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

Molecular pathogenesis and prognosis of light chain amyloidosis

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

vorgelegt von
Susanne Monika Beck

aus
Groß-Gerau

2021

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

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

Contents

List of Figures

List of Tables

List of Codes

Nomenclature

AL	Light Chain Amyloidosis
ALMG	Light Chain Amyloidosis with subentity MGUS
ALMM	Light Chain Amyloidosis with subentity MM
AMM	Asymptomatic Multiple Myeloma
BAM	Binary representation of SAM
BH	Benjamini-Hochberg correction
BMPC	Bone Marrow Plasma Cell
CA	Chromosomal Aberration
chr	Chromosome
CNA	Copy Number Alteration
CoMMpass	relating Clinical outcomes in MM to Personal Assessment of Genetic Profile
CP	CoMMpass
CPM	Counts Per Million
CRAB	Hyper**C**alcemia, **R**enal impairment, **A**nemia and **B**one disease
cTnT	Cardiac Troponin T
DE	Differential Expression
DEG	Differentially Expressed Genes
del	Deletion
del13q14	Deletion of region q14 on chr 13
del17p13	Deletion of region p13 on chr 17
del8p21	Deletion of region p21 on chr 8
diff FLC	Difference between Involved to Uninvolved Free Light Chain
FEA	Functional Enrichment Analysis
FLC	Free Light Chain
gain	Amplification
gain1q21	Gain of a region q21 on chr 1
GEP	Gene Expression Profiling
GPI	Gene Expression based Proliferation Index
GRCh38	Genome Reference Consortium human genome build 38
HC	Heavy Chain
HD	Heidelberg
HDAL	Heidelberg AL Score
HDT	High-Dose Therapy
hg19	human genome build 19
hg38	human genome build 38

HMCL	Human Myeloma Cell Line
HR	Hazard Ratio
HRD	Hyperdiploidy
iFISH	interphase Fluorescence *in situ* Hybridization
IFM	Intergroupe Francophone du Myélome
IFM15	IFM 15-gene score
Ig	Immunoglobulin
IgH	Immunoglobulin Heavy Chain
IMWG	International Myeloma Working Group
InDel	Insertion or Deletion
LC	Light Chain
LFC	Log Fold Change
LfM	Multiple Myeloma Research Laboratory
M-protein	Monoclonal Protein
MAI	Myc-activation Index
MBC	Memory B Cell
MC	Molecular Classification
MGUS	Monoclonal Gammopathy of Undetermined Significance
MM	Multiple Myeloma
NGS	Next Generation Sequencing
NT-ProBNP	N-Terminal Pro-Brain Natriuretic Peptide type-B
OL	Osteolytic Lesions
OR	Odds Ratio
OS	Overall Survival
PCA	Principal Component Analysis
PCD	Plasma Cell Dyscrasia
PCI	Plasma Cell Infiltration
PPC	Polyclonal Plasmablastic Cell
QC	Quality Control
RNA Seq	RNA Sequencing
RS	Risk Score
SAM	Sequence Alignment Map
SKY92	SKYLINE DX 92 gene score
SMM	Smoldering Multiple Myeloma
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
t(11;14)	Translocation of region q13 on chr 11 to chr 14 region q32.3

t(14;16)	Translocation of region q23 on chr 16 to chr 14 region q32.3
t(4;14)	Translocation of region p16.3 on chr 4 to chr 14 region q32.3
t(6;14)	Translocation of region p21 on chr 6 to chr 14 region q32.3
t-SNE	t-distributed Stochastic Neighbor Embedding
TC	Translocation/Cyclin D
Ti	Transition
Tv	Transversion
UAMS	University of Arkansas for Medical Sciences
UAMS70	UAMS 70-gene score
VAF	Variant Allele Frequency
WES	Whole Exome Sequencing

1 Introduction

Primary aim of this thesis is to assess molecular properties and prognosis of patients suffering from light chain amyloidosis in relation to other malignant plasma cell diseases including multiple myeloma and its precursor state. In the following chapter, an introduction is given into different plasma cell dyscrasias, their diagnosis and treatment, followed by an overview about current knowledge regarding molecular pathogenesis of these entities. The chapter closes introducing the detailed aims of the thesis.

1.1 Malignant plasma cell diseases

Malignant plasma cell diseases are a group of entities characterized by the accumulation of malignant plasma cells in the bone marrow. These include, as main entities briefly described in the following, monoclonal gammopathy of undetermined significance (MGUS), multiple myeloma (MM), and light chain amyloidosis (AL). Different entities are defined by tumor mass (number of malignant plasma cells) and whether and which "end organ damage" they cause. For tumor mass, different surrogates are in use, in particular the amount of secreted monoclonal protein (M-protein), free light chains (FLC) as part of monoclonal immunoglobulin (Ig) in serum or urine, and the amount of plasma cell infiltration (PCI) in the bone marrow (for a more detailed description of M-protein or FLC see section 1.2.2, for PCI see section 1.6) [265].
The incidence for malignant plasma cell diseases increases with age [133, 168, 169, 171]. The estimated annual incidence of MGUS in men is 120 per 100,000 population at the age of 50 years and increases to 530 per 100,000 population at the age of 90 years [293]. For MM and AL, incidences are 7.0 and 0.89 per 100,000 inhabitants in the USA [133, 168]. Men are more often affected than women [133, 168, 169, 171, 293].

MGUS and multiple myeloma

The earliest detectable stage in plasma cell diseases is called MGUS, consecutively followed by asymptomatic multiple myeloma (AMM) and MM. The risk of progression from MGUS to symptomatic disease, e.g. MM or AL, is about 1% per year [172]. AMM progresses to MM with a rate of 10% per year [170]. Generally, MGUS precedes MM [171, 174].
In MGUS, per definition, M-protein is detectable, but at an amount below 30 g/L in serum, PCI is below 10% of nucleated cells in the bone marrow, and no disease-related end organ damage is present [142]. A special case of MGUS is the so-called light chain (LC) MGUS with presence of monoclonal kappa or lambda LC in serum and/or urine [142, 242]. PCI or M-protein exceeding the respective threshold classifies the disease

as multiple myeloma [142]. It is further sub-stratified regarding the presence of end organ damage. This is caused by the accumulation of malignant plasma cells in the bone marrow leading to displacement of normal hematopoiesis (e.g. anemia), induction of osteolytic bone lesions (potentially leading to hypercalcemia), and production of a monoclonal protein or parts thereof which can damage the glomerular capillaries in the kidneys (leading to renal insufficiency). These signs and symptoms are summarized as "CRAB" (hyperCalcemia, Renal impairment, Anemia, and Bone disease) [142] by the "International Myeloma Working Group" (IMWG) in 2003. In absence of "CRAB" criteria, multiple myeloma is termed asymptomatic multiple myeloma (AMM) [142].
In 2014, the IMWG revised their classification by introducing "SLiM-CRAB" [242]. These criteria add to CRAB biomarkers thought to surrogate immediate progression, i.e. more than one focal lesion in magnetic resonance imaging (MRI), a serum free light chain ration (FLCR) between involved and uninvolved chain above 100 (see section 1.2.2), and a PCI equal or above 60% [242].
Now, asymptomatic patients who are not MGUS and do not fulfill any of the "SLiM-CRAB" criteria are termed smoldering multiple myeloma (SMM) patients, and patients who fulfill at least one of the criteria are considered to be symptomatic. Previously, the terms AMM and SMM were used synonymously. In this thesis, the term AMM is used for asymptomatic patients according to the 2003 IMWG criteria [142, 242].

Light chain amyloidosis

"Systemic amyloidosis" describes a group of diseases characterized by misfolded proteins, deposited in sites distant from their production and secretion. The most common form is AL [199]. Here, the misfolded proteins are monoclonal LC (defining the M-protein in this case) that are transported via blood from the producing malignant plasma cells in the bone marrow to organs. The deposition causes clinical signs and symptoms. AL is characterized by the type of involved organ. Typically, these are heart (82%), kidneys (68%), liver (14%), and less frequently lung, gastrointestinal tract (8%), or soft tissues (17%) [199]. Brain involvement is not observed [199]. Peripheral neuropathy is detected in 12%, and autonomous neuropathy in 10% of patients [199].
Early signs and symptoms of AL include fatigue, inappetence, unexplained weight loss, and reduced physical fitness [106]. Among clinical signs are unexplained heart failure and proteinuria. As these symptoms and signs are unspecific, common to elderly people and other diseases, they frequently do not raise enough suspicion to specifically diagnose or exclude AL [106, 107].
Diagnosis of AL is made almost always late, due to the unspecific nature of signs and symptoms and concomitant low incidence of 8.9 $(5-13)$ per million persons per year

[168]. For comparison, a progression rate of 1% *per anno* is expected for MGUS to AL or MM [167, 330]. In 10% of myeloma patients AL occurs [142]. Hence, late diagnosis of about one year after the recognition of first symptoms and signs is prevalent [187]. In many patients, end organ damage is present at diagnosis and increases especially early mortality [107].

For earlier diagnosis it is recommended [107] to screen all MGUS patients with abnormal FLCR for amyloid deposits (see section 1.2.2), determine the LC type (see section 1.2.2), and to assess cardiac biomarkers (see section 1.4) [199, 233].

MGUS, AMM and MM are stratified by tumor mass [142, 242], but for AL the tumor mass is not used for classification [200]. In about $45\% - 50\%$ of AL patients the PCI exceeds 10% [217, 264, 330]. In this thesis, AL patients are further subdivided whether their underlying plasma cell disease fulfills the MGUS or MM criteria and are termed ALMG or ALMM, respectively (as described above in part "MGUS and multiple myeloma" in section 1.1) [28].

1.2 Pathophysiology and pathogenesis

In the following section, the pathophysiologic background and molecular pathogenesis of malignant plasma cell diseases in relation to their normal counterpart are described.

1.2.1 Bone marrow plasma cells

Bone marrow plasma cells (BMPC) are a component of the adaptive immune system [156, 219]. Derived from haematopoietic stem cells, they are formed via different precursor stages including naïve B cells [156, 219]. These enter lymphatic organs and differentiate - after encountering their specific antigen and being selected for self tolerance - with the help of T cells to memory B cells (MBC) and early polyclonal plasmablastic cells (PPC) in the so-called "germinal center reaction" [156, 219]. Further maturation to (early) plasma cells enables the production of huge amounts of specific Ig [156, 219]. These are Y-shaped proteins, consisting of two identical heavy chains (HC) of the type α, δ, ε, γ, or μ and two functionally identical LC of the type λ or κ [219]. The HC defines the type of the Ig (A, D, E, G, or M) [219]. PPC migrating via blood circulation enter the bone marrow [156, 241]. They interact with the bone marrow microenvironment and become long-lived BMPC if they can compete successfully and home in a specific "niche" [156, 241]. The fraction of plasma cells in physiological bone marrow is below 1% and remains constant during adult life [156, 241].

1.2.2 Malignant plasma cells

Malignant plasma cells accumulate in the bone marrow due to their ability to proliferate. With increasing tumor mass in the bone marrow, malignant plasma cells displace the normal compartments (including the niches of BMPC), taking over nutrition supply and space [156].

Malignant plasma cells are monoclonal, i.e. of the same clonal origin, and produce the same Ig, referred to as M-protein [156, 265]. The level of M-protein per individual myeloma between different patients however can vary [78–80]. It is a surrogate for tumor mass - more cells produce more protein. Consisting of a full Ig or LC only, the M-protein is measured and determined in serum and urine samples. For quantitative assessment, in case of FLC secretion, the fraction and difference of the involved *versus* the uninvolved LC can be calculated. The term "involved" is used for the monoclonal LC. This fraction also comprises a part of LC of the same general type produced by normal plasma cells. "Uninvolved" LC comprise the sum of all light chains of the other type, completely produced by non-malignant BMPC. The frequency of Ig type varies between the different disease entities. In MGUS, the most common type is a full IgG ($\sim 76\%$) [264]. In MM, IgG accounts for about 59%, IgA for about 20%, and in approximately 18% only LCs are secreted [264]. This is different to AL. Here, mainly the LC Ig subunit is secreted [23]. In MM, the κ LC is more frequently detected ($\sim 66\%$) than the λ LC [264]. In contrast, most patients in AL secrete a λ LC ($\sim 76\% - 80\%$) [23, 199, 264].

Amyloid forming light chains

AL is characterized by the deposition of monoclonal LC. Secreted LC are transported via blood circulation and deposited in different organs [22]. Here, they aggregate and form amyloid fibrils causing organ dysfunctions [22]. The pathophysiologic mechanism of amyloid toxicity is only partially understood, e.g. soluble amyloid can directly induce cellular apoptosis [22, 199].

The presence of amyloid deposits can be verified by Congo red staining of tissue biopsies. Under polarized light, stained amyloid positive tissue shows an apple green birefringence [199]. For diagnosis of amyloid deposits, abdominal fat aspirates are of high reliability and can be carried out by a fast and minimal invasive biopsy [85, 86, 101, 233]. The LC type can be determined and quantified by protein electrophoresis or immunofixation electrophoresis from serum and urine samples [199], the so called serum "free light chain assay", an immunoassays based on polyclonal antibodies [37], or mass spectrometry [206, 207, 271].

1.2.3 Mechanisms of molecular pathogenesis

Malignant plasma cell characteristics can be informative regarding the molecular pathogenesis of the underlying disease and be used to stratify molecular subgroups that are defined by e.g. a specific chromosomal aberration, deregulated gene expression, or nucleotide variants, affecting signal transduction pathways or the cell cycle. The assessment of subgroups provides prognostic and potentially treatment-relevant information.

Chromosomal aberrations

The most important molecular subgroups are defined by IgH translocation (IgH-TL) and hyperdiploidy (HRD), regarding two paths of underlying "primary" genetic alterations leading to myeloma [156, 265]. IgH-TL are considered primary events that appear during the maturation of plasma cells [265, 309]. HRD is defined by gains of several odd numbered chromosomes especially 5, 15, 19 [323]. As IgH-TL, HRD appears during plasma cell differentiation [156, 265]. Both types of underlying aberration appear mostly disjunct [156, 265]. The key difference between malignant and normal plasma cells is the ability to proliferate [156, 265]. Here, a unifying feature of malignant plasma cells is a deregulation, i.e. an over- or aberrant expression of one of the three D-type cyclins (*CCND1*[1], *CCND2*, *CCND3*) [156, 265]. D-type cyclins drive cells from the G_0 phase (cell cycle arrest) into the cell cycle by interaction with CDK4/6 (cyclin-dependent kinases 4/6) [156, 265].

IgH-TL are characterized by translocation of a gene under the control of the IgH super enhancer located at chromosome 14, leading to aberrant or overexpression of the translocated gene [265]. In case of t(11;14), the translocated gene locus is 11q13 (*CCND1*), which is subsequently aberrantly expressed [265]. The t(4;14) leads to an overexpression of *MMSET* (*NSD2*), and in 68% of events to aberrant expression of *FGFR3* [20, 156, 258, 265]. Indirect *CCND2* overexpression is also induced by t(4;14) [265]. Direct overexpression of *CCND2* by IgH-TL is a very rare case of t(12;14) [20, 51, 265]. The translocation t(14;16) causes an upregulation of the proto-oncogene *MAF* [50]. Direct *CCND3* upregulation can be caused by t(6;14) [20, 265]. The most frequent IgH-TL is the t(11;14), detected in $15\% - 20\%$ of MM patients, followed by the t(4;14) with $10\% - 15\%$ [28, 62, 223]. The IgH-TL t(6;14), t(12;14), and t(14;16) affect less than 2% of patients [28, 156, 221, 265].

HRD is likewise associated with aberrant or overexpression of D-type cyclins. In case

[1] In this thesis, all genes are only named by their respective HGCN gene symbol [325] and not by full names, as these are not unique. Genes are written capitalized and italicized, proteins are capitalized. Online resource: `https://www.genenames.org/`

of *CCND2*, the mechanism is indirect, in case of *CCND1* it is mediated by 11q13 gain [156, 265].

Besides IgH-TL and HRD, further aberrations frequently appear; malignant plasma cells usually harbor more than one chromosomal aberration (CA) [34, 62, 223, 265]. In **MM**, the most frequent aberrations are gain of 1q21 (36%), del 13q14 (46%), and del 17p13 (10%) [221, 265]. Gain of 1q21 is copy number dependently associated with higher proliferation rate and adverse prognosis in MM [116]. The del 13q14 affects the tumor suppressor genes *RB1* and *DIS3* [265]. The gene product of *DIS3* has an exonuclease domain and is active in RNA processing and degradation [190]. *DIS3* is frequently mutated in MM and loss of *DIS3* or disruptive mutations may act oncogenic by deregulation of protein translation (see the next but one part in section 1.2.3 on altered gene expression). Deletion of 17p13 is associated with adverse prognosis in MM, and with a higher frequency of mutations in *TP53* gene [35, 265].

In **AL**, the same aberrations occur as in MGUS or MM without any "AL-typical aberration" being detected. Differences however exist in the frequency at which these aberrations appear in the population of affected individuals, most notably t(11;14) and HRD.

Compared to MGUS (30%) [25] and MM (57%) [221], HRD as clonal aberration appears less frequent in AL (11%) [25]. The most common CA in AL is the IgH-TL t(11;14) (61%) [28]. More frequent in MM than in AL is the t(4;14) (4%) [28]. The translocations t(14;16) and t(6;14) are rare in both, AL and MM [23, 265].

Gain of 1q21 (31%), del 13q14 (38%) or del 17p13 (3%) [28], which are associated with poor prognosis in MM, are less frequent in AL compared to MM, but more frequent compared to MGUS [23, 26, 28, 116, 265].

A prognostic relevance of CA in AL is - in contrast to MM - frequently depending on a specific treatment. For example, patients with t(11;14) had a significantly adverse prognosis if treated with bortezomib containing regimes [27], whereas the aberration convey neutral prognosis in MM [267]. Patients with gain 1q21 treated by melphalan/dexamethasone combinations show a comparably adverse prognosis [26].

Different molecular pathways leading to malignant plasma cells imply disease heterogeneity between patients (*inter*-patient heterogeneity, e.g. HRD or IgH-TL) nonetheless leading to a comparable phenotype - plasma cell accumulation [156, 265]. Furthermore, an *intra*-patient heterogeneity can be detected [156, 265]. Despite being a clonal disease (all malignant plasma cells produce the same Ig or parts thereof, i.e. M-protein), a "sub-"structure can be demonstrated for most patients, exemplified by subclonal presence of alterations like del 17p13, and mutations in *NRAS* or *BRAF* [156, 265].

Part of the objectives of this thesis is to assess whether or not different or AL-typical CA can be found, how the pattern of CA in AL relates to MGUS, AMM, and MM, and in as much these are dependent or independent of other molecular and clinical risk factors (see section 1.6).

Single nucleotide variants

In **MM**, several studies [30, 31, 45, 185, 307, 308, 310] reported recurrently mutated genes, none of them being a unifying event: *KRAS* (24%), *NRAS* (20%), *DIS3* (10%), *FAM46C* (11%), *TP53* (7%), *BRAF* (6%), *TRAF3* (5%), *PRDM1* (4%), *CLYD* (3%), and *RB1* (2%) [45, 185, 308]. Thus, also on the level of single nucleotide variants (SNV), *inter*-patient heterogeneity can be found. Some of these mutations are thought to bear disease-driving effects [308], but they likewise appear subclonal (see the discussion in section 4.4.8). Subclonal appearance again exemplifies *intra*-patient heterogeneity.

In **AL**, only few studies with small numbers of patients have investigated alterations like SNV [36, 230, 254], in part only selecting several ($n = 10$) interesting genes [254]. As part of this thesis, the largest and most comprehensive analysis will be undertaken.

Altered- or overexpression of genes

In **MM**, genetic alterations frequently directly or indirectly influence gene expression, as exemplified above in case of t(11;14) and aberrant *CCND1* expression mediating cell cycle entry [20, 156, 265]. Other functional examples comprise increased angiogenesis [127, 156] and bone turnover, the latter leading to osteolytic lesions as hallmarks of bone marrow microenvironment transformation [156, 182, 294]. Malignant plasma cells change the balance between pro-angiogenic and anti-angiogenic cytokines that influence the microenvironment with increasing tumor mass. They express at least one pro-angiogenic gene aberrantly [127]. The Wnt-signaling inhibitor *DKK1* is aberrantly expressed by malignant plasma cells in patients with bone lesions [294] and inhibits osteoblast formation [182]. HGF inhibits the development of osteoblasts [282]. Likewise, mutations in signal transduction pathways are able to alter (downstream) gene expression. Examples include the NF-κB (*BIRC2*, *BIRC3*, *CYLD*, *TRAF3*) and the MAPK/ERK (*KRAS*, *NRAS*, *BRAF*) signaling pathways [45, 185, 307]. Other affected pathways comprise RNA processing, via mutations in genes for RNA binding proteins like *DIS3* and *FAM46C*, [45] or apoptosis via mutations in members of the *BCL2* family of apoptosis regulation genes [44]. Loss of function mutations in the histone demethylase *KDM6A* [45, 302] can interfere chromatin remodeling.

In **AL**, gene expression data were previously only reported in small patient cohorts in

comparison to MM and BMPC [1, 6, 160, 230], i.e. for 24, 16, 9, or 41 AL patients. Besides that, a set of four genes was investigated in 53 AL patients by RT-qPCR, and for 16 patients by DNA microarray [335]. Comparisons of different malignant plasma cell diseases were carried out by significance analysis of gene expression data [1, 160, 230], or in one case, by differential expression analysis [6]. Other than the unifying D-type cyclin expression present as well in MM, no characteristic "AL gene (set)" could be identified. Analysis of gene expression data on the largest cohort with samples from 196 and 124 AL patients for DNA microarray and RNA sequencing will be carried out as part of this thesis.

1.2.4 Risk stratification

Besides CA associated with adverse prognosis (see section 1.2.3), gene expression profiling (GEP) can also be used to *a priori* determine risk, in contrast to defining biological subentities and assessing their risk-association *a posteriori*. Whereas different strategies exist, in principle, genes individually or as set associated with survival are selected and grouped into a "score".

Three main strategies can be distinguished: First, scores that primary surrogate biological variables like proliferation or Myc-activation, which are afterward investigated for association with prognosis. The first is exemplified by the gene expression-based proliferation index (GPI) by Hose *et al.* [128], which assesses low, medium, and high risk of progression by the cumulative expression of genes associated with cell proliferation, serving as a biological surrogate for proliferation. As for the second, Chng *et al.* [55] used a gene set enrichment analysis to find genes associated with cell cycle, proliferation, and Myc-activation, which are overexpressed in MM, creating the Myc-activation index (MAI) [55]. The MAI delineates patients with poor prognosis if it is above a defined threshold [55].

The second strategy is to group myeloma patients regarding molecular plasma cell subentities, and subsequently assess their prognosis. Based on altered gene expression of D-type cyclins and oncogenes dysregulated by IgH-TL, the translocation/cyclin D (TC) classification by Bergsagel *et al.* [20] assigns patients to eight distinct groups [20, 54]. Alternatively, Zhan *et al.* [329] created the molecular classification (MC) of MM with seven subentities based on unsupervised hierarchical clustering analysis. Here, the three subentities MF, MS, and PR combined *versus* all others, delineate a poor prognosis group.

The third strategy is to directly choose genes associated with survival: Clustering methods were used at the University of Arkansas for Medical Sciences (UAMS) encompassing the expression values of 70 survival associated genes in a score (UAMS70) [272].

The score was intended to classify patients regarding high risk disease, i.e. short overall survival [272]. Using k-means clustering, patients were classified in short and long overall survival [272]. Subsequently, a PAM-based predictor was calculated [272]. The Intergroupe Francophone du Myélome created a 15 gene model (IFM15) [68]. For every gene in the expression data set, a univariate cox regression was performed, and the top 15 survival associated genes were chosen to create the score, delineating high risk from low risk regarding overall survival. For creation of the EMC92 by Kuiper *et al.* [161], 92 genes were selected by a PCA analysis with a cross validation strategy. Chosen genes were assigned with a positive or negative weighting factor, summed up and a threshold was determined to separate high risk *versus* standard risk patients. The threshold for high risk of EMC92 was defined by delineating the patients with an overall survival below 24 months. The risk score (RS) by Rème *et al.* [248] selects genes by a running log rank test and estimates the association with prognosis for each gene. Genes selected for the score were assigned a positive weight for poor prognosis or a negative weight for good prognosis and summed up. The score splits myeloma patients into three risk groups: low, medium, and high risk of progression and short survival.
Strategies to assess the risk of progression by GEP are for the first time applied to AL in this thesis.

1.2.5 Gene expression-based risk assessment in extended clinical routine

Given the primary aim of this thesis of assessing molecular properties and prognosis of patients suffering from AL in relation to other malignant plasma cell diseases including MM, prospective target assessment and multimodal prediction of survival for personalized and risk-adapted treatment strategies in MM was conducted in the GMMG-MM5 multicenter trial as part of this thesis, and published (Hose, D.*, Beck, S.* [shared first-authorship] *et al.* [126]). A further important question here, regarding gene expression-based risk assessment and its embedding in multimodal risk assessment (i.e. combination of e.g. expression-based and conventional prognostic factors), is whether this is applicable in extended clinical routine, e.g. in the context of a clinical phase III trial.
Two of the above-described scores (UAMS70 and IFM15), the GPI and the two classifications (TC and MC) were previously combined to a GEP-Report (GEP-R) [198], which could be seen as laying the basis for potential personalized and risk-adapted treatment of single patients [126]. The GEP-R framework normalizes each sample to the original MM reference cohort [198]. Afterwards, it calculates different GEP-based stratifications (GPI, UAMS70, IFM15), assesses the expression of potential target genes associated with adverse survival in MM (*AURKA* [129], *FGFR3* [296],

IGF1R [210]) or targets for potential immunotherapy [137, 259] (*CTAG1*, *MAGEA1*, *MAGEA3*, *HM1.24*, *MUC1*, *SSX2*), predicts the presence of a t(4;14) from gene expression by a PAM-based predictor [121], and finally evaluates the HM metascore (Heidelberg Montpellier meatascore) [198]. The HM metascore [198] summarizes the GEP-based risk scores UAMS70 and IFM15, the GEP-based proliferation assessment GPI, the International staging system [111] (ISS, described below in section 1.3), and t(4;14) assessment into one single risk score.

1.3 Risk assessment of multiple myeloma in clinical routine

Standard for assessment of risk of progression for MM patients is currently the revised ISS (rISS) by Palumbo *et al.* [234]. The rISS includes the conventional ISS score, published by Greipp *et al.* [111], based on the concentration of serum albumin and β_2-microglobulin, the latter seen as a surrogate for tumor mass [111]. The rISS further includes the presence of selected CA associated with adverse prognosis, i.e. del 17p13, t(4;14), and t(14;16) [234].

1.4 Clinical prognostic markers in light chain amyloidosis

In AL, clinical prognostic markers for organ involvement and dysfunction are routinely determined for heart and kidney. These markers can be combined, to delineate patients of different risk groups.

Cardiac failure is the most frequent reason for premature death in AL [199]. Commonly used cardiac biomarkers are cardiac troponin T (cTnT) [71] and N-terminal pro-brain natriuretic peptide type-B (NT-ProBNP) [231].

NT-ProBNP is a biologically inactive prohormone consisting of the 32 amino acid polypeptide BNP secreted attached to a 76 amino acid N-terminal fragment. [33]. It is produced in response to cardiac stress [33]. NT-ProBNP is increased in serum concentration in patients with advanced heart involvement [231]. Different prognostic thresholds are drawn; the most frequently used is at 1800 ng/L [164].

The regulator protein cTnT controls calcium-mediated interaction between actin and myosin [270]. It is a sensitive serum marker for cardiac injury [71]. Kumar *et al.* [164] defined a prognostic threshold at 0.025 ng/mL.

The production of FLC correlates with the total tumor mass [24, 295]. Production of FLC is measured in serum quantitatively, and the fraction and difference of involved *versus* uninvolved chain (diff FLC) is calculated. Whereas different thresholds in serum exist [72, 303], most frequently used is 180 mg/L [164].

Combinations of these three serum parameters are used in three risk assessment models: The "standard" Mayo Score (2004) by Dispenzieri *et al.* [70], the revised Mayo Score (2012) by Kumar *et al.* [164], and the "advanced" Mayo Stage III Euro Score (2013) by Wechalekar *et al.* [315]. For a description, see table 1.1.

Table 1.1: Light chain amyloidosis risk assessment models: the "standard" Mayo Score (2004) by Dispenzieri *et al.* [70], the "revised" Mayo Score (2012) by Kumar *et al.* [164], and the "advanced" Mayo Stage III Euro Score (2013) by Wechalekar *et al.* [315]. NT-ProBNP: N-terminal pro-brain natriuretic peptide type-B; cTnT: cardiac troponin T; diff FLC: difference in free light chains.

Model	Marker and threshold	Stage
standard Mayo 2004	NT-ProBNP $>$ 332 ng/L cTnT $>$ 0.035 ng/mL	I: no marker above threshold II: one marker above threshold III: both marker above threshold
Euro 2013	Mayo 2004 stage III and NT-ProBNP $>$ 8500 ng/L	IIIA: below threshold IIIB: above threshold
revised Mayo 2012	NT-ProBNP $>$ 1800 ng/L cTnT $>$ 0.025 ng/mL diff FLC $>$ 180 mg/L	0: no marker above threshold 1: one marker above threshold 2: two markers above threshold 3: all markers above threshold

Renal involvement is assessed, as in MM, by measuring the creatinine level and especially creatinine clearance [96, 131]. Creatinine is a by-product of muscle metabolism renally excreted [96, 131]. Creatinine production and excretion depends on body mass, age, and sex of the patient, mainly due to differences in muscle mass, and as such decreases with age [131]. Physiologically, in blood of male individuals levels of $0.6-1.2$ mg/dL, and in female individuals $0.5-1.1$ mg/dL can be found [131]. A prognostic threshold indicating severe renal failure is defined at 2 mg/dL [89]. Creatinine clearance, as volume of blood cleared of creatinine per minute, is calculated from the concentration of creatinine in serum and urine, as well as the volume of excreted urine over 24 hours. It approximates the glomerular filtration rate. The physiologic creatinine clearance is $110-150$ mL/min for men, and $100-130$ mL/min for women [110]. Any decrease in creatinine clearance indicates reduced renal function.

1.5 Treatment of multiple myeloma and light chain amyloidosis

Patients progressing to MM require treatment. The intention of treatment is to eliminate malignant plasma cells and prevent patients from myeloma associated organ damages. In general, it is aimed at giving patients an intensive treatment with induction treatment, followed by high-dose chemotherapy (HDT), autologous stem cell transplantation (ASCT), consolidation, and maintenance therapy, as described in the fol-

lowing sections 1.5.1 and 1.5.2. Patients not "fit" enough for intensive treatment, due to general constitution or comorbidities, receive less intensive treatment, i.e. no ASCT and frequently reduced doses of therapeutic agents (see section 1.5.3).
Pathophysiology of AL and patient prognosis are primarily determined by organ damage caused by LC deposition. Median survival of AL patients varies between $12-40$ months after diagnosis [199]. As signs and symptoms of AL are unspecific (see section 1.1), the diagnosis is frequently made late, i.e. when organ dysfunctions hamper intensive treatment (see also section 1.5.1) [107, 199, 232, 330]. Organ transplantation is often required for patients with severe organ damages, especially cardiac failure [107, 199, 232, 330] (see section 1.5.5). Elimination or at least as deep as possible reduction of malignant plasma cell numbers to stop LC production and concomitant deposition is currently the main therapeutic principle [107, 199, 232, 330]. Given that AL is a malignant plasma cell disease (to which extent molecular identical AL is to MM will be determined in this thesis, see aims in section 1.7), treatment schemata of MM are modified and applied [107, 199, 232, 330]. Treatment aims first at hematological response in terms of elimination of malignant plasma cells and concomitant decrease of M-protein, especially FLC in serum and urine [107, 199, 232, 330]. The second aim, taking more time, is the organ response, i.e. recovery of organ function [107, 199, 232, 330]. Pulmonary, renal, and cardiac toxicity of chemotherapeutic agents can reduce the tolerability of treatment schemata especially in AL patients already encountering AL derived toxicities [232]. Generally, as in MM, the basic distinction is made between patients that can be treated intensively, i.e. with HDT and ASCT, and those in which this is not possible [107, 199, 232, 330]. To a certain extent, molecularly defined subgroups (as described in section 1.2.3) can be used to select the appropriate agents [26, 27].

1.5.1 Intensive treatment

This concept uses three to six cycles of induction therapy [4, 107], followed by HDT with 200 mg/m^2 melphalan (HDM) and ASCT. The induction regimen currently most frequently applied includes a proteasome inhibitor, a corticosteroid like dexamethasone, and a third agent. Currently, three different proteasome inhibitors [244] are used in upfront treatment, i.e. bortezomib [166], ixazomib [211], and carfilzomib (GMMG-CONCEPT trial NCT03104842) [58, 285]. As third agent, either immunomodulatory agents (IMiDs) [284] like lenalidomide and pomalidomide, or cytotoxic agents like cyclophosphamide and adriamycin (e.g. GMMG-MM5 trial, see section 1.5.2) are used. IMiDs are frequently less well tolerated in AL compared to MM patients [330]. In recent clinical trials, the administration of monoclonal antibodies against CD38 [264]

(daratumumab [67] NCT03201965, and isatuximab [69] NCT03499808)) or against SLAMF7 (elotuzumab [186] NCT03252600) is tested [330]. In particular, anti-CD38-treatment is promising in AL [150, 330]. An example for an intensive regimen including HDT and ASCT for MM is depicted in the following section.

1.5.2 The GMMG-MM5 clinical trial for previously untreated multiple myeloma patients

In this multicenter randomized phase III trial of the GMMG, (GMMG-MM5 EudraCT no. 2010-019173-16) [191, 203], patients were allocated to four treatment arms (A1, A2, B1, and B2). They received two different induction treatments, either three 4-week cycles PAd (bortezomib, adriamycin, low dose dexamethasone) (A1+B1), or three 3-week cycles VCD (bortezomib, cyclophosphamide, dexamethasone) (A2+B2, current GMMG standard). After induction, patients underwent stem cell mobilization and leukapheresis for subsequent ASCT. Then HDM and ASCT were applied. Patients not achieving a near complete response (nCR) or better receive a second HDT and ASCT. Afterwards, all patients receive two cycles of lenalidomide (25 mg, days 1-21) as consolidation therapy. For maintenance therapy patients receive either lenalidomide (for the first 3 months, 10 mg/day continuously and thereafter 15 mg/day continuously) for 2 years (A1+A2) or until CR (B1+B2). The GMMG-MM5 trial was analyzed in this thesis regarding prospective target assessment and multimodal prediction of survival for personalized and risk-adapted treatment strategies in MM [126] to relate respective findings to those in AL patients.

1.5.3 Non-intensive treatment

For transplantation ineligible patients in MM and AL, bortezomib based treatment regimen have long been considered standard of care [107, 243]. Recently, regimen including the CD38-antibody daratumumab were introduced for MM, both in combination with bortezomib based chemotherapy [193] or lenalidomid [95].

In AL, the application of treatment can improve organ function and transplant ineligible patients can become candidates for ASCT [107, 192]. In very fragile patients, a conventional cytostatic drug like melphalan is combined with dexamethasone [107]. Doses can be reduced, increasing the tolerability but decreasing the response rates [233].

1.5.4 Experimental treatment

Experimental treatment regimen includes the small molecule inhibitor venetoclax, which was tested in relapsed AL patients (NCT03000660) [225]. Venetoclax shows activity especially in patients with increased levels of BCL2, which is associated with the translocation t(11;14) (see section 1.2.3) [165, 330].

Besides targeting malignant plasma cells as in all approaches depicted above, experimental strategies reducing formation and fostering removal of amyloid are currently in clinical testing. This comprises 11-1F4 [134, 278, 330] for removal of amyloid deposits (NCT02245867). Inhibition of amyloid fibril formation is currently tested in multiple clinical trials (NCT02207556, NCT03474458, and NCT03401372) applying doxycycline, a long known antibiotic drug [21, 151].

1.5.5 Organ transplantation

As organ, especially cardiac, function primarily determines early patient survival, organ transplantation is considered for AL patients [107, 233, 330]. Kidney and heart transplantation are preferable for patients with severe single organ involvement [233]. To prevent recurrence of amyloid deposition in the graft, preferably malignant plasma cell accumulations should be controlled before [107].

1.6 Molecular profiling of plasma cells

For molecular profiling of (malignant) plasma cells, bone marrow aspirates were collected from the iliac crest (*spina iliaca posterior superior*) [125]. In brief, $60-80$ ml of bone marrow are aspirated from patients diagnosed as AL, MGUS, AMM, or MM [125, 126]. PCI is determined from bone spicules present in the first draft, the remaining aspirate used for plasma cell purification (see section 2.1.2) [125]. Purified malignant plasma cells are subjected to molecular profiling [125]. For this, malignant plasma cells are spinned on glass slides and used for interphase fluorescence *in situ* hybridization (iFISH). RNA and DNA are extracted from purified malignant plasma cells for GEP with DNA microarray or RNA sequencing (RNA seq), as well as for whole exome sequencing (WES) (see figure 1.1 for an overview). Samples were processed according to the Multiple Myeloma Research Laboratory (LfM) standard operating procedure (SOP) at the University Hospital Heidelberg [125]. Molecular profiling is performed in extended clinical routine at the LfM. A brief outline is described in the following. For a detailed description, see section 2.1.2 and Hose [125].

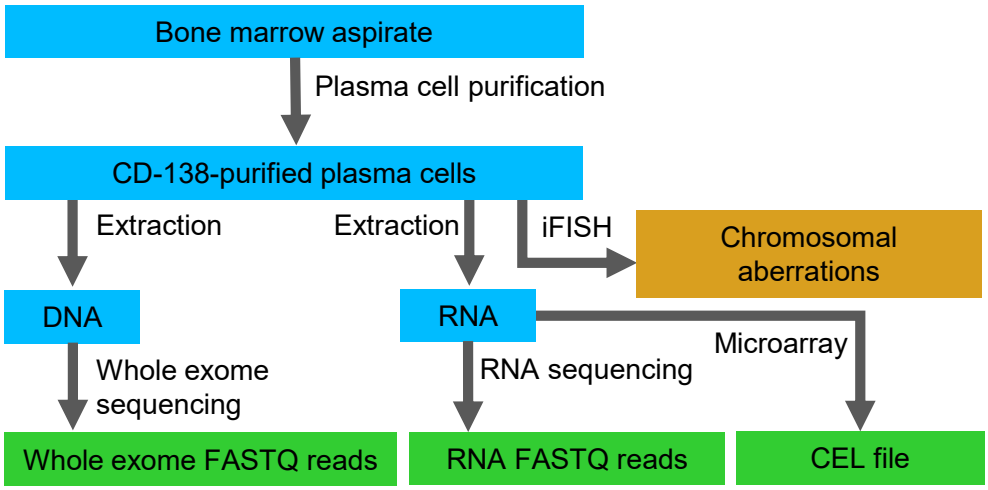

Figure 1.1: Summary of sampling and experimental laboratory strategy. Samples and laboratory processing are depicted in blue colored boxes, derived data in green colored boxes, and the iFISH data set in orange colored box. Key words for methods are indicated besides the gray colored arrows. iFISH: interphase fluorescence *in situ* hybridization

1.6.1 Interphase fluorescence *in situ* hybridization

Genetic alterations in terms of copy number alterations and translocations are assessed by highlighting pre-determined genetic regions with iFISH [125]. For this, purified myeloma cells are spinned and fixed to glass slides, permeabilized, and incubated with fluorescence dye labeled DNA-probes (iFISH probes) [125]. These iFISH probes bind to their pre-defined specific DNA region [125]. Thus, amplifications, deletions, and translocations can be visualized at single cell level [125].
The routine turnaround time is several days only [125]. Largest disadvantages are first that only pre-determined aberrations can be detected, and second that the increase of assessed aberrations considerably increases the needed amount of input cells, as well as the workload [125]. For a detailed description of the method, see section 2.1.2.
The analyses are performed in extended clinical routine in cooperation of the LfM with the "Molekular-zytogenetisches Labor" (Prof. Anna Jauch, department of human genetics, Heidelberg). A full list of used iFISH probes is depicted in section 2.1.2.

1.6.2 DNA microarrays

DNA microarray "chips" are glass slides, on which 54675 different clusters of 25-mer oligonucleotides (cDNA probes) are synthesized (in the case of Affymetrix HG-U133 2.0 plus chip) [3, 125]. Each cDNA probe binds a specific RNA sequence. The cDNA probes on the chip are organized as "match" and "mismatch" probes. This can be used for subsequent determination of "presence" and "absence" of gene expression (see section 2.3). In clinical routine, GEP with DNA microarrays is possible within

four weeks, as shown and published based on this thesis [126], and it is a reasonably expensive method costing in the range of 500 € (in academic setting) to 2000 € (in commercial setting). Main methodological disadvantages are a saturation of cDNA probes, which lead to an upper bound for single expression values for highly expressed genes, and background noise due to unspecific binding to the cDNA probes [125].
In brief, the experimental procedure is conducted as follows. Extracted total RNA is labeled, fragmented, *in vitro* transcribed, amplified, and hybridized to the chip [125]. The amount of labelled RNA binding to a specific cluster is optically measured [125]. Intensities for each probe are stored per sample in a so-called "CEL" file (Affymetrix specific format) [125].
At the LfM GEP with DNA microarrays is performed in extended clinical routine [125, 126]. For details, see section 2.3

1.6.3 Next generation sequencing

"Next generation sequencing" (NGS) is a sequencing-by-synthesis based determination of all (or selected) DNA or RNA sequences in a sample, both in terms of nucleic acid sequence and abundance. It can be applied at relatively low cost (500€ in an academic setting at the LfM), within a clinically reasonable time frame of four weeks [204, 273, 301]. In the following section, the sequencing method underlying the data used in this thesis is briefly described, for more details, see Balasubramanian [15] and Illumina [138, 141]. The approach comprises three subsequent steps: library preparation, cluster generation, and sequencing-by-synthesis.
First, to obtain "something that can actually be sequenced" a "library" needs to be prepared. Here, the first step is "tagmentation" during which DNA is fragmented into pieces of around 200 bp each, called tag. RNA is previously reversely and stoichiometrically transcribed to DNA in case of RNA seq. During the "sample preparation", custom sequencing adapters are ligated to both ends of each tag. Afterwards, two indices for sample and regions complementary to the flow cell oligonucleotides are added. The "flow cell" is a kind of glass slide or reaction chamber on which the actual sequencing is performed. It is sub-structured in lanes, with each lane containing oligonucleotides fixed to the surface. The oligonucleotides added to the tags bind to the complementary oligonucleotides on the flow cell.
In the "cluster generation phase", each bound tag of the library is locally amplified by PCR ("bridge amplification"). By this method, clusters of identical sequences are produced to enhance the fluorescence intensity in the following sequencing step.
The actual "sequencing" is performed by synthesis in cycles. Different strategies apply, but in principle, each of the four bases A, T, G, C in a nucleotide is labeled with a

specific fluorescence dye. After each synthesis step, the build-in bases at each of the clusters are determined by laser excitation and specific fluorescence detection, i.e. a high-resolution image of the flow cell is taken, and position and color of all light signals are recorded. This visualization determines the necessity of more than one event (i.e. incorporation of more than one labeled base) at each cluster, the reason for the need of the previous bridge amplification. The information for each cluster is therefore the incorporated base at each cycle until the desired number of cycles is reached and is called a "read".

In "paired end sequencing", another bridging of the tags is applied for turning the tags and sequencing is performed from the reverse complement strand creating "read two". This results in a multitude of parallel produced sequencing reads each representing a tag.

Sequencing reads are stored in FASTQ format (see section 2.4.2) in text files for each sample. Possible target sequences for NGS are the whole genome, the protein coding exome, or the transcriptome. At the LfM DNA (WES) and RNA are sequenced in extended clinical routine.

RNA sequencing uses RNA reverse transcribed to DNA as input material for NGS [140]. It allows the quantitative assessment of gene expression comparable to DNA microarrays [138, 140]. As advantage compared to DNA microarrays, the whole transcriptome can be assessed without restriction to predefined sequences [138, 140]. This allows, in addition to the assessment of gene expression, the detection of mutations, alternative transcripts or fusion transcripts [138, 140]. There is no saturation for a specific transcript and background noise is almost nonexistent [138, 140]. Technical bias is however given by very high abundance transcripts (as e.g. Ig-transcripts) reducing the detection probability of low abundance transcripts because of a predefined overall number of sequencing reads. This is addressed within normalization [49]. As RNA seq can use less input material (0.01 to 10 ng of RNA), it can be performed in more samples [126, 261]. This is an advantage especially in plasma cell diseases, in which frequently low numbers of purified malignant plasma cells are available [261].

Whole exome sequencing uses exonic DNA as input material. The protein coding exome accounts for approximately $2\% - 3\%$ of the whole genome [114]. To capture only the protein coding part, biotinylated oligonucleotide probes are used with sequences complementary to the coding sequence of the reference genome [139]. During the library preparation, the tags containing fragmented exonic DNA are bound by the probes and enriched with streptavidin beads binding to biotin on the probes [139]. Tags containing non-exonic DNA fragments do not bind to probes and thus can be removed from exonic DNA fragments [139]. Due to the target capturing, the amount of re-

quired input material is larger as compared to RNA seq, i.e. routinely 50 ng of DNA [139].
The principle aim of WES is the assessment of mutations, especially SNV, and small insertions and deletions (InDel). Whereas not all structural alterations can be detected by WES, the method allows detecting chromosomal gains and losses. The major advantage compared to whole genome sequencing are lower costs. To assess alterations specific to malignant cells, called somatic variants, a corresponding non-malignant (germline, normal) sample needs to be sequenced for each tumor sample. In this thesis, "somatic" is defined as only present in the tumor and not in the matched normal sample.
Whole genome sequencing is used for assessment of mutations, also in non-coding regions, and structural variations. As application in extended clinical routine implies a balance of cost and experimental yield. Whole genome sequencing compared to WES is only performed at the LfM in selected circumstances as analysis of paired samples.

1.7 Aim of this thesis

Primary aim of this thesis is to assess molecular properties and prognosis of patients suffering from AL in relation to other malignant plasma cell diseases including MM and its precursor state. This first comprises, as a basis to relate to, prospective target assessment and multimodal prediction of survival for personalized and risk-adapted treatment strategies in MM in the GMMG-MM5 multicenter trial. Second, to determine to what degree prognosis of AL patients is driven by malignant plasma cell factors in contrast to light chain deposition-based factors, and whether both are independent. And third, to assess the molecular properties of malignant plasma cells in patients suffering from AL in comparison to normal bone marrow plasma cells and those of patients with other plasma cell diseases, i.e. MGUS, AMM, and MM (pathophysiology). This comprises:

- **Assessment in MM within the randomized GMMG-MM5 phase III trial for relation to AL**

    1. Are gene expression-based risk assessments determining the malignant plasma cell properties in MM as good as the current standard risk stratifications?
    2. Is a personalized therapeutic recommendation possible by assessing the expression of target genes?

- **Regarding prognosis of AL patients**

    3. What role play malignant plasma cell characteristics *versus* properties associated with amyloid light chain formation and deposition (amyloidogenicity)?
    4. Do myeloma derived malignant plasma cell factors as proliferation or expression-based scores also determine risk in AL?
    5. Is it possible to define an expression-based risk score for AL patients, and does it in turn conveys prognostic significance in MM patients?

- **Regarding pathophysiology of AL**

    6. What are the differences and similarities of malignant plasma cells in AL in relation to MM and to the precursor stages MGUS and AMM?
    7. Do malignant plasma cells in AL represent a unique molecular entity in terms of pathophysiology?
    8. What "molecular age" can be attributed to the malignant plasma cells in AL, i.e. do they resemble myeloma cells, MGUS cells, or earlier precursors?

To answer these questions the worldwide largest cohort of molecular profiled patients classified as AL, MGUS, AMM, and MM in terms of iFISH ($n = 582/306/444/1691$), DNA microarray ($n = 196/64/271/765$), RNA sequencing ($n = 124/51/140/515$) and whole exome sequencing (AL, $n = 113$) provided by the Multiple Myeloma Research Laboratory will be analyzed.
For reaching aims 1. and 2. concerned to forming a basis for relation of AL-based factors to risk assessment of MM, report data generated by the GEP-R [198] for a cohort of 456 MM patients within the GMMG-MM5 clinical trial [191, 203] have to be analyzed as part of this thesis. Results are already published in shared first-authorship (Hose, D.*, Beck, S.* *et al.* [126]).
Five strategies to reach the six aims (3. to 8.) related to AL are pursued:
First, the risk of AL patients in terms of clinical prognostic factors associated with amyloid LC deposition and malignant plasma cell derived factors will be assessed. This includes the application of myeloma derived molecular risk factors to AL.
Second, whether gene expression-based risk assessments can be *de novo* defined for AL patients will be analyzed. No such score has previously been derived. For this, a strategy by Rème *et al.* [248] in collaboration with the LfM to create a risk assessment based on gene expression and survival data will be applied.

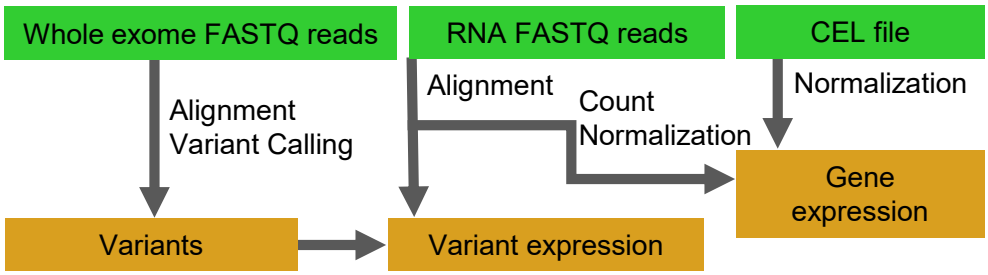

Figure 1.2: Outline of analyses performed in this thesis. Derived data is depicted in green colored boxes, data sets used for analysis in orange colored boxes. Key words for methods are indicated besides the gray colored arrows.

Third, the prognostic impact of the derived risk stratification for asymptomatic and symptomatic myeloma patients will be determined to assess the at which degree it depicts amyloid-specific risk *versus* malignant plasma cell derived risk.
Fourth, to assess pathogenesis, plasma cell properties will be assessed by transcriptome profiling of the different entities. Here, gene expression is assessed by DNA microarrays and by RNA sequencing.
Fifth, for the assessment of somatic variation of malignant plasma cells in AL *versus* other plasma cell diseases, a whole exome sequencing data analysis pipeline will be implemented (see figures 1.1 and 1.2). Subsequently, small genomic variants, i.e. single nucleotide variants, insertions, and deletions, as well as copy number alterations will be analyzed.
Finally, molecular findings will then be interpreted to derive a hypothesis about the "molecular age" and the existence of an AL specific "malignant plasma cell identity".

2 Materials and methods

The following chapter consists of a description of patients and samples, performed computational analysis, applied methods, and tools. For sample processing methods, see section 1.6.

2.1 Patients and samples

Consecutive 3023 patients, presenting at the University Hospital in Heidelberg or being treated within the GMMG-HD4 (ISRCTN64455289) [279] and the GMMG-MM5 (EudraCT no. 2010-019173-16) [191, 203] clinical trials, with available clinical and iFISH data were included in the analysis. For 1296 of these, DNA microarray data were generated. RNA seq data were available for 830 patients and WES data for 141 of these. Table 2.1 gives and overview regarding numbers of patients and samples in the specific data set per entity. Analyses are covered by votes of the ethics committee of the Medical Faculty of the Ruprecht-Karls-University Heidelberg (ethic vote no. 229/2003 and S152/2010) and are in accordance with the Declaration of Helsinki.

Table 2.1: Patients, samples, and investigations for the whole cohort per entity and analysis method: Samples of patients with light chain amyloidosis (AL) with subentity MGUS (ALMG) or multiple myeloma (ALMM), monoclonal gammopathy of undetermined significance (MGUS), asymptomatic multiple myeloma (AMM) or symptomatic multiple myeloma (MM). For comparison the variant table of CoMMpass multiple myeloma patients (MM CP) was used. Memory B cells (MBC), polyclonal plasmablastic cells (PPC), and healthy donor bone marrow plasma cells (BMPC) represent normal compartments. For comparison, human myeloma cell lines (HMCL) were used. Data for analysis were generated from interphase fluorescence *in situ* hybridization (iFISH), DNA microarray, RNA sequencing (RNA seq), variant tables and copy number data from whole exome sequencing (WES).

Entity	iFISH	DNA microarray	RNA seq	WES
ALMG	264	82	57	51
ALMM	318	114	67	62
MGUS	306	64	51	-
AMM	444	271	140	-
MM	1691	765	515	28
MM CP	-	-	-	930
BMPC	-	19	10	-
MBC	-	5	4	-
PPC	-	5	4	-
HMCL	-	54	26	-

2.1.1 Patients characteristics

Clinical parameters describing included AL patients by disease subentity are depicted in table 2.2. For MGUS, AMM, and MM see table 2.3. The distribution of AL specific biomarkers and staging systems is listed in table 2.4. The assessment of organ involvement of AL patients is depicted in table 2.5.

Table 2.2: Clinical characteristics of the assessed AL patient cohort splitted by underlying disease entity ALMG or ALMM. *n*: number of patients, NA: not available, AL: light chain amyloidosis, ALMG: AL with subentity monoclonal gammopathy of undetermined significance, ALMM: AL with subentity multiple myeloma.

		AL		ALMG		ALMM	
Variable	**Level**	***n***	**%**	***n***	**%**	***n***	**%**
Sex	female	229	39.3	110	41.7	119	37.4
	male	353	60.7	154	58.3	199	62.6
	NA	0	0	0	0	0	0
Age	≤60 years	246	42.3	101	38.3	145	45.6
	>60 years	336	57.7	163	61.7	173	54.4
	NA	0	0	0	0	0	0
Type	Bence Jones	311	53.4	147	55.7	164	51.6
	Double gammopathy	3	0.5	1	0.4	2	0.6
	IgA	50	8.6	18	6.8	32	10.1
	IgD	5	0.9	2	0.8	3	0.9
	IgG	176	30.2	72	27.3	104	32.7
	Other	7	1.2	4	1.5	3	0.9
	NA	30	5.2	20	7.6	10	3.1
Amyloid light chain type	Kappa	120	20.6	46	17.4	74	23.3
	Lambda	461	79.2	217	82.2	244	76.7
	NA	1	0.2	1	0.4	0	0
Plasma cell infiltration	<10%	265	45.5	232	87.9	33	10.4
	≥10%	270	46.4	31	11.7	239	75.2
	≥30%	37	6.4	0	0	37	11.6
	≥60%	8	1.4	0	0	8	2.5
	NA	2	0.3	1	0.4	1	0.3
Monoclonal protein	<20 g/L	203	34.9	90	34.1	113	35.5
	≥20 g/L	22	3.8	3	1.1	19	6
	≥30 g/L	8	1.4	0	0	8	2.5
	NA	349	60	171	64.8	178	56
Urinary monoclonal protein	<500 mg/24h	465	79.9	238	90.2	227	71.4
	≥500 mg/24h	75	12.9	6	2.3	69	21.7
	NA	42	7.2	20	7.6	22	6.9
Creatinine	<2 mg/dL	497	85.4	228	86.4	269	84.6
	≥2 mg/dL	85	14.6	36	13.6	49	15.4
	NA	0	0	0	0	0	0

Table 2.3: Clinical characteristics of the assessed patient cohort with MGUS, AMM, and MM. *n*: number of patients, NA: not available, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

		MGUS		AMM		MM	
Variable	**Level**	***n***	**%**	***n***	**%**	***n***	**%**
Sex	female	147	48	197	44.4	694	41
	male	159	52	247	55.6	997	59
	NA	0	0	0	0	0	0
Age	≤60 years	146	47.7	206	46.4	791	46.8
	>60 years	160	52.3	238	53.6	858	50.7
	NA	0	0	0	0	42	2.5
Type	Asecretory	0	0	1	0.2	12	0.7
	Bence Jones	16	5.2	26	5.9	320	18.9
	Double gammopathy	6	2	5	1.1	4	0.2
	Hyposecretory	0	0	0	0	4	0.2
	IgA	51	16.7	95	21.4	356	21.1
	IgD	0	0	0	0	14	0.8
	IgG	233	76.1	316	71.2	978	57.8
	Other	0	0	1	0.2	2	0.1
	NA	0	0	0	0	1	0.1
Plasma cell infiltration	<10%	291	95.1	67	15.1	123	7.3
	≥10%	10	3.3	291	65.5	306	18.1
	≥30%	0	0	64	14.4	423	25
	≥60%	0	0	11	2.5	378	22.4
	NA	5	1.6	11	2.5	461	27.3
Monoclonal protein	<20 g/L	268	87.6	234	52.7	340	20.1
	≥20 g/L	21	6.9	102	23	210	12.4
	≥30 g/L	1	0.3	82	18.5	861	50.9
	NA	16	5.2	26	5.9	280	16.6
Urinary monoclonal protein	<500 mg/24h	293	95.8	379	85.4	697	41.2
	≥500 mg/24h	3	1	35	7.9	406	24
	NA	10	3.3	30	6.8	588	34.8
Creatinine	<2 mg/dL	283	92.5	426	95.9	1454	86
	≥2 mg/dL	19	6.2	11	2.5	203	12
	NA	4	1.3	7	1.6	34	2

Table 2.4: Clinical characteristics and risk stratification of AL patient cohort, splitted by underlying disease entity ALMG or ALMM.*n*: number of patients, NA: not available, AL: light chain amyloidosis, ALMG: AL with subentity monoclonal gammopathy of undetermined significance, ALMM: AL with subentity multiple myeloma, FLC: free light chains, cTNT: cardiac troponin T, NT-ProBNP: N-terminal pro-brain natriuretic peptide type-B.

		AL		ALMG		ALMM	
Variable	**Level**	***n***	**%**	***n***	**%**	***n***	**%**
Difference FLC	<50 mg/L	68	11.7	48	18.2	20	6.3
	≥50 mg/L	140	24.1	82	31.1	58	18.2
	>180 mg/L	293	50.3	96	36.4	197	61.9
	NA	81	13.9	38	14.4	43	13.5
cTnT	<0.025 ng/mL	186	32	97	36.7	89	28
	≥0.025 ng/mL	343	58.9	144	54.5	199	62.6
	NA	53	9.1	23	8.7	30	9.4
NT-ProBNP	<1800 ng/L	210	36.1	119	45.1	91	28.6
	≥1800 ng/L	348	59.8	137	51.9	211	66.4
	NA	24	4.1	8	3	16	5
Mayo staging 2004	1	103	17.7	60	22.7	43	13.5
	2	165	28.4	83	31.4	82	25.8
	3	271	46.6	105	39.8	166	52.2
	NA	43	7.4	16	6.1	27	8.5
Mayo staging 2012	0	70	12	46	17.4	24	7.5
	1	76	13.1	40	15.2	36	11.3
	2	128	22	57	21.6	71	22.3
	3	181	31.1	63	23.9	118	37.1
	NA	127	21.8	58	22	69	21.7
European staging 2013 advanced NT-ProBNP	I	103	17.7	60	22.7	43	13.5
	II	165	28.4	83	31.4	82	25.8
	IIIA	137	23.5	47	17.8	90	28.3
	IIIB	132	22.7	57	21.6	75	23.6
	NA	45	7.7	17	6.4	28	8.8

Table 2.5: Organ involvement of AL patient cohort, splitted by underlying disease entity ALMG or ALMM. *n*: number of patients, AL: light chain amyloidosis, ALMG: AL with subentity monoclonal gammopathy of undetermined significance, ALMM: AL with subentity multiple myeloma.

Variable	Level	AL		ALMG		ALMM	
		n	**%**	***n***	**%**	***n***	**%**
Heart involvement	no	144	24.9	81	30.7	63	20
	yes	435	75.1	183	69.3	252	80
Kidney involvement	no	234	40.3	83	31.4	151	47.8
	yes	346	59.7	181	68.6	165	52.2
Liver involvement	no	460	79.4	211	80.2	249	78.8
	yes	119	20.6	52	19.8	67	21.2
Soft tissue involvement	no	353	61.2	190	72.5	163	51.7
	yes	224	38.8	72	27.5	152	48.3
Peripheral neuropathy	no	496	85.7	216	82.1	280	88.6
	yes	83	14.3	47	17.9	36	11.4
Autonomous neuropathy	no	506	87.4	232	88.2	274	86.7
	yes	73	12.6	31	11.8	42	13.3
Gastrointestinal tract involvement	no	367	63.3	169	64	198	62.7
	yes	213	36.7	95	36	118	37.3
Lung involvement	no	565	97.8	258	98.1	307	97.5
	yes	13	2.2	5	1.9	8	2.5
Number of involved organs	1	122	21.1	61	23.1	61	19.4
	2-4	410	70.8	180	68.2	230	73
	>4	47	8.1	23	8.7	24	7.6

2.1.2 Molecular diagnostics of patients

As described before and depicted in table 2.1, patient samples were processed by four different molecular profiling methods, i.e. iFISH, DNA microarray, RNA seq, and WES. If not otherwise specified, all procedures were performed at the LfM according to the laboratory's SOP [125]. See figure 1.1 for an overview.

Cell purification

After bone marrow aspiration, the bulk of the material was subjected to density gradient centrifugation, and subsequently plasma cells were purified with anti-CD138 microbeads by automated magnetic cell sorting (autoMACS Pro, Miltenyi Biotec) and fluorescence activated cell sorting (FACSAria; Becton Dickinson) [125]. The purity of $CD138^+$ plasma cells was controlled, and at least 80% were required as quality measure [125]. Part of the purified $CD138^+$ plasma cells were used for iFISH [125, 126]. Another fraction was subjected to RNA and DNA extraction and subsequently used for DNA microarray and RNA seq, and WES, respectively [125, 126]. To perform somatic variant calling from WES, additionally DNA from non-malignant cells of the same patient was used.

For comparison to different normal cellular compartments, expression data using DNA microarrays and RNA seq of peripheral CD27$^+$ MBC as well as *in vitro* generated and differentiated PPC [212] (highly proliferative, non-malignant) were generated [125, 264]. As highly proliferating, malignant comparator, human myeloma cell lines (HMCL) were used [125, 264]. The HG-lines HG1, HG3, HG4, HG5, HG6, HG7, HG8, HG9, HG11, HG12, HG13, HG14, HG15, HG17, and HG19 were generated at the LfM [125, 264] and the cell lines XG1, XG2, XG3, XG4, XG5, XG6, XG7, XG10, XG11, XG12, XG13, XG14, XG16, XG19, XG20, XG21, XG22, XG23, and XG24 at Montpellier, France [213, 331]. The cell lines L363, SK-MM-2, LP-1, RPMI-8226, AMO-1, KMS-18, JIM-3, JJN3, KARPAS-620, KMS-12-BM, ANBL-6, KMS-11, MM1S, NCI-H929, KMS-12-PE, KMS-28-BM, OPM-2, MOLP-8, MOLP-2, KMM-1, U266, and EJM were purchased from the German Collection of Microorganisms and Cell Cultures, American Type Cell Culture, or Japan Health Science Research Resources Bank.

Interphase fluorescence *in situ* hybridization

The iFISH panel for extended clinical routine at the LfM comprises iFISH probes for chromosomal regions 1q21, 4p16, 5p15, 5q31, 5q35, 8p21, 9q34, 11q13, 11q22, 11q23, 13q14, 14q32, 15q22, 16q23, 17p13, and 19q13, and IgH-TL t(4;14)(p16.3;q32.3), t(11;14)(q13;q32.3), t(14;16)(q32.3;q23), and t(6;14)(p21;q32.3) (Poseidon Probes, Kreatech) [125]. IgH-TL with unknown partners were identified with the IgH-rearrangement probe (Poseidon Probes, Kreatech). Hybridization of iFISH probes (Kreatech and Meta-Systems) to CD138$^+$ plasma cells was carried out according to manufacturer's instructions.

HRD was assessed according to Wuilleme *et al.* [323]: A sample was termed hyperdiploid if gains in two of the three chromosomes 5, 9, and 15 were detected [323]. The percentage of cells carrying a specific aberration were used to define the clonality of an aberration [222]. If above 60% of cells carry a specific aberration, it was termed clonal, in the range of $20\% - 59\%$ of cells it was termed subclonal [222]. The percentage of malignant plasma cells was estimated by the fraction of cells harboring the most frequent aberration [222].

DNA microarray

RNA was extracted from CD138$^+$ plasma cells samples with the RNeasy Mini Kit (QIAGEN) [125–127, 129, 266]. Labeled cRNA was generated using the small sample labeling protocol vII (Affymetrix), fragmented, and hybridized to DNA microarrays (HG-U133 2.0 plus, Affymetrix) according to manufacturer's instructions [3].

RNA sequencing

For RNA seq, 5 ng of total RNA were used for full-length double-stranded cDNA preparation [259, 261, 262]. Amplification was performed using the SMARTer Ultra Low RNA Kit (Illumina) [259, 261, 262]. The Libraries were prepared from 10ng of fragmented cDNA following the NEBNext Chip-Seq Library Prep protocol (New England Labs) [259, 261, 262]. Sequencing was done on an Illumina Hiseq2000 with $2 * 50$ bp or $2 * 75$ bp paired-end reads [259, 261, 262]. FASTQ files that contain the sequencing reads were prepared with Illumina bcl2fastq software [259, 261, 262].

Whole exome sequencing

WES was performed in extended clinical routine at the LfM by the following method: Exome capture and library generation were performed from $10 - 50$ ng of extracted DNA using the Nextera Flex Exome Enrichment Kit (Illumina) according to the manufacturer's instructions [139]. Libraries were prepared as $2 * 151$ bp paired-end reads and sequenced using an Illumina Novaseq 6000 sequencer. A mean read coverage of 30x was aimed at. The Illumina sequencing software HCS (version 1.4.0) and basecall software RTA (version 3.3.3) were used. Raw sequencing results were processed with the Illumina bcl2fastq software (version 2.20.0.422).

2.2 Computational and statistical methods

Computations were performed on Ubuntu 18.04 and Windows Server 2012 using freely available software. If possible, parallelization of the code was realized using GNU parallel [288] on the Ubuntu 18.04 operating system. A complete list of the used tools is depicted in supplementary table A.1.

Subsequent statistically analysis and graphical interpretation of results were made in R [239]. Within R, pre-build analysis methods are bundled up in packages. For biological questions, the Bioconductor project [105] offers additional R packages. For the list of used R packages, refer to supplementary table A.3.

Statistical methods

Comparisons of two groups with nominal data were performed with Fisher's exact test [5] for 2 *x* 2 contingency tables (see example table 2.6) or Pearson's χ^2 test [5] for larger contingency tables. The magnitude of differences between groups were depicted by odds ratio (OR), see formula 1. The overlap rate (efficiency) was calculated from 2 *x* 2 contingency tables with formula 2. It contains the percentage of overlaps between two groups.

For testing of sorted differences among ordered groups, a Jonckheere-Terpstra test [144, 268] was applied. The number of permutations for the reference distribution was set to 1000 to obtain a permuted p-value if the number of analyzed samples was larger than 100.

Table 2.6: Example of a 2 *x* 2 contingency table for calculating Fisher's exact test, odds ratio, and efficiency.

Number of patients		
	in group Y	**not in group Y**
in group X	*a*	*b*
not in group X	*c*	*d*

$$OddsRatio = \frac{a/b}{c/d} = \frac{a*d}{b*c} \tag{1}$$

$$Efficiency = \frac{a+d}{a+b+c+d} \tag{2}$$

For the depiction of metric data, boxplots were used. The boxes range from the 25% quantile over the median to the 75% quantile. This defines the interquartile range (IQR). The lower and the upper whiskers correspond to 1.5 x IQR. The notches at the median value depict the 90% confidence interval for the median. For the comparison of metric values in the boxplots, Wilcoxon's rank sum test [16] was applied.
Fisher's test (F test) was applied to compare variances of two groups of metric data [97].
A difference was termed significant if the p-value of the respective test was ≤ 0.05. Corrections for multiple testing were performed using the Benjamini-Hochberg (BH) method [18].

