldschmidt, H., volume 183, pages 39–86, Berlin Heidelberg. Springer-Verlag. doi:10.1007/978-3-540-85772-3_3.

[122] Korde, N., Roschewski, M., Zingone, A., Kwok, M., Manasanch, E. E., Bhutani, M., Tageja, N., Kazandjian, D., Mailankody, S., Wu, P., Morrison, C., Costello, R., Zhang, Y., Burton, D., Mulquin, M., Zuchlinski, D., Lamping, L., Carpenter, A., Wall, Y., Carter, G., Cunningham, S. C., Gounden, V., Sissung, T. M., Peer, C., Maric, I., Calvo, K. R., Braylan, R., Yuan, C., Stetler-Stevenson, M., Arthur, D. C., Kong, K. A., Weng, L., Faham, M., Lindenberg, L., Kurdziel, K., Choyke, P., Steinberg, S. M., Figg, W., and Landgren, O. (2015). **Treatment With Carfilzomib-Lenalidomide-Dexamethasone With Lenalidomide Extension in Patients With Smoldering or Newly Diagnosed Multiple Myeloma**. *JAMA oncology*, 1:746–754. doi:10.1001/jamaoncol.2015.2010.

[123] Kostka, D. (2008). **docval Documenting microarray preprocessing 'by value'**, version 1.1.2. URL: `https://storage.googleapis.com/google-code-archive-downloads/v2/code.google.com/gep-r/docval_1.1.2_gcrma.tar.gz`. [Last visited: 17.10.2019].

[124] Kuiper, R., Broyl, A., de Knegt, Y., van Vliet, M. H., van Beers, E. H., van der Holt, B., el Jarari, L., Mulligan, G., Gregory, W., Morgan, G., Goldschmidt, H., Lokhorst, H. M., van Duin, M., and Sonneveld, P. (2012). **A gene expression signature for high-risk multiple myeloma**. *Leukemia*, 26:2406–2413. doi:10.1038/leu.2012.127.

[125] Kumar, S., Dispenzieri, A., Lacy, M. Q., Hayman, S. R., Buadi, F. K., Colby, C., Laumann, K., Zeldenrust, S. R., Leung, N., Dingli, D., Greipp, P. R., Lust, J. A., Russell, S. J., Kyle, R. A., Rajkumar, S. V., and Gertz, M. A. (2012). **Revised prognostic staging system for light chain amyloidosis incorporating cardiac biomarkers and serum free light chain measurements**. *Journal of clinical oncology*, 30:989–995. doi:10.1200/JCO.2011.38.5724.

[126] Kurashige, T., Noguchi, Y., Saika, T., Ono, T., Nagata, Y., Jungbluth, A., Ritter, G., Chen, Y. T., Stockert, E., Tsushima, T., Kumon, H., Old, L. J., and Nakayama, E. (2001). **Ny-ESO-1 expression and immunogenicity associated with transitional cell carcinoma Correlation with tumor grade**. *Cancer Research*, 61:4671–4674.

[127] Kyle, R. A. and Greipp, P. R. (1980). **Smoldering multiple myeloma**. *The New England journal of medicine*, 302:1347–1349. doi:10.1056/NEJM198006123022405.

[128] Kyle, R. A. and Rajkumar, S. V. (2004). **Multiple Myeloma**. *The New England journal of medicine*, 351:1860–1873. doi:10.1056/NEJMra041875.

[129] Kyle, R. A., Remstein, E. D., Therneau, T. M., Dispenzieri, A., Kurtin, P. J., Hodnefield, J. M., Larson, D. R., Plevak, M. F., Jelinek, D. F., Fonseca, R., Melton, L. J., and Rajkumar, S. V. (2007). **Clinical course and prognosis of smoldering (asymptomatic) multiple myeloma**. *The New England journal of medicine*, 356:2582–2590. doi:10.1056/NEJMoa070389.

[130] Kyle, R. A., Therneau, T. M., Rajkumar, S. V., Larson, D. R., Plevak, M. F., Offord, J. R., Dispenzieri, A., Katzmann, J. A., and Melton, L. J. (2006). **Prevalence of monoclonal gammopathy of undetermined significance**. *The New England journal of medicine*, 354:1362–1369. doi:10.1056/NEJMoa054494.

[131] Kyle, R. A., Therneau, T. M., Rajkumar, S. V., Offord, J. R., Larson, D. R., Plevak, M. F., and Melton, L. J. (2002). **A long-term study of prognosis in monoclonal gammopathy of undetermined significance**. *The New England journal of medicine*, 346:564–569. doi:10.1056/NEJMoa01133202.

[132] Landgren, O., Roschewski, M., Mailankody, S., Kwok, M., Manasanch, E. E., Bhutani, M., Tageja, N., Kazandjian, D., Zingone, A., Costello, R., Burton, D., Zhang, Y., Wu, P., Carter, G., Mulquin, M., Zuchlinski, D., Carpenter, A., Gounden, V., Morrison, C., Maric, I., Calvo, K. R., Braylan, R. C., Yuan, C., Stetler-Stevenson, M., Arthur, D. C., Lindenberg, L., Karen, K., Choyke, P., Steinberg, S. M., Figg, W. D., and Korde, N. (2014). **Carfilzomib, Lenalidomide, and Dexamethasone in High-Risk Smoldering Multiple Myeloma Final Results from the NCI Phase 2 Pilot Study**. *Blood*, 124:4746, ASH Annual Meeting Abstract.

[133] Langmead, B. and Salzberg, S. L. (2012). **Fast gapped-read alignment with Bowtie 2**. *Nature methods*, 9:357–359. doi:10.1038/nmeth.1923.

[134] Lanser, A. (2015). **Shrinkage Methods on Cox PH with Too Few Events**. Master thesis, Vanderbilt University, Nashville. URL: `https://pdfs.semanticscholar.org/ac7e/de4cf0aff82f13f12e175dd6f1c48e8c0672.pdf`. [Last visited: 19.09.2019].

[135] Larson, D. and Abbott, T. (2016). **bam-readcount**. The McDonnell Genome Institute, version 0.7.4. URL: `https://github.com/genome/bam-readcount`. [Last visited: 08.01.18].

[136] Lausen, B., Hothorn, T., Bretz, F., and Schumacher, M. (2004). **Assessment of Optimal Selected Prognostic Factors**. *Biometrical journal*, 46:364–374. doi:10.1002/bimj.200310030.

[137] Lausen, B. and Schumacher, M. (1992). **Maximally Selected Rank Statistics**. *Biometrics*, 48:73–85.

[138] Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., and Carey, V. (2013). **Software for Computing and Annotating Genomic Ranges**. *PLoS Computational Biology*, 9:1–10. doi:10.1371/journal.pcbi.1003118.

[139] Lehners, N., Becker, N., Benner, A., Pritsch, M., Löpprich, M., Mai, E. K., Hillengass, J., Goldschmidt, H., and Raab, M.-S. (2018). **Analysis of long-term survival in multiple myeloma after first-line autologous stem cell transplantation Impact of clinical risk factors and sustained response**. *Cancer medicine*, 7:307–316. doi:10.1002/cam4.1283.

[140] Li, B. and Dewey, C. N. (2011). **RSEM Accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC Bioinformatics*, 12:323. doi:10.1186/1471-2105-12-323.

[141] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics*, 25:2078–2079. doi:10.1093/bioinformatics/btp352.

[142] Li, Q., Birkbak, N. J., Gyorffy, B., Szallasi, Z., and Eklund, A. C. (2011). **Jetset Selecting the optimal microarray probe set to represent a gene**. *BMC Bioinformatics*, 12:474. doi:10.1186/1471-2105-12-474.

[143] Ligtenberg, W. (2016). **reactome.db A set of annotation maps for reactome**, version 1.58.0. URL: `http://bioconductor.org/packages/3.4/data/annotation/html/reactome.db.html`. [Last visited: 17.10.2019].

[144] Liu, Y., Zhou, J., and White, K. P. (2014). **RNA-seq differential expression studies More sequence or more replication?** *Bioinformatics*, 30:301–304. doi:10.1093/bioinformatics/btt688.

[145] Lloyd, S. (1982). **Least squares quantization in PCM**. *IEEE Transactions on Information Theory*, 28:129–137. doi:10.1109/TIT.1982.1056489.

[146] Lohr, J. G., Stojanov, P., Carter, S. L., Cruz-Gordillo, P., Lawrence, M. S., Auclair, D., Sougnez, C., Knoechel, B., Gould, J., Saksena, G., Cibulskis, K., McKenna, A., Chapman, M. A., Straussman, R., Levy, J., Perkins, L. M., Keats, J. J., Schumacher, S. E., Rosenberg, M., Multiple

Myeloma Research Consortium, Getz, G., and Golub, T. R. (2014). **Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy**. *Cancer Cell*, 25:91–101. doi:10.1016/j.ccr.2013.12.015.

[147] Lokhorst, H. M., Plesner, T., Laubach, J. P., Nahi, H., Gimsing, P., Hansson, M., Minnema, M. C., Lassen, U., Krejcik, J., Palumbo, A., van de Donk, N. W. C. J., Ahmadi, T., Khan, I., Uhlar, C. M., Wang, J., Sasser, A. K., Losic, N., Lisby, S., Basse, L., Brun, N., and Richardson, P. G. (2015). **Targeting CD38 with Daratumumab Monotherapy in Multiple Myeloma**. *The New England journal of medicine*, 373:1207–1219. doi:10.1056/NEJMoa1506348.

[148] Lonial, S., Dimopoulos, M., Palumbo, A., White, D., Grosicki, S., Spicka, I., Walter-Croneck, A., Moreau, P., Mateos, M.-V., Magen, H., Belch, A., Reece, D., Beksac, M., Spencer, A., Oakervee, H., Orlowski, R. Z., Taniwaki, M., Röllig, C., Einsele, H., Wu, K. L., Singhal, A., San-Miguel, J., Matsumoto, M., Katz, J., Bleickardt, E., Poulart, V., Anderson, K. C., and Richardson, P. (2015). **Elotuzumab Therapy for Relapsed or Refractory Multiple Myeloma**. *The New England journal of medicine*, 373:621–631. doi:10.1056/NEJMoa1505654.

[149] MacQueen, J. (1967). **Some methods for classification and analysis of multivariate observations**. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

[150] Mai, E. K., Bertsch, U., Dürig, J., Kunz, C., Haenel, M., Blau, I. W., Munder, M., Jauch, A., Schurich, B., Hielscher, T., Merz, M., Huegle-Doerr, B., Seckinger, A., Hose, D., Hillengass, J., Raab, M. S., Neben, K., Lindemann, H.-W., Zeis, M., Gerecke, C., Schmidt-Wolf, I. G. H., Weisel, K., Scheid, C., Salwender, H., and Goldschmidt, H. (2015). **Phase III trial of bortezomib, cyclophosphamide and dexamethasone (VCD) versus bortezomib, doxorubicin and dexamethasone (PAd) in newly diagnosed myeloma**. *Leukemia*, 29:1721–1729. doi:10.1038/leu.2015.80.

[151] Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). **Multivariate analysis**. Probability and mathematical statistics. Academic Press.

[152] Mateos, M.-V., Hernández, M.-T., Giraldo, P., de La Rubia, J., de Arriba, F., Corral, L. L., Rosiñol, L., Paiva, B., Palomera, L., Bargay, J., Oriol, A., Prosper, F., López, J., Olavarría, E., Quintana, N., García, J.-L., Bladé, J., Lahuerta, J.-J., and San Miguel, J.-F. (2013). **Lenalidomide plus Dexamethasone for High-Risk Smoldering Multiple Myeloma**. *The New England journal of medicine*, 369:438–447. doi:10.1056/NEJMoa1300439.

[153] Mateos, M.-V. and San Miguel, J.-F. (2015). **Smoldering multiple myeloma When to observe and when to treat?** *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting*, pages e484–92. doi:10.14694/EdBook_AM.2015.35.e484.

[154] McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). **Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation**. *Nucleic acids research*, 40:4288–4297. doi:10.1093/nar/gks042.

[155] Meißner, T., Seckinger, A., Hemminki, K., Bertsch, U., Foersti, A., Haenel, M., Duering, J., Salwender, H., Goldschmidt, H., Morgan, G. J., Hose, D., and Weinhold, N. (2015). **Profound impact of sample processing delay on gene expression of multiple myeloma plasma cells**. *BMC medical genomics*, 8:85. doi:10.1186/s12920-015-0161-6.

[156] Meissner, T., Seckinger, A., Rème, T., Hielscher, T., Möhler, T., Neben, K., Goldschmidt, H., Klein, B., and Hose, D. (2011). **Gene expression profiling in multiple myeloma–reporting of entities, risk, and targets in clinical routine**. *Clinical cancer research*, 17:7240–7247. doi:10.1158/1078-0432.CCR-11-1628.

[157] Merz, M., Salwender, H., Haenel, M., Mai, E. K., Bertsch, U., Kunz, C., Hielscher, T., Blau, I. W., Scheid, C., Hose, D., Seckinger, A., Jauch, A., Hillengass, J., Raab, M. S., Schurich, B., Munder, M., Schmidt-Wolf, I. G. H., Gerecke, C., Lindemann, H.-W., Zeis, M., Weisel, K., Duerig, J., and Goldschmidt, H. (2015). **Subcutaneous versus intravenous bortezomib in two different induction therapies for newly diagnosed multiple myeloma An interim analysis from the prospective GMMG-MM5 trial**. *Haematologica*, 100:964–969. doi:10.3324/haematol.2015.124347.

[158] Metzker, M. L. (2010). **Sequencing technologies - the next generation**. *Nature reviews genetics*, 11:31–46. doi:10.1038/nrg2626.

[159] Mikhael, J. R., Dingli, D., Roy, V., Reeder, C. B., Buadi, F. K., Hayman, S. R., Dispenzieri, A., Fonseca, R., Sher, T., Kyle, R. A., Lin, Y., Russell, S. J., Kumar, S., Bergsagel, P. L., Zeldenrust, S. R., Leung, N., Drake, M. T., Kapoor, P., Ansell, S. M., Witzig, T. E., Lust, J. A., Dalton, R. J., Gertz, M. A., Stewart, A. K., Stewart, K., Rajkumar, S. V., Chanan-Khan, A., and Lacy, M. Q. (2013). **Management of newly diagnosed symptomatic multiple myeloma Updated Mayo Stratification of Myeloma and Risk-Adapted Therapy (mSMART) consensus guidelines 2013**. *Mayo Clinic proceedings*, 88:360–376. doi:10.1016/j.mayocp.2013.01.019.

[160] Mills, F. C., Harindranath, N., Mitchell, M., and Max, E. E. (1997). **Enhancer complexes located downstream of both human immunoglobulin Calpha genes**. *The Journal of experimental medicine*, 186:845–858. doi:10.1084/jem.186.6.845.

[161] Mogensen, U. B., Ishwaran, H., and Gerds, T. A. (2012). **Evaluating Random Forests for Survival Analysis Using Prediction Error Curves**. *Journal of Statistical Software*, 50:1–23. doi:10.18637/jss.v050.i11.

[162] Moreau, P., Cavallo, F., Leleu, X., Hulin, C.and Amiot, M., Descamps, G., Facon, T., Boccadoro, M., Mignard, D., and Harousseau, J. L. (2011). **Phase I study of the anti insulin-like growth factor 1 receptor (IGF-1R) monoclonal antibody, AVE1642, as single agent and in combination with bortezomib in patients with relapsed multiple myeloma**. *Leukemia*, 25:872–874. doi:10.1038/leu.2011.4.

[163] Moreau, P., Hulin, C., Facon, T., Boccadoro, M., Mery-Mignard, D., Deslandes, A., and Harousseau, J.-L. (2007). **Phase I Study of AVE1642 Anti IGF-1R Monoclonal Antibody in Patients with Advanced Multiple Myeloma.** *Blood*, 110:1166–1166. doi:10.1182/blood.V110.11.1166.1166, ASH Annual Meeting Abstract.

[164] Moreaux, J., Cremer, F. W., Reme, T., Raab, M., Mahtouk, K., Kaukel, P., Pantesco, V., de Vos, J., Jourdan, E., Jauch, A., Legouffe, E., Moos, M., Fiol, G., Goldschmidt, H., Rossi, J. F., Hose, D., and Klein, B. (2005). **The level of TACI gene expression in myeloma cells is associated with a signature of microenvironment dependence versus a plasmablastic signature**. *Blood*, 106:1021–1030. doi:10.1182/blood-2004-11-4512.

[165] Moreaux, J., Klein, B., Bataille, R., Descamps, G., Maïga, S., Hose, D., Goldschmidt, H., Jauch, A., Rème, T., Jourdan, M., Amiot, M., and Pellat-Deceunynck, C. (2011). **A high-risk signature for patients with multiple myeloma established from the molecular classification of human myeloma cell lines**. *Haematologica*, 96:574–582. doi:10.3324/haematol.2010.033456.

[166] Moreaux, J., Legouffe, E., Jourdan, E., Quittet, P., Rème, T., Lugagne, C., Moine, P., Rossi, J.-F., Klein, B., and Tarte, K. (2004). **BAFF and APRIL protect myeloma cells from apoptosis induced by interleukin 6 deprivation and dexamethasone**. *Blood*, 103:3148–3157. doi:10.1182/blood-2003-06-1984.

[167] Morgan, G. J., Walker, B. A., and Davies, F. E. (2012). **The genetic architecture of multiple myeloma**. *Nature reviews cancer*, 12:335–348. doi:10.1038/nrc3257.

[168] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nature methods*, 5:621–628. doi:10.1038/nmeth.1226.

[169] Multiple Myeloma Research Foundation Personalized Medicine Initiatives (2018). **CoMMpass IA13 relating Clinical outcomes in Multiple Myeloma to Personal Assessment of Genetic Profile**. URL: `https://research.themmrf.org/`. [Last visited: 20.05.2019].

[170] Murphy, K. M. and Weaver, C. (2017). **Janeway's immunobiology**. GS Garland Science Taylor & Francis Group, 9. edition.

[171] Murrell, P. (2005). **R Graphics**. Chapman & Hall/CRC Press.

[172] Musgrove, E. A., Caldon, C. E., Barraclough, J., Stone, A., and Sutherland, R. L. (2011). **Cyclin D as a therapeutic target in cancer**. *Nature reviews cancer*, 11:558–572. doi:10.1038/nrc3090.

[173] Nagelkerke, N. J. D. (1991). **A note on a general definition of the coefficient of determination**. *Biometrika*, 78:691–692. doi:10.1093/biomet/78.3.691.

[174] Neben, K., Jauch, A., Bertsch, U., Heiss, C., Hielscher, T., Seckinger, A., Mors, T., Müller, N. Z., Hillengass, J., Raab, M. S., Ho, A. D., Hose, D., and Goldschmidt, H. (2010). **Combining information regarding chromosomal aberrations t(4;14) and del(17p13) with the International Staging System classification allows stratification of myeloma patients undergoing autologous stem cell transplantation**. *Haematologica*, 95:1150–1157. doi:10.3324/haematol.2009.016436.

[175] Neben, K., Jauch, A., Hielscher, T., Hillengass, J., Lehners, N., Seckinger, A., Granzow, M., Raab, M. S., Ho, A. D., Goldschmidt, H., and Hose, D. (2013). **Progression in smoldering myeloma is independently determined by the chromosomal abnormalities del(17p), t(4;14), gain 1q, hyperdiploidy, and tumor load**. *Journal of clinical oncology*, 31:4325–4332. doi:10.1200/JCO.2012.48.4923.

[176] Neben, K., Lokhorst, H. M., Jauch, A., Bertsch, U., Hielscher, T., van der Holt, B., Salwender, H., Blau, I. W., Weisel, K., Pfreundschuh, M., Scheid, C., Dührsen, U., Lindemann, W., Schmidt-Wolf, I. G. H., Peter, N., Teschendorf, C., Martin, H., Haenel, M., Derigs, H. G., Raab, M. S., Ho, A. D., van de Velde, H., Hose, D., Sonneveld, P., and Goldschmidt, H. (2012). **Administration of bortezomib before and after autologous stem cell transplantation improves outcome in multiple myeloma patients with deletion 17p**. *Blood*, 119:940–948. doi:10.1182/blood-2011-09-379164.

[177] Neuwirth, E. (2014). **RColorBrewer Color Brewer Palettes**, version 1.1.-2. URL: `https://cran.r-project.org/web/packages/RColorBrewer/index.html`. [Last visited: 17.10.2019].

[178] Niemeyer, P., Türeci, O., Eberle, T., Graf, N., Pfreundschuh, M., and Sahin, U. (2003). **Expression of serologically identified tumor antigens in acute leukemias**. *Leukemia research*, 27:655–660. doi:10.1016/S0145-2126(02)00230-8.

[179] Noble, W. S. (2009). **How does multiple testing correction work?** *Nature biotechnology*, 27:1135–1137. doi:10.1038/nbt1209-1135.

[180] O'Connor, B. P., Raman, V. S., Erickson, L. D., Cook, W. J., Weaver, L. K., Ahonen, C., Lin, L.-L., Mantchev, G. T., Bram, R. J., and Noelle, R. J. (2004). **BCMA is essential for the survival of long-lived bone marrow plasma cells**. *The Journal of experimental medicine*, 199:91–98. doi:10.1084/jem.20031330.

[181] Pagès, H., Carlson, M., Falcon, S., and Li, N. (2017). **AnnotationDbi Annotation Database Interface**, version 1.36.2. URL: `https://bioconductor.org/packages/3.4/bioc/html/AnnotationDbi.html`. [Last visited: 17.10.2019].

[182] Palumbo, A., Avet-Loiseau, H., Oliva, S., Lokhorst, H. M., Goldschmidt, H., Rosinol, L., Richardson, P., Caltagirone, S., Lahuerta, J. J., Facon, T., Bringhen, S., Gay, F., Attal, M., Passera, R., Spencer, A., Offidani, M., Kumar, S., Musto, P., Lonial, S., Petrucci, M. T., Orlowski, R. Z., Zamagni, E., Morgan, G., Dimopoulos, M. A., Durie, B. G. M., Anderson, K. C., Sonneveld, P., San Miguel, J., Cavo, M., Rajkumar, S. V., and Moreau, P. (2015). **Revised International Staging System for Multiple Myeloma A Report From International Myeloma Working Group**. *Journal of clinical oncology*, 33:2863–2869. doi:10.1200/JCO.2015.61.2267.

[183] Palumbo, A., Gay, F., Falco, P., Crippa, C., Montefusco, V., Patriarca, F., Rossini, F., Caltagirone, S., Benevolo, G., Pescosta, N., Guglielmelli, T., Bringhen, S., Offidani, M., Giuliani, N., Petrucci, M. T., Musto, P., Liberati, A. M., Rossi, G., Corradini, P., and Boccadoro, M. (2010). **Bortezomib as induction before autologous transplantation, followed by lenalidomide as consolidation-maintenance in untreated multiple myeloma patients**. *Journal of clinical oncology*, 28:800–807. doi:10.1200/JCO.2009.22.7561.

[184] Paterfield, W. M. (1981). **Algorithm AS 159: An Efficient Method of Generating Random R $\times$ C Tables with GivenRow and Column Totals**. *Applied Statistics*, 30:91–97.

[185] Pearson, K. (1901). **LIII. On lines and planes of closest fit to systems of points in space**. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2:559–572. doi:10.1080/14786440109462720.

[186] Pillarisetti, K., Edavettal, S., Mendonca, M., Li, Y., Tornetta, M., Babich, A., Majewski, N., Husovsky, M., Reeves, D., Walsh, E., Chin, D., Luistro, L., Joseph, J., Chu, G., Packman, K., Shetty, S., Elsayed, Y., Attar, R., and Gaudet, F. (2020). **A T-cell-redirecting bispecific G-protein-coupled receptor class 5 member D x CD3 antibody to treat multiple myeloma.** *Blood*, 9:1232–1243. doi:10.1182/blood.2019003342.

[187] Qiu, Y. (2019). **showtext Using Fonts More Easily in R Graphs**, version 0.5-1. URL: `https://cran.r-project.org/src/contrib/Archive/showtext/showtext_0.5-1.tar.gz`.

[188] R Core Team (2016). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, version 3.3.2. URL: `https://www.r-project.org/`. [Last visited: 15.10.18].

[189] Radbruch, A., Muehlinghaus, G., Luger, E. O., Inamine, A., Smith, K. G. C., Dörner, T., and Hiepe, F. (2006). **Competence and competition: the challenge of becoming a long-lived plasma cell**. *Nature reviews immunology*, 6:741–750. doi:10.1038/nri1886.

[190] Rajkumar, S. V., Dimopoulos, M. A., Palumbo, A., Blade, J., Merlini, G., Mateos, M.-V., Kumar, S., Hillengass, J., Kastritis, E., Richardson, P., Landgren, O., Paiva, B., Dispenzieri, A., Weiss, B., Leleu, X., Zweegman, S., Lonial, S., Rosinol, L., Zamagni, E., Jagannath, S., Sezer, O., Kristinsson, S. Y., Caers, J., Usmani, S. Z., Lahuerta, J. J., Johnsen, H. E., Beksac, M., Cavo, M., Goldschmidt, H., Terpos, E., Kyle, R. A., Anderson, K. C., Durie, B. G. M., and Miguel, J. F. S. (2014). **International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma**. *The Lancet Oncology*, 15:e538–e548. doi:10.1016/S1470-2045(14)70442-5.

[191] Rajkumar, S. V., Gupta, V., Fonseca, R., Dispenzieri, A., Gonsalves, W. I., Larson, D., Ketterling, R. P., Lust, J. A., Kyle, R. A., and Kumar, S. K. (2013). **Impact of primary molecular cytogenetic abnormalities and risk of progression in smoldering multiple myeloma**. *Leukemia*, 27:1738–1744. doi:10.1038/leu.2013.86.

[192] Rajkumar, S. V., Landgren, O., and Mateos, M.-V. (2015). **Smoldering multiple myeloma**. *Blood*, 125:3069–3075. doi:10.1182/blood-2014-09-568899.

[193] Rao, L., De Veirman, K., Giannico, D., Saltarella, I., Desantis, V., Antonia Frassanito, M., Giovanni Solimando, A., Ribatti, D., Prete, M., Harstrick, A., Fiedler, U., De Raeve, H., Racanelli, V., Vanderkerken, K., and Vacca, A. (2018). **Targeting angiogenesis in multiple myeloma by the VEGF and HGF blocking DARPin® protein MP0250: a preclinical study**. *Oncotarget*, 9:13366–13381. doi:https://doi.org/10.18632/oncotarget.24351.

[194] Rapoport, A. P., Stadtmauer, E. A., Binder-Scholl, G. K., Goloubeva, O., Vogl, D. T., Lacey, S. F., Badros, A. Z., Garfall, A., Weiss, B., Finklestein, J., Kulikovskaya, I., Sinha, S. K., Kronsberg, S., Gupta, M., Bond, S., Melchiori, L., Brewer, J. E., Bennett, A. D., Gerry, A. B., Pumphrey, N. J., Williams, D., Tayton-Martin, H. K., Ribeiro, L., Holdich, T., Yanovich, S., Hardy, N., Yared, J., Kerr, N., Philip, S., Westphal, S., Siegel, D. L., Levine, B. L., Jakobsen, B. K., Kalos, M., and June, C. H. (2015). **NY-ESO-1-specific TCR-engineered T cells mediate sustained antigen-specific antitumor effects in myeloma**. *Nature medicine*, 21:914–921. doi:10.1038/nm.3910.

[195] Ratner, B. (2009). **The correlation coefficient Its values range between +1/−1, or do they?** *Journal of Targeting, Measurement and Analysis for Marketing*, 17:139–142. doi:10.1057/jt.2009.5.

[196] Raykar, V. C., Steck, H., Krishnapuram, B., Dehing-Oberije, C., and Lambin, P. (2007). **On Ranking in Survival Analysis Bounds on the Concordance Index**. In: *Neural Information Processing Systems*.

[197] Rème, T., Hose, D., Theillet, C., and Klein, B. (2013). **Modeling risk stratification in human cancer**. *Bioinformatics*, 29:1149–1157. doi:10.1093/bioinformatics/btt124.

[198] Rew, S. B., Peggs, K., Sanjuan, I., Pizzey, A. R., Koishihara, Y., Kawai, S., Kosaka, M., Ozaki, S., Chain, B., and Yong, K. L. (2005). **Generation of potent antitumor CTL from patients with multiple myeloma directed against HM1.24**. *Clinical cancer research*, 11:3377–3384. doi:10.1158/1078-0432.CCR-04-0650.

[199] Ripley, B. D. (1987). **Stochastic simulation**. Wiley series in probability and mathematical statistics Applied probability and statistics. Wiley. doi:10.1002/9780470316726.

[200] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics*, 26:139–140. doi:10.1093/bioinformatics/btp616.

[201] Robinson, M. D. and Oshlack, A. (2010). **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome biology*, 11:R25. doi:10.1186/gb-2010-11-3-r25.

[202] Rosenthal, A., Kumar, S., Hofmeister, C., Laubach, J., Vij, R., Dueck, A., Gano, K., and Stewart, A. K. (2016). **A Phase Ib Study of the combination of the Aurora Kinase Inhibitor Alisertib (MLN8237) and Bortezomib in Relapsed Multiple Myeloma**. *British Journal of Haematology*, 174:323–325. doi:10.1111/bjh.13765.

[203] RStudio Team (2016). **RStudio: Integrated Development Environment for R**. RStudio, Inc., Boston, MA, version 1.1.442. URL: `http://www.rstudio.com/`. [Last visited: 17.10.2019].

[204] Salwender, H., Bertsch, U., Weisel, K., Duerig, J., Kunz, C., Benner, A., Blau, I. W., Raab, M. S., Hillengass, J., Hose, D., Huhn, S., Hundemer, M., Andrulis, M., Jauch, A., Seidel-Glaetzer, A., Lindemann, H.-W., Hensel, M., Fronhoffs, S., Martens, U., Hansen, T., Wattad, M., Graeven, U., Munder, M., Fenk, R., Haenel, M., Scheid, C., and Goldschmidt, H. (2019). **Rationale and design of the German-speaking myeloma multicenter group (GMMG) trial HD6 A randomized phase III trial on the effect of elotuzumab in VRD induction/consolidation and lenalidomide maintenance in patients with newly diagnosed myeloma**. *BMC cancer*, 19:504. doi:10.1186/s12885-019-5600-x.

[205] San Miguel, J. F., Garcia-Sanz, R., Gonzalez, M., Moro, M. J., Hernandez, J. M., Ortega, F., Borrego, D., Carnero, M., Casanova, F., and Jimenez, R. (1995). **A new staging system for multiple myeloma based on the number of S- phase plasma cells**. *Blood*, 85:448–455.

[206] San Miguel, J. F., Paiva, B., and Lasarte, J.-J. (2015). **Engineering Anti-myeloma Responses Using Affinity-Enhanced TCR-Engineered T Cells**. *Cancer Cell*, 28:281–283. doi:10.1016/j.ccell.2015.08.009.

[207] Scheid, C., Reece, D., Beksac, M., Spencer, A., Callander, N., Sonneveld, P., Kalimi, G., Cai, C., Shi, M., Scott, J. W., and Stewart, A. K. (2015). **Phase 2 study of dovitinib in patients with relapsed or refractory multiple myeloma with or without t(4;14) translocation**. *European Journal of Haematology*, 95:316–324. doi:10.1111/ejh.12491.

[208] Schmid, M., Wright, M. N., and Ziegler, A. (2016). **On the use of Harrell's C for clinical risk prediction via random survival forests**. *Expert Systems with Applications*, 63:450–459. doi:10.1016/j.eswa.2016.07.018.

[209] Schmitt, M., Hückelhoven, A. G., Hundemer, M., Schmitt, A., Lipp, S., Emde, M., Salwender, H., Hänel, M., Weisel, K., Bertsch, U., Dürig, J., Ho, A. D., Blau, I. W., Goldschmidt, H., Seckinger, A., and Hose, D. (2017). **Frequency of expression and generation of T-cell responses against antigens on multiple myeloma cells in patients included in the GMMG-MM5 trial**. *Oncotarget*, 8:84847–84862. doi:10.18632/oncotarget.11215.

[210] Schober, P., Boer, C., and Schwarte, L. A. (2018). **Correlation Coefficients Appropriate Use and Interpretation**. *Anesthesia and analgesia*, 126:1763–1768. doi:10.1213/ANE.0000000000002864.

[211] Seckinger, A., Bähr-Ivacevic, T., Benes, V., and Hose, D. (2018). **RNA-Sequencing from Low-Input Material in Multiple Myeloma for Application in Clinical Routine**. *Methods in molecular biology (Clifton, N.J.)*, 1792:97–115. doi:10.1007/978-1-4939-7865-6_7.

[212] Seckinger, A., Delgado, J. A., Moser, S., Moreno, L., Neuber, B., Grab, A., Lipp, S., Merino, J., Prosper, F., Emde, M., Delon, C., Latzko, M., Gianotti, R., Lüoend, R., Murr, R., Hosse, R. J., Harnisch, L. J., Bacac, M., Fauti, T., Klein, C., Zabaleta, A., Hillengass, J., Cavalcanti-Adam, E. A., Ho, A. D., Hundemer, M., San Miguel, J. F., Strein, K., Umaña, P., Hose, D., Paiva, B., and Vu, M. D. (2017). **Target Expression, Generation, Preclinical Activity, and Pharmacokinetics of the BCMA-T Cell Bispecific Antibody EM801 for Multiple Myeloma Treatment**. *Cancer Cell*, 31:396–410. doi:10.1016/j.ccell.2017.02.002.

[213] Seckinger, A., Hillengass, J., Emde, M., Beck, S., Kimmich, C., Dittrich, T., Hundemer, M., Jauch, A., Hegenbart, U., Raab, M.-S., Ho, A. D., Schönland, S., and Hose, D. (2018). **CD38 as Immunotherapeutic Target in Light Chain Amyloidosis and Multiple Myeloma-Association With Molecular Entities, Risk, Survival, and Mechanisms of Upfront Resistance**. *Frontiers in immunology*, 9:1676. doi:10.3389/fimmu.2018.01676.

[214] Seckinger, A. and Hose, D. (2015). **Dissecting the clonal architecture of multiple myeloma**. *EHA 20th Congress*, 9:173–180.

[215] Seckinger, A., Jauch, A., Emde, M., Beck, S., Mohr, M., Granzow, M., Hielscher, T., Rème, T., Schnettler, R., Fard, N., Hinderhofer, K., Pyl, P. T., Huber, W., Benes, V., Marciniak-Czochra, A., Pantesco, V., Ho, A. D., Klein, B., Hillengass, J., and Hose, D. (2016). **Asymptomatic Multiple Myeloma - Background of Progression, Evolution, and Prognosis**. *Blood*, 128:235, ASH Annual Meeting Abstract and Poster.

[216] Seckinger, A., Meissner, T., Moreaux, J., Depeweg, D., Hillengass, J., Hose, K., Rème, T., Rösen-Wolff, A., Jauch, A., Schnettler, R., Ewerbeck, V., Goldschmidt, H., Klein, B., and Hose, D. (2012). **Clinical and prognostic role of annexin A2 in multiple myeloma**. *Blood*, 120:1087–1094. doi:10.1182/blood-2012-03-415588.

[217] Seckinger, A., Meissner, T., Moreaux, J., Goldschmidt, H., Fuhler, G. M., Benner, A., Hundemer, M., Rème, T., Shaughnessy, Jr, JD, Barlogie, B., Bertsch, U., Hillengass, J., Ho, A. D., Pantesco, V., Jauch, A., de Vos, J., Rossi, J. F., Möhler, T., Klein, B., and Hose, D. (2009). **Bone morphogenic protein 6: a member of a novel class of prognostic factors expressed by normal and malignant plasma cells inhibiting proliferation and angiogenesis**. *Oncogene*, 28:3866–3879. doi:10.1038/onc.2009.257.

[218] Seshan, V. E. (2018). **clinfun Clinical Trial Design and Data Analysis Functions**, version 1.0.15. URL: `https://cran.r-project.org/web/packages/clinfun/index.html`. [Last visited: 17.10.2019].

[219] Shaughnessy, J. D., Zhan, F., Burington, B. E., Huang, Y., Colla, S., Hanamura, I., Stewart, J. P., Kordsmeier, B., Randolph, C., Williams, D. R., Xiao, Y., Xu, H., Epstein, J., Anaissie, E., Krishna, S. G., Cottler-Fox, M., Hollmig, K., Mohiuddin, A., Pineda-Roman, M., Tricot, G., van Rhee, F., Sawyer, J., Alsayed, Y., Walker, R., Zangari, M., Crowley, J., and Barlogie, B. (2007). **A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1**. *Blood*, 109:2276–2284. doi:10.1182/blood-2006-07-038430.

[220] Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). **dbSNP The NCBI database of genetic variation**. *Nucleic acids research*, 29:308–311. doi:10.1093/nar/29.1.308.

[221] Song, Y., Milon, B., Ott, S., Zhao, X., Sadzewicz, LisavShetty, A., Boger, E. T., Tallon, L. J., Morell, R. J., Mahurkar, A., and Hertzano, R. (2018). **A comparative analysis of library prep approaches for sequencing low input translatome samples**. *BMC Genomics*, 19:696. doi:10.1186/s12864-018-5066-2.

[222] Sonneveld, P., Schmidt-Wolf, I. G. H., van der Holt, B., el Jarari, L., Bertsch, U., Salwender, H., Zweegman, S., Vellenga, E., Broyl, A., Blau, I. W., Weisel, K. C., Wittebol, S., Bos, G. M. J., Stevens-Kroef, M., Scheid, C., Pfreundschuh, M., Hose, D., Jauch, A., van der Velde, H., Raymakers, R., Schaafsma, M. R., Kersten, M.-J., van Marwijk-Kooy, M., Duehrsen, U., Lindemann, W., Wijermans, P. W., Lokhorst, H. M., and Goldschmidt, H. M. (2012). **Bortezomib induction and maintenance treatment in patients with newly diagnosed multiple myeloma Results of the randomized phase III HOVON-65/ GMMG-HD4 trial**. *Journal of clinical oncology*, 30:2946–2955. doi:10.1200/JCO.2011.39.6820.

[223] Sprynski, A. C., Hose, D., Caillot, L., Réme, T., Shaughnessy, J. D., Barlogie, B., Seckinger, A., Moreaux, J., Hundemer, M., Jourdan, M., Meißner, T., Jauch, A., Mahtouk, K., Kassambara, A., Bertsch, U., Rossi, J. F., Goldschmidt, H., and Klein, B. (2009). **The role of IGF-1 as a major growth factor for myeloma cell lines and the prognostic relevance of the expression of its receptor**. *Blood*, 113:4614–4626. doi:10.1182/blood-2008-07-170464.

[224] Stadtmauer, E. A., Faitg, T. H., Lowther, D. E., Badros, A. Z., Chagin, K., Dengel, K., Iyengar, M., Melchiori, L., Navenot, J.-M., Norry, E., Trivedi, T., Wang, R., Binder, G. K., Amado, R., and Rapoport, A. P. (2019). **Long-term safety and activity of NY-ESO-1 SPEAR T cells after autologous stem cell transplant for myeloma**. *Blood Advances*, 3:2022–2034. doi:10.1182/bloodadvances.2019000194.

[225] Steensma, D. P., Gertz, M. A., Greipp, P. R., Kyle, R. A., Lacy, M. Q., Lust, J. A., Offord, J. R., Plevak, M. F., Therneau, T. M., and Witzig, T. E. (2001). **A high bone marrow plasma cell labeling index in stable plateau-phase multiple myeloma is a marker for early disease progression and death**. *Blood*, 97:2522–2523. doi:10.1182/blood.v97.8.2522.

[226] Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A.,

Rappaport, N., Safran, M., and Lancet, D. (2016). **The GeneCards Suite From Gene Data Mining to Disease Genome Sequence Analyses**. *Current protocols in bioinformatics*, 54:1.30.1–1.30.33. doi:10.1002/cpbi.5.

[227] Terstappen, L. W., Johnsen, S., Segers-Nolten, I. M., and Loken, M. R. (1990). **Identification and characterization of plasma cells in normal human bone marrow by high-resolution flow cytometry**. *Blood*, 76:1739–1747, ASH Annual Meeting Abstract.

[228] Therneau, T. (2015). **Survival: A Package for Survival Analysis in S**, version 2.41-3. URL: `https://cran.r-project.org/src/contrib/Archive/survival/survival_2.41-3.tar.gz`. [Last visited: 17.10.2019].

[229] Therneau, T. M. and Grambsch, P. M. (2000). **Modeling Survival Data: Extending the Cox Model**. Springer.

[230] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). **Diagnosis of multiple cancer types by shrunken centroids of gene expression**. *Proceedings of the national academy of sciences*, 99:6567–6572. doi:10.1073/pnas.082099299.

[231] Treon, S. P., Pilarski, L. M., Belch, A. R., Kelliher, A., Preffer, F. I., Shima, Y., Mitsiades, C. S., Mitsiades, N. S., Szczepek, A. J., Ellman, L., Harmon, D., Grossbard, M. L., and Anderson, K. C. (2002). **CD20-Directed Serotherapy in Patients With Multiple Myeloma Biologic Considerations and Therapeutic Applications**. *Journal of Immunotherapy*, 25:72–81.

[232] Trudel, S., Li, Z. H., Wei, E., Wiesmann, M., Chang, H., Chen, C., Reece, D., Heise, C., and Stewart, A. K. (2005). **CHIR-258, a novel, multitargeted tyrosine kinase inhibitor for the potential treatment of t(4;14) multiple myeloma**. *Blood*, 105:2941–2948. doi:10.1182/blood-2004-10-3913.

[233] Tsuboi, A., Oka, Y., Nakajima, H., Fukuda, Y., Elisseeva, O. A., Yoshihara, S., Hosen, N., Ogata, A., Kito, K., Fujiki, F., Nishida, S., Shirakata, T., Ohno, S., Yasukawa, M., Oji, Y., Kawakami, M., Morita, S., Sakamoto, J., Udaka, K., Kawase, I., and Sugiyama, H. (2007). **Wilms tumor gene WT1 peptide-based immunotherapy induced a minimal response in a patient with advanced therapy-resistant multiple myeloma**. *International journal of hematology*, 86:414–417. doi:10.1532/IJH97.07007.

[234] Usmani, S. Z., Weiss, B. M., Plesner, T., Bahlis, N. J., Belch, A., Lonial, S., Lokhorst, H. M., Voorhees, P. M., Richardson, P. G., Chari, A., Sasser, A. K., Axel, A., Feng, H., Uhlar, C. M., Wang, J., Khan, I., Ahmadi, T., and Nahi, H. (2016). **Clinical efficacy of daratumumab monotherapy in patients with heavily pretreated relapsed or refractory multiple myeloma**. *Blood*, 128:37–44. doi:10.1182/blood-2016-03-705210, ASH Annual Meeting Abstract.

[235] van de Wiel, M. A., Berkhof, J., and van Wieringen, W. N. (2009). **Testing the prediction error difference between 2 predictors**. *Biostatistics*, 10:550–560. doi:10.1093/biostatistics/kxp011.

[236] van Laar, R., Flinchum, R., Brown, N., Ramsey, J., Riccitelli, S., Heuck, C., Barlogie, B., and Shaughnessy, J. D. (2014). **Translating a gene expression signature for multiple myeloma prognosis into a robust high-throughput assay for clinical use**. *BMC medical genomics*, 7:25. doi:10.1186/1755-8794-7-25.

[237] van Rhee, F., Szmania, S. M., Zhan, F., Gupta, S. K., Pomtree, M., Lin, P., Batchu, R. B., Moreno, A., Spagnoli, G., Shaughnessy, J., and Tricot, G. (2005). **NY-ESO-1 is highly expressed in poor-prognosis multiple myeloma and induces spontaneous humoral and cellular immune responses**. *Blood*, 105:3939–3944. doi:10.1182/blood-2004-09-3707.

[238] van Vliet, M., Ubels, J., de Best, L., van Beers, E., and Sonneveld, P. (2015). **The Combination of SKY92 and ISS Provides a Powerful Tool to Identify Both High Risk and Low Risk Multiple Myeloma Cases, Validation in Two Independent Cohorts**. *Blood*, 126:2970–2970. doi:10.1182/blood.V126.23.2970.2970, ASH Annual Meeting Abstract.

[239] Venables, W. N. and Ripley, B. D. (2002). **Modern applied statistics with S**. Springer-Verlag.

[240] Vittinghoff, E. and McCulloch, C. E. (2007). **Relaxing the rule of ten events per variable in logistic and Cox regression**. *American journal of epidemiology*, 165:710–718. doi:10.1093/aje/kwk052.

[241] Wald, A. (1943). **Tests of statistical hypotheses concerning several parameters when the number of observations is large**. *Transactions of the American Mathematical Society*, 54:426. doi:10.1090/S0002-9947-1943-0012401-3.

[242] Walker, B. A., Boyle, E. M., Wardell, C. P., Murison, A., Begum, D. B., Dahir, N. M., Proszek, P. Z., Johnson, D. C., Kaiser, M. F., Melchor, L., Aronson, L. I., Scales, M., Pawlyn, C., Mirabella, F., Jones, J. R., Brioli, A., Mikulasova, A., Cairns, D. A., Gregory, W. M., Quartilho, A., Drayson, M. T., Russell, N., Cook, G., Jackson, G. H., Leleu, X., Davies, F. E., and Morgan, G. J. (2015). **Mutational Spectrum, Copy Number Changes, and Outcome Results of a Sequencing Study of Patients With Newly Diagnosed Myeloma**. *Journal of clinical oncology*, 33:3911–3920. doi:10.1200/JCO.2014.59.1503.

[243] Walker, B. A., Leone, P. E., Chiecchio, L., Dickens, N. J., Jenner, M. W., Boyd, K. D., Johnson, D. C., Gonzalez, D., Dagrada, G. P., Protheroe, R. K. M., Konn, Z. J., Stockley, D. M., Gregory, W. M., Davies, F. E., Ross, F. M., and Morgan, G. J. (2010). **A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value**. *Blood*, 116:e56–65. doi:10.1182/blood-2010-04-279596.

[244] Walker, B. A., Wardell, C. P., Melchor, L., Brioli, A., Johnson, D. C., Kaiser, M. F., Mirabella, F., Lopez-Corral, L., Humphray, S., Murray, L., Ross, M., Bentley, D., Gutiérrez, N. C., Garcia-Sanz, R., San Miguel, J., Davies, F. E., Gonzalez, D., and Morgan, G. J. (2014). **Intraclonal heterogeneity is a critical early event in the development of myeloma and precedes the development of clinical symptoms**. *Leukemia*, 28:384–390. doi:10.1038/leu.2013.199.

[245] Walker, B. A., Wardell, C. P., Melchor, L., Hulkki, S., Potter, N. E., Johnson, D. C., Fenwick, K., Kozarewa, I., Gonzalez, D., Lord, C. J., Ashworth, A., Davies, F. E., and Morgan, G. J. (2012). **Intraclonal heterogeneity and distinct molecular mechanisms characterize the development of t(4;14) and t(11;14) myeloma**. *Blood*, 120:1077–1086. doi:10.1182/blood-2012-03-412981.

[246] Warren, P. (2016). **panp Presence-Absence Calls from Negative Strand Matching Probesets**, version 1.44.0. URL: `https://bioconductor.org/packages/3.4/bioc/html/panp.html`. [Last visited: 17.10.2019].

[247] Warren, P., Taylor, D., Martini, P. G. V., Jackson, J., and Bienkowska, J. (2007). **PANP - a New Method of Gene Detection on Oligonucleotide Expression Arrays**. In: *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, editor IEEE, pages 108–115. IEEE. doi:10.1109/BIBE.2007.4375552.

[248] Waxman, A. J., Mink, P. J., Devesa, S. S., Anderson, W. F., Weiss, B. M., Kristinsson, S. Y., McGlynn, K. A., and Landgren, O. (2010). **Racial disparities in incidence and outcome in multiple myeloma: a population-based study**. *Blood*, 116:5501–5506. doi:10.1182/blood-2010-07-298760.

[249] Wu, J., Irizarry, R., and Gentry, J. M. (2016). **gcrma Background Adjustment Using Sequence Information**, version 2.46.0. URL: `https://bioconductor.org/packages/3.4/bioc/html/gcrma.html`. [Last visited: 17.10.2019].

[250] Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). **A Model-Based Background Adjustment for Oligonucleotide Expression Arrays**. *Journal of the American Statistical Association*, 99:909–917. doi:10.1198/016214504000000683.

[251] Wuilleme, S., Robillard, N., Lodé, L., Magrangeas, F., Beris, H., Harousseau, J.-L., Proffitt, J., Minvielle, S., Avet-Loiseau, H., and Intergroupe Francophone de Myélome (2005). **Ploidy, as detected by fluorescence in situ hybridization, defines different subgroups in multiple myeloma**. *Leukemia*, 19:275–278. doi:10.1038/sj.leu.2403586.

[252] Xu, J., Pfarr, N., Endris, V., Mai, E. K., Md Hanafiah, N. H., Lehners, N., Penzel, R., Weichert, W., Ho, A. D., Schirmacher, P., Goldschmidt, H., Andrulis, M., and Raab, M. S. (2017). **Molecular signaling in multiple myeloma Association of RAS/RAF mutations and MEK/ERK pathway activation**. *Oncogenesis*, 6:e337. doi:10.1038/oncsis.2017.36.

[253] Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A., and Flicek, P. (2018). **Ensembl 2018**. *Nucleic acids research*, 46:D754–D761. doi:10.1093/nar/gkx1098.

[254] Zhan, F., Huang, Y., Colla, S., Stewart, J. P., Hanamura, I., Gupta, S., Epstein, J., Yaccoby, S., Sawyer, J., Burington, B., Anaissie, E., Hollmig, K., Pineda-Roman, M., Tricot, G., van Rhee, F., Walker, R., Zangari, M., Crowley, J., Barlogie, B., and Shaughnessy, Jr, John D (2006). **The molecular classification of multiple myeloma**. *Blood*, 108:2020–2028. doi:10.1182/blood-2005-11-013458.

[255] Zhang, X., Gaillard, J.-P., Robillard, N., Lu, Z., Gu, Z., Jourdan, M., Boiron, J.-M., Bataille, R., and Klein, B. (1994). **Reproducible obtaining of human myeloma cell lines as a model for tumor stem study in human multiple myeloma**. *Blood*, 83:3654–63. doi:10.1182/blood.V83.12.3654.bloodjournal83123654.

7 Contributions and publications

This work was realised within and funded by the Multiple Myeloma Research Laboratory at the University Hospital Heidelberg and supported in parts by grants from the German Federal Ministry of Education (BMBF) "CAMPSIMM" (01ES1103), the BMBF "CLIOMMICS" (01ZX1309), the Deutsche Forschungsgemeinschaft (SFB/TRR79; TP B12; Bonn, Germany) and the 7th EU-framework program "OverMyR". This work was contributed in parts by technicians or members of the following associations:

1 Multiple Myeloma Research Laboratory (Head: PD Dr. Dr. Dirk Hose), Medizinische Klinik V, Universitätsklinikum Heidelberg, Im Neuenheimer Feld 410, D-69120 Heidelberg, Germany

2 Biomathematics and Bioinformatics (Head: Prof. Dr. Dr. Thierry Rème), Department of Biological Hematology, CHRU de Montpellier, Hôpital Saint-Eloi, 80, av. Augustin Fliche, F-34295 Montpellier Cedex 5, France.

3 Molekular-zytogenetisches Labor (Head: Prof. Dr. sc. hum. Anna Jauch), Institut für Humangenetik , Universität Heidelberg, Im Neuenheimer Feld 366, D-69120 Heidelberg, Germany.

4 Institute for Research in Biotherapy (Head: Prof. Dr. Bernard Klein), CHU de Montpellier, Hôpital Saint-Eloi, 80, av. Augustin Fliche, F-34295 Montpellier Cedex 5, France.

5 Genomics Core Facility (Head: Dr. Vladimir Benes), EMBL Heidelberg, Meyerhofstraße 1, D-69117 Heidelberg, Germany.

The parts that others have contributed to are listed in the following table:

Subject	**Item**	**Contribution [Association]**
Classification transfer and creation of the HDHRS	Discussion and interpretation of the results	PD Dr. Dr. Dirk Hose [1] Dr. Anja Seckinger [1] Susanne Beck [1]
	Basic RS algorithm	Prof. Dr. Dr. Thierry Rème [2]
Target analyses	Discussion and interpretation of the results	PD Dr. Dr. Dirk Hose [1] Dr. Anja Seckinger [1] Susanne Beck [1]
Clinical data	Scientific responsibility and data interpretation	PD Dr. Dr. Dirk Hose [1] Dr. Anja Seckinger [1]
	Data preprocessing and cleaning	Susanne Beck [1]
	Data collection	PD Dr. Dr. Dirk Hose [1] Dr. Anja Seckinger [1] Sybille Seyfried [1]
Laboratory performance	Administrative and scientific responsibility	PD Dr. Dr. Dirk Hose [1] Dr. Anja Seckinger [1]
	Performing plasma cell purification	Maria Dörner and Birgit Schneiders et al., Multiple Myeloma Research Laboratory (PD Dr. Dr. Dirk Hose) [1]
	Performing interphase fluorescence *in situ* hybridisation	Technicians, Molekular-zytogenetisches Labor (Prof. Dr. sc. hum. Anna Jauch) [3]
	Performing microarray gene expression profiling	Véronique Pantesco, Institute for Research in Biotherapy (Prof. Dr. Bernard Klein) [4]
	Performing RNA-sequencing	Tomi Bähr-Ivacevic, Multiple Myeloma Research Laboratory (PD Dr. Dr. Dirk Hose) [1] Genomics Core Facility (Dr. Vladimir Benes) [6]
Sampling	Performing bone marrow aspiration	Responsible physicians

Partial results of this thesis were published in advance in the following articles:

[1] Schmitt, M., Hückelhoven, A. G., Hundemer, M., Schmitt, A., Lipp, S., ***Emde, M.***, Salwender, H., Hänel, M., Weisel, K., Bertsch, U., Dürig, J., Ho, A. D., Blau, I. W., Goldschmidt, H., Seckinger, A., and Hose, D. (2017). **Frequency of expression and generation of T-cell responses against antigens on multiple myeloma cells in patients included in the GMMG-MM5 trial**. *Oncotarget*, 8:84847–84862. doi:10.18632/oncotarget.11215.

[2] Seckinger, A., Delgado, J. A., Moser, S., Moreno, L., Neuber, B., Grab, A., Lipp, S., Merino, J., Prosper, F., ***Emde, M.***, Delon, C., Latzko, M., Gianotti, R., Lüoend, R., Murr, R., Hosse, R. J., Harnisch, L. J., Bacac, M., Fauti, T., Klein, C., Zabaleta, A., Hillengass, J., Cavalcanti-Adam, E. A., Ho, A. D., Hundemer, M., San Miguel, J. F., Strein, K., Umaña, P., Hose, D., Paiva, B., and Vu, M. D. (2017). **Target Expression, Generation, Preclinical Activity, and Pharmacokinetics of the BCMA-T Cell Bispecific Antibody EM801 for Multiple Myeloma Treatment**. *Cancer Cell*, 31:396–410. doi:10.1016/j.ccell.2017.02.002.

[3] Seckinger, A., Hillengass, J., ***Emde, M.***, Beck, S., Kimmich, C., Dittrich, T., Hundemer, M., Jauch, A., Hegenbart, U., Raab, M.-S., Ho, A. D., Schönland, S., and Hose, D. (2018). **CD38 as Immunotherapeutic Target in Light Chain Amyloidosis and Multiple Myeloma-Association With Molecular Entities, Risk, Survival, and Mechanisms of Upfront Resistance**. *Frontiers in Immunology*, 9:1676. doi:10.3389/fimmu.2018.01676.

The parts of this thesis, published in the above indicated articles are listed in the following table:

Subject	Section	Figures	Tables	Abstract	Comment
DNA-microarray expression and survival analyses of CTAs (HM1.24, NYESO1/2, MAGEA3, RHAMM, WT1)	3.5.1, 4.4.2.3, 4.4.3	A.9, A.10, A.11, A.14, A.15		[1]	published with altered sample composition
RNA-seq expression analyses of CTAs	3.5.1, 4.4.3	A.9, A.10, A.11, A.14, A.15		[1]	published for a smaller cohort

Correlation of DNA-microarray and RNA-seq expression of CTAs	3.5.1	A.16		[1]	published for a smaller cohort
Microarray expression analysis of BCMA	3.5.1, 4.4.2.1	3.22		[2]	published with altered sample composition
RNA-seq expression analysis of BCMA	3.5.1, 4.4.2.1	3.22		[2]	published for a smaller cohort
DNA-microarray expression and survival analysis of CD38	3.5.1, 4.4.2.2, 4.4.3		A.9	[3]	published with altered sample composition
RNA-seq expression analysis of CD38	3.5.1, 4.4.2.2, 4.4.3		A.9	[3]	published for a smaller cohort
CD38 splice junction analysis	3.5.2.2, 4.4.4.2	3.24, A.17, A.18, A.19	3.19, B.23	[3]	published for a smaller cohort

Partial results of this thesis were published in advance in the following abstract:

[1] ***Emde, M.***, Seckinger, A., Benes, V., Moreaux, J., Beck, S., and Hose, D. (2019). **RNA-Sequencing Based Assessment of Targets, Risk and Long Term Survival for Personalized Treatment of Multiple Myeloma**. *Blood*, 134:1801. doi:10.1182/blood-2019-131159, ASH Annual Meeting Abstract and Poster.