Survival analysis

To assess the effect of a variable in disease progression, survival analysis was performed. Survival data were collected for all patients in the study cohort. Survival data consists of a time span and an outcome status that indicates if the event occurred. The time span is the interval from diagnosis to specific event, e.g. disease progression or death. If no event occurred at the end of the time span, the patient is censored. In overall survival (OS) analysis, the measured status is the death of the patient. For progression free survival (PFS) of MM patients, a relapse of the disease or patient's death

is considered as event. For AMM patients, progression to therapy requiring, symptomatic MM or the patient's death is considered as event.
A formula was constructed between the survival data and the variable of interest to receive an estimate. For survival curves and median time to event, the estimate was computed with nonparametric survival estimates for censored data using the Kaplan-Meier method [98, 149] in R. In the plot, the time period in years is marked on the x-axis, the survival rate in percent is depicted on the y-axis. If a patient had an event, the curve drops, if a patient was censored, a crossing vertical line is drawn at the respective timepoint. The Log-rank test was performed per analysis to test for difference between the curves [120]. A difference was termed significant if the p-value was ≤ 0.05. Median survival time for a variable was measured as the time point at which the curve meets the 50% of the survival rate. Survival rates, describing the percentage of patients not showing an event in accordance with the definition above, were assessed after two and five years.
Based on Cox's proportional hazard model, univariate and multivariate survival regression analysis were performed [8, 292]. As univariate variables, categorical groups with multiple levels were used. In multivariate analysis, several univariate groups were compared. The magnitude of difference between two levels of the groups was calculated as hazard ratio (HR). The hazard describes the risk to experience an event at a time point t and is written $h(t)$. It can be interpreted as actual mortality rate. In formula 3, the calculation of the HR is described for the two groups a and b. Here, a HR > 1 indicates that the risk in group b is higher than in group a, and *vice versa* for a HR < 1. A HR of approximately 1 implies no difference between the mortality rates of the two groups.

$$HazardRatio = \frac{h_b(t)}{h_a(t)} \tag{3}$$

To determine significance of the HR, a Wald test was applied [306]. The reliability of a significant HR is bound to the proportionality between the two assessed groups in the time span and was controlled [8, 292]. In the multivariate Cox regression, a significant HR indicates that the investigated variables are independent predictive.
An integrated Brier score was calculated to evaluate the prediction accuracy for prognostic risk assessments [208, 260, 297]. As input, Cox's proportional hazard models are suggested. For cross validation, a subsampling parameter of $2/3$ and a bootstrapping parameter of $1/3$ of the complete test cohort are recommended [208]. Significance of the Brier score was determined by the van de Wiel test [297].

2.3 Gene expression profiling with DNA microarray

GEP data from DNA microarray were available for 196 patients with AL, 64 with MGUS, 271 with AMM, and 765 with MM (see table 2.1). As a reference, GEP for HMCL ($n = 54$), BMPC ($n = 19$), PPC ($n = 5$), and MBC ($n = 5$) were used.

Preprocessing was performed using the `just.gcrma` function from the gcrma [321, 322] package in R. The function applies a background correction, normalization, and log transformation to the basis of 2, converting the raw intensities to expression values for every gene. It uses the robust multiarray average method (RMA), in awareness of the GC content. For background correction, experimental gained reference values for non-specific binding are implemented to determine the amount of background noise for every probe. The values are normalized to all CEL files in the investigated cohort and subsequently log-transformed to the basis of 2, with results being termed "expression values".

After normalization, a batch correction with ComBat [143], which uses a nonparametric Bayesian approach to filter batch effects of known batches, was applied. This was necessary due to the usage of different IVT labelling kits for preparation of the different microarrays over time as implied by the manufacturer due to changes in the availability of kits.

The expression values of the gcrma normalized microarray data are subject to background noise. To distinguish whether an observed expression of a gene is plausible, the "presence and absence of negative strand matching probes" method (PANP) [312, 313] was applied. "Negative strand matching probesets", i.e. those being inherently without a hybridization partner, are used to calculate a matrix of p-values for the expression values in the given expression data. From the population of p-values, a loose and a tight cutoff value are determined for every gene. In this thesis, the loose cutoff was used to determine if an expression value was present or absent. A probe with an expression value above the cutoff was termed "present" and below the cutoff as "absent".

A different normalization with the `mas5` function from the affy R package [32, 103] was performed, as being necessary for calculation of the UAMS70 score [272], the Myc-activation index [55], and the molecular classification [329] (described in section 2.3.2). Mas5 strictly subtracts the values of the mismatch probes from the perfect match probes. Mismatch probes have one exchanged nucleotide and should therefore not bind (or at least less) specifically [3]. In contrast to the gcrma normalization, mas5 normalizes every sample independently and does not perform a log transformation.

The gcrma and the mas5 normalized expression values of the complete cohort were used for calculating expression-based risk scores and classifications, as outlined in section 2.3.2, and for dimension reduction methods, described in section 2.5.1.

2.3.1 Clinical risk assessment by the GEP-Report

Analyses of reports generated by the GEP-R framework [198] were published as part of this thesis [126]. Normalization to a reference cohort of 262 MM samples, by the GEP-R, improves comparability and enables to report prospective GEP-based risk stratification to each patient in a clinically reasonable time frame considered as during the first cycle of induction therapy [126, 198].

The risk assessments calculated by the GEP-R were, among others, compared by the Brier score according to their prediction accuracy (as described above in section 2.2). For this, a subset of 451 patients with complete prediction information (GEP-R and rISS) was used within a 69/73 month period for evaluating the performance of risk assessment regarding PFS/OS. As input for cross validation, a subsampling parameter of 301 and a bootstrapping parameter of 150 were chosen, based on the sample size (described in section 2.2).

2.3.2 Myeloma derived gene expression-based risk assessments

Normalized GEP data were used to calculate risk scores and molecular classifications previously published for MM. Six different scores and two classifications were used: to assess risk by biological variables, the gene expression-based proliferation index (GPI) by Hose *et al.* [128], created at the LfM, and the Myc-activation index (MAI) by Chng *et al.* [55] were calculated. For classifications, the translocation/cyclin D (TC) by Chng *et al.* [54] and the molecular classification (MC) of multiple myeloma by Zhan *et al.* [329] were computed. The respective risk scores are the UAMS 70-gene score (UAMS70; University of Arkansas for Medical Sciences) by Shaughnessy *et al.* [272], the IFM 15-gene score (IFM15; Intergroupe Francophone du Myélome) by Decaux *et al.* [68], the EMC92 by Kuiper *et al.* [161], and the Risk-score (RS) by Rème *et al.* [248]. For more details on the scores and classifications, see section 1.2.4.

GPI, TC, MC, UAM70, and IFM15 are all combined in the GEP-R reporting tool, previously developed by Meissner *et al.* [198] at the LfM and used in the GMMG-MM5-trial. Application and respective analysis are part of this thesis and being published in shared first authorship Hose, D.*, Beck, S.* *et al.* [126].

In this thesis, the analysis comprises a comparison of the proportions of the score and classification assessments in AL *versus* MGUS, AMM, and MM, as well as AL underlying subentities ALMG and ALMM. Additionally, the assessments were compared to AL specific clinical characteristics, i.e. presence of heart involvement, NT-ProBNP level, difference of FLC level, AL type, and creatinine level substratified by AL suben-

tities ALMG and ALMM. Finally, for the AL patient cohort, association of every score and classification with OS was assessed.

2.3.3 New risk assessment for light chain amyloidosis

The Heidelberg AL score (HDAL) [17] was created at the LfM within this thesis, according to a strategy by Rème *et al.* [248]. The score categorizes patients into three distinct groups based on gene expression and clinical outcome. For the score, 59 prognostic genes were selected, 15 of them associated with good prognosis and 44 with poor prognosis in the AL patient samples of the training group. Information on the prognostic genes are described in supplementary table A.12. The score was analyzed in analogy to the MM scores, and its prognostic value was tested against the current standard AL stagings. The list of prognostic genes was analyzed according to their role in plasma cells and cellular pathways by gene set enrichment analysis with metascape [337] (see section 2.5.6). Using this approach, it takes about one minute to classify a new patient sample. The code is depicted in supplement B.1.

2.4 Next generation sequencing

In this thesis, RNA seq data of 830 and WES data of 141 patient samples were analyzed (see table 2.1). Sequencing data were generated as described in section 1.6.3 and 2.1.2.

2.4.1 Human reference genome

The human reference genome FASTA file [2] (GRCh38; release 77) was downloaded from Ensembl [328]. It was the latest available version of human reference genome, is approximately 3.6 billion base pairs long, and contains annotations for more than twenty thousand coding genes. Only the primary assembly, which is not influenced by the changes of the routinely released patches, was used for sequence alignments.

2.4.2 Sequencing file formats and quality assessment

FASTQ format

The sequencing reads were preprocessed by the Illumnia bcl2fastq software to FASTQ format [57] and saved in compressed text files.
Every read in a FASTQ file consists of four text lines:
1. Sequence read identifier, starting with an @ symbol

[2] Ensembl: human reference genome FASTA; Online resource: `ftp://ftp.ensembl.org/pub/release-77/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz`; Status: 2015-10-01, 13:46

2. Sequence (the four bases A, G, T, and C and N for "unsure" bases)
3. Quality score identifier line (contains only a + sign)
4. Base quality score (Phred score; one symbol encoding quality for each base by Illumina 1.9 encoding)

See short example read in code 2.1. Base quality is depicted by Illumina 1.9 encoding and is explained below and in table 2.7.

```
1 @NS500188:53:H3J5VBGXX:1:11101:25611:1537 1:N:0:GCTACGCT}
2 CCTCAAAACACCTGAGTTTAGTTCTTGCCCAGACCTGGAAAGCTAAGGACACACCTCCGATTTT}
3 +
4 AAAAAFFFF.FFFFAFFAF.FAAFF7FFFFAAFFFFFFFFFFFF.FFFFFAFFFFFFFF.FFFF
```

Code 2.1: Example sequence read in FASTQ format. For explanation, see text and table 2.7

Base quality

Within the sequencing process, the base is identified by the intensity of a fluorescence signal in the cluster of the read sequence [141]. Afterwards, the Illumina sequencing software calculates the expected base and a base quality, the Phred scaled base quality score [92, 93]. The Phred score determines the error probability for a base call [92, 93]. Reads with a mean base quality below 10 were dismissed in the subsequent analyses. For overview of the error probability of the different Phred scores, see table 2.7. In every read, one symbol for a Phred score is encoded by Illumina 1.9 encoding [92, 93, 181]. See the last row of the example read in code 2.1.

Table 2.7: Error probability of different Phred scaled base quality scores and the resulting base call accuracy in percent.

Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Sequence alignment format

The alignment is stored in the sequence alignment map (SAM; human readable) format or its compressed binary version (BAM) [181]. Read information from the FASTQ file [57] are enriched with information from the alignment step [181] (further described in section 2.4.3 and the alignment part in section 2.4.4). The SAM/BAM file consists of an optional header section and an alignment section [181]. The header lines start with an "@" symbol and contain information on the preprocessing of the alignment and the

used programs [181]. In the alignment section, every mapped read is written in one line with single fields tab separated [181]. Eleven fields are mandatory, describing the read, the alignment and further information for subsequent analysis tools [181].
The aligner determines a best mapping position for every read in the reference genome, calculates a mapping quality per read, and saves the information about the read mapping direction, mismatching bases, insertions, and deletions [181].
The mapping quality is Phred scaled like the base quality [92, 93, 181]. It reflects the probability that a read is misplaced, e.g. a multiple mapping read receives a mapping quality of zero [181].

2.4.3 RNA sequencing analysis

The RNA seq analysis pipeline was generated for MM patient samples by Martina Emde [90] at the LfM and applied to AL patient samples in this thesis. Quality of all RNA seq FASTQ files was controlled and files of insufficient-quality were re-performed [90]. For the alignment of the RNA seq reads, the STAR [76] aligner was used. First, genome index files were generated by STAR, using the human reference genome FASTA file (see section 2.4.1) and an additional GTF (Gene Transfer Format) file [3] (GRCh38; release 82), which were downloaded from Ensembl [328]. The GTF file includes annotations to all known genes and is used to map the gene name to the position in the reference genome for further read counting. Alignments in BAM format were created using STAR with default options. STAR uses HTSeq [7] internally to count the reads per gene with the union method. Read counts of technically replicated BAM files were summed up per patient sample. Then, read counts were saved together in a matrix, with each column representing a sample and each row representing a gene. Afterwards, unstranded read counts were normalized with edgeR [49, 195, 251] for expression analysis. Normalization factors were determined with the trimmed mean of M-values method [49, 195, 251]. The raw read counts were adjusted to library size and afterwards normalized with the "counts per million" (CPM), and the "reads per million per kilobase" (RPKM) methods [49, 195, 251]. The RPKM method takes the gene length into account, which is an advantage for comparisons between different genes, but not between different samples [252]. As expression values per gene, normalized counts were log transformed to the basis of 2 with a prior count of 1.
Both, i.e. CPM and RPKM normalized expression values, were used as gene expression of mutated genes in the variant table, described below in part "final variants" in section 2.4.4. CPM normalized expression values were used for dimension reduction

[3] Ensembl: human reference genome GTF; Online resource: `ftp://ftp.ensembl.org/pub/release-82/gtf/homo_sapiens/Homo_sapiens.GRCh38.82.gtf.gz`; Status: 2015-10-05, 09:19

analysis, described in section 2.5.1, analysis of Ig gene expression, described in section 2.5.3, and to determine if gene expression is altered in CNA, described in section 2.5.4. Read counts were used for differential expression analysis as described in section 2.5.2. For references of used tools in the RNA seq pipeline, see supplementary table A.1 and A.3.

2.4.4 Variant calling pipeline

The following section describes the variant calling pipeline created in this thesis and applied to 113 AL patient samples. The complete pipeline from raw FASTQ files to the final table of variants is depicted in figure 2.1. To speed up computations, individual steps were executed in parallel using GNU parallel [288]. For references of applied tools used in this pipeline, see supplementary table A.1.

Quality control and read trimming

The quality control of sequencing data is important, as for confident variant calling, e.g. in mutation detection, a high quality of sequencing reads is essential. Sequencing reads of low quality, adapter sequences, and common sequencing errors can be filtered or cut out. For downstream analysis tools, for example the aligner, a higher read quality and no or a low sequencing adapter content led to a more precise alignment. The quality of the sequencing reads was controlled with the Fastqc [9] and fastp [47] software. Raw FASTQ files were used as input and a quality profiling was implemented before and after read trimming. Reads were trimmed and filtered with fastp [47] using default options, according to Chen *et al.* [47] (supplementary code B.2 line 28). Sequencing adapters were automatically detected and trimmed [47]. "Bad" reads, which were too short, of too low quality, or had too many N bases in the read, were filtered out [47]: a read was determined as too short, if its length was below 15 bases [47]. If a base had a Phred scaled quality score below 15, it was discarded [47]. A single read could contain up to 40% of low-quality bases before being discarded [47]. Reads were filtered out if a read contained more than 5 N bases [47].

The assessment of sequencing quality was continued regarding the alignments (see figure 2.1). Reports and metrics were generated with Picard [38] (supplementary code B.3 line 46), GATK [196] (supplementary code B.3 line 60), samtools [181] idxstats (supplementary code B.4 line 16), and samtools [181] stats (supplementary code B.4 line 18) for alignments. The coverage of the targeted exome for every alignment was assessed with Alfred [246] (supplementary code B.4 line 30). Alfred uses coordinates for all genes from the reference genome as exome target [246].

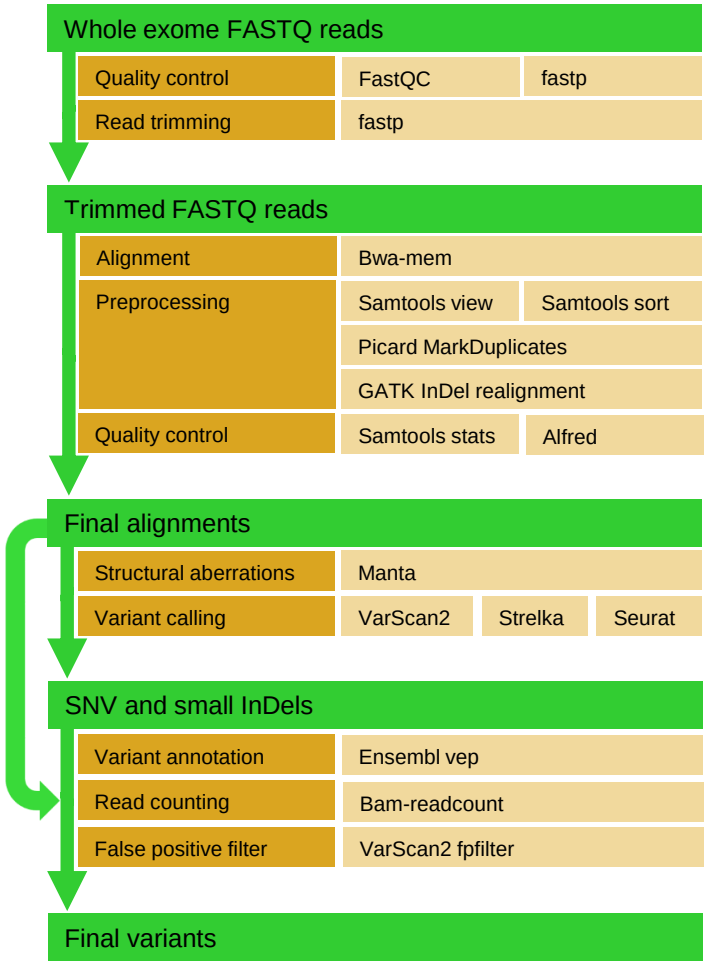

Figure 2.1: Flowchart of used variant calling pipeline. Input, intermediate, and final data are depicted in green colored boxes, the respective processing step, described in section 2.4.4, is depicted in yellow color, and the name of the tools in light yellow color. For references of applied tools used in this pipeline, see supplementary table A.1

With the MultiQC [91] tool, quality reports from multiple files from different report generating tools were collapsed to a summary report on all samples.

Alignment

After preprocessing the reads, the remaining high quality trimmed reads were aligned (see figure 2.1). Alignments were saved in the SAM format [181] (see section 2.4.2). Reads were mapped to a reference with the objective of the "best" match [179, 180]. Taking mismatching bases, insertions, and deletions into account, the aligner determines where each read maps best to the reference [179, 180].

The alignment of FASTQ files to the human reference genome was performed with bwa [179, 180], more specifically with the `bwa mem` command (see supplemental code B.3 line 39). The `bwa mem` aligner is designed for reads in a range of 70 bp to 1 Mbp to a known reference genome [179, 180]. It is suggested as ideal for the subsequent calling of variants like SNV and InDel [179, 180]. It is based on the "Burrows-Wheeler transformation" [40], a block sorting data compression method, and on the "finding super-maximal exact matches" (SMEM) algorithm [178, 179].

For a fast and accurate alignment, the generation of index files from the reference genome is helpful. The `bwa index` command uses the "Burrow-Wheeler Transformation Smith-Waterman" (BWT-SW) method [178, 179]. This creates a bidirectional FM-index [180] (Ferragina-Manzini), which speeds up computations for the price of an increased memory requirement (see code in supplement B.3 line 36) [178, 179].

The `bwa mem` aligner is the most often recommended and used aligner for WES [102], including large-scale sequencing projects, e.g. 1000Genomes [14].

Processing alignments

Variant calling algorithms recommend preprocessing of alignments before calling variants [102, 154] to reduce bias introduced by sequencing. The decision whether a mismatch base is a true mutation or a sequencing error depends on several criteria, e.g. base quality, mapping quality, or position in the read (see in section 2.4.2 the parts on sequencing quality).

The bwa alignment in SAM format was directly piped to samtools to skip unnecessary memory operations. Reads were compressed to BAM format with `samtools view` [181] and sorted with `samtools sort` (supplemental code B.3 line 39). Compression to BAM format and sorting saves hard disk space and computing time of downstream tools.

As next step, read duplications were marked. Duplicated reads arise either during the PCR amplification, or the enrichment phase before library preparation (PCR duplicate), or a single large cluster is reported as two clusters by the sequencing software (optical duplicate) (see introduction to NGS in section 1.6.3). The marking of duplicates was carried out by the java (version 1.8) application Picard [38]. With the `Picard`

`MarkDuplicates` command, duplicated reads were marked in the BAM files. Additionally, index files (BAI) were created for every BAM file (see supplemental code B.3 line 46).

As last step of preprocessing, InDel realignment was performed with GATK [196]. Mismatching bases often arise during alignment in the direct neighborhood of short insertions and deletions. A local realignment around these regions minimizes the number of mismatching bases. For this, targets were created only for chromosomes $1-22$, X, and Y with `GATK RealignerTargetCreator` (see supplemental code B.3 line 57). This command scans all given BAM files at once and creates a list of potential misaligned target positions. Then, BAM files were realigned with `GATK IndelRealigner` (see supplemental code B.3 line 60). In addition to the realignment, unmapped reads, reads marked as secondary alignment, or as duplicate were removed from the resulting BAM files with the `-rf` option, to save hard disk space. The full code can be found in supplemental code B.3.

Variant calling

Somatic variant calling was performed with sequencing data from tumor samples (malignant plasma cells) in relation to the respective germline samples for each patient. The somatic variants were called by three different variant callers. For parameter adjustment, a minimum base quality of 13 and a minimum read mapping quality of 10 was applied. A variant was kept if it was called and passed the internal quality filters of at least two of the three callers. This increased precision [112].

The first caller is VarScan2 [158], which runs within java (version 1.8.) (see supplemental code B.5). VarScan2 uses samtools [181] pileup files. A combined pileup file was created from the germline and the tumor BAM file pairs with the `samtools mpileup` command (see supplemental code B.5 line 28). Then, the `VarScan2 somatic` command was used (supplemental code B.5 line 30). It performs a pairwise comparison of tumor versus germline by base calls and normalized sequence depth from the pileup file. A heuristic algorithm detects variants and determines the statistical significance by Fisher's exact test. The variants are classified as SNV, insertion or deletion. VarScan2 detects, somatic variants, loss of heterozygosity (LOH), and germline variants. Only somatic variants of high confidence were used for downstream analysis as assessed by the `VarScan2 somaticFilter` command. In addition to the filter step with a p-value threshold of ≤ 0.05 and a minimum variant allele frequency (VAF) of 10%, SNV near InDel were re-evaluated.

The second caller used is the java (version 1.6) application Seurat [56] (see supplemental code B.6), which uses the GATK [196] routines internally. It is based on a gen-

eralized Bayesian analysis framework for somatic variant calling from pairs of tumor and germline BAM files. Seurat calls LOH and somatic variants, which are classified as SNV, insertion, or deletion.

The third caller is Strelka [154] (see supplemental code B.7). Strelka is a python command line tool, which runs in two steps. First, the parameters for variant calling are automatically fitted per sample and an internal score is produced. Second, the workflow runs with the fitted parameters and determines the variants. Strelka uses a mixture-model-based estimation. Thereby, a variant probability model is supplemented by a final empirical variant scoring step. Before variant calling with Strelka, structural variants were called with Manta [48] (see supplemental code B.7 line 23). These were necessary for the Strelka parameter fitting.

All used variant callers return their results in variant call format (VCF) files. VCF files were merged per patient sample with an R script (see supplemental code B.8). Variants were kept if they were identified as somatic and passed the caller intern filter criteria (see supplemental code B.8 line 34 to 59 and 110 to 126). The variant tables were reduced to one entry if a variant was found by more than one caller (see supplemental code B.8 line 95 to 107 and 166 to 176). BED (browser extensible data) files with only the variant positions for read counting, as well as tables with variant positions and the alleles for false positive filtering were generated.

Variant filtering

Read counts of the DNA and the RNA tumor BAM files for each variant at the respective variant position were counted using bam-readcount [175] (see supplemental code B.9 line 25 and line 28). For bam-readcount, criteria for base quality (13) and mapping quality (10) in DNA were the same as for the variant call, while in RNA a mapping and base quality above 1 was considered sufficient.

Bam-readcount sums up the reads for the variant and the reference base and calculates quality measurements like the average mapping quality, the average base quality, the number of reads on the forward and on the reverse strand, and the average position of the variant in the read. The subsequently used VarScan2 false positive filter `fpfilter` controls these measures, related to the reliability of the detected variants. A complete list of measurements and the respective filter criteria is depicted in table 2.8. DNA read counts of variants were utilized by the VarScan2 `fpfilter` function to exclude "probably false positive" variants and to calculate the VAF. The RNA read counts were used to determine which of the variants detected in DNA were expressed and to estimate a VAF for RNA (see supplemental code B.9 line 30 and line 32).

Table 2.8: Quality measurements for called variants calculated by bam-readcount used to exclude probably false positive variants with VarScan2 false positive filter

Measurement	Description	Filter Criterion
Count	Number of reads supporting the variant base	4
Mapping quality	Minimum average mapping quality	30
Base quality	Minimum average base quality	30
Strandedness	Fraction of supporting reads from the forward strand	1%
Read position	Average variant position in supporting reads from the ends of the read as fraction of the read length	10%
Mapping quality difference	Maximum difference in average mapping quality between reference and variant reads	50
Read length difference	Maximum difference in average trimmed read length between reference and variant reads	25
Mismatch quality difference	Maximum difference in average mismatch quality sum between reference and variant reads	50
VAF	Minimum variant allele frequency	10%

Variant annotation

To determine whether a variant is located in a protein coding part of the genome, variant annotation is necessary. Variant annotation was performed with the Ensembl variant effect predictor (vep) [197] and Ensembl [328] annotations (GRCh38; release 94)[4] (see supplemental code B.11). Besides the HGNC [325] gene symbol, variant consequence by Sequence Ontology [87] and default annotations, the SIFT [304] prediction, the PolyPhen [2] prediction, the Ensembl variation database for known genetic variation, and information about allelic frequencies from different large scale sequencing projects were included. SIFT [304] (Sorting Intolerant From Tolerant) and PolyPhen [2] (Polymorphism Phenotyping) are prediction tools for amino acid substitutions in protein coding regions. A calculated score predicts the impact of a mutation on the resulting protein, and its function and is subsequently graded into categories by both tools [2, 304].

Final variants

Annotated variants without a HGNC [325] gene symbol were excluded, and a filtering according to the calculated consequences, defined by Sequence Ontology [87] was

[4]Ensembl: human reference variant annotation; Online resource: `ftp://ftp.ensembl.org/pub/release-94/variation/VEP/homo_sapiens_merged_vep_94_GRCh38.tar.gz`; Status: 2018-10-11, 14:12

performed (see supplemental code B.12 line 27 and 29). Only variants with the following annotated consequences were included: `splice_acceptor_variant`, `splice_donor_variant`, `stop_gained`, `frameshift_variant`, `stop_lost`, `start_lost`, `inframe_insertion`, `inframe_deletion`, `missense_variant`, `splice_region_variant`, `start_retained_variant`, `stop_retained_variant`, and `synonymous_variant`. If a variant was assigned to multiple annotations, the one with the most severe consequence, as estimated by Ensembl [328], was used for downstream analyses (see supplemental code B.12 line 34 to 44).
Annotation data were merged with read counts and false positive filtering results. Read counts were used to calculate the final VAF for DNA and RNA (see supplemental code B.10 line 28 to 31). The used formula for calculating the VAF is depicted below (4). Only variants annotated with a coding consequence and a minimum VAF of 10% were kept. Variation in the frequency of appearance of mutations within one patient sample, (e.g. between different mutations) give evidence for presence of different subclones, i.e. *intra*-patient heterogeneity. To clarify, if a mutation is clonal in a diploid organism its VAF is 50% for a heterozygous and 100% for a homozygous mutation.

$$Variant\ Allele\ Frequency = \frac{\#Variant\ Reads}{(\#Reference\ Reads + \#Variant\ Reads)} * 100 \qquad (4)$$

For every gene affected by a variant in a sample, the expression value of RNA (assessed as described in section 2.4.3) was used as assessment whether the respective gene is expressed, and thus taken as an indicator for the actual influence of the mutation (see supplemental code B.12 line 58 to 70).
Per patient variant tables were merged with R (see supplemental code B.13).

2.4.5 Copy number calling

By copy number analysis, the ploidy status of chromosomal regions can be analyzed from WES similar to the assessment of numerical CA by iFISH (described in section 1.2.3), but with the advantage of assessing the whole exome or genome instead of targeted regions. Parallel to calling variants, copy number calls were produced with VarScan2 [158] (see supplemental code B.14). For this, the VarScan2 function `copynumber` suggests positions for copy number alterations (CNA) between germline and tumor in the joint mpileup file. Afterwards, the VarScan2 function `copyCaller` calls the copy number in these positions and assesses whether being normal (diploid), or if a deletion or a gain is found. Raw copy number positions created by VarScan2 were processed with the DNAcopy [269] package in R (see supplemental code B.15).

DNAcopy merges the copy number to larger segments. These segments were plotted with R.
Additionally, the segmentation data were used as input to GISTIC2 [201] (see supplemental code B.16). GISTIC2 identifies genes in the respective CNA and produces a score for each alteration. This score considers the amplitude and the frequency of occurrence of a CNA in the cohort of input sample data [201]. The higher the GISTIC score, the more distinct is the CNA in the analyzed cohort [201]. For comparison of AL to MM, CNA were processed in the same way from the 28 MM WES samples.

2.5 Analysis of final data sets

The following section comprises a description of the analytical methods used to assess gene expression, copy number, and variant data.

2.5.1 Dimension reduction of gene expression

Gene expression data are multidimensional: each sample contains expression values for thousands of genes, which can be interpreted as dimensions.
To receive a measure of similarity between the different entities or groups, RV coefficients [250] were calculated by the `cia` function of the made4 package [65] with gene expression data from gcrma normalized DNA microarrays (described in section 2.3). The RV coefficient is a multivariate generalization of the squared Pearson correlation coefficient, interpretable as a matrix correlation coefficient [277]. It ranges from 0 to 1, with 1 indicating the highest degree of similarity between the compared groups [64].
For better comprehensible representation, dimension reduction was performed. Two commonly applied methods were used: principal component analysis (PCA) [236] and t-distributed stochastic neighbor embedding (t-SNE) [298, 300].
By PCA, main variables are linearly combined to principal components. These principal components are ordered by the magnitude of variance they explain in the data. The highest amount of explained variance is given by the first principal component. Plotting only the first two principal components, data can be visualized in two dimensions, by keeping the maximum of information on the variance in the data.
T-SNE is a machine learning algorithm for visualizing high dimensional data. It iteratively reduces the dimensions to two vectors, which can be plotted in a scatter plot. The parameter for the number of iterations is freely adjustable. The optimization function in t-SNE contains a random initialization. With the R function `set.seed` [239], the result of random sampling is conserved by a defined starting point for the random number generator. For the assessment of the final result, it is advised to choose a re-

sult with a minimal Kullback-Leibler divergence (relative entropy measure) [299]. It can be minimized by increasing the number of iterations and by varying the perplexity parameter. By increasing the number of iterations, the Kullback-Leibler divergence decreases. The perplexity parameter is an information measure interpretable as number of nearest neighbors [299]. It is recommended by van der Maaten [299] to vary the perplexity only between $5-50$ and to raise the perplexity if more samples are included [314]. Too high perplexity values lead to a thorough mixing of the resulting visualization and too low values to false clustering.
Both methods were applied to gene expression values from gcrma normalized DNA microarrays and edgeR normalized RNA seq counts. As input, expression values per probeset/gene and per sample were used. For PCA, following the suggestion in the R stats package [239], expression data were scaled to have uniform variance before performing the analysis. In t-SNE, a perplexity value of 50 was chosen, due to large sample size in expression data. The iterations parameter was set at 5000, due to a converging Kullback-Leibler divergence. To account for continuity, the same parameters were applied to both gene expression data.

2.5.2 Differential expression analysis

Differential gene expression analysis with RNA seq data was performed with edgeR [49, 195, 251]. For this, samples were grouped by entities: BMPC, AL, MGUS, AMM, and MM. With this grouping design, the negative binomial (NB) dispersion from the normalization step, and the raw counts, a robust NB generalized linear model (glm) with the `glmQLFit` function was fitted. Input genes were filtered to a subset of 27341 genes with an expression value above one CPM normalized count in at least three samples [49].
To test for significantly differentially expressed genes (DEG), the `glmQLTest` function, which uses a quasi-likelihood F-test, was applied. Significant DEG were defined by a BH-adjusted [18] p-value of ≤ 0.05. Genes were sorted according to their log fold changes (LFC) by the `topTags` function. The LFC quantifies the ratio of a gene's median expression height between two groups. Using the `glmTreat` function, genes significantly above a chosen LFC threshold can be detected. Here, a LFC threshold of 1 was applied, equivalent to a twofold difference between the groups.
The lists of DEG were annotated with the biomaRt R package [82] and the Ensembl reference version 82 [328], of the same release as the GTF file used for annotation of raw read counts. Comparisons between the disease entities and the BMPC are listed in table 2.9

Table 2.9: Differential expression analysis: Performed comparisons between disease entities AL, MGUS, AMM, and MM, and normal plasma cells (BMPC) divided into group 1 and group 2. BMPC: bone marrow plasma cells, AL: light chain amyloidosis, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma

Comparison	**group 1**	**group 2**
BMPC *vs.* AL	BMPC	AL
BMPC *vs.* MGUS	BMPC	MGUS
BMPC *vs.* AMM	BMPC	AMM
BMPC *vs.* MM	BMPC	MM
AL *vs.* MGUS	AL	MGUS
AL *vs.* AMM	AL	AMM
AL *vs.* MM	AL	MM

2.5.3 Immunoglobulin gene expression

For analysis of Ig gene expression, a list of respective gene segments was downloaded from HGNC [5] [177, 325]. Of the 431 Ig gene segments in the list, 383 were detected as expressed in RNA seq. They are grouped to Ig HC genes ($n = 190$, $\alpha = 2$, $\delta = 1$, $\varepsilon = 3$, $\gamma = 5$, $\mu = 1$), Ig LC λ genes ($n = 92$) or Ig LC κ genes ($n = 101$). A heatmap was created from the expression values (described in section 2.4.3) of the detected Ig gene segments for the samples of 124 AL, 51 MGUS, 140 AMM, and 515 MM patients with the `heatmap.2` function of the gplots package [311]. Internally, unsupervised hierarchical clusterings were performed for samples and gene segments. As clustering method, `ward.2D` was used, which uses the agglomerative Ward's [220] minimum variance method to find compact spherical clusters. For distance measuring, the Euclidean distance was chosen. Clustering dendrograms were colored for patient samples by disease entity, and for Ig gene segments by Ig gene group. Afterwards, visually distinguishable clusters were cut according to their height in the clustering dendrogram, and subsequently analyzed. As cutting height 400 for Ig gene segments, and 200 for patient samples was applied.

2.5.4 Assessment of copy number alterations

After CNA calling, as described in section 2.4.5, calls were analyzed in threefold manner. First, the disease entities AL and MM were compared regarding CNA present in both entities. For this, odds ratios (as described in the statistics part of section 2.2) between AL and MM for each CNA present in both disease entities were calculated.
Second, for testing if CNA translate to changed gene expression, RNA seq gene expression data (see section 2.4.3) of the respective samples and genes were screened

[5] HGNC: Gene group: Immunoglobulins; Online resource: `https://www.genenames.org/cgi-bin/genegroup/download?id=348&type=branch`; Status: 2019-05-03, 10:55

for up- or downregulation. For this comparison, only genes present in the expression table of the 113 AL patient samples with median expression above 1 log transformed and normalized count across all samples were used. This threshold was applied to avoid false positive detection. To determine if gene expression is altered, one-sided Wilcoxon's rank sum tests, with alternative "greater" for genes inside a gained region and alternative "less" for genes in a deleted region, with the expression values of samples with the respective CNA *versus* samples without were performed. P-values were adjusted by BH correction [18]. The negative log10 ($-log10$) values of the adjusted p-values were used for graphical depiction in a boxplot.
Third, for all AL patient samples, CNA were compared to CA assessed by iFISH for the same cytoband. The frequency of presence of an alteration in both instances (CNA and CA) of a sample (overlap rate, see formula 2 in section 2.2) was calculated for all 113 AL patients.

2.5.5 Analysis of variants

The final variant table, created by the pipeline described in section 2.4.4, was analyzed in relation to the mutational load, indicated by the median number of variants per sample in each group. The different variant annotations, described in the variant annotation part of section 2.4.4, were analyzed for all AL patient samples.
Likewise, the two possible cases of base substitution patterns, i.e. transition (Ti) and transversion (Tv), were assessed, with the following background: purine bases adenine and guanine consist of a fused-ring structure, pyrimidine bases cytosine and thymine of a simple-ring. A base substitution can lead to an exchange of a base with the same underlying ring structure (purine to purine and pyrimidine to pyrimidine) or to an exchange with a different ring structure (purine to pyrimidine and *vice versa*). The first is called Ti, the latter Tv. Considering that there are twice as many possibilities for Tv than for Ti, one might expect more Tv, whereas in practice the opposite is the case; Ti are more frequent [152, 305, 332], likely due to steric similarities of the bases [63]. For simplification, the four possible Ti are summarized together, e.g. C>T stands for C>T and G>A, because the second Ti is complementary to the first. The same procedure is used for Tv. Here, four of the six possible Tv can be summarized in two, e.g. T>G stands for T>G and A>C. The function `titv` in the R package maftools [194] was used for estimating the so-called "Ti-Tv bias".
Assessing whether mutational load, variant type distribution, or Ti-Tv bias are entity specific, they were compared to MM variant data of the Multiple Myeloma Research Foundation's CoMMpass trial ("relating Clinical outcomes in Multiple Myeloma to Personal Assessment of Genetic Profile") [61, 218]. The current CoMMpass release,

IA13, includes somatic, coding, and non-synonymous variants detected in 930 newly diagnosed MM patients (see table 2.1). The variants of the CoMMpass cohort were detected in alignments against the hg19 reference genome that is a previous version to the hg38 and contains less information. They differ in the exact positioning of variants and genes, caused by adjustments from the inclusion of new information. Hence, a direct comparison of the variants by position is impossible and only mutated genes regarding gene name and annotated variants can be compared. For this comparison, only previously untreated patients of the CoMMpass cohort were used to avoid detection of mutations potentially induced by treatment.

Further comparisons of the variant table of AL were made by screening for occurrence of variants in 63 potential myeloma driver genes suggested by Walker *et al.* [308], and for overlap with variant tables from small previously published studies [36, 229, 254] including AL patient samples.

The variant table contains annotations for known genetic variation, i.e. the variant had previously been detected and analyzed. The annotation was performed by the Ensembl vep tool in the variant calling pipeline (see the variant annotation part in section 2.4.4). For these variants, unique and stable identifier exists, e.g. from dbSNP [275] and COSMIC [100]. Information regarding the variation represented by these identifiers were retrieved from dbSNP[6] and COSMIC[7] and subsequently analyzed for selected genes (i.e. NRAS, KRAS, and BRAF).

2.5.6 Functional enrichment analysis

Functional enrichment analysis (FEA) groups genes to clusters sharing properties that may lead to association e.g. with disease pathogenesis. The analyses were performed with metascape [8] [337] (see supplementary table A.1). This tool uses a variety of databases for annotation of genes with names, translation to protein names, molecular functions, gene ontologies, and pathways, and performs a FEA. In the following, database entries are referred as "terms". By metascape, all input genes are annotated to Entrez Gene ID (ENTREZID) for subsequent annotation and analysis. This reduces the number of input genes if no ENTREZID is available or multiple genes map the same ENTREZID [337].

Functional annotation terms, used in the FEA are from gene ontology (GO) [11, 290]

[6] dbSNP (build 153): Online resource: `https://www.ncbi.nlm.nih.gov/snp/`; Status: 2019-10-09, 09:31

[7] COSMIC (Release v90): Online resource: `https://cancer.sanger.ac.uk/cosmic`; Status: 2019-10-09, 09:45

[8] metascape: functional enrichment analysis; Online resource: `http://metascape.org/gp/index.html#/main/step1`; Status: 2019-09-13, 16:33

biological process (BP) with 4436 gene sets, and from the molecular signatures database (MSigDB) [183, 184, 209] subsets Canonical Pathways [286] and Hallmark [183]. Canonical Pathways includes 1329 curated gene sets from pathway databases like KEGG [146–148], Reactome [94] and CORUM [255]. Furthermore, kinase class by UniProt [247], subcellular location and protein function by Protein Atlas [135] were annotated. A detailed table of used databases and versions is depicted in supplementary table A.2.

In enrichment analysis, a list of genes is tested against a background of genes by a hypergeometric test [326, 337]. For FEA, all genes in the genome (i.e. GRCh38) were used as enrichment background. The then calculated enrichment factor is the ratio of hits for a term against the hits expected by chance [337]. Test results were adjusted for multiple testing by BH correction [18, 337]. Enriched terms with a p-value < 0.01, a minimum of three hits, and an enrichment factor > 1.5 were grouped into clusters [337]. These clusters contain all terms that can be grouped together by a parent term [337]. The term within a cluster with the smallest p-value represents the cluster [337]. FEA were performed for four sets of gene names from different analyses: First, the overlap of genes found differentially expressed in RNA seq between BMPC *versus* AL and BMPC *versus* MM (see supplementary table A.15). Second, four gene lists with all DEG between BMPC *versus* AL, BMPC *versus* MGUS, BMPC *versus* AMM and BMPC *versus* MM (for a detailed description see section 2.5.2). Third, the 59 prognostic genes of the HDAL (as described in section 2.3.3 and supplementary table A.12). Fourth, all genes, excluding Ig genes, for which at least one somatic, non-synonymous, and expressed variant was detected in AL by WES (outlined in section 2.4.4). For comparison, variant table of MM from the CoMMpass cohort were analyzed. Here, the gene list was reduced to genes with at least three somatic, non-synonymous, and expressed variants (for a description of the CoMMpass cohort, see section 2.5.5).

Terms were sorted by p-value in decreasing order, depicting the top 20 terms. Significantly enriched terms were depicted in a heatmap for sets of gene lists. For heatmaps, unsupervised hierarchical clustering was performed with Kappa scores [59] as similarity metric. For graphical depiction, the negative log10 values ($-log10$) of the p-values were used.

3 Results

This chapter comprises: First results regarding WES measurements, i.e. quality control information derived during variant calling, and variant type analyses (see section 3.1). Second, the evaluation of individual GEP-based risk assessments in MM (see section 3.2) for relation and subsequent analysis in AL. And third, analyses regarding clinical parameters, chromosomal aberrations assessed by iFISH, gene expression assessed by DNA microarrays and RNA seq, and mutational characteristics assessed by WES that focus on prognosis and pathophysiology of AL (see sections 3.3, 3.4, 3.5, and 3.6).

3.1 Whole exome sequencing

This section first summarizes quality measurements. Subsequently, an analysis of the variant calling process for 113 AL patient samples is presented. Then, variants from the CoMMpass cohort of 930 MM patients are described. Analysis of the detected variant types is outlined at the end of the section.

3.1.1 Quality control

Table 3.1: Quality control (QC) results of FASTQ files and alignments by five tools as described in section 2.4.4 and supplementary table A.1.

Measurement	**Value**	**Tool**
Per base sequence quality	mean 35.3 (sd 0.3)	Fastqc
Minimal sequence quality score	33	Fastqc
Per base sequence content	Unusual, strong bias in base 1-15	Fastqc
Overrepresented sequences	None	Fastqc
Per base N content	Low	Fastqc
Read length	151	Fastqc
GC content	48%-50%	Fastqc
Adapter Sequence detected	"Nextera Transposase Sequence"	Fastqc
Trimmed reads with adapter sequences	mean 11,316,561 reads	fastp
Trimmed bases in reads with adapter sequences	mean 286,718,792 bases	fastp
Filtered reads per FASTQ file	mean 1.4%	fastp
Sequence duplication per FASTQ	33.6% - 69.4%	Fastqc
Sequence duplication per FASTQ pair	3.1% - 10.3%	fastp
Sequence duplication per alignment	5.2% - 18.6%	Picard
Percentage of unmapped reads	< 0.1%	samtools
Insert size per FASTQ pair	mean 193.7 bases (range 157 - 220)	fastp
Insert size per alignment	median 219 bases (range 179 - 260)	Alfred
Covered coding genes in GRCh38	mean 52.7% (range 48.8% - 55.2%)	Alfred

Quality control was performed regarding raw FASTQ files by Fastqc and fastp, and regarding alignments by samtools, Picard, and Alfred at different steps of the WES pipeline. An overview of the results is summarized in table 3.1.
The "per base sequence content" was reported as "unusual" by Fastqc in all FASTQ files. The reason for this was a bias in the base distribution from the start of all reads up to base 15. Read trimming with fastp eliminated this bias. In all FASTQ files, "Nextera Transposase Sequence" adapter were detected at the end of the reads. They originate from library preparation and should in principle already be removed by the Illumina software. In read pairs with insert size below 151 bases (because of too short tags during sequence tagmentation, see section 1.6.3) they were not recognized by the Illumina software. These adapters were trimmed with fastp as well. Fastp removed only a low percentage of complete reads (mean:1.4%, see table 3.1) per FASTQ file and all due to insufficient quality. All FASTQ files have a similar "GC content distribution". With a mean of 49.6% (48% – 50%), it was detected as "abnormal" to the theoretical distribution, i.e. in a normal random library a nearly normal distribution would be expected. The theoretical distribution is this normal distribution with the peak corresponding to the overall GC content of the underlying data, see [9]. The "Sequence duplication level" was detected as "slightly abnormal" by Fastqc with a mean of 52.5% and ranges from 34% – 69%. This was further monitored by fastp and Picard, which both detect low (usual) levels of duplicated sequences (3% – 10% and 5% – 18%).
No file needed to be removed from further analysis or had to be resequenced due to quality issues.

3.1.2 Variant calling and filtering

Variants were called with three different variant callers (see section 2.4.4). A total number of 1388995 potential variants in 113 AL patient samples were detected. Each variant either being of type SNV or InDel. VarScan2 called 544132 variants, Strelka called 735920 and Seurat called 837629 possible variants. In figure 3.1, the overlap of called variants is depicted. Half of the variants detected by Seurat or Strelka were discarded, as only detected by one caller. In sum, 446929 variants were called by two or more callers, i.e. 32% of all called variants. Table 3.2 lists the respective variant statistics.
The first step, applying VarScan2 `fpfilter` (see the respective part on variant filtering in section 2.4.4), reduces the number of variants from 446929 to 253642 (57%, see table 3.2). After annotation, 56216 variants were detected as located in the coding part of the exome and of these, 30051 code for non-synonymous mutations. Among these

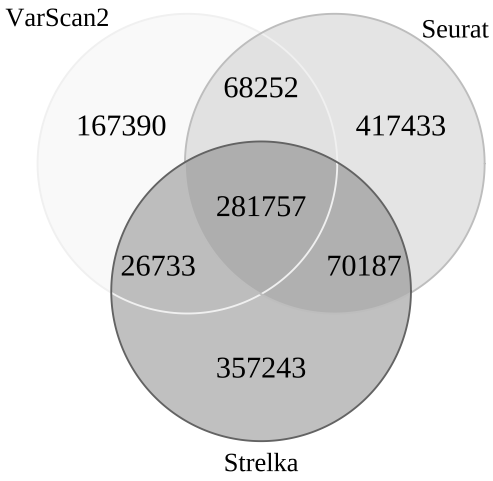

Figure 3.1: Numbers of all called variants by overlap of variant callers VarScan2, Seurat, and Strelka.

are 370 variants in 118 Ig genes of the constant, variable, diversity and joining region for HC and for LC, which were analyzed separately.

Table 3.2: Summary statistics of called variants in light chain amyloidosis (AL). The second column depicts the total number of variants in the group, the following columns contain summary statistics of variants. The first row comprises all jointly called variants in 113 AL samples, each following row contains values for a subset of the variants, described in the first column and related to the previous row. 25% Q value of the 25% percentile (first quartile), 75% Q value of the 75% percentile (third quartile).

Variants	All	Mean	Median	Min	25% Q	75% Q	Max
all	1388995	12292	8277	2776	6400	10551	94542
passed filter	253642	2244.6	384	115	298	512	34033
coding	56216	497.5	33	8	25	50	7190
non-synonymous	30051	265.9	25	3	19	39	3791
without Ig	29681	262.7	22	1	16	36	3785
expressed	4552	48.4	4.5	1	3	8.75	1100

A total number of 9842 mutated genes is detected with 29681 variants. Of these, 4552 variants were likewise detected as expressed in RNA seq in a total of 2909 genes.

3.1.3 Multiple myeloma variants for comparison

The variants and their frequency detected in AL were compared to variants detected in 930 MM patient samples (from the CoMMpass cohort, described in section 2.5.5). The median number of variants in MM is 43 ranging from $1 - 1963$ variants per sample. A summary on the variants is given in table 3.3.
To compare the CNA detected in AL to MM, WES samples of 28 MM patients were assessed in the same way as the AL patient samples at the LfM.

Table 3.3: Summary statistics of variants in multiple myeloma (MM). The second column depicts the total number of variants in the 930 samples, the following columns summary statistics of variants. The first row comprises all non-synonymous variants in CoMMpass cohort variant table version IA13, every following row contains values for a subset of variants, described in the first column, to the previous row. 25% Q value of the 25% percentile (first quartile), 75% Q value of the 75% percentile (third quartile)

Variants	All	Mean	Median	Min	25% Q	75% Q	Max
non-synonymous	64954	69.8	56	2	45	71	1980
without Ig	52956	56.9	43	2	33	56	1963
expressed	18442	25.3	19	1	13	56	985

3.1.4 Variant types

The proportions of the detected variants and CNA divided by variant types for AL and MM are depicted in table 3.4 **a** and **b**. The majority of variants (97%) are SNV appearing in both AL and MM (cf table 3.4 **a**). A difference in frequency of SNV between ALMG and ALMM is not detectable. Of the genes detected as altered in copy number, 87% genes are gained in AL and 82% in MM (cf table 3.4 **b**). Same as for SNV and InDel, a difference in frequency of CNA type between ALMG and ALMM is marginal.

Table 3.4: Number of variants by variant type. **a** Number of single nucleotide variants (SNV), short deletions (Del) and insertions (Ins) in 113 light chain amyloidosis (AL) and 930 multiple myeloma (MM) patients detected by variant calling. **b** Number of genes affected by copy number alterations (CNA) separated as gains (Gain) and deletions (Del) in 113 AL and 28 MM. ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, *n*: number

a

Vartype	AL	ALMG	ALMM	MM
n	29681	19276	10405	64954
SNV	28880 (97.3%)	18757 (97.3%)	10123 (97.3%)	62828 (96.7%)
Del	477 (1.6%)	307 (1.6%)	170 (1.6%)	1370 (2.1%)
Ins	324 (1.1%)	212 (1.1%)	112 (1.1%)	756 (1.2%)

b

CNA	AL	ALMG	ALMM	MM
n	109765	44504	65261	1671
Gain	95812 (87.3%)	38356 (86.2%)	57456 (88%)	1375 (82.3%)
Del	13953 (12.7%)	6148 (13.8%)	7805 (12%)	296 (17.7%)

3.2 Individual gene expression-based risk assessment in multiple myeloma

In this section, results from the GEP-R framework assessing individual risk and biological entity in MM within the randomized GMMG-MM5 phase III trial are presented for relation to AL regarding the two main questions: "Are gene expression-based risk assessments estimating malignant plasma cell properties in MM as good as the current

standard risk stratifications?" and "Is a personalized therapeutic recommendation possible by assessing the expression of target genes?" Parts of the analyses have already been published in the article Hose, D., Beck, S. *et al.* [126].

From 573 (95%) of the 604 patients, included in the GMMG-MM5-trial, bone marrow aspirates were available and for 559 (97.6%) of these patients and malignant plasma cells could be purified. GEP by DNA microarray was feasible in 456 (81.9%) patients and CA could be assessed by iFISH in 99.5% of these 559 patients. All samples with GEP passed quality control of the GEP-R [126].

Table 3.5: Univariate Cox regression analyses regarding progression free survival (PFS) of patients classified as multiple myeloma (MM) with gene expression profiling (GEP) by DNA microarray included in the GMMG-MM5-trial ($n = 456$, adapted from Hose, D., Beck, S. *et al.* [126]). Events: number of events per score grouping level, HR: hazard ratio, CI: 95% confidence interval for HR, p: p-value of Wald test, Median time: median survival time in months, Survival rate: percentage of patients not progressing after 2 and 5 years for the respective grouping level, n.r.: not reached.

						Median	**Survival rate [%]**	
Group	**Level**	**Events**	**HR**	**CI**	***p***	**time**	**2-year**	**5-year**
GPI	low	117				49	80	42
	medium	136	1.78	1.4-2.3	< .001	33	60	23
	high	31	3.59	2.4-5.4	< .001	18	36	9
UAMS70	low risk	197				43	74	36
	high risk	87	1.91	1.5-2.5	< .001	23	50	17
IFM15	low risk	199				44	73	37
	high risk	85	1.96	1.5-2.5	< .001	25	52	15
predicted	0	245				40	70	33
t(4;14)	1	39	1.58	1.1-2.2	.008	26	53	17
HM	low	22				n.r.	93	55
metascore	medium	222	2.20	1.4-3.4	< .001	39	68	31
	high	40	5.89	3.5-10	< .001	15	35	5
ISS	I	123				48	82	41
	II	126	1.43	1.1-1.8	.005	39	68	32
	III	117	1.90	1.5-2.4	< .001	26	53	24
rISS	I	76				54	84	44
	II	200	1.65	1.3-2.2	< .001	37	67	30
	III	56	2.66	1.9-3.8	< .001	22	47	18

The GPI is predictive for the 456 patients for OS ($p < 0.001$) and PFS ($p < 0.001$), 229/191/36 patients were classified as low/medium/high proliferative by the GPI (see figure 3.2). A median PFS of 49/33/18 months and for OS of not reached (n.r.)/74/33 was found for low/medium/high GPI (see table 3.5, 3.6) [126].

UAM70 and IFM15 are predictive for OS and PFS (all $p < 0.001$) and showed comparable results, classifying 343 and 350 patients as high risk (see figure 3.3). Survival rates for both scores (UAMS70/IFM15) are equal with 26%/27% of low risk,

Table 3.6: Univariate Cox regression analyses regarding overall survival (OS) of patients classified as multiple myeloma (MM) with gene expression profiling (GEP) by DNA microarray included in the GMMG-MM5-trial ($n = 456$, adapted from Hose, D., Beck, S. *et al.* [126]). Events: number of events per score grouping level, HR: hazard ratio, CI: 95% confidence interval for HR, p: p-value of Wald test, Median time: median survival time in months, Survival rate: percentage of patients being alive after 2 and 5 years for the respective grouping level, n.r.: not reached.

Group	Level	Events	HR	CI	p	Median time	Survival rate [%]	
							2-year	5-year
GPI	low	38				n.r.	94	83
	medium	79	2.71	1.8-4	< .001	74	86	57
	high	22	6.13	3.6-10.4	< .001	33	59	33
UAMS70	low risk	81				n.r.	92	74
	high risk	58	2.67	1.9-3.7	< .001	54	77	48
IFM15	low risk	87				n.r.	92	73
	high risk	52	2.46	1.7-3.5	< .001	58	76	50
predicted t(4;14)	0	113				n.r.	88	70
	1	26	1.89	1.2-2.9	.004	56	85	49
HM metascore	low	2				n.r.	98	98
	medium	106	9.88	2.4-40	.001	n.r.	89	68
	high	31	35.55	8.5-148.8	< .001	41	64	25
ISS	I	42				n.r.	94	81
	II	61	1.86	1.3-2.8	.002	n.r.	89	67
	III	75	3.25	2.2-4.8	< .001	62	78	51
rISS	I	22				n.r.	95	86
	II	99	2.50	1.6-4	< .001	n.r.	87	65
	III	41	5.84	3.5-9.8	< .001	42	75	40

Table 3.7: Integrated Brier score for GEP-R assessed scores regarding progression free survival (PFS) and overall survival (OS) of patients classified as multiple myeloma (MM) with gene expression profiling (GEP) by DNA microarray included in the GMMG-MM5-trial ($n = 456$, adapted from Hose, D., Beck, S. *et al.* [126]). Score: Integrated Brier score, p: p-value.

Group	Brier PFS		Brier OS	
	Score	p	Score	p
GPI	0.189	.08	0.135	.008
UAMS70	0.191	.06	0.139	.02
IFM15	0.192	.12	0.140	.058
predicted t(414)	0.196	.41	0.145	.12
HM metascore	0.186	.02	0.132	< .001
ISS	0.187	.03	0.137	.009
rISS	0.186	.02	0.137	.003

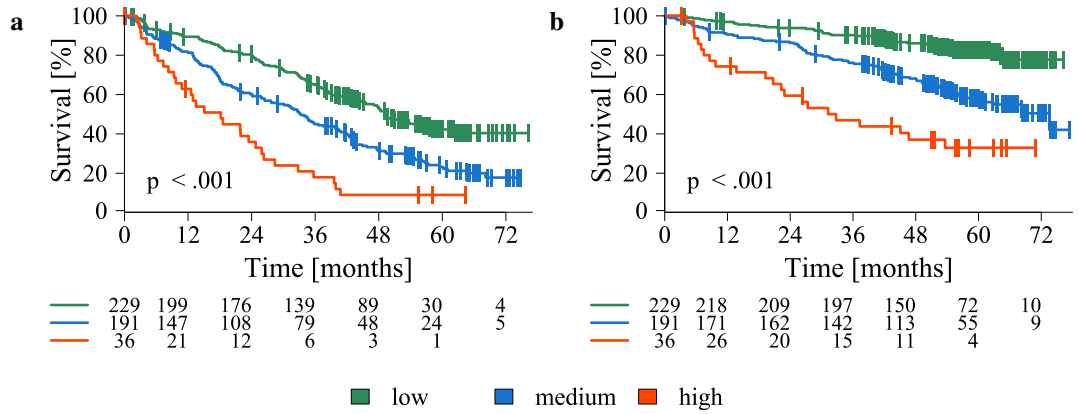

Figure 3.2: Survival analyses of patients classified as multiple myeloma (MM) with gene expression profiling (GEP) by DNA microarray included in the GMMG-MM5-trial ($n = 456$) grouped by: **a** progression free survival (PFS) for GPI **b** overall survival (OS) for GPI. Difference between curves was tested using Log-rank test and was termed significant if the p-value was ≤ 0.05. (adapted from Hose, D., Beck, S. *et al.* [126])

50%/48% of high risk patients progressing after 2 years (see table 3.5) and 92% of low risk, 77%/76% of patients being alive after 2 years (see 3.6) [126].
A t(4;14) is predicted from gene expression for 53 patients. In 5 patients, the prediction does not correspond to the iFISH results. Predicted presence of t(4;14) is significantly associated with adverse PFS ($p = 0.008$) and OS ($p = 0.003$), see figure 3.4 [126].
The HM metascore, as a combination of GEP-based risk assessments and conventional serum-based risk stratifications, is predictive for PFS and OS (both $p < 0.001$), and classifies 58/352/46 as low/medium/high risk (see figure 3.5). The HM metascore has the smallest Brier score for PFS (0.186) and OS (0.132), indicating the best prediction accuracy of all stratification methods (see table 3.7) [126].
For comparison to clinical standard stratification ISS and rISS were assessed. Both are predictive for PFS and OS (all $p < 0.001$, see figure 3.6), with median time to OS of 48/39/26 and 54/37/22 months for the stages I/II/III of ISS and rISS (see table 3.6). The Brier scores of ISS and rISS are 0.187 and 0.186 in PFS and 0.137 for both in OS. They are larger than the Brier scores for the HM metascore (see table 3.7) [126].
The GEP-R provides an assessment of expression of defined target genes (for a list of genes, see section 1.2.5). In the analyzed cohort of 456 MM patients, 197 express AURKA, 151 IGF1R, and 50 FGFR3, respectively [126].

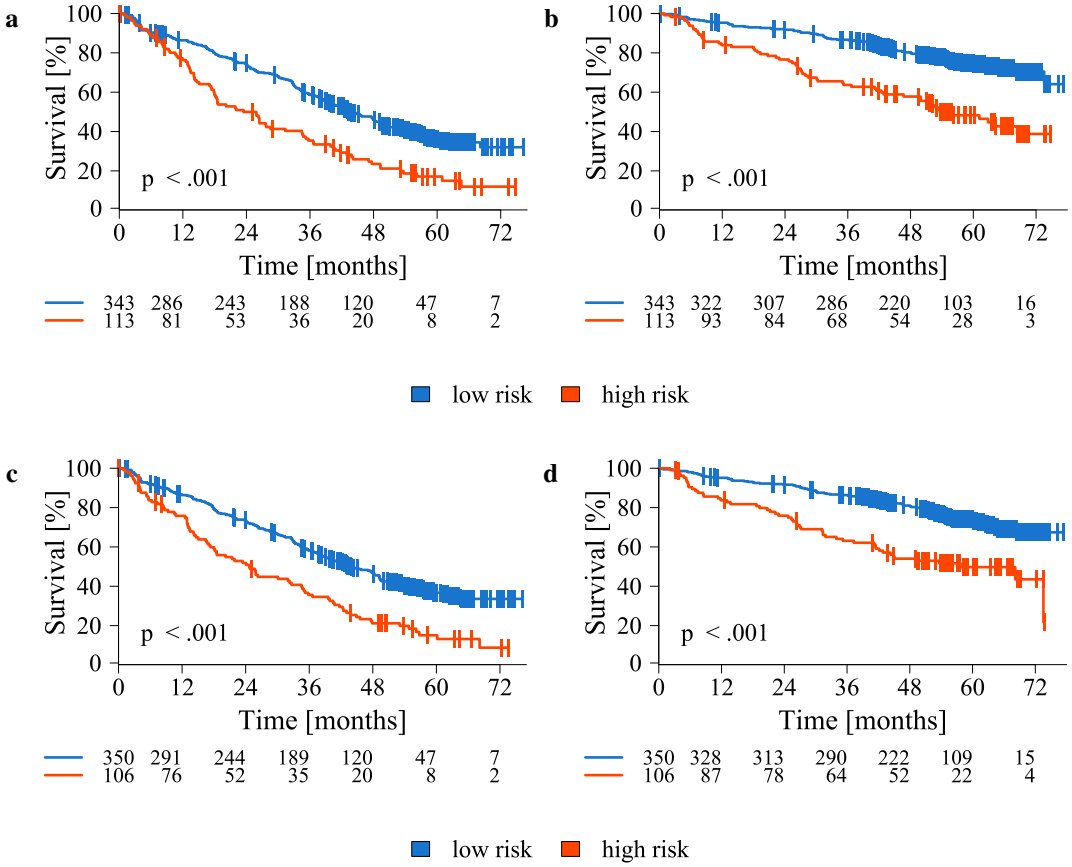

Figure 3.3: Survival analyses of patients classified as multiple myeloma (MM) with gene expression profiling (GEP) by DNA microarray included in the GMMG-MM5-trial ($n = 456$) grouped by: **a** progression free survival (PFS) for UAM70 **b** overall survival (OS) for UAMS70 **c** PFS for IFM15 **d** OS for IFM15. Difference between curves was tested using Log-rank test and was termed significant if the p-value was ≤ 0.05. (adapted from Hose, D., Beck, S. *et al.* [126])

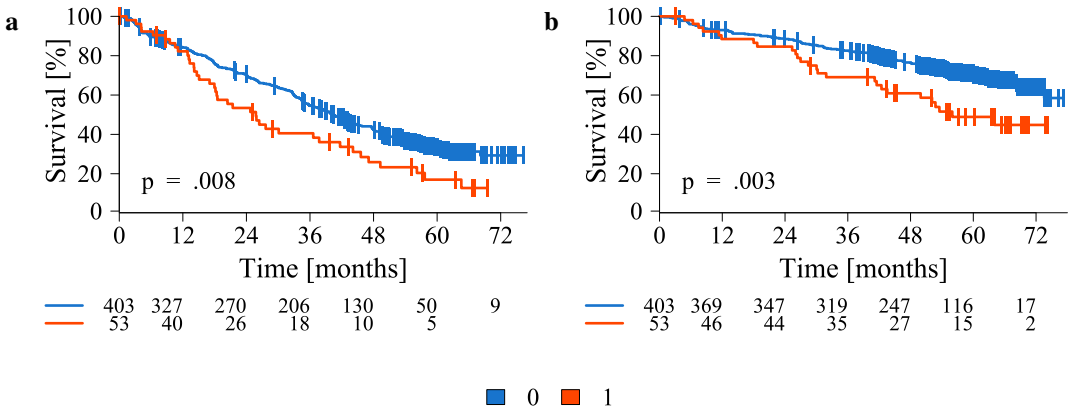

Figure 3.4: Survival analyses of patients classified as multiple myeloma (MM) with gene expression profiling (GEP) by DNA microarray included in the GMMG-MM5-trial ($n = 456$) grouped by: **a** progression free survival (PFS) for predicted t(4;14) (1) *versus* no predicted t(4;14) (0), **b** overall survival (OS) for predicted t(4;14) (1) *versus* no predicted t(4;14) (0). Difference between curves was tested using Log-rank test and was termed significant if the p-value was ≤ 0.05. (adapted from Hose, D., Beck, S. *et al.* [126])

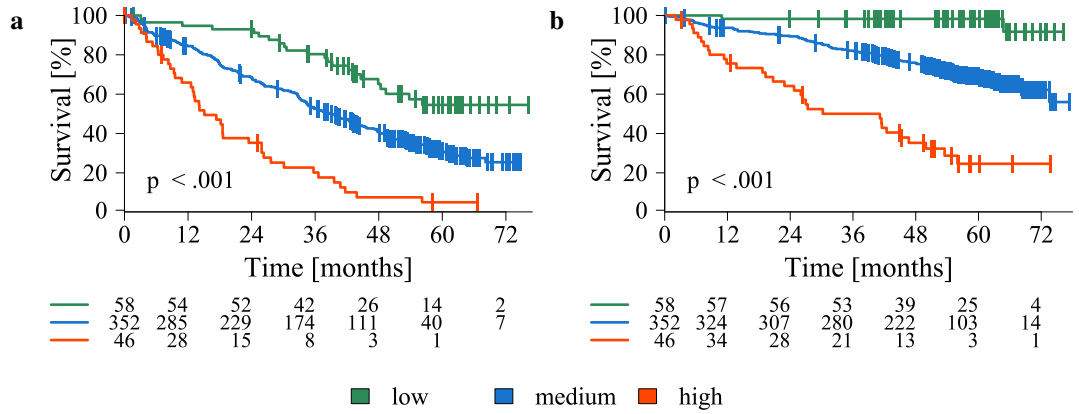

Figure 3.5: Survival analyses of patients classified as multiple myeloma (MM) with gene expression profiling (GEP) by DNA microarray included in the GMMG-MM5-trial ($n = 456$) grouped by: **a** progression free survival (PFS) for HM metascore **b** overall survival (OS) for HM metascore. Difference between curves was tested using Log-rank test and was termed significant if the p-value was ≤ 0.05. (adapted from Hose, D., Beck, S. *et al.* [126])

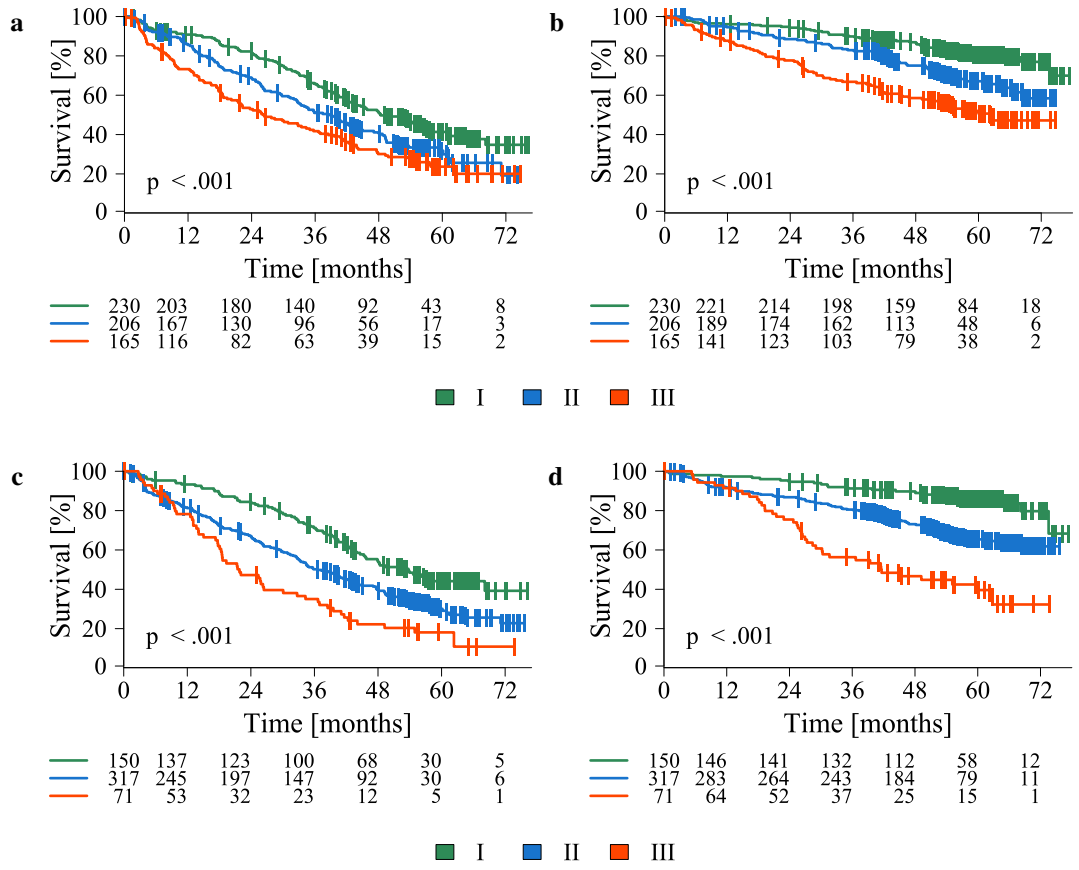

Figure 3.6: Survival analyses of patients classified as multiple myeloma (MM) with gene expression profiling by DNA microarray included in the GMMG-MM5-trial ($n = 456$) grouped by: **a** progression free survival (PFS) for ISS **b** overall survival (OS) for ISS **c** PFS for rISS **d** OS for rISS. Difference between curves was tested using Log-rank test and was termed significant if the p-value was ≤ 0.05. (adapted from Hose, D., Beck, S. *et al.* [126])

3.3 Prognostic role of malignant plasma cell characteristics *versus* amyloidogenicity

In the following section the third aim "What role play malignant plasma cell characteristics *versus* properties associated with amyloid light chain formation and deposition (amyloidogenicity)" is addressed. OS analyses were performed for AL patients regarding clinical parameters, including organ involvement, serum parameters, staging systems, and tumor load. Second, regarding malignant plasma cell characteristics CA detected by iFISH.