The parts of this thesis, published in the above indicated abstract are listed in the following table:

Subject	**Section**	**Figures**	**Tables**	**Comment**
Survival analyses of transferred proliferation based stratification	3.2.1, 4.6	3.4a, 3.4a		published with merged TG, VG and TeG
Survival analyses of novel risk stratification	3.3, 4.6	3.14, 3.21, A.7		published with merged TG, VG and TeG

Further own publications:

[1] Hose, D., Beck, S., Salwender, H., ***Emde, M.***, Bertsch, U., Kunz, C., Scheid, C., Hänel, M., Weisel, K., Hielscher, T., Raab, M. S., Goldschmidt, H., Jauch, A., Moreaux, J., and Seckinger, A. (2019). **Prospective target assessment and multimodal prediction of survival for personalized and risk-adapted treatment strategies in multiple myeloma in the GMMG-MM5 multicenter trial**. *Journal of hematology & oncology*, 12:65. doi:10.1186/s13045-019-0750-5.

[2] Xue, J., Schmidt, S. V., Sander, J., Draffehn, A., Krebs, W., Quester, I., de Nardo, D., Gohel, T. D., ***Emde, M.***, Schmidleithner, L., Ganesan, H., Nino-Castro, A., Mallmann, M. R., Labzin, L., Theis, H., Kraut, M., Beyer, M., Latz, E., Freeman, T. C., Ulas, T., and Schultze, J. L. (2014). **Transcriptome-based network analysis reveals a spectrum model of human macrophage activation**. *Immunity*, 40:274-288. doi:10.1016/j.immuni.2014.01.006.

Further published abstracts:

[1] Beck, S., ***Emde, M.***, Moreaux, J., Seckinger, A., and Hose, D. (2019). **Prediction of Malignant Plasma Cell Biology Related Survival in AL-Amyloidosis**. *Blood*, 134:3078, ASH Annual Meeting Abstract and Poster.

[2] Seckinger, A., Hegenbart, U., Beck, S., ***Emde, M.***, Bochtler, T., Kimmich, C., Müller-Tidow, C., Jauch, A., Schönland, S., and Hose, D. (2018). **AL Amyloidosis - Pathogenesis and Prognosis Are Determined By the Amyloidogenic Potential of the Light Chain and the Molecular Characteristics of Malignant Plasma Cells**. *Blood*, 132:187, ASH Annual Meeting Abstract and Poster.

[3] Seckinger, A., Salwender, H. J., Martin, H., Scheid, C., Hielscher, T., Bertsch, U., Hummel, M., Jauch, A., Knauf, W., ***Emde, M.***, Beck, S., Neben, K., Lokhorst, H. M., van der Holt, B., Duehrsen, U., Dürig, J., Lindemann, H.-W., Schmidt-Wolf, I., Haenel, M., Lathan, B., Raab, M. S., Müller-Tidow, C., Sonneveld, P., Blau, I. W., Hillengass, J., Weisel, K., Goldschmidt, H., and Hose, D. (2018). **Treatment Response and Long-Term Survival in Multiple Myeloma in the GMMG-HD4 Trial - Neither Profit All Molecular Entities Alike, Nor Are Remissions to Different Regimen Equal**. *Blood*, 132:4485, ASH Annual Meeting Abstract and Poster.

[4] Seckinger, A., Jauch, A., ***Emde, M.***, Beck, S., Mohr, M., Granzow, M., Hielscher, T., Réme, T., Schnettler, R., Fard, N., Hinderhofer, K., Pyl, P. T., Huber, W., Benes, V., Marciniak-Czochra, A., Pantesco, V., Ho, A. D., Klein, B., Hillengass, J., and Hose, D. (2016). **Asymptomatic Multiple Myeloma - Background of Progression, Evolution, and Prognosis**. *Blood*, 128:235, ASH Annual Meeting Abstract and Poster.

[5] Stroh, J., Seckinger, A., Heider, M., Eichner, R., ***Emde, M.***, Salweder, H., Bertsch, U., Goldschmidt, H., Weisel, K., Scheid, C., Hose, D., and Bassermann, F. (2019). **MCT1 As Molecularly Validated Predictive Marker for Lenalidomide-Maintenance Therapy in Multiple Myeloma**. *Blood*, 134:3187. ASH Annual Meeting Abstract and Poster.

Appendix

A Supplementary Figures

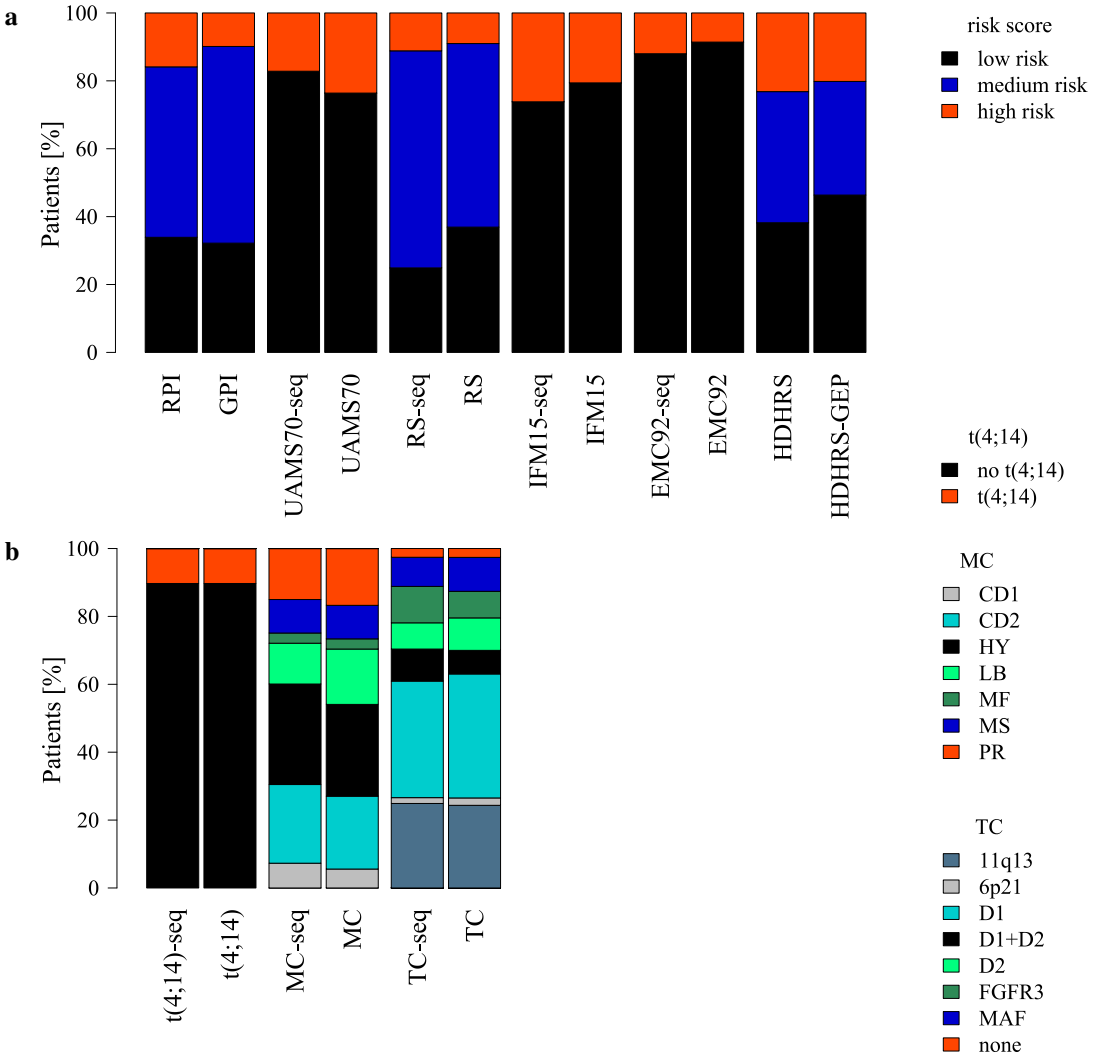

Figure A.1: Proportions of patients regarding risk stratifications and classifications(test group (TeG)). Shown are the different **a** risk stratification and **b** molecular classification proportions for DNA-microarrays in comparison to RNA-sequencing based analysis. Depicted are t(4;14) prediction assessed by RNA-seq (t(4;14)-seq) and DNA-microarrays (t(4;14)), molecular classification (MC) and TC. For further explanations of the groups see section 1.4.2.3.

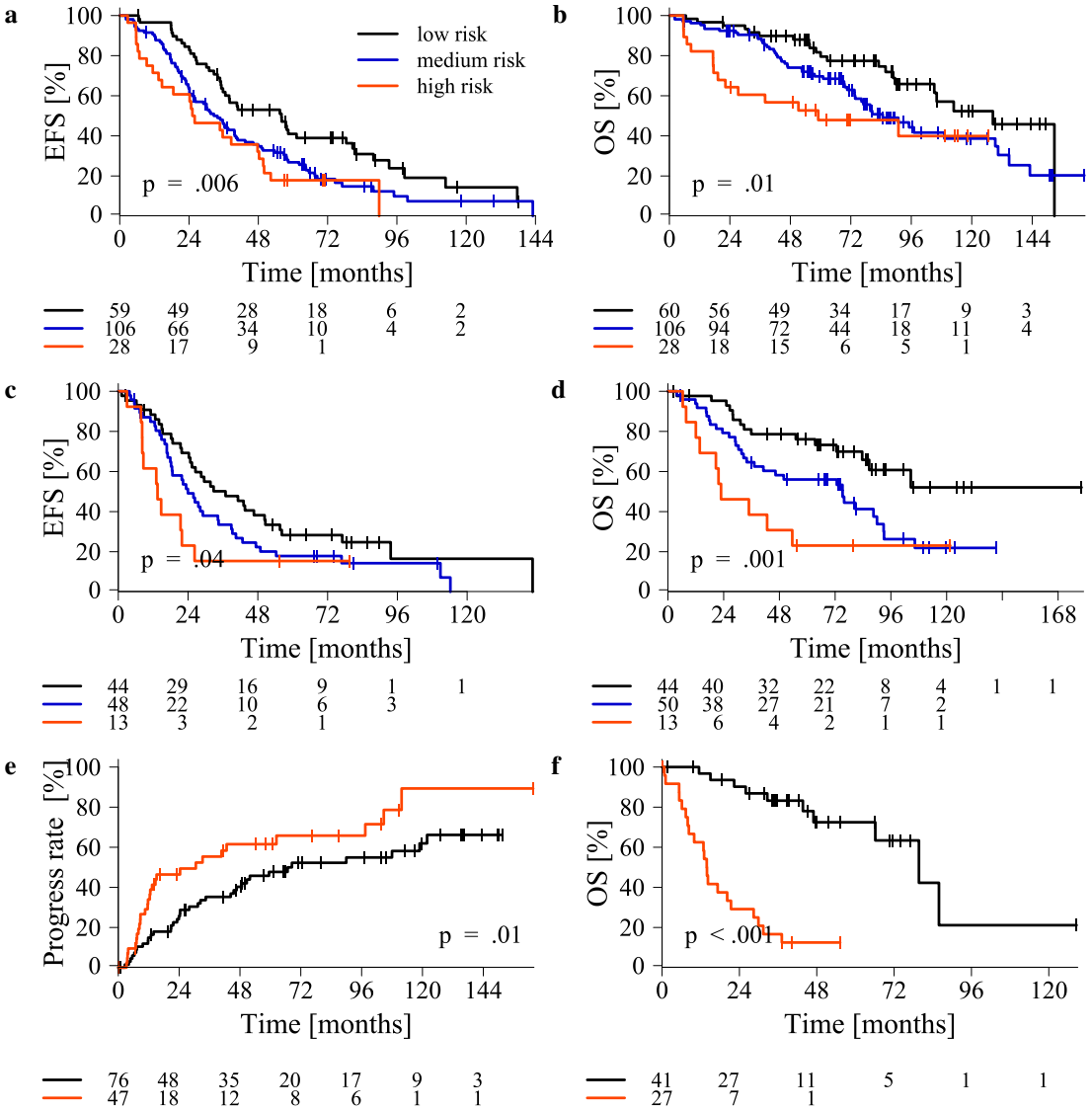

Figure A.2: Survival analysis regarding RNA-seq based proliferation index (RPI) for training group (TG), validation group (VG), as well as for asymptomatic multiple myeloma (AMM) and relapsed myeloma (MMR) patients. Performance of the RPI in event free survival (EFS) and overall survival (OS) for the TG (**a**, **b**) and VG (**c**, **d**). Progression free survival of AMM patients is depicted in subfigure **e** and OS of MMR patients in subfigure **f**.

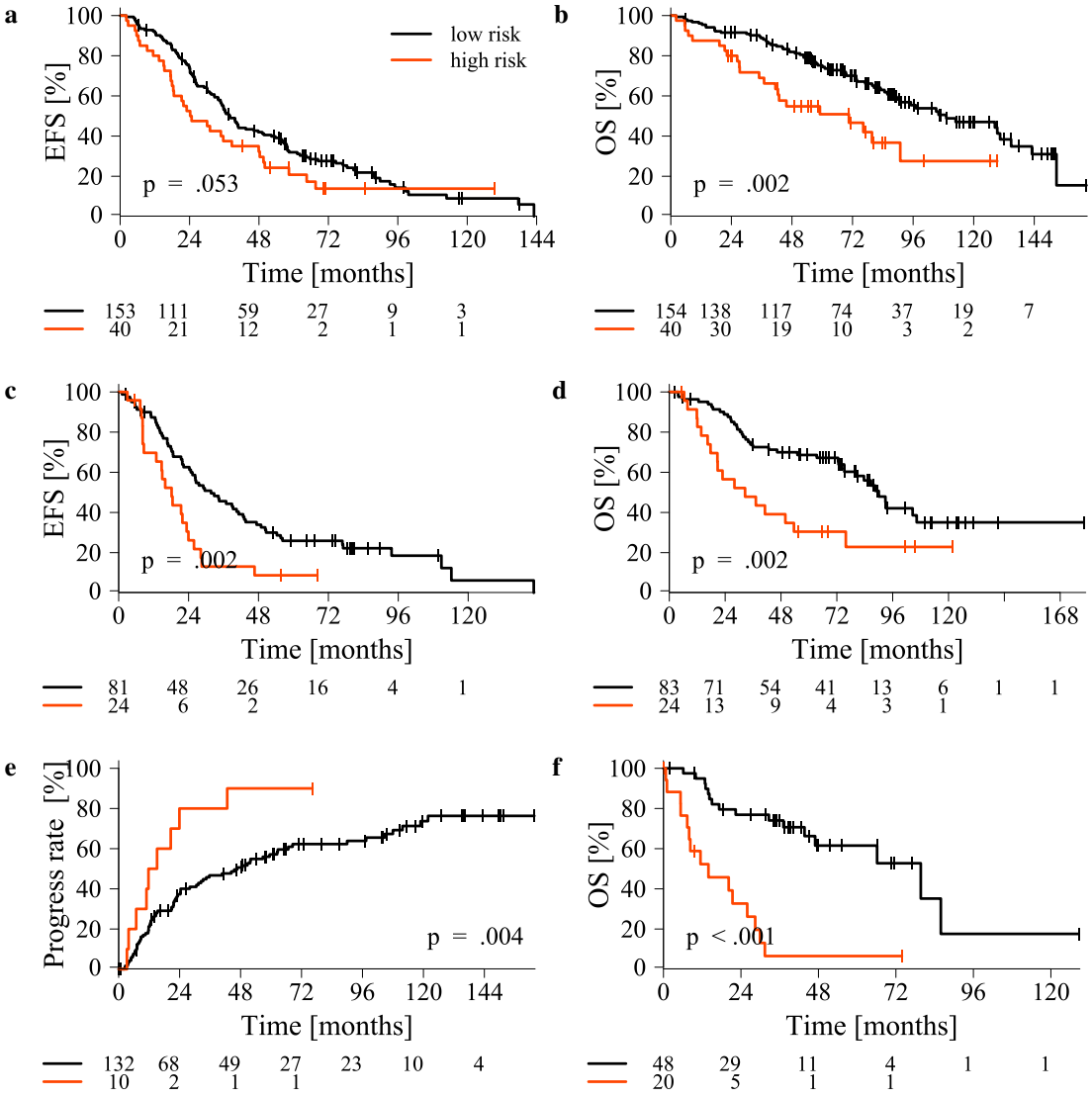

Figure A.3: Survival analysis regarding UAMS70-seq for training group (TG), validation group (VG), as well as for asymptomatic multiple myeloma (AMM) and relapsed myeloma (MMR) patients. Performance of the UAMS70-seq in event free survival (EFS) and overall survival (OS) for the TG (**a**, **b**) and VG (**c**, **d**). Progression free survival of AMM patients is depicted in subfigure **e** and OS of MMR patients in subfigure **f**.

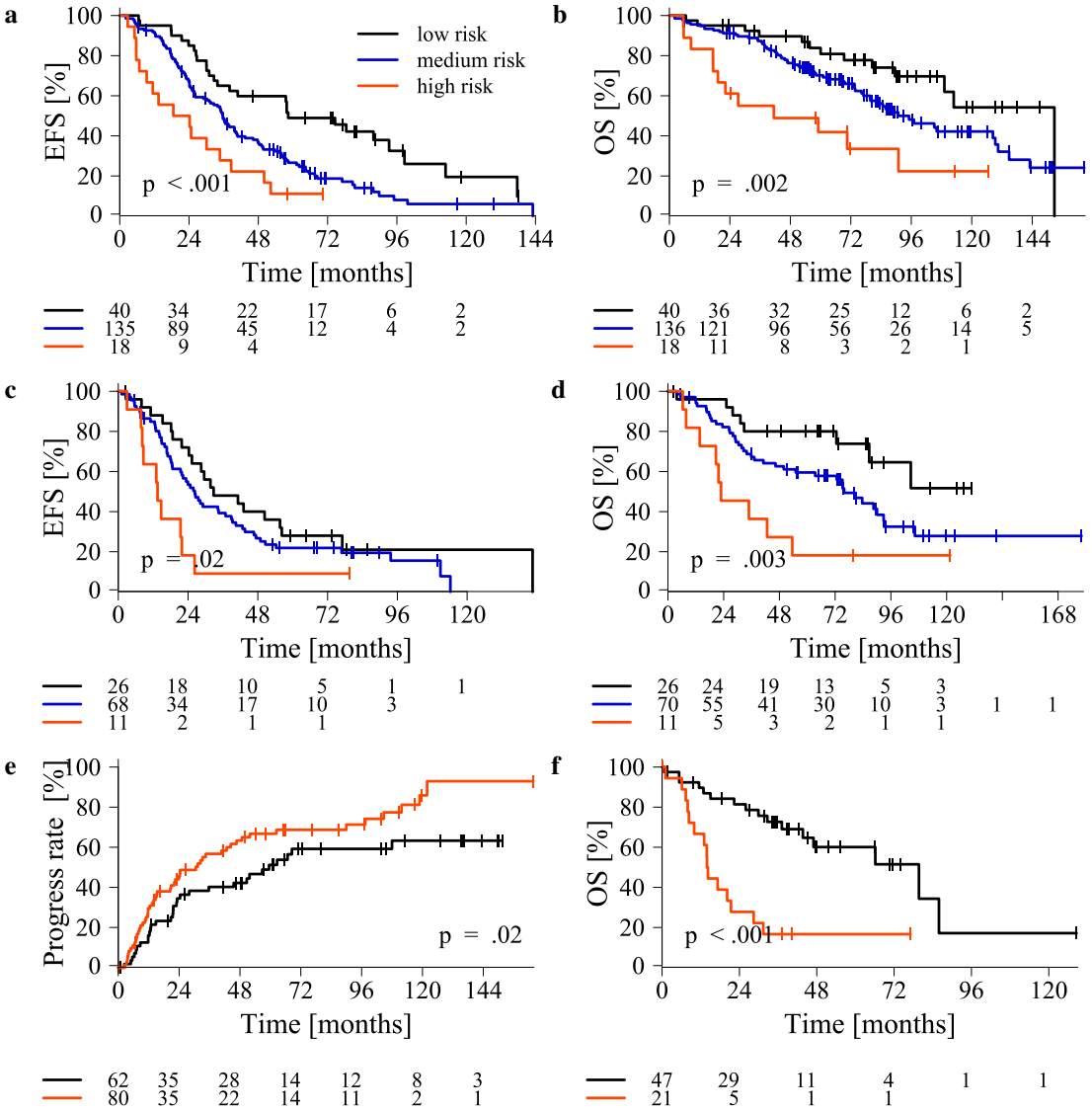

Figure A.4: Survival analysis regarding RS-seq for training group (TG), validation group (VG), as well as for asymptomatic multiple myeloma (AMM) and relapsed myeloma (MMR) patients. Performance of the RS-seq in event free survival (EFS) and overall survival (OS) for the TG (**a**, **b**) and VG (**c**, **d**). Progression free survival of AMM patients is depicted in subfigure **e** and OS of MMR patients in subfigure **f**.

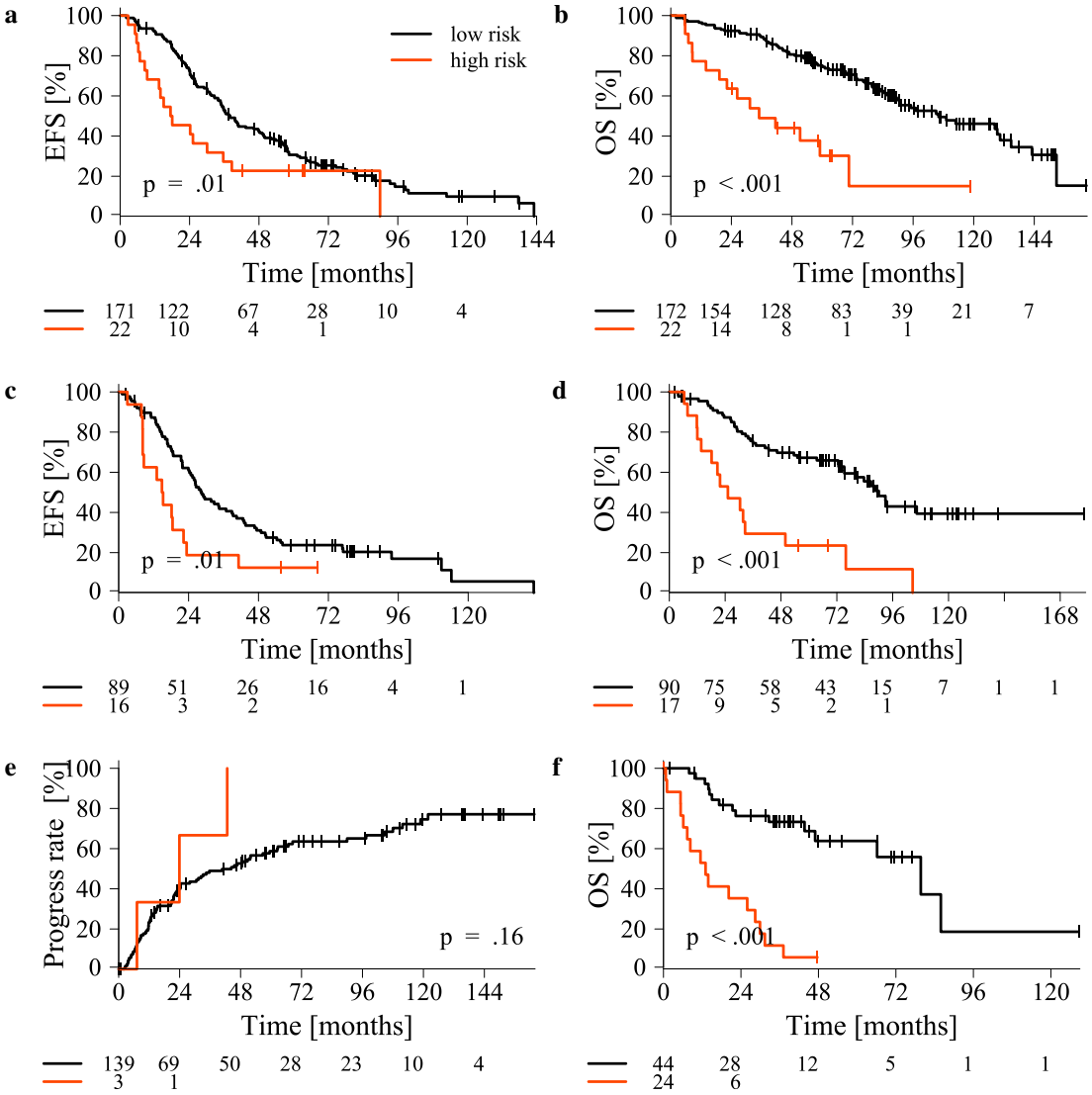

Figure A.5: Survival analysis regarding EMC92-seq for training group (TG), validation group (VG), as well as for asymptomatic multiple myeloma (AMM) and relapsed myeloma (MMR) patients. Performance of the EMC92-seq in event free survival (EFS) and overall survival (OS) for the TG (**a**, **b**) and VG (**c**, **d**). Progression free survival of AMM patients is depicted in subfigure **e** and OS of MMR patients in subfigure **f**.

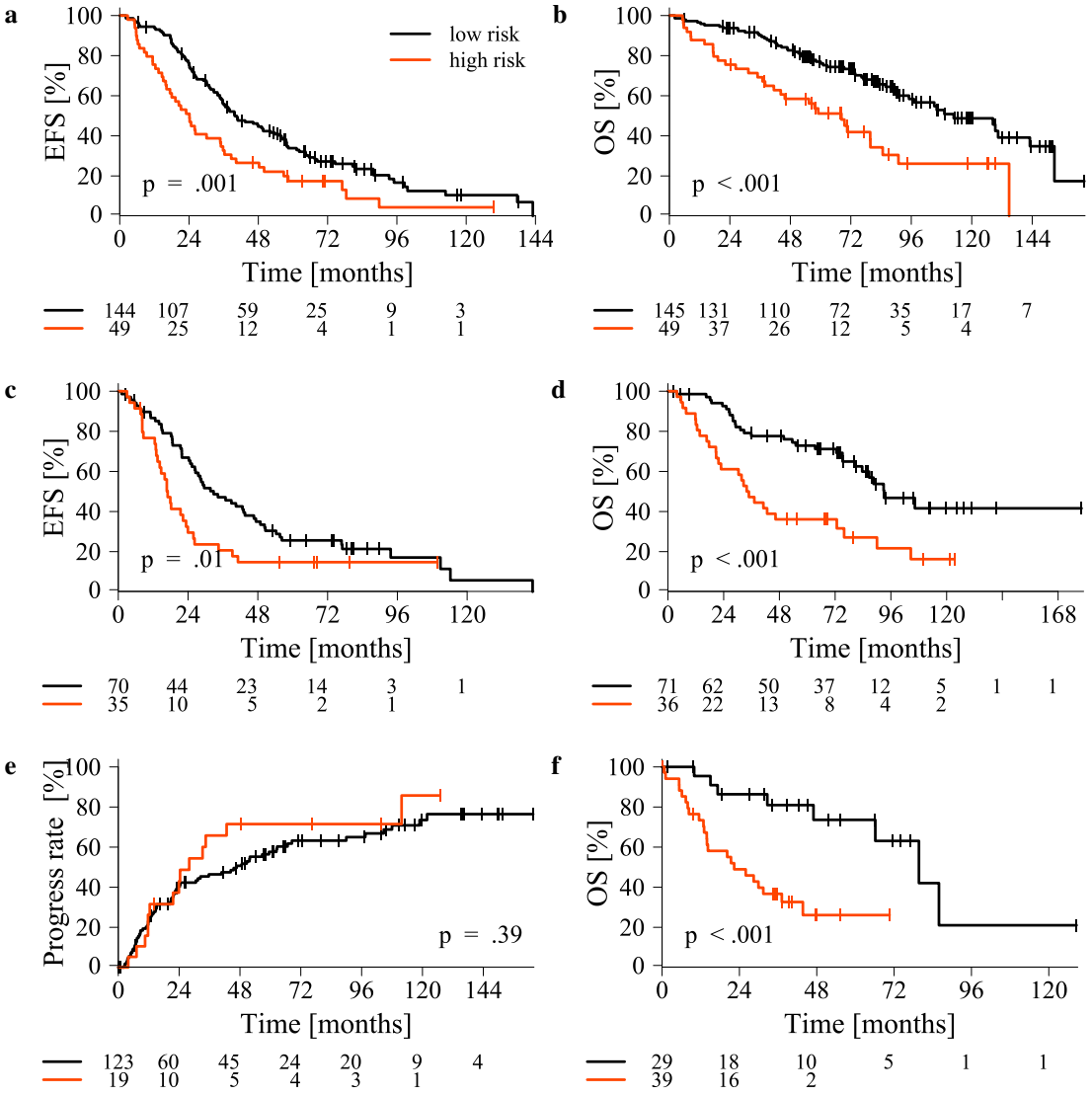

Figure A.6: Survival analysis regarding IFM15-seq for training group (TG), validation group (VG), as well as for asymptomatic multiple myeloma (AMM) and relapsed myeloma (MMR) patients. Performance of the IFM15-seq in event free survival (EFS) and overall survival (OS) for the TG (**a**, **b**) and VG (**c**, **d**). Progression free survival of AMM patients is depicted in subfigure **e** and OS of MMR patients in subfigure **f**.

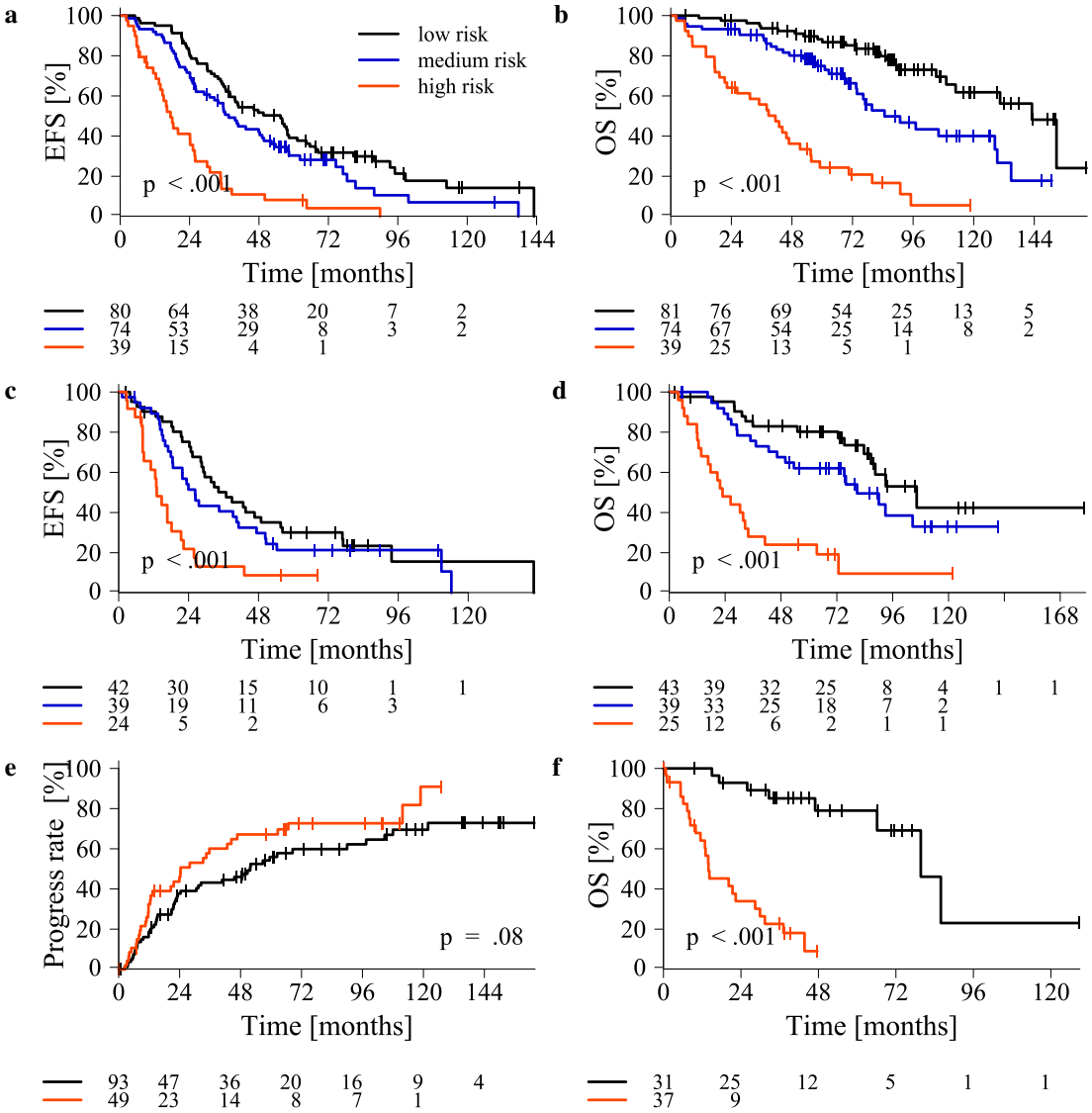

Figure A.7: Survival analysis regarding HDHRS for training group (TG), validation group (VG), as well as for asymptomatic multiple myeloma (AMM) and relapsed myeloma (MMR) patients. Performance of the HDHRS in event free survival (EFS) and overall survival (OS) for the TG (**a**, **b**) and VG (**c**, **d**). Progression free survival of AMM patients is depicted in subfigure **e** and OS of MMR patients in subfigure **f**.

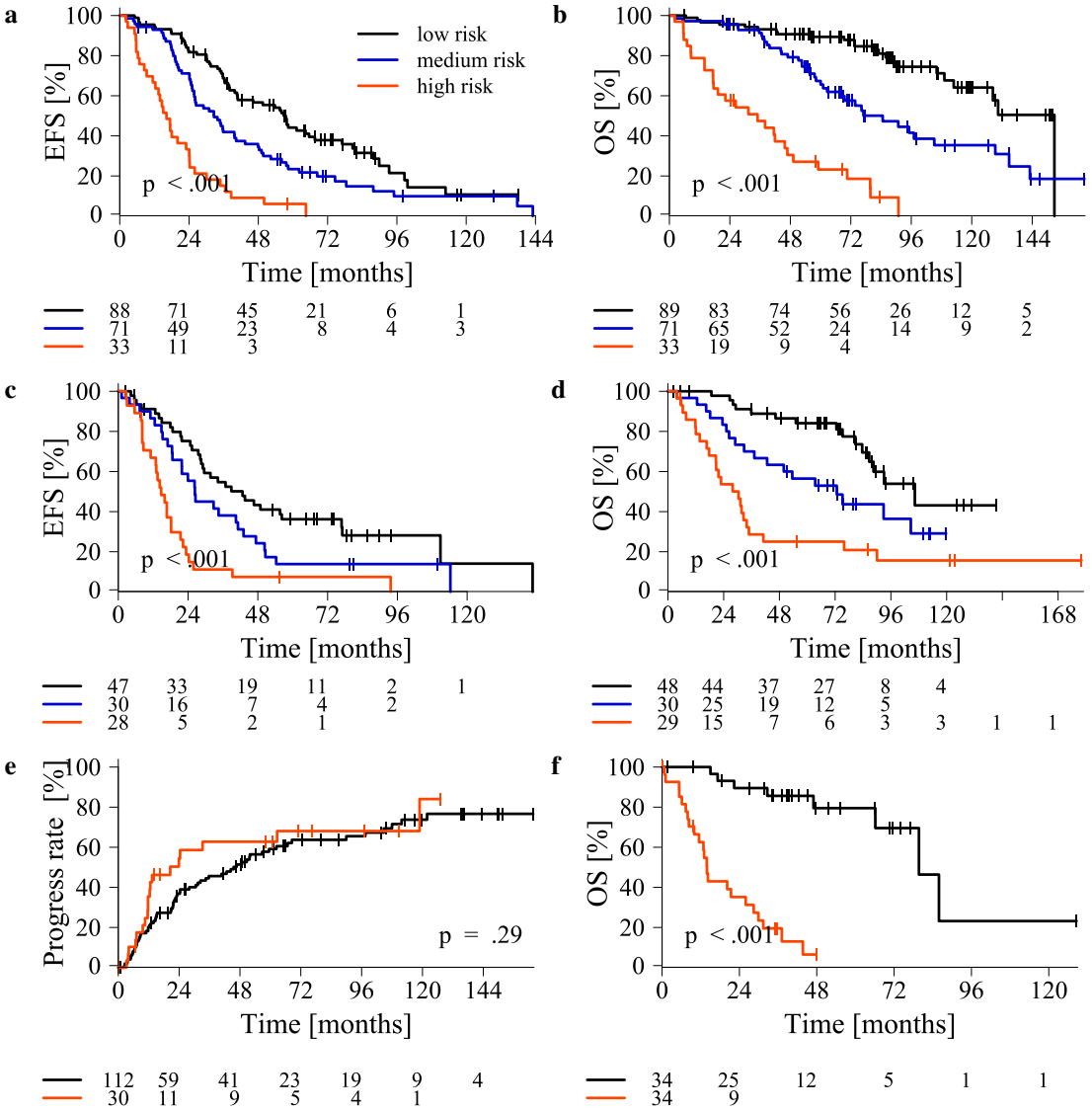

Figure A.8: Survival analysis regarding HDHRS-GEP for training group (TG), validation group (VG), as well as for asymptomatic multiple myeloma (AMM) and relapsed myeloma (MMR) patients. Performance of the HDHRS-GEP in event free survival (EFS) and overall survival (OS) for the TG (**a**, **b**) and VG (**c**, **d**). Progression free survival of AMM patients is depicted in subfigure **e** and OS of MMR patients in subfigure **f**.

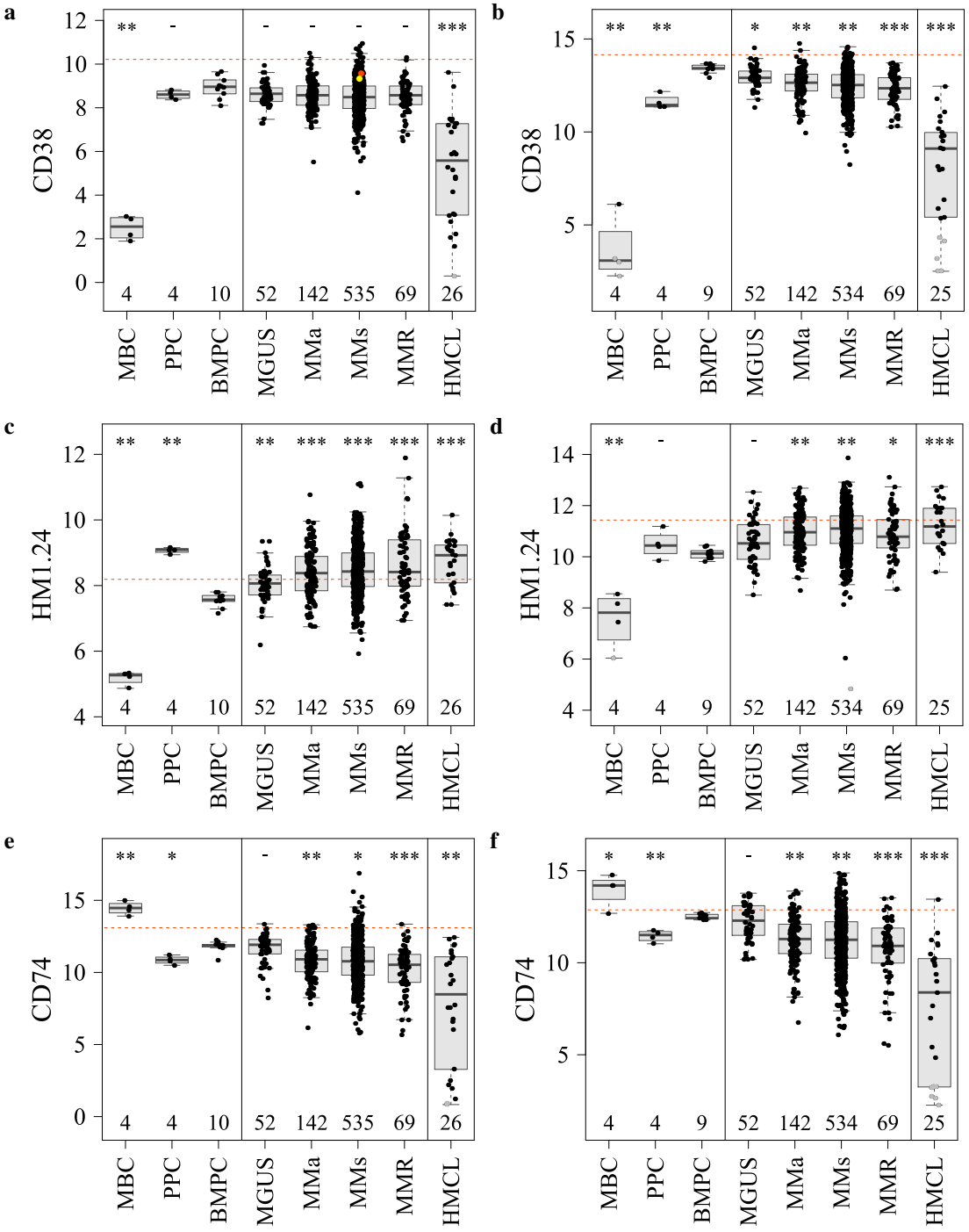

Figure A.9: Expression of *CD38*, *HM1.24* and *CD74*. The whole cohort is assessed. Expression is grouped by disease entity, compared to non-malignant cells and precursors (memory B cells (MBC), polyclonal plasmablastic cells (PPC), bone marrow plasma cells (BMPC)) and human myeloma cell lines (HMCL). Significant differences in expression compared to BMPCs are indicated by 1, 2 or 3 asterisks, for p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. RNA-seq expression is depicted for **a** *CD38* (published with a smaller cohort in Seckinger, ..., Emde et al., Frontiers in Immunology 2018 [213]) and **c** *HM1.24* (published with a smaller cohort in Schmitt, ..., Emde et al., Oncotarget 2017 [209]) and **e** *CD74*. Microarray expression is depicted for **b** *CD38* (published with altered sample composition in Seckinger, ..., Emde et al., Frontiers in Immunology 2018 [213]), **d** *HM1.24* (published with altered sample composition in Schmitt, ..., Emde et al., Oncotarget 2017 [209]) and **f** *CD74*. The red dashed line depicts the threshold for overexpression, defined as median expression in BMCPs plus 3 times the standard deviation. Two exemplary patients are highlighted with red (patient 1) and yellow (patient 2) dot for *CD38* (in panel **a**). MGUS, monoclonal gammopathy of undetermined significance; AMM, asymptomatic multiple myeloma, MM, symptomatic multiple myeloma; MMR, relapsed myeloma.

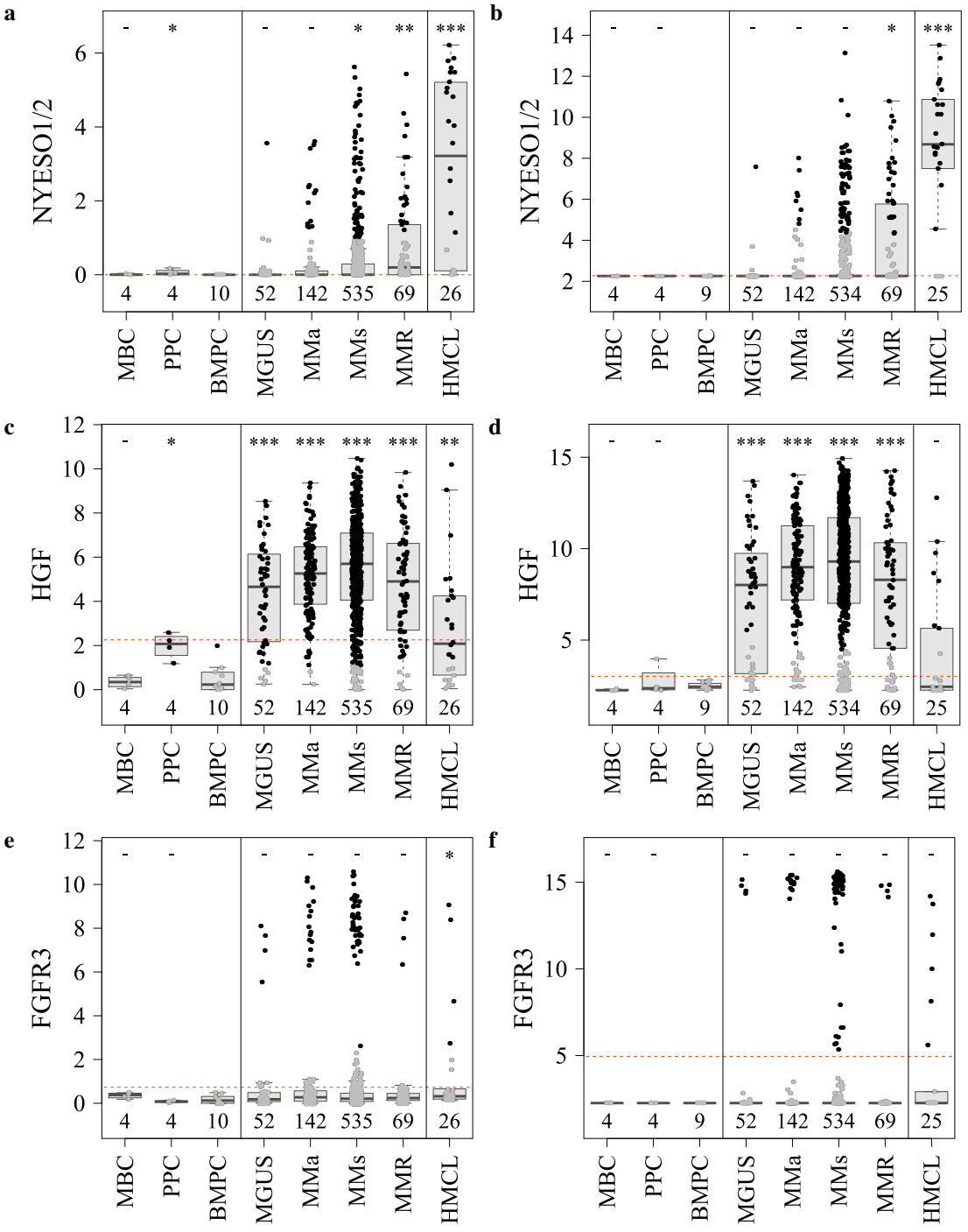

Figure A.10: Expression of *NYESO1/2*, *HGF* and *FGFR3*. The whole cohort is assessed. Expression is grouped by disease entity, compared to non-malignant cells and precursors (memory B cells (MBC), polyclonal plasmablastic cells (PPC), bone marrow plasma cells (BMPC)) and human myeloma cell lines (HMCL). Significant differences in expression compared to BMPCs are indicated by 1, 2 or 3 asterisks, for p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. RNA-seq expression is depicted for **a** *NYESO1/2* (published with smaller cohort in Schmitt, ..., Emde et al., Oncotarget 2017 [209]), **c** *HGF* and **e** *FGFR3*. Microarray expression is depicted for **b** *NYESO1/2* (published with altered sample cohort in Schmitt, ..., Emde et al., Oncotarget 2017 [209]), **d** *HGF* and **f** *FGFR3*. The red dashed line depicts the threshold for overexpression, defined as median expression in BMCPs plus 3 times the standard deviation. MGUS, monoclonal gammopathy of undetermined significance; AMM, asymptomatic multiple myeloma, MM, symptomatic multiple myeloma; MMR, relapsed myeloma.

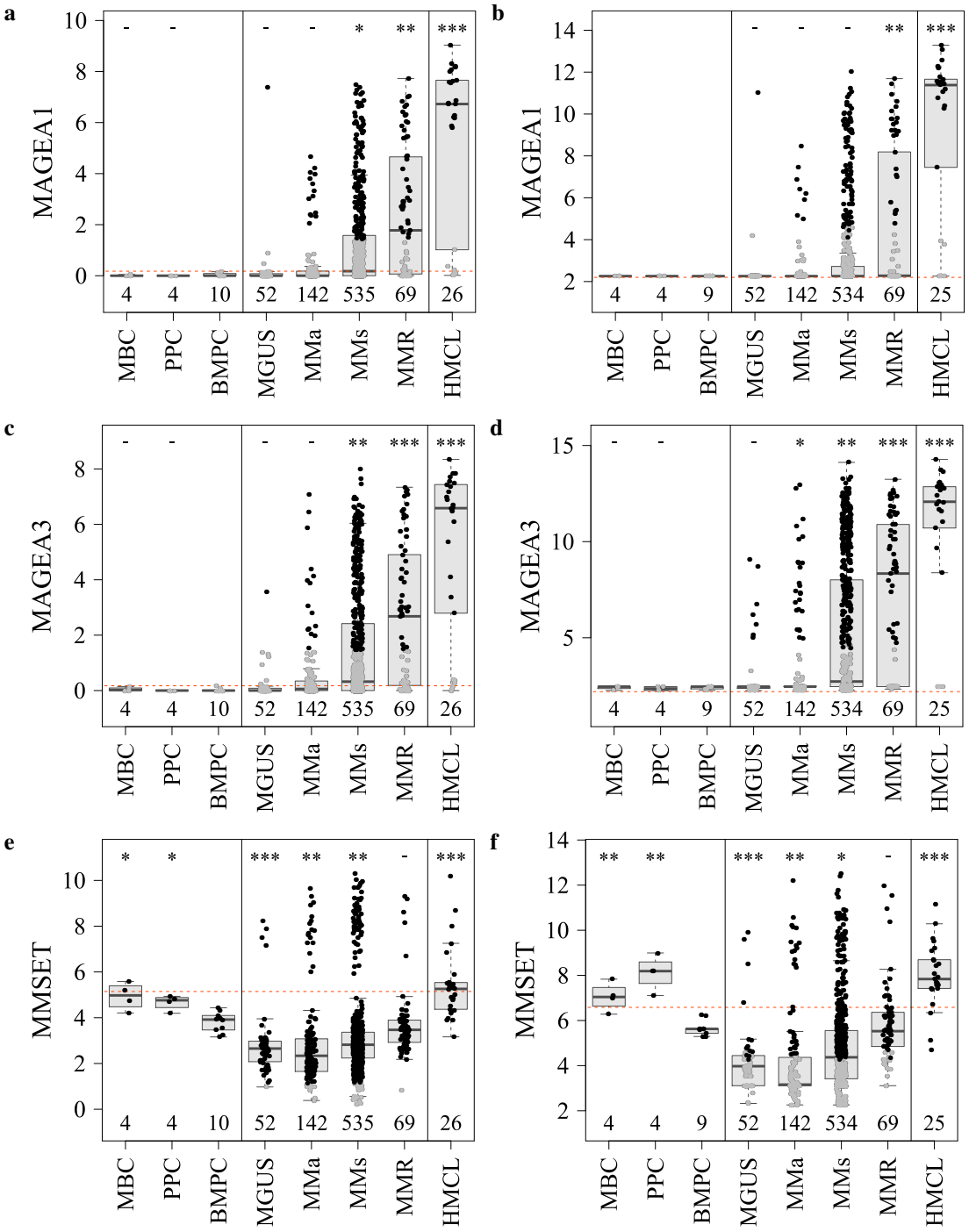

Figure A.11: Expression of *MAGEA1*, *MAGEA3* and *MMSET*. The whole cohort is assessed. Expression is grouped by disease entity, compared to non-malignant cells and precursors (memory B cells (MBC), polyclonal plasmablastic cells (PPC), bone marrow plasma cells (BMPC)) and human myeloma cell lines (HMCL). Significant differences in expression compared to BMPCs are indicated by 1, 2 or 3 asterisks, for p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. RNA-seq expression is depicted for **a** *MAGEA1*, **c** *MAGEA3* (published with a smaller cohort in Schmitt, ..., Emde et al., Oncotarget 2017 [209]) and **e** *MMSET*. Microarray expression is depicted for **b** *MAGEA1*, **d** *MAGEA3* (published with an altered sample cohort in Schmitt, ..., Emde et al., Oncotarget 2017 [209]) and **f** *MMSET*. The red dashed line depicts the threshold for overexpression, defined as median expression in BMCPs plus 3 times the standard deviation. MGUS, monoclonal gammopathy of undetermined significance; AMM, asymptomatic multiple myeloma, MM, symptomatic multiple myeloma; MMR, relapsed myeloma.

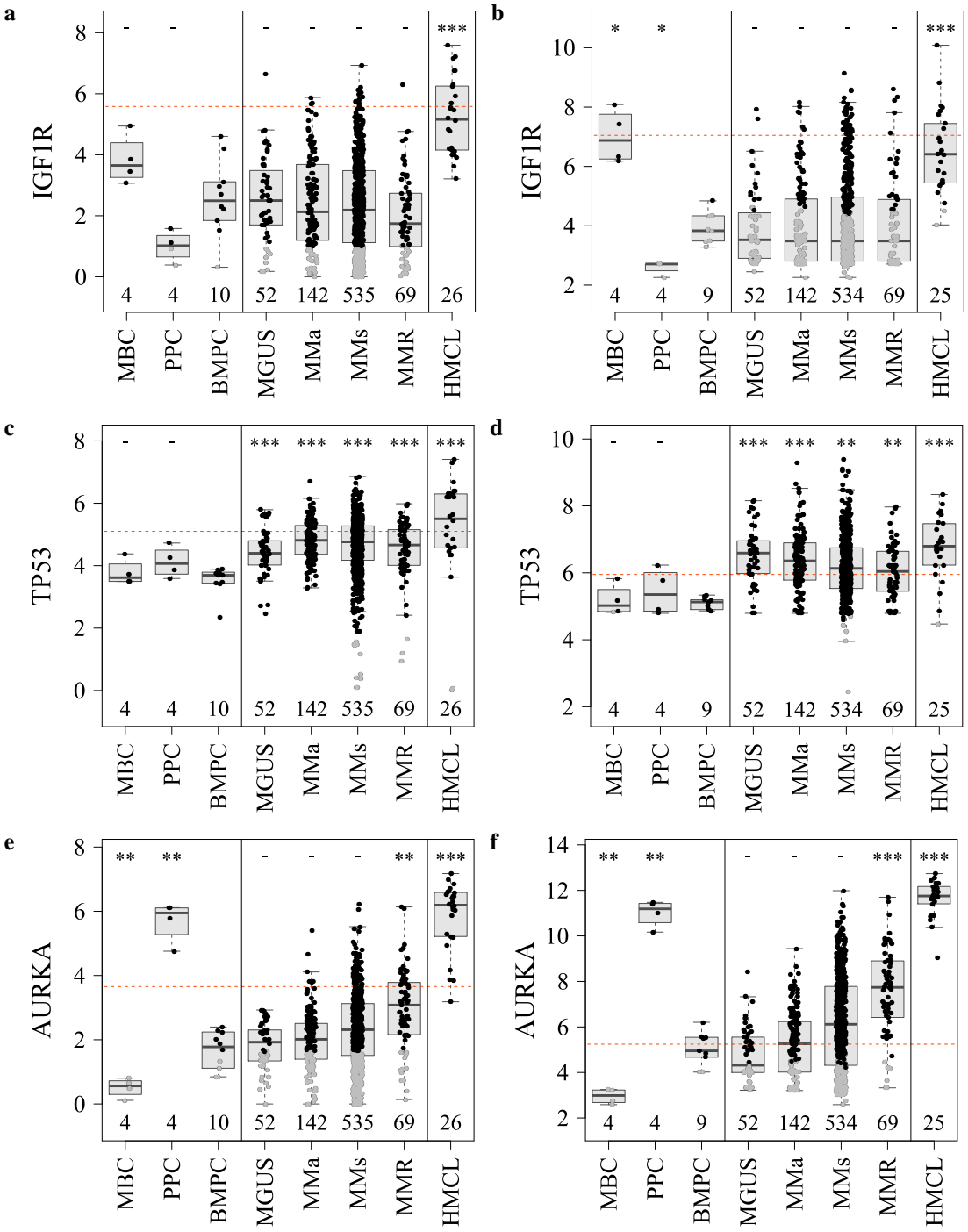

Figure A.12: Expression of *IGF1R*, *TP53* and *AURKA*. The whole cohort is assessed. Expression is grouped by disease entity, compared to non-malignant cells and precursors (memory B cells (MBC), polyclonal plasmablastic cells (PPC), bone marrow plasma cells (BMPC)) and human myeloma cell lines (HMCL). Significant differences in expression compared to BMPCs are indicated by 1, 2 or 3 asterisks, for p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. RNA-seq expression is depicted for **a** *IGF1R*, **c** *TP53* and **e** *AURKA*. Microarray expression is depicted for **b** *IGF1R*, **d** *TP53* and **f** *AURKA*. The red dashed line depicts the threshold for overexpression, defined as median expression in BMCPs plus 3 times the standard deviation. MGUS, monoclonal gammopathy of undetermined significance; AMM, asymptomatic multiple myeloma, MM, symptomatic multiple myeloma; MMR, relapsed myeloma.