3.3.1 Amyloidogenicity

Prognostic impact regarding involvement of different organs is as expected, most severely for heart *versus* kidney involvement ($p < 0.001$, see figure 3.7). NT-ProBNP and cTnT levels, split by published thresholds (see section 1.4) are predictive as expected (see figure 3.8).

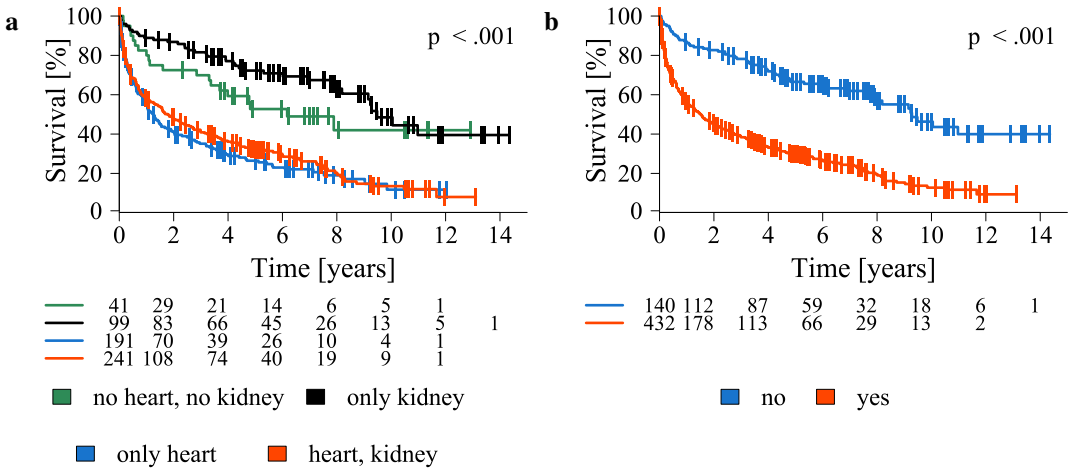

Figure 3.7: Overall survival (OS) of patients classified as light chain amyloidosis (AL) grouped by: **a** heart and kidney involvement **b** heart involvement. Difference between curves was tested using Log-rank test and was termed significant if the p-value was ≤ 0.05. The legend beyond each plot indicates the different delineated levels.

All three different staging methods assessed, i.e. the standard Mayo Score (2004), the advanced Mayo Stage III Euro Score (2013), and the revised Mayo Score (2012) (see section 1.4), all based on serum parameters, are predictive for OS (all $p < 0.001$), see figure 3.9.

Co-occurrence of involved organs is frequently seen, as depicted in figure 3.10. Only heart and kidney involvement frequently occur as single involved organ. These are the two organs that are also to the largest proportion associated with diff FLC (see table 3.8).

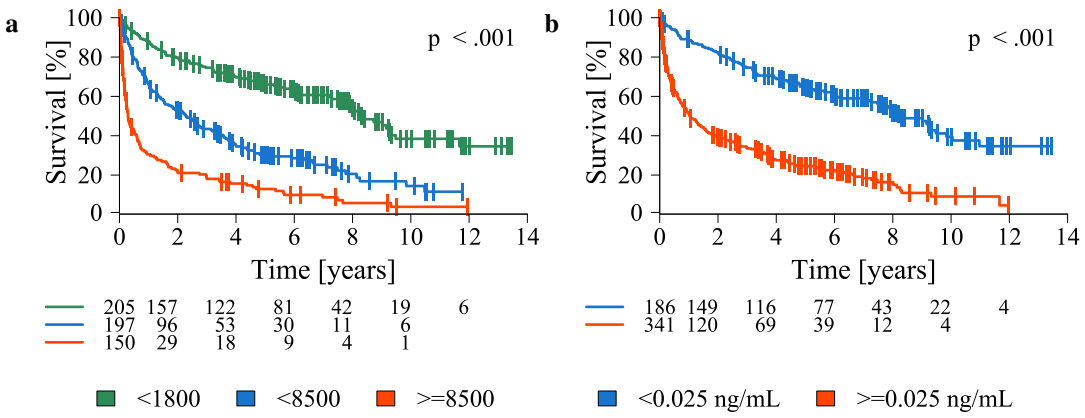

Figure 3.8: Overall survival (OS) of patients classified as light chain amyloidosis (AL) grouped by: **a** N-terminal pro-brain natriuretic peptide type-B (NT-ProBNP) **b** cardiac troponin T (cTnT). Difference between curves was tested using Log-rank test and was termed significant if the p-value was ≤ 0.05. The legend beyond each plot indicates the different delineated levels.

Table 3.8: Involved organs and measurements of diff FLC (difference in free light chains) in patients classified as light chain amyloidosis.

		<50 mg/L		**≥50 mg/L**		**>180 mg/L**	
Variable	**Level**	***n***	***%***	***n***	***%***	***n***	***%***
Involved Organ	no heart, no kidney	6	8.8	8	5.7	18	6.2
	only kidney	35	51.5	26	18.6	24	8.2
	only heart	4	5.9	34	24.3	129	44.2
	heart, kidney	23	33.8	72	51.4	121	41.4
Number of involved organs	1	27	39.7	24	17.1	50	17.1
	2-4	34	50.0	104	74.3	216	73.7
	>4	7	10.3	12	8.6	27	9.2

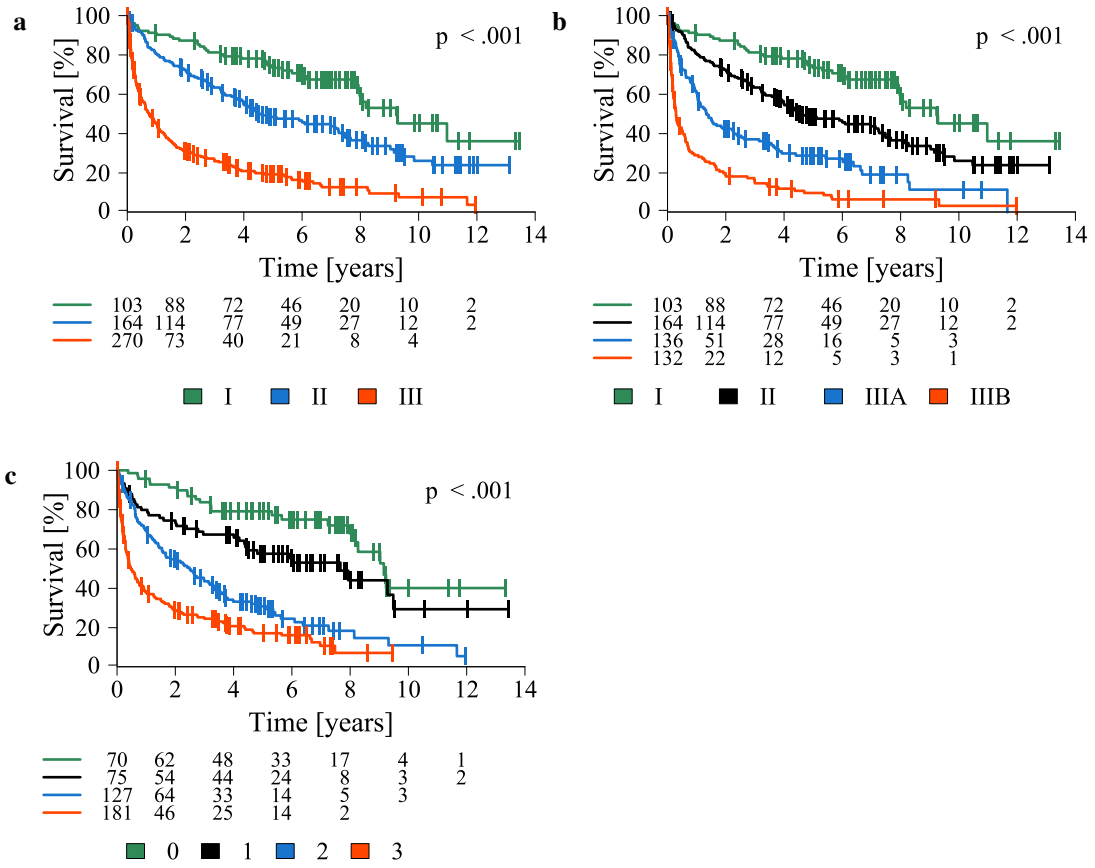

Figure 3.9: Overall survival (OS) of patients classified as light chain amyloidosis (AL) grouped by: **a** standard Mayo Score (2004) **b** advanced Mayo Stage III Euro Score (2013) **c** revised Mayo Score (2012). Difference between curves was tested using Log-rank test and was termed significant if the p-value was ≤ 0.05. The legend beyond each plot indicates the different delineated levels.

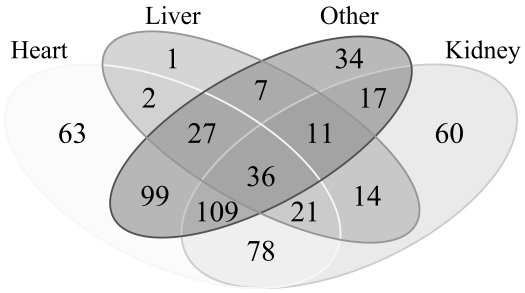

Figure 3.10: Co-occurrence of involved organs, heart, kidney, liver, and others (soft tissue and gastrointestinal tract) in patients classified as light chain amyloidosis.

3.3.2 Tumor load

Surrogated measures for tumor mass, i.e. PCI ($p = 0.002$), diff FLC ($p < 0.001$), and M-protein ($p = 0.01$) are prognostic for OS of AL patients (see figure 3.11).

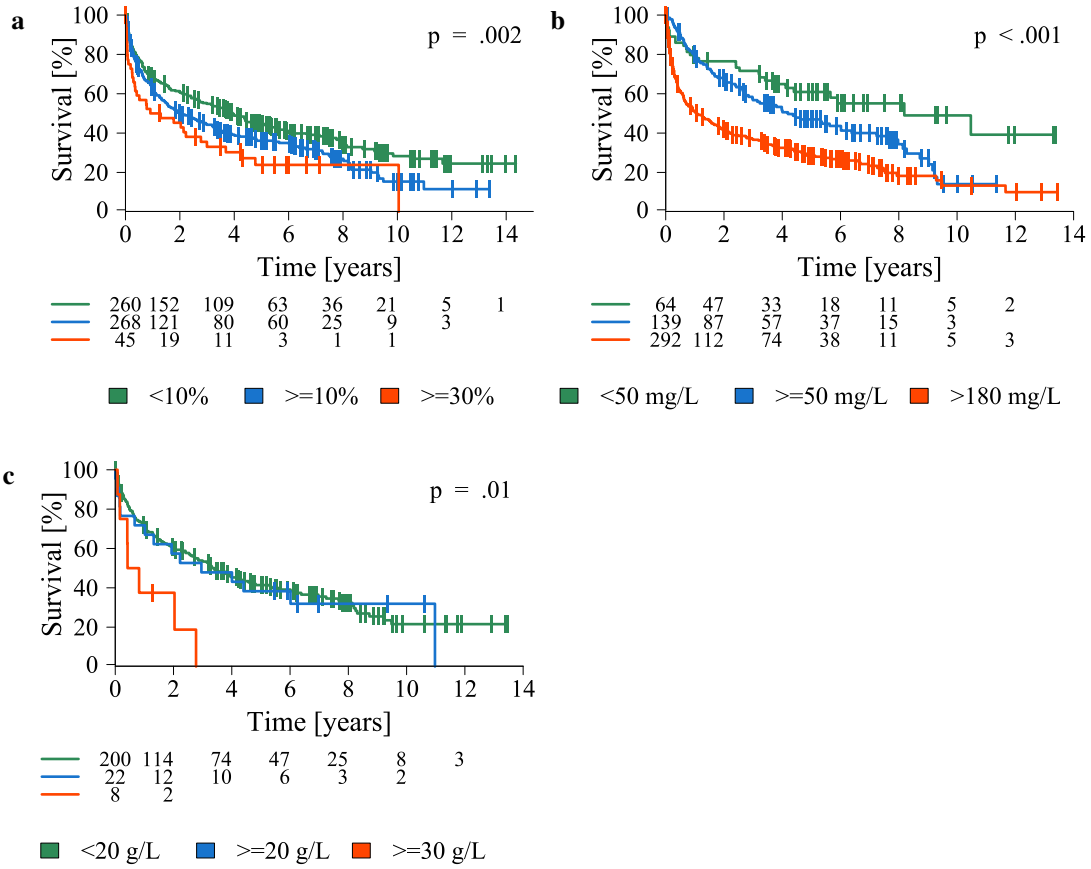

Figure 3.11: Overall survival (OS) of patients classified as light chain amyloidosis (AL) grouped by: **a** plasma cell infiltration (PCI) **b** difference between involved and uninvolved free light chains (diff FLC) **c** monoclonal protein (M-protein). Difference between curves was tested using Log-rank test and was termed significant if the p-value was ≤ 0.05. The legend beyond each plot indicates the different delineated levels.

3.3.3 Chromosomal aberrations

CA were analyzed individually regarding impact on OS. In the assessed cohort, comprising patients under different treatment regimen, none of the CA alone is associated with survival (see figures 3.12, 3.13).

A higher proliferation rate, as assessed by the GPI, is significantly associated with gain 1q21 in MM ($p < 0.001$) but not in AL (see figure 3.14).

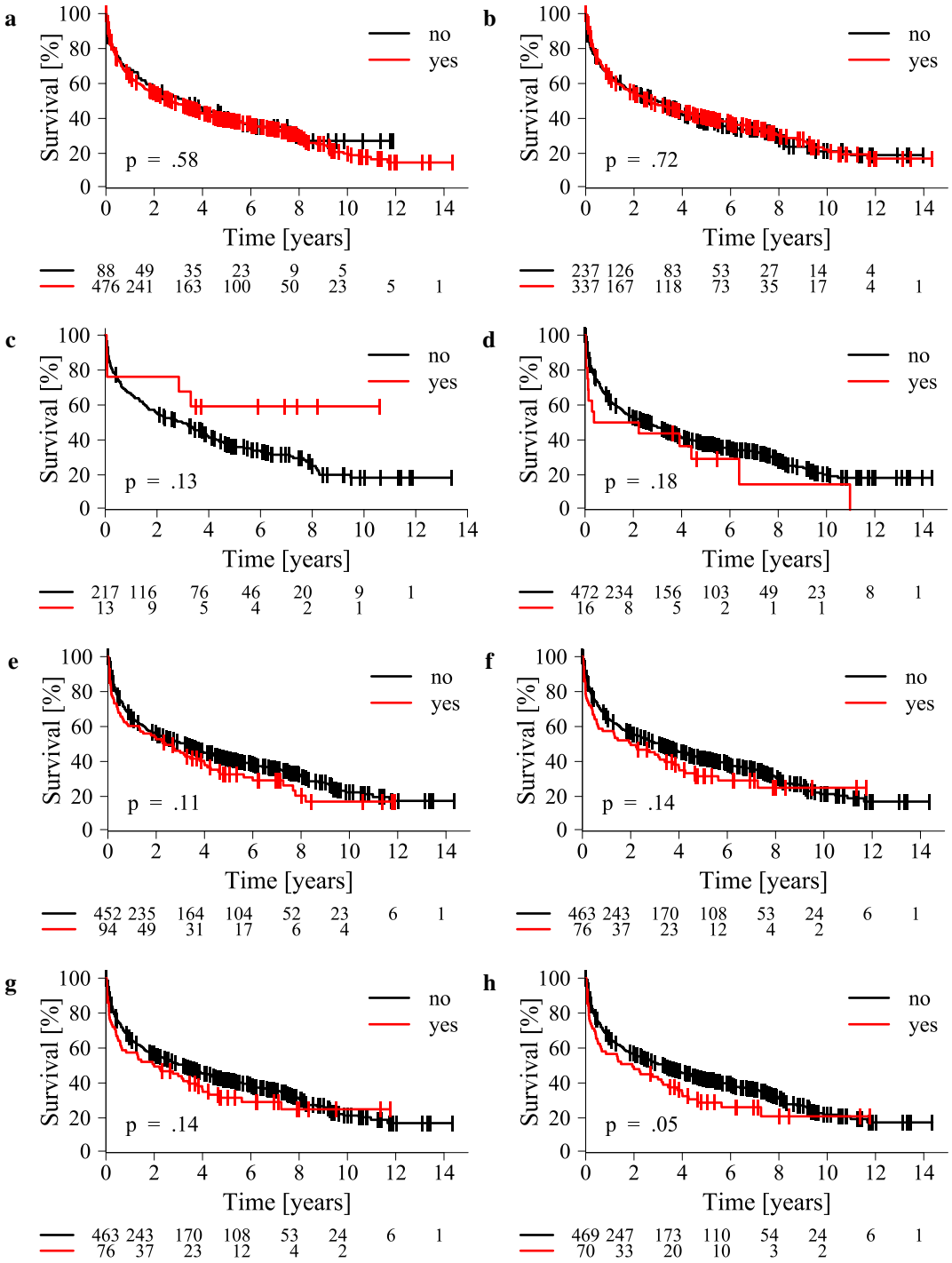

Figure 3.12: Overall survival (OS) of patients classified as light chain amyloidosis (AL) grouped by: **a** IgH-rearrangement (presence/absence) **b** t(11;14) (presence/absence) **c** t(14;16) (presence/absence) **d** t(4;14) (presence/absence) **e** hyperdiploidy (presence/absence) **f** 5q31/5q35 (gain *versus* no gain) **g** 5p15 (gain *versus* no gain) **h** 15q22 (gain *versus* no gain). Difference between curves was tested using Log-rank test and was termed significant if the p-value was ≤ 0.05.

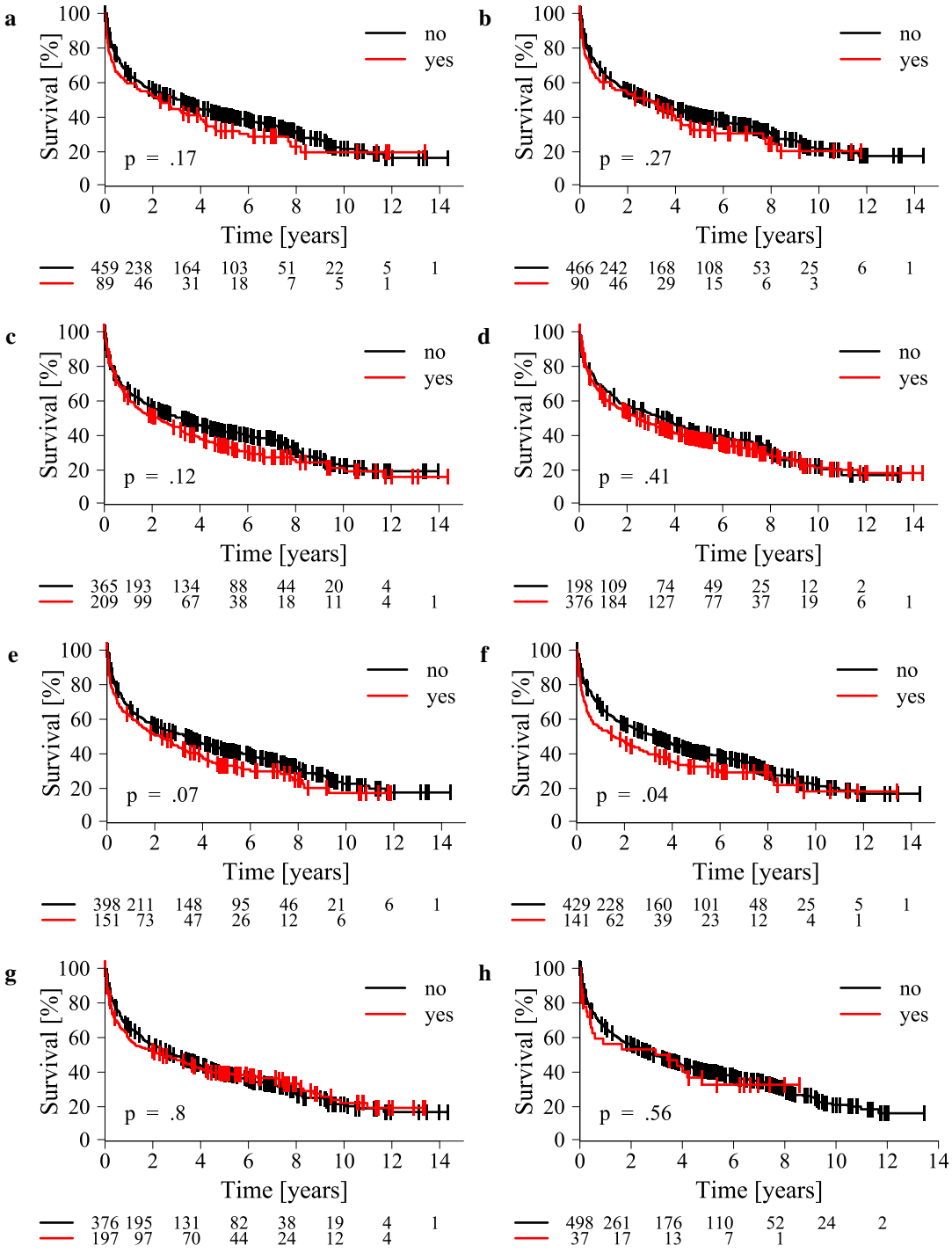

Figure 3.13: Overall survival (OS) of patients classified as light chain amyloidosis (AL) grouped by: **a** 19q13 (gain *versus* no gain) **b** 11q22/11q23 (gain *versus* no gain) **c** 11q13 (gain *versus* no gain) **d** 9q34 (gain *versus* no gain) **e** 1q21 (gain *versus* no gain) **f** 13q14 (deletion *versus* no deletion) **g** 8p21 (deletion *versus* no deletion) **h** 17p13 (deletion *versus* no deletion). Difference between curves was tested using Log-rank test and was termed significant if the p-value was ≤ 0.05.

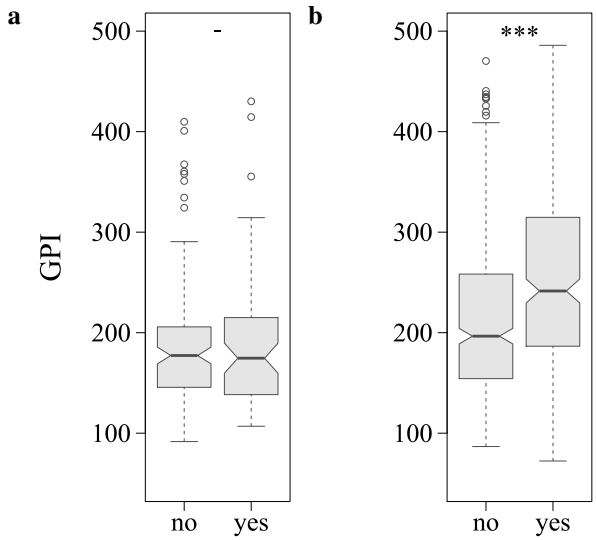

Figure 3.14: GPI by presence (yes) or no presence (no) of gain 1q21 in **a** AL (n=196) and **b** MM (n=765). Significant differences, indicated by Wilcoxon's rank sum test, are illustrated as ***, representing a significant p-value < 0.001, - indicates no significant difference. AL: light chain amyloidosis, MM: multiple myeloma

3.4 Gene expression-based assessment of biological variables and risk in light chain amyloidosis

In this section the fourth aim, i.e. "Do myeloma derived malignant plasma cell factors as proliferation or expression-based scores also determine risk in AL?" is addressed. For this, GEP-based scores and classifications created for MM were assessed in terms of differences in distribution between the entities and prognostic impact. As described in section 1.2.4, three main strategies had been used in their creation. Briefly, the first strategy based on surrogating biological variables is exemplified by GPI and MAI. The second strategy is outlined by classification of myeloma (and in this case AL) in molecular subentities as exemplified by TC and MC classifications. The third strategy is represented by different algorithms of selecting genes associated with survival, i.e. UAMS70, IFM15, EMC92, and RS scores. In the following, assessments are described in the order GPI, MAI, TC, MC, UAMS70, IFM15, EMC92, and RS.

Table 3.9: Univariate Cox regression analyses of overall survival (OS) of patients classified as light chain amyloidosis (AL) by GEP-based surrogates of biological variables and risk stratifications. Events: number of events per score grouping level, HR: hazard ratio, CI: 95% confidence interval for HR, *p*: p-value of Wald test, GEP: gene expression profiling by DNA microarray.

Variable	Level	Events	HR	CI	*p*
GPI	low risk	35			
	medium risk	75	1.24	0.8-1.9	.3
	high risk	8	3.58	1.6-7.8	.001
MAI	$\leq$1	73			
	>1	45	1.42	1-2.1	.07
UAMS70	low risk	106			
	high risk	12	2.38	1.3-4.3	.005
IFM15	low risk	111			
	high risk	7	1.30	0.6-2.8	.5
RS	low risk	88			
	medium risk	24	1.17	0.7-1.8	.5
	high risk	6	4.63	2-10.7	< .001

Table 3.10: Univariate Cox regression analyses of overall survival (OS) of patients classified as light chain amyloidosis (AL) by GEP-based surrogates of biological variables, classifications and risk stratifications. Median time: median survival time in months, Survival rate: percentage of patients being alive after 2 and 5 years for the respective grouping level, n.r.: not reached, GEP: gene expression profiling by DNA microarray.

		Median	**Survival rate [%]**	
Variable	**Level**	**time**	**2-year**	**5-year**
GPI	low risk	39	64	47
	medium risk	43	59	43
	high risk	3	21	21
MAI	$\leq$1	52	64	47
	>1	20	48	36
TC	11q13	52	63	46
	6p21	31	75	47
	D1	27	52	18
	D1+D2	41	56	44
	D2	20	46	37
	FGFR3	27	50	n.r.
	MAF	72	66	66
MC	CD1	n.r.	74	74
	CD2	48	62	45
	HY	27	60	20
	LB	30	53	39
	MF	n.r.	80	80
	MS	3	40	n.r.
	PR	7	40	40
UAMS70	low risk	41	61	45
	high risk	6	29	19
IFM15	low risk	39	59	44
	high risk	32	56	33
RS	low risk	41	62	45
	medium risk	43	51	44
	high risk	3	18	n.r.

3.4.1 Gene expression-based assessment of biological variables

Proliferation assessed by GPI

Using the GPI, 6% of 196 AL patients with available GEP and survival data are classified as high risk compared to 16% of MM patients. AL high risk patients have a very adverse median survival time; eight of eleven patients died during the first year (see table 3.10). Ten of these eleven patients are classified as having concomitantly being diagnosed as myeloma (ALMM, see figure 3.15 **c**, supplementary table A.4). The median OS of patients grouped in low and medium risk is 39 and 43 months, respectively (see table 3.10). The HR of low to high risk is 3.58 ($p = 0.001$) (see table 3.9).
Between different subgroups of AL patients, including presence *versus* absence of heart involvement, levels of NT-ProBNP, TNT, or diff FLC (for thresholds see section 1.4), no significant difference in the proportions of GPI assessment could be detected (see 3.15 **a**). In contrast, a significant difference in the proportions was found between AL and MGUS, AMM and MM (see figure 3.15 **c**). No significant difference was detected between AL with different underlying plasma cell disease, i.e. ALMG and ALMM.

Myc-activation assessed by MAI

A MAI above a threshold of 1 is associated with adverse survival in MM. This was significantly more often found in the ALMM compared to the ALMG subgroup ($p < 0.01$) (see figure 3.16 **c**). With 43% of patients, the group of patients with a MAI > 1 and subentity ALMM is nearly two times larger than in the ALMG subgroup with 22% ($p < 0.01$). The proportion of AL patients with a MAI > 1 is 34% compared to 55% within MM ($p < 0.001$) and 17% in MGUS ($p < 0.05$, see figure 3.16 **c**). In AL, MAI is not associated to OS (see figure 3.16 **b** and table 3.9).

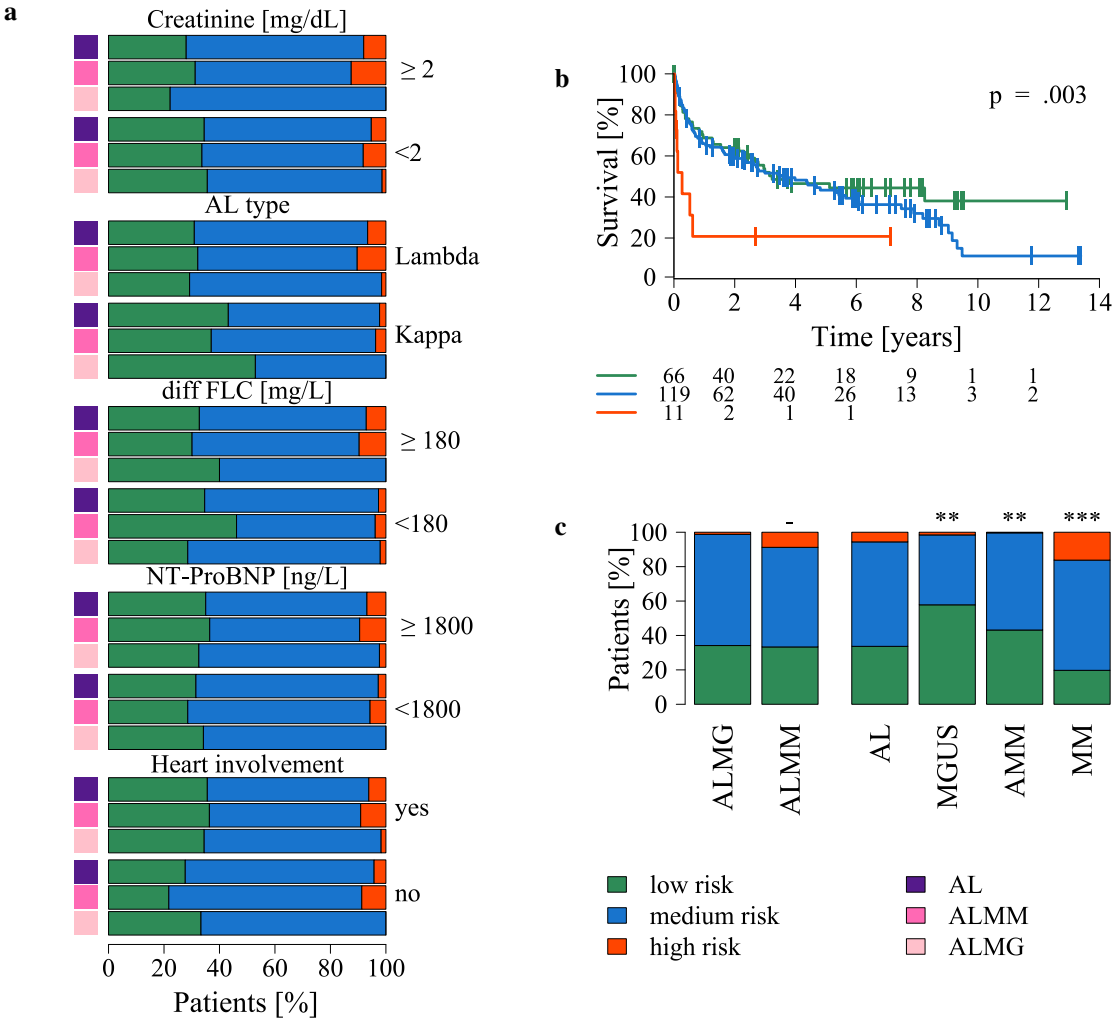

Figure 3.15: Gene expression-based proliferation index (GPI). **a** Distribution of AL associated clinical prognostic markers including heart involvement regarding GPI proliferation rate levels. **b** Overall survival (OS) of AL patients delineated by GPI proliferation rate levels. **c** Distribution of GPI proliferation rate levels low, medium, and high for patients classified as ALMG, ALMM, AL, MGUS, AMM, or MM. Significant differences, indicated by Pearson's χ^2 test, are illustrated as *, **, and ***, representing a significant p-value < .05, .01, and .001, respectively. The legend at the bottom right side of the figure depicts GPI proliferation rate levels and AL entities. See supplementary table A.4 for frequencies regarding **a** and **c**. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

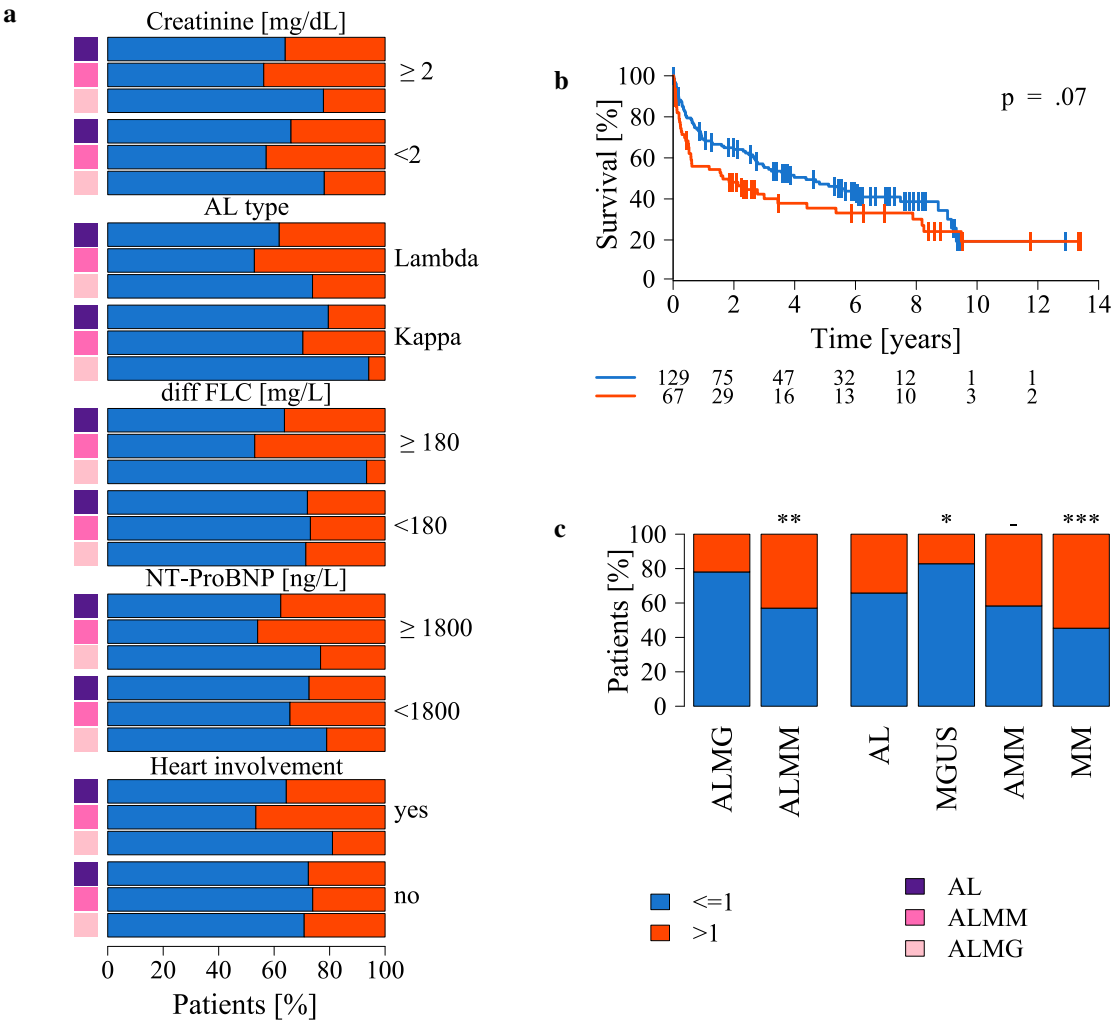

Figure 3.16: Myc-activation index (MAI). **a** Distribution of AL associated clinical prognostic markers including heart involvement regarding MAI levels. **b** Overall survival (OS) of AL patients delineated by MAI levels. **c** Distribution of MAI levels >1 and ≤1 for patients classified as ALMG, ALMM, AL, MGUS, AMM, or MM. Significant differences, indicated by Pearson's χ^2 test, are illustrated as *, **, and ***, representing a significant p-value < .05, .01, and .001, respectively. The legend at the bottom right side of the figure depicts MAI levels and AL entities. See supplementary table A.7 for frequencies regarding **a** and **c**. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

3.4.2 Gene expression-based classifications

TC classification

The proportions of TC classes are significantly different in AL compared to MGUS, AMM, and MM (all $p < 0.001$) (see figure 3.17 **c**). Most striking difference was detected for the 11q13, accounting for 50% of AL patients, in contrast to MGUS (9%), AMM, and MM (both 19%) (see figure 3.17 **c**, supplementary table A.5). The MAF group in AL comprises 10% of the patients, similar to AMM (11%) and MM (8%) but less than within MGUS (28%). The proportions of the D1 and D2 group in AL are different compared to MGUS, AMM, and MM. The D1 group comprises 8% of AL patients compared to 17% in MGUS, 29% in AMM, and 37% in MM. An inverse distribution was found for the D2 group, being larger in AL (21%) compared to MM (12%). OS is not associated with any of the TC classes either in AL or MM (see figure 3.17 **b**). No significant difference was detected regarding AL specific clinical factors between TC classes (see figure 3.17 **a**).

MC

The distribution of groups within the MC is significantly different between AL and MGUS, AMM, and MM (all $p < 0.001$, see figure 3.18 **c**). The CD2 group is the largest group with 47% of AL patients, in contrast to all other entities (see figure 3.18 **c**, supplementary table A.6). From MGUS to AMM and MM the group size of CD2 decreases. The HY group size with 3% in AL is the smallest from all entities. The proportion of patients in the LB group in AL (37%) is comparable to the in MGUS (34%) and consists of twice as much patients as in MM (17%). The MF group size in AL (3%) is closer to MM (4%) than to AMM (7%) or MGUS (16%). Very few AL patients are classified as either MS (3%) or PR (3%), resembling the low to zero frequencies in MGUS or AMM (see figure 3.18 **c**). None of the MC groups is associated with adverse OS (see figure 3.18 **b**). No difference was detected for AL-associated clinical factors or the underlying disease subentity (see figure 3.18 **a**, **c**).

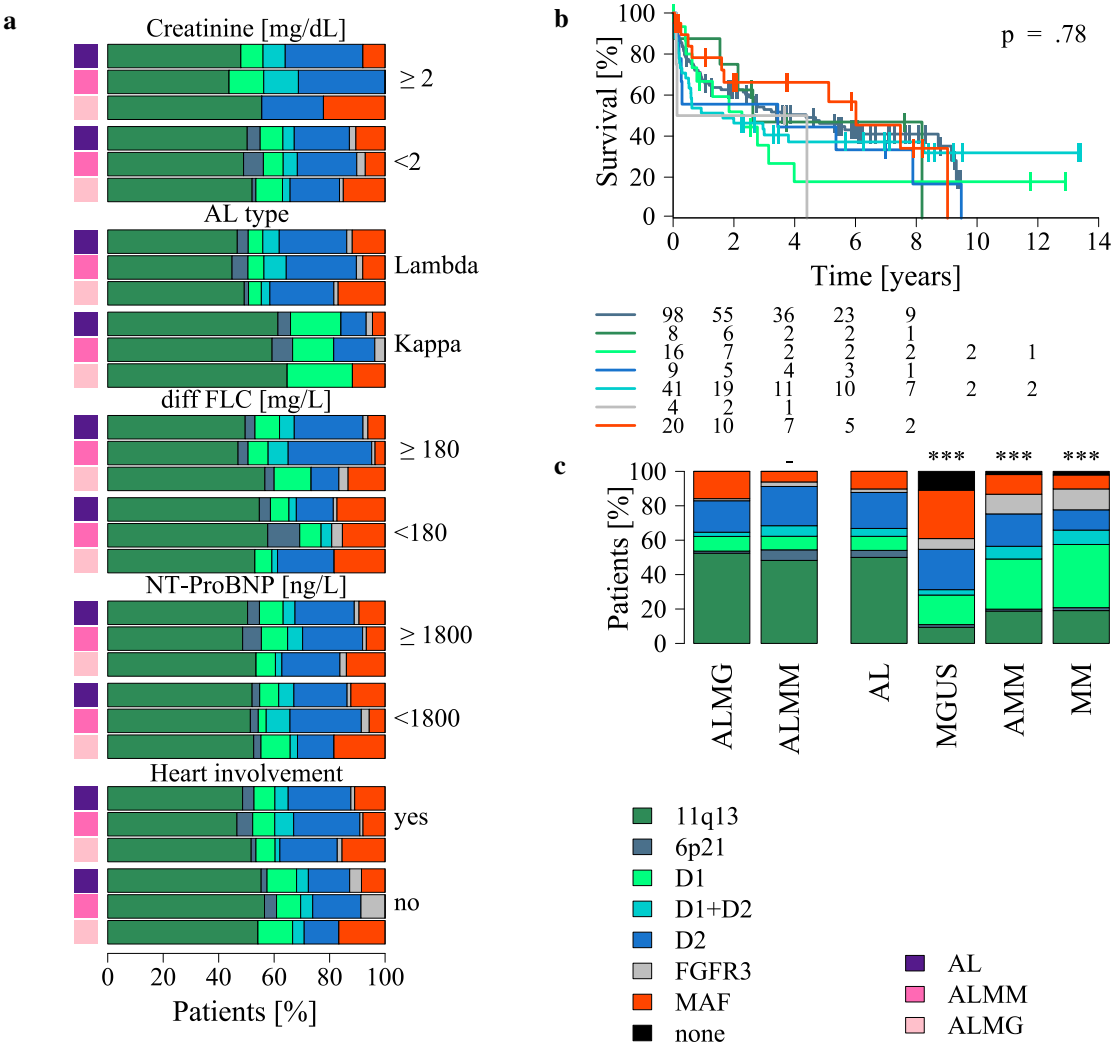

Figure 3.17: Translocation/cyclin D (TC) classification. **a** Distribution of AL associated clinical prognostic markers including heart involvement regarding TC classes. **b** Overall survival (OS) of AL patients delineated by TC classes. **c** Distribution of TC classes for patients classified as ALMG, ALMM, AL, MGUS, AMM, or MM. Significant differences, indicated by Pearson's χ^2 test, are illustrated as *, **, and ***, representing a significant p-value < .05, .01, and .001, respectively. The legend at the bottom right side of the figure depicts TC classes and AL entities. See supplementary table A.5 for frequencies regarding **a** and **c**. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

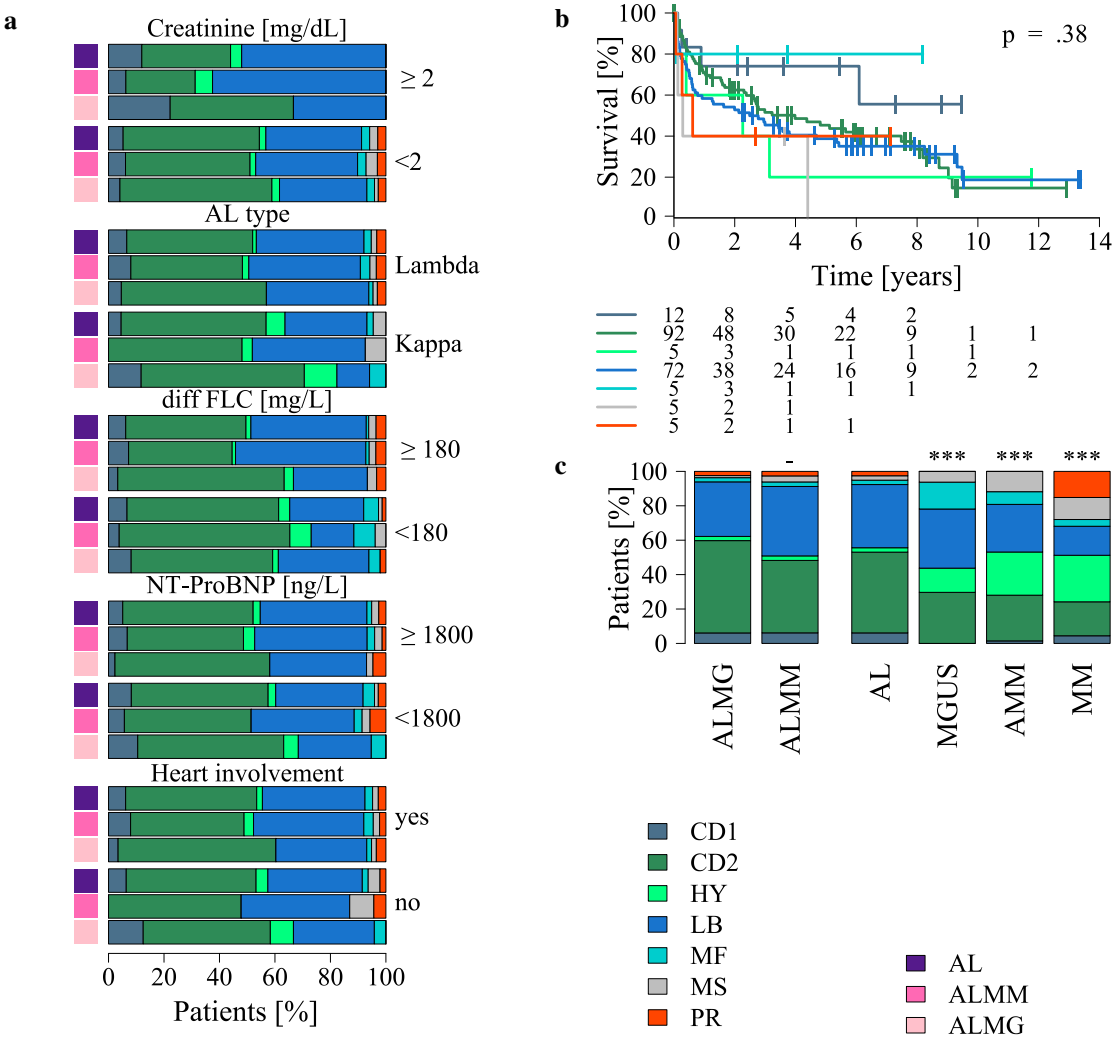

Figure 3.18: Molecular classification (MC). **a** Distribution of AL associated clinical prognostic markers including heart involvement regarding MC classes. **b** Overall survival (OS) of AL patients delineated by MC classes. **c** Distribution of MC groups for patients classified as ALMG, ALMM, AL, MGUS, AMM, or MM. Significant differences, indicated by Pearson's χ^2 test, are illustrated as *, **, and ***, representing a significant p-value < .05, .01, and .001, respectively. The legend at the bottom right side of the figure depicts MC classes and AL entities. See supplementary table A.6 for frequencies regarding **a** and **c**. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

3.4.3 Gene expression-based assessment of risk

UAMS70

In figure 3.19 the UAMS70 score as calculated for AL is depicted; Fifteen (8%) AL patients are stratified as high risk, significantly fewer patients compared to MM (25%, $p < 0.001$, see figure 3.19 **c**). Thirteen of the fifteen high risk AL patients are of the ALMM subentity (see figure 3.19 **c**, supplementary table A.8). Patients identified as high risk had a significant shorter OS compared to low risk patients ($p = 0.004$) (see figure 3.19 **b**, table 3.9, 3.10). Median OS is 41 months for patients in the low risk group compared to 6 months in the high risk group (see table 3.10). The HR of low to high risk is 2.38 (see table 3.9). AL specific parameters (see section 1.4) are not significantly differentially distributed in the proportions between low and high risk group (see figure 3.19 **a**).

IFM15

The IFM15 score classifies 12 (6%) AL patients as high risk, in contrast to 25% in MM ($p < 0.001$, see figure 3.20 **c**, supplementary table A.9). An association of the IFM15 to the OS of AL patients was not found (see figure 3.20 **b** and table 3.9). The median time of OS is 39 months in the low risk group and 32 months in the high risk group (see table 3.10). A difference in the proportions of AL specific factors in relation to IFM15 was not detected (see figure 3.20 **a**).

EMC92

The EMC92 classifies all 196 AL patients as standard risk, while it classifies 74 (10%) MM patients as high risk.

RS

Seven AL patients (4%), all classified as ALMM, were identified as being high risk by the RS. This proportion is significantly smaller compared to the RS high risk group in MM (8%, $p < 0.001$; see figure 3.21 **c**, supplementary table A.10). The median OS in the high risk group is 3 months (see table 3.10). All seven patients died within the first 3 years of follow-up (see figure 3.21 **b**). The median survival time in the low and the medium risk group is not significantly different with 41 and 43 months, respectively. The HR of low to high risk indicates a 4.6 fold higher risk of death for high risk patients (see table 3.9). A difference in the proportions of RS groups by AL specific factors was not found (see figure 3.21 **a**).

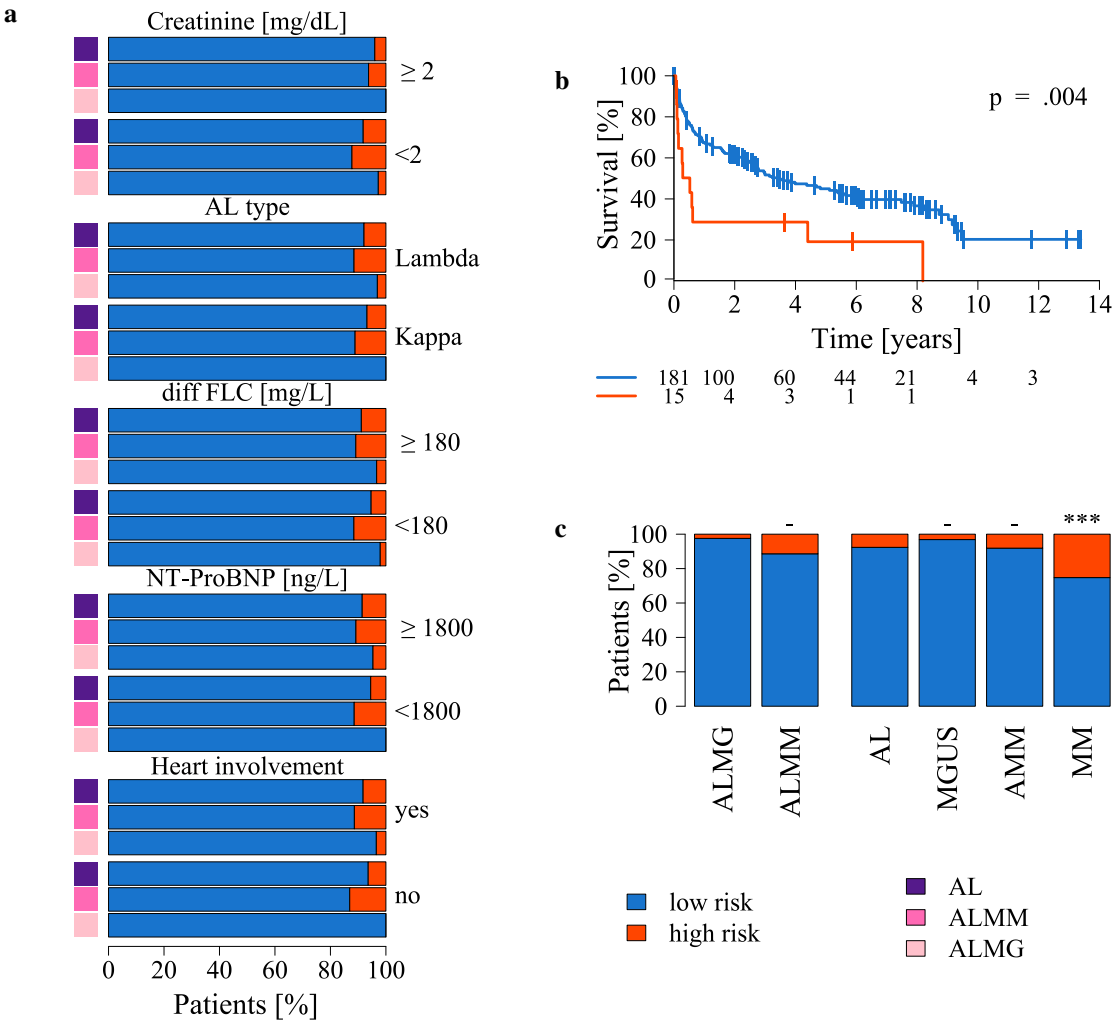

Figure 3.19: UAMS 70-gene score (UAMS70). **a**Distribution of AL associated clinical prognostic markers including heart involvement regarding UAMS70 risk stratification. **b** Overall survival (OS) of AL patients delineated by UAMS70 groups. **c** Distribution of UAMS70 high and low risk groups for patients classified as ALMG, ALMM, AL, MGUS, AMM, or MM. Significant differences, indicated by Pearson's χ^2 test, are illustrated as *, **, and ***, representing a significant p-value < .05, .01, and .001, respectively. The legend at the bottom right side of the figure depicts UAMS70 groups and AL entities. See supplementary table A.8 for frequencies regarding **a** and **c**. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

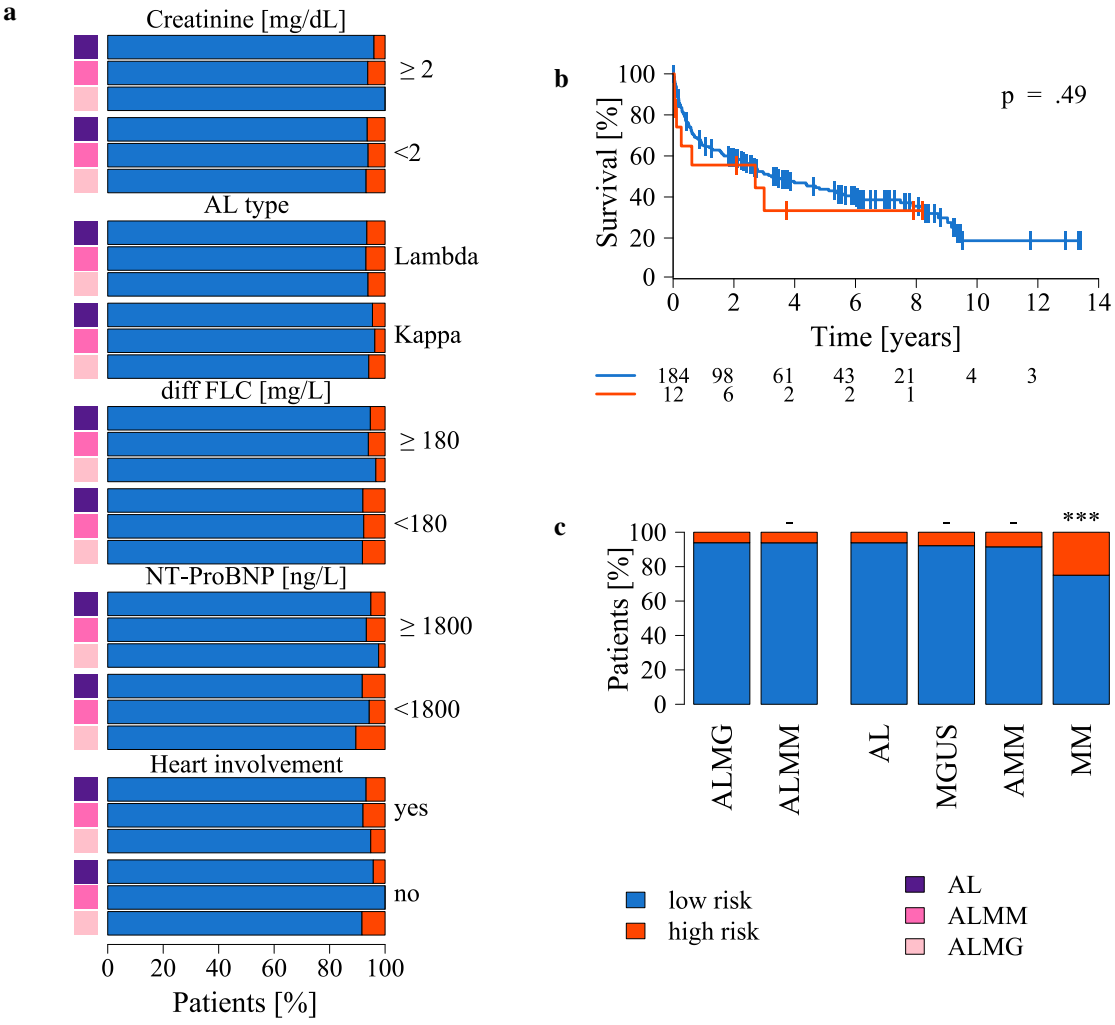

Figure 3.20: Intergroup Francophone du Myélome 15-gene score (IFM15). **a** Distribution of AL associated clinical prognostic markers including heart involvement regarding IFM15 risk stratification. **b** Overall survival (OS) of AL patients delineated by IFM15 risk groups. **c** Distribution of IFM15 low and high risk groups for patients classified as ALMG, ALMM, AL, MGUS, AMM, or MM. Significant differences, indicated by Pearson's χ^2 test, are illustrated as *, **, and ***, representing a significant p-value < .05, .01, and .001, respectively. The legend at the bottom right side of the figure depicts IFM15 risk groups and AL entities. See supplementary table A.9 for frequencies regarding **a** and **c**. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

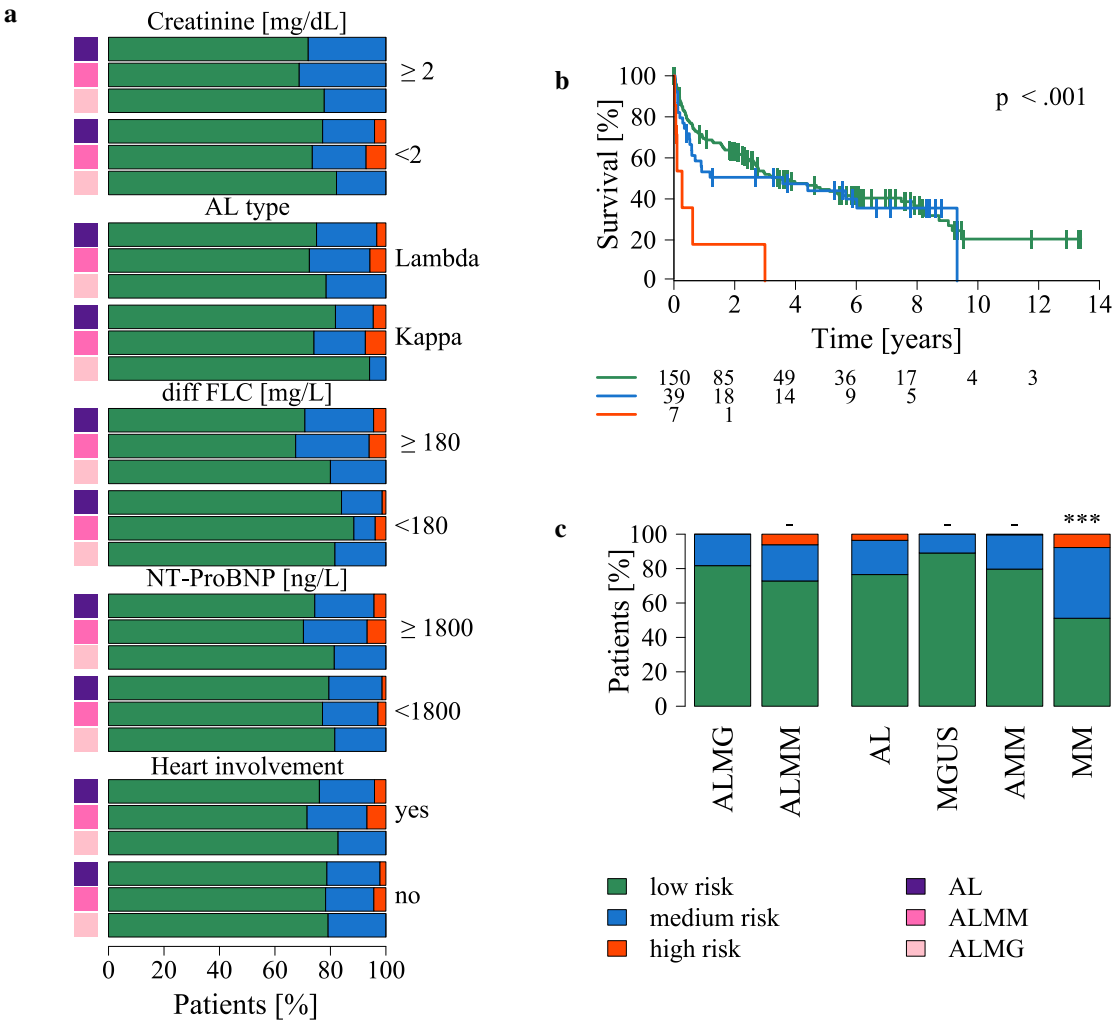

Figure 3.21: Risk score (RS). **a** Distribution of AL associated clinical prognostic markers including heart involvement regarding RS risk stratification. **b** Overall survival (OS) of AL patients delineated by RS groups. **c** Distribution of RS low, medium, and high risk groups for patients classified as ALMG, ALMM, AL, MGUS, AMM, or MM. Significant differences, indicated by Pearson's χ^2 test, are illustrated as *, **, and ***, representing a significant p-value < .05, .01, and .001, respectively. The legend at the bottom right side of the figure depicts HDAL groups and AL entities. See supplementary table A.10 for frequencies regarding **a** and **c**. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

3.5 New gene expression-based risk assessment for light chain amyloidosis

In the following section the fifth aim is addressed, i.e. "Is it possible to define an expression-based risk score for AL patients, and does it in turn conveys prognostic significance in MM patients?". For this, the HDAL score was created at the LfM as part of this thesis (see section 2.3.3), and applied here regarding its prognostic impact and potential interrelation with clinical risk stratification and different patterns of organ involvement in AL patients. Subsequently, the assessments based on gene expression data of myeloma patients described in the previous section 3.4 were compared to the HDAL score.

3.5.1 HDAL score

The categorical HDAL stratifies 46 (23.5%) of all AL patients as high risk. Of these seven, all with heart involvement and λ LC, were sub-stratified as ALMG (see figure 3.22 **c**, supplementary table A.11). Compared to the HDAL stratification of MGUS, AMM, and MM, the HDAL groups in AL are significantly different (see figure 3.22 **c**). With increasing stage (MGUS to AMM to MM), the percentage of high risk patients significantly grows (Jonckheere-Terpstra test, $p < 0.001$). If AL patients are sub-stratified for concomitant presence of MGUS or MM, the latter cohort shows significantly more high risk patients (Jonckheere-Terpstra test, $p < 0.001$).

The ability of the derived HDAL score to predict survival in AL patients had been validated in an independent group ($n = 97$, see figure 3.22 **b** and table 3.11). It significantly delineates low from high risk patients with a HR of 3.8 (p<0.001, see table 3.11). The median survival time is $72/33/6$ months for low, medium, and high risk, respectively. It is interesting to denote that the HDAL score, derived on a cohort of AL patients, is likewise predictive for AMM and MM (both $p < 0.001$, see figure 3.23).

Table 3.11: Univariate Cox regression analysis with overall survival (OS) of patients classified as light chain amyloidosis (AL) included in the validation group ($n = 97$) by Heidelberg AL score (HDAL). Events: number of events per score grouping level, HR: hazard ratio, CI: 95% confidence interval for HR, p: p-value of Wald test, Median time: median survival time in months, Survival rate: percentage of patients being alive after 2 and 5 years for the respective grouping level, n.r.: not reached.

					Median	Survival rate [%]	
Level	**Events**	**HR**	**CI**	p	**time**	**2-year**	**5-year**
low risk	22				72	67	54
medium risk	16	1.62	0.8-3.1	.14	33	69	29
high risk	14	3.80	1.9-7.6	< .001	6	34	n.r.

Table 3.12: Multivariate Cox regression analyses with overall survival (OS) of patients classified as light chain amyloidosis (AL) included in the validation group ($n = 97$) regarding Heidelberg AL score (HDAL), AL stagings, and serum parameters. HR: hazard ratio, CI: 95% confidence interval for HR, p: p-value of Wald test.