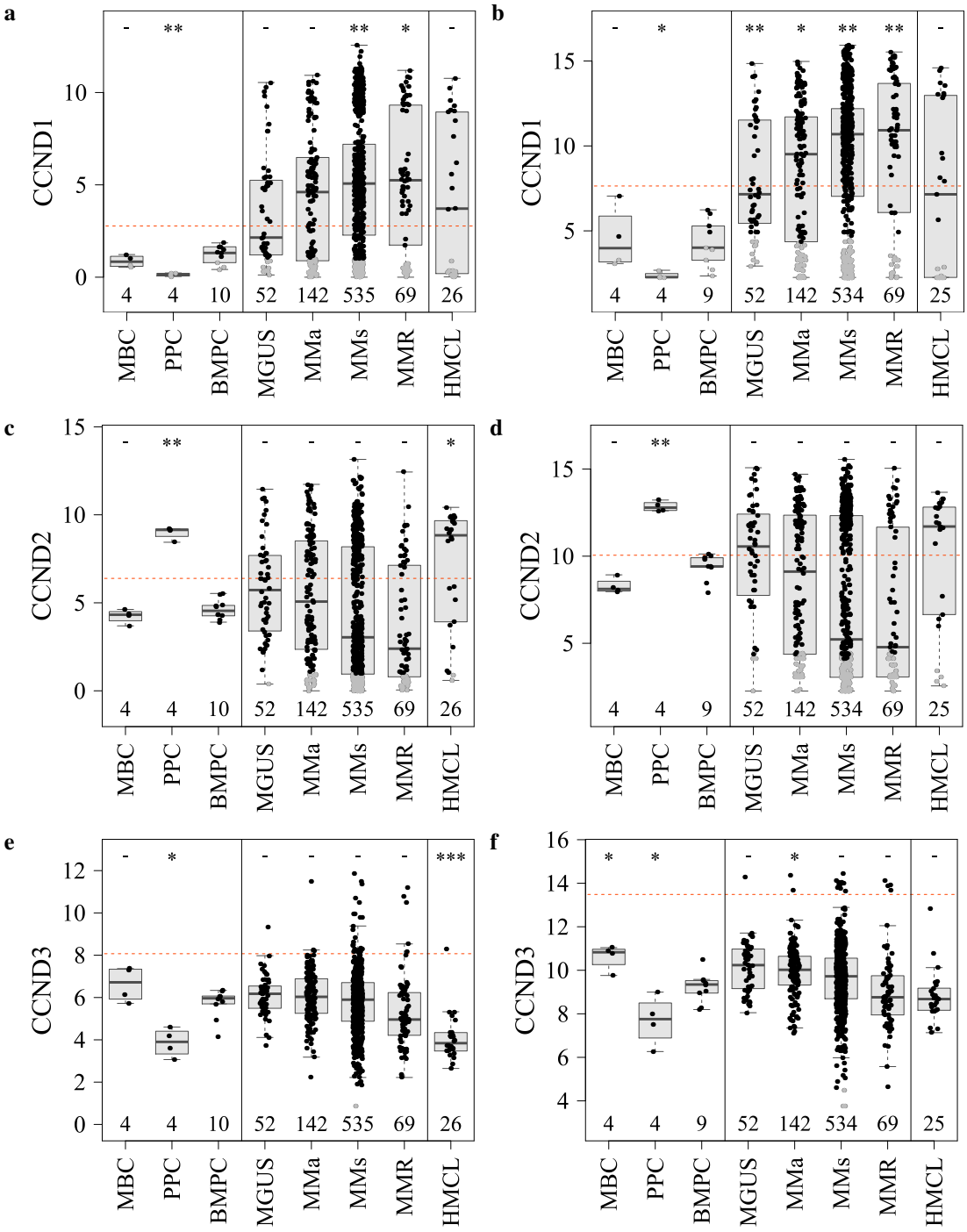

Figure A.13: Expression of *CCND1*, *CCND2* and *CCND3*. The whole cohort is assessed. Expression is grouped by disease entity, compared to non-malignant cells and precursors (memory B cells (MBC), polyclonal plasmablastic cells (PPC), bone marrow plasma cells (BMPC)) and human myeloma cell lines (HMCL). Significant differences in expression compared to BMPCs are indicated by 1, 2 or 3 asterisks, for p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. RNA-seq expression is depicted for **a** *CCND1*, **c** *CCND2* and **e** *CCND3*. Microarray expression is depicted for **b** *CCND1*, **d** *CCND2* and **f** *CCND3*. The red dashed line depicts the threshold for overexpression, defined as median expression in BMCPs plus 3 times the standard deviation. MGUS, monoclonal gammopathy of undetermined significance; AMM, asymptomatic multiple myeloma, MM, symptomatic multiple myeloma; MMR, relapsed myeloma; HMCL, human myeloma cell line.

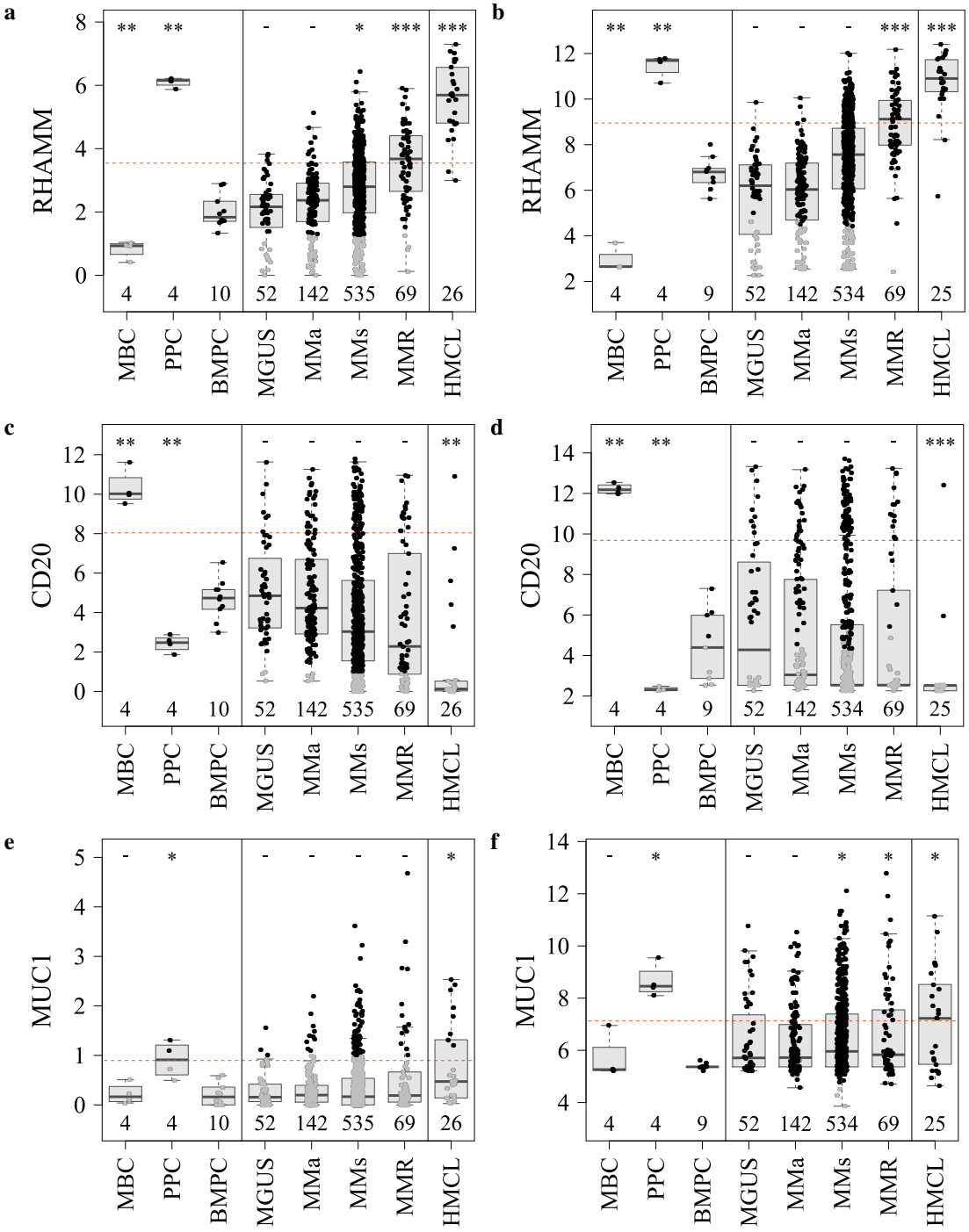

Figure A.14: Expression of *RHAMM*, *CD20* and *MUC1*. The whole cohort is assessed. Expression is grouped by disease entity, compared to non-malignant cells and precursors (memory B cells (MBC), polyclonal plasmablastic cells (PPC), bone marrow plasma cells (BMPC)) and human myeloma cell lines (HMCL). Significant differences in expression compared to BMPCs are indicated by 1, 2 or 3 asterisks, for p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. RNA-seq expression is depicted for **a** *RHAMM* (published with a smaller cohort in Schmitt, ..., Emde et al., Oncotarget 2017 [209]), **c** *CD20* and **e** *MUC1*. Microarray expression is depicted for **b** *RHAMM* (published with altered sample composition in Schmitt, ..., Emde et al., Oncotarget 2017 [209]), **d** *CD20* and **f** *MUC1*. The red dashed line depicts the threshold for overexpression, defined as median expression in BMCPs plus 3 times the standard deviation. MGUS, monoclonal gammopathy of undetermined significance; AMM, asymptomatic multiple myeloma, MM, symptomatic multiple myeloma; MMR, relapsed myeloma.

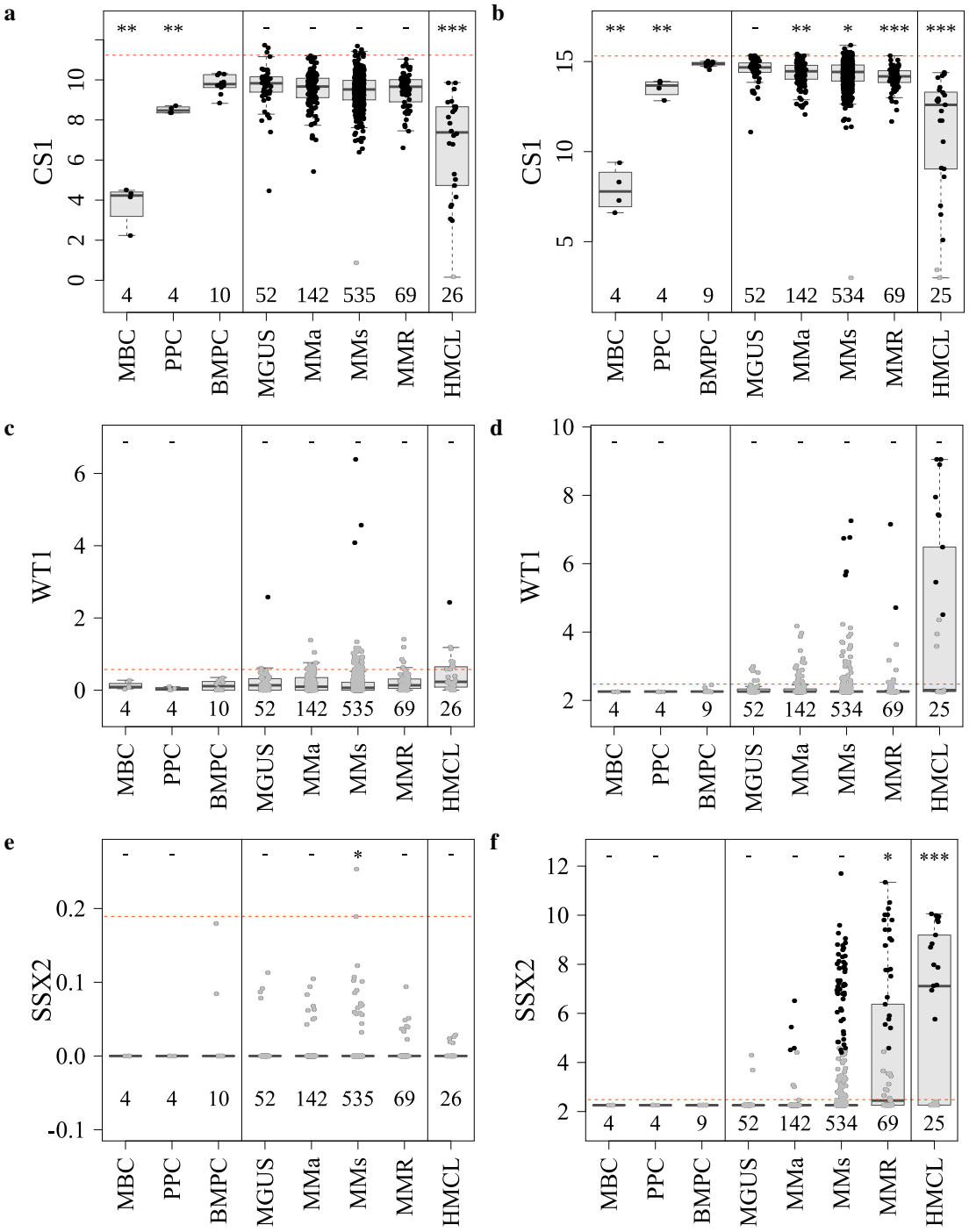

Figure A.15: Expression of *CS1*, *WT1* and *SSX2*. The whole cohort is assessed. Expression is grouped by disease entity, compared to non-malignant cells and precursors (memory B cells (MBC), polyclonal plasmablastic cells (PPC), bone marrow plasma cells (BMPC)) and human myeloma cell lines (HMCL). Significant differences in expression compared to BMPCs are indicated by 1, 2 or 3 asterisks, for p-values (p) smaller than 0.05, 0.01 and 0.001, respectively. RNA-seq expression is depicted for **a** *CSF1*, **c** *WT1* (published with a smaller cohort in Schmitt, ..., Emde et al., Oncotarget 2017 [209]) and **e** *SSX2*. Microarray expression is depicted for **b** *CSF1*, **d** *WT1* (published with altered sample composition in Schmitt, ..., Emde et al., Oncotarget 2017 [209]) and **f** *SSX2*. The red dashed line depicts the threshold for overexpression, defined as median expression in BMCPs plus 3 times the standard deviation. MGUS, monoclonal gammopathy of undetermined significance; AMM, asymptomatic multiple myeloma, MM, symptomatic multiple myeloma; MMR, relapsed myeloma.

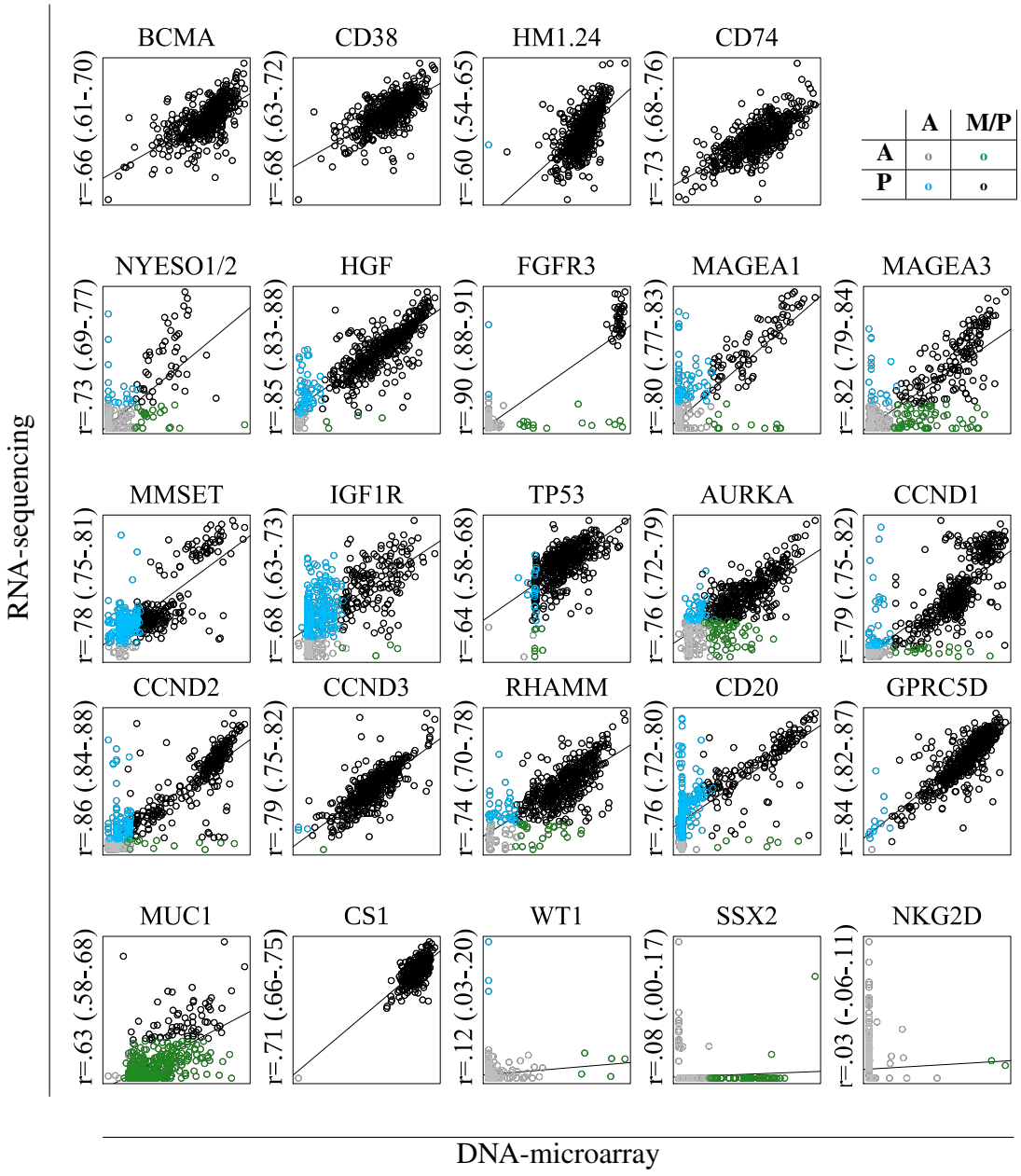

Figure A.16: Correlation of target expressions on RNA-sequencing (RNA-seq) and microarrays for all 534 symptomatic multiple myeloma patients (MM). *HM1.24*, *MAGEA3*, *RHAMM* and *WT1* are published with a smaller cohort in Schmitt, ..., Emde et al., Oncotarget 2017 [209]. The Pearson correlation coefficient (r) is displayed with its confidence interval. As the correlation and not the expression height is in the focus of this figure, axes per plot are not shown. Present (P), marginal present (M) and absent (A) expression is displayed. Absence in RNA-seq and microarray is depicted in grey, presence (P in RNA-seq and M or P in microarray) in black. Expression values only present in microarrays are depicted in green, expression values only defined present in RNA-seq in blue.

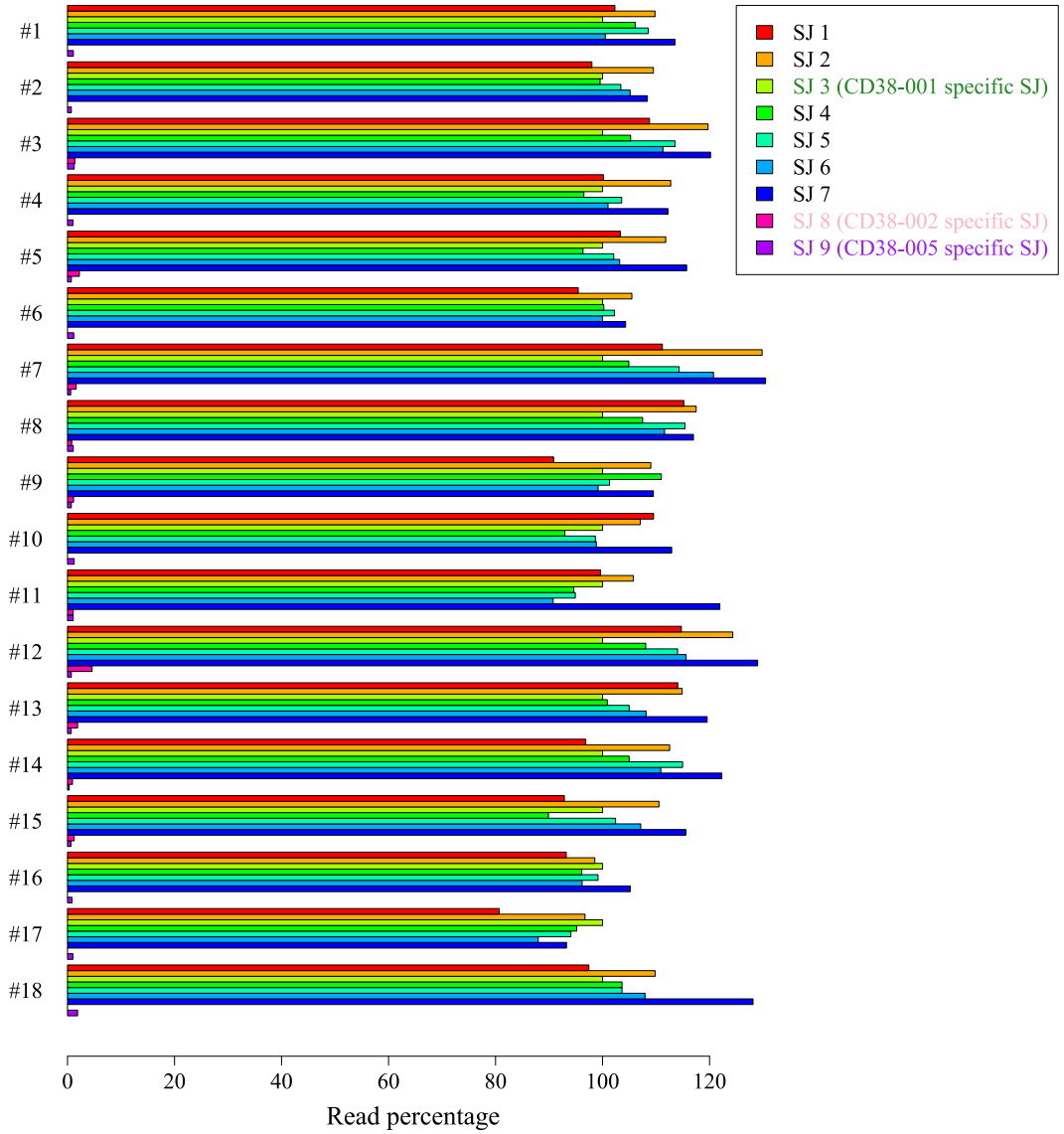

Figure A.17: Relative abundance of reads per splice junction for 18 exemplary patients. Depicted are the nine splice junctions of CD38 for 18 patients with present (spanned by at least 10 reads) CD38-005 specific splice junction (lilac) (adapted from Seckinger, ..., Emde et al., Frontiers in Immunology 2018 [213]). Shown is the relative abundance of the reads per splice junction in comparison to the CD38-001 specific splice junction. The maximum observed frequency of the alternatively spliced transcript CD38-005 is is 1.8%.

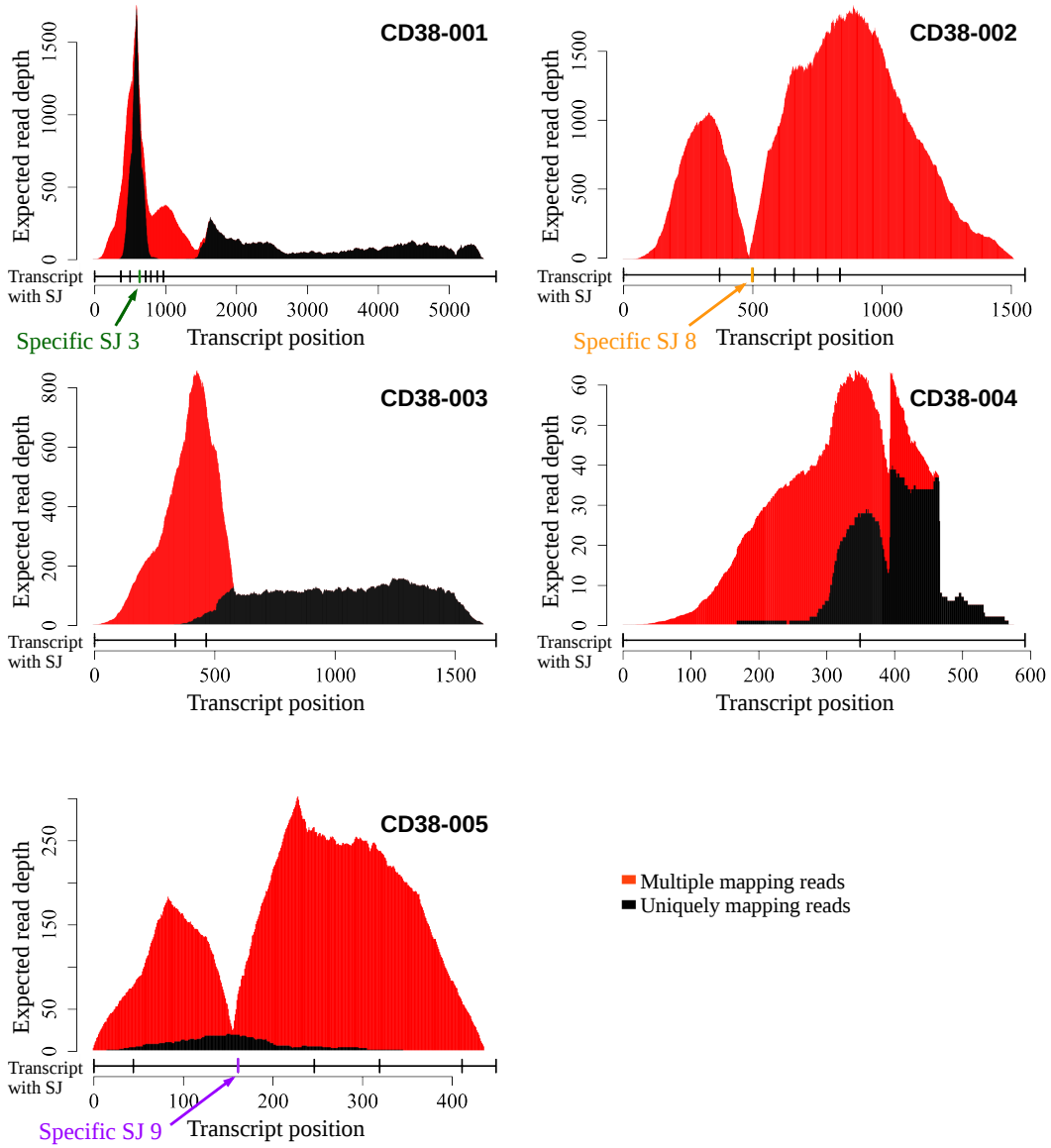

Figure A.18: CD38 transcript expression, exemplary patient 1. (adapted from Seckinger, ..., Emde et al., Frontiers in Immunology 2018 [213]). Depicted are multiple (red colour) and uniquely (black colour) mapping reads in a histogram per transcript. The transcript specific splice junctions per transcript are highlighted in coloured font. Exemplary patient 1 shows alternative splicing in low frequency. SJ: splice junction.