Variable	**Level**	**HR**	**CI**	p
HDAL & Euro	medium risk	1.62	0.8-3.2	.15
	high risk	3.21	1.6-6.5	.001
	II	2.31	0.7-7.5	.16
	IIIA	3.27	1.1-9.9	.04
	IIIB	5.40	1.8-16.1	.002
HDAL & Mayo 2012	medium risk	1.87	1-3.6	.06
	high risk	2.85	1.4-6	.006
	1	1.37	0.3-7.1	.71
	2	4.31	1-18.7	.051
	3	4.34	1-18.8	.05
HDAL & Mayo 2004	medium risk	1.73	0.9-3.4	.1
	high risk	3.36	1.7-6.8	< .001
	2	2.35	0.7-7.6	.15
	3	4.19	1.5-12	.008
HDAL & NT-ProBNP	medium risk	1.70	0.9-3.3	.11
	high risk	3.28	1.6-6.6	< .001
	$\geq$1800 ng/L	2.84	1.4-5.6	.003
HDAL & cTnT	medium risk	1.91	1-3.7	.055
	high risk	3.17	1.5-6.5	.002
	$\geq$0.025 ng/mL	2.65	1.2-6	.02
HDAL & diff FLC	medium risk	1.63	0.9-3.1	.14
	high risk	3.29	1.6-7	.002
	$\geq$180 mg/L	1.37	0.7-2.6	.34
HDAL & Creatinine	medium risk	1.64	0.9-3.1	.14
	high risk	3.95	1.9-8.1	< .001
	$\geq$2 mg/dL	0.83	0.3-2	.68

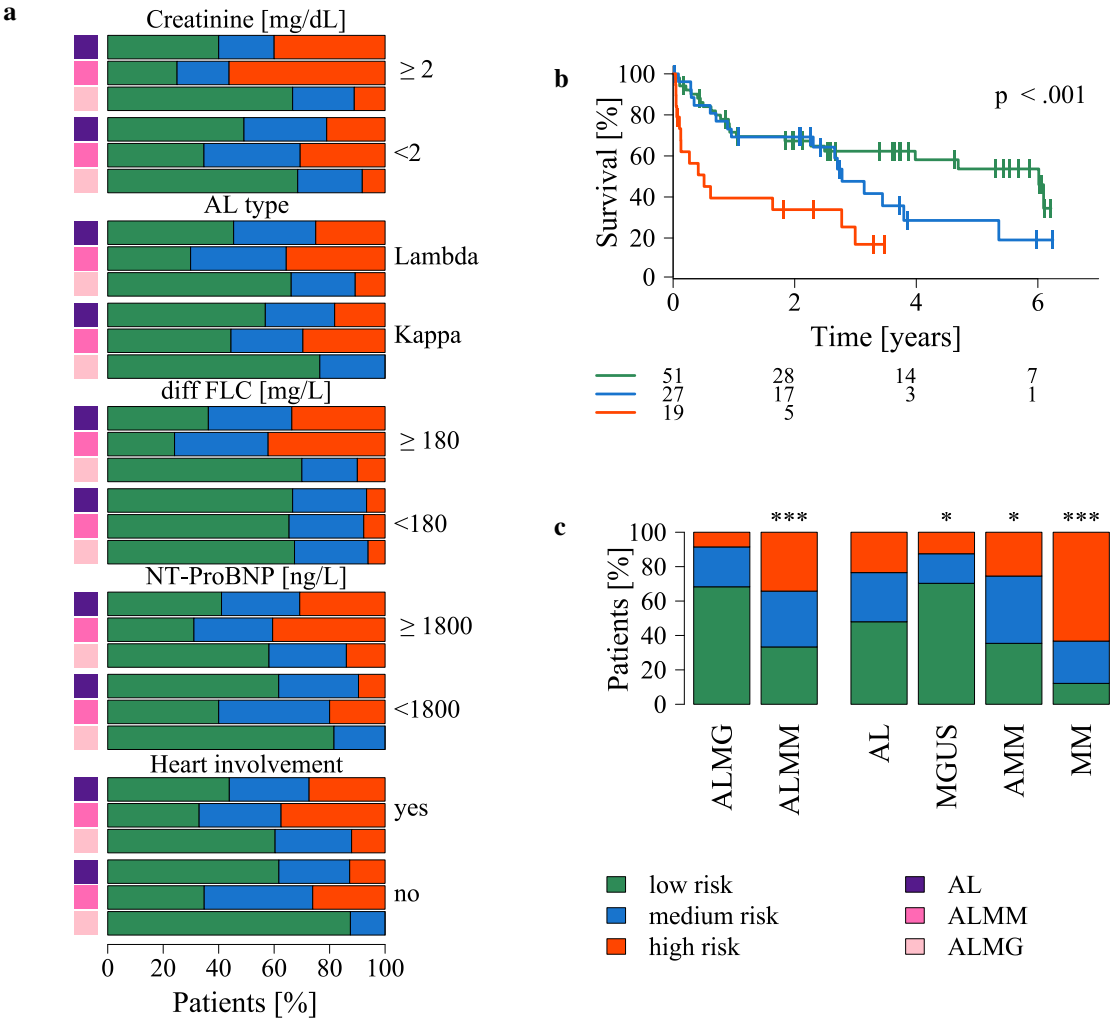

Figure 3.22: Heidelberg AL score (HDAL). **a** Distribution of AL associated clinical prognostic markers including heart involvement regarding HDAL risk stratification. **b** Overall survival (OS) of the validation group ($n = 97$) delineated by HDAL groups. **c** Distribution of HDAL low, medium, and high risk groups for patients classified as ALMG, ALMM, AL, MGUS, AMM, and MM. Significant differences, indicated by Pearson's χ^2 test, are illustrated as *, **, and ***, representing a significant p-value < .05, .01, and .001, respectively. The legend at the bottom right side of the figure depicts HDAL groups and AL entities. See supplementary table A.11 for frequencies regarding **a** and **c**. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

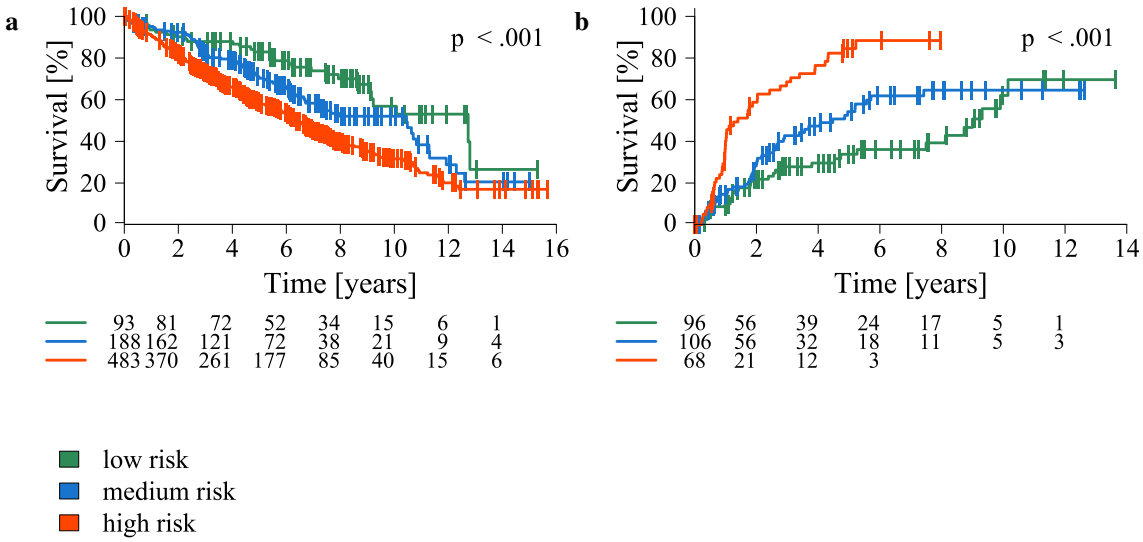

Figure 3.23: **a** Overall survival (OS) of multiple myeloma (MM) patients and **b** progression free survival (PFS) of asymptomatic multiple myeloma (AMM) patients for high and low risk groups as determined by Heidelberg AL score (HDAL) risk stratification. Difference between curves was tested using Log-rank test and was termed significant if the p-value was ≤ 0.05. The legend at the bottom depicts HDAL groups.

A significant difference of HDAL low and high risk groups was found in AL if subdivided regarding diff FLC ≥ 180 *versus* < 180 mg/L ($p < 0.001$), or NT-ProBNP levels of ≥ 1800 *versus* < 1800 ng/L ($p = 0.004$), see figure 3.22 **a**.
In multivariate Cox regression analyses, the categorical HDAL score is independently predictive if tested with either Mayo or Euro scores, as well as from the biomarkers NT-ProBNP, cTnT, diff FLC, and creatinine (see table 3.12, for definition see section 1.4). The continuous HDAL does not correlate with any biomarker (NT-ProBNP, cTnT, diff FLC, and creatinine) and the hazard of it significantly increases over time ($p < 0.001$). The 59 microarray IDs for the prognostic genes of the HDAL score were translated to 64 unique ENTREZIDs. These were assessed by FEA, as described in section 2.5.6. Six terms were found significantly enriched: "phagocytosis, recognition" (GO:0006910), "Protein processing in endoplasmic reticulum" (hsa04141), "ATP biosynthetic process" (GO:0006754), "negative regulation of ion transport" (GO:0043271), "immune response-regulating cell surface receptor signaling pathway involved in phagocytosis" (GO:0002433), and "mRNA 3'-end processing" (R-HSA-72187). The largest group of genes, with six genes, is annotated and enriched to the GO term "phagocytosis, recognition". HDAL genes that are Ig genes are enriched in this term.

3.5.2 Delineation of high risk by gene expression-based scores

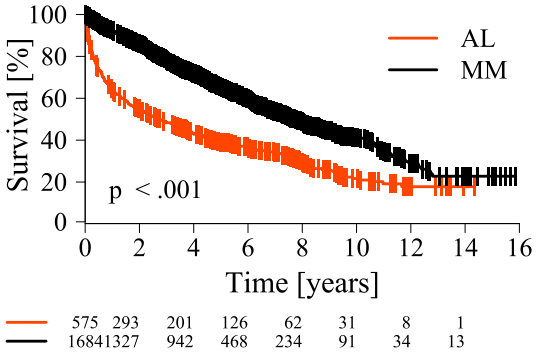

Figure 3.24: **a** Overall survival (OS) of patients suffering from light chain amyloidosis (AL) or multiple myeloma (MM). Difference between curves was tested using Log-rank test and was termed significant if the p-value was ≤ 0.05.

Prognosis of AL patients is much poorer than prognosis for MM, especially during the first 12 months (survival rate of 64% *versus* 92%, see figure 3.24).
Regarding GEP-based scores predictive for survival in AL patients (GPI, UMAS70, RS, and HDAL), 54 patients were classified as high risk (see figure 3.25 **a** for depiction of overlaps). Of 46 patients classified as high risk by the HDAL score, 33 were exclusively identified by it, i.e. neither by GPI, UMAS70, nor RS. *Vice versa*, 8 (15%) AL patients were not identified by HDAL but either by GPI, UAMS70 or RS. Of all AL patients, with survival data and GEP, succumbing to their disease within the first 12 months ($n = 68$), 13 (19%) were identified by MM-based scores, 32 (47%) by HDAL. Of these, 8 were identified by the GPI, 10 by UAMS70, 5 by RS, and 34 (50%) by none of the scores. Of these, 22 (32%) patients were identified by standard Mayo staging, and 18 (26%) by revised Mayo staging. In the MM cohort, HDAL exclusively identified 286 patients as high risk (53%) and only 10% of patients were not identified by HDAL (see figure 3.25 **b**).
The frequency of patients classified as high risk by all considered GEP-based scores for the AL patients is different compared to the frequency of the MM disease entities (see table 3.13). But the larger frequency of patients classified as high risk in the MM cohort compared to MGUS and AMM is "transferred" to the AL subentities ALMM compared to ALMG.

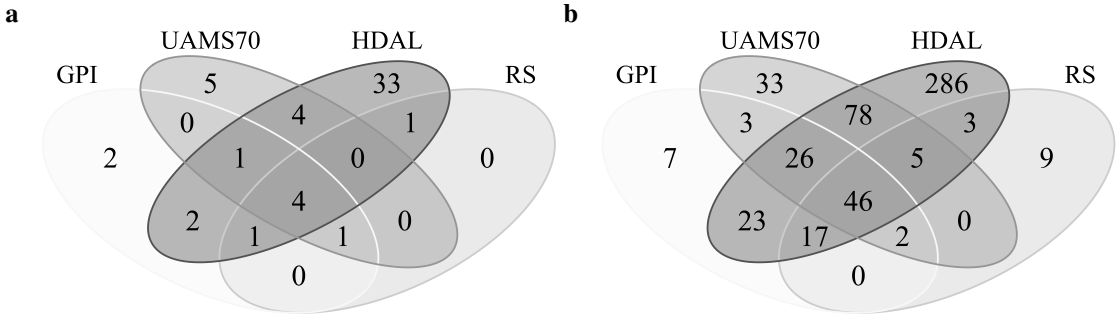

Figure 3.25: Comparison and overlap of high risk by different GEP-based risk assessments. Patients defined as high risk suffering from **a** light chain amyloidosis (AL) or **b** multiple myeloma (MM). Depicted scores are GPI, UAMS70, RS, and HDAL. GEP: gene expression profiling by DNA microarray.

Table 3.13: Frequency of patients classified as high risk by GEP-based risk assessments in the different disease entities. The respective risk assessments are GPI, UAMS70, RS, IFM15, MAI, and HDAL. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma, GEP: gene expression profiling by DNA microarray.

Entity	GPI	UAMS70	Rs	IFM15	MAI	HDAL
ALMG	1 (1.2%)	2 (2.4%)	0 (0%)	5 (6.1%)	18 (22%)	7 (8.5%)
ALMM	10 (8.8%)	13 (11.4%)	7 (6.1%)	7 (6.1%)	49 (43%)	39 (34.2%)
AL	11 (5.6%)	15 (7.7%)	7 (3.6%)	12 (6.1%)	67 (34.2%)	46 (23.5%)
MGUS	1 (1.6%)	2 (3.1%)	0 (0%)	5 (7.8%)	11 (17.2%)	8 (12.5%)
AMM	1 (0.4%)	22 (8.1%)	1 (0.4%)	23 (8.5%)	113 (41.7%)	69 (25.5%)
MM	124 (16.2%)	193 (25.2%)	82 (7.8%)	191 (25%)	418 (54.6%)	484 (63.3%)

3.6 Pathogenetic role of malignant plasma cell characteristics in AL in comparison to MGUS, AMM, and MM

This section addresses the sixth aim of this thesis, i.e. "What are the differences and similarities of malignant plasma cells in AL in relation to MM and to the precursor stages MGUS and AMM?", and specifically, the last two aims, i.e. "Do malignant plasma cells in AL represent a unique molecular entity in terms of pathophysiology?" and "What 'molecular age' can be attributed to the malignant plasma cells in AL, i.e. do they resemble myeloma cells, MGUS cells, or earlier precursors?". Here, malignant plasma cells of AL patients were compared to malignant plasma cells of other disease entities and normal BMPC.

3.6.1 Chromosomal aberrations as assessed by iFISH

The distribution of CA in the plasma cell samples of 582 AL patients was analyzed in comparison to MGUS ($n = 306$), AMM ($n = 444$), MM ($n = 1691$), and regarding simultaneous presence of MGUS or MM in AL patients, ALMG ($n = 264$) and ALMM ($n = 318$). Depicted in figure 3.26, AL contains a significant larger proportion of patients with presence of t(11;14) (58%) and consequently of IgH-TL (IgH-rearrangement) (84%) in comparison to MGUS (20%/62%), AMM (22%/60%) and MM (20%/63%). This stage dependent difference is not present between ALMG and ALMM (see figure 3.27).

HRD and the underlying gains of odd numbered chromosomes (5, 9, 15, and 19, see the respective part in section 2.1.2 for definition) are substantially rarer in AL (17%) compared to MGUS (26%), AMM (43%), and MM (52%) (see figure 3.26). The smallest difference is found between AL and MGUS. This also applies to the comparison of AL subentities ALMG (12%) and ALMM (22%), driven by the difference in gain of 19q13 (see figure 3.27).

Patterns of co-occurrence of CA slightly differ in AL compared to MM (see figure 3.28), mainly driven by the different frequency of t(11;14) and hyperdiploidy (see figure 3.26). The presence of t(11;14) and hyperdiploidy is significantly disjunct in AL ($p < 0.001$, see figure 3.28), being detectable simultaneously in 3% of AL patients. A fraction of 39 (7%) AL patients neither harbors an IgH-TL nor are HRD. Nonetheless, 24 (62%) of these patients show presence of at least one other "myeloma typical aberration", in this case defined as presence of an alteration detected by any of the investigated iFISH-probes (1q21, 5p15, 5q31, 5q35, 8p21, 11q13, 13q14, 15q22, 17p13, or 19q13, see section 1.6.1, 2.1.2, 1.2.3).

The frequencies of patients harboring a gain 1q21 or a del 13q14 are both higher in AL

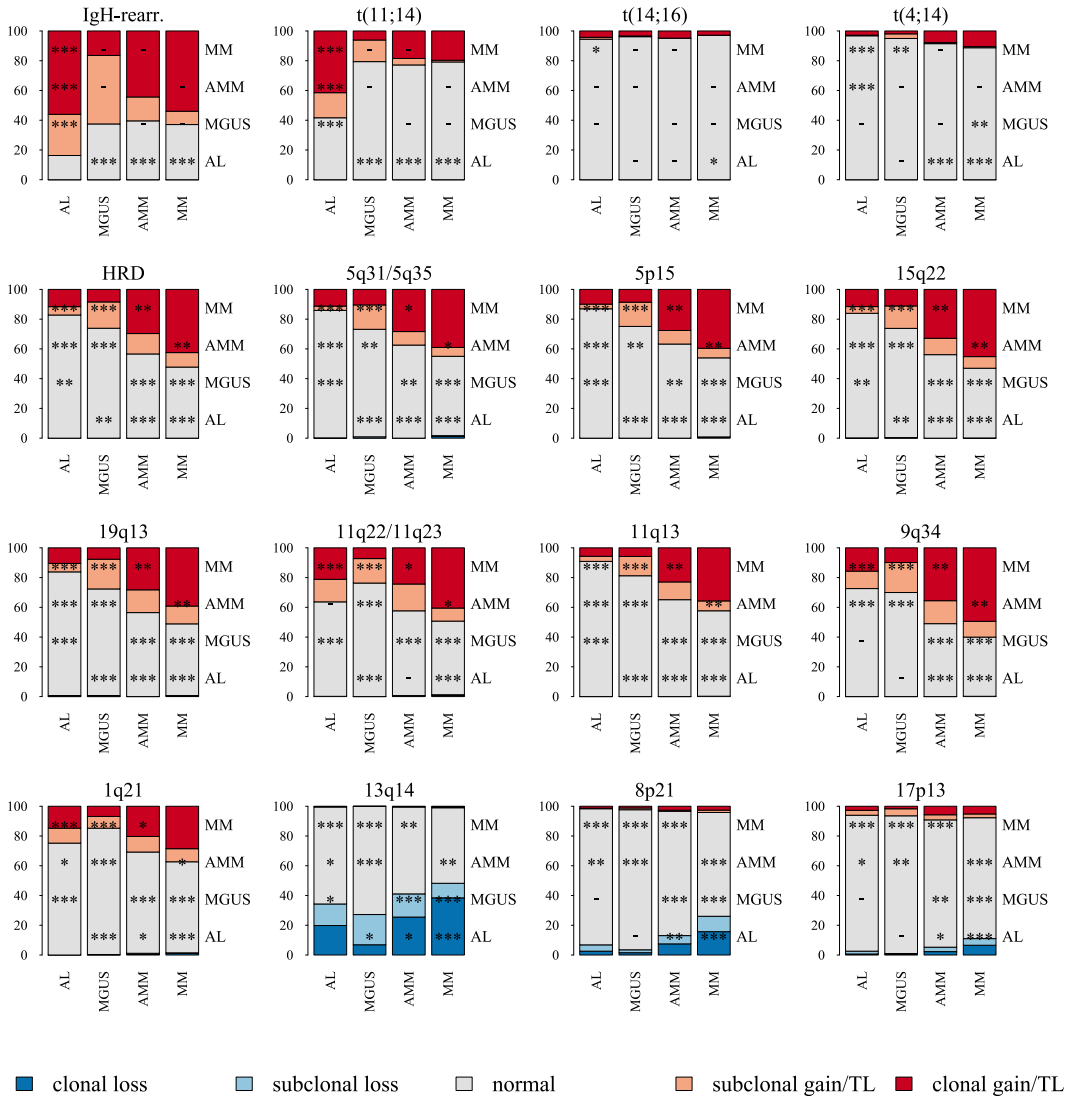

Figure 3.26: Frequency of chromosomal aberrations (CA) in different disease entities. Frequencies are depicted in percent on the y-axis. Significant differences, indicated by Fisher's exact test, are illustrated as *, **, and ***, representing a significant p-value < .05, .01, and .001, respectively. AL: light chain amyloidosis, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma, TL: translocation, IgH-rearr.: IgH-rearrangement

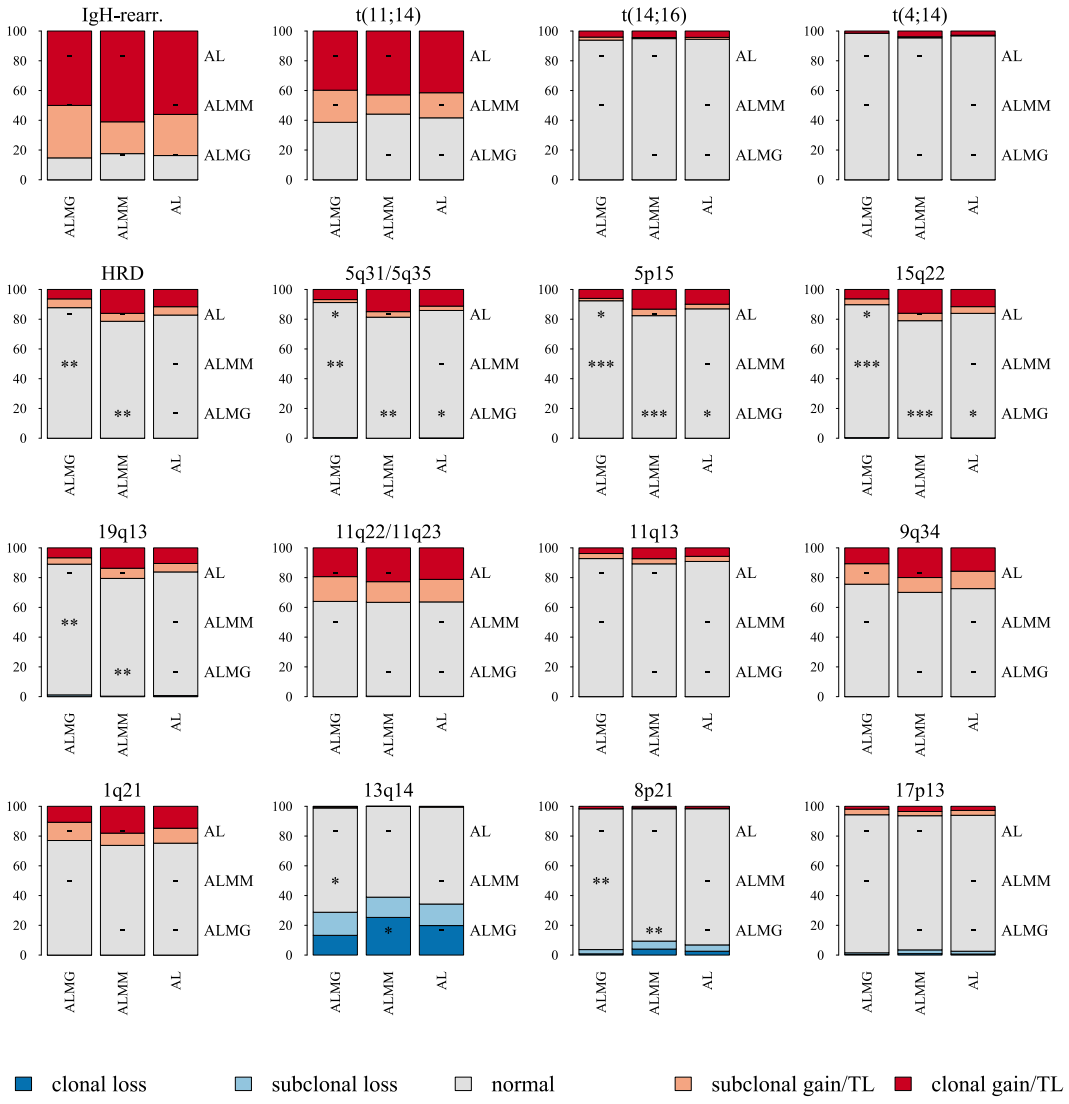

Figure 3.27: Frequency of chromosomal aberrations (CA) in AL, ALMG and ALMM. Frequencies are depicted in percent on the y-axis. Significant differences, indicated by Fisher's exact test, are illustrated as *, **, and ***, representing a significant p-value < .05, .01, and .001, respectively. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, MM: multiple myeloma, TL: translocation, IgH-rearr.: IgH-rearrangement

(25%/34%) than in MGUS (15%/27%), but lower compared to AMM (31%/41%) or MM (37%/48%). Proportions of patients with either del 8p21 or del 17p13 in AL (7%/2.5%) are equal to the proportions in MGUS (3.5%/1%) and significantly smaller compared with AMM (13%/5%) or MM (26%/11%). Within AL patients, ALMM shows significantly higher fractions of presence of del 13q14 and del 8p21 but not del 17p13 (39%/9%/3.5%) compared with ALMG (29%/4%/1.5%) (see figure 3.27).

In contrast to MM (67%), most AL (47%) patients do not harbor more than one of the CA 1q21, 13q14, 17p13, 8p21, HRD, t(4;14), and t(11;14). Corresponding to the proportions detected in MGUS (26%) there is a higher frequency of patients harboring more than one CA in ALMM (54%) compared to ALMG (38%) (see table 3.14).

Table 3.14: Frequency of chromosomal aberrations in different disease entities. Patients harboring 0 or 1 aberration and patients harboring ≥ 2 aberrations. Summed up aberrations: 1q21, 13q14, 17p13, 8p21, HRD, t(4;14), t(11;14). AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma

Number	AL	ALMG	ALMM	MGUS	AMM	MM
0	10.2	14	7.3	37.3	10.1	4.4
1	43	48.2	39.2	36.3	39.9	28.8
>=2	46.8	37.8	53.5	26.4	50	66.8

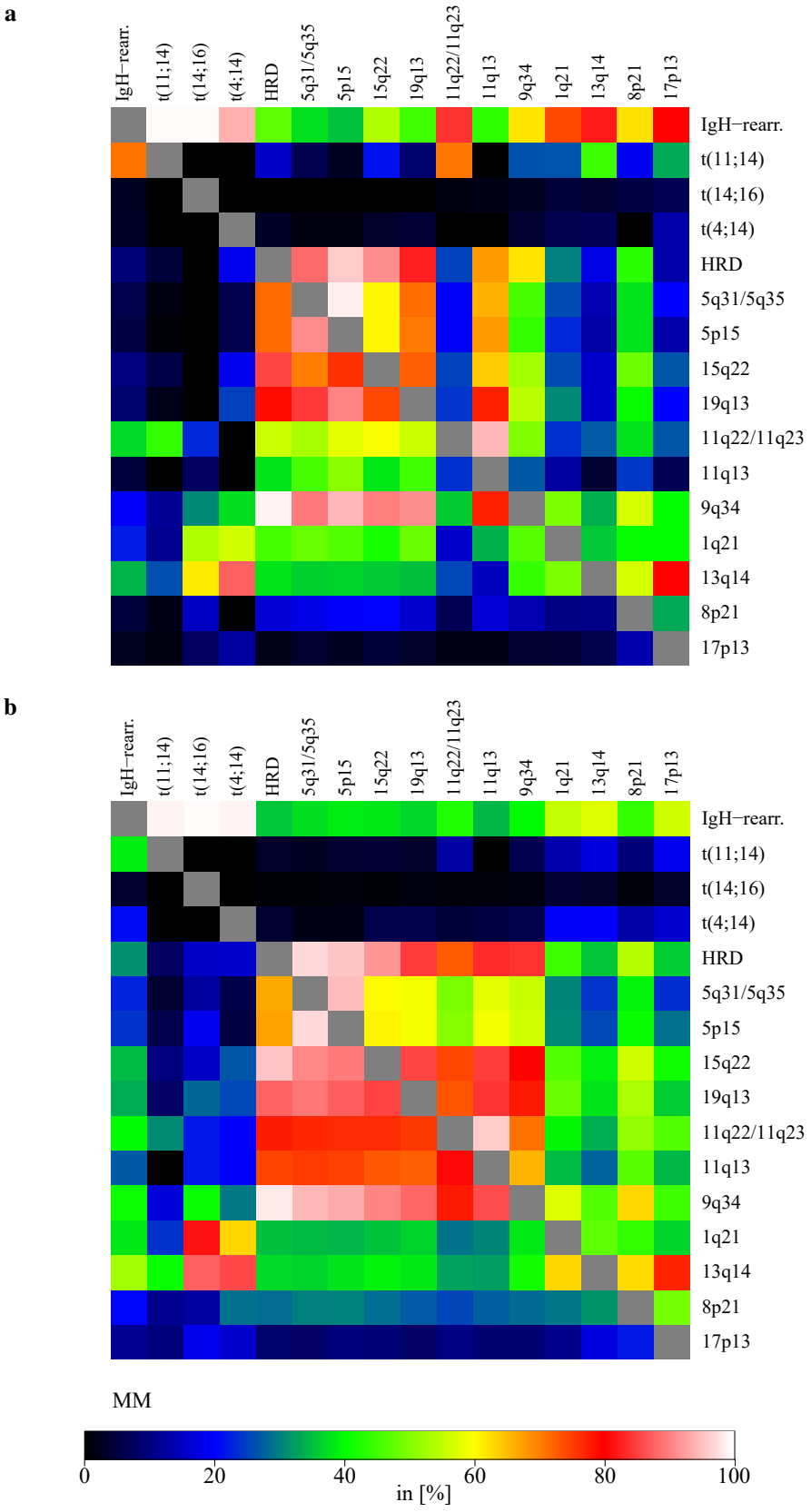

Figure 3.28: Co-occurrence of chromosomal aberrations detected by interphase fluorescence *in situ* hybridization (iFISH) for **a** light chain amyloidosis (AL) and **b** multiple myeloma (MM).

3.6.2 Copy number alterations as assessed by WES

The copy number segments of 113 AL and 28 MM WES samples were created as described in section 2.4.5 and analyzed as outlined in 2.5.4.
In figure 3.29, all alterations (raw copy number segments) per sample are depicted for the AL patients. Large gains of chromosomal regions (spanning the whole chromosome or a chromosome arm) are more frequent than large deletions in AL. This can be exemplarily visualized for alterations involving chromosomes 9 and 11 (figure 3.29).

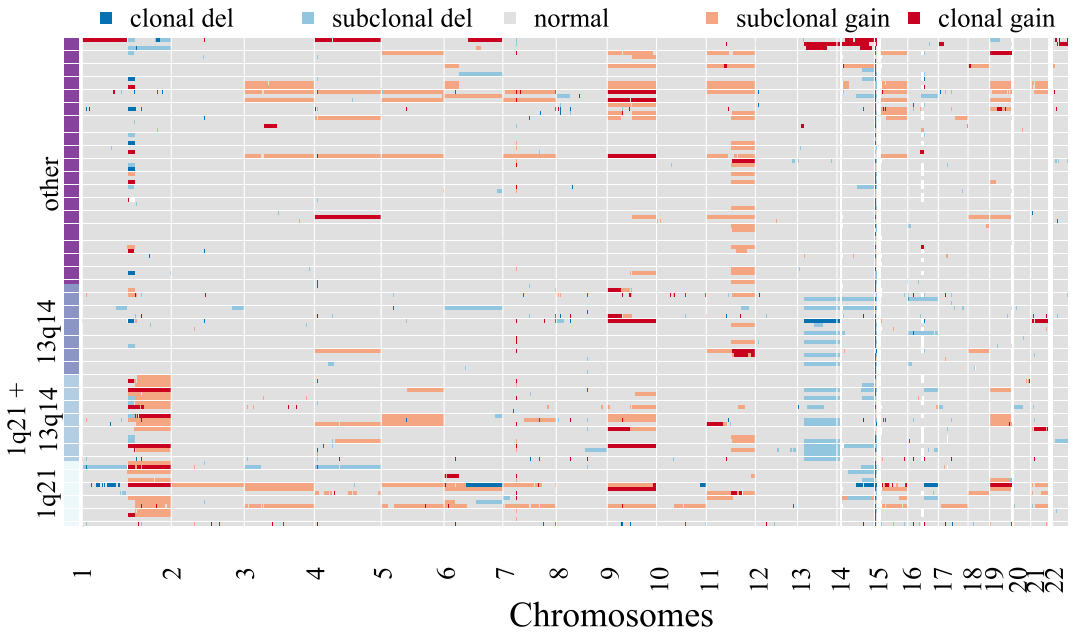

Figure 3.29: Copy number alterations (CNA) assessed by whole exome sequencing of 113 AL patients. Every row represents CNA data of one patient sample, sorted according to presence of gain 1q21 and deletion 13q14. The color legend above the plot indicates for copy number status, clonal or subclonal alteration or normal, diploid copy number status. The copy number status is calculated as the mean log ratio at the respective chromosomal segment. For clonal and subclonal deletions the log ratio is below -1 and -0.5, for normal, diploid status it is between -0.5 and 0.25, for subclonal gain and clonal gain it is above 0.25 and 0.5.

A higher number of different cohort-wide CNA can be found in AL (50) compared to MM (25), see table 3.15. If choosing a random set of 28 AL patients to adjust for the number of tested patients, this difference remains (47 *versus* 25). AL patients harbor a minimum of one CNA, MM patients a minimum of five. The median number of CNA per patient is 9 in AL and 12.5 in MM, see table 3.15. The four CNA provided with the highest GISTIC scores in the AL cohort are gain of 7q34, gain of 14q11.2, deletion of 14q32.33 and deletion of 22q11.22 (for details on the method see section 2.4.5). In the following the comparison of CNA detected in AL and MM is described. Afterward, the analysis on the influence of the CNA on gene expression is outlined. Complete methods are described in the respective section 2.5.4.

Table 3.15: Cohort-wide copy number alterations (CNA) in 113 light chain amyloidosis (AL) and 28 multiple myeloma (MM) patients: **a** Total number of different cohort-wide CNA. **b** Summary statistics calculated per cohort for the disease entities AL and MM. Del: deletion, 25% Q value of the 25% percentile (first quartile), 75% Q value of the 75% percentile (third quartile)

a

Entity	All	Gain	Del
AL	50	20	30
MM	25	16	9

b

Entity	Mean	Median	Min	25% Q	75% Q	Max
AL	9.3	9	1	5	13	28
MM	12.5	12.5	5	10	15	20

Copy number alterations in light chain amyloidosis *versus* in multiple myeloma

Table 3.16: Copy number alterations present in 113 light chain amyloidosis (AL) and 28 multiple myeloma (MM) patients. Comparison by odds ratio (OR) regarding the fraction of patients harboring a specific alteration. Number of patients per disease entity is given in percent and the number of genes altered is given for AL. CI: 95% confidence interval for OR, *p*: Fisher's exact test p-value

Cytoband	Alteration	OR	95% CI	*p*	AL [%]	MM [%]	Genes AL
19q13.42	Gain	120.1	17.8-5022.4	< .001	17.7	96.4	8
2p11.2	Del	24.25	8-83.2	< .001	10.6	75	3
1p36.33	Gain	4.83	1.7-13.5	< .001	13.3	42.9	2
22q11.22	Del	3.66	1.4-9.8	.004	67.3	35.7	4
7q34	Gain	2.8	1.1-7.5	.02	38.9	64.3	1
14q32.33	Del	2.91	1-8	.03	84.1	64.3	10
14q11.2	Gain	2.65	0.9-7.5	.054	85	67.9	1
4q12	Del	4.21	0.3-60.6	.18	1.8	7.1	1

The four CNA with the highest GISTIC scores in AL are among the eight CNA also detected in the 28 MM patient samples. In six of the eight CNA a significant minimum 3-fold difference in the frequency of appearance in the cohort, indicated by OR of AL to MM, is present (see table 3.16). Gain of 19q13.42, gain of 1p36.33, gain of 7q34 are significantly more frequent in MM, gain of 14q11.2 is more frequent in AL. Among the deletions, 2p11.2 is significantly more frequent in MM, but del 22q11.22 and del 14q32.33 are significantly more frequent in AL.

Expression of affected genes in copy number alterations

Of the 3153 genes potentially altered by any detected CNA in AL, 2853 genes could be annotated to an ensemble gene id. In the 113 AL RNA seq samples, 1211 of these genes are expressed, i.e. present in the expression table (for present definition, see section 2.5.4), and were analyzed for altered gene expression. Of these genes, 1187 belong to the five CNA (gain 1q21.1, gain 11q13.4, del 13q14.2, del 16q24.3, and del 22q11.22) with significantly altered gene expression depicted in figure 3.30.

Gene expression of 308 of the 1087 genes at the 1q21.1 locus is significantly higher in samples harboring a gain 1q21.1 CNA ($n = 35$) compared to the samples without the gain. These represent only a sub-fraction of 28% of the genes located at 1q21.1 (median below the significance threshold of $p < 0.05$, see figure 3.30). The list of genes at 1q21 comprises several genes, previously had been suggested as "disease drivers" (see section 2.5.5), including ARID1A, FAM46C, CDKN2C, FUBP1, and NRAS.
In the 41 samples with a gain 11q13.4, gene expression of 3 (ARAP1, ATG16L2, STARD10) of the 4 genes in the gained region is significantly higher (FCHSD2 did not exceed the p-value threshold).
For the three deletions 13q14.2, 16q24.3 and 22q11.22, the gene expression of $77/1/1$ genes of total $93/2/1$ genes is significantly lower in $41/31/76$ samples. Regarding del 13q14.2, RB1 is among the significantly downregulated genes. In case of del 22q11 (comprising 4 genes), the only expressed gene is the super-enhancer *IGLL5*, showing significantly lower expression in samples with the deletion compared to samples without (see figure 3.30).

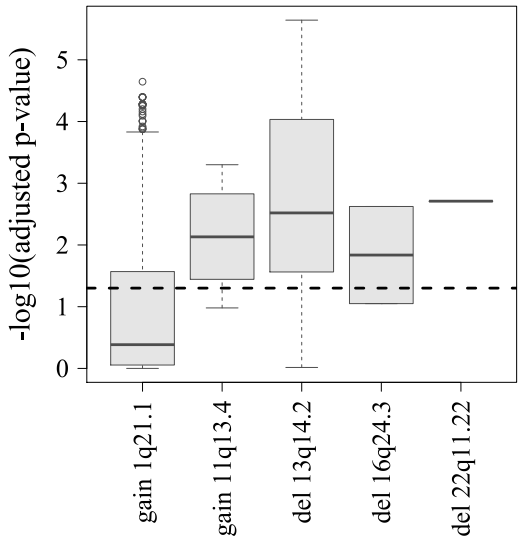

Figure 3.30: Gene expression in RNA sequencing data in relation to copy number alteration (CNA) in whole exome sequencing data of 113 light chain amyloidosis patients. The expression values of genes were tested for significant lower or higher expression in CNA altered samples compared to samples without the CNA. For four of the five depicted CNA, a subset of altered genes shows deregulated gene expression. Only in del 22q11.22 the only expressed gene *IGLL5* is significantly downregulated. This is indicated by a significant p-value in a Wilcoxon's rank sum test. For multiple testing correction the p-values were adjusted by the Benjamini-Hochberg (BH) method and plotted as negative log10 value at the y-axis. The dashed line depicts the significance threshold for a BH-adjusted p-value of 0.05.

3.6.3 Overlap between copy number alterations and chromosomal aberrations

Table 3.17: Comparison of copy number alterations (CNA) detected by whole exome sequencing (WES) to chromosomal aberrations (CA) detected by interphase fluorescence *in situ* hybridization (iFISH, iFISH probe at the same or a nearby cytoband to WES) in 113 light chain amyloidosis (AL) patients. *n*: total number of patient samples affected by CA or CNA, %: percentage of patient samples affected by the CA or CNA, Type: the type of alteration Gain or Del (deletion), Genes: number of genes in the respective CNA, Overlap: Number of patient samples being of same type in CNA and CA, i.e. showing both alterations or none, Rate: rate of efficient overlap between the CNA and the CA, NA: number of patient samples not classified by iFISH.

CA			CNA							
iFISH	***n***	**%**	**Cytoband**	***n***	**%**	**Type**	**Genes**	**Overlap**	**Rate**	**NA**
1q21	38	33.9	1q21.1	29	25.7	Gain	2	94	83.9	1
1q21	38	33.9	1q21.1	35	31.0	Gain	2699	92	82.1	1
1q21	38	33.9	1q21.1	37	32.7	Gain	1	94	83.9	1
4p16	2	6.2	4p16.1	11	9.7	Del	14	28	87.5	81
8p21	7	6.5	8p23.1	18	15.9	Del	3	91	85.0	6
9q34	37	32.7	9q21.11	36	31.9	Gain	8	92	81.4	0
11q13	7	6.2	11q13.4	41	36.3	Gain	6	79	69.9	0
13q14	46	40.7	13q11	38	33.6	Del	2	91	80.5	0
13q14	46	40.7	13q14.2	41	36.3	Del	274	92	81.4	0
14q32	40	35.4	14q11.2	96	85.0	Gain	1	47	41.6	0
14q32	12	10.6	14q32.33	95	84.1	Del	10	28	24.8	0
15q22	18	15.9	15q11.2	23	20.4	Gain	1	96	85.0	0
15q22	18	15.9	15q14	17	15.0	Gain	7	100	88.5	0
16q23	6	21.4	16q23.1	29	25.7	Del	1	20	71.4	85
16q23	6	21.4	16q24.3	31	27.4	Del	5	20	71.4	85
19q13	20	17.7	19q13.42	20	17.7	Gain	8	105	92.9	0

Of the 50 detected CNA in AL, sixteen CNA are spanning the same or a near cytoband as ten iFISH probes. A pairwise comparison of the samples between the CNA and the iFISH results was performed (for description see section 2.5.4). Table 3.17 depicts the comparison of CNA to the respective CA assessed by iFISH. Within the region of gain 1q21 (CA), three different CNA are found. The two smaller gain 1q21 CNA, with 1 and 2 genes, lie within the gain 1q21 CNA encompassing 2699 genes. An efficient overlap of $82-84\%$ exists for all three gain 1q21 CNA with the respective CA detected by iFISH. High overlap rates are present for del 8p23.1 to 8p21 (85%), gain 9q21.11 to 9q34 (81%), del 13q11 (81%) and 13q13.2 (81%) to 13q14, gain 15q11.2 (85%) and 15q14 (89%) to 15q22, and with the highest overlap rate regarding gain 19q13.42 to 19q13 with 93% (see table 3.17). Gain of 11q13 CNA is less consistent to the iFISH detected CA (70%). The CA 14q32 compared to gain 14q11.2 and del 14q32.33 CNA show low rates of overlap of 42% and 25% (see table 3.17). For the three CNA del 4p16.1, del 16q23.1 and del 16.24.3 the respective iFISH probe is not determined in 81/85/85 patient samples (see table 3.17).

3.6.4 Entity specific alterations of gene expression in malignant plasma cells - similarities and differences between malignant plasma cell populations and comparator populations

Table 3.18: RV coefficient of gene expression data from DNA microarray AL: light chain amyloidosis, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma, GEP: gene expression profiling by DNA microarray.

Entity	MGUS	AMM	MM	BMPC	HMCL	MBC	PPC
AL	0.55	0.63	0.72	0.22	0.15	0.07	0.10
MGUS		0.76	0.61	0.14	0.13	0.05	0.08
AMM			0.74	0.17	0.15	0.06	0.09
MM				0.25	0.21	0.11	0.15
BMPC					0.07	0.04	0.06
HMCL						0.03	0.06
MBC							0.08

Next, entity specific alterations of gene expression in malignant plasma cells were assessed and similarities and differences between malignant plasma cell populations and comparator populations investigated. This was performed to assess the degree of similarity between different malignant plasma cell populations, i.e. between AL, MGUS, AMM, and MM. The similarity was then compared to other cell populations, i.e. BMPC, MBC (non-malignant, non-proliferating), PPC (non-malignant, proliferating), and HMCL (malignant, proliferating). To do so, two dimension reduction methods, i.e. PCA and t-SNE and a method to receive a correlation coefficient, i.e. RV-coefficient, were applied (see section 2.5.1 for a description of methodology).
As depicted by the RV-coefficient (see table 3.18) malignant plasma cell entities show a greater similarity among each other ($0.55-0.76$) than to BMPC ($0.14-0.25$), MBC ($0.05-0.11$), PPC ($0.08-0.15$), or HMCL ($0.13-0.21$). Figures 3.31 and 3.33 depict PCA and t-SNE for malignant plasma cell populations only (i.e., without comparator populations). Using both dimension reduction methods, the center of gravity of all four malignant plasma cell populations overlaps (yellow, red, green, and purple colored circles at the right in figure 3.31 for DNA microarray, 3.33 for RNA seq). Comparing this variance to other populations, in both analyses, MBC, PPC, HMCL, and BMPC are distinct from malignant plasma cells (see figure 3.32 for DNA microarray, 3.34 for RNA seq). In all analyses, BMPC are distinct from all other populations (see the center of gravity circles at the right in figure 3.32 for DNA microarray, 3.34 for RNA seq). In PCA, BMPC are a distinct entity closest neighboring malignant plasma cell populations; in t-SNE, they are located between MBC as normal precursor-counterpart of BMPC, and malignant plasma cell populations, as in PCA. In each case, malignant and non-malignant proliferating populations, i.e. PPC and HMCL, respectively, are

grouped separately, but most closely together (see the center of gravity circles at the right in figure 3.32 for DNA microarray, 3.34 for RNA seq).
Quantitatively, the calculated explained variance in the PCA without B-cell lineage (MBC, PPC, BMPC) and HMCL is 5.7% for principal component 1 (PC1) and 4.4% for PC2 (figure 3.31 **a**). For PCA with these comparator populations, PC1 remains 5.7% and PC2 is 4.3% (figure 3.32 **a**). In RNA seq, PC1 and PC2 comprise more explained variance then in DNA microarrays and are as similar as with (PC1 = 16.3%, PC2 = 2.8%, figure 3.34 **a**) and without (PC1 = 17%, PC2 = 2.5%, figure 3.33 **a**) comparator populations.

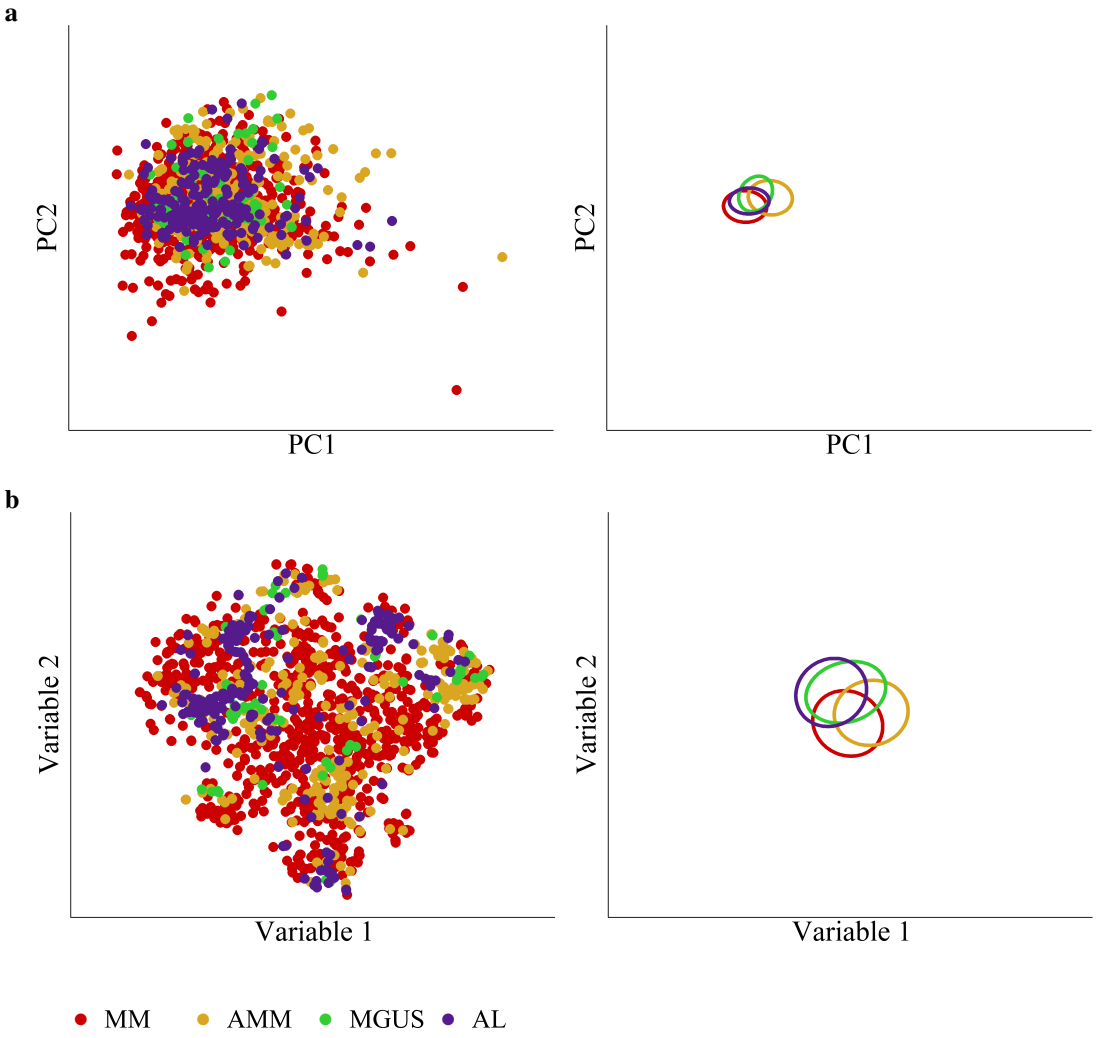

Figure 3.31: Similarities and differences in gene expression between and within different malignant plasma cell disease entities. Dimension reduction of gene expression data from DNA microarrays using **a** Principal component analysis (PCA) and **b** t-distributed stochastic neighbor embedding (t-SNE). Left side: individual data points, right side: center of each group is depicted as ellipse of 10% variance around the group's mean value. Groups are color coded, see legend below the figure. Patients samples: light chain amyloidosis (AL), monoclonal gammopathy of undetermined significance (MGUS), asymptomatic multiple myeloma (AMM), or symptomatic multiple myeloma (MM). PC: principal component

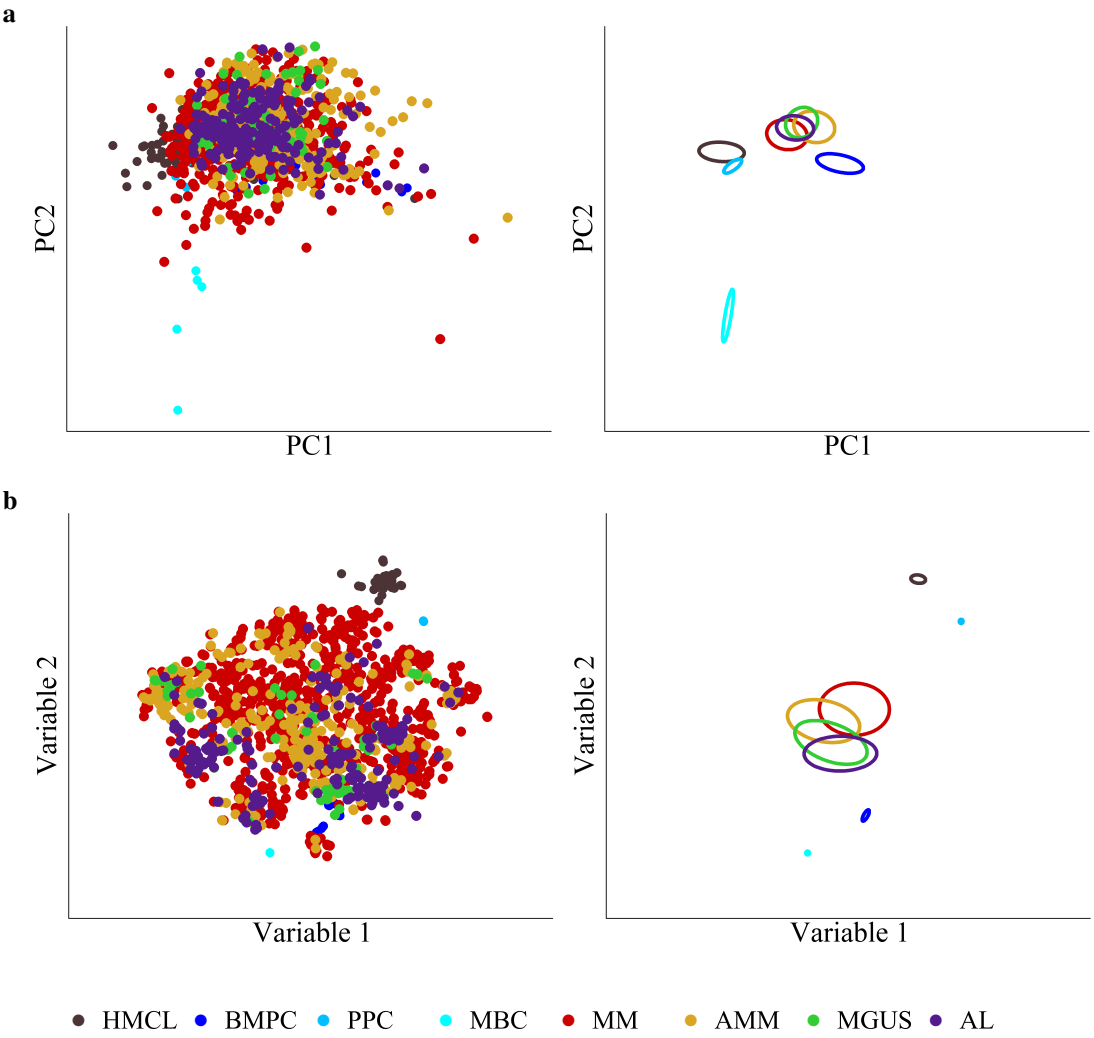

Figure 3.32: Similarities and differences in gene expression between and within different malignant plasma cell disease entities, comparator populations, and within these. Dimension reduction of gene expression data from DNA microarrays using **a** Principal component analysis (PCA) and **b** t-distributed stochastic neighbor embedding (t-SNE). Left side: individual data points, right side: center of each group is depicted as ellipse of 10% variance around the group's mean value. Groups are color coded, see legend below the figure. Patients samples: light chain amyloidosis (AL), monoclonal gammopathy of undetermined significance (MGUS), asymptomatic multiple myeloma (AMM) or symptomatic multiple myeloma (MM). For comparison, memory B cells (MBC), polyclonal plasmablastic cells (PPC) and healthy normal donor bone marrow plasma cells (BMPC), human myeloma cell lines (HMCL) are depicted. PC: principal component

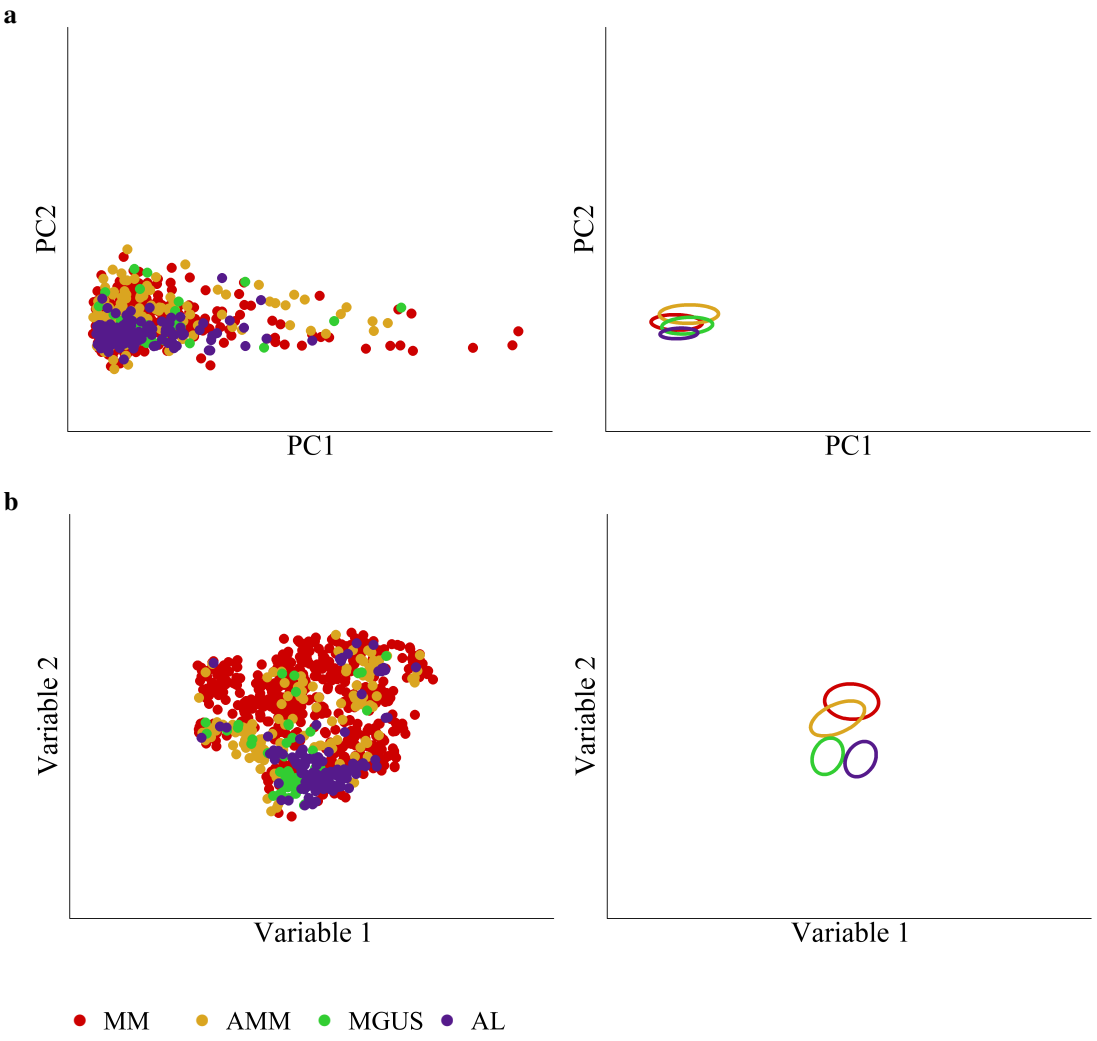

Figure 3.33: Similarities and differences in gene expression between and within different malignant plasma cell disease entities. Dimension reduction of gene expression data from RNA sequencing using **a** Principal component analysis (PCA) and **b** t-distributed stochastic neighbor embedding (t-SNE). Left side: individual data points, right side: center of each group is depicted as ellipse of 10% variance around the group's mean value. Groups are color coded, see legend below the figure. Patients samples: light chain amyloidosis (AL), monoclonal gammopathy of undetermined significance (MGUS), asymptomatic multiple myeloma (AMM) or symptomatic multiple myeloma (MM). PC: principal component

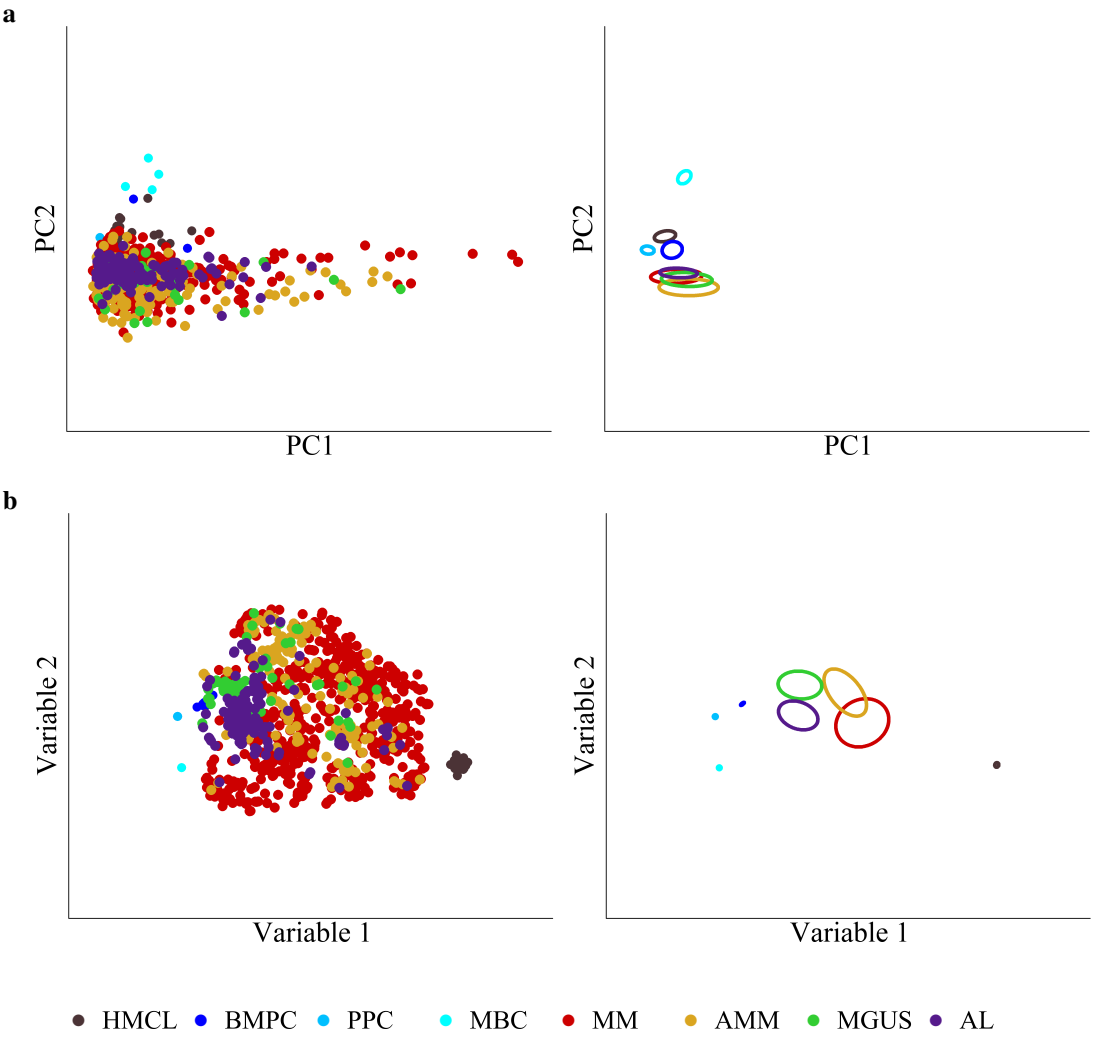

Figure 3.34: Similarities and differences in gene expression between and within different malignant plasma cell disease entities, comparator populations, and within these. Dimension reduction of gene expression data from RNA sequencing using **a** Principal component analysis (PCA) and **b** t-distributed stochastic neighbor embedding (t-SNE). Left side: individual data points, right side: center of each group is depicted as ellipse of 10% variance around the group's mean value. Groups are color coded, see legend below the figure. Patients samples: light chain amyloidosis (AL), monoclonal gammopathy of undetermined significance (MGUS), asymptomatic multiple myeloma (AMM) or symptomatic multiple myeloma (MM). For comparison, memory B cells (MBC), polyclonal plasmablastic cells (PPC) and healthy normal donor bone marrow plasma cells (BMPC), human myeloma cell lines (HMCL) are depicted. PC: principal component

3.6.5 Differential gene expression

Whereas dimension reduction methods were applied to gain a general impression and quantification of similarities and differences of populations, differential gene expression was assessed to obtain first a numerical estimation of DEG, and secondly to list these genes in detail. Differential expression was assessed by RNA seq as most comprehensive method. Seven differential gene expression analyses were performed (see table 2.9 in section 2.5.2 for an overview). Numbers of significantly DEG per analysis are summarized in table 3.19 and table 3.20. Each list of DEG was subjected to a p-value threshold of ≤ 0.05 and a LFC > 1 (2-fold difference between groups, see tables 3.19, 3.20, second row).

Table 3.19: Differentially expressed genes (DEG) between normal and malignant plasma cell samples of each of the four investigated entities AL, MGUS, AMM, and MM. Depicted are number (*n*) and percentage (%) of DEG per comparison and subset group. DE: differentially expressed; DOWN: downregulated compared to first in comparison; UP: upregulated compared to first in comparison. First row: all significantly DEG per comparison with an adjusted p-value $p \leq 0.05$; second row: applied log fold change (LFC) of above 1, as a subset of the first row; third row: immunoglobulin (Ig) genes as a subset of genes of the second row; fourth row: genes with ENTREZID as a subset of second row; fifth row: protein coding genes as a subset of second row. See figure 3.35 **a** for overlap of DE genes between the comparisons. BMPC: bone marrow plasma cells, AL: light chain amyloidosis, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

		Number of genes					
		DE		**DOWN**		**UP**	
Comparison	**Group**	***n***	**%**	***n***	**%**	***n***	**%**
BMPC *vs.* AL	$p \leq 0.05$	2466	9	1310	53.1	1156	46.9
	lfc >1	106	4.3	49	46.2	57	53.8
	Ig genes	8	7.5	8	100	0	0
	ENTREZID	71	67	23	32.4	48	67.6
	protein coding	67	63.2	19	28.4	48	71.6
BMPC *vs.* MGUS	$p \leq 0.05$	2391	8.7	1024	42.8	1367	57.2
	lfc >1	55	2.3	26	47.3	29	52.7
	Ig genes	1	1.8	1	100	0	0
	ENTREZID	44	80	19	43.2	25	56.8
	protein coding	42	76.4	18	42.9	24	57.1
BMPC *vs.* AMM	$p \leq 0.05$	5834	21.3	2495	42.8	3339	57.2
	lfc >1	551	9.4	222	40.3	329	59.7
	Ig genes	34	6.2	34	100	0	0
	ENTREZID	374	67.9	132	35.3	242	64.7
	protein coding	350	63.5	121	34.6	229	65.4
BMPC *vs.* MM	$p \leq 0.05$	5327	19.5	1776	33.3	3551	66.7
	lfc >1	529	9.9	144	27.2	385	72.8
	Ig genes	49	9.3	49	100	0	0
	ENTREZID	314	59.4	66	21	248	79
	protein coding	284	53.7	54	19	230	81

Table 3.20: Differentially expressed genes (DEG) between malignant plasma cell samples from AL patients and of each of the three investigated malignant plasma cell disease entities, i.e. MGUS, AMM, and MM. Depicted are number (*n*) and percentage (%) of DEG per comparison and subset group. DE: differentially expressed; DOWN: downregulated compared to first in comparison; UP: upregulated compared to first in comparison. First row: all significantly DEG per comparison with an adjusted p-value $p \leq 0.05$; second row: applied log fold change (LFC) of above 1, as a subset of the first row; third row: immunoglobulin (Ig) genes as a subset of genes of the second row; fourth row: genes with ENTREZID as a subset of second row; fifth row: protein coding genes as a subset of second row. See figure 3.35 **b** for overlap of DE genes between the comparisons. AL: light chain amyloidosis, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

		Number of genes					
		DE		**DOWN**		**UP**	
Comparison	**Group**	***n***	**%**	***n***	**%**	***n***	**%**
AL *vs.* MGUS	$p \leq 0.05$	3051	11.2	1497	49.1	1554	50.9
	lfc >1	71	2.3	22	31	49	69
	Ig genes	20	28.2	4	20	16	80
	ENTREZID	35	49.3	9	25.7	26	74.3
	protein coding	31	43.7	8	25.8	23	74.2
AL *vs.* AMM	$p \leq 0.05$	9062	33.1	3535	39	5527	61
	lfc >1	484	5.3	263	54.3	221	45.7
	Ig genes	55	11.4	18	32.7	37	67.3
	ENTREZID	366	75.6	232	63.4	134	36.6
	protein coding	348	71.9	231	66.4	117	33.6
AL *vs.* MM	$p \leq 0.05$	11017	40.3	3546	32.2	7471	67.8
	lfc >1	825	7.5	100	12.1	725	87.9
	Ig genes	66	8	36	54.5	30	45.5
	ENTREZID	421	51	57	13.5	364	86.5
	protein coding	364	44.1	54	14.8	310	85.2

Overlap of differentially expressed genes

The lists of DEG (see tables 3.19, 3.20) were subsequently analyzed regarding the intersections of DEG between the different comparisons. The Venn diagrams in figure 3.35 depict the numbers of overlapping DEG with a LFC > 1. In the supplemental tables A.13, A.14 and A.15 lists of overlapping DEG are depicted.

Between BMPC and the four malignant plasma cell disease entities, i.e. AL, MGUS, AMM, and MM, 31 DEG overlap (figure 3.35 **a**). Of these, 26 genes are up- and 5 genes downregulated in malignant plasma cell diseases compared to BMPC (see supplementary table A.15 for genes, LFC, and expression height). The direction (higher or lower) of alteration of expression is the same for all analyzed malignant plasma cell diseases.

The same holds true for the 80 genes overlapping between the comparisons BMPC *versus* AL and BMPC *versus* MM, 57 are up- and 23 downregulated (see supplementary table A.15). Of these, 32 genes are likewise overlapping for MGUS, and 73 in the comparison to AMM (see supplementary table A.15). Nine genes are unique to the comparison BMPC *versus* AL (*IGHD6-19*, *H1FX-AS1*, *PI4KAP1*, *AC159540.1*, *RP11-58H15.1*, *HSH2D*, *RPL35AP32*, *IGHD7-27*, and *NFKBID*), two of which genes are protein coding, i.e. *HSH2D* and *NFKBID* (see supplemental table A.13).

Thirteen genes show differential expression between AL *versus* each of the other malignant plasma cell diseases MGUS, AMM, and MM, i.e. *HES1*, *FOLH1*, *RASD1*, *IGHD2-8*, *BARX2*, *PAGE1*, *HTR1D*, *SCARNA22*, *RP11-66N7.2*, *SSTR1*, *IGHV1OR15-2*, *IGLV6-57*, and *HBE1*. Eight are protein coding genes, three are Ig genes, one is in antisense and one is scaRNA (see supplementary table A.14). Of these, six genes are higher expressed and seven are lower expressed in AL (see figure 3.35 **b**).

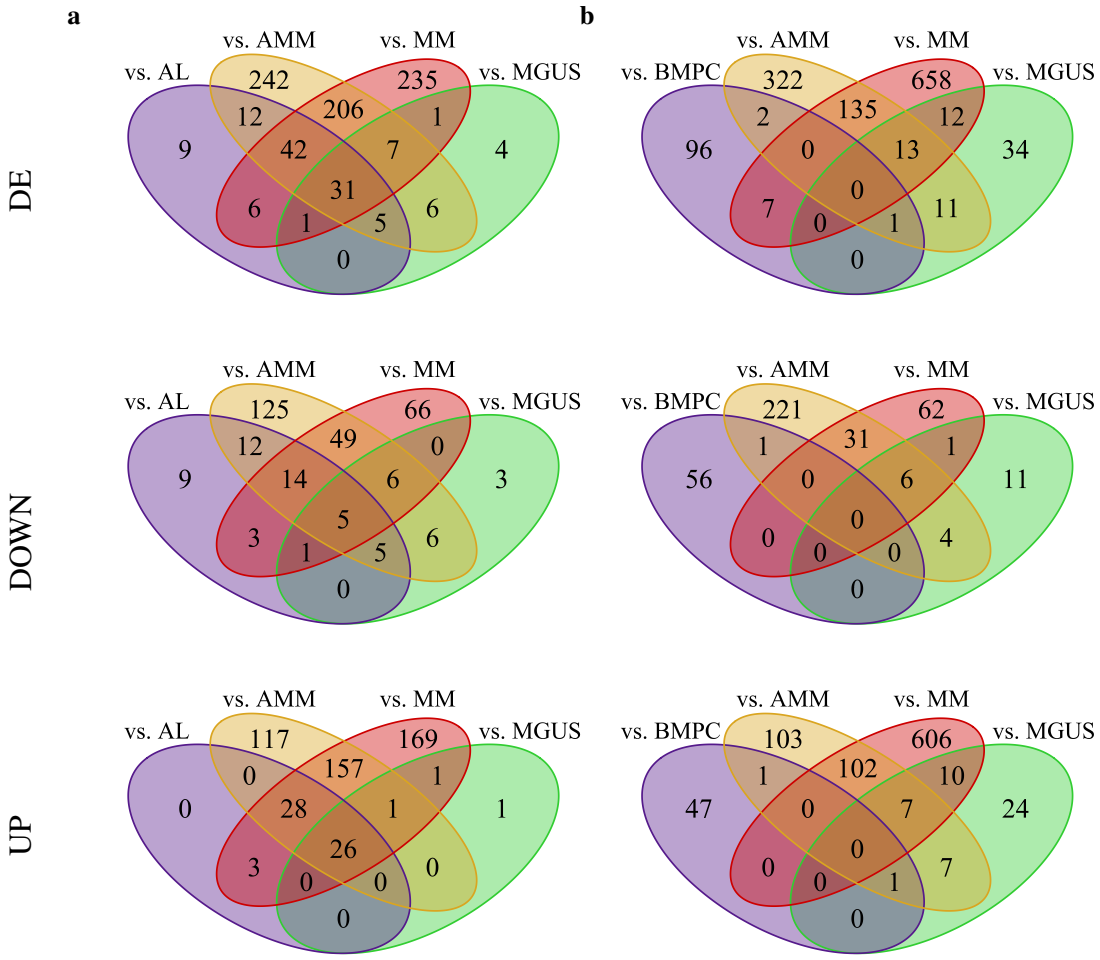

Figure 3.35: Overlap of significant differentially expressed genes (DEG) with a log fold change >1 from differential expression analyses on RNA sequencing data. **a** BMPC *versus* entities AL, MGUS, AMM and MM and **b** AL *versus* BMPC, MGUS, AMM and MM. The top Venn diagram shows all differentially expressed (DE) genes, in the middle are only downregulated (DOWN) genes and the bottom depicts only upregulated (UP) genes. BMPC: bone marrow plasma cells, AL: light chain amyloidosis, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

Comparison to published differential gene expression analyses

For the comparison of the DEG to previous published DEG, gene lists of Abraham *et al.* [1] and Kryukov *et al.* [160] for the analysis of AL *versus* MM, and from Paiva *et al.* [230] for the AL *versus* BMPC comparison were available.

Table 3.21: Overlap of DEG between AL *versus* MM with DEG detected by Abraham *et al.* [1] and Kryukov *et al.* [160]. Negative LFC indicates a lower expression in AL compared to MM. DEG: differentially expressed genes, adj. *p*: adjusted p-value, LFC: log fold change, BMPC: bone marrow plasma cell, AL: light chain amyloidosis, MM: multiple myeloma.

	adj.		log counts per million			
Gene	*p*	**LFC**	**BMPC**	**AL**	**MM**	**Citation**
JUN	< .001	2.01	8.54	8.29	6.29	Abraham *et al.* [1]
CXCL12	.003	1.77	7.76	8.33	6.57	Abraham *et al.* [1]
APOE	.01	1.66	1.54	2.05	0.99	Abraham *et al.* [1]
MYC	< .001	-1.71	7.01	6.16	7.85	Abraham *et al.* [1]
TXN	< .001	-1.34	6.28	6.98	8.31	Abraham *et al.* [1]
RPS21	< .001	-1.36	5.47	5.90	7.25	Kryukov *et al.* [160]
RPL28	.005	-1.28	6.53	6.95	8.22	Kryukov *et al.* [160]
RPL35	.01	-1.25	7.18	7.96	9.20	Kryukov *et al.* [160]

From the 47 DEG of Abraham *et al.* [1], 32 are differentially expressed with a subset of 5 genes having a LFC >1 and being altered in the same direction (i.e. increased or decreased expression in both analyses, respectively). *MYC* and *TXN* are lower expressed, *JUN*, *CXCL12* and *APOE* significantly higher expressed in AL. Of 100 DEG described by Kryukov *et al.* [160], 72 genes overlap. Of these, the three ribosomal protein encoding genes *RPS21*, *RPS28* and *RPS35* are significantly differentially expressed by a LFC > 1 in AL *versus* MM. Four of the 60 genes that had been detected as differentially expressed between AL and BMPC by Paiva *et al.* [230] were detected as differentially expressed in the comparison AL *versus* BMPC, none of them with a LFC > 1. Table 3.21 depicts the 8 overlapping genes with a LFC > 1. Ninety-six overlapping DEG show a LFC < 1, with 29 genes demonstrating a opposite direction compared to the analysis depicted above.

Pathway analysis of differentially expressed genes

Using the described RNA seq data, for two different gene sets of DEG a FEA was performed (described in section 2.5.6): First, for all common DEG between BMPC *versus* AL and BMPC *versus* MM, which were annotated to an ENTREZID (DEG $n = 70$) (see supplementary table A.15). Ten terms, all belonging to GO "biological process" (BP) "origin", are significantly enriched in these genes (see supplementary table A.16).

Second, the four gene lists with all DEG between BMPC *versus* AL (DEG $n = 73$),

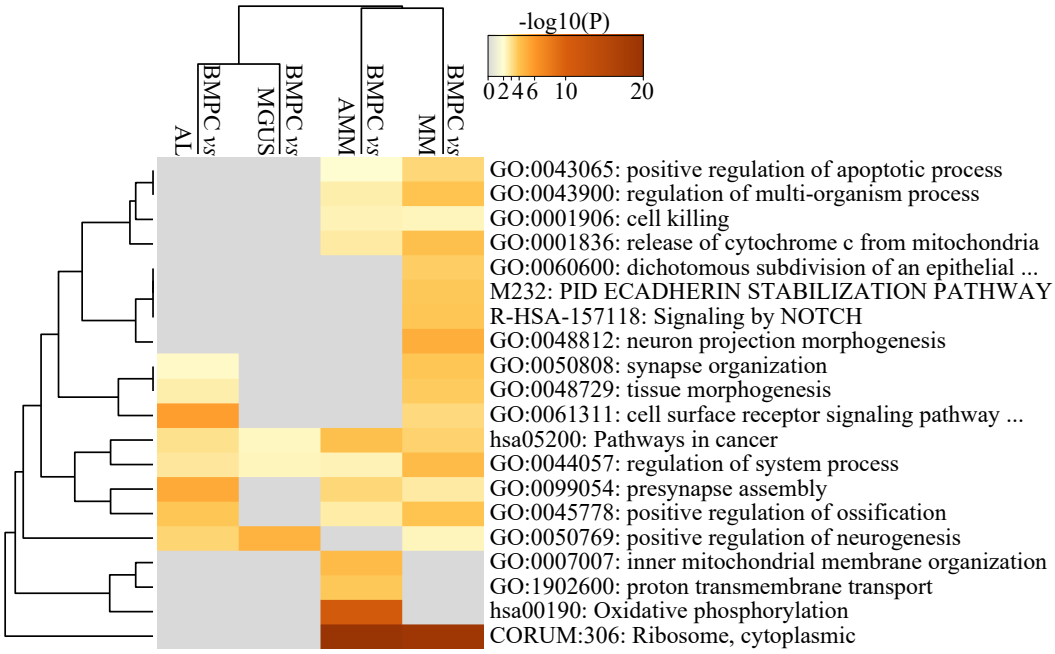

Figure 3.36: Functional enrichment analyses of differentially expressed genes (DEG) between normal bone marrow plasma cells (BMPC) and each of the malignant plasma cell populations, i.e. BMPC *versus* AL, BMPC *versus* MGUS, BMPC *versus* AMM and BMPC *versus* MM. Depicted are the top 20 terms (rows) enriched in the individual DEG lists (columns). For better representation the negative log10 value of the adjusted p-value is depicted. The highest value presents the lowest p-value. A value of 2 equates to an adjusted p-value of 0.01. BMPC: bone marrow plasma cells, AL: light chain amyloidosis, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma

Table 3.22: Two enriched terms and the corresponding genes overlapping in the four DEG lists as detected by functional enrichment analysis. Terms are GO:0044057 from GO (gene ontology) and hsa05200 from KEGG (Kyoto Encyclopedia of Genes and Genomes)

Term	Description	Genes
GO:0044057	regulation of system process	*ADM, CYP2J2, TYMP, EDNRB, HGF, IGF1, KIT, PRKG1, RELN, TGFB2, HOMER1, CALCRL, KCNMB2, SEMA3A, BVES, HEY2, NLGN4X, MLIP, MTG1, RNF207, RHOC, BDKRB1, BMP4, CTSH, MAP2, RPS19, CCL7, WNT5A, RTN4, SEMA3D, RPLP1, TXN, UBA52, MAPK8IP3, GPRC5D, NOD2, PTHLH, RASGRF1, TNFSF10, DKK1, INHBE, UTS2B*
hsa05200	pathways in cancer	*CCND1, BMP4, CKS2, COL4A3, EDNRB, HGF, IGF1, JUN, KIT, LAMA5, NFKB2, PLCB4, PTEN, ELOB, TGFB2, WNT5A, LPAR5, RELN, RASGRF1, VAV3, PIM1, FOSB, HLA-A, COL4A5*

BMPC *versus* MGUS (DEG $n = 44$), BMPC *versus* AMM (DEG $n = 383$) and BMPC *versus* MM (DEG $n = 317$) were analyzed. The result of the enrichment analysis is depicted by a heatmap of the top twenty enriched terms for crossing all four gene lists (see figure 3.36). Enriched terms include GO BP terms, KEGG and Reactome pathway terms.

Two terms are enriched in all four DEG lists: "Regulation of system process" (GO:0044057) and "pathways in cancer" (hsa05200). The respective genes enriched in these terms are listed in table 3.22. The BMPC *versus* MGUS comparison DEG list shares three top enriched terms with the other three DEG lists. The BMPC *versus* AL DEG list shares eight terms with BMPC *versus* MM and four with BMPC *versus* AMM (see figure 3.36).

3.6.6 Immunoglobulin gene expression by RNA sequencing

Gene expression of Ig genes was evaluated as described in section 2.5.3. Figure 3.37 shows a heatmap of the expression values of 383 Ig genes. The largest cluster comprises Ig genes which are rarely or not at all expressed (cluster A, 267 genes, see table 3.23 **a**)). A cluster of highly expressed Ig genes (cluster C, Ig HC and IG LC λ, see table 3.23 **a**) can be detected. This cluster overlaps with cluster 2 of patient samples, showing higher expression in more Ig genes than the other patient samples. Cluster 2 refers to 137 samples of patients and harbors the largest group of AL with 35% (see table 3.23 **d** and **e**).

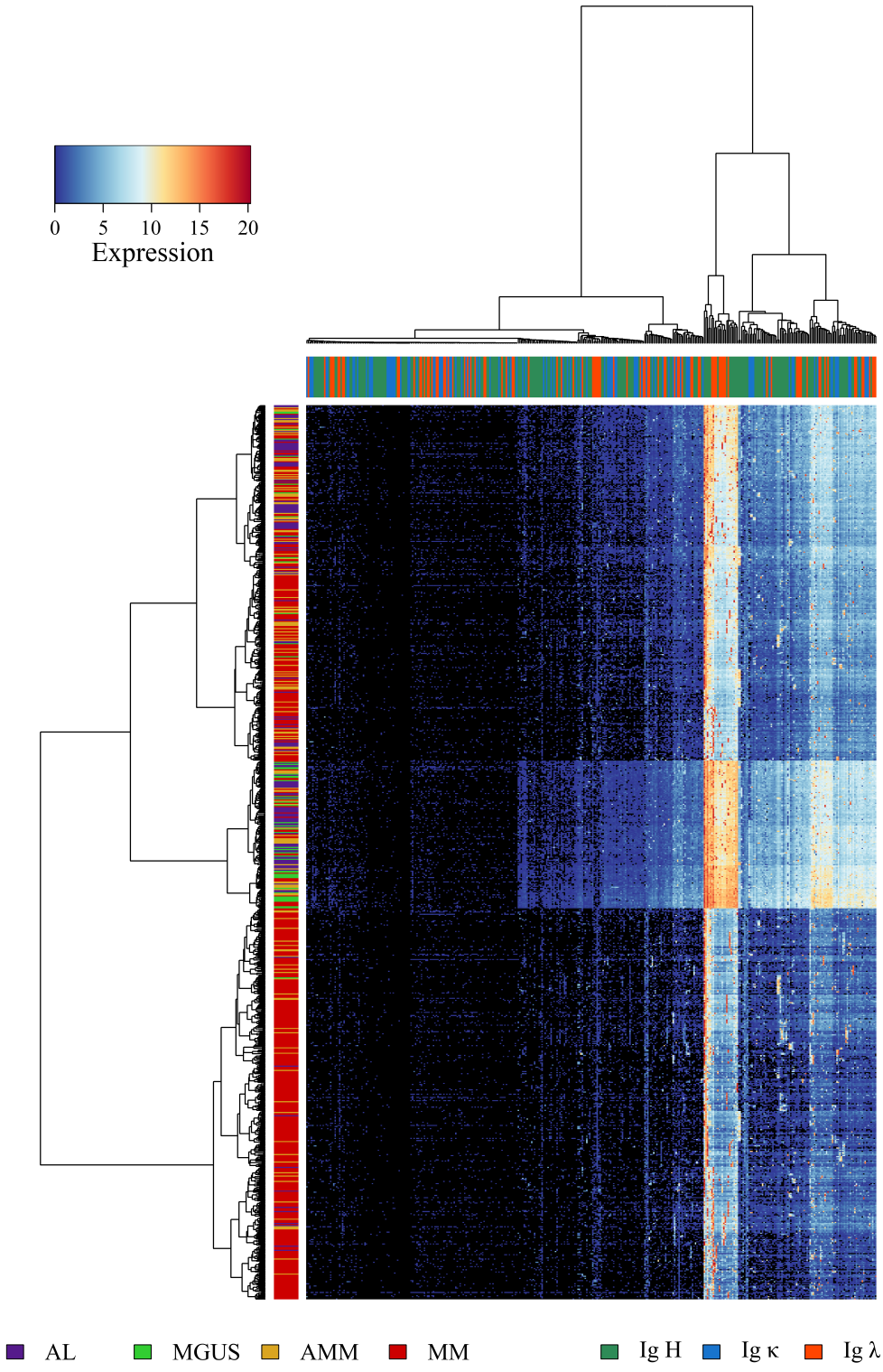

Figure 3.37: Gene expression of immunoglobulin (Ig) genes in different entities analyzed by unsupervised Ward clustering. Entities are in rows and genes in columns. The legend below depicts the disease entities AL: light chain amyloidosis, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma, and the Ig gene groups Ig H: Ig heavy chain gene, Ig κ: Ig light chain gene of type κ, Ig λ: Ig light chain gene of type λ. RNA expression intensity is color coded (upper left corner, expression height of 0 is depicted in black, expression values increase from blue to red color).

Table 3.23: Allocation per cluster for immunoglobulin (Ig) genes and patient samples per entity **a** Total number of Ig genes per cluster **b** Percentage of Ig genes per cluster **c** Percentage of Ig genes per Ig group **d** Total number of patient samples **e** Percentage of patient samples per cluster **f** Percentage of patient samples per entity. The disease entities AL: light chain amyloidosis, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma, and the Ig gene groups Ig H: Ig heavy chain gene, Ig κ: Ig light chain gene of type κ, Ig λ: Ig light chain gene of type λ.

a

Cluster	Ig H	Ig LC κ	Ig LC λ
A	125	79	63
B	32	9	7
C	13	0	10
D	20	13	12

b

Cluster	Ig H	Ig LC κ	Ig LC λ
A	46.8	29.6	23.6
B	66.7	18.8	14.6
C	56.5	0	43.5
D	44.4	28.9	26.7

c

Cluster	Ig H	Ig LC κ	Ig LC λ
A	65.8	78.2	68.5
B	16.8	8.9	7.6
C	6.8	0	10.9
D	10.5	12.9	13

d

Cluster	AL	MGUS	AMM	MM
1	11	1	36	315
2	48	36	25	28
3	18	3	42	115
4	47	11	37	57

e

Cluster	AL	MGUS	AMM	MM
1	3	0.3	9.9	86.8
2	35	26.3	18.2	20.4
3	10.1	1.7	23.6	64.6
4	30.9	7.2	24.3	37.5

f

Cluster	AL	MGUS	AMM	MM
1	8.9	2.0	25.7	61.2
2	38.7	70.6	17.9	5.4
3	14.5	5.9	30	22.3
4	37.9	21.6	26.4	11.1

3.6.7 SNVs and InDels in light chain amyloidosis

The following section describes the analysis of the variants detected by the variant calling pipeline outlined in section 2.4.4. Variants detected in 113 AL patient samples were analyzed in comparison to variants detected in MM as well as previously published studies.

Summary statistics

The median number of variants in AL is 22 (range: $1-3785$) per sample, being similar in ALMG (22, range: $1-3785$) and ALMM (23, range: $7-3656$). In terms of the number of detected variants, two patterns can be distinguished: the majority of patients (94%) shows between 1 and 938 variants, whereas 7 patient samples (2 with ALMM, 5 with ALMG) contrasts from the rest of the group with a median of 3609 variants per sample (see figure 3.38, figure 3.39).

To put this into perspective, the number of variants per sample of the 113 AL patient samples was compared to the 930 samples from the CoMMpass cohort (newly diagnosed MM patients). The median number of variants per sample in AL is significantly lower ($p < 0.001$, figure 3.39), the maximum number of observed variants in a sample is however higher in AL (see table 3.2, 3.3 and figure 3.39, 3.38). The heterogeneity in the number of variants, which is indicated by the wider peak in the

density distribution in figure 3.38, the variance of $754,962/35,178$ *versus* $9,048/2,418$ for non-synonymous/expressed variants (F test, both $p < 0.001$), and the range (see table 3.2, 3.3), is much larger in AL than in MM .

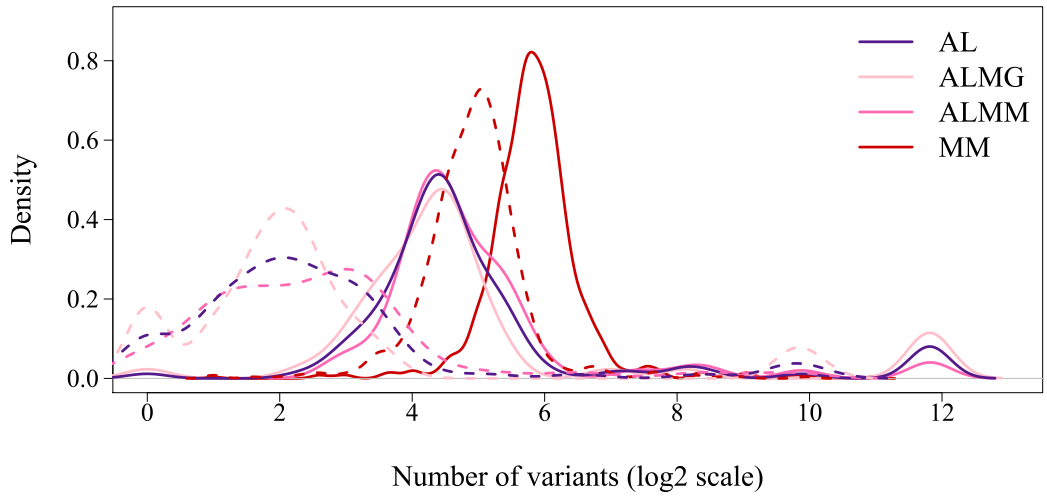

Figure 3.38: Distribution of variants per patient sample in AL (n=113), splitted by underlying subentity ALMG (n=51) and ALMM (n=62), and MM (n=930, from CoMMpass cohort). Variants in immunoglobulin (Ig) genes were excluded and analyzed separately, see last part in section 3.6.7. Depicted on the x-axis are the number of variants per sample after log2 transformation. Solid lines: non-synonymous variants, dashed lines: expressed variants. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, MM: multiple myeloma.

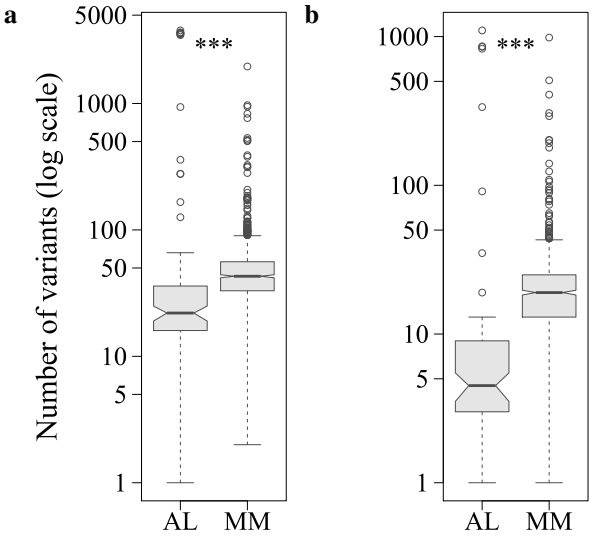

Figure 3.39: Numbers of variants per patient sample in AL (n=113) and MM (n=930). Variants in immunoglobulin (Ig) genes were excluded and analyzed separately, see section 3.6.7. **a** non-synonymous variants, **b** expressed variants. Data on the y-axis is shown on a log scale. Significant differences, indicated by Wilcoxon's rank sum test, are illustrated as ***, representing a significant p-value < .001. AL: light chain amyloidosis, MM: multiple myeloma

Transitions and transversions

The composition of the SNV depicted in the variant tables was analyzed by type of base substitution and frequency of Ti to Tv, called Ti-Tv bias (see figure 3.40). The Ti-Tv frequencies are comparable in AL (see figure 3.40 **a**) and MM (see figure 3.40 **b**), with a median of 60% transitions in AL and 59% in MM, representing a median Ti-Tv ratio of 1.5 in AL and 1.4 in MM. The base substitution C>T, with a median of 43.4% in AL and 45.2% in MM, is the most frequent base substitution and T>A is the least frequent with a median of 5.9% in AL and 7.5% in MM.

Variant annotations

Most of the detected coding, non-synonymous variants are missense mutations ($\sim$ 95%), in AL as well as in MM, see table 3.24 **a**. In AL, these are associated with a moderate impact ($>$ 94%, see table 3.24 **b**), a tolerated SIFT score ($>$ 61%, see table 3.24 **c**) or are classified as benign by PolyPhen ($>$ 67%, see table 3.24 **d**). Most (110/112) of the 113 AL patients have at least one deleterious/damaging variant classified by SIFT/PolyPhen. For both the median is 9 variants per patient in comparably equal ranges of $1-762$ for SIFT and $1-750$ for PolyPhen.

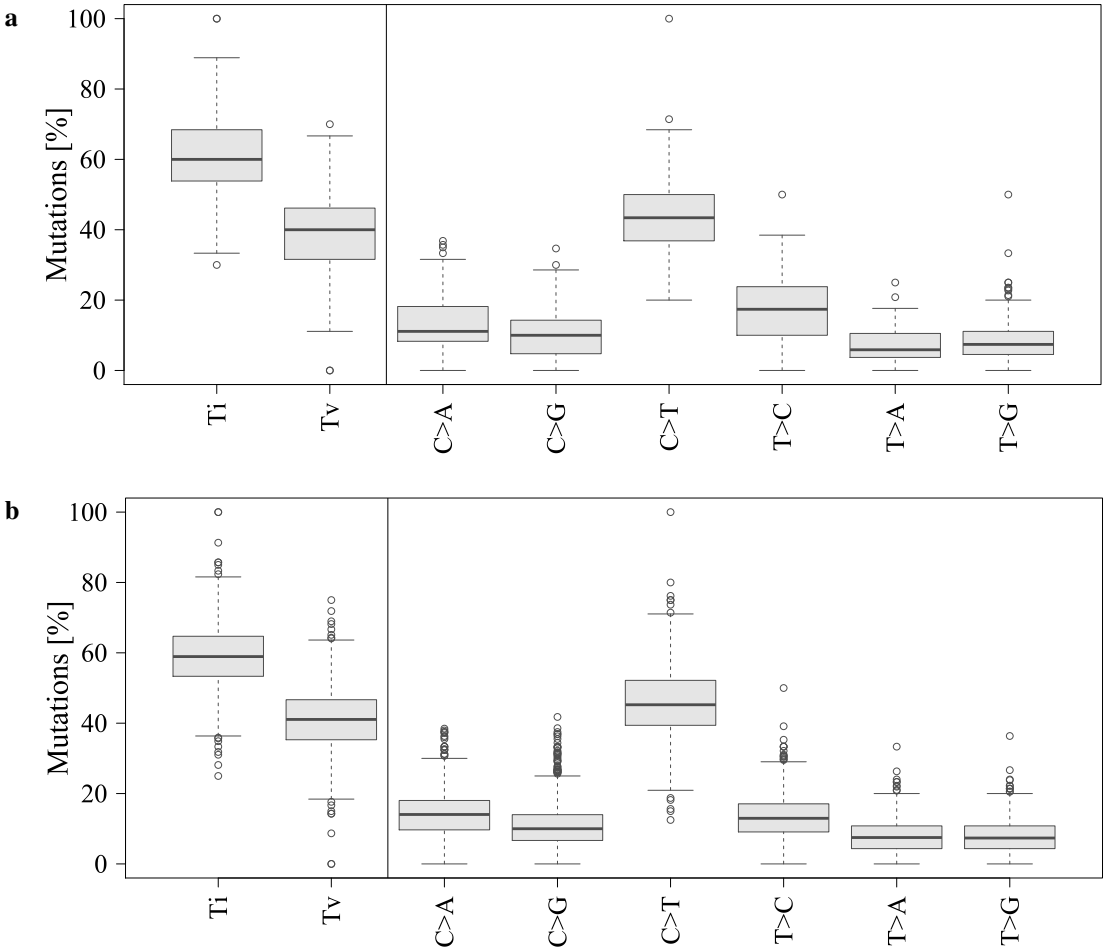

Figure 3.40: Number of transition (Ti), transversion (Tv) and the respective base substitution possibilities (C>A, C>G, C>T, T>C, T>A, T>G) of coding, non-synonymous single nucleotide variants for **a** 113 light chain amyloidosis (AL) samples and **b** 930 multiple myeloma (MM) samples. Percentage of mutations is depicted on the y-axis.

Table 3.24: Predicted coding consequences for variants detected by variant calling. Assessment of variants by different impact prediction algorithms and databases, in total numbers (*n*) and percentages (%) for **a** coding consequence by Sequence Ontology in 113 AL and MM 930, **b** impact prediction by vep (variant effect predictor), **c** SIFT (Sorting Intolerant From Tolerant) in 113 AL and **d** PolyPhen (Polymorphism Phenotyping) in 113 AL. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, MM: multiple myeloma.

a

Consequence	**AL**	**ALMG**	**ALMM**	**MM**
n	29681	19276	10405	52355
Missense	28328 (95.4%)	18393 (95.4%)	9935 (95.5%)	49499 (94.5%)
Frame Shift Del	185 (0.6%)	121 (0.6%)	64 (0.6%)	920 (1.8%)
Frame Shift Ins	171 (0.6%)	115 (0.6%)	56 (0.5%)	493 (0.9%)
In Frame Del	291 (1%)	186 (1%)	105 (1%)	248 (0.5%)
In Frame Ins	150 (0.5%)	95 (0.5%)	55 (0.5%)	49 (0.1%)
Splice Site	556 (1.9%)	366 (1.9%)	190 (1.8%)	1146 (2.2%)

b

Impact	**AL**	**ALMG**	**ALMM**
n	29681	19276	10405
Low	553 (1.9%)	365 (1.9%)	188 (1.8%)
Moderate	28160 (94.9%)	18320 (95%)	9840 (94.6%)
High	968 (3.2%)	591 (3.1%)	377 (3.6%)

c

SIFT	**AL**	**ALMG**	**ALMM**
n	29681	19276	10405
-	2838 (9.6%)	1821 (9.5%)	1017 (9.8%)
tolerated	18796 (63.3%)	12456 (64.6%)	6340 (60.9%)
tolerated low confidence	3010 (10.1%)	1986 (10.3%)	1024 (9.8%)
deleterious low confidence	1658 (5.6%)	1018 (5.3%)	640 (6.2%)
deleterious	8047 (27.1%)	4999 (25.9%)	3048 (29.3%)

d

PolyPhen	**AL**	**ALMG**	**ALMM**
n	29681	19276	10405
-	1925 (6.5%)	1221 (6.3%)	704 (6.8%)
benign	20381 (68.7%)	13432 (69.7%)	6949 (66.8%)
possibly damaging	3075 (10.4%)	1950 (10.1%)	1125 (10.8%)
probably damaging	3327 (11.2%)	2054 (10.7%)	1273 (12.2%)
unknown	973 (3.3%)	619 (3.2%)	354 (3.4%)

Enriched gene set terms

Enrichment analysis were performed with the list of genes harboring an expressed mutation for AL and MM patient samples. The resulting list of top 20 enriched terms is depicted in the heatmap in figure 3.41, comparing AL to MM. Terms are listed in increasing order of p-value of the respective hypergeometric test in the FEA, i.e. the top term has the lowest p-value. The top 20 terms include "DNA repair" (GO:0006281) and "apoptotic signaling pathway" (GO:0097190). Terms associated with plasma cells in the top 100 contain "cell activation involved in immune response" (GO:0002263), "Adaptive Immune System" (R-HSA-1280218) and "B cell activation" (GO:0042113).

Regarding the list with the top 100 enriched terms 83% of these are enriched in both gene lists, 10 are only significant in MM, and 7 in AL only (data not shown).

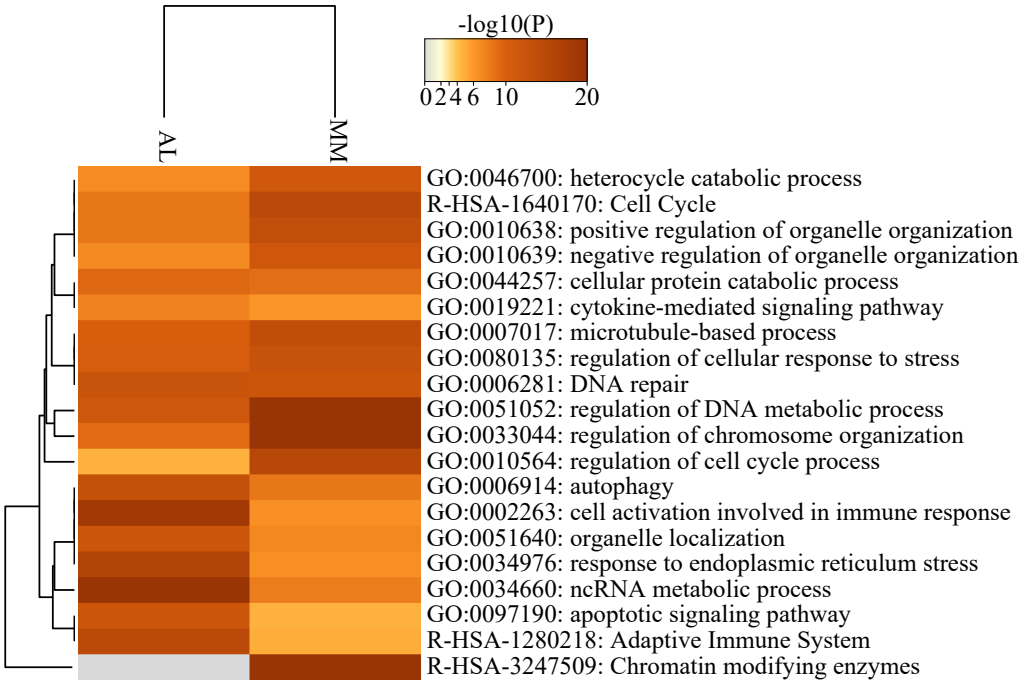

Figure 3.41: Heatmap of functional enrichment analysis of genes carrying expressed, non-synonymous variants per disease entity. Depicted are the top 20 terms (rows), enriched in the list of mutated genes from variant calling of 113 light chain amyloidosis (AL) and 930 multiple myeloma (MM) patient samples. For representation the negative log10 value of the adjusted p-value is depicted. The highest value thus represents the lowest p-value. A value of 2 equates to an adjusted p-value of 0.01.

Known genetic variation

Variants in the table of 113 AL patient samples were annotated by genetic variation databases in the variant calling pipeline (see variant annotation part in section 2.4.4). Annotated identifiers, which represent a known single nucleotide polymorphism (SNP), were analyzed as described in section 2.5.5.

All 8 detected *KRAS* mutations are located at 4 different positions and are annotated to known SNPs (see table 3.25 **a**). The SNP with the identifier rs17851045 was detected 3 times in 3 samples, the one by T>A substitution had previously been associated with MM, the other two (of type T>G) were previously detected in other cancer entities (acute myeloid lymphoma, non-small cell lung cancer). The remaining three SNPs rs121913530, rs121913529, and rs121913240 were also previously detected in other cancer entities, e.g. lung cancer, bladder cancer, and ovarian neoplasms.

In *NRAS*, five different missense mutations were detected (see table 3.25 **b**). Three had been associated with MM (rs121434595, rs11554290, rs121913254), the other

two (COSM1666991, rs121913255) to other cancer entities, e.g. myelomonocytic leukemia, cutaneous melanoma, acute myeloid lymphoma, and non-small cell lung cancer.

Regarding *BRAF*, 3 missense mutations in 3 samples were detected, two being V600E (rs113488022, COSM476). One of the two was expressed (VAF in RNA 57%, see table 3.25 **c**) with a concurrent expression of the BRAF gene.

Table 3.25: Known genetic variation for variants in **a** *KRAS*, **b** *NRAS*, and **c** *BRAF* detected by variant calling in 113 samples with light chain amyloidosis (AL). Chr: chromosome, Pos: base position on chromosome, SNP: known genetic variation identifier, VAF: variant allele frequency, VAF RNA: VAF in RNA, Expr: expression in RNA

a

Chr	Position	Consequence	Allele	SNP	VAF	VAF RNA	Expr
12	25245351	missense	C/A	rs121913530	10.2	4.9	yes
12	25227341	missense	T/G	rs17851045	36.2	21.4	yes
12	25227341	missense	T/A	rs17851045	81.2	75.5	yes
12	25227341	missense	T/G	rs17851045	37.4	50.9	yes
12	25245350	missense	C/T	rs121913529	48.5	3.9	yes
12	25245350	missense	C/T	rs121913529	38	85.7	yes
12	25227342	missense	T/A	rs121913240	39.9		no
12	25227342	missense	T/A	rs121913240	27.1	27.7	yes

b

Chr	Position	Consequence	Allele	SNP	VAF	VAF RNA	Expr
1	114713900	missense	A/T	COSM1666991	12.2	10.6	yes
1	114716124	missense	C/G	rs121434595	33.9	11.5	yes
1	114713907	missense	T/G	rs121913255	16.7	4.5	yes
1	114713908	missense	T/C	rs11554290	26.1		no
1	114713909	missense	G/T	rs121913254	16.5		no

c

Chr	Position	Consequence	Allele	SNP	VAF	VAF RNA	Expr
7	140753336	missense	A/T	rs113488022	12.7		no
7	140753336	missense	A/T	rs113488022	39	57.1	yes
7	140753334	missense	T/C	rs121913364	52.6	100	yes

Immunoglobulin genes

Detected variants affecting Ig genes, in the variant table of 113 AL patient samples, were separately analyzed. The 370 variants are in 118 different Ig genes of the constant, variable, diversity and joining region for heavy or for light chain; 64.5% of detected variants are expressed and detectable in RNA seq (see table 3.26). Of the 370 variants 296 (80%) are located within genes of the variable regions. The overlap to the MM variant table is 75%, i.e. of the 118 Ig genes 89 are also mutated in MM. Genesymbols of Ig genes are depicted in supplementary table A.18.

Table 3.26: Summary statistics of immunoglobulin (Ig) gene variants in light chain amyloidosis (AL).

Ig gene group	Variants	Genes	Expressed	Samples
Heavy Chain Constat Alpha	5	2	4	5
Heavy Chain Constat Delta	13	6	6	11
Heavy Chain Constat Gamma	28	4	23	19
Heavy Chain Joining	13	4	12	13
Heavy Chain Constat Mu	1	1	1	1
Heavy Chain Variable	151	30	76	71
Light Chain Kappa Constant	1	1	1	1
Light Chain Kappa Variable	31	17	18	18
Light Chain Lambda Constant	10	3	8	10
Light Chain Lambda Joining	3	2	0	2
Light Chain Lambda Varaible	114	23	90	59

Mutated genes

Variant table (SNV, InDel) and CNA were compared with a list of 63 potential myeloma driver genes by Walker *et al.* [308]. Of these 63, 44 genes are detected as mutated in the 113 investigated AL samples. In total, variants and CNA in these genes affect 76 (67.3%) of the 113 patient samples, while 27 patients are only affected by CNA. Of these 44 genes, 4 are only affected by CNA and not by SNV or InDel. These are the *RB1* gene, which is only interfered by del 13q14 CNA in 41 samples and 3 genes (*CDKN2C*, *FAM46C*, *FUBP1*) that are only affected by gain of 1q21 CNA in 35 samples. Additionally, *NRAS* and *ARID1A* are amplified by gain 1q21 in these 35 samples. Nine samples harbor variants in these two genes (see figure 3.42 and supplementary table A.17). Overall, in 64 (56.6%) samples none of the 63 listed genes was found to harbor SNVs or InDels. In 28 (24.8%) samples one of these genes was mutated, 14 (12.4%) samples showed presence of $2-4$ mutated genes, and 7 (6.2%) samples $6-9$ mutated genes.

Regarding SNVs and InDels, mutations in *KRAS*, *DIS3*, *NRAS* and *TP53* appear disjunct, see figure 3.42. *RASA2* of the *RAS* family is mutated twice in samples harboring a KRAS mutation. All *NRAS*, *KRAS*, *RASA2*, *TP53* and *DIS3* variants have a median VAF below 50% (see figure 3.43). For 14 of the 44 genes (31.8%) the median VAF is above 50%. The VAF of variants in 9 different genes (*XBP1*, *KDM6A*, *NF1*, *TRAF3*, *HUWE1*, *KMT2B*, *EP300*, *ZNF292*, *CCND1*) were above 90% in at least one sample (see figure 3.43). Variants in *NRAS* and *KRAS* in MM do appear with VAF over 50% in 26 and 27 of 930 samples, but the median VAF in MM is 35% and 30%, i.e. subclonal. In AL the median VAF is 17% and 38% for *NRAS* and *KRAS* and only one sample showed a clonal variant with a VAF of 81% in *KRAS*. Variants in *NRAS*, *KRAS*, *DIS3* and *TP53* occur coincidentally in MM, but are mainly disjunct like in AL.

Five members of the *BCL2* family harbor 12 missense mutations in 5 samples, see table

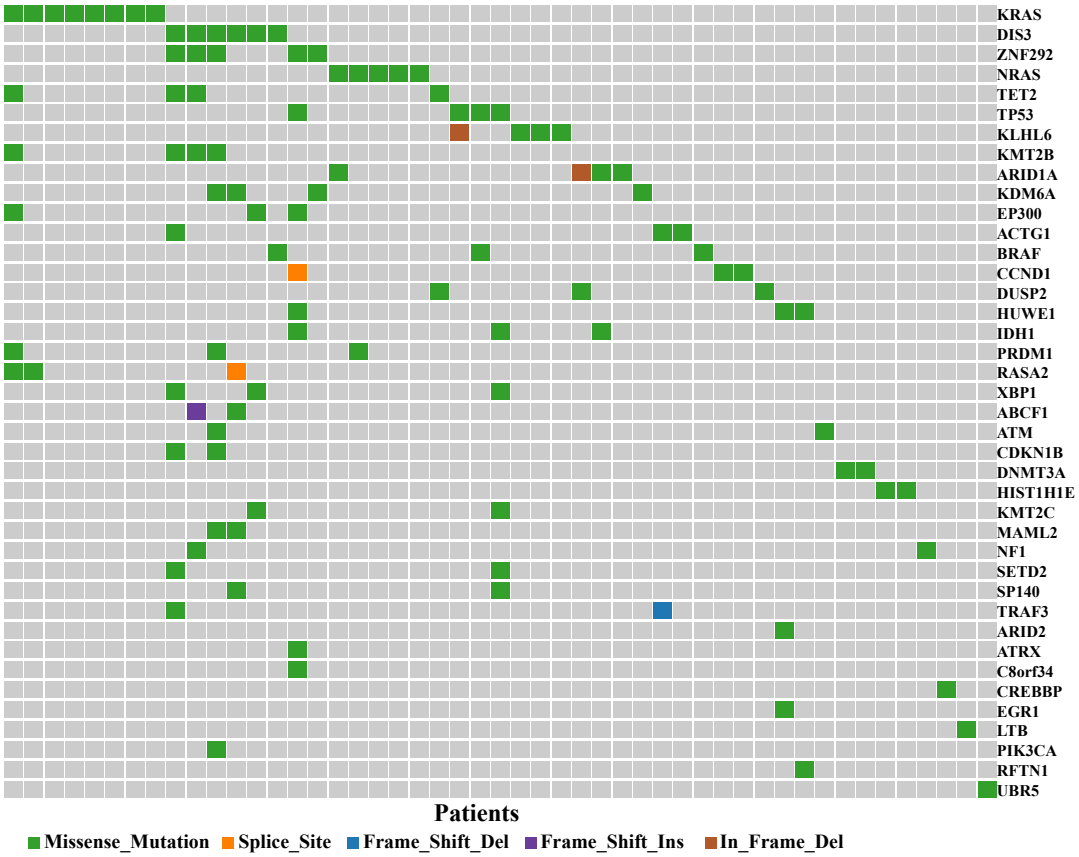

Figure 3.42: Incidence of variants in 63 genes published as potential "driver genes" of multiple myeloma (MM) in light chain amyloidosis (AL). In total, 40 of the 63 genes variants (single nucleotide variant, insertion, or deletion) were detected in the 113 AL patient samples. Every row depicts a gene and every column a sample. If a non-synonymous variant was detected, the color of the box is changed to the respective color for the variant type as referenced in the legend below the figure.

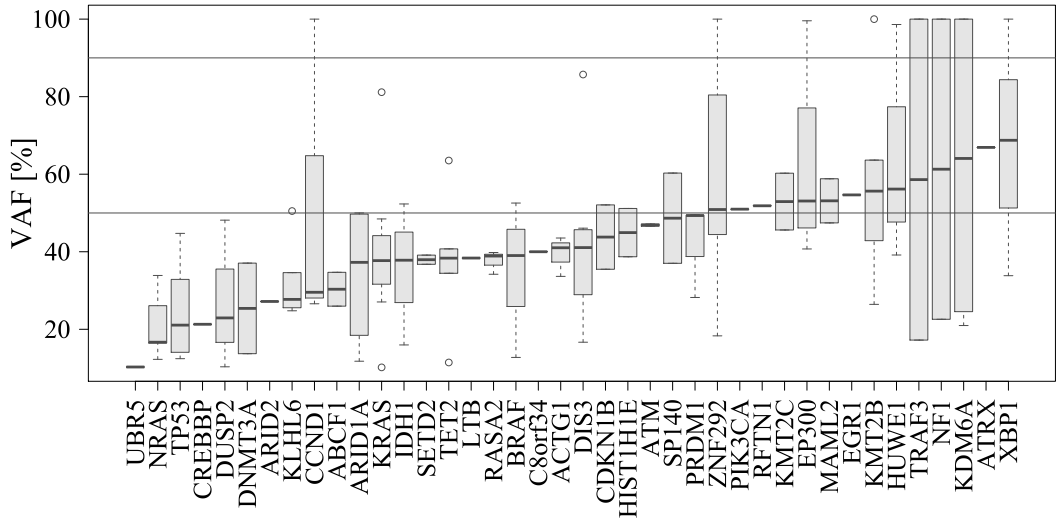

Figure 3.43: Variant allele frequency (VAF) of variants in 63 genes published as potential "driver genes" of multiple myeloma (MM) in light chain amyloidosis (AL). In total, 40 of the 63 genes variants (single nucleotide variant, insertion, or deletion) were detected in the 113 AL patient samples. At the y-axis the VAF is given in percent. Guiding lines are depicted at 50% and 90% VAF.

3.27. These variants occur in pro-apoptotic (*BCL2L13*, *BCL2L14*) and anti-apoptotic (*BCL2A1*, *BCL2L10*, *BCL2L2*) members of the family. Two variants are also expressed (*BCL2L13*, *BCL2A1*).

Table 3.27: Variants in genes of the *BCL2* detected by variant calling in 113 samples with light chain amyloidosis (AL). SNP: known genetic variation identifier, VAF: variant allele frequency, VAF RNA: VAF in RNA.

Gene	Sample	VAF	VAF RNA	SNP
BCL2L10	N1493	100	0	rs2231292
BCL2L13	N1493	59.2	53.8	rs9306198
BCL2L14	N1530	52.8	0	rs61739220
BCL2A1	N1531	100	100	rs3826007,CM064994
BCL2A1	N1531	100	100	rs1138358
BCL2A1	N1531	100	100	rs1138357
BCL2L2	N1551	15.2	0	rs10149339
BCL2L14	N1597	39.7	0	rs150190776
BCL2L14	N1597	40.4	0	rs138650437
BCL2A1	N1597	100	0	rs3826007,CM064994
BCL2A1	N1597	100	0	rs1138358
BCL2A1	N1597	100	0	rs1138357

The variant table was assessed for overlap of mutated genes in relation to three previously published lists of detected mutations in AL (see supplementary table A.17 for a list of these genes). The overlap of mutated genes as regards to Boyle *et al.* [36] is 10 of 13 genes, to those of Rossi *et al.* [254] is 5 of 5 genes, and third to Paiva *et al.* [230] is 65 of 93 genes.

Difference in the incidence of mutations in potential "drivers"

Odds ratios were computed for the 44 genes previously indicated as potential "driver genes" that were detected in both variant tables (AL and MM). The difference of incidence of variants between AL and MM in these genes was assessed by odds ratio if genes are at least once mutated in both cohorts (see figure 3.44). *NRAS* and *KRAS* are significantly more frequently mutated in MM. In AL, *KDM6A* is more frequently mutated (3.5% *versus* 0.5%) but the frequency is low in both cohorts. For most genes, no significant different frequency of mutation exists between the two disease entities. *FAM46C* mutations, detected in 89 (9.6%) samples in MM were not detected in AL.

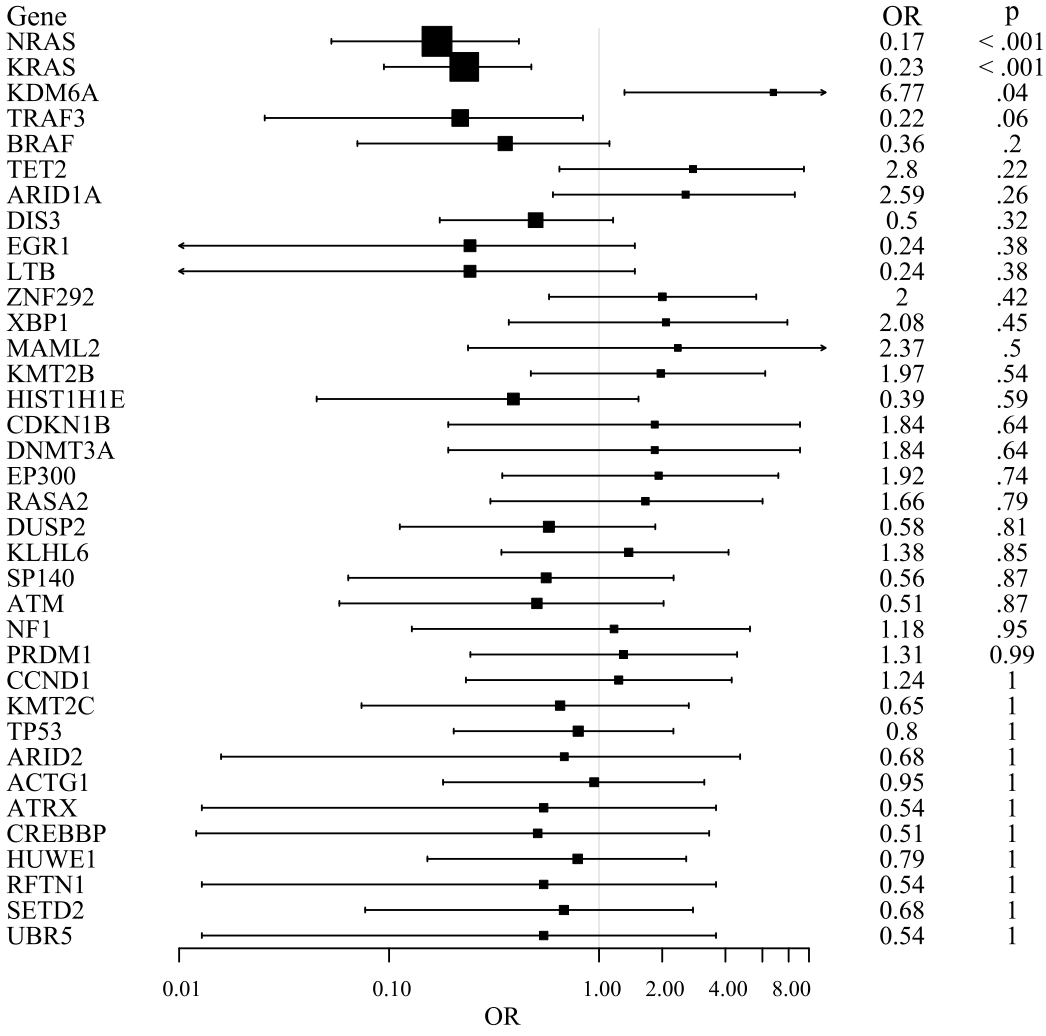

Figure 3.44: Odds ratio (OR) for the number of variants of previously described potential MM "driver genes" in AL *versus* MM. An OR < 1 indicates a greater incidence for mutations in this genes in MM, an OR ~ 1 indicates no difference of frequency in any entity and an OR > 1 indicates a greater incidence in AL. Significant differences, indicated by Fisher's exact test, are illustrated by a BH-adjusted p-value ≤ 0.05. AL: light chain amyloidosis, MM: multiple myeloma, BH: Benjamini-Hochberg, OR: odds ratio, p: adjusted p-value.

4 Discussion

The discussion of this thesis is organized in four parts: first, a methodological section on the WES pipeline. Second, the evaluation of the personalized GEP-R. Third, a part on prognosis of AL patients and fourth, findings in connection with the malignant plasma cells in AL regarding the aims of the thesis. Lastly, the thesis aims are discussed, a conclusion is drawn, and an outlook is given.

4.1 Methodological discussion of WES pipeline

In the following, the implementation of a WES pipeline and analysis of WES data are discussed regarding reproducibility, quality control, variant calling strategy, and adaptability.

4.1.1 Reproducibility of WES pipeline

To ensure reproducibility of the thesis WES pipeline, a comprehensive description of the analysis steps is given in the materials and methods chapter, see section 2.4.4. Versions of all used tools (see supplementary table A.1) and reference files (see section 2.4.1) are documented and referenced for online access. All code is available in supplement B.

4.1.2 Quality of sequencing data

QC of WES sequencing files was performed during analysis by five tools, see section 2.4.4, the flowchart in figure 2.1 for the processing scheme, and supplementary table A.1 for a list of used tools. QC results are described in section 3.1.1 and briefly summarized in table 3.1.
Two categories of results, potentially identifying quality issues need to be discussed, sequence duplication and GC content.
Sequence duplication rates are available from Fastqc, fastp, and Picard. Fastqc reports high rates of sequence duplication. In 69% of the FASTQ files, Fastqc displays $\geq 50\%$ of sequences as duplicated. In principle, this could indicate a problem, because sequence duplication would imply low sequence diversity. To further assess this point, sequence duplication rates determined with different tools, i.e. fastp and Picard, showed the expected ranges of $3\% - 10\%$ and $5\% - 19\%$ (see table 3.1) corresponding to typical technical variation [9, 38, 47]. The explanation could be attributed to Fastqc scanning only the first $100,000$ reads [9], in contrast, fastp and picard assessing all reads.

The GC content is indicated by Fastqc as "abnormal", pointing out an unusual distribution of GC content in comparison to the normal distribution. This fits to the unusual per base sequence content, detected in all files by Fastqc, showing a variation in the first 15 bases. Per base sequence content is detected as unusual if the frequency of two different bases at any position differs in more than 20%. Given that this bias is restricted to the first 15 bases (and otherwise normal) it can be attributed to technical bias introduced in the library generation as previously described [9, 117]. With fastp GC content was calculated per base position over all reads. In all FASTQ files the first 15 bases show a mean GC content varying between $31\% - 73\%$ per position in each sample, while the latter base positions show an expected GC content of 50%. The automatic trimming by fastp improves the GC content and the per base sequence content a bit. No consequences were expected by the bias for subsequent analyses steps as previously described [117].

4.1.3 Sequencing analysis strategy

In the presented analysis strategy for WES data, somatic variant calling and copy number calling is implemented. For both analysis methods, pairs of tumor sample and corresponding germline need to be sequenced. By this direct comparison, assessment of germline variants can be omitted during variant calling [112]. For copy number calling, germline data are used for estimation of the read depth for physiological diploid status.
Somatic variants were called in WES data (DNA-level) and validated using RNA seq (mRNA expression). RNA seq data were also used to assess the expression height and abundance of altered transcripts compared to non-altered ones (see sections 3.1.2, 3.6.7). Analyses regarding CNA detected by copy number calling are described in section 3.6.2.

Variant calling and verification in DNA

The crucial step in variant calling is to decide whether mismatching bases in alignments are based on mutations or sequencing errors (see section 1.6.3 and base quality in section 2.4.2). Seemingly a straight-forward assessment, variant calling results are impacted by alignment, the calling method, filtering of "possible" variants, and subsequent further validation of "high confidence" variants.
A variety of variant callers is available. Comparative studies tend not to suggest "the" method [226, 227, 324] but support selection of different tools for analysis pipelines (see the description of the analysis pipeline in section 2.4.4 and figure 2.1). The strategy applied in this thesis is to use more than one variant caller [227] and applying

different strategies and statistical models to generate an encompassing list of potential variants first, accepting potential false positive detections. To do so, the three callers VarScan2 [158], Strelka [154] and Seurat [56] were chosen based on the following criteria: first, availability of a comprehensive and up-to-date documentation [227], second, different underlying statistical models as basis for variant calling, and third, the possibility to install and implement the caller. All chosen callers needed to have performed well in comparative studies [227, 324]. Strelka and Seurat had been used to call variants for the CoMMpass cohort [61, 218] used as MM comparator samples to AL in this thesis (see section 3.1.3 , discussed in section 4.1.3).
The raw variant table from the analysis of 113 patients in this thesis contains 446929 variants, 32% detected by at least two callers (see figure 3.1). The number of variants detected by only one caller passing the false positive filtering is smaller than the one detected by two or more callers (37% *versus* 57%). In a second step, the number of hits was decreased, and the likelihood of true variants was increased by using the intersection of the three variant calls [112]. After the initial step of assessing an upper limit for the number of potential variants (by the least stringent criteria, i.e. the sum of variants found by the different callers), this step was used to obtain highly confident hits for further biological and systematic analysis. To validate this process, patterns of known, frequently occurring sequencing errors [158] can be used. For example, if all variant supporting reads being obtained from the same strand, they are almost always based on PCR amplification errors [158]. If the average variant position is at the edge of supporting reads, methodologically the chance of a sequencing error is higher than that of a true variant [158]. To account for this, metrics of the reads in the alignments were assessed at the respective variant positions using bam-readcount. These metrics were used to filter all detected variants. Read counts of the remaining variants were used to calculate the VAF.
As a further threshold for variants to be assessed in subsequent analyses, a VAF of $\geq 10\%$ was chosen to further reduce false positives and to focus on those variants involved in prior progression or initiation of the diseases (which are conservatively assessed as those being present at least in a subclonal fraction, defined here as in iFISH [222]). For other analyses as e.g. longitudinal assessment of two samples, a different strategy would need to be applied. Further reason for the threshold (as in iFISH) are possible contaminations by normal BMPC in low frequencies. Preprocessing of alignments before calling variants further improves the quality of the VAF. It has previously been shown that the marking of optical and PCR duplicated reads with picard MarkDuplicates enhanced the precision of the VAF for SNV [38]. Improvement of

the VAF for InDels can be gained by realignment around possible InDels with GATK IndelRealigner before variant calling [112].
Variant types can be distinguished in synonymous variants having no effect on encoded amino acids, and the non-synonymous variants, resulting in a different amino acid.

Orthogonal validation of variants by RNA

A further validation of biological impact of variants detected by WES is their expression on RNA-level. First, detection of the same variant by an independent method can be seen as "orthogonal validation": Variants confirmed by RNA have a high confidence to be no sequencing error, because RNA seq is a different sequencing method (in case of input material and library preparation) and the sequences are determined by different analysis pipelines compared to genome or exome sequencing [112] (see section 1.6.3). RNA seq is of further benefit in validation of alterations detected by WES as in malignant plasma cell diseases other conventional validations strategies such as conventional sequencing is not possible due to low amount of tumor input material [126] (see the RNA seq and WES parts in section 1.6.3 and section 2.1.2). Secondly, to a certain extent, the biological impact of this alteration can be assessed by determining the frequency of a particular alteration within all reads at a specific location. E.g., a potentially driver mutation will less likely act as one if either not expressed or expressed in a minor subfraction of reads. The relation of reads in WES and RNA seq can furthermore be taken as indication for the clonal fraction harboring a mutation (in WES) in relation to the expression level of the altered allele in comparison to the wild-type allele.

Comparison AL variants to MM

The obtained variants in AL were compared to MM variants detected in the CoMMpass cohort (see summaries on AL variant table, section 3.1.2, CoMMpass cohort, section 3.1.3). The underlying question here is whether AL-typical variants exist, or whether the same variants could be found in both malignant plasma cell diseases. In turn, a potentially unique AL-specific pathogenetic mechanism could be investigated. A possible confounder in this comparison is that CoMMpass samples were aligned to the hg19 reference genome and the AL to the up-to date hg38 version (see section 2.5.5 for description of CoMMpass). As FASTQ files are not available for CoMMpass, no new alignment could be performed. According to Pan *et al.* [235] this introduces only a small technical bias (1%). To correct for ambiguities potentially introduced by differing reference genomes, variants were only compared regarding gene names, not variant positions (see section 2.5.5). Further methodological differences cannot be excluded:

plasma cell purification methods, sequencing strategy (WES and WGS), sequencing kits, and the bioinformatical pipelines can induce technical variation. Given the size of the cohorts (especially COMMpass where less information about purity of samples is available), it is highly unlikely that major alterations in terms of frequent mutation would have been overlooked. Impurities (e.g. contaminations by normal plasma cells) would dilute the VAF, but not to an extent omitting specifically aberrant hits. In turn, purity was well controlled in the AL-cohort. Here, purity was assessed by flow cytometry as well as minimum size of clonal alteration by iFISH (see section 2.1.2 [222]). A further potential source of variation is the ethnic constitution of both cohorts. Whereas the AL assessed in this thesis mainly includes samples of German patients, those in CoMMpass are mainly from US, Canadian, Italian and Spanish patients. Global quality measurements, e.g. Variant type (97% SNV, 95% missense variants in both, see 3.6.7) and Ti/Tv ratio (1.5 and 1.4, see 3.6.7) are equal between both cohorts (and therefore also between AL and MM), suggesting a presumably small variation by all these confounders.

Copy number analysis

The copy number analysis was possible using with VarScan2, DNAcopy and GISTIC2 as described in section 2.4.5. This strategy was validated using assessment by iFISH [23, 25, 26, 39, 116, 221–223]. Whereas assessment of ploidy by WES is considered possible using the strategy implemented as part of this thesis, limitations also become clear. In the meantime, publications refer to an inferiority as compared to analysis by WGS [145, 327, 333, 334]. The results of this thesis lead to the implementation of a low coverage WGS strategy (using a low amount of input material necessary in malignant plasma cell diseases) complementing the RNA seq/WES strategy described here. Results of the analysis are described in section 3.6.2 and 3.6.3 and discussed in section 4.4.3.

4.1.4 Adaptability and variability of sequencing pipelines

Sequencing analysis pipelines are, as molecular profiling in malignant plasma cell disease and other tumor entities in general, not standardized. This especially applies to library preparation and sequencing strategy, but likewise to bioinformatic analysis. Different strategies, especially in terms of bioinformatic tools used (see also section 3.1.2 and figure 3.1) lead to different numbers of variants found [226]. This "technical bias" needs to be taken into account comparing different studies, especially if the composition in terms of entities differs. Sequencing analysis pipelines therefore need to be variable and should be easily up-datable or expandable. An analysis of sequencing

data should be considered a dynamic process, and along these lines there are even recommendations to iteratively reanalyzed sequencing data every $1-2$ years, to benefit from the advantage of new analysis tools, bug fixes and updated annotation sources [60, 320]. The difficulties of not doing this can be seen for the CoMMpass cohort still adhering to hg19, making comparisons more error prone (see section 2.5.5). A versatile pipeline structure is of additional need for comparability of samples (e.g. between different study groups) or addition of patient data to reference cohorts. The applicability of the sequencing pipeline described in this thesis was tested for RNA seq data in the GMMG-MM5-multicenter-trial, for which the LfM served as central laboratory.

As an example for the described adaptability of the pipeline, the analysis of the BRAF V600E mutation was included, for which clinical grade inhibitors are available [10, 29] (see section 3.6.7 and 4.4.8).

4.2 Evaluation and prospective application of the GEP-R within the GMMG-MM5-multicenter trial

This section comprises the discussion of the first two aims, regarding risk assessment in MM for relation to AL and particularly, are GEP-based risk assessments determining the malignant plasma cell properties in MM as good as the current standard risk stratifications? And is a personalized therapeutic recommendation possible by assessing the expression of target genes? Results are presented in section 3.2. Parts of the analyses have already been published in shared first-authorship in the article Hose, D., Beck, S. *et al.* [126].

Bone marrow aspirates were obtained from $573/604$ patients (95%) and could be CD138-purified in $559/573$ (97.6%) [126]. Of these, iFISH-analysis was possible in 556 (99.5%), GEP in 458 (82%) [126]. This compares favorably to iFISH results reported by the EMN02/HOVON95-trial (74.1%) [43], IFM2009-trial (73.6%) [12], DSMM XI-trial (73.7%) [88], SWOG S0777-trial (60.2%) [81], or a pooled analysis of three PETHEMA/GEM clinical trials, i.e., GEM2000, GEM2005MENOS65, and GEM2010MAS65 (60.8%) [173], which were conducted during a comparable time frame [126]. It was therefore for the first time validated prospectively in a randomized phase III multicenter trial the possibility to perform not only cytogenetic (including rISS) but also GEP-based risk stratification and reporting in $> 80\%$ of patients during the first cycle of induction chemotherapy as - potentially - molecular risk-adapted, personalized treatment strategy [126].

Risk stratification using an integrated approach, i.e., HM metascore, delineated $10/77/13\%$ of patients as high/medium/low risk, transmitting into significantly differ-

ent median progression-free survival (PFS) of 15 vs. 39 months vs. not reached (NR; $p < 0.001$) and median overall survival (OS) of 41 months vs. NR vs. NR ($p < 0.001$), see figure 3.5 and tables 3.5, 3.6 [126]. Five-year PFS and OS rates were $5/31/54\%$ and $25/68/98\%$, respectively (see tables 3.5, 3.6) [126]. Survival prediction by HM metascore (Brier score 0.132, $p < 0.001$) is superior compared with the current gold standard, i.e., revised ISS score (0.137, $p = 0.005$), see table 3.7 [126].

Besides risk stratification which can be done by both FISH and GEP [99], the specific benefit of GEP, either by DNA microarrays as in this trial or by RNA seq, lies in the additional ability to identify target gene expression. In the analysis, this was intended for immunological targets and those for which small molecules or antibodies existed, e.g., AURKA (VX-680 [129]), IGF1-receptor (e.g., AVE1642 [210]), or FGF3R (e.g., CHIR-258 [296]). Targets were reported as expressed; for IGF1R in 33.1%, AURKA in 43.2%, and FGFR3 in 11% of patients [126]. AURKA was selected at the time when the GEP-R was developed and it had previously be shown to be expressed in approximately 30% of previously untreated myeloma patients and is associated with adverse survival [129]. IGF1R-inhibition was selected due to its importance as myeloma growth factor and impact on patient survival [280, 281]. Here, it was the intention to give the proof-of-principle for prospective advanced molecular diagnostics of targets and reporting in clinical routine; in this way, the GEP-R is depicted and should be interpreted. Novel targets for which clinical grade inhibitors become available or immunological targets can be added to the assessment due to the adaptable surface of the reporting tool (GEP-R) [198]. Without doubt, actual implementation necessitates standardization and either commercial or academic development of an actual molecular diagnostics test. It is shown that such a strategy is in principle feasible.

Alongside the principal possibility of running advanced molecular profiling and depiction of potential targets for individualized treatment, assessment of risk was the first objective of this thesis. Here, many prognostic factors have been described with the ongoing discussion of which to include [53, 99]. This leaves the treating physician with a plethora of information that is difficult to consolidate intending counseling patients and drawing a clinical conclusion. Metascoring appears as an appropriate strategy [52, 162, 198, 274] to overcome this, and that this is likewise possible in a randomized clinical trial setting was shown in this thesis. Regarding the molecular techniques used in the metascore, i.e., iFISH and GEP, it was chosen to include both due to in part non-overlapping prognostic information, e.g., it is not possible to predict del17p13 at a high-enough accuracy by GEP [336].

The actual (good) prediction result of the HM metascore per se is thereby not the main focus of this analysis. Nonetheless, even with this "2010 choice" regarding risk

factors and target genes, metascoring including GEP-based risk assessment is superior in numbers to rISS, although not statistically so. The same however holds true for a comparison of rISS to ISS even on the comparably large cohort of patients used in this thesis.

4.3 Prognosis of AL patients

This section comprises the discussion for the third, fourth, and fifth research questions of the thesis regarding prognosis of AL patients. The prognosis of AL patients, independent of a specific risk assessment, during the first 12 months is significantly poorer compared to previously untreated myeloma patients (see figure 3.24).

Determinants of prognosis in AL can be grouped in two broad categories. First, prognostic markers that reflect end organ damage caused by amyloid deposition (see also section 1.4), and secondly, malignant plasma cell characteristics. Of course, both are interconnected, and the exact delineation depends on the pathogenic point of view. An example is the serum FLC-level that is determining FLC organ deposition and thus, is an organ damage factor, but likewise depends on the number of cells and individual LC production rate which in turn are malignant plasma cell factors. Traditionally, plasma cellular factors on the one hand are described as corresponding to the pathogenic LC and underlying clonal bone marrow disorder. On the other hand there are organ biomarkers, which reflect the end organ damage caused by the toxic FLC and the deposited amyloid fibrils [73, 75]. As mentioned and discussed in the following, the two categories are connected. In this thesis, the delineation is made between amyloid deposition based prognostic factors, and those corresponding to malignant plasma cell characteristics impacting in "non-amyloid producing" plasma cell diseases, i.e. multiple myeloma.

As a rule of thumb, end organ damage associated clinical prognostic markers define most of early prognosis of AL patients [75], whereas especially if the heart is involved, the prognosis is poor [75, 107].

Malignant plasma cell characteristics, impact later in terms of response to treatment and time until disease recurrence (e.g. such as proliferation rate of malignant plasma cells, in this thesis for the first time assessed for AL, see section 3.4). Prognostic malignant plasma cell factors have traditionally been restricted to detection of CA in malignant plasma cells and the amount of FLC by the cells (see section 3.3).