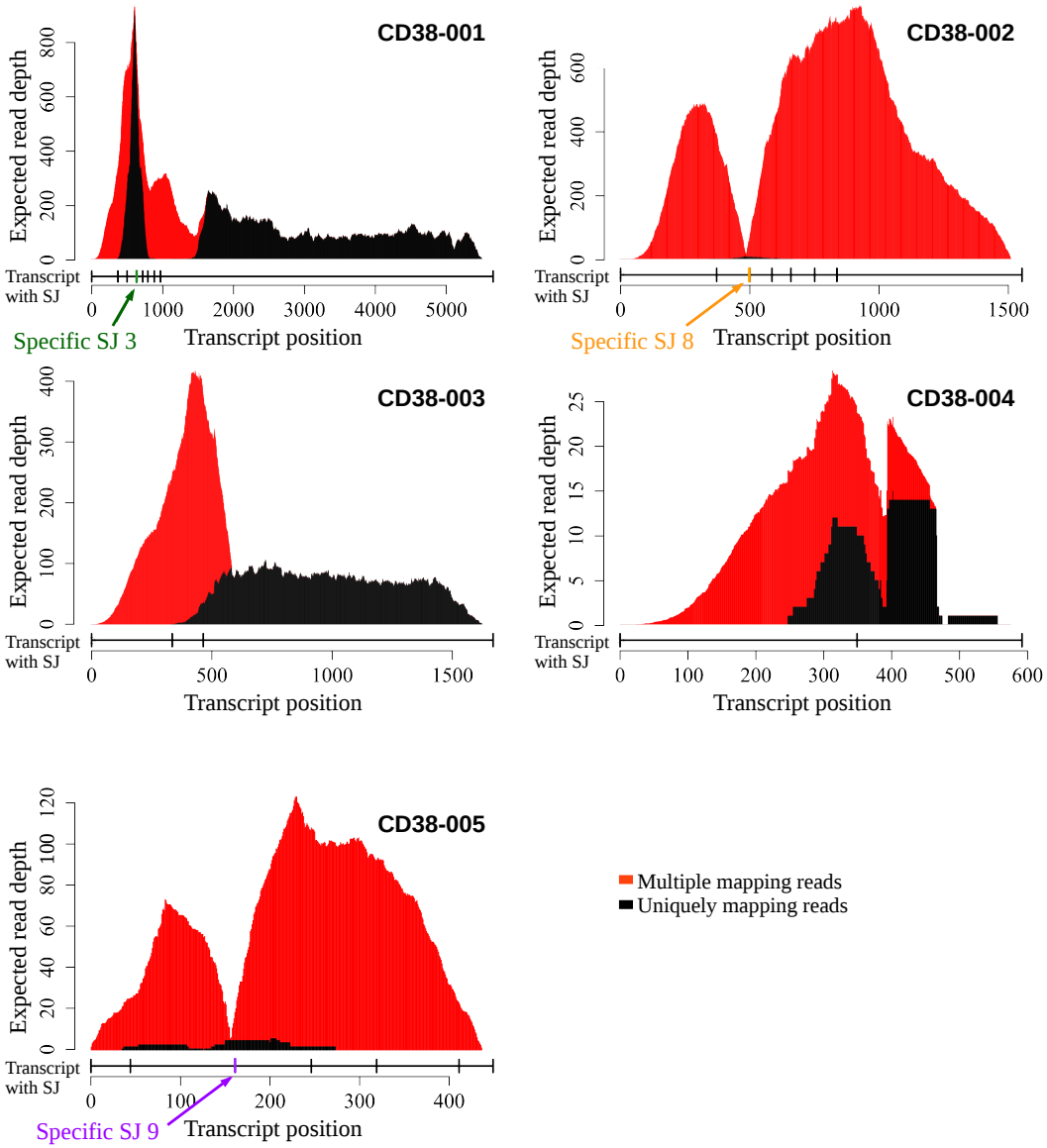

Figure A.19: CD38 transcript expression, exemplary patient 2. (adapted from Seckinger, ..., Emde et al., Frontiers in Immunology 2018 [213]). Depicted are multiple (red colour) and uniquely (black colour) mapping reads in a histogram per transcript. The transcript specific splice junctions per transcript are highlighted in coloured font. Exemplary patient 2 shows expression of CD38-001 only. SJ: splice junction.

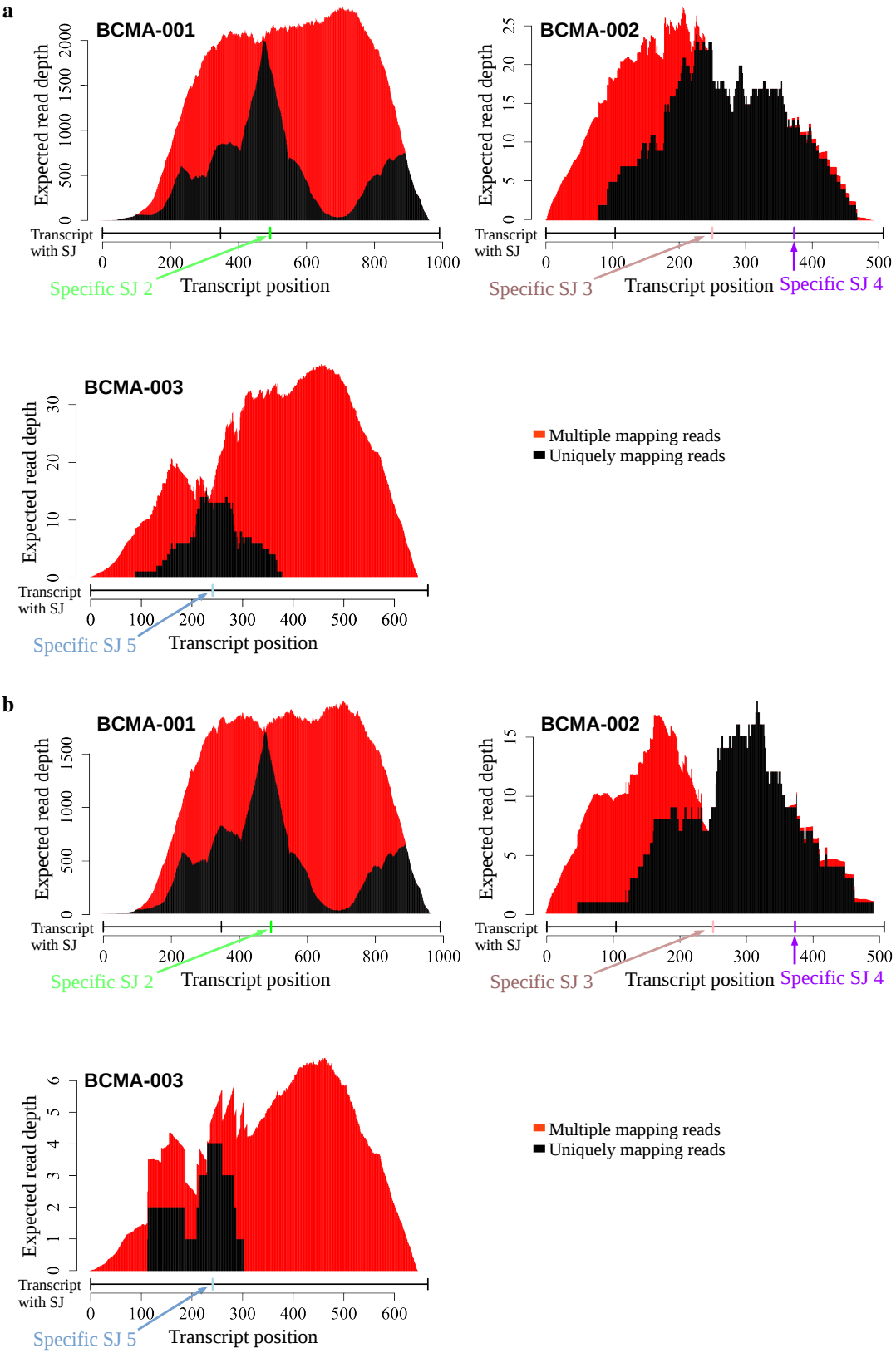

Figure A.20: BCMA transcript expression, exemplary patient 1. Depicted are multiple (red colour) and uniquely (black colour) mapping reads in a histogram per transcript. The transcript specific splice junctions per transcript are highlighted in coloured font. **a** patient 1 shows alternative splicing in low frequency. **b** patient 2 shows expression of transcript BCMA-001 only. SJ, splice junction.

B Supplementary Tables

Table B.4: Download links of used data. Overview of all downloaded data with corresponding links. GRCh38=Genome Reference Consortium Human Build 38; GTF=Gene Transfer Format

Data	Format	Release	Citation	Download	
				date	link
GRCh38 sequence	FASTA	77	Zerbino *et al.* [253]	2015.10.01, 13:46	`ftp://ftp.ensembl.org/pub/release-77/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz`
GRCh38 annotation	GTF	82	Zerbino *et al.* [253]	2015.10.05, 09:19	`ftp://ftp.ensembl.org/pub/release-82/gtf/homo_sapiens/Homo_sapiens.GRCh38.82.gtf.gz`
CoMMpass data	various	12	Craig *et al.* [51], MMRF *et al.* [169]	2019.05.09, 12:54	`https://research.themmrf.org`

Table B.5: List of applied tools. Tools are depicted with version and corresponding citation.

Tool	Version	Citation and/or link
bam-readcount	1.52.0	Gautier *et al.* [70]
Bioconductor	3.4	Gentleman *et al.* [72]
BLAT	97	Kent [119] (`https://www.ensembl.org/Homo_sapiens/Tools/Blast`)
Ensembl	97	Zerbino *et al.* [253] (`https://www.ensembl.org/`)
FastQC	0.11.4	Conesa *et al.* [49]
GeneAnnot	2.2	Chalifa-Caspi *et al.* [33, 34], Ferrari *et al.* [64] (`https://genecards.weizmann.ac.il/cgi-bin/geneannot/GA_search.pl`)
GeneCards	3.08	Stelzer *et al.* [226] (`https://www.genecards.org/`)
GEP-R	3.08	Meissner *et al.* [156] ()
htseq-count	0.6.0	Anders *et al.* [5]
R	3.3.2	R Core Team [188]
R-Studio	1.1.442	RStudio Team [203]
RSEM		Li and Dewey [140]
samtools	1.2	Li *et al.* [141]
STAR	2.4.2	Dobin *et al.* [58]

Table B.6: List of applied R packages. Overview of the R packages, their versions and references (R base packages are not listed).

Package	Version	Citation
clinfun	1.0.15	Seshan [218]
docval	1.1.2	Kostka [123]
edgeR	3.16.5	Chen *et al.* [38]
forestplot	1.7.2	Gordon and Lumley [78]
gcrma	2.46.0	Wu *et al.* [249]
GenomicRanges	1.26.4	Lawrence *et al.* [138]

hgu133plus2.db	3.2.3	Carlson [29]
jetset	3.4.0	Li *et al.* [142]
maxstat	0.7-25	Hothorn and Lausen [105]
pamr	1.55	Hastie *et al.* [90]
panp	1.44.0	Warren [246]
pec	2.5.4	Mogensen *et al.* [161]
RColorBrewer	1.1-2	Neuwirth [177]
reactome.db	1.58.0	Ligtenberg [143]
rms	5.1-2	Harrell [87]
S4Vectors	0.12.2	Pagès *et al.* [181]
showtext	0.5-1	Qiu [187]
survival	2.41-3	Therneau [228]
xtable	1.8-2	Dahl *et al.* [53]

Table B.7: Patient characteristics TG, VG, TeG. Depicted are the patient characteristics for the 535 symptomatic multiple myeloma patients (MM), divided in TG, VG and TeG. None of the groups showed significant differences regarding the depicted characteristics. TG: training group; VG: validation group; TeG: test group; ISS: International staging system; R-ISS: revised ISS; NA: not available; NR: not reached; n: number of patients.

		TG		**VG**		**TeG**	
Variable	**Level**	**n**	**%**	**n**	**%**	**n**	**%**
Sex	female	78	40.20	51	47.20	89	38.20
	male	116	59.80	57	52.80	144	61.80
Age [years]	≤60	108	55.70	62	57.40	117	50.20
	>60	86	44.30	46	42.60	116	49.80
Monoclonal protein [g/l]	<20	39	20.10	19	17.60	44	18.90
	≥20	27	13.90	11	10.20	29	12.40
	≥30	105	54.10	58	53.70	117	50.20
	NA	23	11.90	20	18.50	43	18.50
ISS stage	1	76	39.20	50	46.30	99	42.50
	2	63	32.50	29	26.90	69	29.60
	3	50	25.80	25	23.10	59	25.30
	NA	5	2.60	4	3.70	6	2.60
R-ISS stage	I	37	19.10	31	28.70	47	20.20
	II	88	45.40	38	35.20	100	42.90
	III	23	11.90	19	17.60	28	12.00
	NA	46	23.70	20	18.50	58	24.90
meta score risk	low	11	5.70	9	8.30	25	10.70
	medium	156	80.40	82	75.90	177	76.00
	high	21	10.80	13	12.00	25	10.70
	NA	6	3.10	4	3.70	6	2.60

Table B.8: Proportions of risk stratifications per patient cohort. The proportions are depicted regarding RNA-seq for symptomatic multiple myeloma patients (MM) in the test group (TeG), MM patients in the CoMMpass cohort (CP), asymptomatic MM patients (AMM) and relapsed MM patients (MMR). For comparison, the microarray stratification proportions are shown for the TeG. See also figure 3.16.

Stratification	Group	low risk [%]	medium risk [%]	high risk [%]
GPI	TeG	32.19	57.94	9.87
RPI	TeG	33.91	50.21	15.88
	CP	38.33	42.76	18.90
	AMM	64.79	33.10	2.11
	MMR	8.70	50.72	40.58
UAMS70	TeG	76.39	-	23.61
UAMS70-seq	TeG	82.83	-	17.17
	CP	78.10	-	21.90
	AMM	92.96	-	7.04
	MMR	71.01	-	28.99
RS	TeG	36.91	54.08	9.01
RS-seq	TeG	24.89	63.95	11.16
	CP	24.25	62.45	13.30
	AMM	43.66	54.93	1.41
	MMR	4.35	63.77	31.88
EMC92	TeG	79.40	-	20.60
EMC92-seq	TeG	73.82	-	26.18
	CP	69.88	-	30.12
	AMM	86.62	-	13.38
	MMR	42.03	-	57.97
IFM15	TeG	91.42	-	8.58
IFM15-seq	TeG	87.98	-	12.02
	CP	86.96	-	13.04
	AMM	97.89	-	2.11
	MMR	63.77	-	36.23
HDHRS-GEP	TeG	46.35	33.48	20.17
HDHRS	TeG	38.20	38.63	23.18
	CP	35.20	41.72	23.08
	AMM	65.49	28.87	5.63
	MMR	11.59	34.78	53.62

Table B.9: Translation of GPI genes. In this table the 50 genes of GPI are listed as microarray probeset and corresponding Ensembl Gene Identifiers (ENSG), determined with hgu133plus2.db in R. Correlation (r) and the percentage of present expression determined with the present/absent (PA) call (PA-seq) for RNA-sequencing (RNA-seq) and absent PA call for microarrays (nCO_1, see section 2.2.2) are listed in the table. Excluded genes are listed with exclusion reason: 1 is a correlation $r < 0.15$, 2 is a correlation $r < 0.4$ and a non-overlapp $nCO_1 > 0.3$, and 3 is an inconsistent translation (for details see section 2.5).

Probeset	ENSG	r	nCO_1 [%]	exclusion reason
200886_s_at	ENSG00000171314	0.36	0	-
205339_at	ENSG00000123473	0.47	55	-
221923_s_at	ENSG00000181163	0.48	0	-
213008_at	ENSG00000140525	0.52	0	-
204531_s_at	ENSG00000012048	0.52	9	-
205394_at	ENSG00000149554	0.53	0	-
204887_s_at	ENSG00000142731	0.56	0	-

221520_s_at	ENSG00000134690	0.57	3	-
209464_at	ENSG00000178999	0.57	0	-
212021_s_at	ENSG00000148773	0.58	6	-
204240_s_at	ENSG00000136824	0.58	2	-
38158_at	ENSG00000135476	0.58	2	-
219306_at	ENSG00000163808	0.59	3	-
219918_s_at	ENSG00000066279	0.59	20	-
205167_s_at	ENSG00000158402	0.59	2	-
203145_at	ENSG00000076382	0.6	3	-
209408_at	ENSG00000142945	0.6	1	-
201930_at	ENSG00000076003	0.61	1	-
220651_s_at	ENSG00000065328	0.61	0	-
204444_at	ENSG00000138160	0.61	25	-
204170_s_at	ENSG00000123975	0.62	0	-
212789_at	ENSG00000151503	0.62	0	-
203755_at	ENSG00000156970	0.62	10	-
218662_s_at	ENSG00000109805	0.63	7	-
203418_at	ENSG00000145386	0.63	1	-
203554_x_at	ENSG00000164611	0.63	0	-
204162_at	ENSG00000080986	0.64	6	-
210052_s_at	ENSG00000088325	0.64	1	-
222077_s_at	ENSG00000161800	0.64	1	-
218755_at	ENSG00000112984	0.65	1	-
203764_at	ENSG00000126787	0.65	2	-
218542_at	ENSG00000138180	0.66	16	-
209714_s_at	ENSG00000100526	0.66	0	-
204318_s_at	ENSG00000075218	0.66	0	-
203967_at	ENSG00000094804	0.67	0	-
219588_s_at	ENSG00000146918	0.68	3	-
202954_at	ENSG00000175063	0.68	2	-
204026_s_at	ENSG00000122952	0.68	8	-
202705_at	ENSG00000157456	0.68	3	-
202870_s_at	ENSG00000117399	0.71	0	-
202095_s_at	ENSG00000089685	0.71	1	-
206102_at	ENSG00000101003	0.72	4	-
214710_s_at	ENSG00000134057	0.72	1	-
201897_s_at	ENSG00000173207	0.72	0	-
204092_s_at	ENSG00000087586	0.73	3	-
204641_at	ENSG00000117650	0.73	13	-
201202_at	ENSG00000132646	0.73	0	-
203213_at	ENSG00000170312	0.74	3	-
203362_s_at	ENSG00000164109	0.74	1	-
209642_at	ENSG00000169679	0.78	1	-

Table B.10: Translation of UAMS70 genes. In this table, the 70 genes comprising the UAMS70 score are listed as microarray probeset and corresponding Ensembl Gene Identifiers (ENSG), determined with hgu133plus2.db in R. Correlation (r) and the percentage of present expression determined with the present/absent (PA) call (PA-seq) for RNA-sequencing (RNA-seq) and absent PA call for microarrays (nCO_1, see section 2.2.2) are listed in the table. Excluded genes are listed with exclusion reason: 1 is a correlation $r < 0.15$, 2 is a correlation $r < 0.4$ and a non-overlapp $nCO_1 > 0.3$, and 3 is an inconsistent translation (for details see section 2.5).

Probeset	**ENSG**	**r**	**nCO_1 [%]**	**exclusion reason**
227547_at	-	-	-	-
237964_at	-	-	-	-
1557277_a_at	ENSG00000249346	-0.03	0	1
244686_at	ENSG00000070814	-0.07	0	1
1554736_at	ENSG00000137962	0	0	1
210334_x_at	ENSG00000089685	0.18	0	-
200634_at	ENSG00000108518	0.22	0	-
204016_at	ENSG00000011376	0.24	0	-
200638_s_at	ENSG00000164924	0.27	0	-
201921_at	ENSG00000242616	0.29	100	2
216194_s_at	ENSG00000105254	0.3	0	-
48106_at	ENSG00000211584	0.31	0	-
200966_x_at	ENSG00000149925	0.31	0	-
226954_at	ENSG00000107341	0.32	0	-
201091_s_at	ENSG00000122565	0.32	0	-
213607_x_at	ENSG00000008130	0.33	0	-
200850_s_at	ENSG00000168710	0.35	1	-
218947_s_at	ENSG00000107951	0.36	0	-
208117_s_at	ENSG00000001497	0.37	0	-
1555274_a_at	ENSG00000138018	0.4	0	-
224523_s_at	ENSG00000184220	0.41	0	-
243011_at	ENSG00000144815	0.43	0	-
225834_at	ENSG00000263513	0.44	68	3
224200_s_at	ENSG00000070950	0.45	0	-
210460_s_at	ENSG00000159352	0.45	0	-
213535_s_at	ENSG00000103275	0.46	0	-
208931_s_at	ENSG00000129351	0.46	0	-
205235_s_at	ENSG00000138182	0.46	0	-
225834_at	ENSG00000196550	0.47	64	-
200750_s_at	ENSG00000132341	0.47	0	-
209717_at	ENSG00000067208	0.49	1	-
58696_at	ENSG00000178896	0.5	0	-
209740_s_at	ENSG00000006757	0.5	0	-
213310_at	ENSG00000123908	0.5	0	3
225082_at	ENSG00000119203	0.51	0	-
225834_at	ENSG00000188610	0.51	58	3
206364_at	ENSG00000118193	0.51	8	-
1565951_s_at	ENSG00000203668	0.51	23	-
201947_s_at	ENSG00000166226	0.52	0	-
204023_at	ENSG00000163918	0.52	0	-
222417_s_at	ENSG00000089006	0.53	0	-
220789_s_at	ENSG00000136270	0.53	0	-
238952_x_at	ENSG00000185869	0.55	2	3
201105_at	ENSG00000100097	0.55	0	-
1555864_s_at	ENSG00000131828	0.55	0	-

201231_s_at	ENSG00000074800	0.55	0	-
222495_at	ENSG00000215717	0.56	0	-
218924_s_at	ENSG00000117151	0.57	0	-
230192_at	ENSG00000204977	0.58	0	-
213628_at	ENSG00000121940	0.58	0	-
212435_at	ENSG00000197323	0.58	4	-
226936_at	ENSG00000203760	0.58	10	-
202838_at	ENSG00000179163	0.58	0	-
203432_at	ENSG00000120802	0.59	0	-
219918_s_at	ENSG00000066279	0.59	20	-
225834_at	ENSG00000215784	0.6	68	-
212533_at	ENSG00000166483	0.61	40	-
227278_at	ENSG00000197780	0.62	6	-
201614_s_at	ENSG00000175792	0.64	0	-
242488_at	ENSG00000133019	0.66	5	-
202729_s_at	ENSG00000049323	0.66	4	-
204033_at	ENSG00000071539	0.69	2	-
206513_at	ENSG00000163568	0.69	0	-
225582_at	ENSG00000148841	0.71	5	-
211576_s_at	ENSG00000173638	0.71	2	-
213194_at	ENSG00000169855	0.71	20	-
202345_s_at	ENSG00000164687	0.72	7	-
200916_at	ENSG00000158710	0.72	0	-
201897_s_at	ENSG00000173207	0.73	0	-
204092_s_at	ENSG00000087586	0.73	2	-
221970_s_at	ENSG00000130935	0.74	0	-
217901_at	ENSG00000046604	0.74	8	-
206332_s_at	ENSG00000163565	0.77	0	-

Table B.11: Translation of RS genes. In this table, the 19 genes comprising the RS score are listed as microarray probeset and corresponding Ensembl Gene Identifiers (ENSG), determined with hgu133plus2.db in R. Correlation (r) and the percentage of present expression determined with the present/absent (PA) call (PA-seq) for RNA-sequencing (RNA-seq) and absent PA call for microarrays (nCO_1, see section 2.2.2) is listed in the table. Excluded genes are listed with exclusion reason: 1 is a correlation $r < 0.15$, 2 is a correlation $r < 0.4$ and a non-overlapp $nCO_1 > 0.3$, and 3 is an inconsistent translation (for details see section 2.5).

Probeset	**ENSG**	**r**	**nCO_1 [%]**	**exclusion reason**
233660_at	ENSG00000103966	0.26	0	3
204031_s_at	ENSG00000197111	0.34	0	-
218460_at	ENSG00000164818	0.38	7	-
235353_at	ENSG00000091490	0.41	0	3
225687_at	ENSG00000101447	0.42	7	-
219978_s_at	ENSG00000137804	0.45	1	-
214464_at	ENSG00000143776	0.46	2	-
225272_at	ENSG00000141504	0.48	0	-
234672_s_at	ENSG00000058804	0.54	0	-
221520_s_at	ENSG00000134690	0.57	4	-
218726_at	ENSG00000123485	0.57	1	-
203358_s_at	ENSG00000106462	0.58	24	-
226936_at	ENSG00000203760	0.58	10	-
229553_at	ENSG00000165434	0.61	24	-
203755_at	ENSG00000156970	0.62	10	-
226980_at	ENSG00000035499	0.65	9	-
203764_at	ENSG00000126787	0.66	3	-
220945_x_at	ENSG00000111261	0.76	8	-
219855_at	ENSG00000196368	0.81	4	-

Table B.12: Translation of EMC92 genes. In this table, the 92 genes comprising the EMC92 score are listed as microarray probeset and corresponding Ensembl Gene Identifiers (ENSG), determined with hgu133plus2.db in R. Correlation (r) and the percentage of present expression determined with the present/absent (PA) call (PA-seq) for RNA-sequencing (RNA-seq) and absent PA call for microarrays (nCO_1, see section 2.2.2) are listed in the table. Excluded genes are listed with exclusion reason: 1 is a correlation $r < 0.15$, 2 is a correlation $r < 0.4$ and a non-overlapp $nCO_1 > 0.3$, and 3 is an inconsistent translation (for details see section 2.5).

Probeset	ENSG	r	nCO_1 [%]	exclusion reason
243018_at	-	-	-	-
231989_s_at	ENSG00000183889	-0.02	2	1,3
231989_s_at	ENSG00000233024	-0.06	2	1,3
233399_x_at	ENSG00000196922	-0.19	3	1
213350_at	ENSG00000142534	0.03	0	1,3
230034_x_at	ENSG00000182154	0.04	0	1,3
216473_x_at	ENSG00000281591	0.04	100	1,2,3
223811_s_at	ENSG00000164828	0.05	0	1,3
215181_at	ENSG00000149654	0.07	99	1,2
223811_s_at	ENSG00000239857	0.1	95	1,2
217548_at	ENSG00000242498	0.11	0	1,3
216473_x_at	ENSG00000260596	0.12	100	1,2,3
242180_at	ENSG00000130167	0.13	85	1,2
221755_at	ENSG00000173442	0.13	50	1,2
231989_s_at	ENSG00000237296	0.16	2	-
208232_x_at	ENSG00000157168	0.16	79	2,3
231989_s_at	ENSG00000183793	0.18	1	3
210334_x_at	ENSG00000089685	0.18	0	-
231989_s_at	ENSG00000180747	0.19	1	-
231989_s_at	ENSG00000185864	0.2	1	3
201102_s_at	ENSG00000141959	0.2	0	-
214482_at	ENSG00000089775	0.2	0	-
200933_x_at	ENSG00000198034	0.2	0	-
221677_s_at	ENSG00000159147	0.21	73	2
202542_s_at	ENSG00000164022	0.24	0	-
200775_s_at	ENSG00000165119	0.25	0	-
201398_s_at	ENSG00000067167	0.26	0	-
231989_s_at	ENSG00000243716	0.27	1	3
217732_s_at	ENSG00000136156	0.3	0	-
202884_s_at	ENSG00000137713	0.32	0	-
208967_s_at	ENSG00000004455	0.33	0	-
218365_s_at	ENSG00000117593	0.34	0	-
219510_at	ENSG00000051341	0.34	5	-
208904_s_at	ENSG00000233927	0.34	0	-
231989_s_at	ENSG00000198064	0.35	2	3
238662_at	ENSG00000134146	0.36	0	-
208732_at	ENSG00000104388	0.36	0	-
208942_s_at	ENSG00000008952	0.36	0	-
239054_at	ENSG00000163935	0.38	0	-
205046_at	ENSG00000138778	0.39	0	-
214150_x_at	ENSG00000113732	0.39	0	-
211714_x_at	ENSG00000196230	0.4	0	-
209026_x_at	ENSG00000196230	0.4	0	-
217824_at	ENSG00000198833	0.41	0	-
208667_s_at	ENSG00000100380	0.44	0	-
226742_at	ENSG00000152700	0.44	0	-
222713_s_at	ENSG00000183161	0.44	1	-

225366_at	ENSG00000169299	0.46	0	-
212788_x_at	ENSG00000087086	0.46	0	-
233437_at	ENSG00000109158	0.47	0	-
221826_at	ENSG00000174606	0.47	0	-
219550_at	ENSG00000154134	0.48	0	-
231210_at	ENSG00000168070	0.48	0	-
238116_at	ENSG00000168589	0.49	0	-
217852_s_at	ENSG00000134108	0.5	0	-
202728_s_at	ENSG00000049323	0.51	4	-
202842_s_at	ENSG00000128590	0.52	0	-
226217_at	ENSG00000162695	0.52	0	-
208747_s_at	ENSG00000182326	0.52	0	-
202553_s_at	ENSG00000117614	0.52	0	-
202813_at	ENSG00000059588	0.54	26	-
238780_s_at	ENSG00000120457	0.56	4	-
213002_at	ENSG00000277443	0.56	2	3
212055_at	ENSG00000134779	0.57	0	-
221041_s_at	ENSG00000119899	0.57	0	-
203145_at	ENSG00000076382	0.57	1	-
213007_at	ENSG00000140525	0.57	1	-
38158_at	ENSG00000135476	0.58	2	-
200875_s_at	ENSG00000101361	0.58	0	-
228416_at	ENSG00000121989	0.59	0	-
215177_s_at	ENSG00000091409	0.6	0	-
207618_s_at	ENSG00000074582	0.6	0	-
201930_at	ENSG00000076003	0.61	1	-
202532_s_at	ENSG00000228716	0.61	18	-
222154_s_at	ENSG00000196141	0.62	2	-
218662_s_at	ENSG00000109805	0.63	8	-
225601_at	ENSG00000029993	0.64	1	-
212282_at	ENSG00000109084	0.64	1	-
218355_at	ENSG00000090889	0.66	1	-
226218_at	ENSG00000168685	0.66	5	-
201292_at	ENSG00000131747	0.67	1	-
200701_at	ENSG00000119655	0.67	0	-
211963_s_at	ENSG00000162704	0.68	0	-
AFFX-HUMISGF3A/M97935_MA_at	ENSG00000115415	0.68	0	-
204026_s_at	ENSG00000122952	0.69	8	-
221606_s_at	ENSG00000198157	0.69	2	-
217728_at	ENSG00000197956	0.69	0	-
222680_s_at	ENSG00000143476	0.69	1	-
209683_at	ENSG00000197872	0.7	1	-
202322_s_at	ENSG00000152904	0.71	0	-
231738_at	ENSG00000113212	0.72	14	-
201795_at	ENSG00000143815	0.72	0	-
202107_s_at	ENSG00000073111	0.72	4	-
220351_at	ENSG00000129048	0.72	4	-
223381_at	ENSG00000143228	0.73	2	-
201555_at	ENSG00000112118	0.74	1	-
214612_x_at	ENSG00000197172	0.76	9	-
201307_at	ENSG00000138758	0.82	1	-
206204_at	ENSG00000115290	0.82	1	-
224009_x_at	ENSG00000073737	0.85	1	-
204379_s_at	ENSG00000068078	0.92	3	-

Table B.13: Translation of IFM15 genes. In this table, the 15 genes comprising IFM15 score are listed as microarray probeset and corresponding Ensembl Gene Identifiers (ENSG), determined with hgu133plus2.db in R. Correlation (r) and the percentage of present expression determined with the present/absent (PA) call (PA-seq) for RNA-sequencing (RNA-seq) and absent PA call for microarrays (nCO_1, see section 2.2.2) are listed in the table. Excluded genes are listed with exclusion reason: 1 is a correlation $r < 0.15$, 2 is a correlation $r < 0.4$ and a non-overlapp $nCO_1 > 0.3$, and 3 is an inconsistent translation (for details see section 2.5).