For the first time, in this thesis, molecular characteristics by gene expression regarding prognosis of AL were analyzed. Results are depicted in section 3.4 and 3.5 and are discussed below.

4.3.1 Amyloid deposition based prognostic factors

Amyloid deposition can be seen as depending on the height of FLC (or diff FLC), which in turn depends on the production rate of FLC by an individual malignant plasma cell, and the number of malignant plasma cells. An important question lies in how much the deposition pattern of amyloid is driven by the absolute height of FLC or if it is an individual characteristic of a specific LC. For the latter, in this thesis, the term "amyloidogenicity" is used.

As stated, LC production rate is determined by the amount of malignant plasma cells and their individual production rate [24, 295]. Regarding the latter, normal plasma cells produce a high amount of intact Ig [156] and only a low rate of unpaired FLC as identified by the respective serum levels in healthy individuals compared to MM or AL-patients [79, 80, 256, 257]. Malignant plasma cells produce less Ig, as the derangement leads to a general deviation from BMPC state [79, 80, 256, 257] (e.g. acquisition of proliferation by malignant plasma cell instead of being a Ig-factory or by a CA, like t(11;14), affecting the Ig locus). Derangement of malignant plasma cells, leading to lower production of complete Ig simultaneously implicate a higher relative and frequently absolute production of unpaired/free LC as in the physiological state.

A higher number of plasma cells, indicated by high PCI and M-protein [295], is in turn related to the production of a higher amount of unpaired LC for a given level of aberrant LC production per cell. It is thus logical that PCI and M-protein levels are significantly associated with adverse survival ($p = 0.002$ and $p = 0.01$, see figure 3.11 **a** and **c**). Given that the tumor mass is a rather indirect measure, and the FLC deposition directly impacting on the amount of deposited LC, it is understandable that the grade of significance is lower compared to the diff FLC measurement ($p < 0.001$). As the total FLC production is the product of an individual production rate per malignant cell and the number of cells (i.e. surrogated by PCI or to a lesser degree M-protein), the total impact of plasma cell number will also impact in prognosis via the absolute number. In other words, it is more difficult given fractional cell killing by chemotherapeutic approaches to lower the tumor mass under a critical threshold of FLC production where amyloid deposition is prevented if a higher tumor mass is present at the beginning, i.e. PCI or M-protein. In as much the deposition pattern of amyloid is driven by the absolute height of FLC or its "amyloidogenicity".

If analyzing the pattern of organ involvement, a large overlap is found, i.e. most frequently several organs are involved (see figure 3.10 in section 3.3). For patients with heart involvement (75%, see table 2.5), the most frequent organ location, 23% likewise show kidney, and 20% liver involvement. This has also been shown assessing the differences in involved and non-involved FLC, i.e. diff FLC: High ratios are as-

sociated with heart involvement low ratios more frequently with kidney involvement - and consecutively better prognosis [24, 74, 163, 205]. Thus, at least part the amyloid deposition is driven by the height of FLC (either assessed absolutely or as diff FLC). This is exemplified in low diff FLC ratios more frequently affecting kidney, whereas high ratios are more prone to affect heart, or both organs (see table 3.8). The presence of a small proportion of patients with one involved organ (21%, see table 2.5), e.g. 8% with kidney only involvement despite high diff FLC (see table 3.8) however implies that amyloidogenicity varies at least to a certain degree between different LC. It is also well understandable that organ involvement determines almost exclusively the early prognosis of treated patients, and that this is especially the case for heart involvement as most critical feature (see figure 3.7), and in turn, excellent prognosis if kidneys are the only involved organs (see figure 3.7 **a**). In this thesis, we show for the first time that a malignant plasma cell factor, i.e. the proliferation rate, also highly impacts on this early prognosis (see section 3.4).
Frequently used serum markers depicting heart damage are NT-ProBNP and cTnT. Especially cTnT is already very sensitive to small damages [70]. Both markers were found to be prognostic [71, 164, 231], with NT-ProBNP $>=$ 8500 [315] detecting a group of patients with high risk of premature death (see figure 3.8 **a**).
Previous research on these prognostic serum parameters, application of predictive thresholds and the combination of serum parameters introduced risk assessment models, i.e. the "standard" Mayo Score (2004) by Dispenzieri *et al.* [70], the revised Mayo Score (2012) by Kumar *et al.* [164], and the "advanced" Mayo Stage III Euro Score (2013) by Wechalekar *et al.* [315] (for definition see section 1.4). The combination of clinical prognostic serum parameters representing different disease associated factors increases the accuracy of survival prediction, especially late survival [216]. All three risk assessment models are analyzed in this thesis (see figure 3.9).

4.3.2 Malignant plasma cell characteristics

The second group of factors analyzed in this thesis is associated with malignant plasma cell characteristics, which are in the following discussed regarding their prognostic impact. CA assessed by iFISH have previously been described. They are the same as assessed in MM. In this thesis, special focus is laid on addressing gene expression. For the first time, proliferation of malignant plasma cells as biological but likewise prognostic factor is assessed. GEP-based myeloma risk scores are addressed for AL, and a novel AL-based score derived and tested *vice versa* for prognostic impact in MM.

Chromosomal aberrations as detected by iFISH

AL has a lower fraction of patients harboring CA associated with high risk (as gain 1q21, del 17p, t(4;14)). In contrast to MM, in which CA are to a large extent prognostic independent of applied treatment, this is not the case in AL. Examples are the negative prognostic impact of t(4;14), t(14;16), del 17p or 1q21 gain, i.e. all features with high risk in MM [156, 202, 221, 265]. In AL, none of the CA alone has a prognostic impact (see figures 3.12, 3.13). This is *a priori* surprising, but then CA have been shown to be only prognostic regarding a specific therapy scheme [26, 27, 75, 77, 155, 215, 245] (cf. section 1.2.3). Examples are patients with t(11;14), the most frequent CA in AL, having a significantly adverse prognosis if treated with bortezomib containing regimes [27, 77, 215, 245]. Patients with t(11;14) however respond particularly well to HDT and ASCT [75, 108] (c.f. section 1.5.1). Combinations of daratumumab/dexamethasone and ASCT, likewise show a (slightly) better prognosis in event-free survival of patients with t(11;14) [75, 155]. Better PFS was found for this group of patients with venetoclax [237] (c.f. section 1.5.4). The latter in agreement with data in MM showing a significantly higher activity of venetoclax in t(11;14) patients [165]. In contrast, treatment with melphalan/dexamethasone combinations or daratumumab show a relatively adverse prognosis in patients with gain 1q21 [26, 155, 215]. The investigation in our large cohort of 582 patients treated with different regimen thus implies that the prognostic impact of CA in AL is almost completely depending on the treatment regimen.

This is the more surprising as GEP-based features remain significant (see section 3.4 and the discussion below). This likewise holds true for biological variables as proliferation.

MM-derived GEP-based risk factors

As discussed, malignant plasma cell derived factors impact on prognosis of AL patients in the mixed cohort of this thesis. This impact manifests mainly after the early phase of treatment, which is mainly derived by organ involvement (see above and reviewed in [75]). In this thesis GEP-based risk scores derived for MM, as well as different molecular classifications of myeloma are assessed for the first time (see also discussion on pathogenesis in section 4.4).

For a description of the risk scores and classifications see section 1.2.4 and the methods sections 2.3.2. The results are depicted for the individual scores in section 3.4.

The idea behind prognostic scores is to select - by different strategies - genes associated with prognosis. All respective scores have been initially determined for previously untreated MM patients undergoing treatment.

Same as risk associated CA all three investigated myeloma derived GEP-based risk scores (UAMS70, RS, IFM15), show generally a lower proportion of high risk patients in AL compared to MM (see table 3.13). In contrast to iFISH-based factors showing prognostic impact late, GEP-based risk factors however already impact very early (i.e. in the first 6 months) heavily on patient outcome (see figures 3.19, 3.21 **b**).

Proliferation

Proliferation assessed by gene expression is a strong prognostic factor in MM and AL. Here, as GEP-based score, the GPI strongly delineates early differences in patient survival (see figure 3.15 **b**). In the latter, AL patients compared to MM, a significantly lower percentage of malignant plasma cells show either medium or high GPI (66% *versus* 80%, $p < 0.001$).

This is in line with the pathogenic findings of a more adverse (risk prone) molecular profile in MM compared to AL. This holds true for high-risk CA as well as for all GEP-based risk factors (see table 3.13, figure 3.26 and below).

GEP-based myeloma risk scores

The GEP-based myeloma risk scores depict a smaller fraction of patients as high risk compared to previously untreated myeloma patients. The UAMS70 depicts 8% of AL patients (compared to 25% of previously untreated MM). These patients however show a very adverse risk (see figure 3.19 **b**). The same general observation holds true for the RS (4%/8% AL/MM, see figure 3.21). For the EMC92, no patient is classified as high risk. Taken together, myeloma derived high risk features (if present) imply adverse prognosis in AL. Interestingly, this can be detected already early (c.f. fraction of ALMG and ALMM in table 3.13). Therefore, malignant plasma cell factors can impact on early mortality in AL, as it is the case in MM.

Next, it was aimed to develop a risk score for AL. Basically, two questions were followed: Is such a score specific for AL, i.e. would AL-specific alterations in gene expression dominate, or would the score at the same time be prognostic in MM, i.e. be built on malignant plasma cell specific factors in contrast to AL-specific characteristics.

AL-derived GEP-based risk assessments

The HDAL was created using the methodology developed by Rème *et al.* [248] in collaboration with the LfM (RS as described in section 2.3.3). Genes for the HDAL had been selected by prognostic relevance and optimal thresholds for stratification were determined, both by running Log-rank tests.

The HDAL is the first gene expression-based risk assessment on AL, which is independently predictive and stratifies patients, as intended, in three groups with significant different median survival (6/33/72 months, see figure 3.22 **b** and table 3.11). Likewise, it is prognostic for AMM and MM ($p < 0.001$, see figure 3.23). In MM, a larger proportion of high risk is present (63% *versus* 23% in AL, see figure 3.22 **c**). Risk score determination on either AL or MM thus delineates malignant plasma cell characteristics. The HDAL is independently predictive for low risk to high risk in multivariate Cox regressions with risk models based on the amount of FLC production and organ involvement, i.e. "standard" Mayo Score (2004) by Dispenzieri *et al.* [70], revised Mayo Score (2012) by Kumar *et al.* [164], and "advanced" Mayo Stage III Euro Score (2013) by Wechalekar *et al.* [315] (see table 3.12). These three risk assessment models are the best currently available risk assessments for AL patients based on amyloid deposition and particularly on heart involvement (as described in section 1.4). Here, it is shown for the first time that amyloidogenicity and malignant plasma cell factors independently determine prognosis in AL patients. This holds true for the HDAL specifically developed on AL patients, but likewise for other - MM-based - risk factors or proliferation as biological variable (see above).

Both classes of factors play out their main impact during different phases of the disease. As indicated at the beginning of this section, amyloid deposition and organ involvement based factors mainly determine survival in the first $6-12$ months of the disease, whereas malignant plasma cell disease based factors become important with chemotherapeutic treatment. This has been shown for chromosomal aberrations [26, 27, 155]. Gene expression-based risk factors also impact early during the course of the disease, if they are present. The latter is the case in a significantly fewer proportion of AL patients. A specific case is proliferation. One notable exception of this rule is high proliferation. If present, patients show a dismal prognosis in first three months (see figure 3.15 **b** and table 3.10). One potential explanation is that, as diagnosis is made late in AL [106, 107], patients with fast accumulation of malignant plasma cells are potentially diagnosed later during their natural course of disease.

The transferability of the HDAL to AMM and MM is a strong indication that malignant plasma cells in AL do not determine a unique entity and that they do not differ to the malignant plasma cells in AMM and MM (in addition to the findings discussed below on molecular pathogenesis in section 4.4).

4.4 Molecular pathogenesis of AL

The following section discusses the last three research questions of the thesis aims regarding pathogenesis of AL. The discussion revolves around the question whether

malignant plasma cells in AL are characterized by specific, i.e. "AL-specific" features. To do so, similarities and differences of AL to other disease entities, i.e. MGUS, AMM, and MM are assessed (as described in section 3.4 and 3.6). It starts with gene expression-based assessments of plasma cell characteristics. Afterwards, the results as determined by iFISH were discussed. These were compared to copy number alterations determined from WES data. Subsequently, results from expression analysis obtained from RNA seq are discussed. Finally, results from variant calling determined from WES were discussed.

4.4.1 Gene expression-based assessments

It is not surprising to see that GEP-based risk scores show generally a lower proportion of high-risk patients in AL compared to MM (see table 3.13). Regarding plasma cell derived prognostic factors, these patients fall in between the categories MGUS, AMM, and MM, implying that an advanced malignant plasma cell disease like in MM is less frequent in AL. But this cannot be interpreted as that all in AL patients plasma cells are more like the early stage MGUS than MM: on a single patient level, the malignant plasma cells in AL do resemble the characteristics either in MGUS, AMM, or MM. Different molecular subtypes do contribute in altered proportions.
The group proportions of GPI, MAI, TC and MC in AL are different from the proportions of MGUS and AMM (for example see figure 3.15 **c**): AL as an entity fits in an own "molecular age" between MGUS and AMM.

Myc-activation index

Activation of Myc, as surrogated by the MAI, is found in all malignant plasma cell diseases. A MAI above 1 is more frequent in AL than in MGUS but less frequent than in MM (see figure 3.16 **c**). A significantly higher proportion of patients carry a MAI > 1 in ALMM compared to ALMG ($p < 0.01$). Thus, Myc-activation in AL mirrors the stage dependent increase of Myc-activation in plasma cell diseases.

Gene expression-based classifications

GEP-based classifications stratify patients based on gene expression and attribution to different molecular entities. Groups are in part, but not completely, driven by underlying chromosomal aberrations, as e.g. t(11;14) or t(4;14). The different distribution of molecular entities regarding GEP-based classifications thus mirrors the observed and known differences in underlying chromosomal aberrations. Consecutively, the TC classification applied to AL shows the 11q13 group as largest (see figure 3.17 **c**), corresponding to t(11;14) CA involving the 11q13 locus being the most frequent underlying

CA in AL. This was detected in the same way for the MC. Here, the CD2 group is the largest group with 47% of AL patients (see figure 3.18 **c**). Analogously, the small proportion of HRD in AL is visible by the small HY group in MC.
In summary, GEP-based classification mirror the distribution of underlying chromosomal aberrations and, in case of MC, the distribution of a (in AL small) "high proliferation" group.

4.4.2 Chromosomal aberrations assessed by iFISH

For evaluation of CA the largest cohort of iFISH data with samples of 3023 patients is used (see section 3.6.1). The specific advantage of this cohort is, beside its size, that all patient populations (i.e. AL, MGUS, AMM, and MM) have been analyzed in the same way.
The distribution of CA in AL patients (see figure 3.27) has previously been described, and the findings in this thesis are consistence with the previously published results [23, 25, 28]. The most striking difference to MM in frequency of CA is the higher percentage of IgH-TL in AL, driven by the frequency of t(11;14) (see figure 3.28 **a**). Only 3% of AL patients with presence of t(11;14) concurrently fulfill the definition of HRD in agreement with previous findings on smaller patient cohorts [25]. AL patients show a significant lower number of high-risk aberrations, i.e. t(4;14), del 17p, gain 1q21 (especially three or more copies), or t(14;16) compared to MM (see figure 3.26). In the mean, malignant plasma cell in AL have less aggressive features compared to MM. This holds true for proliferation or high-risk CA or GEP-based risk determination. If however AL patients show high-risk MM-features, their prognosis is especially adverse (c.f. section 4.3).
Regarding the number of CA as assessed by iFISH, malignant plasma cells in AL patients mostly harbor one or more of the investigated aberrations (90% *versus* 10%, as depicted in table 3.14). This can be explained first by the high proportion of patients harboring a t(11;14) (with concomitantly fewer aberrations, especially those determining HRD that were counted as one), see figure 3.28 and section 1.2.3.

4.4.3 Copy number alterations assessed by WES

CNA were assessed using WES data of 113 AL and 28 MM patients. The analysis of CNA is outlined in section 2.5.4 and comprises a comparison of AL to MM by the assessed CNA and an analysis of expression of altered genes. See section 3.6.2 for depiction of results.
The presence of large gains on chromosome 9 and 11 seems to be the most frequent numerical aberration finding in AL patients, with presence in 32% and 36% of patients,

see figure 3.29 (Note: translocations like t(11;14) cannot be assessed by WES). This is somewhat a surprising finding as HRD (defined by the gain of two of the three chromosomes 5,9,15 [323]) is significantly less frequent in AL compared to MM (see section 4.4.2 above), but still remain the relatively most frequently detected aberrations. The total number deletions in AL, assessed by GISTIC, is higher than the number of gains (see table 3.15 **a**). But the number of genes in amplified regions is larger than the number in deleted regions (see table 3.4 **b**).

The CNA of AL patients were compared to MM. Both entities show significantly different results. The median number of CNA per patient sample in AL (9) is lower compared to MM (12.5) (see table 3.15 **b**). Whereas in AL more deletions than gains are found, the number of gains is higher compared to losses in MM (see table 3.15 **a**). This finding reflects the larger proportion of HRD (defined by gains) in MM compared to AL (as detected by iFISH, see section 3.6.1). Of the 50 cohort-wide CNA in AL, only eight (2%) are recurrent CNA, which are also present in MM (see table 3.16). This is likely driven first by the lack of recurrent CNA, and the relatively small comparator cohort in MM. Although the difference between AL and MM remained in a comparison with subsampling to 28 randomly chosen AL patient samples. The GISTIC score considers the amplitude and the frequency of occurrence of a CNA in the cohort [201]. The higher the score, the more distinct is the CNA. The CNA with the highest GISTIC scores, are among the overlapping CNA between the entities. Six of the eight CNA common to AL and MM show significant different frequencies (see table 3.16): Gain 19q13.42, del 2p11.2, gain 1p36.33 and gain 7q34 are significantly less frequent in AL than in MM. Del 22q11.22 and del 14q32.33 are more frequent in AL compared to MM.

The expression of the genes altered by a CNA in the AL samples was subsequently analyzed (see figure 3.30). Interestingly, expression of none of the genes suggested to be myeloma driver genes [308], i.e. *ARID1A*, *CDKN2C*, *FAM46C*, *FUBP1*, is altered in samples with gain 1q21 CNA compared to samples without, nor is *NRAS*.

A significant downregulation of the expression of the *RB1* gene in samples with a del 13q14.2 CNA is present. RB1 is a negative regulator of the cell cycle and as this a tumor suppressor [9] [283]. In MM, lower *RB1* expression was previously found to be associated with del 13q14 assessed by iFISH and a higher proliferation rate [128].

Another potential myeloma associated gene, either mutated or showing higher expression as consequence of the (rare) t(14;22) translocation [316], is *IGLL5* (located at the 22q11 cytoband). It is downregulated in expression in samples with del 22q11. *IGLL5*

[9] Genecards: RB Transcriptional Corepressor 1; Online resource: `https://www.genecards.org/cgi-bin/carddisp.pl?gene=RB1&keywords=RB1`; Status: 2019-09-20, 15:20

encodes an immunoglobulin polypeptide that does not require somatic rearrangement [10] [283]. Its upregulation in myeloma is bound to super enhancement by a translocation with the IgH locus [188] (described in the CA part in section 1.2.3).

4.4.4 Copy number alterations *versus* chromosomal aberrations

For validation of CNA, results in AL were compared to CA detected by iFISH in the same samples (see section 3.6.3). Here, sixteen CNA correspond to ten loci investigated with iFISH probes (see table 3.17). The overlap rate, indicated by the calculated efficiency between the iFISH results and the respective CNA is 83% (1q21.1), 85% (8p23.1), 81% (9q21.11), 80% (13q11), 81% (13q14.2), 85% (15q11.2), 89% (15q14), and 93% (19q13.42) (see table 3.17). A high overlap to CA at the same locus or at a near cytoband validates the assessment of CNA from WES data and is implicitly an indicator of sequencing quality. CNA involved in translocations, i.e. 11q13.4, 14q11.2, and 14q32.33 show less consistent overlap due to location in break-point regions [19]. For the remaining 37 CNA no corresponding iFISH probes were used routinely.

4.4.5 Differences in global gene expression patterns between AL and other malignant plasma cell entities

To identify patterns and generate hypotheses in multidimensional gene expression data, dimensions were reduced by PCA and t-SNE (see detailed description of methods in section 2.5.1 and results in section 3.6.4). Both methods are to a certain degree complementary. In homogenous data (e.g. cell lines, PPC, MBC, BMPC), the inter-sample variance is enlarged in PCA but not in t-SNE. For comparison between disease entities that are more heterogeneous, t-SNE shows larger variance between patient samples than PCA does.

PCA transforms data to as many PCs as variables (genes) exist, ordered by the variance they explain. Therefore, the first components hold the most variance. All input variables contribute to all PCs. Single PCs cannot be attributed to single input variables. In the analysis presented in this thesis, 10% of variance is explained by PC1 (5.7%) and PC2 (4.4%) in gene expression data from DNA microarray and 19% from RNA seq.

Taking into account all PCs, 100% variance is explained. In homogenous data, e.g. the same cell line, the first PC contains a higher amount of explained variance than in heterogeneous data e.g. patient samples. Here, more PCs are needed to explain

[10] Genecards: Immunoglobulin Lambda Like Polypeptide 5; Online resource: `https://www.genecards.org/cgi-bin/carddisp.pl?gene=IGLL5&keywords=IGLL5`; Status: 2019-09-20, 15:20

the same degree of variance. By t-SNE, all variance in the data is transformed and compressed into two dimensions. No variance, even low degree, is excluded from analysis. This makes t-SNE more applicable to heterogeneous data.

In the analyzed dataset, three main observations were made: first, methodologically, t-SNE is more applicable to gene expression data, analyzed in this thesis, than PCA. Within the sample groups HMCL, BMPC, MBC and PPC, a low biological variability was expected (scattering of points in the plot) and detected (see figure 3.32, 3.34 **b**). Likewise, the expected larger biological variability between the disease entities (AL, MGUS, AMM, and MM) and BMPC is visible. Between the disease entities and HMCL, a larger variance was detected. PPC and HMCL are grouped closer together than all other groups. Unifying biological factor between both is a high proliferation rate. Whether proliferation is induced by malignant transformation (as in HMCL) or physiologically (PPC) therefore has less impact than the biological process per se. In PCA this is more difficult to see (figure 3.32, 3.34 **a**).

Second, a large sample variance inside a disease entity is present. Ellipses for each disease entity are remarkably large compared to the ellipses for HMCL, BMPC, MBC and PPC (see the right side in figure 3.32, 3.34 **a** and **b**). This is likewise seen in PCA.

Third, the most interesting finding, is that only a small scattering distance is found between the compared malignant plasma cell disease entities, i.e. MGUS, AMM, MM, and AL. In PCA, ellipses for all disease entities overlap, indicating a high similarity (figure 3.31, 3.32, 3.33, 3.34 **a**, right side). In t-SNE AL and MM have tangential ellipses, likewise indication a high similarity between the disease entities (figure 3.32 **b**). This finding is in remarkable contrast to the (never proven) theory of ongoing clonal evolution in MM [115, 214]. When taking into account that the delineation of malignant plasma cell diseases by the IMWG [142] is based on tumor mass (and end organ damage), it is well conceivable that not the molecular background, but the tumor mass drives end organ damage. Thus, the main discerning factor between the entities is tumor mass. This does, of course, not exclude that different frequency of e.g. high-risk CA or more altered gene expression is present in a cohort of MM patients. The similarity between these entities is more remarkable as the proportion of underlying subentities is different between AL and the other malignant plasma cell disease entities, as exemplified by the frequency of (t11;14) and HRD. Of course, already a different proportion of t(11;14) between two disease groups could drive a certain difference in PCA or t-SNE assessments. It is thus conceivable that the found difference for PCA is smaller than for t-SNE.

The findings in this thesis disproof the previously published assumption that overall gene expression profiling of AL is different from other groups (BMPC, MGUS,

AMM, and MM) [160]. Here, Kryukov *et al.* [160] used unsupervised clustering and multi-dimensional scaling analysis on a small cohort of 16 AL, 15 BMPC, 21 MGUS, 24 AMM, and 69 MM samples. The applied method for group identification, hierarchical clustering, shows a large chance for potential pitfalls [253]. These are first, the dependence of results on the applied similarity measure, e.g. manhattan or euclidean distance. Second, hierarchical clusterings will always define clusters - even if significant variability in the data is missing [124]: the distance between the clusters is independent of the variability of the data. And third, input data is assumed to be of hierarchical origin, but this is not proven for malignant plasma cell disease entities. It is likely that this is not fully overcome by the use of multi-dimensional scaling by Kryukov *et al.* [160]. A further criticism is the lack of correction for molecular entities (e.g. accounting for different frequencies of t(11;14)). The main limitation of Kryukov *et al.* [160] can be seen however in the small sample size of the cohort (including the "increase of the molecular differences problem") [124, 153]. In summary, the dimension reduction analyses applied in this thesis on a large cohort of 1296 samples do not show distinct molecular profiles. This finding is in agreement with the AL-based GEP score being prognostic in MM, and *vice versa* (see discussion above in section 4.3.2).

4.4.6 Differential gene expression

In this thesis, differential gene expression analyses were performed on RNA seq data using the largest available cohort of 124 AL patients. A brief description of compared groups is depicted in section 2.5.2 in table 2.9. The results are described in section 3.6.5. Of all four comparisons between BMPC and the four malignant plasma cell disease entities (AL, MGUS, AMM, and MM), most DEG overlap with at least one other entity, indicating that the differences of malignant plasma cell diseases to healthy BMPC are mainly the same (see the Venn diagram in figure 3.35 **a**). This is in line with overlapping genes sharing the same direction of regulation i.e. being in all cases either up or downregulated compared to BMPC in both AL and MM.

The list of overlapping genes between BMPC and all four malignant plasma cell diseases and subentities (see supplementary table A.15) contains genes like *DKK1* and *HGF*, which are already known to be aberrantly expressed by malignant plasma cells in MM [130, 156, 282, 294] (see the part on altered gene expression in section 1.2.3). Here, "aberrant" indicates absence of expression in BMPC but not in malignant plasma cells.

The detected downregulation of *MYC* expression in AL to the level of BMPC expression (see table 3.21) is in agreement with a low rate of Myc-activation in AL compared to MM indicated by the MAI (see table A.7 and discussion on MAI in section 4.4.1).

A MAI > 1, indicating an enhanced Myc-activation but also induction of proliferation and cell cycle, is detected in 67 AL patients, this is 20% less frequent than in MM, but more than in MGUS.
MYC[11] codes for a unspecific transcriptional enhancer and proto oncogene. Translocations with *MYC* [11] are known to be relatively frequent in myeloma [307]. Myc enhances angiogenesis by enhancing the transcription of *VEGFA* [11] [276]. Both indicate a potential activation of Myc in AL, but at lower frequencies than in MM.
A large biological variation of gene expression in AL is indicated by comparing the DEG of AL *versus* MM to previous published lists of DEG with the same comparators. The DEG of this thesis show only a small overlap of 5 [1] and 3 [160] DEG with a LFC > 1 (see table 3.21). No overlap was detected to a previous comparison of AL *versus* BMPC. Besides biological variation, this is very likely due to small sample sizes in these previous studies and therefore an overfitting (see section 1.2.3): 9 [230], 16 [160], 24 [1] or 41 [6] AL patients were assessed, few for the study of heterogeneous diseases. For differential gene expression analysis from patient samples in contrast to samples from cell lines usage of a LFC threshold and large cohorts is beneficial (see section 2.5.2). Because small differences in expression height between groups could not be discerned from inter-patient variability by low fold changes and in small groups [49].

4.4.7 Immunoglobulin gene expression

Given the small difference in gene expression between AL and the other malignant plasma cell disease entities, it was assessed whether a specific difference exists in the expression of Ig genes. In other words, whether a specific expression pattern of light chain genes might explain the amyloid deposition (pattern) in AL in contrast to MM.
In section 3.6.6 the expression of 383 Ig genes is depicted regarding Ig type and malignant plasma cell entity. As depicted in figure 3.37, the expression of most Ig genes is low or not present in agreement with a polyclonal immune suppression by malignant plasma cells. As the accumulation of malignant (monoclonal) plasma cells takes place, polyclonal plasma cells are suppressed and outcompeted from their niches. Concomitantly, in assessment of purified plasma cells in bone marrow aspirates, expression of other than the dominant (monoclonal) light chain is less frequently found and depends on the overall tumor mass (i.e., number of malignant vs. normal (polyclonal) plasma cells). Ig genes from polyclonal plasma cells form the largest cluster of Ig genes (cluster A). For cluster 2, comprising a large proportion of AL patient samples (35%, see

[11] Genecards: MYC Proto-Oncogene; Online resource: `https://www.genecards.org/cgi-bin/carddisp.pl?gene=MYC&keywords=MYC`; Status: 2019-09-20, 15:20 [283]

table 3.23 **e**) a larger variance in expression of Ig genes is found. Thus, the degree of remaining polyclonal plasma cells drives the clustering. This is in agreement with AL patients showing a lower bone marrow infiltration. These patients form a subset with the cluster C of Ig genes. This is the Ig cluster with the highest expression values and does not consists of any Ig LC κ gene. The large proportion of AL patient samples showing high expression values in the Ig genes of cluster C is likely explained by the large proportion of AL patients expressing amyloid of λ LC type (79%, c.f. table 2.2). No disease entity did overlap with a single cluster. Cluster 1 is the most homogeneous with a proportion of 87% MM samples (see table 3.23 **e**), in agreement with the above discussed polyclonal suppression. In conclusion, no pattern of Ig expression driven by amyloid deposition as opposed to MM could be discerned, with the likely explanation of polyclonal gene expression driving clustering.

4.4.8 SNVs and InDels

In the following section the results of the variant call described in section 3.6.7 are discussed. All results for AL were analyzed in comparison to MM.

Mutational load

The mutational load indicated by the median number of variants in AL (22) is smaller than in MM (43) (see table 3.2 and table 3.3). The median number of variants in ALMG and ALMM are 22 and 23 and do not differ between the subentities. This is comparable to previously published studies: Paiva *et al.* [230] detected a median of 15 SNV per sample in 5 AL samples, Boyle *et al.* [36] detected 39 variants per sample in a cohort of 24 AL samples. Sample size and differences in applied VAF threshold of 5% [36] compared to 10% in this analysis likely explain this variation [36, 230]. Furthermore, differences in the applied WES method exists and underlying selection bias, i.e. the percentage of AL patients that can be assessed by WES depends on the number of malignant plasma cells purified and the amount of input DNA in WES, as already mentioned in the part on MM comparison in section 4.1.3. In seven AL samples more than 3000 non-synonymous variants were detected. In four of these samples more than 23% of variants are expressed, in one sample 10% are expressed and in two only a small percentage of 0.4% is expressed. The presence of hyper-mutated samples was not reported by the previous studies, which is an effect of the low sample number.

Similarities and differences between AL and MM

Variant tables of AL and MM were subsequently compared regarding Ti-Tv bias, variant types, and functional enrichment.

The Ti-Tv bias, i.e. the number of Ti and Tv in base substitutions, of AL and MM, depicted in figure 3.40, differ only slightly between the two entities. The most frequent (C>T, 43%/45%) and least frequent (T>A, 6%/8%) base substitution possibilities appear at comparable frequencies in both entities (AL/MM). The ratio of Ti-Tv is equal between AL and MM.

For both entities, the most common variant type was SNV (97%, see table 3.4 **a**) and the most common coding consequence was missense mutation (95%, see table 3.24 **a**).

In the FEA, most significantly enriched terms (83%) are shared between AL and MM for each of the investigated subsets of expressed, non-synonymous coding variants. Among the top 20 terms only one (R-HSA-3247509: Chromatin modifying enzymes) was unique to MM as depicted in figure 3.41.

In conclusion, general variant statistics between the variants assessed in the CoMMpass cohort for MM and in this thesis cohort for AL are equal. This is more striking as sample processing pipelines show considerable differences between both analyses.

Specific genes of interest

Three genes frequently mutated in MM were also found mutated in the AL cohort: *NRAS*, *KRAS* and *BRAF* (with the "actionable" V600E mutation). All three genes do harbor SNPs that were previously known to be associated to MM, again indication that the malignant plasma cells in AL and MM are comparable.

The clinical actionable *BRAF* mutation V600E (rs113488022, COSM476) has been detected in 2 AL patient samples on DNA level and once likewise as expressed on RNA level. V600E as well as V600K can be targeted by clinically available compounds (vermurafenib and dabrafenib [29, 122]) [10, 45].

A large overlap exists for genes previously suggested as potential "driver genes" (better termed frequently mutated genes as the relation of being driver has not been discerned). Out of 63 genes presented by Walker *et al.* [308] analyzing a cohort of 1273 MM patients, 44 (70%) genes were likewise found in the AL cohort analyzed in this thesis. Sixty-seven percent (76 of 113) of patients are affected by a variant or alteration in at least one of these genes. The incidence of mutations in these genes significantly differs between AL and MM only for three of the 44 genes (see figure 3.44 and the respective part on mutations in potential drivers in section 3.6.7). *NRAS* and *KRAS* are significantly more frequently mutated in MM. *BRAF* mutations is also more frequent in MM, but not significant. *KDM6A* is significantly more frequently mutated in AL. This implies that the mutational spectrum of AL and MM are essentially the same.

4.4.9 Molecular age of malignant plasma cells in AL

Malignant plasma cells in AL resemble plasma cell characteristics detected in all other malignant plasma cell disease entities (MGUS, AMM, and MM). GEP-based assessments of proliferation (GPI) and Myc-activation (MAI) can be applied to AL same as in MM, resulting in different frequencies of underlying group proportions (see section 4.4.1). The same CA as in the other entities do occur, but in AL-specific frequencies (see section 4.4.2). CA associated with high risk in MM, are significantly less frequent in AL. As shown in section 4.4.5 the inter-sample variance is larger than the variance between the different malignant plasma cell disease entities. The analysis of DEG showed that the different disease entities share a large overlap of genes regulated in the same direction compared to BMPC (c.f. section 4.4.6). And finally, genes which are frequently mutated in MM are also detected as mutated in AL (see section 4.4.8). AL seems to be no distinct molecular entity. The median "molecular age" in terms of the state during increasingly malignant behavior from BMPC to MGUS to AMM to MM to HMCL can be attributed as between MGUS and AMM.

4.5 Discussion of thesis aims

The primary aim of this thesis was to assess the molecular background of malignant plasma cell diseases, as described in section 1.7, especially considering three instances.
First, evaluate reports from prospective personalized risk assessment of MM patients generated within a clinical trial, which can be used for individualized and risk-adapted treatment strategies for MM patients.
Second, investigate to what degree prognosis of AL is determined by amyloidogenicity or to which extent by the properties of the malignant plasma cells.
Third, assess the similarities and differences of AL to MGUS, AMM, and MM by the malignant plasma cell characteristics, determine if AL is a unique molecular entity and if not, place the underlying malignant plasma cell disease of AL in relation to MM and the myeloma precursor stages AMM and MGUS to determine its "molecular age".
These aims mainly focus on assessing the pathogenic background (discussed in section 4.3 and 4.4). Nevertheless, besides the analytical part the analyses comprised a necessary computational part by the implementation of a WES analysis pipeline (described in section 2.4.4 and discussed in section 4.1.3), which was achieved.
Regarding the first aim of the thesis, i.e. the first two research questions, personalized risk assessment is possible in a reasonable time frame, shown by the evaluation of the MM5 clinical trial with 604 patients, for which GEP-based reports could been handed to treating physician in 456 cases.

The three research questions belonging to the second part could be answered in that the molecular properties of malignant plasma cells can determine prognosis as well as the independent parameters assessing the amyloidogenicity, e.g. Mayo and Euro scores (described in section 1.4, analyzed in section 3.3, and discussed in section 4.3.1). By risk assessment based on gene expression, the molecular properties were determined. AL showed the same underlying pattern compared to the other disease entities with small but significantly prognostic high risk groups (see section 3.4 and 4.3.2).

The hypothesis that the same malignant plasma cell factors determine prognosis in AL was validated by application of a risk score (HDAL) on AL and testing its prognostic significance on two independent groups of AMM and MM (see section 3.5.1 and 4.3.2). This also indicates that the underlying malignant plasma cell component in AL and MM is similar.

The research questions regarding the third aim could be answered in that on a patient level, malignant plasma cells in AL do not show main molecular differences to MGUS, AMM, or MM. Thus, AL is not a distinct molecular entity regarding malignant plasma cell characteristics. AL likewise do not resemble a stage earlier than MGUS. The median "molecular age" can be placed - as a result of this thesis - between MGUS and AMM. By dimension reduction the sample variance inside the disease entities and the variance between the entities was visualized, showing a larger variance between the samples than between the disease entities. Furthermore, by analyzing the overlap of DEG from the comparisons between BMPC and the disease entities similarities and differences were assessed, showing a large number of DEG shared between the disease entities and a smaller number of genes at a low magnitude explicit to the AL *versus* BMPC comparison (see section 3.6.5 and 4.4.6). Finally, the comparison of mutated genes in AL and MM was performed. Both entities harbor mutations in the same genes partially in varying frequencies, but mutational patterns of AL and MM were similar. Here, we found consistent results between all entities in case of gene expression and mutations. However differences exist on a population level, i.e. the large proportion of patients with t(11;14) (see section 3.6.1) or the small group of highly proliferative samples (see section 3.4) or the lower mutational load in AL compared to MM (see sections 3.1.3 and 3.6.7 and figure 3.39).

4.6 Conclusion

The conclusion of this thesis is separated in a methodological section, a part on personalized risk assessments for MM, and a part on risk assessment and pathogenesis in AL, just like the discussion above.

Conclusions on methodology

A WES analysis pipeline was implemented to call somatic variants in the hitherto largest group of 113 AL samples. The focus was on the detection of probably protein altering variants. By a versatile filtering strategy, including false positive filtering and validation by RNA, potential difficulties due to low input of sample material were overcome. The development strategy for the pipeline was intended to be applicable for research as well as clinical requirements, exemplified by the detection of specific mutations, e.g. V600E in *BRAF*.

Conclusions on personalized risk reports

The application of the GEP-R within the randomized phase III GMMG multicenter MM5 clinical trial proofs that it is possible to perform prospective target assessment and survival prediction for personalized and risk-adapted treatment in over 90% (iFISH) and 80% (GEP) of patients, respectively. The personalized report can be made available during the first cycle of induction, a time frame seen clinically reasonable. Risk assessment by GEP (i.e. HM metascore) compared to the international standard of rISS, shows a better Brier score (0.132 *versus* 0.137) and a delineation of more adverse five-year overall survival rate of 98%/68%/25% for low/medium/high risk HM metascore *versus* 86%/65%/40% for rISS stages I/II/III, respectively. A main benefit can be seen in the potential assessment of targets. (see section 4.2 above).

Conclusions on risk and pathogenesis

Till now, early prognosis and risk for AL patients was seen as defined by amyloid deposition and FLC secretion [75]. Malignant plasma cell characteristics, represented by CA assessed by iFISH, were only considered regarding treatment decisions and as such defining prognosis during the course of disease (primarily late prognosis) [75]. Within this thesis, it is shown that early prognosis in AL is also defined by molecular factors of plasma cells, i.e. proliferation (GPI) or GEP-based risk scores that resemble these factors (UAMS70, RS).

The genetic landscape of AL was assessed by a plethora of methods (gene expression, molecular risk stratification, CA, CNA and somatic variants), all indicating that the malignant plasma cells in AL resemble the characteristics of other malignant plasma cell diseases like MGUS and MM. This was hypothesized previously by Seckinger *et al.* [263] at the LfM and others [36, 230, 254] and could now be verified on a large cohort of patient samples by diverse methods. Thereby, the hypothesis that AL is a molecular distinct entity from MM, with a unique molecular signature [1], can be rejected.

The comparison of dimension reduction methods and the little overlap with the published DEGs indicate that the AL plasma cells hold a great inter-patient variability in case of gene expression, similar to MM. The wide variety of variants and CNA in patient samples is further in line.
Given that all MM subgroups are also present in AL, although at different frequencies, the ability of malignant plasma cells to produce amyloid can appear on every molecular background. Although, it happens more frequently in case of t(11;14). As expected, no unifying mutation in AL was detected, similar to MM, as previously suggested by Boyle *et al.* [36] and Paiva *et al.* [230].
In a nutshell, in respect to gene expression, molecular risk stratification, CA, CNA, and mutations AL is still a malignant plasma cell disease in the mean between MGUS and AMM - atop an "unlucky" LC [123].

4.7 Outlook

Regarding pathogenesis and prognosis of AL, two main directions of future research can be envisioned, both regarding amyloidogenicity of LC. First, the actual *de novo* alignment of LC and HC gene segments from the RNA sequencing data, i.e. a reconstruction of the complete LC and (in case of co-presence of intact Ig) HC. Second, the individual assessment of the protein subunits (M-protein) that form the amyloid fibrils by mass spectrometry methods like the MALDI-TOF assay MASS-FIX [206, 207, 271]. These analyses showed to be more sensitive than the present methods protein electrophoresis and immunofixation electrophoresis to identify, isotype and quantify the M-protein [206, 207, 271]. Recently, amyloid fibrils were analyzed by electron microscopy to assess the exact structure of amyloids [240, 287] in the expectation to identify disease relevant characteristics. A further direction possible based on the findings and work described in this thesis is the implementation of prospective target assessment and survival prediction for personalized and risk-adapted treatment - as shown for MM in the GMMG-MM5 trial - for AL.

5 Summary

Light chain amyloidosis is a malignant plasma cell disease characterized by the production and secretion of immunoglobulin light chains, aggregating as amyloid and causing end organ damage, most frequently in heart and kidney.
The primary aim of this thesis was to assess the molecular background and prognosis of light chain amyloidosis in relation to other malignant plasma cell diseases.
CD138-positive purified malignant plasma cells from patients diagnosed with light chain amyloidosis, monoclonal gammopathy of undetermined significance, asymptomatic and symptomatic multiple myeloma were subjected to interphase fluorescence *in situ* hybridization ($n = 582/306/444/1691$), gene expression profiling by DNA-microarrays ($n = 196/64/271/765$), RNA sequencing ($n = 124/51/140/515$), and whole exome sequencing (light chain amyloidosis $n = 113$). Clinical and survival data were collected.
First, it was shown that for multiple myeloma risk assessment by the gene expression-based HM metascore compared to the current gold-standard (rISS) shows superior delineation of five-year overall survival rate of 98%/68%/25% for low/medium/high risk HM metascore versus 86%/65%/40% for rISS stages I/II/III, respectively.
Second, it was investigated to what degree prognosis of light chain amyloidosis is determined by free light chain production and amyloid deposition and to what degree by properties of the malignant plasma cells. Till now, early prognosis and risk for patients was seen as defined by amyloid deposition and free light chain secretion, with risk assessment based on serum parameters of cardiac involvement and the difference in free light chains in the serum. Malignant plasma cell characteristics in light chain amyloidosis, represented by chromosomal aberrations, were considered regarding treatment decisions and as defining prognosis rather late during the course of disease.
In this thesis, it was shown that prognosis of light chain amyloidosis patients is likewise driven by malignant plasma cell factors accessible by gene expression profiling independent of light chain deposition-based factors. This especially holds true for gene expression-based assessment of proliferation (GPI), myeloma-based risk scores (UAMS70, RS), and the *de novo* for light chain amyloidosis patients generated HDAL score. The hypothesis that the same malignant plasma cell factors determine prognosis in light chain amyloidosis and myeloma was validated by testing the HDAL score's prognostic significance on two independent groups of asymptomatic and symptomatic multiple myeloma patients. This indicates that the underlying malignant plasma cell component in both diseases is similar. Thus, early prognosis is also defined by molecular factors of plasma cells.

Third, the similarities and differences regarding malignant plasma cell characteristics were assessed, to determine if light chain amyloidosis is a unique molecular entity and if not, place the underlying malignant plasma cell disease of it in relation to multiple myeloma and its precursor stages to determine a "molecular age".
On a patient level, malignant plasma cells in light chain amyloidosis do not show mayor molecular differences to myeloma. This was shown by dimension reduction, analysis of differentially expressed and mutated genes. A larger variance was detected within samples than between different disease entities. Thus, light chain amyloidosis is not a distinct molecular entity regarding malignant plasma cell characteristics, and it likewise does not resemble a stage earlier than monoclonal gammopathy of undetermined significance. Differences exist on a population level, i.e. in light chain amyloidosis a large proportion of patients with translocation t(11;14), a small group of highly proliferative samples, and a lower mutational load on average compared to multiple myeloma. Given that all myeloma subgroups are also present in light chain amyloidosis, although at different frequencies, the ability of malignant plasma cells to produce amyloid generating light chains can appear on every myeloma associated molecular background.
In a nutshell, in respect to gene expression, molecular risk stratification, chromosomal aberrations, copy number alterations, and mutations light chain amyloidosis is a malignant plasma cell disease with a median "molecular age" between monoclonal gammopathy of undetermined significance and asymptomatic myeloma - atop an "unlucky" light chain.

6 Zusammenfassung

Leichtketten Amyloidose ist eine maligne Plasmazellerkrankung charakterisiert durch die Produktion und Sekretion von Immunglobulin Leichtketten, welche als Amyloid zusammenlagern und dadurch Organschäden verursachen, am häufigsten in Herz und Niere.

Ziel dieser Dissertation war es, molekularen Hintergrund und Prognose der Leichtketten Amyloidose zu bestimmen und zu anderen malignen Plasmazellerkrankungen in Bezug zu setzen.

CD138-positive maligne Plasmazellen von Patienten mit Leichtketten Amyloidose, Monoklonaler Gammopathie unklarer Signifikanz, asymptomatischem und symptomatischem Multiplen Myelom wurden aufgereinigt und mittels Interphase Fluoreszenz *in situ* Hybridisierung ($n = 582/306/444/1691$), Genexpressionsanalyse mit DNA-microarrays ($n = 196/64/271/765$), RNA Sequenzierung ($n = 124/51/140/515$), und Exome Sequenzierung (Leichtketten Amyloidose $n = 113$) untersucht. Klinische und Überlebenszeitdaten wurden erhoben.

Als erstes wurde gezeigt, dass hinsichtlich Risikobewertung beim Multiple Myelom der Genexpressions-basierte HM Metascore im Vergleich zum aktuellen Goldstandard rISS eine bessere Auftrennung bezüglich des Gesamtüberleben zeigt. Mit einer jeweiligen Rate von 98%/68%/25% beim HM metascore für niedriges/mittleres/hohes Risiko gegen 86%/65%/40% für rISS Stadien I/II/III nach fünf Jahren.

Zweitens wurde untersucht, wie weit die Prognose der Leichtketten Amyloidose durch die Produktion freier Leichtketten und der Deposition von Amyloid bestimmt wird und in wie weit durch molekulare Eigenschaften der malignen Plasmazellen. Bisher wurde angenommen, dass die frühe Prognose für die Patienten durch die Produktion der freien Leichtketten und die Amyloid Ablagerungen bestimmt wird. Daher basiert die Risikobewertung auf Serum Parametern der Herzbeteiligung und der Differenz der freien Leichtketten im Serum. Maligne Plasmazellcharakteristika, wie chromosomalen Aberrationen, werden im Hinblick auf Therapieentscheidungen berücksichtigt und werden als eher spät im Krankheitsverlauf Einfluss auf die Prognose nehmend angesehen.

In dieser Dissertation wurde gezeigt, dass die Prognose bei Leichtketten Amyloidose Patienten unabhängig von Leichtketten Depositions-basierten Faktoren auch durch molekulare maligne Plasmazellfaktoren bestimmt wird, die durch Genexpressionsanalysen erfassbar sind. Dies beinhaltet die Genexpression-basierte Messung von Proliferation (GPI), Myelom-basierte Risikobewertungen (UAMS70, RS), sowie um die für Leichtketten Amyloidose Patienten neuentwickelte HDAL Stratifikation. Zur Bestätigung der Hypothese, dass die gleichen malignen Plasmazellfaktoren die Prognose

in der Leichtketten Amyloidose und dem Myelom bestimmen, wurde der HDAL auf zwei unabhängigen Kohorten von asymptomatischen und symptomatischen Multiplen Myelom Patienten getestet und seine prognostische Signifikanz gezeigt. Dies bedeutet, dass die grundlegende maligne Plasmazellkomponente in beiden Erkrankungen gleich ist. Die Prognose im frühen Krankheitsverlauf ist somit auch durch die malignen Plasmazellfaktoren bestimmt.
Drittens wurden Gemeinsamkeiten und Unterschiede der molekularen Charakteristika maligner Plasmazellen untersucht, um zu analysieren ob die Leichtketten Amyloidose eine eigene molekulare Entität darstellt, und falls nicht, die ihr zu Grunde liegende maligne Plasmazellerkrankung in Relation zum Myelom und seinen Vorstadien zu stellen, und dadurch ein "molekulares Alter" bestimmen zu können.
Auf Ebene eines individuellen Patienten gibt es kein Anzeichen für molekulare Unterschiede maligner Plasmazellen zwischen Leichtketten Amyloidose und Multiplen Myelom. Das konnten mittels Dimensionsreduktionsverfahren sowie der Analyse von differentiell exprimierten und mutierten Genen gezeigt werden. Es wurde eine größere Varianz zwischen einzelnen Patientenproben innerhalb einer Entität als zwischen den Krankheitsentitäten gefunden. Die Leichtketten Amyloidose ist somit keine eigenständige molekulare Entität in Hinblick auf maligne Plasmazellcharakteristika. Sie ist ebenfalls keine Vorstufe der Monoklonalen Gammopathie unklarer Signifikanz. Auf Populationsebene sind Unterschiede sichtbar. Die Leichtketten Amyloidose ist charakterisiert durch einen größeren Anteil an Patienten mit Translokation t(11;14), einer kleineren Gruppe mit hoher Proliferationsrate, und einer niedrigeren durchschnittlichen Mutationslast im Vergleich zum Multiplen Myelom. Vor dem Hintergrund, dass alle aus dem Myelom bekannten Subgruppen auch bei der Leichtketten Amyloidose vorkommen, wenn auch mit unterschiedlichen Häufigkeiten, kann die Produktion zu Amyloid aggregierenden Leichtketten durch maligne Plasmazellen vor jedem Myelom-assoziierten molekularen Hintergrund auftreten.
In der Quintessenz ist die Leichtketten Amyloidose bezüglich Genexpression, molekularer Risikostratifizierung, chromosomalen Aberrationen, Veränderungen der Kopienzahl, und Mutationen eine maligne Plasmazellerkrankung mit einem medianen "molekularen Alter" zwischen Monoklonaler Gammopathie unklarer Signifikanz und asymptomatischen Myelom - die eine "unglückliche" Leichtkette produziert.

7 References

[1] Abraham, R. S., Ballman, K. V., Dispenzieri, A., Grill, D. E., Manske, M. K., Price-Troska, T. L., Paz, N. G., Gertz, M. A., and Fonseca, R. (2005). **Functional gene expression analysis of clonal plasma cells identifies a unique molecular profile for light chain amyloidosis**. *Blood*, 105:794–803. doi:10.1182/blood-2004-04-1424.

[2] Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). **A method and server for predicting damaging missense mutations**. *Nature Methods*, 7:248 EP –. doi:10.1038/nmeth0410-248.

[3] Affymetrix (2020). **Data Sheet: Gene Chip Human Genome U133 Array: The Most Comprehensive Coverage of the Human Genome in TwoFlexible Formats: Single-array Cartridges and Multi-array Plates**. URL: `https://assets.thermofisher.com/TFS-Assets/LSG/brochures/hgu133arrays_datasheet.pdf`. [Last visited: 13.02.20].

[4] Afrough, A., Saliba, R. M., Hamdi, A., Honhar, M., Varma, A., Cornelison, A. M., Rondon, G., Parmar, S., Shah, N. D., Bashir, Q., Hosing, C., Popat, U., Weber, D. M., Thomas, S., Orlowski, R. Z., Champlin, R. E., and Qazilbash, M. H. (2018). **Impact of Induction Therapy on the Outcome of Immunoglobulin Light Chain Amyloidosis after Autologous Hematopoietic Stem Cell Transplantation**. *Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation*, 24:2197–2203. doi:10.1016/j.bbmt.2018.07.010.

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

[6] Alameda, D., Vicari, M., Lara-Astiaso, D., Lasa, M., Puig, N., Cedena Romero, M. T., Rodriguez, I., Alignani, D., Vilas-Zornoza, A., Goicoechea, I., Sarvide, S., Ocio, E. M., Lecumberri, R., García de Coca, A., Labrador, J., Garcia, E. G., Palomera, L., Gironella, M., Cabañas, V., Casanova, M., Oriol, A., Krsnik, I., Pérez, A., Martinez Lopez, J., Mateos, M.-V., Lahuerta, J.-J., Prosper, F., Weiner, A., Amit, I., San-Miguel, J. F., and Paiva, B. (2018). **Understanding the Cellular Origin and Pathogenic Transcriptional Programs in Multiple Myeloma (MM) and Light-Chain Amyloidosis (AL) through the Dissection of the Normal Plasma Cell (PC) Development**. *Blood*, 132:188. doi:10.1182/blood-2018-188, ASH Annual Meeting Abstract.

[7] 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.

[8] 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.

[9] 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].

[10] 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.

[11] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). **Gene ontology Tool for the unification of biology. The Gene Ontology Consortium**. *Nature genetics*, 25:25–29. doi:10.1038/75556.

[12] Attal, M., Lauwers-Cances, V., Hulin, C., Leleu, X., Caillot, D., Escoffre, M., Arnulf, B., Macro, M., Belhadj, K., Garderet, L., Roussel, M., Payen, C., Mathiot, C., Fermand, J. P., Meuleman, N., Rollet, S., Maglio, M. E., Zeytoonjian, A. A., Weller, E. A., Munshi, N., Anderson, K. C., Richardson, P. G., Facon, T., Avet-Loiseau, H., Harousseau, J.-L., and Moreau, P. (2017). **Lenalidomide, Bortezomib, and Dexamethasone with Transplantation for Myeloma**. *The New England journal of medicine*, 376:1311–1320. doi:10.1056/NEJMoa1611750.

[13] Auguie, B. (2017). **gridExtra Miscellaneous Functions for Grid Graphics**, version 2.3.

[14] Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). **A global reference for human genetic variation**. *Nature*, 526:68–74. doi:10.1038/nature15393.

[15] Balasubramanian, S. (2011). **Sequencing nucleic acids From chemistry to medicine**. *Chemical communications (Cambridge, England)*, 47:7281–7286. doi:10.1039/c1cc11078k.

[16] 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.

[17] Beck, S., Emde, M., Moreaux, J., Seckinger, A., and Hose, D. (2019). **Prediction of Malingant Plasma Cell Biology Related Survival in AL-Amyloidosis**. *Blood*, 134:3078. doi:10.1182/blood-2019-131161, ASH Annual Meeting Abstract and Poster.

[18] 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*, 57:289–300.

[19] Bergsagel, P. L. and Kuehl, W. M. (2001). **Chromosome translocations in multiple myeloma**. *Oncogene*, 20:5611–5622. doi:10.1038/sj.onc.1204641.

[20] 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.

[21] Biolo, A., Ramamurthy, S., Connors, L. H., O'Hara, C. J., Meier-Ewert, H. K., Soo Hoo, P. T., Sawyer, D. B., Seldin, D. C., Seldin, D. S., and Sam, F. (2008). **Matrix metalloproteinases and their tissue inhibitors in cardiac amyloidosis Relationship to structural, functional myocardial changes and to light chain amyloid deposition**. *Circulation. Heart failure*, 1:249–257. doi:10.1161/CIRCHEARTFAILURE.108.788687.

[22] Blancas-Mejía, L. M. and Ramirez-Alvarado, M. (2013). **Systemic amyloidoses**. *Annual review of biochemistry*, 82:745–774. doi:10.1146/annurev-biochem-072611-130030.

[23] Bochtler, T., Hegenbart, U., Cremer, F. W., Heiss, C., Benner, A., Hose, D., Moos, M., Bila, J., Bartram, C. R., Ho, A. D., Goldschmidt, H., Jauch, A., and Schonland, S. O. (2008). **Evaluation of the cytogenetic aberration pattern in amyloid light chain amyloidosis as compared with monoclonal gammopathy of undetermined significance reveals common pathways of karyotypic instability**. *Blood*, 111:4700–4705. doi:10.1182/blood-2007-11-122101.

[24] Bochtler, T., Hegenbart, U., Heiss, C., Benner, A., Cremer, F., Volkmann, M., Ludwig, J., Perz, J. B., Ho, A. D., Goldschmidt, H., and Schonland, S. O. (2008). **Evaluation of the serum-free light chain test in untreated patients with AL amyloidosis**. *Haematologica*, 93:459–462. doi:10.3324/haematol.11687.

[25] Bochtler, T., Hegenbart, U., Heiss, C., Benner, A., Moos, M., Seckinger, A., Pschowski-Zuck, S., Kirn, D., Neben, K., Bartram, C. R., Ho, A. D., Goldschmidt, H., Hose, D., Jauch, A., and Schonland, S. O. (2011). **Hyperdiploidy is less frequent in AL amyloidosis compared with monoclonal gammopathy of undetermined significance and inversely associated with translocation t(11;14)**. *Blood*, 117:3809–3815. doi:10.1182/blood-2010-02-268987.

[26] Bochtler, T., Hegenbart, U., Kunz, C., Benner, A., Seckinger, A., Dietrich, S., Granzow, M., Neben, K., Goldschmidt, H., Ho, A. D., Hose, D., Jauch, A., and Schönland, S. O. (2014). **Gain of chromosome 1q21 is an independent adverse prognostic factor in light chain amyloidosis patients treated with melphalan/dexamethasone**. *Amyloid*, 21:9–17. doi:10.3109/13506129.2013.854766.

[27] Bochtler, T., Hegenbart, U., Kunz, C., Granzow, M., Benner, A., Seckinger, A., Kimmich, C., Goldschmidt, H., Ho, A. D., Hose, D., Jauch, A., and Schönland, S. O. (2015). **Translocation t(11;14) is associated with adverse outcome in patients with newly diagnosed AL amyloidosis when treated with bortezomib-based regimens**. *Journal of clinical oncology*, 33:1371–1378. doi:10.1200/JCO.2014.57.4947.

[28] Bochtler, T., Merz, M., Hielscher, T., Granzow, M., Hoffmann, K., Krämer, A., Raab, M.-S., Hillengass, J., Seckinger, A., Kimmich, C., Dittrich, T., Müller-Tidow, C., Hose, D., Goldschmidt, H., Hegenbart, U., Jauch, A., and Schönland, S. O. (2018). **Cytogenetic intraclonal heterogeneity of plasma cell dyscrasia in AL amyloidosis as compared with multiple myeloma**. *Blood advances*, 2:2607–2618. doi:10.1182/bloodadvances.2018023200.

[29] 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.

[30] Bolli, N., Avet-Loiseau, H., Wedge, D. C., van Loo, P., Alexandrov, L. B., Martincorena, I., Dawson, K. J., Iorio, F., Nik-Zainal, S., Bignell, G. R., Hinton, J. W., Li, Y., Tubio, J. M., McLaren, S., O' Meara, S., Butler, A. P., Teague, J. W., Mudie, L., Anderson, E., Rashid, N., Tai, Y.-T., Shammas, M. A., Sperling, A. S., Fulciniti, M., Richardson, P. G., Parmigiani, G., Magrangeas, F., Minvielle, S., Moreau, P., Attal, M., Facon, T., Futreal, P. A., Anderson, K. C., Campbell, P. J., and Munshi, N. C. (2014). **Heterogeneity of genomic evolution and mutational profiles in multiple myeloma**. *Nature communications*, 5:2997. doi:10.1038/ncomms3997.

[31] Bolli, N., Biancon, G., Moarii, M., Gimondi, S., Li, Y., de Philippis, C., Maura, F., Sathiaseelan, V., Tai, Y.-T., Mudie, L., O'Meara, S., Raine, K., Teague, J. W., Butler, A. P., Carniti, C., Gerstung, M., Bagratuni, T., Kastritis, E., Dimopoulos, M., Corradini, P., Anderson, K. C., Moreau, P., Minvielle, S., Campbell, P. J., Papaemmanuil, E., Avet-Loiseau, H., and Munshi, N. C. (2018). **Analysis of the genomic landscape of multiple myeloma highlights novel prognostic markers and disease subgroups**. *Leukemia*, 32:2604–2616. doi:10.1038/s41375-018-0037-9.

[32] Bolstad, B. M., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R. A., and Speed, T. P. (2005). **Quality Assessment of Affymetrix GeneChip Data**. Springer New York. doi:10.1007/0-387-29362-0_3.

[33] Boomsma, F. and van den Meiracker, A. H. (2001). **Plasma A- and B-type natriuretic peptides Physiology, methodology and clinical use**. *Cardiovascular research*, 51:442–449.

[34] 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.

[35] Boyd, K. D., Ross, F. M., Tapper, W. J., Chiecchio, L., Dagrada, G., Konn, Z. J., Gonzalez, D., Walker, B. A., Hockley, S. L., Wardell, C. P., Gregory, W. M., Child, J. A., Jackson, G. H., Davies, F. E., and Morgan, G. J. (2011). **The clinical impact and molecular biology of del(17p) in multiple myeloma treated with conventional or thalidomide-based therapy**. *Genes, chromosomes & cancer*, 50:765–774.

[36] Boyle, E. M., Ashby, C., Wardell, C. P., Rowczenio, D., Sachchithanantham, S., Wang, Y., Johnson, S. K., Bauer, M. A., Weinhold, N., Kaiser, M. F., Johnson, D. C., Jones, J. R., Pawlyn, C., Proszek, P., Schinke, C., Facon, T., Dumontet, C., Davies, F. E., Morgan, G. J., Walker, B. A., and Wechalekar, A. D. (2018). **The genomic landscape of plasma cells in systemic light chain amyloidosis**. *Blood*, 132:2775–2777. doi:10.1182/blood-2018-08-872226.

[37] Bradwell, A. R., Carr-Smith, H. D., Mead, G. P., Tang, L. X., Showell, P. J., Drayson, M. T., and Drew, R. (2001). **Highly sensitive, automated immunoassay for immunoglobulin free light chains in serum and urine**. *Clinical chemistry*, 47:673–680.

[38] Broad Institute (2017). **Picard Tools java 1.8**. Broad Institute of MIT and Harvard, Cambridge, MA, version 2.10.10. URL: `http://broadinstitute.github.io/picard/`. [Last visited: 29.08.18].

[39] Bryce, A. H., Ketterling, R. P., Gertz, M. A., Lacy, M., Knudson, R. A., Zeldenrust, S., Kumar, S., Hayman, S., Buadi, F., Kyle, R. A., Greipp, P. R., Lust, J. A., Russell, S., Rajkumar, S. V., Fonseca, R., and Dispenzieri, A. (2009). **Translocation t(11;14) and survival of patients with light chain (AL) amyloidosis**. *Haematologica*, 94:380–386. doi:10.3324/haematol.13369.

[40] Burrows, M. and Wheeler, D. (1994). **A Block-sorting lossless data compression algorithm**. *Technical report 124. Palo Alto, CA: Digital Equipment Corporation*.

[41] Carlson, M. (2016). **hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2)**, version 3.2.3.

[42] Carlson, M. (2016). **org.Hs.eg.db Genome wide annotation for Human**, version 3.4.0.

[43] Cavo, M., Beksac, M., Dimopoulos, M. A., Pantani, L., Gay, F., Hájek, R., Testoni, N., Mellqvist, U.-H., Patriarca, F., Montefusco, V., Galli, M., Johnsen, H. E., Ludwig, H., Zweegman, S., Wester, R., Wu, K. L., Driessen, C., Troia, R., Cornelisse, P., van der Holt, B., Palumbo, A., and Sonneveld, P. (2016). **Intensification Therapy with Bortezomib-Melphalan-Prednisone Versus Autologous Stem Cell Transplantation for Newly Diagnosed Multiple Myeloma: An Intergroup, Multicenter, Phase III Study of the European Myeloma Network (EMN02/HO95 MM Trial)**. *Blood*, 128:673. doi:10.1182/blood.V128.22.673.673.

[44] Chao, D. T. and Korsmeyer, S. J. (1998). **BCL-2 family Regulators of cell death**. *Annual review of immunology*, 16:395–419. doi:10.1146/annurev.immunol.16.1.395.

[45] 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.

[46] Chen, H. (2018). **VennDiagram Generate High-Resolution Venn and Euler Plots**, version 1.6.20. URL: `https://CRAN.R-project.org/package=VennDiagram`.

[47] Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). **fastp An ultra-fast all-in-one FASTQ preprocessor**. *Bioinformatics*, 34:i884–i890. doi:10.1093/bioinformatics/bty560.

[48] Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., and Saunders, C. T. (2016). **Manta Rapid detection of structural variants and indels for germline and cancer sequencing applications**. *Bioinformatics*, 32:1220–1222. doi:10.1093/bioinformatics/btv710.

[49] Chen, Y., McCarthy, D., Ritchie, M., Robinson, M., and Smyth, G. (2019). **edgeR: differential expression analysis of digital gene expression data User's Guide**, version 13 August 2019. URL: `https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf`. [Last visited: 08.10.2019].

[50] Chesi, M., Bergsagel, P. L., Shonukan, O. O., Martelli, M. L., Brents, L. A., Chen, T., Schröck, E., Ried, T., and Kuehl, W. M. (1998). **Frequent dysregulation of the c-maf proto-oncogene at 16q23 by translocation to an Ig locus in multiple myeloma**. *Blood*, 91:4457–4463.

[51] Chesi, M., Nardini, E., Lim, R. S., Smith, K. D., Kuehl, W. M., and Bergsagel, P. L. (1998). **The t(4;14) translocation in myeloma dysregulates both FGFR3 and a novel gene, MMSET, resulting in IgH/MMSET hybrid transcripts**. *Blood*, 92:3025–3034.

[52] Chng, W. J., Chung, T.-H., Kumar, S., Usmani, S., Munshi, N., Avet-Loiseau, H., Goldschmidt, H., Durie, B., and Sonneveld, P. (2016). **Gene signature combinations improve prognostic stratification of multiple myeloma patients**. *Leukemia*, 30:1071–1078. doi:10.1038/leu.2015.341.

[53] 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.

[54] 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.

[55] Chng, W.-J., Huang, G. F., Chung, T. H., Ng, S. B., Gonzalez-Paz, N., Troska-Price, T., Mulligan, G., Chesi, M., Bergsagel, P. L., and Fonseca, R. (2011). **Clinical and biological implications of MYC activation A common difference between MGUS and newly diagnosed multiple myeloma**. *Leukemia*, 25:1026–1035. doi:10.1038/leu.2011.53.

[56] Christoforides, A., Carpten, J. D., Weiss, G. J., Demeure, M. J., Hoff, D. D. V., and Craig, D. W. (2013). **Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs**. *BMC genomics*, 14:302. doi:10.1186/1471-2164-14-302.

[57] 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.

[58] Cohen, A. D., Landau, H., Scott, E. C., Liedtke, M., Kaufman, J. L., Rosenzweig, M., Gasparetto, C., Vesole, D. H., Sanchorawala, V., Lentzsch, S., Gomes, C. L., Comenzo, R. L., and Durie, B. G. (2016). **Safety and Efficacy of Carfilzomib (CFZ) in Previously-Treated Systemic Light-Chain (AL) Amyloidosis**. *Blood*, 128:645, ASH Annual Meeting Abstract.

[59] Cohen, J. (1960). **A Coefficient of Agreement for Nominal Scales**. *Educational and Psychological Measurement*, 20:37–46. doi:10.1177/001316446002000104.

[60] Costain, G., Jobling, R., Walker, S., Reuter, M. S., Snell, M., Bowdin, S., Cohn, R. D., Dupuis, L., Hewson, S., Mercimek-Andrews, S., Shuman, C., Sondheimer, N., Weksberg, R., Yoon, G., Meyn, M. S., Stavropoulos, D. J., Scherer, S. W., Mendoza-Londono, R., and Marshall, C. R. (2018). **Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing**. *European journal of human genetics : EJHG*, 26:740–744. doi:10.1038/s41431-018-0114-6.

[61] 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.

[62] 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.

[63] Crozier, R. H. and Crozier, Y. C. (1993). **The mitochondrial genome of the honeybee Apis mellifera Complete sequence and genome organization**. *Genetics*, 133:97–117.

[64] Culhane, A. C., Perrière, G., and Higgins, D. G. (2003). **Cross-platform comparison and visualisation of gene expression data using co-inertia analysis**. *BMC Bioinformatics*, 4:59. doi:10.1186/1471-2105-4-59.