Probeset	ENSG	r	nCO_1 [%]	exclusion reason
202470_s_at	ENSG00000111605	0.21	0	-
202486_at	ENSG00000141385	0.33	28	-
228677_s_at	ENSG00000105122	0.41	87	-
200779_at	ENSG00000128272	0.44	0	-
204072_s_at	ENSG00000073910	0.47	1	-
217752_s_at	ENSG00000133313	0.61	0	-
228737_at	ENSG00000124191	0.62	5	-
202951_at	ENSG00000112079	0.68	3	-
203657_s_at	ENSG00000174080	0.69	9	-
208644_at	ENSG00000143799	0.69	0	-
209683_at	ENSG00000197872	0.7	1	-
212098_at	ENSG00000152127	0.7	32	-
200783_s_at	ENSG00000117632	0.72	0	-
201425_at	ENSG00000111275	0.73	2	-
231736_x_at	ENSG00000008394	0.8	2	-

Table B.14: Translation of HDHRS genes. In this table, the 53 genes comprising the HDHRS score are listed as Ensembl Gene Identifiers (ENSG) and corresponding microarray probesets and gene symbols (SYMBOL), determined with hgu133plus2.db in R. Best probesets determined with jetset package in R are depicted in bold. Additionally, the HDHRS factor, correlation (r) and the percentage of present expression determined with the present/absent (PA) call (PA-seq) for RNA-sequencing (RNA-seq) and absent PA call for microarrays (nCO_1, see section 2.2.2) are listed in the table.

ENSG	SYMBOL	Probeset	Factor	r	nCO_1 [%]
ENSG00000123965	-	-	1	-	-
ENSG00000203819	-	-	1	-	-
ENSG00000211592	-	-	-1	-	-
ENSG00000234722	-	-	1	-	-
ENSG00000236675	-	-	1	-	-
ENSG00000240086	-	-	1	-	-
ENSG00000272525	-	-	1	-	-
ENSG00000273002	-	-	1	-	-
ENSG00000166415	WDR72	1563874_at, 227174_at, 236741_at, **238253_at**	1	-0.03	0
ENSG00000237424	FOXD2-AS1	**224456_s_at**, 224457_at	1	0.01	0
ENSG00000247774	PCED1B-AS1	**241947_at**	-1	0.16	0
ENSG00000276234	TADA2A	209938_at, **210537_s_at**	1	0.23	0
ENSG00000151239	TWF1	**201745_at**, 214007_s_at, 214008_at, 243033_at, 244199_at	-1	0.3	0
ENSG00000158427	TMSB15B	1570039_at, **205347_s_at**, **214051_at**	1	0.3	85
ENSG00000188092	GPR89B	1562412_at, **220642_x_at**, 222140_s_at, 223531_x_at, 225463_x_at	1	0.3	17
ENSG00000100629	CEP128	1557755_at, 1557756_a_at, 232635_at, **233859_at**, 244033_at	1	0.32	0
ENSG00000143379	SETDB1	**203155_at**, 214197_s_at	1	0.44	1
ENSG00000105438	KDELR1	**1555575_a_at**, 200922_at	-1	0.45	0

ENSG00000250317	SMIM20	**225014_at**	-1	0.46	0
ENSG00000172057	ORMDL3	**223259_at**, 235136_at	-1	0.48	0
ENSG00000167077	MEI1	1554208_at, 1564621_a_at, **230011_at**	-1	0.49	0
ENSG00000133678	TMEM254	**218174_s_at**, 222545_s_at	1	0.5	0
ENSG00000203668	CHML	1565947_a_at, 1565949_x_at, **1565951_s_at**, 1566337_x_at, 206079_at, 226350_at	1	0.51	23
ENSG00000091483	FH	203032_s_at, 203033_x_at, **214170_x_at**	1	0.53	0
ENSG00000149809	TM7SF2	**210130_s_at**	-1	0.53	0
ENSG00000168275	COA6	**225638_at**	1	0.53	0
ENSG00000172954	LCLAT1	**226996_at**	1	0.54	0
ENSG00000174456	C12orf76	1556299_s_at, **226583_at**, 229679_at	1	0.54	5
ENSG00000149218	ENDOD1	212570_at, **212573_at**	1	0.51	55
ENSG00000163689	C3orf67	**239697_x_at**	1	0.55	2
ENSG00000146192	FGD2	**1553906_s_at**, 1559091_s_at, 1565751_at, 1565752_at, 1565754_x_at, 215602_at, 242632_at	-1	0.56	5
ENSG00000143179	UCK2, MIR3658	209825_s_at, **225722_at**	1	0.58	0
ENSG00000163468	CCT3	**200910_at**	1	0.58	0
ENSG00000103540	CCP110	1569353_at, **204662_at**	1	0.59	1
ENSG00000143390	RFX5	**202963_at**, 202964_s_at	1	0.59	0
ENSG00000147535	PLPP5	**223568_s_at**, 223569_at, 226150_at, 226384_at	-1	0.59	0
ENSG00000221944	TIGD1	**1553099_at**	1	0.59	16
ENSG00000143157	POGK	**218229_s_at**, 222564_at, 239392_s_at	1	0.6	3
ENSG00000152270	PDE3B	208591_s_at, 214582_at, **222317_at**, 231276_at	1	0.62	4
ENSG00000097046	CDC7	**204510_at**	1	0.63	0
ENSG00000124243	BCAS4	1569005_at, 220588_at, **228787_s_at**, 231584_s_at, 239278_at	-1	0.63	3
ENSG00000140455	USP3	221654_s_at, **226652_at**	-1	0.63	0
ENSG00000188343	FAM92A1	**228011_at**, 235391_at, 237910_x_at	1	0.63	1
ENSG00000160710	ADAR	**201786_s_at**	1	0.66	0
ENSG00000131778	CHD1L	**1556988_s_at**, 207645_s_at, 212539_at, 238070_at, 244848_at	1	0.67	2
ENSG00000170522	ELOVL6	**204256_at**, 210868_s_at, 227491_at	1	0.68	3
ENSG00000117625	RCOR3	218344_s_at, **222605_at**, 241433_at	1	0.7	0
ENSG00000164647	STEAP1	**205542_at**	1	0.71	4
ENSG00000117650	NEK2	**204641_at**, 211080_s_at	1	0.73	26
ENSG00000148175	STOM	201060_x_at, **201061_s_at**, 201062_at	-1	0.79	0
ENSG00000158164	TMSB15A	**205347_s_at**, **214051_at**	1	0.83	6
ENSG00000238269	PAGE2B	**231307_at**	1	0.84	2
ENSG00000234068	PAGE2B	**231307_at**	1	0.95	1

Table B.15: Confusion matrices for TC-seq and TC classification on the training group (TG) and validation group (VG). Depicted are the number (and percentage) of patients per TC-seq group in rows and per TC group in columns. **a** TG and **b** VG. In the top left the percentage of consistency (*CO*) is depicted.

a

		TC 2007							
	CO = 75.1%	**11q13**	**6p21**	**D1**	**D1+D2**	**D2**	**FGFR3**	**MAF**	**none**
TC2007-seq	**11q13**	*28 (15%)*	0	7 (4%)	2 (1%)	1 (<1%)	0	1 (<1%)	0
	6p21	0	*2 (1%)*	1 (<1%)	0	0	0	1 (1%)	0
	D1	2 (1%)	1 (<1%)	*56 (30%)*	0	1 (<1%)	1 (<1%)	0	1 (<1%)
	D1+D2	2 (1%)	0	5 (3%)	*15 (8%)*	4 (2%)	0	3 (2%)	0
	D2	0	0	2 (1%)	0	*16 (8%)*	0	0	1 (<1%)
	FGFR3	2 (1%)	0	0	1 (<1%)	0	*14 (7%)*	0	0
	MAF	0	0	6 (3%)	0	1 (<1%)	0	*11 (6%)*	0
	none	0	1 (<1%)	0	0	0	0	0	*0*

b

		TC 2007							
	CO = 75.1%	**11q13**	**6p21**	**D1**	**D1+D2**	**D2**	**FGFR3**	**MAF**	**none**
TC2007-seq	**11q13**	*11 (10%)*	0	0	1 (1%)	1 (1%)	1 (1%)	0	0
	6p21	0	*2 (2%)*	0	0	0	0	0	0
	D1	0	2 (2%)	*32 (30%)*	0	1 (1%)	1 (1%)	1 (1%)	0
	D1+D2	1 (1%)	0	1 (1%)	*6 (6%)*	1 (1%)	0	2 (2%)	1 (1%)
	D2	1 (1%)	0	0	0	*10 (10%)*	0	1 (1%)	0
	FGFR3	0	0	1 (1%)	0	2 (2%)	*9 (9%)*	0	0
	MAF	3 (3%)	0	0	1 (1%)	3 (3%)	0	*9 (9%)*	0
	none	0	0	0	0	0	0	0	*0*

Table B.16: Confusion matrices for MC-seq and MC classification on the training group (TG) and validation group (VG). Depicted are the number (and percentage) of patients per MC-seq group in rows and per MC group in columns. **a** TG and **b** VG. In the top left the percentage of consistency (*CO*) is depicted.

a

MC-seq \ MC	*CO = 83.9%*	**CD1**	**CD2**	**HY**	**LB**	**MF**	**MS**	**PR**
CD1	*15 (8%)*	1 (<1%)	0	0	0	0	3 (2%)	
CD2	1 (<1%)	*28 (14%)*	3 (2%)	3 (2%)	0	2 (1%)	2 (1%)	
HY	0	0	*50 (26%)*	3 (2%)	0	1 (<1%)	0	
LB	0	1 (<1%)	2 (1%)	*21 (11%)*	0	0	1 (<1%)	
MF	0	0	1 (<1%)	0	*8 (4%)*	0	0	
MS	0	1 (<1%)	0	1 (<1%)	0	*17 (9%)*	0	
PR	1 (<1%)	1 (<1%)	0	3 (2%)	0	0	*23 (12%)*	

b

MC-seq \ MC	*CO = 82.4%*	**CD1**	**CD2**	**HY**	**LB**	**MF**	**MS**	**PR**
CD1	*4 (4%)*	0	0	2 (2%)	0	1 (<1%)	2 (2%)	
CD2	1 (<1%)	*15 (14%)*	0	2 (2%)	0	1 (<1%)	0	
HY	0	0	*24 (22%)*	3 (3%)	0	0	1 (<1%)	
LB	1 (<1%)	0	1 (<1%)	*14 (13%)*	0	0	1 (<1%)	
MF	0	0	0	0	*7 (6%)*	0	0	
MS	0	0	1 (<1%)	0	0	*13 (12%)*	0	
PR	0	1 (<1%)	1 (<1%)	0	0	0	*12 (11%)*	

Table B.17: Confusion matrices of t(4;14) prediction on the training group (TG) and validation group (VG) regarding RNA-seq, microarray and iFISH. Depicted are the number (and percentage) of patients with and without predicted t(4;14) on RNA-seq in rows and **a** and **c** predicted on microarray or **b** and **d** determined with iFISH in columns. **a** and **b** TG, **c** and **d** VG. In the top left the percentage of consistency (*CO*) is depicted. iFISH: interphase fluorescence *in-situ* hybridisation.

a

t(4;14)-seq \ t(4;14)-microarray	*CO = 97.4%*	**no t(4;14)**	**t(4;14)**
no t(4;14)	*171 (89%)*	3 (2%)	
t(4;14)	2 (1%)	*17 (9%)*	

b

t(4;14)-seq \ t(4;14)-iFISH	*CO = 99%*	**no t(4;14)**	**t(4;14)**
no t(4;14)	*172 (90%)*	1 (<1%)	
t(4;14)	1 (<1%)	*18 (9%)*	

c

t(4;14)-seq \ t(4;14)-microarray	*CO = 97.3%*	**no t(4;14)**	**t(4;14)**
no t(4;14)	*91 (84%)*	2 (2%)	
t(4;14)	1 (<1%)	*14 (13%)*	

d

t(4;14)-seq \ t(4;14)-iFISH	*CO = 95.3%*	**no t(4;14)**	**t(4;14)**
no t(4;14)	*90 (83%)*	3 (3%)	
t(4;14)	2 (2%)	*13 (12%)*	

Table B.18: Median survival of AMM, MM, MMR and CoMMpass patients regarding assessed risk stratifications. Depicted are RNA-sequencing (RNA-seq) based risk stratifications, DNA-microarray based risk stratifications, the international staging system (ISS) and the revised ISS (R-ISS). EFS: event free survival; OS: overall survival; AMM: asymptomatic multiple myeloma; MM: multiple myeloma; CP: CoMMpass cohort; MMR: relapsed myeloma patients.

Stratificaion	**Group**	**Median EFS [months]**			**Median OS [months]**		
		AMM	MM	CP	MM	CP	MMR
RPI	low	68.5	37.52	39.1	103.46	NR	-
	medium	30.59	32.59	31.21	91.86	63.87	79.7
	high	-	25.2	18.17	54.93	49.28	13.9
GPI	low	89.95	35.71	-	126.65	-	-
	medium	24.31	32.59	-	82.27	-	79.7
	high	-	16.72	-	32.99	-	13.86
UAMS70-seq	low	46.69	35.61	36.63	103.46	NR	79.7
	high	13.42	16.56	20.11	46.85	38.44	13.93
UAMS70	low	48.76	36.44	-	109.83	-	66.17
	high	11.73	23.43	-	52.4	-	15.01
RS-seq	low	59.47	40.18	38.64	129.81	NR	-
	medium	30.59	31.93	35.94	83.32	63.87	79.7
	high	-	16.72	14.75	37.19	31.21	14.08
RS	low	56.77	39.69	-	126.65	-	-
	medium	14.65	29.40	-	70.11	-	79.7
	high	-	16.56	-	32.72	-	13.45
IFM15-seq	low	46.69	36.37	35.94	103.46	NR	79.7
	high	27.99	20.96	23.62	65.97	58.78	22.37
IFM15	low	44.35	34.86	-	103.46	-	79.7
	high	21.62	23.43	-	44.85	-	24.16
EMC92-seq	low	44.35	35.35	35.94	103.46	NR	79.7
	high	23.92	13.63	13.86	29.83	20.04	13.04
EMC92	low	42.74	34.66	-	103.46	-	66.17
	high	15.56	16.33	-	28.88	-	8.31
HDHRS	low	51.78	39.03	39.92	147.55	NR	-
	medium	24.38	31.74	32.92	82.27	63.87	79.7
	high	-	18.2	20.14	36.76	52.24	13.93
HDHRS-GEP	low	45.54	39.06	-	147.55	-	-
	medium	20.5	28.7	-	70.08	-	79.7
	high	-	20.67	-	37.19	-	13.93
ISS	1	40.71	37.78	43.66	129.81	NR	NR
	2	22.51	27.89	29.21	70.08	NR	30.54
	3	51.78	23.85	23.49	54.21	54.74	46.98
R-ISS	I	48.76	41.4	46.16	NR	NR	NR
	II	27.99	26.71	30.13	78.49	NR	46.98
	III	22.49	19.84	18.99	33.71	NR	5.21

Table B.19: Targets - jetset and GeneAnnot database results. In this table 24 exemplary targets are listed. Depicted are gene gene symbbol (SYMBOL), Ensembl gene identifiers (ENSG) and Affymetrix U133 2.0 plus GeneChip probesets. Percentage of consistency of present and absent (PA) expression determination for RNA-seq (PA-seq call) and for microarrays (PA call) is depcited in the collumn "*CO* PA [%]". The column "Jetset max." includes if the depicted ENSG-probeset-pair has the maximum Jetset score for the SYMBOL. Sensitivity and specificity are extracted from the GeneAnnot databse. Selected probesets or ENGS are depicted in bold. - indicates no entry in the Jetset or GeneAnnot databse. Presented are only probesets and ENSGs present in the expression tables. The probeset 210546_x_at of *NY-ESO 1/2* used in the GEP-R, matches to three ENSGs and three SYMBOLs: *CTAG2*, *CTAG1B* and *CTAG1A*.

Name	**SYMBOL**	**ENSG**	**probeset**	*r*	***CO* PA [%]**	**Jetset max.**	**Sensitivity**	**Specificity**
BCMA	TNFRSF17	**ENSG00000048462**	**206641_at**	0.66	100	TRUE	0.909	1
CD38	CD38	**ENSG00000004468**	**205692_s_at**	0.68	100	TRUE	1	1
HM1.24	BST2	**ENSG00000130303**	**201641_at**	0.60	100	TRUE	0.909	1
CD74	CD74	**ENSG00000019582**	1567627_at	0.54	31	FALSE	-	-
			1567628_at	0.38	2	FALSE	0.909	1
			209619_at	0.73	100	TRUE	1	1
NYESO1/2	CTAG2	**ENSG00000126890**	**207337_at**	0.58	90	TRUE	0.909	1
			215733_x_at	0.77	93	FALSE	1	0.758
			210546_x_at	0.73	92	-	0.636	0.407
			211674_x_at	0.72	93	-	0.636	0.407
	CTAG1A	**ENSG00000268651**	**210546_x_at**	0.22	87	FALSE	0.636	0.296
			211674_x_at	0.21	87	FALSE	0.636	0.296
			217339_x_at	0.26	94	FALSE	0.545	0.333
	CTAG1B	**ENSG00000184033**	**210546_x_at**	0.20	87	FALSE	0.636	0.296
			211674_x_at	0.19	87	FALSE	0.636	0.296
			217339_x_at	0.21	94	FALSE	0.545	0.333
HGF	HGF	**ENSG00000019991**	209960_at	0.79	89	FALSE	0.273	1
			209961_s_at	0.71	91	FALSE	1	1
			210755_at	0.81	66	FALSE	1	1
			210997_at	0.85	88	TRUE	0.818	1
			210998_s_at	0.83	83	FALSE	1	1
FGFR3	FGFR3	**ENSG00000068078**	**204379_s_at**	0.90	97	TRUE	1	1
			204380_s_at	0.76	96	FALSE	1	1
MAGEA1	MAGEA1	**ENSG00000198681**	**207325_x_at**	0.80	86	TRUE	0.545	0.889
MAGEA3/ MAGEA6	MAGEA3	**ENSG00000221867**	**209942_x_at**	0.82	85	FALSE	1	0.462
			214612_x_at	0.80	84	-	0.818	0.386
	MAGEA6	ENSG00000197172	**209942_x_at**	0.79	85	FALSE	0.909	0.439
			214612_x_at	0.80	85	FALSE	0.909	0.432
MMSET	WHSC1	**ENSG00000109685**	209052_s_at	0.66	10	FALSE	1	0.955
			209053_s_at	0.78	50	FALSE	1	1
			209054_s_at	0.76	96	TRUE	1	1
			222777_s_at	0.78	81	FALSE	0.727	1
			222778_s_at	0.79	45	FALSE	0.818	1
			223472_at	0.73	12	FALSE	0.636	1
IGF1R	IGF1R	**ENSG00000140443**	203627_at	0.50	28	FALSE	0.818	1
			203628_at	0.32	24	FALSE	0.455	1
			208441_at	-0.01	23	FALSE	-	-
			225330_at	0.68	52	TRUE	0.818	1
			243358_at	-0.04	23	FALSE	0.727	1

TP53	TP53	**ENSG00000141510**	**201746_at**	0.64	95	TRUE	1	1
			211300_s_at	0.28	4	FALSE	1	1
AURKA	AURKA	**ENSG00000087586**	204092_s_at	0.78	72	FALSE	1	1
			208079_s_at	0.76	81	TRUE	1	1
			208080_at	0.00	29	FALSE	0.364	1
CCND1	CCND1	**ENSG00000110092**	208711_s_at	0.81	82	FALSE	1	1
			208712_at	0.79	89	TRUE	1	1
			214019_at	0.04	16	FALSE	0.545	1
CCND2	CCND2	**ENSG00000118971**	200951_s_at	0.85	65	FALSE	1	1
			200952_s_at	0.62	43	FALSE	1	1
			200953_s_at	0.86	78	TRUE	0.909	1
			231259_s_at	0.79	59	FALSE	-	-
CCND3	CCND3	**ENSG00000112576**	1562028_at	0.08	0	FALSE	-	-
			201700_at	0.79	99	TRUE	1	1
RHAMM	HMMR	**ENSG00000072571**	**207165_at**	0.74	88	TRUE	1	1
			209709_s_at	0.69	57	FALSE	1	1
CD20	MS4A1	**ENSG00000156738**	210356_x_at	0.73	37	FALSE	1	1
			217418_x_at	0.71	35	FALSE	0.909	1
			228592_at	0.78	65	FALSE	1	1
			228599_at	0.76	44	TRUE	1	1
			231418_at	0.61	30	FALSE	-	-
MUC1	MUC1	**ENSG00000185499**	207847_s_at	0.59	89	FALSE	1	1
			211695_x_at	0.20	87	FALSE	0.909	1
			213693_s_at	0.63	14	TRUE	1	1
GPRC5D	GPRC5D	**ENSG00000111291**	**221297_at**	0.84	98	TRUE	1	1
CS1	SLAMF7	**ENSG0000026751**	219159_s_at	0.06	100	FALSE	1	1
			222838_at	0.71	100	TRUE	0.818	1
			234306_s_at	0.56	100	FALSE	1	1
WT1	WT1	**ENSG00000184937**	**206067_s_at**	0.12	99	TRUE	1	1
			216953_s_at	0.02	99	FALSE	1	1
SSX2 / SSX2B	SSX2	**ENSG00000241476**	207493_x_at	0.28	99	FALSE	1	0.311
			210497_x_at	0.08	90	FALSE	1	0.37
			215881_x_at	0.38	100	FALSE	0.636	0.246
			216471_x_at	0.10	92	FALSE	0.909	0.294
	SSX2B	ENSG00000268447	207493_x_at	0.24	99	FALSE	-	-
			210497_x_at	0.22	90	FALSE	-	-
			215881_x_at	0.20	100	FALSE	-	-
			215885_at	0.25	100	FALSE	-	-
			216471_x_at	0.18	92	FALSE	-	-
NKG2D	KLRK1	**ENSG00000213809**	**1555691_a_at**	0.03	100	-	1	0.47
			205821_at	0.04	97	-	0.909	0.47

Table B.20: Presence and absence of targets assessed by RNA-sequencing (RNA-seq) and microarrays. Depicted is the consistency of presence and absence of target expression determined on RNA-seq (P and A) and microarrays (MP and A). Present expression values on both platforms are depicted in black, absent expression values are depicted in grey. Expression values only present in microarray are depicted in green, expression values only present in RNA-seq in blue.

RNA-sequencing

	A	MP	A	MP	A	MP	A	MP	A	MP
	BCMA		**CD38**		**HM1.24**		**CD74**			
A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		
P	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00		
	NYESO1/2		**HGF**		**FGFR3**		**MAGEA1**		**MAGEA3**	
A	0.84	0.04	0.05	0.01	0.89	0.03	0.71	0.02	0.56	0.11
P	0.04	0.08	0.11	0.83	0.00	0.07	0.11	0.15	0.04	0.29
	MMSET		**IGF1R**		**TP53**		**AURKA**		**CCND1**	
A	0.04	0.00	0.22	0.01	0.01	0.01	0.19	0.11	0.12	0.04
P	0.50	0.46	0.47	0.30	0.04	0.95	0.08	0.62	0.07	0.77
	CCND2		**CCND3**		**RHAMM**		**CD20**		**GPRC5D**	
A	0.24	0.02	0.00	0.00	0.07	0.05	0.16	0.01	0.00	0.00
P	0.20	0.54	0.01	0.99	0.07	0.82	0.55	0.28	0.02	0.98
	MUC1		**CSF1**		**WT1**		**SSX2**		**NKG2D**	
A	0.01	0.86	0.75	0.00	0.99	0.01	0.90	0.10	1.00	0.00
P	0.00	0.13	0.25	0.00	0.01	0.00	0.00	0.00	0.00	0.00

DNA-microarray

Table B.21: Proportion of overexpressed targets on RNA-sequencing (RNA-seq) and microarray. Depicted is the consistency of overexpressed targets (up) determined by RNA-seq and microarray. Overexpressed values in both samples are depicted in black, normal expressed values (-) in both samples are depicted in grey. Overexpressed samples only in microarray are depicted in green, overexpressed samples only in RNA-seq in blue.

RNA-sequencing

	–	up	–	up	–	up	–	up	–	up
	BCMA		**CD38**		**HM1.24**		**CD74**			
–	0.44	0.01	0.98	0.01	0.34	0.03	0.85	0.10		
up	0.46	0.10	0.01	0.01	0.33	0.29	0.01	0.04		
	NYESO1/2		**HGF**		**FGFR3**		**MAGEA1**		**MAGEA3**	
–	0.87	0.02	0.11	0.01	0.90	0.02	0.74	0.02	0.58	0.09
up	0.04	0.07	0.06	0.82	0.00	0.08	0.12	0.12	0.04	0.29
	MMSET		**IGF1R**		**TP53**		**AURKA**		**CCND1**	
–	0.79	0.07	0.91	0.07	0.22	0.36	0.31	0.55	0.12	0.04
up	0.01	0.12	0.01	0.01	0.03	0.39	0.00	0.13	0.07	0.77
	CCND2		**CCND3**		**RHAMM**		**CD20**		**GPRC5D**	
–	0.57	0.05	0.95	0.00	0.67	0.06	0.82	0.03	0.00	0.00
up	0.02	0.36	0.03	0.02	0.10	0.17	0.03	0.12	0.02	0.98
	MUC1		**CSF1**		**WT1**		**SSX2**		**NKG2D**	
–	0.54	0.28	0.84	0.00	0.99	0.01	0.92	0.08	1.00	0.00
up	0.03	0.15	0.16	0.00	0.01	0.00	0.00	0.00	0.00	0.00

DNA-microarray

Table B.22: Log-rank test results of target survival analysis. Survival analyses were performed by dividing patients in two expression groups by maxstat test or present/absent (PA) determination. The depicted log-rank p-values are adjusted per column for multiple testing. * *CD38* was previously published at the LfM [213] as associated with survival on microarray, with borderline significant p-value of 0.02 in overall survival (OS). In this thesis, using a subset of patients and a differing normalisation method, *CD38* is not associated with OS, indicated by borderline (undadjusted) p-values of 0.08 (adjusted: 0.11) for microarray and 0.055 (adjusted: 0.08) for RNA-sequencing (RNA-seq). OS: overall survival; EFS: event free survival.

Name	EFS maxs. (microarray)	EFS panp (microarray)	EFS maxs. (RNA-seq)	EFS PA (RNA-seq)	OS maxs. (microarray)	OS panp (microarray)	OS maxs. (RNA-seq)	OS PA (RNA-seq)	survival association
BCMA	.72	NA	.11	NA	.66	NA	.21	NA	no
CD38	.27	NA	.33	NA	.12	NA	.08	NA	yes*
HM1.24	.18	.04	.54	NA	.18	.54	.39	NA	no
CD74	.003	NA	.07	NA	.009	NA	< .001	NA	yes
NYESO1/2	< .001	< .001	.001	< .001	< .001	< .001	< .001	< .001	yes
HGF	.73	.41	.79	.32	.83	.73	.93	.68	no
FGFR3	.002	.004	.01	.01	.007	.01	.01	.01	yes
MAGEA1	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	yes
MAGEA3	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	yes
CCND1	.002	.003	.12	.08	< .001	< .001	< .001	.01	yes
CCND2	< .001	.02	< .001	.87	< .001	.052	< .001	0.96	yes
CCND3	.73	.63	.8	.31	.42	.76	.51	.56	no
MMSET	< .001	< .001	< .001	.76	< .001	< .001	< .001	.16	yes
IGF1R	< .001	.008	.008	.1	< .001	< .001	.006	.07	yes
TP53	.73	.84	.17	.08	.83	.94	.47	.002	no
AURKA	< .001	.008	< .001	< .001	< .001	< .001	< .001	< .001	yes
RHAMM	< .001	.05	< .001	.04	< .001	.1	< .001	.02	yes
CD20	.18	.43	.04	.54	.45	.74	.07	.54	NA
GPRC5D	.02	.64	.02	.94	.13	.52	.21	.15	NA
MUC1	< .001	.6	.004	.13	< .001	.76	.03	.14	yes
CS1	.09	.43	.03	.4	.43	0.18	< .001	.17	NA
WT1	.06	.59	.83	.4	.24	.18	.13	.18	no
SSX2	< .001	< .001	NA	NA	< .001	< .001	NA	NA	NA
NKG2D	.02	.57	.19	NA	.27	.18	.9	NA	NA

Table B.23: Splice junction analysis for CD38 and BCMA exemplified for two patients. **a**Patient 1 (left part) shows expression of CD38-001. All seven splice junctions (SJs) are present. Patient 2 shows additional low expression of SJ 9 specific for CD38-005. **b** Patient 1 shows expression of BCMA-001. Both SJs are present. Patient 2 shows additional low expression of SJ 3, SJ 4 and SJ 14 specific for BCMA-002 and BCMA-003. (Seckinger, ..., Emde et al., Frontiers in Immunology 2018 [213])

a

		Patient 1			Patient 2		
Splice Junction	**Specific for CD38 transcript**	**uniquely mapping**	**multiple mapping**	**over-hang**	**uniquely mapping**	**multiple mapping**	**over-hang**
SJ 1	-	964	1	38	1774	4	38
SJ 2	-	1079	0	38	1904	0	38
SJ 3	001	914	1	38	1734	1	38
SJ 4	-	983	0	38	1840	1	38
SJ 5	-	1017	0	38	1882	2	38
SJ 6	-	919	0	38	1743	1	38
SJ 7	-	1002	0	38	1969	0	38
SJ 8	002	0	0	0	0	0	0
SJ 9	005	0	0	0	19	0	37

b

		Patient 1			Patient 2		
Splice junction	**Specific for BCMA transcript**	**uniquely mapping**	**multiple mapping**	**over-hang**	**uniquely mapping**	**multiple mapping**	**over-hang**
SJ 1	-	1667	0	38	1881	1	38
SJ 2	001	1674	2	38	2010	3	38
SJ 3	002	0	0	-	22	0	37
SJ 4	002	0	0	-	13	0	38
SJ 5	003	0	0	-	14	0	37

C Supplementary Code

```
#!/bin/bash

# 1. Create a directory for the index:
mkdir References/StarIndex

# 2. Step into the directory and download the genome FASTA file and the GTF file:

# Download genome and uncompress it:
# ftp://ftp.ensembl.org/pub/release-77/fasta/homo_sapiens/dna/
# Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz

# Download Ensembl GTF and uncompress it:
# ftp://ftp.ensembl.org/pub/release-82/gtf/homo_sapiens/Homo_sapiens.GRCh38.82.gtf.gz

# 3. Build the genome index:
STAR  --runMode genomeGenerate --runThreadN 15 --genomeDir References/StarIndex
      --genomeFastaFiles
        References/StarIndex/Homo_sapiens.GRCh38.dna.primary_assembly.fa
      --sjdbGTFfile References/Homo_sapiens.GRCh38.82.gtf

# 4. Alignment and read count per gene and splice junction:
STAR --runThreadN 15 --genomeDir References/StarIndex
     --readFilesIn $sample_1.fa.gz $sample_2.fa.gz
     --readFilesCommand gzip -cd --outFileNamePrefix $sample
     --outSAMtype BAM SortedByCoordinate
     --quantMode GeneCounts --outReadsUnmapped Fastx --chimSegmentMin 15
     --genomeLoad LoadAndKeep --limitBAMsortRAM 31532137230

# 5. Generating a bam index:
samtools index $sample_Aligned.sortedByCoord.out.bam

# 6. Read count per base
bam-readcount -f References/StarIndex/Homo_sapiens.GRCh38.dna.primary_assembly.fa
              $sample_Aligned.sortedByCoord.out.bam 7:140753336-140753336 > V600counts
                   .txt
```

Code C.1: Alignment and read count using STAR. The sample `$sample` is aligned against the human genome and the number of reads are counted per gene and splice junction with STAR. The reads per base are counted with bam-readcount. (For references, see section 2.3.2.1)

```
#!/bin/bash

# 1. Create a directory for the index:
mkdir References/RSEMIndex

# 2. Step into the directory

# 3. Build the index:
rsem-prepare-reference  --gtf References/Homo_sapiens.GRCh38.82.gtf
             --star References/StarIndex/Homo_sapiens.GRCh38.dna.primary_assembly.fa
             rsem_ref/human_ensembl

# 4. Alignment:
STAR --runThreadN 15 --genomeDir References/RSEMIndex
     --readFilesIn $sample_1.fa.gz $sample_2.fa.gz
     --readFilesCommand gzip -cd --outFileNamePrefix $sample
     --outSAMtype BAM Unsorted
     --quantMode TranscriptomeSAM
     --genomeLoad LoadAndKeep --limitBAMsortRAM 31532137230

# 5. RSEM quantification
rsem-calculate-expression  --paired-end
                           --num-threads 15
                           --quiet
                           --bam $sample_Aligned.toTranscriptome.out.bam
                           References/rsem_ref/human_ensembl
                           $sample_quant

# 6. RSEM plot
echo -e "ENST00000053243\nENST00000396495\nENST00000562385" > inputBCMA.txt
rsem-plot-transcript-wiggles $sample_quant inputBCMA.txt
                             --show-unique $sample_quant_BCMA.pdf
```

Code C.2: Alignment and read count using RSEM. The sample `$sample` is aligned against the human genome and the number of reads are counted per transcript with STAR. The expression is quantified and visualised with RSEM. (For references, see section 2.3.2)

```
#!/bin/bash

# Quality control
# 1. before alignment: FastQC
fastqc $sample_1.fa.gz --threads 15 --outdir $sample_1_fastqc_reports
fastqc $sample_2.fa.gz --threads 15 --outdir $sample_2_fastqc_reports

# Evaluation of the HTML report as described in chapter 2.3.2.5
```

Code C.3: Quality control with FastQC. The quality report of both files of the sample `$sample` is controlled.

```
# R
# Quality control

# 2. After alignment: at least 60\% mapped reads
logs <- read.csv($sample*Log.final.out, sep=c("|"))
# Column 27 contains the percentage unmapped reads due to too many mismatches
# Column 28 contains the percentage unmapped reads due to too short reads
# Column 29 contains the percentage unmapped reads due to other reasons
UnmappedReads <- sum(as.numeric(
                     c(str_replace(unlist(str_split(logs[27,2], "\t"))[2], "%", ""),
                       str_replace(unlist(str_split(logs[28,2], "\t"))[2], "%", ""),
                       str_replace(unlist(str_split(logs[29,2], "\t"))[2], "%", ""))))
if(UnmappedReads < 40, "keep", "remove")

# 3. After summing up technical replicates: Library size >= 1000000
counts <- read.table("$sample*ReadsPerGene.out.tab")
# The first four rows contain a mapping summary
# column 2 contains the counts for unstranded RNA-seq
librarysize <- sum(ct[-c(1:4),2])
if(librarysize >= 10000000, "keep", "remove")

# merge all count tables in one (count.table)
```

Code C.4: Quality control of the number of mapping reads and the library size. The sample `$sample` has to fullfill the following criteria: At least 60% of the reads have to map to the genome and the library size has to be >10000000 reads.

```
# 1. Sum technical replicates:
ct <- aggregate(t(count.table), by=list(colnames(count.table)), sum)

# 2. Exclude genes with zero counts in all samples:
zero <- apply(ct, 1, function(x) all(x==0))
ct2 <- ct[!zero,]

# 3. Normalisation
y <- DGEList(counts=ct2, group=entities, genes=rownames(ct2))
y1 <- calcNormFactors(y) # method=TMM
nc <- cpm(y1, normalized.lib.sizes=TRUE)

# 4. Log2 transformation:
nc.log <- log2(nc+1)
```

Code C.5: EdgeR normalisation. The merged read counts from the STAR output files (`countfiles`) are summed per patient, normalised with EdgeR and log2 transformed.

```
add.normalise <- function(ct, entity="MMs", ct.TG=ct.TG, e.TG=entities.TG){
  # 1. Order the genes of the ct and add the new ct to ct.TG:
  ct.add <- cbind(ct.TG, ct[match(row.names(ct.TG), row.names(ct)),])
  colnames(ct.add)[ncol(ct.add)] <- colnames(ct)

  # 2. Adds the new entity to e.TG:
  entity.add <- c(e.TG$entity.1, as.character(entity))

  # 3. Normalisation:
  y <- DGEList(counts=ct.add, group=entity.add, genes=rownames(ct.add))
  y1 <- calcNormFactors(y)
  nc <- cpm(y1, normalized.lib.sizes=TRUE) # nc: normalized counts

    # 4. Generate nice output:
  nc <- data.frame(nc[, ncol(nc)])
  colnames(nc) <- colnames(ct)
  nc
}
```

Code C.6: Normalisation function. Input parameters are the raw counts of the new sample (`ct`) and of the training group (`ct.TG`) and the entities of both (`entity` and `entities.TG`).

```
RPI.seq <- function(nc.log, entity="MMs", genes=gpi.genes, glc=gene.length.cutoff){
  # 1. Vector with the genes translated to Ensembl gene IDs: genes

  # 2. Cutting nc.log to 50 gpi_genes and do present/absent call
  nc.RPI <- nc.log[genes, ncol(nc.log)]
    glc.nc <- cbind(glc[genes], nc.RPI)
  nc.RPI.pa <- t(apply(glc.nc, 1, function(x) ifelse(x[-1]>=x[1], 1, 0)))
  nc.RPI.pa <- nc.RPI*nc.RPI.pa

  # 3. Score estimation by summing up the 50 genes
  score <- sum(nc.RPI.pa)

  # 4. Division in 3 groups
  low.cut=121.9601
  high.cut=202.7359
  risk <- ifelse(score<=low.cut, "low risk",
                 ifelse(score<=high.cut, "medium risk", "high risk"))

  data.frame(ID=colnames(nc.log), score=score, risk=risk, entity=entity)
}
```

Code C.7: RPI function. Input parameters are the normalised and log2 transformed counts (`nc.log`), the entity, the genes of the RPI (`gpi_genes`), depicted in table B.9 and the gene length cutoff (`gene.length.cutoff`).

```
UAMS70.seq <- function(nc.log, entity="MMs", td=traindata.UAMS70){
  # 1. Extract the genes, translated to Ensembl gene IDs
  uams.genes <- row.names(td)
  sel = intersect(uams.genes, rownames(nc.log))
  gep70.log2.exp = as.matrix(nc.log[sel,])

  # 2. Centre probe sets using the TG expression averages
  centred.gep70.log2.exp = gep70.log2.exp - td[,"TrainingMeans"]

  # 3. weighting of the expression, using a weighting matrix and a vector
  weighted.ave = t(td[,7:12]) %*% centered_gep70_log2_exp
  score = t(weighted.ave) %*% c(-0.193970199, -0.102403915, 0.048411741,
                                0.043861307, 0.181940086, 0.171153439)

  # 4. Estimate the squared distance to the group mean risk
  squared.dist = (matrix(score, nrow=length(score),ncol=3) -
                  matrix(c(-6.083014, -5.190644, -4.096769),
                           nrow=length(score), ncol=3, byrow=TRUE))^2

  # 5. Classify the sample
  risk = apply(squared.dist, 1, which.min)
  risk = ifelse(risk==1, "low risk", ifelse(risk==2, "medium risk", "high risk"))
  names(risk) <- row.names(score)

  data.frame(ID=colnames(nc.log), score=score, risk=risk, entity=entity)
}
```

Code C.8: UAMS70-seq function. Input parameters are the normalised and log2 transformed counts (`nc.log`), the entity and the data of the training cohort (`traintab`).

```
EMC92.seq <-function(nc.log, entity="MMs", td=traindata.EMC92){
  # 1. Match translated genes
  nc.log <- nc.log[match(row.names(td),row.names(nc.log)),]

  # 2. Mean variance standardisation
  nc.EMC <- t((nc.log-td$trainmean)/td$trainsd)

  # 3. Estimate score by matrix-weighting score-product estimation
  score <- nc.EMC %*% td$weight_EMC

  # 4. Division in 2 groups
  threshold=1.014283
  risk <- ifelse(score>threshold, "high risk", "standard risk")

  data.frame(ID=colnames(nc.log), score=score, risk=risk, entity=entity)
}
```

Code C.9: EMC92-seq function. Input parameters are the normalised and log2 transformed counts (`nc.log`), the entity and the data of the training cohort (`traindata`).

```
RS.seq <- function(nc.log, entity="MMs", genes=rs.genes){
  # 1. RS genes translated to Ensembl gene IDs: genes

  # 2. Multiply genes by a factor
  RS.factor <- c(rep(1, 14),rep(-1, 3))
  nc.RS <- nc.log[genes,]*as.numeric(rs.factor)

  # 3. Score estimation by summing up the gene expression
  score <- sum(nc.RS)

  # 4. Division in 3 groups
  low.cut=9.01033587115606
  high.cut=29.3685251672214
  risk <- ifelse(score<=low.cut, "low risk",
                 ifelse(score<=high.cut, "medium risk", "high risk"))

  data.frame(ID=colnames(nc.log), score=score, risk=risk, entity=entity)
}
```

Code C.10: RS-seq function. Input parameters are the normalised and log2 transformed counts (`nc.log`), the entity and the genes of the RS-seq (`RS_genes`), depicted in table B.11.

```
HDHRS.seq <- function(nc.log, entity="MMs", genes=HDHRS.genes){
  # 1. Vector of genes of HDHRS (Ensemble gene IDs): genes

  # 2. Multiply genes by a factor
  HDHRS.factor <- c(rep(-1, 13), rep(1, 40))
  nc.HDHRS <- nc.log[genes,]*as.numeric(HDHRS.factor)

  # 3. Score estimation by summing up the 53 genes
  score <- sum(nc.HDHRS)

  # 4. Division in 3 groups
  low.cut=3.86805329889207
  high.cut=24.0962532436869
  risk <- ifelse(score<=low.cut, "low risk",
                 ifelse(score<=high.cut, "medium risk", "high risk"))

  data.frame(ID=colnames(nc.log), score=score, risk=risk, entity=entity)
}
```

Code C.11: HDHRS function. Input parameters are the normalised and log2 transformed counts (`nc.log`), the entity and the genes of the HDHRS (`HDHRS_genes`), depicted in table B.14.

```
TC2007.seq = function(nc, entity="MMs") {
  co=c(FGFR3=72.19458, WHSC1=39.63992, CCND3=459.58197, CCND2.x=48.84759,
       CCND2.y=68.41890, CCND1.2=2.90520, ITGB7=104.57858, CX3CR1 =14.05472,
       MAF=24.63713, MAFB=13.50716, CCND1.1=333.19385, mpI=132.79614)

  # 1. Estimate geometric mean of the ten genes (called macrophage index)
  p = c("ENSG00000120708", "ENSG00000170458", "ENSG00000177575", c("ENSG00000162747",
        "ENSG00000203747"), "ENSG00000073754", "ENSG00000131495", "ENSG00000275385",
        c("ENSG00000113141", "ENSG00000113119"))
  expr <- c(nc[p[c(1:3,6:8)],], sum(nc[p[c(4:5)],]), sum(nc[p[c(9:10)],]))
  macrophageIndex = exp(mean(log2(expr+1),na.rm=TRUE))

  # 2. Class prediction
  TC.1 <- ifelse(nc["ENSG00000068078",] > co["FGFR3"], 1,
                 ifelse(nc["ENSG00000109685",] > co["WHSC1"], 1,0))
  TC.2 <- ifelse(nc["ENSG00000112576",] > co["CCND3"] &
                 (nc["ENSG00000118971",] < co["CCND2.x"] &
                 nc["ENSG00000110092",] < co["CCND1.2"]), 1,0)
  TC.3 <- ifelse((nc["ENSG00000139626",] > co["ITGB7"] &
                  nc["ENSG00000168329",] > co["CX3CR1"]), 1,0)
  TC.32<- ifelse((nc["ENSG00000139626",] > co["ITGB7"] |
                  nc["ENSG00000178573",] > co["MAF"] |
                  nc["ENSG00000204103",] > co["MAFB"] &
                  macrophageIndex < co["mpI"] &
                  nc["ENSG00000118971",] > co["CCND2.y"]), 1,0)
  TC.4 <- ifelse(nc["ENSG00000110092",] >  co["CCND1.1"], 1,0)
  TC.5 <- ifelse(nc["ENSG00000110092",] >  co["CCND1.2"] &
                 nc["ENSG00000118971",] <= co["CCND2.x"], 1,0)
  TC.6 <- ifelse(nc["ENSG00000110092",] >  co["CCND1.2"] &
                 nc["ENSG00000118971",] >  co["CCND2.x"], 1,0)
  TC.7 <- ifelse(nc["ENSG00000110092",] <= co["CCND1.2"] &
                 nc["ENSG00000118971",] >  co["CCND2.x"], 1,0)
  TC.8 <- ifelse(nc["ENSG00000110092",] <= co["CCND1.2"] &
                 nc["ENSG00000118971",] <= co["CCND2.x"], 1,0)

  # 3. Choose right class
    TC.mat = c("FGFR3"=TC.1, "6p21"=TC.2, "MAF"=max(TC.3, TC.32), "11q13"=TC.4,
               "D1"=TC.5, "D1+D2"=TC.6, "D2"=TC.7, "none"=TC.8)
    class <- names(TC.mat)[which(TC.mat==1)][1]

  data.frame(ID=colnames(nc), class=class, entity=entity)

}
```

Code C.12: TC2007-seq function. Input parameters are the normalised counts (`nc`) and the entity.

```
IFM15.seq <- function(nc.log, entity="MMs", genes=IFM15.genes){
  # 1. Vector with IFM genes translated to Ensembl gene IDs and
  #    named with the microarray probe set: genes

  # 2. Estimate IFM score by equation
  score = nc.log[genes["208644_at"],]   * 0.27578783 +
          nc.log[genes["202470_s_at"],] * 0.26987655 +
          nc.log[genes["202951_at"],]   * 0.29530369 +
          nc.log[genes["200783_s_at"],] * 0.31490195 -
          nc.log[genes["201425_at"],]   * 0.13137903 +
          nc.log[genes["231736_x_at"],] * 0.17772804 +
          nc.log[genes["217752_s_at"],] * 0.38697337 +
          nc.log[genes["202486_at"],]   * 0.30371178 +
          nc.log[genes["212098_at"],]   * 0.25043791 -
          nc.log[genes["209683_at"],]   * 0.29483393 +
          nc.log[genes["228677_s_at"],] * 0.19243758 -
          nc.log[genes["200779_at"],]   * 0.2491429  -
          nc.log[genes["203657_s_at"],] * 0.17822457 +
          nc.log[genes["204072_s_at"],] * 0.21255699 +
          nc.log[genes["228737_at"],]   * 0.21956366

  # 3. Classify in 2 groups
  threshold=7.672949
  risk = factor(ifelse(score > threshold,"high risk","low risk"),
                levels=c("low risk", "high risk"))

  data.frame(ID=colnames(nc.log), score=score, risk=risk, entity=entity)
}
```

Code C.13: IFM15-seq function. Input parameters are the normalised and log2 transformed counts (`nc.log`), the entity and the genes of the IFM-seq (`IFM_genes`), depicted in table B.13.

```
MC.pam.seq <- function(nc.log, entity="MMs", td=traindata.MC){
  # 1. Filtering the expression table
  nc.MC <- data.frame(nc.log[match(row.names(td$centroids), row.names(nc.log)),])
  row.names(nc.MC) <- row.names(td$centroids)

  # 2. Prediction
  threshold = 0.2505557
  prediction <- pamr.predict(fit = td, newx = nc.MC, threshold=threshold)

  data.frame(ID=colnames(nc.log), prediction=as.character(prediction), entity=entity)
}
```

Code C.14: MC-seq function. Input parameters are the normalised and log2 transformed counts (`nc.log`), the entity and the parameters of the training cohort (`pam.train`).

```
t414.pam.seq <- function(nc.log, entity="MMs", td=traindata.t414){
  # 1. Filtering the expression table
  nc.t414 <- data.frame(nc.log[match(row.names(td$centroids), row.names(nc.log)),])
  row.names(nc.t414) <- row.names(td$centroids)

  # 2. Prediction
  threshold = 6.556106
  prediction <- pamr.predict(fit = td, newx = nc.t414, threshold=threshold)

  data.frame(ID=colnames(nc.log), prediction=as.character(prediction), entity=entity)
}
```

Code C.15: t(4;14)-seq prediction function. Input parameters are the normalised and log2 transformed counts (`nc.log`), the entity and the parameters of the training cohort (`pam.train`).

```
# 1. Read splice junction table
counts.SJ <- read.table("$sample*SJ.out.tab")
# column 1: chromosome,
# column 2: first base of the intron (1-based)
# column 3: last base of the intron (1-based)
# ...

# 2. Count splice junctions of e.g. BCMA at chromosome 16
BCMA.SJ <- counts.SJ[which(counts.SJ[,1]==16 & counts.SJ[,2]>=11965107 &
                           counts.SJ[,3]<=11968068),]

# 3. Filter: at least 10 uniquely mapping reads crossing the junction
BCMA.SJ <- BCMA.SJ[which(BCMA.SJ[,7]>=10),]

# Positions of the splice junctions per transcript
BCMA.001 <- c(SJ.1="11965455_11966194", SJ.2="11966342_11967569")
BCMA.002 <- c(SJ.1="11965455_11966194", SJ.3="11966342_11967019",SJ.4="
    11967144_11967569")
BCMA.003 <- c(SJ.5="11965455_11967569")

# 4. Test if transcripts are expressed
if(all(BCMA.001 %in% paste(BCMA.SJ[,2], BCMA.SJ[,3], sep="_")), "present", "absent")
if(all(BCMA.002 %in% paste(BCMA.SJ[,2], BCMA.SJ[,3], sep="_")), "present", "absent")
if(all(BCMA.003 %in% paste(BCMA.SJ[,2], BCMA.SJ[,3], sep="_")), "present", "absent")
```

Code C.16: Splice junction analysis. The table containing the counts per splice junction is preprocessed and filtered. The expressed transcripts are determined.

```
pa.call.seq <- function(nc, glc=gene.length.cutoff){
  # 1. load median transcript length per gene divided by 1000 (gene length cutoff)

  # 2. estimate present (P) and absent (A) expression per gene
  nc.pa <- data.frame(ifelse(nc[,1]>=glc[row.names(nc)], "P", "A"))
  colnames(nc.pa) <-  colnames(nc)
  nc.pa
}
```

Code C.17: PA call function for RNA-seq. Input parameters are the normalised counts (`nc`) and the gene length cutoff (`gene.length.cutoff`).

```
overexpression.call.seq <- function(nc.log, Bc=BMPC.log.cutoff){
  # 1. the overexpression cutoff is the median expression of the BMPCs per gene
  # multiplied with three times the standard deviation of the BMPC expressions

  # 2. estimate overexpression per gene
  nc.ov <- data.frame(ifelse(nc.log[,1]>=BC[row.names(nc.log)], "UP", "-"))
  colnames(nc.ov) <-  colnames(nc.log)
  nc.ov
}
```

Code C.18: Overexpression function for RNA-seq. Input parameters are the normalised and log transformed counts (`nc.log`) and a estimated cutoff, using the expression of BMPCs (`BMPC.log.cutoff`).

```
# 1. Read table with counts per base
counts.base <- read.table("$sample*counts.txt")
# column 1, 2, 3, 4: chr, position, reference_base, depth
# column 6, 7, 8, 9: base information for A, C, G, T :
# base:count:avg_mapping_quality:avg_basequality:avg_se_mapping_quality:num_plus_
    strand:num_minus_strand:avg_pos_as_fraction: ...

# 2. Split information per base e.g. G
counts.G <- as.numeric(str_split(counts.base[,8], ":")[[1]][-1])

# 3. Count and filter reads per base e.g. G
# i) at least two reads covering the mutation
# ii) a mapping quality of 255
# iii) a base quality of at least 30
# iv) at least one read in each strand direction
# v) an average base position in the intermediate 85% of the nucleotides
# vi) variant allel frequency of at least 10%
ifelse(counts.G[1]>=2 & counts.G[2]==255 & counts.G[3]>=30 &
       (counts.G[5]>=1 & counts.G[6]>=1) & counts.G[7]>=0.85 &
       counts.G[1]/as.numeric(counts.base[,4])>=0.1, "G present", "G absent")
```

Code C.19: Mutation analysis. The table containing the counts per base is preprocessed, filtered and the number of mutated bases (e.g. "G" at position V600) is determined.

```
# First part: Preprocessing

# 1.1 The alignment, the index and the read count per base is performed as
# described in code C.1, lines 18 to 25

# 1.2 For quality control see code C.3 and C.4 lines 1 to 20

# Second part: Normalisation, risk stratification and molecular classification

# 2.1 Normalisation with TG 1 and TG 2
ct <- matrix(counts[-c(1:4),2], dimnames=list(counts[-c(1:4),1], "sample.name"))
nc1 <- add.normalize(ct=ct, entity="MMs", ct.TG=counttable.TG, e.TG=entities.TG)
nc2 <- add.normalize(ct=ct, entity="MMs", ct.TG=counttable.TG.M, e.TG=entities.TG.M)
# log2 transformation
nc1.log <- log2(nc1+1)
nc2.log <- log2(nc2+1)

# 2.2 Estimate stratifications and classifications:
RPI.Seq(nc.log=nc1.log, entity="MMs")
UAMS.Seq(nc.log=nc2.log, entity="MMs")
RS.Seq(nc.log=nc2.log, entity="MMs")
EMC92.Seq(nc.log=nc2.log, entity="MMs")
IFM15.Seq(nc.log=nc2.log, entity="MMs")
HDHRS.Seq(nc.log=nc2.log, entity="MMs")
TC2007.Seq(nc=nc2, entity="MMs")
MC.Seq(nc.log=nc2.log, entity="MMs")
t414.Seq(nc.log=nc1.log, entity="MMs")

# Third part: Target analysis

# 3.1 Target expression PA call and overexpression
pa.call.seq(nc, glc=gene.length.cutoff)
overexpression.call.seq(nc.log, Bc=BMPC.log.cutoff)

# 3.2 Splice junction analysis e.g. BCMA is described in code C.16

# 3.3 Mutation analysis, e.g. G at V600E in BRAF is described in code C.19
```

Code C.20: Example of using RNA-seq stratification, classification and target assessment analysis pipeline.

Acknowledgements

This dissertation was challenging and has contributed to considerably evolve my scientific and none the less my personal abilities. My grateful thanks go to all the supporters, colleagues and friends, without whom my scientific work would not have been possible.

First of all, I would like to thank my supervisor PD Dr. Dr. Dirk Hose for mentoring my dissertation and offering me the opportunity to work in his group.

Furthermore, I would like to thank PD Dr. Dr. Dirk Hose and Dr. Anja Seckinger for continuous support and for their professional, helpful advice. Although the work group finally has to be dissolved, they have kept out relevant problems in this field from me and offered me the possibility to finish my thesis.

I also would like to thank Prof. Dr. Dr. Thierry Rème who explained me the RS score algorithm and kindly provided his scripts.

Grateful thanks go to my colleague and friend Susanne Beck, as well writing her thesis in the LFM, for informatical and statistical assistance, proofreading of my dissertation and hours of discussions. She was always ready to listen and found calm words in disturbing circumstances.

Finally, I thank my family and my friends for useful comments and critical reading. They supported me wherever possible and found encouraging words in every kind of situation.

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema

   "**Transcriptome Profiling Assessing Pathogenesis and Prognosis of Plasma Cell Dyscrasias – Bioinformatic Basis for Clinical Application**"

   handelt es sich um meine eigenständige erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärung bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort, Datum

Unterschrift