[65] Culhane, A. C., Thioulouse, J., Perrière, G., and Higgins, D. G. (2005). **MADE4: an R package for multivariate analysis of gene expression data**. *Bioinformatics*, 21:2789–2790. doi:10.1093/bioinformatics/bti394.

[66] Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J. (2018). **xtable: Export Tables to LaTeX or HTML**, version 1.8-3. URL: `https://CRAN.R-project.org/package=xtable`.

[67] 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.

[68] 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 Clinical Oncology*, 26:4798–4805.

[69] 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.

[70] Dispenzieri, A., Gertz, M. A., Kyle, R. A., Lacy, M. Q., Burritt, M. F., Therneau, T. M., Greipp, P. R., Witzig, T. E., Lust, J. A., Rajkumar, S. V., Fonseca, R., Zeldenrust, S. R., McGregor, C. G. A., and Jaffe, A. S. (2004). **Serum cardiac troponins and N-terminal pro-brain natriuretic peptide A staging system for primary systemic amyloidosis**. *Journal of clinical oncology*, 22:3751–3757. doi:10.1200/JCO.2004.03.029.

[71] Dispenzieri, A., Kyle, R. A., Gertz, M. A., Therneau, T. M., Miller, W. L., Chandrasekaran, K., McConnell, J. P., Burritt, M. F., and Jaffe, A. S. (2003). **Survival in patients with primary systemic amyloidosis and raised serum cardiac troponins**. *The Lancet*, 361:1787–1789. doi:10.1016/S0140-6736(03)13396-X.

[72] Dispenzieri, A., Lacy, M. Q., Katzmann, J. A., Rajkumar, S. V., Abraham, R. S., Hayman, S. R., Kumar, S. K., Clark, R., Kyle, R. A., Litzow, M. R., Inwards, D. J., Ansell, S. M., Micallef, I. M., Porrata, L. F., Elliott, M. A., Johnston, P. B., Greipp, P. R., Witzig, T. E., Zeldenrust, S. R., Russell, S. J., Gastineau, D., and Gertz, M. A. (2006). **Absolute values of immunoglobulin free light chains are prognostic in patients with primary systemic amyloidosis undergoing peripheral blood stem cell transplantation**. *Blood*, 107:3378–3383. doi:10.1182/blood-2005-07-2922.

[73] Dittrich, T., Benner, A., Kimmich, C., Siepen, F. A. d., Veelken, K., Kristen, A. V., Bochtler, T., Katus, H. A., Müller-Tidow, C., Hegenbart, U., and Schönland, S. O. (2019). **Performance analysis of AL amyloidosis cardiac biomarker staging systems with special focus on renal failure and atrial arrhythmia**. *Haematologica*, 104:1451–1459. doi:10.3324/haematol.2018.205336.

[74] Dittrich, T., Bochtler, T., Kimmich, C., Becker, N., Jauch, A., Goldschmidt, H., Ho, A. D., Hegenbart, U., and Schönland, S. O. (2017). **AL amyloidosis patients with low amyloidogenic free light chain levels at first diagnosis have an excellent prognosis**. *Blood*, 130:632–642. doi:10.1182/blood-2017-02-767475.

[75] Dittrich, T., Kimmich, C., Hegenbart, U., and Schönland, S. O. (2020). **Prognosis and Staging of AL Amyloidosis**. *Acta haematologica*, 143:388–400. doi:10.1159/000508287.

[76] 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.

[77] Dumas, B., Yameen, H., Sarosiek, S., Sloan, J. M., and Sanchorawala, V. (2020). **Presence of t(11;14) in AL amyloidosis as a marker of response when treated with a bortezomib-based regimen**. *Amyloid : the international journal of experimental and clinical investigation : the official journal of the International Society of Amyloidosis*, 27:244–249. doi:10.1080/13506129.2020.1778461.

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

[79] Durie, B. G. and Salmon, S. E. (1975). **Cellular kinetics staging, and immunoglobulin synthesis in multiple myeloma**. *Annual Review of Medicine*, 26:283–288. doi:10.1146/annurev.me.26.020175.001435.

[80] 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.

[81] Durie, B. G. M., Hoering, A., Abidi, M. H., Rajkumar, S. V., Epstein, J., Kahanic, S. P., Thakuri, M., Reu, F., Reynolds, C. M., Sexton, R., Orlowski, R. Z., Barlogie, B., and Dispenzieri, A. (2017). **Bortezomib with lenalidomide and dexamethasone versus lenalidomide and dexamethasone alone in patients with newly diagnosed myeloma without intent for immediate autologous stem-cell transplant (SWOG S0777): a randomised, open-label, phase 3 trial**. *Lancet (London, England)*, 389:519–527. doi:10.1016/S0140-6736(16)31594-X.

[82] Durinck, S. and Huber, W. (2017). **biomaRt**. Bioconductor. doi:10.18129/B9.BIOC.BIOMART.

[83] Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., de Moor, B., Brazma, A., and Huber, W. (2005). **BioMart and Bioconductor A powerful link between biological databases and microarray data analysis**. *Bioinformatics*, 21:3439–3440. doi:10.1093/bioinformatics/bti525.

[84] Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt**. *Nature Protocols*, 4:1184–1191. doi:10.1038/nprot.2009.97.

[85] Duston, M. A., Skinner, M., Meenan, R. F., and Cohen, A. S. (1989). **Sensitivity, specificity, and predictive value of abdominal fat aspiration for the diagnosis of amyloidosis**. *Arthritis and rheumatism*, 32:82–85. doi:10.1002/anr.1780320114.

[86] Duston, M. A., Skinner, M., Shirahama, T., and Cohen, A. S. (1987). **Diagnosis of amyloidosis by abdominal fat aspiration. Analysis of four years' experience**. *The American Journal of Medicine*, 82:412–414. doi:10.1016/0002-9343(87)90439-6.

[87] Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005). **The Sequence Ontology A tool for the unification of genome annotations**. *Genome biology*, 6:R44. doi:10.1186/gb-2005-6-5-r44.

[88] Einsele, H., Engelhardt, M., Tapprich, C., Müller, J., Liebisch, P., Langer, C., Kropff, M., Mügge, L. O., Jung, W., Wolf, H.-H., Metzner, B., Hart, C., Gramatzki, M., Hertenstein, B., Pfreundschuh, M., Rösler, W., Fischer, T., Maschmeyer, G., Kanz, L., Hess, G., Jäger, E., Bentz, M., Dürk, H. A., Salwender, H., Hebart, H., Straka, C., and Knop, S. (2017). **Phase II study of bortezomib, cyclophosphamide and dexamethasone as induction therapy in multiple myeloma: DSMM XI trial**. *British journal of haematology*, 179:586–597. doi:10.1111/bjh.14920.

[89] Eleutherakis-Papaiakovou, V., Bamias, A., Gika, D., Simeonidis, A., Pouli, A., Anagnostopoulos, A., Michali, E., Economopoulos, T., Zervas, K., and Dimopoulos, M. A. (2007). **Renal failure in multiple myeloma Incidence, correlations, and prognostic significance**. *Leukemia & lymphoma*, 48:337–341. doi:10.1080/10428190601126602.

[90] Emde, M. (2020). **Assessment of Pathogenesis and Prognosis of Plasma Cell Dyscrasias – Basic Research and Clinical Application**. PhD thesis, Inaugural Dissertation, Ruprecht-Karls-Universität, Heidelberg. Unpublished.

[91] Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). **MultiQC Summarize analysis results for multiple tools and samples in a single report**. *Bioinformatics*, 32:3047–3048. doi:10.1093/bioinformatics/btw354.

[92] Ewing, B. and Green, P. (1998). **Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities**. *Genome research*, 8:186–194. doi:10.1101/gr.8.3.186.

[93] Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). **Base-calling of automated sequencer traces using phred. I. Accuracy assessment**. *Genome research*, 8:175–185. doi:10.1101/gr.8.3.175.

[94] Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio,

P. (2018). **The Reactome Pathway Knowledgebase**. *Nucleic acids research*, 46:D649–D655. doi:10.1093/nar/gkx1132.

[95] Facon, T., Kumar, S., Plesner, T., Orlowski, R. Z., Moreau, P., Bahlis, N., Basu, S., Nahi, H., Hulin, C., Quach, H., Goldschmidt, H., O'Dwyer, M., Perrot, A., Venner, C. P., Weisel, K., Mace, J. R., Raje, N., Attal, M., Tiab, M., Macro, M., Frenzel, L., Leleu, X., Ahmadi, T., Chiu, C., Wang, J., van Rampelbergh, R., Uhlar, C. M., Kobos, R., Qi, M., and Usmani, S. Z. (2019). **Daratumumab plus Lenalidomide and Dexamethasone for Untreated Myeloma**. *The New England journal of medicine*, 380:2104–2115. doi:10.1056/NEJMoa1817249.

[96] Feher, J. (2017). **7.4 - Tubular Reabsorption and Secretion**. In: *Quantitative Human Physiology (Second Edition)*, editor Feher, J., pages 719–729, Boston. Academic Press. doi:10.1016/B978-0-12-800883-6.00072-0. URL: `http://www.sciencedirect.com/science/article/pii/B9780128008836000720`.

[97] Fisher, R. A., Bennett, J. H., and Yates, F., editors (1990). **Statistical methods, experimental design, and scientific inference A re-issue of Statistical methods for research workers, The design of experiments, and Statistical methods and scientific inference**. Oxford Univ. Press. URL: `http://www.loc.gov/catdir/enhancements/fy0602/90006726-d.html`.

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

[99] 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.

[100] Forbes, S. A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J. W., Futreal, P. A., and Stratton, M. R. (2008). **The Catalogue of Somatic Mutations in Cancer (COSMIC)**. *Current protocols in human genetics*, Chapter 10:Unit 10.11. doi:10.1002/0471142905.hg1011s57.

[101] Garcia, Y., Collins, A. B., and Stone, J. R. (2018). **Abdominal fat pad excisional biopsy for the diagnosis and typing of systemic amyloidosis**. *Human pathology*, 72:71–79. doi:10.1016/j.humpath.2017.11.001.

[102] GATK Dev Team (09.01.2019). **Introduction to the GATK Best Practices**. URL: `https://software.broadinstitute.org/gatk/best-practices/`. [Last visited: 12.02.2019].

[103] 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.

[104] Gentleman, R., Carey, V., Huber, W., and Hahne, F. (2016). **genefilter Genefilter: methods for filtering genes from high-throughput experiments**, version 1.60.0.

[105] 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.

[106] Gertz, M. A. (2013). **Immunoglobulin light chain amyloidosis 2013 update on diagnosis, prognosis, and treatment**. *American journal of hematology*, 88:416–425. doi:10.1002/ajh.23400.

[107] Gertz, M. A. (2018). **Immunoglobulin light chain amyloidosis diagnosis and treatment algorithm 2018**. *Blood cancer journal*, 8:44. doi:10.1038/s41408-018-0080-9.

[108] Gertz, M. A., Dispenzieri, A., and Muchtar, E. (2017). **Importance of FISH genetics in light chain amyloidosis**. *Oncotarget*, 8:81735–81736. doi:10.18632/oncotarget.21052.

[109] Gordon, M. and Lumley, T. (2019). **forestplot: Advanced Forest Plot Using 'grid' Graphics**, version 1.9. URL: `https://CRAN.R-project.org/package=forestplot`.

[110] Gowda, S., Desai, P. B., Kulkarni, S. S., Hull, V. V., Math, A. A. K., and Vernekar, S. N. (2010). **Markers of renal function tests**. *North American journal of medical sciences*, 2:170–173.

[111] 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.

[112] Griffith, M., Miller, C. A., Griffith, O. L., Krysiak, K., Skidmore, Z. L., Ramu, A., Walker, J. R., Dang, H. X., Trani, L., Larson, D. E., Demeter, R. T., Wendl, M. C., McMichael, J. F., Austin, R. E., Magrini, V., McGrath, S. D., Ly, A., Kulkarni, S., Cordes, M. G., Fronick, C. C., Fulton, R. S., Maher, C. A., Ding, L., Klco, J. M., Mardis, E. R., Ley, T. J., and Wilson, R. K. (2015). **Optimizing cancer genome sequencing and analysis**. *Cell systems*, 1:210–223. doi:10.1016/j.cels.2015.08.015.

[113] Guido van Rossum (2018). **Python**. Python Software Foundation, version 2.7.15. URL: `www.python.org`. [Last visited: 19.02.2019].

[114] 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.

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

[116] Hanamura, I., Stewart, J. P., Huang, Y., Zhan, F., Santra, M., Sawyer, J. R., Hollmig, K., Zangarri, M., Pineda-Roman, M., van Rhee, F., Cavallo, F., Burington, B., Crowley, J., Tricot, G., Barlogie, B., and Shaughnessy, Jr, John D (2006). **Frequent gain of chromosome band 1q21 in plasma-cell dyscrasias detected by fluorescence in situ hybridization: incidence increases from MGUS to relapsed myeloma and is related to prognosis and disease progression following tandem stem-cell transplantation**. *Blood*, 108:1724–1732. doi:10.1182/blood-2006-03-009910.

[117] Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). **Biases in Illumina transcriptome sequencing caused by random hexamer priming**. *Nucleic acids research*, 38:e131. doi:10.1093/nar/gkq224.

[118] Harrell, JR, F. E. (2018). **rms Regression Modeling Strategies**, version 5.1-2. URL: `https://CRAN.R-project.org/package=rms`.

[119] Harrell, JR, F. E., Dupont, w. c. f. C., and others, m. (2018). **Hmisc Harrell Miscellaneous**, version 4.1-1. URL: `https://CRAN.R-project.org/package=Hmisc`.

[120] 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.

[121] Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G. (2014). **pamr: Pam: prediction analysis for microarrays**, version 1.55. URL: `http://CRAN.R-project.org/package=pamr`, R package version 1.55.

[122] 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.

[123] Hayman, S. R., Bailey, R. J., Jalal, S. M., Ahmann, G. J., Dispenzieri, A., Gertz, M. A., Greipp, P. R., Kyle, R. A., Lacy, M. Q., Rajkumar, S. V., Witzig, T. E., Lust, J. A., and Fonseca, R. (2001). **Translocations involving the immunoglobulin heavy-chain locus are possible early genetic events in patients with primary systemic amyloidosis**. *Blood*, 98:2266–2268. doi:10.1182/blood.V98.7.2266.

[124] Holmes, S. and Huber, W. (2019). **Modern Statistics for Modern Biology**. Cambridge University Press, Cambridge, United Kingdom, 1 edition. doi:10.1017/9781108551441.

[125] Hose, D. (2015). **Asymptomatic multiple myeloma - molecular background of progression, evolution, and prognosis**. PhD thesis, 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].

[126] 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.

[127] 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.

[128] 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.

[129] 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.

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

[131] Hosten, A. O. (1990). **Clinical Methods: The History, Physical, and Laboratory Examinations BUN and Creatinine**. In: *Clinical Methods: The History, Physical, and Laboratory Examinations // Clinical methods*, editors Walker, H. K., Hall, W. D., and Hurst, J. W., Boston. Butterworths.

[132] Hothorn, T. (2017). **maxstat Maximally Selected Rank Statistics**, version 0.7-25. URL: `https://CRAN.R-project.org/package=maxstat`. [Last visited: 24.10.2019].

[133] 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.].

[134] Hrncic, R., Wall, J., Wolfenbarger, D. A., Murphy, C. L., Schell, M., Weiss, D. T., and Solomon, A. (2000). **Antibody-Mediated Resolution of Light Chain-Associated Amyloid Deposits**. *The American journal of pathology*, 157:1239–1246. doi:10.1016/S0002-9440(10)64639-1.

[135] Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). **Bioinformatics enrichment tools Paths toward the comprehensive functional analysis of large gene lists**. *Nucleic acids research*, 37:1–13. doi:10.1093/nar/gkn923.

[136] Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oles, A. K., Pages, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., and Morgan, M. (2015). **Orchestrating high-throughput genomic analysis with Bioconductor**. *Nature methods*, 12:115–121. doi:10.1038/nmeth.3252.

[137] 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.

[138] Illumina (2011). **Transitioning from Microarrays to mRNA-Seq**. *illumina-marketing*.

[139] Illumina (2015). **Data Sheet: Nextera Rapid Capture Exomes: A rapid workflow and comprehensive exome content, with unparalleled flexibility**. URL: www.illumina.com. [Last visited: 01.06.2016].

[140] Illumina (2015). **RNA-Seq Technology | Comparison vs. Microarrays**. URL: http://www.illumina.com/technology/next-generation-sequencing/mrna-seq.html. [Last visited: 23.06.2015].

[141] Illumina (2017). **An introduction to Next-Generation Sequencing Technology**. URL: https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf. [Last visited: 13.02.20].

[142] 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.

[143] Johnson, W. E., Li, C., and Rabinovic, A. (2007). **Adjusting batch effects in microarray expression data using empirical Bayes methods**. *Biostatistics (Oxford, England)*, 8:118–127. doi:10.1093/biostatistics/kxj037.

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

[145] Kadalayil, L., Rafiq, S., Rose-Zerilli, M. J. J., Pengelly, R. J., Parker, H., Oscier, D., Strefford, J. C., Tapper, W. J., Gibson, J., Ennis, S., and Collins, A. (2015). **Exome sequence read depth methods for identifying copy number changes**. *Briefings in bioinformatics*, 16:380–392. doi:10.1093/bib/bbu027.

[146] Kanehisa, M. (2019). **Toward understanding the origin and evolution of cellular organisms**. *Protein science : a publication of the Protein Society*. doi:10.1002/pro.3715.

[147] Kanehisa, M. and Goto, S. (2000). **KEGG Kyoto encyclopedia of genes and genomes**. *Nucleic acids research*, 28:27–30. doi:10.1093/nar/28.1.27.

[148] Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). **New approach for understanding genome variations in KEGG**. *Nucleic acids research*, 47:D590–D595. doi:10.1093/nar/gky962.

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

[150] Kaufman, G. P., Schrier, S. L., Lafayette, R. A., Arai, S., Witteles, R. M., and Liedtke, M. (2017). **Daratumumab yields rapid and deep hematologic responses in patients with heavily pretreated AL amyloidosis**. *Blood*, 130:900–902. doi:10.1182/blood-2017-01-763599.

[151] Keeling, J. and Herrera, G. A. (2005). **Matrix metalloproteinases and mesangial remodeling in light chain-related glomerular damage**. *Kidney international*, 68:1590–1603. doi:10.1111/j.1523-1755.2005.00571.x.

[152] Keller, I., Bensasson, D., and Nichols, R. A. (2007). **Transition-transversion bias is not universal A counter example from grasshopper pseudogenes**. *PLoS genetics*, 3:e22. doi:10.1371/journal.pgen.0030022.

[153] Ketchen Jr., D. J. and Shook, C. L. (1996). **THE APPLICATION OF CLUSTER ANALYSIS IN STRATEGIC MANAGEMENT RESEARCH: AN ANALYSIS AND CRITIQUE**. *Strategic Management Journal*, 17:441–458. doi:10.1002/(SICI)1097-0266(199606)17:6%3C441::AID-SMJ819%3E3.0.CO;2-G.

[154] Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., and Saunders, C. T. (2018). **Strelka2 Fast and accurate calling of germline and somatic variants**. *Nature Methods*, 15:591–594. doi:10.1038/s41592-018-0051-x.

[155] Kimmich, C. R., Terzer, T., Benner, A., Dittrich, T., Veelken, K., Carpinteiro, A., Hansen, T., Goldschmidt, H., Seckinger, A., Hose, D., Jauch, A., Wörner, S., Beimler, J., Müller-Tidow, C., Hegenbart, U., and Schönland, S. O. (2020). **Daratumumab for systemic AL amyloidosis: prognostic factors and adverse outcome with nephrotic-range albuminuria**. *Blood*, 135:1517–1530. doi:10.1182/blood.2019003633.

[156] 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.

[157] Knaus, B. J. and Grünwald, N. J. (2017). **vcfr A package to manipulate and visualize variant call format data in R**. *Molecular ecology resources*, 17:44–53. doi:10.1111/1755-0998.12549.

[158] Koboldt, D. C., Larson, D. E., and Wilson, R. K. (2013). **Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection**. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, 44:15.4.1–15.4.17. doi:10.1002/0471250953.bi1504s44.

[159] Krijthe, J. H. (2015). **Rtsne T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation**, version 0.15. URL: `https://github.com/jkrijthe/Rtsne`. [Last visited: 2019.10.17].

[160] Kryukov, F., Kryukova, E., Brozova, L., Kufova, Z., Filipova, J., Growkova, K., Sevcikova, T., Jarkovsky, J., and Hajek, R. (2016). **Does AL amyloidosis have a unique genomic profile? Gene expression profiling meta-analysis and literature overview**. *Gene*, 591:490–498. doi:10.1016/j.gene.2016.06.017.

[161] 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.

[162] Kuiper, R., van Duin, M., van Vliet, M. H., Broijl, A., van der Holt, B., el Jarari, L., van Beers, E. H., Mulligan, G., Avet-Loiseau, H., Gregory, W. M., Morgan, G., Goldschmidt, H., Lokhorst, H. M., and Sonneveld, P. (2015). **Prediction of high- and low-risk multiple myeloma**

based on gene expression and the International Staging System. *Blood*, 126:1996–2004. doi:10.1182/blood-2015-05-644039.

[163] Kumar, S., Dispenzieri, A., Katzmann, J. A., Larson, D. R., Colby, C. L., Lacy, M. Q., Hayman, S. R., Buadi, F. K., Leung, N., Zeldenrust, S. R., Ramirez-Alvarado, M., Clark, R. J., Kyle, R. A., Rajkumar, S. V., and Gertz, M. A. (2010). **Serum immunoglobulin free light-chain measurement in primary amyloidosis: prognostic value and correlations with clinical features**. *Blood*, 116:5126–5129. doi:10.1182/blood-2010-06-290668.

[164] 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.

[165] Kumar, S., Kaufman, J. L., Gasparetto, C., Mikhael, J., Vij, R., Pegourie, B., Benboubker, L., Facon, T., Amiot, M., Moreau, P., Punnoose, E. A., Alzate, S., Dunbar, M., Xu, T., Agarwal, S. K., Enschede, S. H., Leverson, J. D., Ross, J. A., Maciag, P. C., Verdugo, M., and Touzeau, C. (2017). **Efficacy of venetoclax as targeted therapy for relapsed/refractory t(11;14) multiple myeloma**. *Blood*, 130:2401–2409. doi:10.1182/blood-2017-06-788786.

[166] Kupperman, E., Lee, E. C., Cao, Y., Bannerman, B., Fitzgerald, M., Berger, A., Yu, J., Yang, Y., Hales, P., Bruzzese, F., Liu, J., Blank, J., Garcia, K., Tsu, C., Dick, L., Fleming, P., Yu, L., Manfredi, M., Rolfe, M., and Bolen, J. (2010). **Evaluation of the proteasome inhibitor MLN9708 in preclinical models of human cancer**. *Cancer research*, 70:1970–1980. doi:10.1158/0008-5472.CAN-09-2766.

[167] Kyle, R. A., Larson, D. R., Therneau, T. M., Dispenzieri, A., Kumar, S., Cerhan, J. R., and Rajkumar, S. V. (2018). **Long-Term Follow-up of Monoclonal Gammopathy of Undetermined Significance**. *The New England journal of medicine*, 378:241–249. doi:10.1056/NEJMoa1709974.

[168] Kyle, R. A., Linos, A., Beard, C. M., Linke, R. P., Gertz, M. A., O'Fallon, W. M., and Kurland, L. T. (1992). **Incidence and natural history of primary systemic amyloidosis in Olmsted County, Minnesota, 1950 through 1989 [see comments]**. *Blood*, 79:1817–1822. doi:10.1182/blood.V79.7.1817.bloodjournal7971817.

[169] Kyle, R. A. and Rajkumar, S. V. (2006). **Monoclonal gammopathy of undetermined significance**. *British journal of haematology*, 134:573–589. doi:10.1111/j.1365-2141.2006.06235.x.

[170] 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.

[171] 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.

[172] 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.

[173] Lahuerta, J.-J., Paiva, B., Vidriales, M.-B., Cordón, L., Cedena, M.-T., Puig, N., Martinez-Lopez, J., Rosiñol, L., Gutierrez, N. C., Martín-Ramos, M.-L., Oriol, A., Teruel, A.-I., Echeveste, M.-A., de Paz, R., de Arriba, F., Hernandez, M. T., Palomera, L., Martinez, R., Martin, A., Alegre, A., de La Rubia, J., Orfao, A., Mateos, M.-V., Blade, J., and San-Miguel, J. F. (2017). **Depth of Response in Multiple Myeloma: A Pooled Analysis of Three PETHEMA/GEM Clinical Trials**. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 35:2900–2910. doi:10.1200/JCO.2016.69.2517.

[174] 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.

[175] Larson, D. and Abbott, T. (2016). **bam-readcount**. The McDonnell Genome Institute, version 0.8.0. URL: `https://github.com/genome/bam-readcount`. [Last visited: 08.01.18].

[176] 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.

[177] Lefranc, M.-P. (2014). **Immunoglobulin and T Cell Receptor Genes IMGT(®) and the Birth and Rise of Immunoinformatics**. *Frontiers in immunology*, 5:22. doi:10.3389/fimmu.2014.00022.

[178] Li, H. (2012). **Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly**. *Bioinformatics*, 28:1838–1844. doi:10.1093/bioinformatics/bts280.

[179] Li, H. (2013). **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM**. *arXiv.org*, arXiv:1303.3997:1–3.

[180] Li, H. and Durbin, R. (2009). **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics*, 25:1754–1760. doi:10.1093/bioinformatics/btp324.

[181] 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.

[182] Li, J., Sarosi, I., Cattley, R. C., Pretorius, J., Asuncion, F., Grisanti, M., Morony, S., Adamu, S., Geng, Z., Qiu, W., Kostenuik, P., Lacey, D. L., Simonet, W. S., Bolon, B., Qian, X., Shalhoub, V., Ominsky, M. S., Zhu Ke, H., Li, X., and Richards, W. G. (2006). **Dkk1-mediated inhibition of Wnt signaling in bone results in osteopenia**. *Bone*, 39:754–766. doi:10.1016/j.bone.2006.03.017.

[183] Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). **The Molecular Signatures Database (MSigDB) hallmark gene set collection**. *Cell systems*, 1:417–425. doi:10.1016/j.cels.2015.12.004.

[184] Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics*, 27:1739–1740. doi:10.1093/bioinformatics/btr260.

[185] 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.

[186] 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.

[187] Lousada, I., Comenzo, R. L., Landau, H., Guthrie, S., and Merlini, G. (2015). **Light Chain Amyloidosis Patient Experience Survey from the Amyloidosis Research Consortium**. *Advances in therapy*, 32:920–928. doi:10.1007/s12325-015-0250-0.

[188] Lovén, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., Bradner, J. E., Lee, T. I., and Young, R. A. (2013). **Selective inhibition of tumor oncogenes by disruption of super-enhancers**. *Cell*, 153:320–334. doi:10.1016/j.cell.2013.03.036.

[189] Lucas, A. (2018). **amap Another Multidimensional Analysis Package**, version 0.8-16.

[190] Lykke-Andersen, S., Tomecki, R., Jensen, T. H., and Dziembowski, A. (2011). **The eukaryotic RNA exosome Same scaffold but variable catalytic subunits**. *RNA biology*, 8:61–66. doi:10.4161/rna.8.1.14237.

[191] 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.

[192] Manwani, R., Hegenbart, U., Mahmood, S., Sachchithanantham, S., Kyriakou, C., Yong, K., Popat, R., Rabin, N., Whelan, C., Dittrich, T., Kimmich, C., Hawkins, P., Schönland, S., and Wechalekar, A. (2018). **Deferred autologous stem cell transplantation in systemic AL amyloidosis**. *Blood cancer journal*, 8:101. doi:10.1038/s41408-018-0137-9.

[193] Mateos, M.-V., Cavo, M., Blade, J., Dimopoulos, M. A., Suzuki, K., Jakubowiak, A., Knop, S., Doyen, C., Lucio, P., Nagy, Z., Pour, L., Cook, M., Grosicki, S., Crepaldi, A., Liberati, A. M., Campbell, P., Shelekhova, T., Yoon, S.-S., Iosava, G., Fujisaki, T., Garg, M., Krevvata, M., Chen, Y., Wang, J., Kudva, A., Ukropec, J., Wroblewski, S., Qi, M., Kobos, R., and San-Miguel, J. (2020). **Overall survival with daratumumab, bortezomib, melphalan, and prednisone in newly diagnosed multiple myeloma (ALCYONE): a randomised, open-label, phase 3 trial**. *Lancet (London, England)*, 395:132–141. doi:10.1016/S0140-6736(19)32956-3.

[194] Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., and Koeffler, H. P. (2018). **Maftools Efficient and comprehensive analysis of somatic variants in cancer**. *Genome research*, 28:1747–1756. doi:10.1101/gr.239244.118.

[195] 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.

[196] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). **The Genome Analysis Toolkit A MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome research*, 20:1297–1303. doi:10.1101/gr.107524.110.

[197] McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., and Cunningham, F. (2016). **The Ensembl Variant Effect Predictor**. *Genome biology*, 17:122. doi:10.1186/s13059-016-0974-4.

[198] 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.

[199] Merlini, G. (2017). **AL amyloidosis From molecular mechanisms to targeted therapies**. *Hematology. American Society of Hematology. Education Program*, 2017:1–12. doi:10.1182/asheducation-2017.1.1.

[200] Merlini, G. and Stone, M. J. (2006). **Dangerous small B-cell clones**. *Blood*, 108:2520–2530. doi:10.1182/blood-2006-03-001164.

[201] Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). **GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers**. *Genome biology*, 12:R41. doi:10.1186/gb-2011-12-4-r41.

[202] Merz, M., Jauch, A., Hielscher, T., Bochtler, T., Schönland, S. O., Seckinger, A., Hose, D., Bertsch, U., Neben, K., Raab, M. S., Hillengass, J., Salwender, H., Blau, I. W., Lindemann, H.-W., Schmidt-Wolf, I. G. H., Scheid, C., Haenel, M., Weisel, K. C., and Goldschmidt, H. (2018). **Prognostic significance of cytogenetic heterogeneity in patients with newly diagnosed multiple myeloma**. *Blood advances*, 2:1–9. doi:10.1182/bloodadvances.2017013334.

[203] 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.

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

[205] Milani, P., Basset, M., Russo, F., Foli, A., Merlini, G., and Palladini, G. (2017). **Patients with light-chain amyloidosis and low free light-chain burden have distinct clinical features and outcome**. *Blood*, 130:625–631. doi:10.1182/blood-2017-02-767467.

[206] Milani, P., Murray, D. L., Barnidge, D. R., Kohlhagen, M. C., Mills, J. R., Merlini, G., Dasari, S., and Dispenzieri, A. (2017). **The utility of MASS-FIX to detect and monitor monoclonal proteins in the clinic**. *American journal of hematology*, 92:772–779. doi:10.1002/ajh.24772.

[207] Mills, J. R., Kohlhagen, M. C., Dasari, S., Vanderboom, P. M., Kyle, R. A., Katzmann, J. A., Willrich, M. A. V., Barnidge, D. R., Dispenzieri, A., and Murray, D. L. (2016). **Comprehensive Assessment of M-Proteins Using Nanobody Enrichment Coupled to MALDI-TOF Mass Spectrometry**. *Clinical chemistry*, 62:1334–1344. doi:10.1373/clinchem.2015.253740.

[208] 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.

[209] Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). **PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nature genetics*, 34:267–273. doi:10.1038/ng1180.

[210] Moreau, P., Cavallo, F., Leleu, X., Hulin, C., 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.

[211] Moreau, P., Masszi, T., Grzasko, N., Bahlis, N. J., Hansson, M., Pour, L., Sandhu, I., Ganly, P., Baker, B. W., Jackson, S. R., Stoppa, A.-M., Simpson, D. R., Gimsing, P., Palumbo, A., Garderet, L., Cavo, M., Kumar, S., Touzeau, C., Buadi, F. K., Laubach, J. P., Berg, D. T., Lin, J., Di Bacco, A., Hui, A.-M., van de Velde, H., and Richardson, P. G. (2016). **Oral Ixazomib, Lenalidomide, and Dexamethasone for Multiple Myeloma**. *The New England journal of medicine*, 374:1621–1634. doi:10.1056/NEJMoa1516282.

[212] 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.

[213] 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.

[214] 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.

[215] Muchtar, E., Dispenzieri, A., Kumar, S. K., Ketterling, R. P., Dingli, D., Lacy, M. Q., Buadi, F. K., Hayman, S. R., Kapoor, P., Leung, N., Chakraborty, R., Gonsalves, W., Warsame, R., Kourelis, T. V., Russell, S., Lust, J. A., Lin, Y., Go, R. S., Zeldenrust, S., Kyle, R. A., Rajkumar, S. V., and Gertz, M. A. (2017). **Interphase fluorescence in situ hybridization in untreated AL amyloidosis has an independent prognostic impact by abnormality type and treatment category**. *Leukemia*, 31:1562–1569. doi:10.1038/leu.2016.369.

[216] Muchtar, E., Therneau, T. M., Larson, D. R., Gertz, M. A., Lacy, M. Q., Buadi, F. K., Dingli, D., Hayman, S. R., Kapoor, P., Gonsalves, W., Kourelis, T. V., Warsame, R., Fonder, A., Hobbs, M., Hwa, Y. L., Leung, N., Russell, S., Lust, J. A., Lin, Y., Go, R. S., Zeldenrust, S., Kyle, R. A., Rajkumar, S. V., Kumar, S. K., and Dispenzieri, A. (2019). **Comparative analysis of staging systems in AL amyloidosis**. *Leukemia*, 33:811–814. doi:10.1038/s41375-018-0370-z.

[217] Müller, A. M. S., Geibel, A., Neumann, H. P. H., Kühnemund, A., Schmitt-Gräff, A., Böhm, J., and Engelhardt, M. (2006). **Primary (AL) amyloidosis in plasma cell disorders**. *The oncologist*, 11:824–830. doi:10.1634/theoncologist.11-7-824.

[218] 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].

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

[220] Murtagh, F. and Legendre, P. (2014). **Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?** *J. Classif.*, 31:274–295. doi:10.1007/s00357-014-9161-z.

[221] 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.

[222] 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**. *J Clin Oncol*, 31:4325–4332. doi:10.1200/JCO.2012.48.4923.

[223] 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.

[224] Neuwirth, E. (2014). **RColorBrewer Color Brewer Palettes**, version 1.1.-2. URL: `https://CRAN.R-project.org/package=RColorBrewer`.

[225] Oltersdorf, T., Elmore, S. W., Shoemaker, A. R., Armstrong, R. C., Augeri, D. J., Belli, B. A., Bruncko, M., Deckwerth, T. L., Dinges, J., Hajduk, P. J., Joseph, M. K., Kitada, S., Korsmeyer, S. J., Kunzer, A. R., Letai, A., Li, C., Mitten, M. J., Nettesheim, D. G., Ng, S., Nimmer, P. M., O'Connor, J. M., Oleksijew, A., Petros, A. M., Reed, J. C., Shen, W., Tahir, S. K., Thompson, C. B., Tomaselli, K. J., Wang, B., Wendt, M. D., Zhang, H., Fesik, S. W., and Rosenberg, S. H. (2005). **An inhibitor of Bcl-2 family proteins induces regression of solid tumours**. *Nature*, 435:677–681. doi:10.1038/nature03579.

[226] O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W. E., Wei, Z., Wang, K., and Lyon, G. J. (2013). **Low concordance of multiple variant-calling pipelines Practical implications for exome and genome sequencing**. *Genome medicine*, 5:28. doi:10.1186/gm432.

[227] Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., and Trajanoski, Z. (2014). **A survey of tools for variant analysis of next-generation genome sequencing data**. *Briefings in bioinformatics*, 15:256–278. doi:10.1093/bib/bbs086.

[228] Pagès, H., Carlson, M., Falcon, S., and Li, N. (2017). **AnnotationDbi Annotation Database Interface**, version 1.40.0.

[229] Paiva, B., Almeida, J., Pérez-Andrés, M., Mateo, G., López, A., Rasillo, A., Vén, López-Berges, M. F. S., and Orfao, A. (2010). **Utility of flow cytometry immunophenotyping in multiple myeloma and other clonal plasma cell-related disorders**. *Cytometry B Clin Cytom*, 78:239–252. doi:10.1002/cyto.b.20512.

[230] Paiva, B., Martinez-Lopez, J., Corchete, L. A., Sanchez-Vega, B., Rapado, I., Puig, N., Barrio, S., Sanchez, M.-L., Alignani, D., Lasa, M., García de Coca, A., Pardal, E., Oriol, A., Garcia, M.-E. G., Escalante, F., González-López, T. J., Palomera, L., Alonso, J., Prosper, F., Orfao, A., Vidriales, M.-B., Mateos, M.-V., Lahuerta, J.-J., Gutierrez, N. C., and San Miguel, J. F. (2016). **Phenotypic, transcriptomic, and genomic features of clonal plasma cells in light-chain amyloidosis**. *Blood*, 127:3035–3039. doi:10.1182/blood-2015-10-673095.

[231] Palladini, G., Campana, C., Klersy, C., Balduini, A., Vadacca, G., Perfetti, V., Perlini, S., Obici, L., Ascari, E., d'Eril, G. M., Moratti, R., and Merlini, G. (2003). **Serum N-terminal pro-brain natriuretic peptide is a sensitive marker of myocardial dysfunction in AL amyloidosis**. *Circulation*, 107:2440–2445. doi:10.1161/01.CIR.0000068314.02595.B2.

[232] Palladini, G., Dispenzieri, A., Gertz, M. A., Kumar, S., Wechalekar, A., Hawkins, P. N., Schönland, S., Hegenbart, U., Comenzo, R., Kastritis, E., Dimopoulos, M. A., Jaccard, A., Klersy, C., and Merlini, G. (2012). **New criteria for response to treatment in immunoglobulin light chain amyloidosis based on free light chain measurement and cardiac biomarkers Impact on survival outcomes**. *Journal of clinical oncology*, 30:4541–4549. doi:10.1200/JCO.2011.37.7614.

[233] Palladini, G. and Merlini, G. (2016). **What is new in diagnosis and management of light chain amyloidosis?** *Blood*, 128:159–168. doi:10.1182/blood-2016-01-629790.

[234] 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.

[235] Pan, B., Kusko, R., Xiao, W., Zheng, Y., Liu, Z., Xiao, C., Sakkiah, S., Guo, W., Gong, P., Zhang, C., Ge, W., Shi, L., Tong, W., and Hong, H. (2019). **Similarities and differences between variants called with human reference genome HG19 or HG38**. *BMC Bioinformatics*, 20:101. doi:10.1186/s12859-019-2620-0.

[236] 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.

[237] Premkumar, V. J., Lentzsch, S., Pan, S., Bhutani, D., Richter, J., Jagannath, S., Liedtke, M., Jaccard, A., Wechalekar, A. D., Comenzo, R., Sanchorawala, V., Royer, B., Rosenzweig, M., Valent, J., Schönland, S., Fonseca, R., Wong, S., and Kapoor, P. (2021). **Venetoclax induces deep hematologic remissions in t(11;14) relapsed/refractory AL amyloidosis**. *Blood cancer journal*, 11:10. doi:10.1038/s41408-020-00397-w.

[238] Qiu, Y. (2019). **showtext Using Fonts More Easily in R Graphs**, version 0.6. URL: `https://CRAN.R-project.org/package=showtext`.

[239] R Core Team (2018). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, version 3.4.4. URL: `https://www.r-project.org/`. [Last visited: 09.01.19].

[240] Radamaker, L., Lin, Y.-H., Annamalai, K., Huhn, S., Hegenbart, U., Schönland, S. O., Fritz, G., Schmidt, M., and Fändrich, M. (2019). **Cryo-EM structure of a light chain-derived amyloid fibril from a patient with systemic AL amyloidosis**. *Nature communications*, 10:1103. doi:10.1038/s41467-019-09032-0.

[241] 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.

[242] 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.

[243] Rajkumar, S. V. and Kumar, S. (2020). **Multiple myeloma current treatment algorithms**. *Blood cancer journal*, 10:94. doi:10.1038/s41408-020-00359-2.

[244] Rajkumar, S. V., Richardson, P. G., Hideshima, T., and Anderson, K. C. (2005). **Proteasome inhibition as a novel therapeutic target in human cancer**. *Journal of clinical oncology*, 23:630–639. doi:10.1200/JCO.2005.11.030.

[245] Ratermann, K., Steinbach, M., Caballero, K., Cowley, J., Nativi-Nicolau, J., and Kovacsovics, T. (2019). **Retrospective study of AL-amyloid patients with t(11;14) treated with daratumumab**. *Clinical Lymphoma Myeloma and Leukemia*, 19:e328. doi:10.1016/j.clml.2019.09.539.

[246] Rausch, T., Fritz, M. H.-Y., Korbel, J. O., and Benes, V. (2018). **Alfred Interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing**. *Bioinformatics*. doi:10.1093/bioinformatics/bty1007.

[247] Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). **g:Profiler-a web server for functional interpretation of gene lists (2016 update)**. *Nucleic acids research*, 44:W83–9. doi:10.1093/nar/gkw199.

[248] 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.

[249] Ritchie, M. E., Phipson, B., Di Wu, Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). **limma powers differential expression analyses for RNA-sequencing and microarray studies**. *Nucleic acids research*, 43:e47. doi:10.1093/nar/gkv007.

[250] Robert, P. and Escoufier, Y. (1976). **A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient**. *Applied Statistics*, 25:257. doi:10.2307/2347233.

[251] 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.

[252] 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.

[253] Ronan, T., Qi, Z., and Naegle, K. M. (2016). **Avoiding common pitfalls when clustering biological data**. *Science signaling*, 9:re6. doi:10.1126/scisignal.aad1932.

[254] Rossi, A., Voigtlaender, M., Janjetovic, S., Thiele, B., Alawi, M., März, M., Brandt, A., Hansen, T., Radloff, J., Schön, G., Hegenbart, U., Schönland, S., Langer, C., Bokemeyer, C., and Binder, M. (2017). **Mutational landscape reflects the biological continuum of plasma cell dyscrasias**. *Blood cancer journal*, 7:e537. doi:10.1038/bcj.2017.19.

[255] Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2010). **CORUM The comprehensive resource of mammalian protein complexes–2009**. *Nucleic acids research*, 38:D497–501. doi:10.1093/nar/gkp914.

[256] Salmon, S. E. (1971). **Immunoglobulin synthesis and tumour cell number and the natural history of multiple myeloma**. *British medical journal*, 2:321. doi:10.1136/bmj.2.5757.321.

[257] Salmon, S. E. and Smith, B. A. (1970). **Immunoglobulin synthesis and total body tumor cell number in IgG multiple myeloma**. *The Journal of clinical investigation*, 49:1114–1121. doi:10.1172/JCI106327.

[258] Santra, M., Zhan, F., Tian, E., Barlogie, B., and Shaughnessy, J. (2003). **A subset of multiple myeloma harboring the t(4;14)(p16;q32) translocation lacks FGFR3 expression but maintains an IGH/MMSET fusion transcript**. *Blood*, 101:2374–2376. doi:10.1182/blood-2002-09-2801.

[259] 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.

[260] Schumacher, M., Graf, E., and Gerds, T. (2003). **How to assess prognostic models for survival data: a case study in oncology**. *Methods of information in medicine*, 42:564–571.

[261] 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*, 1792:97–115. doi:10.1007/978-1-4939-7865-6_7.

[262] 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.

[263] 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.

[264] 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.

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

[266] 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.

[267] 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. doi:10.1182/blood-2018-99-113284, ASH Annual Meeting Abstract and Poster.

[268] Seshan, V. E. (2014). **clinfun: Clinical Trial Design and Data Analysis Functions.** R, version 1.0.6. URL: `https://cran.r-project.org/web/packages/clinfun/index.html`.

[269] Seshan, V. E. and Olshen, A. (2017). **DNAcopy DNA copy number data analysis**, version 1.52.0.

[270] Sharma, S., Jackson, P. G., and Makan, J. (2004). **Cardiac troponins**. *Journal of clinical pathology*, 57:1025–1026. doi:10.1136/jcp.2003.015420.

[271] Sharpley, F. A., Manwani, R., Mahmood, S., Sachchithanantham, S., Lachmann, H. J., Gillmore, J. D., Whelan, C. J., Fontana, M., Hawkins, P. N., and Wechalekar, A. D. (2019). **A novel mass spectrometry method to identify the serum monoclonal light chain component in systemic light chain amyloidosis**. *Blood cancer journal*, 9:16. doi:10.1038/s41408-019-0180-1.

[272] 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.

[273] Shendure, J. and Ji, H. (2008). **Next-generation DNA sequencing**. *Nature Biotechnology*, 26:1135–1145. doi:10.1038/nbt1486.

[274] Sherborne, A. L., Begum, D. B., Price, A., Johnson, D. C., Ellis, S., Smith, C., Mirabella, F., Menezes, K., Kimber, S., Jones, J. R., Pawlyn, C., Houlston, R. S., Russell, N. H., Jenner, M. W.,

Cook, G., Striha, A., Collett, C., Waterhouse, A., Gregory, W. M., Cairns, D. A., Drayson, M. T., Owen, R. G., Davies, F. E., Morgan, G. J., Jackson, G. H., and Kaiser, M. F. (2016). **Identifying Ultra-High Risk Myeloma By Integrated Molecular Genetic and Gene Expression Profiling**. *Blood*, 128:4407. doi:10.1182/blood.V128.22.4407.4407.

[275] 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.

[276] Shi, Y., Xu, X., Zhang, Q., Fu, G., Mo, Z., Wang, G. S., Kishi, S., and Yang, X.-L. (2014). **tRNA synthetase counteracts c-Myc to develop functional vasculature**. *eLife*, 3:e02349. doi:10.7554/eLife.02349.

[277] Smilde, A. K., Kiers, H. A. L., Bijlsma, S., Rubingh, C. M., and van Erk, M. J. (2009). **Matrix correlations for high-dimensional data: the modified RV-coefficient**. *Bioinformatics*, 25:401–405. doi:10.1093/bioinformatics/btn634.

[278] Solomon, A., Weiss, D. T., and Wall, J. S. (2003). **Therapeutic Potential of Chimeric Amyloid-reactive Monoclonal Antibody 11-1F4**. *Clinical Cancer Research*, 9:3831s–3838s.

[279] 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.

[280] 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.

[281] Sprynski, A. C., Hose, D., Kassambara, A., Vincent, L., Jourdan, M., Rossi, J. F., Goldschmidt, H., and Klein, B. (2010). **Insulin is a potent myeloma cell growth factor through insulin/IGF-1 hybrid receptor activation**. *Leukemia*, 24:1940–1950. doi:10.1038/leu.2010.192.

[282] Standal, T., Abildgaard, N., Fagerli, U.-M., Stordal, B., Hjertner, O., Borset, M., and Sundan, A. (2007). **HGF inhibits BMP-induced osteoblastogenesis Possible implications for the bone disease of multiple myeloma**. *Blood*, 109:3024–3030. doi:10.1182/blood-2006-07-034884.

[283] 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.

[284] Stewart, A. K. (2014). **Medicine. How thalidomide works against cancer**. *Science (New York, N.Y.)*, 343:256–257. doi:10.1126/science.1249543.

[285] Stewart, A. K., Rajkumar, S. V., Dimopoulos, M. A., Masszi, T., Špička, I., Oriol, A., Hájek, R., Rosiñol, L., Siegel, D. S., Mihaylov, G. G., Goranova-Marinova, V., Rajnics, P., Suvorov, A., Niesvizky, R., Jakubowiak, A. J., San-Miguel, J. F., Ludwig, H., Wang, M., Maisnar, V., Minarik, J., Bensinger, W. I., Mateos, M.-V., Ben-Yehuda, D., Kukreti, V., Zojwalla, N., Tonda, M. E., Yang, X., Xing, B., Moreau, P., and Palumbo, A. (2015). **Carfilzomib, lenalidomide, and dexamethasone for relapsed multiple myeloma**. *The New England journal of medicine*, 372:142–152. doi:10.1056/NEJMoa1411321.

[286] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). **Gene set enrichment analysis A knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America*, 102:15545–15550. doi:10.1073/pnas.0506580102.

[287] Swuec, P., Lavatelli, F., Tasaki, M., Paissoni, C., Rognoni, P., Maritan, M., Brambilla, F., Milani, P., Mauri, P., Camilloni, C., Palladini, G., Merlini, G., Ricagno, S., and Bolognesi, M. (2019). **Cryo-EM structure of cardiac amyloid fibrils from an immunoglobulin light chain AL amyloidosis patient**. *Nature communications*, 10:1269. doi:10.1038/s41467-019-09133-w.

[288] Tange, O. (2018). **GNU Parallel 2018**. Ole Tange and CERN European Organization for Nuclear Research, 1. edition. doi:10.5281/zenodo.1146014.

[289] The Bioconductor Dev Team (2014). **BSgenome.Hsapiens.NCBI.GRCh38: Full genome sequences for Homo sapiens (GRCh38)**, version 1.3.1000.

[290] The Gene Ontology Consortium (2019). **The Gene Ontology Resource 20 years and still GOing strong**. *Nucleic acids research*, 47:D330–D338. doi:10.1093/nar/gky1055.

[291] Therneau, T. (2015). **Survival: A Package for Survival Analysis in S**, version 2.43-2. URL: `http://CRAN.R-project.org/package=survival`.

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

[293] Therneau, T. M., Kyle, R. A., Melton, L. J., Larson, D. R., Benson, J. T., Colby, C. L., Dispenzieri, A., Kumar, S., Katzmann, J. A., Cerhan, J. R., and Rajkumar, S. V. (2012). **Incidence of monoclonal gammopathy of undetermined significance and estimation of duration before first clinical recognition**. *Mayo Clinic proceedings*, 87:1071–1079. doi:10.1016/j.mayocp.2012.06.014.

[294] Tian, E., Zhan, F., Walker, R., Rasmussen, E., Ma, Y., Barlogie, B., and Shaughnessy, J. D. (2003). **The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma**. *The New England journal of medicine*, 349:2483–2494. doi:10.1056/NEJMoa030847.

[295] Tovar, N., Rodríguez-Lobato, L. G., Cibeira, M. T., Magnano, L., Isola, I., Rosiñol, L., Bladé, J., and Fernández de Larrea, C. (2018). **Bone marrow plasma cell infiltration in light chain amyloidosis: impact on organ involvement and outcome**. *Amyloid : the international journal of experimental and clinical investigation : the official journal of the International Society of Amyloidosis*, 25:79–85. doi:10.1080/13506129.2018.1443439.

[296] 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.

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

[298] van der Maaten, L. (2014). **Accelerating t-SNE using Tree-Based Algorithms**. *Journal of Machine Learning Research*, 15:3221–3245.

[299] van der Maaten, L. (2019). **t-SNE FAQ**. URL: `https://lvdmaaten.github.io/tsne/`. [Last visited: 06.03.2019].

[300] van der Maaten, L. and Hinton, G. E. (2008). **Visualizing High-Dimensional Data Using t-SNE**. *Journal of Machine Learning Research*, 9:2579–2605.

[301] van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). **Ten years of next-generation sequencing technology**. *Trends in genetics : TIG*, 30:418–426. doi:10.1016/j.tig.2014.07.001.

[302] van Haaften, G., Dalgliesh, G. L., Davies, H., Chen, L., Bignell, G., Greenman, C., Edkins, S., Hardy, C., O'Meara, S., Teague, J., Butler, A., Hinton, J., Latimer, C., Andrews, J., Barthorpe, S., Beare, D., Buck, G., Campbell, P. J., Cole, J., Forbes, S., Jia, M., Jones, D., Kok, C. Y., Leroy, C., Lin, M.-L., McBride, D. J., Maddison, M., Maquire, S., McLay, K., Menzies, A., Mironenko, T., Mulderrig, L., Mudie, L., Pleasance, E., Shepherd, R., Smith, R., Stebbings, L., Stephens, P., Tang, G., Tarpey, P. S., Turner, R., Turrell, K., Varian, J., West, S., Widaa, S., Wray, P., Collins, V. P., Ichimura, K., Law, S., Wong, J., Yuen, S. T., Leung, S. Y., Tonon, G., DePinho, R. A., Tai, Y.-T., Anderson, K. C., Kahnoski, R. J., Massie, A., Khoo, S. K., Teh, B. T., Stratton, M. R., and Futreal, P. A. (2009). **Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer**. *Nature genetics*, 41:521–523. doi:10.1038/ng.349.

[303] van Rhee, F., Bolejack, V., Hollmig, K., Pineda-Roman, M., Anaissie, E., Epstein, J., Shaughnessy, J. D., Zangari, M., Tricot, G., Mohiuddin, A., Alsayed, Y., Woods, G., Crowley, J., and Barlogie, B. (2007). **High serum-free light chain levels and their rapid reduction in response to therapy define an aggressive multiple myeloma subtype with poor prognosis**. *Blood*, 110:827–832. doi:10.1182/blood-2007-01-067728.

[304] Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2015). **SIFT missense predictions for genomes**. *Nature Protocols*, 11:1 EP –. doi:10.1038/nprot.2015.123.

[305] Wakeley, J. (1996). **The excess of transitions among nucleotide substitutions New methods of estimating transition bias underscore its significance**. *Trends in ecology & evolution*, 11:158–162.

[306] 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.

[307] 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.

[308] Walker, B. A., Mavrommatis, K., Wardell, C. P., Ashby, T. C., Bauer, M., Davies, F. E., Rosenthal, A., Wang, H., Qu, P., Hoering, A., Samur, M., Towfic, F., Ortiz, M., Flynt, E., Yu, Z., Yang, Z., Rozelle, D., Obenauer, J., Trotter, M., Auclair, D., Keats, J., Bolli, N., Fulciniti, M., Szalat, R., Moreau, P., Durie, B., Stewart, A. K., Goldschmidt, H., Raab, M. S., Einsele, H., Sonneveld, P., San Miguel, J., Lonial, S., Jackson, G. H., Anderson, K. C., Avet-Loiseau, H., Munshi, N., Thakurta, A., and Morgan, G. J. (2018). **Identification of novel mutational drivers reveals oncogene dependencies in multiple myeloma**. *Blood*, 132:587–597. doi:10.1182/blood-2018-03-840132.

[309] Walker, B. A., Wardell, C. P., Johnson, D. C., Kaiser, M. F., Begum, D. B., Dahir, N. B., Ross, F. M., Davies, F. E., Gonzalez, D., and Morgan, G. J. (2013). **Characterization of IGH locus breakpoints in multiple myeloma indicates a subset of translocations appear to occur in pregerminal center B cells**. *Blood*, 121:3413–3419. doi:10.1182/blood-2012-12-471888.

[310] 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.

[311] Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. (2019). **gplots Various R Programming Tools for Plotting**, version 3.0.1.1.

[312] Warren, P. (2016). **panp Presence-Absence Calls from Negative Strand Matching Probesets**, version 1.48.0.

[313] 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.

[314] Wattenberg, M., Viégas, F., and Johnson, I. (2016). **How to Use t-SNE Effectively**. *Distill*. doi:10.23915/distill.00002.

[315] Wechalekar, A. D., Schonland, S. O., Kastritis, E., Gillmore, J. D., Dimopoulos, M. A., Lane, T., Foli, A., Foard, D., Milani, P., Rannigan, L., Hegenbart, U., Hawkins, P. N., Merlini, G., and Palladini, G. (2013). **A European collaborative study of treatment outcomes in 346 patients with cardiac stage III AL amyloidosis**. *Blood*, 121:3420–3427. doi:10.1182/blood-2012-12-473066.

[316] White, B. S., Lanc, I., O'Neal, J., Gupta, H., Fulton, R. S., Schmidt, H., Fronick, C., Belter, E. A., Fiala, M., King, J., Ahmann, G. J., DeRome, M., Mardis, E. R., Vij, R., DiPersio, J. F., Levy, J., Auclair, D., and Tomasson, M. H. (2018). **A multiple myeloma-specific capture sequencing platform discovers novel translocations and frequent, risk-associated point mutations in IGLL5**. *Blood cancer journal*, 8:35. doi:10.1038/s41408-018-0062-y.

[317] Wickham, H. (2009). **ggplot2 Elegant Graphics for Data Analysis**. Springer-Verlag New York. URL: `http://ggplot2.org`.

[318] Wickham, H. (2011). **The Split-Apply-Combine Strategy for Data Analysis**. *Journal of Statistical Software*, 40:1–29.

[319] Wickham, H. (2018). **stringr Simple, Consistent Wrappers for Common String Operations**, version 1.3.1. URL: `https://CRAN.R-project.org/package=stringr`.

[320] Wright, C. F., McRae, J. F., Clayton, S., Gallone, G., Aitken, S., FitzGerald, T. W., Jones, P., Prigmore, E., Rajan, D., Lord, J., Sifrim, A., Kelsell, R., Parker, M. J., Barrett, J. C., Hurles, M. E., FitzPatrick, D. R., Firth, H. V., and Study, o. b. o. t. D. (2018). **Making new genetic diagnoses with old data Iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders**. *Genetics in Medicine*, 20:1216–1223. doi:10.1038/gim.2017.246.

[321] Wu, J., Irizarry, R., and Gentry, J. M. (2016). **gcrma Background Adjustment Using Sequence Information**, version 2.50.0.

[322] 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.

[323] 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.

[324] Xu, C. (2018). **A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data**. *Computational and structural biotechnology journal*, 16:15–24. doi:10.1016/j.csbj.2018.01.003.

[325] Yates, B., Braschi, B., Gray, K. A., Seal, R. L., Tweedie, S., and Bruford, E. A. (2017). **Genenames.org The HGNC and VGNC resources in 2017**. *Nucleic acids research*, 45:D619–D625. doi:10.1093/nar/gkw1033.

[326] Zar, J. H. (1999). **Biostatistical analysis**. Prentice Hall, 4. edition, pp. 523.

[327] Zare, F., Dow, M., Monteleone, N., Hosny, A., and Nabavi, S. (2017). **An evaluation of copy number variation detection tools for cancer using whole exome sequencing data**. *BMC Bioinformatics*, 18:286. doi:10.1186/s12859-017-1705-x.

[328] 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.

[329] 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.

[330] Zhang, K. W., Stockerl-Goldstein, K. E., and Lenihan, D. J. (2019). **Emerging Therapeutics for the Treatment of Light Chain and Transthyretin Amyloidosis**. *JACC. Basic to translational science*, 4:438–448. doi:10.1016/j.jacbts.2019.02.002.

[331] 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.

[332] Zhang, Z. and Gerstein, M. (2003). **Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes**. *Nucleic acids research*, 31:5338–5348. doi:10.1093/nar/gkg745.

[333] Zhao, L., Liu, H., Yuan, X., Gao, K., and Duan, J. (2020). **Comparative study of whole exome sequencing-based copy number variation detection tools**. *BMC Bioinformatics*, 21:97. doi:10.1186/s12859-020-3421-1.

[334] Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). **Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives**. *BMC Bioinformatics*, 14 Suppl 11:S1. doi:10.1186/1471-2105-14-S11-S1.

[335] Zhou, P., Hoffman, J., Landau, H., Hassoun, H., Iyer, L., and Comenzo, R. L. (2012). **Clonal plasma cell pathophysiology and clinical features of disease are linked to clonal plasma cell expression of cyclin D1 in systemic light-chain amyloidosis**. *Clinical lymphoma, myeloma & leukemia*, 12:49–58. doi:10.1016/j.clml.2011.09.217.

[336] Zhou, Y., Zhang, Q., Stephens, O., Heuck, C. J., Tian, E., Sawyer, J. R., Cartron-Mizeracki, M.-A., Qu, P., Keller, J., Epstein, J., Barlogie, B., and Shaughnessy, J. D. (2012). **Prediction of**

cytogenetic abnormalities with gene expression profiles. *Blood*, 119:e148–50. doi:10.1182/blood-2011-10-388702.

[337] Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., Benner, C., and Chanda, S. K. (2019). **Metascape provides a biologist-oriented resource for the analysis of systems-level datasets**. *Nature communications*, 10:1523. doi:10.1038/s41467-019-09234-6.

8 Contributions and publications

This work was realized 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 B1; Bonn, Germany) and the 7th EU-framework program "OverMyR".
Contributions to this work were made by members of the following institutions:

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

2 Amyloidose-Zentrum (Speaker: Prof. Dr. med. Ute Hegenbarth), Medizinische Klinik V, Universitätsklinikum Heidelberg, Im Neuenheimer Feld 410, D-69120 Heidelberg, Germany

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 and Proteomics Core Facility, DKFZ (Head: Dr. Stephan Wolf), DKFZ Heidelberg, Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany.

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

Contributions are listed in the following table:

Subject	Item	Contribution [institution]
Pathogenetic analyses	Discussion	PD Dr. Dr. Dirk Hose [1] PD Dr. Anja Seckinger [1] Martina Emde [1]
	HDAL score	Martina Emde [1]
Statistical analyses	Discussion	PD Dr. Dr. Dirk Hose [1] PD Dr. Anja Seckinger [1] Martina Emde [1]
Clinical data	Scientific responsibility, data collection, analysis and interpretation	PD Dr. Dr. Dirk Hose [1] PD Dr. Anja Seckinger [1]
	Clinical data collection (light chain amyloidosis)	Prof. Dr. med. Ute Hegenbarth [2] Prof. Dr. med. Stefan Schönland [2]
	Data preprocessing and cleaning	Martina Emde [1]
Laboratory performance	Scientific, administrative and organizational responsibility	PD Dr. Dr. Dirk Hose [1] PD Dr. Anja Seckinger [1]
	Performing plasma cell purification	Maria Dörner and Birgit Schneiders *et al.*, Multiple Myeloma Research Laboratory (Head: PD Dr. Dr. Dirk Hose) [1]
	Performing interphase fluorescence *in situ* hybridisation	Technicians, Molekular-zytogenetisches Labor (Head: Prof. Dr. sc. hum. Anna Jauch) [3]
	Performing microarray gene expression profiling	Véronique Pantesco, Institute for Research in Biotherapy (Head: Prof. Dr. Bernard Klein) [4]
	Performing whole exome sequencing sequencing	Tomi Bähr-Ivacevic, Multiple Myeloma Research Laboratory (Head: PD Dr. Dr. Dirk Hose), [1] Genomics and Proteomics Core Facility (Head: Dr. Stephan Wolf) [5]
	Performing RNA sequencing	Tomi Bähr-Ivacevic, Multiple Myeloma Research Laboratory (Head: PD Dr. Dr. Dirk Hose), [1] Genomics Core Facility (Head: Dr. Vladimir Benes) [6]
Sampling	Performing bone marrow aspiration	Responsible physicians

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

[1] Hose, D., ***Beck, S.***, Salwender, H., Emde, M., Bertsch, U., Kunz, C., Scheid, C., Hänel, M., Weisel, K., Hielscher, T., Raab, M., Goldschmidt, H., Jauch, A., Moreaux, J., 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.

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

[1] ***Beck, S.***, Emde, M., Moreaux, J., Seckinger, A., and Hose, D. (2019). **Prediction of Malingant 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.

Further own publications:

[1] Bolomsky, A., Hose, D., Schreder, M., Seckinger, A., ***Lipp, S.***, Klein, B., Heintel, D., Ludwig, H., and Zojer, N. (2015). **Insulin like growth factor binding protein 7 (IGFBP7) expression is linked to poor prognosis but may protect from bone disease in multiple myeloma**. *Journal of hematology & oncology*, 8:10. doi:10.1186/s13045-014-0105-1.

[2] Hájek, R., Sandecka, V., Špička, I., Raab, M., Goldschmidt, H., ***Beck, S.***, Minařík, J., Pavlíček, P., Radocha, J., Heindorfer, A., Jelínek, T., Stejskal, L., Brožová, L., Ševčíková, S., Straub, J., Pika, T., Pour, L., Maisnar, V., Seckinger, A., and Hose, D. (2020). **Identification of patients with smouldering multiple myeloma at ultra-high risk of progression using serum parameters: the Czech Myeloma Group model**. *British journal of haematology*, 190:189–197. doi:10.1111/bjh.16572.

[3] Meier, J. K., Schnetz, M., ***Beck, S.***, Schmid, T., Dominguez, M., Kalinovic, S., Daiber, A., Brüne, B., and Jung, M. (2021). **Iron-Bound Lipocalin-2 Protects Renal Cell Carcinoma from Ferroptosis**. *Metabolites*, 11:329. doi:10.3390/metabo11050329.

[4] 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.

[5] 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.

[6] 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.

Further published abstracts:

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

[2] 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.

[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.

Appendix

A Supplementary Tables

Table A.1: List of all used tools and programming languages with version and reference resource.

Tool	Version	Resource
GNU parallel	3.36.5	Tange [288]
Python	2.7.15	Guido van Rossum [113]
R	3.4.4	R Core Team [239]
bioconductor	3.6	Gentleman *et al.* [105]
Fastqc	0.11.4	Andrews [9]
fastp	0.19.4	Chen *et al.* [47]
bwa	0.7.12	Li [179], Li and Durbin [180]
samtools	1.5	Li *et al.* [181]
Alfred	0.1.13	Rausch *et al.* [246]
Picard	2.10.10	Broad Institute [38]
GATK	3.8	McKenna *et al.* [196]
VarScan2	2.4.3	Koboldt *et al.* [158]
Seurat	2.5	Christoforides *et al.* [56]
Strelka	2.9.6	Kim *et al.* [154]
Manta	1.4.0	Chen *et al.* [48]
bam-readcount	0.8.0	Larson and Abbott [175]
Ensembl vep	0.19.4	McLaren *et al.* [197]
GISTIC 2.0	2.0.23	Mermel *et al.* [201]
HTSeq	0.11.2	Anders *et al.* [7]
STAR	2.4	Dobin *et al.* [76]
MultiQC	1.6	Ewels *et al.* [91]
metascape	3.5 (data base 2019-08-14)	Zhou *et al.* [337]

Table A.2: List of all used databases for functional enrichment analysis with metascape.

Database	URL	**Resource**
GO	`http://geneontology.org`	[11, 290]
MSigDB	`http://www.broadinstitute.org/gsea/msigdb`	[183, 184, 209]
Canonical Pathways	`http://www.broadinstitute.org/gsea/msigdb`	[286]
Hallmark	`http://www.broadinstitute.org/gsea/msigdb`	[183]
KEGG	`http://www.genome.jp/kegg`	[146–148]
Reactome	`http://www.reactome.org`	[94]
CORUM	`http://mips.helmholtz-muenchen.de/corum`	[255]
UniProt	`http://www.uniprot.org`	[247]
Protein Atlas	`http://www.proteinatlas.org`	[135]

Table A.3: List of all used R packages with version and reference resource.

Package	**Version**	**Citation**
stats	3.4.4	R Core Team [239]
rms	5.1-2	Harrell [118]
survival	2.43-3	Therneau [291], Therneau and Grambsch [292]
Hmisc	4.1-1	Harrell *et al.* [119]
maxstat	0.7-25	Hothorn [132]
clinfun	1.0.15	Seshan [268]
pec	2018.07.26	Mogensen *et al.* [208]
amap	1.58.0	Lucas [189]
made4	1.52.0	Culhane *et al.* [65]
Rtsne	0.15	Krijthe [159]
edgeR	3.20.9	Chen *et al.* [49], McCarthy *et al.* [195], Robinson *et al.* [251]
limma	3.34.9	Ritchie *et al.* [249]
affy	1.58.0	Gautier *et al.* [103]
gcrma	2.50.0	Wu *et al.* [321]
panp	1.48.0	Warren [312], Warren *et al.* [313]
DNAcopy	1.52.0	Seshan and Olshen [269]
vcfR	1.8.0	Knaus and Grünwald [157]
IRanges	2.12.0	Lawrence *et al.* [176]
plyr	1.8.4	Wickham [318]
stringr	1.3.1	Wickham [319]
AnnotationDbi	1.40.0	Pagès *et al.* [228]
Biobase	2.38.0	Huber *et al.* [136]
BiocGenerics	0.24.0	Huber *et al.* [136]
hgu133plus2.db	3.2.3	Carlson [41]
org.Hs.eg.db	3.5.0	Carlson [42]
BiomaRt	2.34.2	Durinck *et al.* [83, 84]
BSgenome.Hsapiens.NCBI.GRCh38	1.3.1000	The Bioconductor Dev Team [289]
genefilter	1.60.0	Gentleman *et al.* [104]
ggplot2	3.1.0	Wickham [317]
VennDiagram	1.6.20	Chen [46]
forestplot	1.9	Gordon and Lumley [109]
gplots	3.0.1.1	Warnes *et al.* [311]
gridExtra	2.3	Auguie [13]
maftools	1.4.28	Mayakonda *et al.* [194]
RColorBrewer	1.1-2	Neuwirth [224]
showtext	0.6	Qiu [238]
xtable	1.8-3	Dahl *et al.* [66]

Table A.4: Gene expression based proliferation index (GPI) distribution: Number of patients (*n*) and percentages (%) grouped by GPI low risk, medium risk and high risk. Separation by subentities ALMG and ALMM and AL disease specific variables and disease entities AL, MGUS, AMM, and MM. See figure 3.15 **a** and **c** for graphical presentation. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

			low risk		**medium risk**		**high risk**	
Variable	**Level**	**Entity**	***n***	**%**	***n***	**%**	***n***	**%**
		AL	66	33.67	119	60.71	11	5.61
		MGUS	37	57.81	26	40.62	1	1.56
		AMM	117	43.17	153	56.46	1	0.37
		MM	151	19.74	490	64.05	124	16.21
Heart involvement	no	ALMG	8	33.33	16	66.67	0	0
		ALMM	5	21.74	16	69.57	2	8.7
		AL	13	27.66	32	68.09	2	4.26
	yes	ALMG	20	34.48	37	63.79	1	1.72
		ALMM	32	36.36	48	54.55	8	9.09
		AL	52	35.62	85	58.22	9	6.16
NT-ProBNP [ng/L]	<1800	ALMG	13	34.21	25	65.79	0	0
		ALMM	10	28.57	23	65.71	2	5.71
		AL	23	31.51	48	65.75	2	2.74
	≥1800	ALMG	14	32.56	28	65.12	1	2.33
		ALMM	27	36.49	40	54.05	7	9.46
		AL	41	35.04	68	58.12	8	6.84
Difference FLC [mg/L]	<180	ALMG	14	28.57	34	69.39	1	2.04
		ALMM	12	46.15	13	50	1	3.85
		AL	26	34.67	47	62.67	2	2.67
	≥180	ALMG	12	40	18	60	0	0
		ALMM	25	30.12	50	60.24	8	9.64
		AL	37	32.74	68	60.18	8	7.08
AL type	Kappa	ALMG	9	52.94	8	47.06	0	0
		ALMM	10	37.04	16	59.26	1	3.7
		AL	19	43.18	24	54.55	1	2.27
	Lambda	ALMG	19	29.23	45	69.23	1	1.54
		ALMM	28	32.18	50	57.47	9	10.34
		AL	47	30.92	95	62.5	10	6.58
Creatinine [mg/dL]	<2	ALMG	26	35.62	46	63.01	1	1.37
		ALMM	33	33.67	57	58.16	8	8.16
		AL	59	34.5	103	60.23	9	5.26
	≥2	ALMG	2	22.22	7	77.78	0	0
		ALMM	5	31.25	9	56.25	2	12.5
		AL	7	28	16	64	2	8

Table A.5: Translocation/cyclinD (TC) classification distribution: Number of patients (*n*) and percentages (%) grouped by TC classification. Separation by AL disease specific variables and disease entities AL, MGUS, AMM, and MM and subentities ALMG and ALMM. See figure 3.17 **a** and **c** for graphical presentation. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

			11q13		6p21		D1		D1+D2		D2		FGFR3		MAF		none	
Variable	**Level**	**Entity**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
		AL	98	50	8	4.08	16	8.16	9	4.59	41	20.92	4	2.04	20	10.2	0	0
		MGUS	6	9.38	1	1.56	11	17.19	2	3.12	15	23.44	4	6.25	18	28.12	7	10.94
		AMM	51	18.82	3	1.11	79	29.15	20	7.38	51	18.82	31	11.44	31	11.44	5	1.85
		MM	146	19.08	13	1.7	281	36.73	64	8.37	90	11.76	93	12.16	62	8.1	16	2.09
Heart involvement	no	ALMG	13	54.17	0	0	3	12.5	1	4.17	3	12.5	0	0	4	16.67	0	0
		ALMM	13	56.52	1	4.35	2	8.7	1	4.35	4	17.39	2	8.7	0	0	0	0
		AL	26	55.32	1	2.13	5	10.64	2	4.26	7	14.89	2	4.26	4	8.51	0	0
	yes	ALMG	30	51.72	1	1.72	4	6.9	1	1.72	12	20.69	1	1.72	9	15.52	0	0
		ALMM	41	46.59	5	5.68	7	7.95	6	6.82	21	23.86	1	1.14	7	7.95	0	0
		AL	71	48.63	6	4.11	11	7.53	7	4.79	33	22.6	2	1.37	16	10.96	0	0
NT-ProBNP [ng/L]	<1800	ALMG	20	52.63	1	2.63	4	10.53	1	2.63	5	13.16	0	0	7	18.42	0	0
		ALMM	18	51.43	1	2.86	1	2.86	3	8.57	9	25.71	1	2.86	2	5.71	0	0
		AL	38	52.05	2	2.74	5	6.85	4	5.48	14	19.18	1	1.37	9	12.33	0	0
	≥1800	ALMG	23	53.49	0	0	3	6.98	1	2.33	9	20.93	1	2.33	6	13.95	0	0
		ALMM	36	48.65	5	6.76	7	9.46	4	5.41	16	21.62	1	1.35	5	6.76	0	0
		AL	59	50.43	5	4.27	10	8.55	5	4.27	25	21.37	2	1.71	11	9.4	0	0
Difference FLC [mg/L]	<180	ALMG	26	53.06	0	0	3	6.12	1	2.04	10	20.41	0	0	9	18.37	0	0
		ALMM	15	57.69	3	11.54	2	7.69	1	3.85	0	0	1	3.85	4	15.38	0	0
		AL	41	54.67	3	4	5	6.67	2	2.67	10	13.33	1	1.33	13	17.33	0	0
	≥180	ALMG	17	56.67	1	3.33	4	13.33	0	0	3	10	1	3.33	4	13.33	0	0
		ALMM	39	46.99	3	3.61	6	7.23	6	7.23	25	30.12	1	1.2	3	3.61	0	0
		AL	56	49.56	4	3.54	10	8.85	6	5.31	28	24.78	2	1.77	7	6.19	0	0
AL type	Kappa	ALMG	11	64.71	0	0	4	23.53	0	0	0	0	0	0	2	11.76	0	0
		ALMM	16	59.26	2	7.41	4	14.81	0	0	4	14.81	1	3.7	0	0	0	0
		AL	27	61.36	2	4.55	8	18.18	0	0	4	9.09	1	2.27	2	4.55	0	0
	Lambda	ALMG	32	49.23	1	1.54	3	4.62	2	3.08	15	23.08	1	1.54	11	16.92	0	0
		ALMM	39	44.83	5	5.75	5	5.75	7	8.05	22	25.29	2	2.3	7	8.05	0	0
		AL	71	46.71	6	3.95	8	5.26	9	5.92	37	24.34	3	1.97	18	11.84	0	0
Creatinine [mg/dL]	<2	ALMG	38	52.05	1	1.37	7	9.59	2	2.74	13	17.81	1	1.37	11	15.07	0	0
		ALMM	48	48.98	7	7.14	7	7.14	5	5.1	21	21.43	3	3.06	7	7.14	0	0
		AL	86	50.29	8	4.68	14	8.19	7	4.09	34	19.88	4	2.34	18	10.53	0	0
	≥2	ALMG	5	55.56	0	0	0	0	0	0	2	22.22	0	0	2	22.22	0	0
		ALMM	7	43.75	0	0	2	12.5	2	12.5	5	31.25	0	0	0	0	0	0
		AL	12	48	0	0	2	8	2	8	7	28	0	0	2	8	0	0

Table A.6: Molecular classification (MC) distribution: Number of patients (*n*) and percentages (%) grouped by molecular classification. Separation by AL disease specific variables and disease entities AL, MGUS, AMM, and MM and subentities ALMG and ALMM. See figure 3.18 **a** and **c** for graphical presentation. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

Variable	Level	Entity	CD1		CD2		HY		LB		MF		MS		PR	
			n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
		AL	12	6.12	92	46.94	5	2.55	72	36.73	5	2.55	5	2.55	5	2.55
		MGUS	0	0	19	29.69	9	14.06	22	34.38	10	15.62	4	6.25	0	0
		AMM	4	1.48	72	26.57	68	25.09	75	27.68	20	7.38	32	11.81	0	0
		MM	34	4.44	151	19.74	207	27.06	129	16.86	30	3.92	98	12.81	116	15.16
Heart involvement	no	ALMG	3	12.5	11	45.83	2	8.33	7	29.17	1	4.17	0	0	0	0
		ALMM	0	0	11	47.83	0	0	9	39.13	0	0	2	8.7	1	4.35
		AL	3	6.38	22	46.81	2	4.26	16	34.04	1	2.13	2	4.26	1	2.13
	yes	ALMG	2	3.45	33	56.9	0	0	19	32.76	1	1.72	1	1.72	2	3.45
		ALMM	7	7.95	36	40.91	3	3.41	35	39.77	3	3.41	2	2.27	2	2.27
		AL	9	6.16	69	47.26	3	2.05	54	36.99	4	2.74	3	2.05	4	2.74
NT-ProBNP [ng/L]	<1800	ALMG	4	10.53	20	52.63	2	5.26	10	26.32	2	5.26	0	0	0	0
		ALMM	2	5.71	16	45.71	0	0	13	37.14	1	2.86	1	2.86	2	5.71
		AL	6	8.22	36	49.32	2	2.74	23	31.51	3	4.11	1	1.37	2	2.74
	≥1800	ALMG	1	2.33	24	55.81	0	0	15	34.88	0	0	1	2.33	2	4.65
		ALMM	5	6.76	31	41.89	3	4.05	30	40.54	2	2.7	2	2.7	1	1.35
		AL	6	5.13	55	47.01	3	2.56	45	38.46	2	1.71	3	2.56	3	2.56
Difference FLC [mg/L]	<180	ALMG	4	8.16	25	51.02	1	2.04	16	32.65	2	4.08	0	0	1	2.04
		ALMM	1	3.85	16	61.54	2	7.69	4	15.38	2	7.69	1	3.85	0	0
		AL	5	6.67	41	54.67	3	4	20	26.67	4	5.33	1	1.33	1	1.33
	≥180	ALMG	1	3.33	18	60	1	3.33	8	26.67	0	0	1	3.33	1	3.33
		ALMM	6	7.23	31	37.35	1	1.2	39	46.99	1	1.2	2	2.41	3	3.61
		AL	7	6.19	49	43.36	2	1.77	47	41.59	1	0.88	3	2.65	4	3.54
AL type	Kappa	ALMG	2	11.76	10	58.82	2	11.76	2	11.76	1	5.88	0	0	0	0
		ALMM	0	0	13	48.15	1	3.7	11	40.74	0	0	2	7.41	0	0
		AL	2	4.55	23	52.27	3	6.82	13	29.55	1	2.27	2	4.55	0	0
	Lambda	ALMG	3	4.62	34	52.31	0	0	24	36.92	1	1.54	1	1.54	2	3.08
		ALMM	7	8.05	35	40.23	2	2.3	35	40.23	3	3.45	2	2.3	3	3.45
		AL	10	6.58	69	45.39	2	1.32	59	38.82	4	2.63	3	1.97	5	3.29
Creatinine [mg/dL]	<2	ALMG	3	4.11	40	54.79	2	2.74	23	31.51	2	2.74	1	1.37	2	2.74
		ALMM	6	6.12	44	44.9	2	2.04	36	36.73	3	3.06	4	4.08	3	3.06
		AL	9	5.26	84	49.12	4	2.34	59	34.5	5	2.92	5	2.92	5	2.92
	≥2	ALMG	2	22.22	4	44.44	0	0	3	33.33	0	0	0	0	0	0
		ALMM	1	6.25	4	25	1	6.25	10	62.5	0	0	0	0	0	0
		AL	3	12	8	32	1	4	13	52	0	0	0	0	0	0

Table A.7: Myc activation index (MAI) distribution: Number of patients (n) and percentages (%) grouped by MAI $\leq$1 and >1. Separation by AL disease specific variables and disease entities AL, MGUS, AMM, and MM and subentities ALMG and ALMM. See figure 3.16 **a** and **c** for graphical presentation. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

			$\leq$**1**		**>1**	
Variable	**Level**	**Entity**	***n***	**%**	***n***	**%**
		AL	129	65.82	67	34.18
		MGUS	53	82.81	11	17.19
		AMM	158	58.3	113	41.7
		MM	347	45.36	418	54.64
Heart involvement	no	ALMG	17	70.83	7	29.17
		ALMM	17	73.91	6	26.09
		AL	34	72.34	13	27.66
	yes	ALMG	47	81.03	11	18.97
		ALMM	47	53.41	41	46.59
		AL	94	64.38	52	35.62
NT-ProBNP [ng/L]	<1800	ALMG	30	78.95	8	21.05
		ALMM	23	65.71	12	34.29
		AL	53	72.6	20	27.4
	$\geq$1800	ALMG	33	76.74	10	23.26
		ALMM	40	54.05	34	45.95
		AL	73	62.39	44	37.61
Difference FLC [mg/L]	<180	ALMG	35	71.43	14	28.57
		ALMM	19	73.08	7	26.92
		AL	54	72	21	28
	$\geq$180	ALMG	28	93.33	2	6.67
		ALMM	44	53.01	39	46.99
		AL	72	63.72	41	36.28
AL type	Kappa	ALMG	16	94.12	1	5.88
		ALMM	19	70.37	8	29.63
		AL	35	79.55	9	20.45
	Lambda	ALMG	48	73.85	17	26.15
		ALMM	46	52.87	41	47.13
		AL	94	61.84	58	38.16
Creatinine [mg/dL]	<2	ALMG	57	78.08	16	21.92
		ALMM	56	57.14	42	42.86
		AL	113	66.08	58	33.92
	$\geq$2	ALMG	7	77.78	2	22.22
		ALMM	9	56.25	7	43.75
		AL	16	64	9	36

Table A.8: UAMS 70-gene score (UAMS70) distribution: Number of patients (*n*) and percentages (%) grouped by UAMS70 low risk and high risk. entities and subentities. Separation by AL disease specific variables and disease entities AL, MGUS, AMM, and MM and subentities ALMG and ALMM. See figure 3.19 **a** and **c** for graphical presentation. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

			low risk		**high risk**	
Variable	**Level**	**Entity**	***n***	**%**	***n***	**%**
		AL	181	92.35	15	7.65
		MGUS	62	96.88	2	3.12
		AMM	249	91.88	22	8.12
		MM	572	74.77	193	25.23
Heart involvement	no	ALMG	24	100	0	0
		ALMM	20	86.96	3	13.04
		AL	44	93.62	3	6.38
	yes	ALMG	56	96.55	2	3.45
		ALMM	78	88.64	10	11.36
		AL	134	91.78	12	8.22
NT-ProBNP [ng/L]	<1800	ALMG	38	100	0	0
		ALMM	31	88.57	4	11.43
		AL	69	94.52	4	5.48
	≥1800	ALMG	41	95.35	2	4.65
		ALMM	66	89.19	8	10.81
		AL	107	91.45	10	8.55
Difference FLC [mg/L]	<180	ALMG	48	97.96	1	2.04
		ALMM	23	88.46	3	11.54
		AL	71	94.67	4	5.33
	≥180	ALMG	29	96.67	1	3.33
		ALMM	74	89.16	9	10.84
		AL	103	91.15	10	8.85
AL type	Kappa	ALMG	17	100	0	0
		ALMM	24	88.89	3	11.11
		AL	41	93.18	3	6.82
	Lambda	ALMG	63	96.92	2	3.08
		ALMM	77	88.51	10	11.49
		AL	140	92.11	12	7.89
Creatinine [mg/dL]	<2	ALMG	71	97.26	2	2.74
		ALMM	86	87.76	12	12.24
		AL	157	91.81	14	8.19
	≥2	ALMG	9	100	0	0
		ALMM	15	93.75	1	6.25
		AL	24	96	1	4

Table A.9: IFM 15-gene score (IFM15) distribution: Number of patients (*n*) and percentages (%) grouped by IFM15 low risk and high risk. Separation by AL disease specific variables and disease entities AL, MGUS, AMM, and MM and subentities ALMG and ALMM. See figure 3.20 **a** and **c** for graphical presentation. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

			low risk		**high risk**	
Variable	**Level**	**Entity**	***n***	**%**	***n***	**%**
		AL	184	93.88	12	6.12
		MGUS	59	92.19	5	7.81
		AMM	248	91.51	23	8.49
		MM	574	75.03	191	24.97
Heart involvement	no	ALMG	22	91.67	2	8.33
		ALMM	23	100	0	0
	yes	AL	45	95.74	2	4.26
		ALMG	55	94.83	3	5.17
		ALMM	81	92.05	7	7.95
		AL	136	93.15	10	6.85
NT-ProBNP [ng/L]	<1800	ALMG	34	89.47	4	10.53
		ALMM	33	94.29	2	5.71
	≥1800	AL	67	91.78	6	8.22
		ALMG	42	97.67	1	2.33
		ALMM	69	93.24	5	6.76
		AL	111	94.87	6	5.13
Difference FLC [mg/L]	<180	ALMG	45	91.84	4	8.16
		ALMM	24	92.31	2	7.69
	≥180	AL	69	92	6	8
		ALMG	29	96.67	1	3.33
		ALMM	78	93.98	5	6.02
		AL	107	94.69	6	5.31
AL type	Kappa	ALMG	16	94.12	1	5.88
		ALMM	26	96.3	1	3.7
	Lambda	AL	42	95.45	2	4.55
		ALMG	61	93.85	4	6.15
		ALMM	81	93.1	6	6.9
		AL	142	93.42	10	6.58
Creatinine [mg/dL]	<2	ALMG	68	93.15	5	6.85
		ALMM	92	93.88	6	6.12
	≥2	AL	160	93.57	11	6.43
		ALMG	9	100	0	0
		ALMM	15	93.75	1	6.25
		AL	24	96	1	4

Table A.10: Risk score (RS) distribution: Number of patients (*n*) and percentages (%) grouped by Rs low risk, medium risk and high risk. Separation by AL disease specific variables and disease entities AL, MGUS, AMM, and MM and subentities ALMG and ALMM. See figure 3.21 **a** and **c** for graphical presentation. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

			low risk		**medium risk**		**high risk**	
Variable	**Level**	**Entity**	***n***	**%**	***n***	**%**	***n***	**%**
		AL	150	76.53	39	19.9	7	3.57
		MGUS	57	89.06	7	10.94	0	0
		AMM	216	79.7	54	19.93	1	0.37
		MM	541	51.13	435	41.12	82	7.75
Heart involvement	no	ALMG	19	79.17	5	20.83	0	0
		ALMM	18	78.26	4	17.39	1	4.35
		AL	37	78.72	9	19.15	1	2.13
	yes	ALMG	48	82.76	10	17.24	0	0
		ALMM	63	71.59	19	21.59	6	6.82
		AL	111	76.03	29	19.86	6	4.11
NT-ProBNP [ng/L]	<1800	ALMG	31	81.58	7	18.42	0	0
		ALMM	27	77.14	7	20	1	2.86
		AL	58	79.45	14	19.18	1	1.37
	≥1800	ALMG	35	81.4	8	18.6	0	0
		ALMM	52	70.27	17	22.97	5	6.76
		AL	87	74.36	25	21.37	5	4.27
Difference FLC [mg/L]	<180	ALMG	40	81.63	9	18.37	0	0
		ALMM	23	88.46	2	7.69	1	3.85
		AL	63	84	11	14.67	1	1.33
	≥180	ALMG	24	80	6	20	0	0
		ALMM	56	67.47	22	26.51	5	6.02
		AL	80	70.8	28	24.78	5	4.42
AL type	Kappa	ALMG	16	94.12	1	5.88	0	0
		ALMM	20	74.07	5	18.52	2	7.41
		AL	36	81.82	6	13.64	2	4.55
	Lambda	ALMG	51	78.46	14	21.54	0	0
		ALMM	63	72.41	19	21.84	5	5.75
		AL	114	75	33	21.71	5	3.29
Creatinine [mg/dL]	<2	ALMG	60	82.19	13	17.81	0	0
		ALMM	72	73.47	19	19.39	7	7.14
		AL	132	77.19	32	18.71	7	4.09
	≥2	ALMG	7	77.78	2	22.22	0	0
		ALMM	11	68.75	5	31.25	0	0
		AL	18	72	7	28	0	0

Table A.11: Heidelberg AL score (HDAL) distribution: Number of patients (*n*) and percentages (%) grouped by HDAL low risk, medium risk and high risk. Separation by AL disease specific variables and disease entities AL, MGUS, AMM, and MM and subentities ALMG and ALMM. See figure 3.22 **a** and **c** for graphical presentation. AL: light chain amyloidosis, ALMG: AL with subentity MGUS, ALMM: AL with subentity MM, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

			low risk		**medium risk**		**high risk**	
Variable	**Level**	**Entity**	***n***	**%**	***n***	**%**	***n***	**%**
		AL	94	47.96	56	28.57	46	23.47
		MGUS	45	70.31	11	17.19	8	12.5
		AMM	96	35.42	106	39.11	69	25.46
		MM	93	12.16	188	24.58	484	63.27
Heart involvement	no	ALMG	21	87.5	3	12.5	0	0
		ALMM	8	34.78	9	39.13	6	26.09
		AL	29	61.7	12	25.53	6	12.77
	yes	ALMG	35	60.34	16	27.59	7	12.07
		ALMM	29	32.95	26	29.55	33	37.5
		AL	64	43.84	42	28.77	40	27.4
NT-ProBNP [ng/L]	<1800	ALMG	31	81.58	7	18.42	0	0
		ALMM	14	40	14	40	7	20
		AL	45	61.64	21	28.77	7	9.59
	≥1800	ALMG	25	58.14	12	27.91	6	13.95
		ALMM	23	31.08	21	28.38	30	40.54
		AL	48	41.03	33	28.21	36	30.77
Difference FLC [mg/L]	<180	ALMG	33	67.35	13	26.53	3	6.12
		ALMM	17	65.38	7	26.92	2	7.69
		AL	50	66.67	20	26.67	5	6.67
	≥180	ALMG	21	70	6	20	3	10
		ALMM	20	24.1	28	33.73	35	42.17
		AL	41	36.28	34	30.09	38	33.63
AL type	Kappa	ALMG	13	76.47	4	23.53	0	0
		ALMM	12	44.44	7	25.93	8	29.63
		AL	25	56.82	11	25	8	18.18
	Lambda	ALMG	43	66.15	15	23.08	7	10.77
		ALMM	26	29.89	30	34.48	31	35.63
		AL	69	45.39	45	29.61	38	25
Creatinine [mg/dL]	<2	ALMG	50	68.49	17	23.29	6	8.22
		ALMM	34	34.69	34	34.69	30	30.61
		AL	84	49.12	51	29.82	36	21.05
	≥2	ALMG	6	66.67	2	22.22	1	11.11
		ALMM	4	25	3	18.75	9	56.25
		AL	10	40	5	20	10	40

Table A.12: Heidelberg AL score (HDAL) prognostic genes: Affymetrix probeID, prognostic factor, Gene symbol, position in genome and protein biotypes. These are PC: protein coding, AS: antisense, IGD: Immunoglobulin constant gene, IGHV: Immunoglobulin heavy variable, lR: long intergenic noncoding RNA, lncR: long noncoding RNA, PP: processed pseudogene, PT: processed transcript, UP: unprocessed pseudogene.

ProbeID	Factor	Gene	Position	Biotype
1552613_s_at	good	*CDC42SE2*	5q31.1	PC
1556072_at	good	*LINC00528*	22q11.21	lR
201658_at	good	*ARL1*	12q23.2	PC
201976_s_at	good	*MYO10*	5p15.1	PC
208612_at	good	*PDIA3*	15q15.3	PC
210749_x_at	good	*DDR1*	6p21.33	PC
213122_at	good	*TSPYL5*	8q22.1	PC
213257_at	good	*SARM1*	17q11.2	PC
213674_x_at	good	*IGHD*	14q32.33	IGC
214083_at	good	*PPP2R5C*	14q32.31	PC
216510_x_at	good	*IGHV3-33*	14q32.33	IGHV
221648_s_at	good	*AGMAT*	1p36.21	PC
228624_at	good	*TMEM144*	4q32.1	PC
233021_at	good	*RBM26-AS1*	13q31.1	AS
238822_at	good			
1555288_s_at	poor	*FBF1,*	17q25.1	PC
1555971_s_at	poor	*RP11-552F3.12 FBXO28*	1q42.11	PC
1566106_at	poor	*AK091277*		
200822_x_at	poor	*TPI1, TPI1P1*	12p13.31, 1p31.1	PC, PP
201584_s_at	poor	*DDX39A*	19p13.12	PC
202212_at	poor	*PES1*	22q12.2	PC
202248_at	poor	*E2F4*	16q22.1	PC
202325_s_at	poor	*ATP5J*	21q21.3	PC
202910_s_at	poor	*ADGRE5*	19p13.12	PC
203340_s_at	poor	*SLC25A12*	2q31.1	PC
203448_s_at	poor	*TERF1P4, TERF1P5, RP11-311P8.2, TERF1, TERF1P1*	Xq21.1, 13q11, Xq13.3, 8q21.11, 21q11.2	PP, PC
205407_at	poor	*RECK*	9p13.3	PC
207104_x_at	poor	*LILRB1*	19q13.42	PC
210152_at	poor	*LILRB4*	19q13.42	PC
210252_s_at	poor	*MADD*	11p11.2	PC
211336_x_at	poor	*LILRB1*	19q13.42	PC
214149_s_at	poor	*ATP6V0E1*		
214882_s_at	poor	*SRSF2*	17q25.1	PC
214931_s_at	poor	*SRPK2*	7q22.3	PC
218115_at	poor	*ASF1B*	19p13.12	PC
219172_at	poor	*UBTD1*	10q24.1	PC

219816_s_at	poor	*RBM23*	14q11.2	PC
221042_s_at	poor	*CLMN*	14q32.13	PC
221209_s_at	poor	*OTOR*	20p12.1	PC
221418_s_at	poor	*MED16*	19p13.3	PC
222989_s_at	poor	*UBQLN1*	9q21.32	PC
223252_at	poor	*HDGFRP2*	19p13.3	PC
225456_at	poor	*MED1*	17q12	PC
227587_at	poor	*KRI1*	19p13.2	PC
227875_at	poor	*KLHL13*	Xq24	PC
228146_at	poor	*C17orf51*, RP11-822E23.8	17p11.2	PC, PT
228361_at	poor	*E2F2*	1p36.12	PC
229804_x_at	poor	*CBWD7, RP11-15J10.1, CBWD1, CBWD3, CBWD5, CBWD2*	9p11.2, 9q21.11, 9p24.3, 2q14.1	PC, UP
231930_at	poor	*ELMOD1*	11q22.3	PC
232351_at	poor	*AK001012*		
234875_at	poor	*RPL7AP10*	19p12	PP
235394_at	poor	*PLAA*	9p21.2	PC
235848_x_at	poor	*ATL2*	2p22.1	PC
238703_at	poor	*FAM207BP*	13q11	PP
238821_at	poor	*CSTF2*	Xq22.1	PC
243020_at	poor	*FAM13A-AS1*	4q22.1	lncR
243618_s_at	poor	*ZNF827*	4q31.22	PC
34225_at	poor	*NELFA*	4p16.3	PC
56821_at	poor	*SLC38A7*	16q21	PC

Table A.13: Genes only differentially expressed in the comparison BMPC *versus* AL. The LFC indicates the magnitude of difference between BMPC and AL. Mean expression values from RNA sequencing are depicted as log counts per million. LFC: log fold change, adj. *p*: BH adjusted p-value, Protein biotypes for a gene are PC: protein coding, AS: antisense, IGD: Immunoglobulin diversity gene, lR: lincRNA, PP: processed pseudogene, TPP: transcribed processed pseudogene, TUP: transcribed unprocessed pseudogene. Sample entities are BMPC: bone marrow plasma cells, AL: light chain amyloidosis, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

		adj.		log counts per million				
Gene	**Biotype**	*p*	**LFC**	**BMPC**	**AL**	**MGUS**	**AMM**	**MM**
IGHD6-19	IGD	.01	3.44	0.42	0.04	0.10	0.08	0.09
H1FX-AS1	AS	.03	2.30	2.23	0.80	1.10	1.19	1.09
PI4KAP1	TUP	.03	2.69	1.11	0.23	0.27	0.33	0.38
AC159540.1	lR	.04	2.93	0.41	0.05	0.08	0.09	0.13
RP11-58H15.1	TPP	.04	2.50	2.42	0.82	0.85	0.99	1.09
HSH2D	PC	.04	1.96	8.09	6.15	7.04	6.58	6.42
RPL35AP32	PP	.04	2.75	0.40	0.06	0.12	0.10	0.13
IGHD7-27	IGD	.05	2.89	0.13	0.01	0.04	0.02	0.04
NFKBID	PC	.05	2.51	2.73	0.99	1.55	1.66	1.33

Table A.14: Genes only differentially expressed between AL *versus* MGUS, AMM and MM. The log fold change indicates the magnitude of difference between AL and MGUS, AMM and MM. Mean expression values from RNA sequencing are depicted as log counts per million. Protein biotypes for a gene are PC: protein coding, AS: antisense, IGD: Ig diversity gene, IGV: Ig variable gene, scR: scaRNA, Ig: Immunoglobulin. Sample entities are AL: light chain amyloidosis, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

		log fold change			log counts per million				
Gene	**Biotype**	**MGUS**	**AMM**	**MM**	**BMPC**	**AL**	**MGUS**	**AMM**	**MM**
HES1	PC	-2.26	-3.78	-1.70	0.95	0.27	1.00	1.96	0.75
FOLH1	PC	2.38	1.85	2.29	0.15	0.85	0.20	0.29	0.22
RASD1	PC	2.49	3.62	1.93	3.95	4.34	2.14	1.36	2.59
IGHD2-8	IGD	2.91	4.68	3.24	0.56	1.08	0.20	0.05	0.15
BARX2	PC	-2.98	-2.11	-2.76	0.38	1.21	3.50	2.74	3.31
PAGE1	PC	-3.35	-2.74	-4.34	0.01	0.42	2.21	1.75	3.04
HTR1D	PC	3.10	1.99	2.08	0.06	0.73	0.10	0.21	0.20
SCARNA22	scR	-2.82	-3.77	-2.88	0.03	0.02	0.20	0.38	0.21
RP11-669N7.2	AS	-3.29	-3.65	-5.29	0.04	0.08	0.68	0.83	1.78
SSTR1	PC	-3.27	-3.19	-2.26	0.07	0.17	1.19	1.15	0.71
IGHV1OR15-2	IGV	3.61	5.38	6.64	3.31	5.27	2.03	0.93	0.47
IGLV6-57	IGV	3.84	5.17	6.70	11.34	15.54	11.70	10.36	8.84
HBE1	PC	-3.70	-2.24	-4.59	0.06	0.07	0.75	0.32	1.18

Table A.15: Genes differentially expressed in the comparison BMPC *versus* AL, MGUS, AMM and MM. The log fold change indicates the magnitude of difference between BMPC and AL, MGUS, AMM and MM. If log fold change is missing the gene is not differentially expressed. Mean expression values from RNA sequencing are depicted as log counts per million. Protein biotypes for a gene are PC: protein coding, AS: antisense, IGC: Ig constant gene, IGV: Ig variable gene, lR: lincRNA, PP: processed pseudogene, PT: processed transcript, sR: snRNA, TPP: transcribed processed pseudogene, TUP: transcribed unprocessed pseudogene, UP: unprocessed pseudogene, Ig: Immunoglobulin. Sample entities are BMPC: bone marrow plasma cells, AL: light chain amyloidosis, MGUS: monoclonal gammopathy of undetermined significance, AMM: asymptomatic multiple myeloma, MM: multiple myeloma.

		log fold change				**log counts per million**				
Gene	**Biotype**	**AL**	**MGUS**	**AMM**	**MM**	**BMPC**	**AL**	**MGUS**	**AMM**	**MM**
LINC00173	PT	2.71			2.47	0.94	0.18	0.24	0.23	0.22
RBMS1	PC	-2.03		-2.09	-2.30	2.98	4.88	4.55	4.92	5.13
IGHV3-64	IGV	3.53			4.20	7.32	3.89	5.30	6.50	3.27
SCN9A	PC	-3.45	-3.70	-3.94	-3.72	0.75	3.08	3.30	3.52	3.33
HEY2	PC	-2.49			-2.59	1.70	3.78	3.05	3.47	3.87
PRKRIRP3	PP	-3.66		-3.89	-4.01	0.05	0.58	0.45	0.66	0.70
IGHV7-56	IGV	2.92		3.91	4.68	1.60	0.34	0.45	0.18	0.11
RP11-669M16.1	lR	-2.72		-3.15	-2.88	0.45	1.80	1.52	2.12	1.91
ADGRB3	PC	-4.82		-4.29	-4.67	0.33	3.07	2.09	2.61	2.94
TRAPPC13P1	PP	-5.86		-5.83	-5.60	0.00	0.38	0.40	0.37	0.32
HSD17B7P2	TUP	2.43	2.65	2.37	2.41	2.42	0.85	0.75	0.88	0.86
C1orf21	PC	-3.89		-4.05	-3.54	1.03	4.05	3.24	4.20	3.73
MAB21L1	PC	-7.25		-6.23	-6.61	0.06	3.02	2.05	2.17	2.48
RP11-582J16.5	AS	2.77		2.61	2.47	0.77	0.13	0.15	0.15	0.17
RP3-425C14.4	TEC	-2.28		-2.81	-2.41	1.30	3.02	3.03	3.50	3.14
UGT8	PC	-3.44		-3.47	-3.51	1.19	3.91	3.19	3.93	3.97
CCDC144CP	TPP	-4.45			-3.77	0.30	2.61	2.28	1.91	2.07
DKK1	PC	-4.81	-5.10	-5.90	-7.17	0.94	4.74	5.03	5.80	7.06
ANO5	PC	-3.05		-3.06	-3.17	0.99	3.20	2.97	3.21	3.30
GOLGA8S	PC	-5.15		-5.15	-5.67	0.08	1.73	1.35	1.73	2.12
RP3-460G2.2	lR	-2.80	-3.17	-3.97	-4.27	0.38	1.65	1.91	2.54	2.79
PDZRN4	PC	-3.23		-3.15	-2.94	2.19	5.10	4.85	5.03	4.83
KCNS3	PC	-7.06		-6.91	-7.26	0.11	3.69	3.59	3.56	3.88
PRR15	PC	-4.82		-5.26	-4.72	0.45	3.55	3.25	3.96	3.46
PDXDC2P	PT	2.72		2.60	2.45	2.08	0.57	0.64	0.61	0.67
MAGEC3	PC	-5.29		-5.85	-6.02	0.04	1.28	0.65	1.63	1.75
MAGEC2	PC	-7.68		-8.79	-9.05	0.03	2.65	1.73	3.63	3.87
DTX3L	PC	-2.42		-2.60	-2.38	3.54	5.86	5.69	6.03	5.82
BMP4	PC	-5.89	-6.30	-6.59	-6.47	0.20	3.36	3.73	4.00	3.89
TGFB2	PC	-6.74		-6.31	-6.22	0.10	3.26	2.54	2.87	2.80
WNT5A	PC	-4.72		-4.31	-4.70	0.87	4.52	3.76	4.13	4.51
MTG1	PC	1.97	2.27		1.74	2.18	0.92	0.79	1.07	1.04
FAM65C	PC	2.26		2.37	2.57	3.53	1.68	1.73	1.60	1.47
LSAMP	PC	-4.13	-4.10	-4.65	-4.76	1.13	4.45	4.42	4.95	5.06
PARP9	PC	-2.52		-2.68	-2.43	3.76	6.19	5.99	6.35	6.10
FMN1	PC	-3.42		-2.76	-2.85	1.26	4.01	3.54	3.39	3.48
NLGN4X	PC	-6.07		-5.48	-5.45	0.92	5.94	4.62	5.36	5.33
RP11-734I18.1	lR	-4.70	-4.51	-5.24	-5.88	0.01	0.47	0.42	0.65	0.92
CNTN5	PC	-4.83	-6.01	-6.25	-6.66	0.52	3.77	4.88	5.12	5.52
CPXM1	PC	3.33		3.45	3.06	1.93	0.35	0.41	0.32	0.41
DHRS9	PC	-6.09	-5.58	-5.99	-4.87	1.47	6.93	6.42	6.83	5.73
PRDM5	PC	-3.35		-3.40	-3.54	2.07	5.07	4.20	5.12	5.26
U47924.31	AS	2.78	2.88	2.51	2.35	0.99	0.19	0.18	0.23	0.25
DMRT2	PC	-5.89	-5.67	-6.10	-6.23	0.04	1.49	1.36	1.63	1.72
STXBP6	PC	-3.95	-3.50	-3.98	-4.15	0.72	3.47	3.07	3.50	3.66

AC098828.3	PP	-4.57	-4.05	-4.94	-4.69	0.05	0.98	0.74	1.17	1.04
VPREB1	PC	3.51			3.79	2.41	0.46	1.57	1.64	0.39
DQX1	PC	3.51		3.23	3.19	2.31	0.43	0.68	0.51	0.52
PLA2G4A	PC	-3.15		-3.27	-3.21	1.33	3.85	3.35	3.97	3.91
KIAA1683	PC	2.95		3.51	3.33	3.89	1.48	1.74	1.14	1.25
RP11-325P15.2	PP	3.22		3.37	2.90	2.50	0.58	1.08	0.53	0.69
FCRLB	PC	-4.55			-3.57	1.36	5.24	3.73	3.79	4.29
RP11-395G23.3	lR	-3.34	-3.70	-4.11	-4.31	0.12	0.96	1.14	1.38	1.51
FBLN2	PC	-5.87	-5.18	-5.94	-6.49	0.35	4.13	3.49	4.19	4.72
CNTN1	PC	-5.56	-4.76	-5.24	-5.51	1.21	5.97	5.19	5.67	5.93
CCND1	PC	-9.31		-7.04	-7.83	1.23	9.74	6.98	7.48	8.26
MAGEC1	PC	-9.36	-9.39	-10.35	-11.01	0.02	3.50	3.53	4.43	5.07
NPEPL1	PC	2.27	2.57	2.56	2.17	2.87	1.21	1.04	1.05	1.26
RNU1-85P	sR	3.54		3.18	3.78	0.94	0.10	0.31	0.13	0.09
KIT	PC	-6.78	-6.19	-7.00	-6.56	0.82	6.41	5.83	6.63	6.19
TMEM52B	PC	-5.68	-5.01	-6.09	-5.89	0.45	4.32	3.69	4.71	4.52
FAM171B	PC	-4.37		-3.99	-4.20	0.30	2.54	1.66	2.24	2.41
PTPRK	PC	-3.99	-3.37	-3.68	-4.09	1.08	4.23	3.65	3.94	4.33
ESRRG	PC	-4.55	-4.05	-4.92	-4.70	0.72	4.04	3.57	4.39	4.18
ACVR1C	PC	-3.53		-2.59	-2.86	1.01	3.67	2.66	2.83	3.07
IGLV7-35	IGV	3.64		3.45	5.46	1.92	0.29	0.65	0.32	0.08
CALCRL	PC	-3.92	-3.10	-4.34	-4.21	1.44	4.75	3.96	5.15	5.03
MFAP3L	PC	-4.67	-3.83	-3.81	-3.54	1.56	5.67	4.85	4.83	4.57
HGF	PC	-6.35	-6.77	-6.97	-7.69	0.61	5.46	5.88	6.08	6.78
SLC39A8	PC	-2.65		-2.12	-2.28	4.19	6.78	5.81	6.25	6.41
RBFOX2	PC	-2.68	-2.23	-2.58	-2.60	2.21	4.60	4.17	4.51	4.52
VCPKMT	PC	2.27		2.20	2.33	5.28	3.15	3.43	3.21	3.09
IGHEP1	IGC	4.05		3.28	5.70	3.45	0.67	0.95	1.01	0.25
IGHE	IGC	4.46		3.12	6.01	8.88	4.48	5.66	5.79	3.06
FAM133A	PC	-7.24	-6.41	-7.73	-8.83	0.03	2.47	1.83	2.88	3.88
NDNF	PC	-5.37	-4.61	-5.03	-5.23	0.94	5.29	4.56	4.96	5.16
RP5-857K21.6	UP	3.99		3.32	2.57	3.51	0.72	1.81	1.03	1.46
TEX14	PC	3.65	3.11	3.35	3.40	4.05	1.16	1.49	1.34	1.31
IGHV7-34-1	IGV	4.76	3.77	4.43	4.91	1.56	0.09	0.19	0.12	0.09
EDA2R	PC	-5.76	-5.62	-6.35	-6.54	0.15	2.89	2.76	3.40	3.58

Table A.16: Gene ontology (GO) terms enriched in 70 differentially expressed genes (DEG) from both comparisons BMPC *versus* AL and BMPC *versus* MM detected by functional enrichment analysis.

Term	**Description**
GO:0061311	cell surface receptor signaling pathway involved in heart development
GO:0030902	hindbrain development
GO:0010811	positive regulation of cell-substrate adhesion
GO:0030218	erythrocyte differentiation
GO:0050731	positive regulation of peptidyl-tyrosine phosphorylation
GO:0034330	cell junction organization
GO:0007219	Notch signaling pathway
GO:0051962	positive regulation of nervous system development
GO:0051098	regulation of binding
GO:0099054	presynapse assembly

Table A.17: Overlap of genes harboring variants in 113 AL samples and in previously published lists of genes by **A**: Boyle *et al.* [36] **B**: Rossi *et al.* [254] **C**: Paiva *et al.* [230] **D**: Walker *et al.* [308]. Mean expression values are the mean of the log2 transformed, normalized counts from the RNA sequencing data of the samples harboring at least one variant in the gene. AL: light chain amyloidosis, VAF: variant allele frequency.

Gene	Variants	Samples	Median VAF DNA	Expressed in RNA	Mean expression	A	B	C	D
KRAS	8	8	37.71	7	4.97		x	x	x
KLHL6	5	4	27.71	5	6.54	x			x
MIA2	9	6	55.00	4	2.02			x	
TP53	6	4	21.07	4	4.74		x		x
ASCC3	7	6	42.31	3	6.07			x	
DIS3	7	6	41.07	3	4.26	x		x	x
FAM208A	5	2	55.56	3	4.32			x	
NRAS	5	5	16.67	3	7.05	x	x		x
TET2	6	4	38.35	3	5.05	x			x
XBP1	3	3	68.75	3	11.23				x
ZNF292	8	5	50.89	3	5.39	x			x
ACTG1	3	3	41.03	2	8.98				x
BIRC6	3	3	48.72	2	5.53			x	
BRAF	3	3	39.02	2	3.82		x		x
CCND1	3	3	29.55	2	10.05		x		x
EP300	4	3	53.09	2	3.68	x			x
KMT2B	5	4	55.63	2	2.04	x			x
NBR1	2	2	51.13	2	6.41			x	
PRDM1	3	3	49.33	2	7.94				x
TRAF3	2	2	58.61	2	3.90	x			x
TTLL4	2	2	79.03	2	2.67			x	
AHNAK	6	4	44.58	1	4.93			x	
ALKBH4	1	1	36.33	1	3.45			x	
ARID2	1	1	27.17	1	5.33				x
CDKN1B	2	2	43.78	1	5.72				x
CSF3R	3	2	47.86	1	3.10			x	
FYCO1	14	5	65.70	1	2.95			x	
HIST1H1E	2	2	44.94	1	3.75				x
IDH1	3	3	37.82	1	7.13				x
ORC4	2	2	55.71	1	3.74			x	
RAD51D	1	1	39.83	1	2.88			x	
RASA2	3	3	38.89	1	5.41				x
SETD2	2	2	37.95	1	4.09				x
SP140	2	2	48.66	1	5.17				x
ABCF1	2	2	30.33	0					x
AP3B2	1	1	62.50	0				x	
ARID1A	4	4	37.26	0					x
ASIC4	1	1	18.00	0				x	
ATM	2	2	46.88	0					x
ATRX	1	1	66.91	0					x
BSN	5	4	43.24	0				x	
BTBD17	2	2	48.67	0				x	
C14orf39	11	4	47.25	0				x	
C8orf34	1	1	40.00	0					x
CACNA1I	2	2	82.41	0				x	
CCDC17	3	2	50.98	0				x	
CCDC39	1	1	44.64	0				x	
CELSR1	10	3	80.69	0				x	
CPXCR1	6	5	100.00	0				x	
CREBBP	1	1	21.29	0					x

CTNNAL1	1	1	34.62	0				x	
DCAF12L2	3	3	100.00	0				x	
DNAH5	12	8	45.10	0				x	
DNAH9	12	10	50.16	0				x	
DNMT3A	2	2	25.39	0					x
DSEL	4	3	33.08	0				x	
DUSP2	3	3	22.93	0		x			x
EGR1	1	1	54.66	0		x			x
EML6	1	1	42.31	0				x	
FBXO15	2	2	31.44	0				x	
HECW1	2	2	43.16	0				x	
HUWE1	3	3	56.16	0					x
KCNT1	1	1	64.71	0				x	
KDM6A	4	4	64.06	0					x
KMT2C	2	2	52.94	0					x
KRT28	1	1	55.66	0				x	
LAMA3	4	4	32.04	0				x	
LTB	1	1	38.37	0					x
MAML2	2	2	53.13	0					x
MYO18B	11	8	46.15	0				x	
NAALAD2	4	4	41.44	0				x	
NBEAL1	1	1	26.32	0				x	
NF1	2	2	61.29	0					x
OCA2	3	3	48.96	0				x	
PCLO	14	10	37.58	0				x	
PDE8B	1	1	29.57	0				x	
PIK3CA	1	1	50.98	0					x
PKD1	15	4	46.59	0				x	
PLCB4	2	2	100.00	0				x	
PRPF4B	1	1	29.56	0				x	
PSCA	2	2	73.68	0				x	
QPCT	1	1	53.28	0				x	
RBP3	2	2	38.78	0				x	
RFTN1	1	1	51.85	0					x
SALL2	3	3	48.28	0				x	
SERPINA5	3	3	65.00	0				x	
SI	8	8	36.76	0				x	
SLX4	6	3	49.46	0				x	
SPATA31D1	1	1	47.46	0				x	
SPOCK1	1	1	36.16	0				x	
STPG2	4	2	77.22	0				x	
SWSAP1	1	1	25.69	0				x	
TTN	70	14	45.06	0				x	
TTR	1	1	50.59	0				x	
UBR5	1	1	10.27	0					x
USP54	1	1	25.51	0				x	
ZFYVE1	1	1	47.79	0				x	
ZNF519	6	5	77.66	0				x	
ZNF729	3	3	29.31	0				x	

Table A.18: Immunoglobulin (Ig) genes in which variants were detected in light chain amyloidosis separated by Ig gene group.

Ig gene group	**Genes**
Heavy Chain Constat Alpha	*IGHA1, IGHA2*
Heavy Chain Constat Delta	*IGHD, IGHD3-16, IGHD3-10, IGHD2-2, IGHD6-6, IGHD6-25*
Heavy Chain Constat Gamma	*IGHG1, IGHG2, IGHG3, IGHG4*
Heavy Chain Joining	*IGHJ6, IGHJ4, IGHJ5, IGHJ3*
Heavy Chain Constat Mu	*IGHM*
Heavy Chain Variable	*IGHV1-18, IGHV2-70, IGHV2-70D, IGHV3-53, IGHV1-69D, IGHV3-11, IGHV4-34, IGHV1OR15-9, IGHV3OR15-7, IGHV5-10-1, IGHV3-13, IGHV3-16, IGHV3-20, IGHV3-23, IGHV1-45, IGHV3-49, IGHV1-58, IGHV3-64, IGHV1-2, IGHV3-30, IGHV3-35, IGHV4-31, IGHV3-74, IGHV3-38, IGHV5-51, IGHV1-69, IGHV7-4-1, IGHV3-43, IGHV3-66, IGHV4-28*
Light Chain Kappa Constant	*IGKC*
Light Chain Kappa Variable	*IGKV2-24, IGKV5-2, IGKV2-30, IGKV2D-29, IGKV3-7, IGKV1-16, IGKV2D-26, IGKV1D-16, IGKV3-15, IGKV4-1, IGKV3D-11, IGKV1-8, IGKV1-27, IGKV1D-33, IGKV1-39, IGKV1D-37, IGKV6-21*
Light Chain Lambda Constant	*IGLC2, IGLC3, IGLC7*
Light Chain Lambda Joining	*IGLJ3, IGLJ2*
Light Chain Lambda Varaible	*IGLV4-60, IGLV3-1, IGLV4-3, IGLV4-69, IGLV5-45, IGLV7-46, IGLV3-25, IGLV3-12, IGLV6-57, IGLV3-22, IGLV3-10, IGLV2-8, IGLV3-21, IGLV3-19, IGLV5-37, IGLV1-44, IGLV1-47, IGLV2-18, IGLV2-14, IGLV10-54, IGLV2-23, IGLV3-16, IGLV8-61*

B Supplementary Code

```
# ------------------------------------------------------------------------ #
# R Code                                                                   #
# HDAL                                                                     #
# ------------------------------------------------------------------------ #


# Load docval functions for normalization of one CEL file to the training group
source("../HDAL/Scripts/docval.R")

# norm.external.params
# Function options:
# cel:      CEL file name
# path:     path to CEL file
# params:   normalization parameters of training group

norm.external.params <- function(cel=celfiles, path=path, params=params){

  for (i in 1:length(cel)) {
    cel.file <- cel[i]
    external <- ReadAffy(filenames=cel.file, celfile.path=path)

    exprs.external.gcrma <- wrap.val.add(external, params, method="gcrma")
    don <- data.frame(exprs(exprs.external.gcrma))
    if (i==1) ra <- don else ra <- cbind(ra,don)
    colnames(ra)[i] <- as.character(cel[i])
    print(paste(i, "done"))
  }
  ra
}

# HDAL
# Function options:
# cel:      CEL file name
# path:     path to CEL file
# entity:   could be: AL (light chain amyloidosis),
#           MGUS (monoclonal gammopathy of undetermined significance),
#           AMM (asymptomatic multiple myeloma),
#           MM (symptomatic multiple myeloma)

HDAL <- function(cel, entity="AL", path=""){

  # prognostic genes
  al_genes <-  c("1552613_s_at", "1556072_at",  "201658_at",   "201976_s_at",
                 "208612_at",    "210749_x_at", "213122_at",   "213257_at",
                 "213674_x_at",  "214083_at",   "216510_x_at", "221648_s_at",
                 "228624_at",    "233021_at",   "238822_at",   "1555288_s_at",
                 "1555971_s_at", "1566106_at",  "200822_x_at", "201584_s_at",
```

```
                "202212_at",    "202248_at",   "202325_s_at", "202910_s_at",
                "203340_s_at",  "203448_s_at", "205407_at",   "207104_x_at",
                "210152_at",    "210252_s_at", "211336_x_at", "214149_s_at",
                "214882_s_at",  "214931_s_at", "218115_at",   "219172_at",
                "219816_s_at",  "221042_s_at", "221209_s_at", "221418_s_at",
                "222989_s_at",  "223252_at",   "225456_at",   "227587_at",
                "227875_at",    "228146_at",   "228361_at",   "229804_x_at",
                "231930_at",    "232351_at",   "234875_at",   "235394_at",
                "235848_x_at",  "238703_at",   "238821_at",   "243020_at",
                "243618_s_at",  "34225_at",    "56821_at")

  al_factor <- c(rep(-1, 15),rep(1, 44))

  # 1. Normalization
  (load(".../HDAL/Data/HDAL_params.txt")) # load parameters
  nc <- norm.external.params(cel=cel, path=path, params=params)

  # 2. stop if not all al_genes are available in the counttable
  if(any(!(al_genes %in% row.names(nc)))){
    stop("Not all 59 prognostic genes are in count table!")
  }

  # 3. score estimation by summing up the 59 genes multiplied by a factor and
      splitting in 3 groups
  nc.al <- nc[al_genes,]  # cutting nc to 59 al_genes
  score <- sum(nc.al*as.numeric(al_factor))

  # 4. low and high cutoff
  lcut=149.385060831708
  hcut=161.402300016636

  risk <- ifelse(score<=lcut, "low risk",
                 ifelse(score<=hcut, "medium risk", "high risk"))

  data.frame(ID=colnames(nc), score=score, risk=risk, entity=as.character(entity))
}
```

Code B.1: HDAL: Estimation of new CEL files

```
# ------------------------------------------------------------------------ #
# Quality control                                                          #
# ------------------------------------------------------------------------ #

# Quality report with FastQC:
# FastQC version: 0.11.7
#
# options:
# --threads number of computing threads
```

```
mkdir $reportdir
./fastqc --threads $cpu $seqdir/*.txt.gz --outdir $reportdir

# Automatic trimming of reads with fastp
# fastp version: 0.19.4
#
# options:
# -i input reads one for paired end
# -I input reads two for paired end
# -o trimmed output reads one for paired end
# -O trimmed output reads two for paired end
# --dont_overwrite stop if quality reports already exists
# -V
# -j name for the quality control report JSON file
# -h name for the quality control report HTML file
# -R Sample name for report

./fastp -i $reads1 -I $reads2 -o $sequence1 -O $sequence2 --dont_overwrite -V
    -j $samplename.r.json -h $samplename.r.html -R $samplename
```

Code B.2: Quality control FASTQ files

```
# ------------------------------------------------------------------------ #
# Alignment and preprocessing                                              #
# ------------------------------------------------------------------------ #

# Reference genome GRCh38 download link:
# ftp://ftp.ensembl.org/pub/release-77/fasta/homo_sapiens/dna/
    Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz

# Alignment with bwa mem and processing with samtools, picard and GATK
#
# bwa version: 0.7.12
# samtools version: 1.5 using htslib 1.5
# java version: 8
# picard version: 2.10.10 snapshot
# bwa options:
# index:
# -a BWT construction algorithm bwtsw
# -p index prefix
# mem:
# -t number of computing threads
# -R readgroup
# samtools options
# -@ number of computing threads
# view:
# -b b for bam
# -T reference sequence FASTA file
# sort:
```

```
# -m maximum memory per thread
# -O output format BAM
# -o output name
# picard options:
# VALIDATION_STRINGENCY
# REMOVE_DUPLICATES
# MAX_RECORDS_IN_RAM

# before aligning to it build index for reference genome once.
bwa index -a bwtsw -p $reference Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz

# alignment, saving to bam format and sorting bam file
bwa mem -t $cpu -R $RG $reference $sequence1 $sequence2 | samtools view -@ $cpu
-bT $reference.fa - | samtools sort -m 20G -@ $cpu -O BAM -o $bam

# index bam file
samtools index -@ $cpu $bam

# mark duplicate reads
$mdbam=$(ls $bam | grep trim.bam |  sed "s/_trim.bam/_trim_md.bam/")
java -XX:+UseParallelOldGC -XX:ParallelGCThreads=4 -jar picard MarkDuplicates I=$out
    O=$mdbam M=$bam.metrics.txt VALIDATION_STRINGENCY=LENIENT CREATE_INDEX=true
    REMOVE_DUPLICATES=false VERBOSITY=ERROR MAX_RECORDS_IN_RAM=50000000

# merge technical replicates
samtools merge $bam $bam1 $bam2 -@ 14

# new index
ls *bam | parallel -j 16 "samtools index {}"

# InDel realignment with GATK
# create targets for InDel realignment
ls *.bam > files.list
java -Xms8g -Xmx24g -jar GATK -T RealignerTargetCreator -R $reference.fa -I files.list
     -o ir.intervals -S LENIENT -L $regions -nt 10 -U ALLOW_N_CIGAR_READS

# InDel realignment
java -Xmx12g -jar GATK -T IndelRealigner -I $bam -R $reference.fa
    --targetIntervals ir.intervals -S LENIENT --maxReadsForRealignment 50000
    -U ALLOW_N_CIGAR_READS -rf NotPrimaryAlignment -rf DuplicateRead
    -rf UnmappedRead -log $(echo $bam | sed "s/1_trim_md.bam/realignment.log/")
    -o $(echo $bam | sed "s/1_trim_md/realigned/")
```

Code B.3: Alignment

```
# ------------------------------------------------------------------------ #
# Quality measurements                                                     #
# ------------------------------------------------------------------------ #
```

```
#!/bin/bash
set -e

alignment=$1
name=$(echo $alignment | cut -d "_" -f 1 | cut -d"/" -f 6)

# samtools idxstats and samtools stats
# samtools version: 1.5 using htslib 1.5

samtoolsf=$reportdir/samtools

samtools idxstats $alignment>$samtoolsf/$name-idxstats.txt

samtools stats $alignment>$samtoolsf/$name-stats.txt



# Alfred version: 0.1.13

# Reference genome GRCh38 download link:
# ftp://ftp.ensembl.org/pub/release-77/fasta/homo_sapiens/dna/
    Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz
alfred=../alfred/src/alfred
reference=../References/hg38
rbed=../alfred/maps/exonic.hg38.bed.gz
alfredf=$reportdir/alfred

$alfred qc -r $reference.fa -b $rbed -o $alfredf/$name.qc.tsv.gz $alignment
```

Code B.4: Quality measurements alignment

```
# ------------------------------------------------------------------------ #
# Variant Calling with VarScan2                                            #
# ------------------------------------------------------------------------ #

# Variant Calling with VarScan2
# samtools version: 1.5 using htslib 1.5
# Java version: 8
# VarScan2 version: 2.4.3
#
# samtools options:
# -C adjust mapping quality; recommended: 50
# -B disable BAQ (per-Base Alignment Quality)
# -q skip alignments with mapping quality smaller than
# -Q skip alignments with base quality smaller than
# -f fasta indexed reference sequence file
# VarScan2 options:
# --min-coverage Minimum coverage in normal and tumor to call variant
# --min-coverage-normal Minimum coverage in normal to call somatic
# --min-coverage-tumor Minimum coverage in tumor to call somatic
```

```
# -mpileup Input is samtools mpileup
# --indel-file File of indes for filtering nearby SNVs
# --min-reads2 Minimum supporting reads for a variant
# --min-var-freq Minimum variant allele frequency threshold
# --p-value Default p-value threshold for calling variants

samtools mpileup \
-C50 -B -q 10 -Q 13 -f $reference \
$normalbam $tumorbam>$pileup

java -jar VarScan2 somatic \
$pileup \
$patient \
--min-coverage 5 \
--min-coverage-normal 4 \
--min-coverage-tumor 2 \
--mpileup 1

java -jar VarScan2 somaticFilter \
$patient.snp -indel-file $patient.indel \
-output-file $patient.snp.filtered \
--min-coverage 5 \
--min-reads2 2 \
--min-var-freq 0.1 \
--p-value 5e-02
```

Code B.5: Variant call with VarScan2

```
# ------------------------------------------------------------------------ #
# Variant Calling with Seurat                                              #
# ------------------------------------------------------------------------ #

# Variant Calling with Seurat
# Seurat version: 2.5
# Java version: 6
#
# Seurat options:
# --indels enables somatic indel calling
# --structvar enables structural variant detection
# -Q Minimum phred scale for reported events
# -mmq Minimum mapping quality for reads to be considered in the pileup
# -mbq Minimum base quality required to consider a base for calling
# -mcv Minimum per-sample coverage required to attempt a call at a locus
# -L  restrict variant calling to targeted regions by a provided list

java -jar Seurat -T Seurat -R $reference -I:dna_normal $normalbam
    -I:dna_tumor $tumorbam -o $patient.seurat_somatic_variants.vcf
                          -go $patient.seurat_large_events.txt --indels --structvar -Q
    10 -mmq 10             -mbq 13 -mcv 5 -L $regions -log Seurat_log_vc_$patient
```

Code B.6: Variant call with Seurat

```
# ------------------------------------------------------------------------ #
# Variant Calling with Strelka                                             #
# ------------------------------------------------------------------------ #

# Variant Calling with Strelka and Manta
# Strelka version: 2.9.6
# Manta version: 1.4.0
#
# Manta options:
# --exome approciate settings for WES; disables high depth filters
# --callRegions restrict variant calling to targeted regions by a provided BED file
# Strelka options:
# --indelCandidates Candidate indel list by Manta
# -m execution on a single node with local
# -j parallelization across multiple number of cores

mantares=./Manta/Manta_$patient
strelkares=./Strelka/Strelka_$patient

mkdir $mantares
mkdir $strelkares

Manta --normalBam $normalbam --tumorBam $tumorbam --referenceFasta $reference  --exome
      --callRegions $regions --runDir $mantares

#$mantares/runWorkflow.py -h # help
$mantares/runWorkflow.py -m local -j $cpu


Strelka --normalBam $normalbam --tumorBam $tumorbam --referenceFasta $reference
     --exome --callRegions $regions
     --indelCandidates $mantares/results/variants/candidateSmallIndels.vcf.gz
     --runDir $strelkares

#$strelkares/runWorkflow.py -h # help
$strelkares/runWorkflow.py -m local -j $cpu
```

Code B.7: Variant call with Strelka

```
# ------------------------------------------------------------------------ #
# R Code                                                                   #
# Merge VCF files                                                          #
# Create BED file for read counting                                        #
# ------------------------------------------------------------------------ #
args <- commandArgs(TRUE)
p = args[1]
```

```
library(vcfR)
library(plyr)

# Remove unused factor levels
subdf <- function(df) {
  i.row.names <- row.names(df)
  df2 <- as.data.frame(
    lapply(df,
           function(x) if(is.factor(x)) factor(x) else x
    )
  )
  row.names(df2) <- i.row.names
  df2
}

  # Read in VCFs
  VS.snv <- read.vcfR(paste0("/Data/VariantCalls/VarScan/", p, ".snp.vcf"))
  VS.ind <- read.vcfR(paste0("/Data/VariantCalls/VarScan/", p, ".indel.vcf"))
  St.snv <- read.vcfR(paste0("/Data/VariantCalls/Strelka/Strelka_", p, "/results/
      variants/somatic.snvs.vcf.gz"))
  St.ind <- read.vcfR(paste0("/Data/VariantCalls/Strelka/Strelka_", p, "/results/
      variants/somatic.indels.vcf.gz"))
  Se.all <- read.vcfR(paste0("/Data/VariantCalls/Seurat/", p, ".seurat_somatic_
      variants.vcf"))

  # Filter and subset VCFs
  ## SNVs, somatic, passed filter, at CHR 1:22,X,Y
  ### VarScan
  VS.snv.red <- data.frame(cbind(VS.snv@fix, VS.snv@gt))
  # somatic variants
  VS.snv.red <- VS.snv.red[grep("SS=2", VS.snv.red[,"INFO"]),]
  # only which passed filter
  VS.snv.red <- VS.snv.red[grep("PASS", VS.snv.red[,"FILTER"]),]
  # at CHR 1:22, X, Y
  VS.snv.red <- VS.snv.red[which(VS.snv.red[,"CHROM"]%in%c(1:22,"X", "Y")),]
  VS.snv.red$Caller <- "VarScan2"
  ### Strelka
  St.snv.red <- data.frame(cbind(St.snv@fix, St.snv@gt))
  # all somatic variants
  St.snv.red <- St.snv.red[grep("SOMATIC", St.snv.red[,"INFO"]),]
  # only which passed filter
  St.snv.red <- St.snv.red[grep("PASS", St.snv.red[,"FILTER"]),]
  # at CHR 1:22, X, Y
  St.snv.red <- St.snv.red[which(St.snv.red[,"CHROM"]%in%c(1:22,"X", "Y")),]
  St.snv.red$Caller <- "Strelka"
  ### Seurat
  # all somatic variants
  Se.snv.red <- Se.all@fix[grep("somatic_SNV", getINFO(Se.all)),]
```

```
# only which passed filter
Se.snv.red <- Se.snv.red[grep("PASS", Se.snv.red[,"FILTER"]),]
# at CHR 1:22, X, Y
Se.snv.red <- Se.snv.red[which(Se.snv.red[,"CHROM"]%in%c(1:22,"X", "Y")),]
Se.snv.red <- data.frame(Se.snv.red)
Se.snv.red[,c("FORMAT", "NORMAL", "TUMOR")] <- NA
Se.snv.red$Caller <- "Seurat"

SNV <- rbind(VS.snv.red, St.snv.red, Se.snv.red)
SNV <- subdf(SNV)
SNV$POS <- as.numeric(as.character(SNV$POS))

# Coordinates creation Ensembl vep
SNV$InDType <- "sub"
SNV$REF2 <- as.character(SNV$REF)
SNV$ALT2 <- as.character(SNV$ALT)
SNV$Start <- SNV$POS
SNV$End <- SNV$POS
SNV$Alleles <- paste(SNV$REF,SNV$ALT,sep = "/")
SNV$Strand <- "*"

# Coordinates creation bam-readcount
SNV$REF3 <- as.character(SNV$REF)
SNV$ALT3 <- as.character(SNV$ALT)
SNV$POSRC <- SNV$POS

# Index creation
SNV$Position <- paste(SNV$CHROM,SNV$POSRC,sep = ":")
SNV$Index <- paste(p, SNV$Position, paste0(SNV$REF3,"/", SNV$ALT3), sep = ".")

# sort SNV
SNV$Chrom <- as.numeric(mapvalues(SNV$CHROM, c("X", "Y"), 23:24))
SNV <- SNV[order(SNV$POS),]
SNV <- SNV[order(SNV$Chrom),]

# Variant caller column
SNV$VarScan <- ifelse(SNV$Caller=="VarScan2", T, F)
SNV$Strelka <- ifelse(SNV$Caller=="Strelka", T, F)
SNV$Seurat <- ifelse(SNV$Caller=="Seurat", T, F)

# Mark variants found by more than one caller
snv.split <- split(SNV[, "Caller"], SNV$Index)
snv.split <- lapply(snv.split, function(x) paste(x, collapse = ", "))
SNV$Caller2 <- NA
SNV$Caller2 <- sapply(SNV$Index, function(idx) x <- unlist(snv.split[[idx]]))
SNV$Caller2 <- unlist(SNV$Caller2)
SNV$VarScan <- ifelse(grepl("VarScan", SNV$Caller2),T, F)
SNV$Strelka <- ifelse(grepl("Strelka", SNV$Caller2),T, F)
SNV$Seurat  <- ifelse(grepl("Seurat", SNV$Caller2),T, F)
```

```
SNV$CallerCount <- ifelse(SNV$Caller2 == "Strelka, Seurat" | SNV$Caller2 == "VarScan2, Seurat" | SNV$Caller2 == "VarScan2, Strelka", 2, ifelse(SNV$Caller2 == "VarScan2, Strelka, Seurat", 3, 1))

# Remove all variants found by more than one caller exept the first entry
SNV.red <- SNV[-which(duplicated(SNV$Index)),]

## InDel, somatic, passed filter, at CHR 1:22,X,Y
VS.ind.red <- data.frame(cbind(VS.ind@fix, VS.ind@gt))
VS.ind.red <- VS.ind.red[grep("SS=2", VS.ind.red[,"INFO"]),]
VS.ind.red <- VS.ind.red[grep("PASS", VS.ind.red[,"FILTER"]),]
VS.ind.red <- VS.ind.red[which(VS.ind.red[,"CHROM"]%in%c(1:22,"X", "Y")),]
VS.ind.red$Caller <- "VarScan2"
### Strelka
St.ind.red <- data.frame(cbind(St.ind@fix, St.ind@gt))
St.ind.red <- St.ind.red[grep("SOMATIC", St.ind.red[,"INFO"]),]
St.ind.red <- St.ind.red[grep("PASS", St.ind.red[,"FILTER"]),]
St.ind.red <- St.ind.red[which(St.ind.red[,"CHROM"]%in%c(1:22,"X", "Y")),]
St.ind.red$Caller <- "Strelka"
### Seurat
Se.ind.red <- Se.all@fix[grep("somatic_insertion|somatic_deletion", getINFO(Se.all)),]
Se.ind.red <- Se.ind.red[grep("PASS", Se.ind.red[,"FILTER"]),]
Se.ind.red <- Se.ind.red[which(Se.ind.red[,"CHROM"]%in%c(1:22,"X", "Y")),]
Se.ind.red <- data.frame(Se.ind.red)
Se.ind.red[,c("FORMAT", "NORMAL", "TUMOR")] <- NA
Se.ind.red$Caller <- "Seurat"

IND <- rbind(VS.ind.red, St.ind.red, Se.ind.red)
IND <- subdf(IND)
IND$POS <- as.numeric(as.character(IND$POS))
IND$REF <- as.character(IND$REF)
IND$ALT <- as.character(IND$ALT)

# Coordinates creation Ensembl vep
IND$InDType <- ifelse(nchar(IND$REF)<nchar(IND$ALT), "ins", "del")
IND$REF2 <- as.character(IND$REF)
IND$ALT2 <- as.character(IND$ALT)
IND$REF2[which(IND$InDType=="del")] <- substr(IND$REF[which(IND$InDType=="del")], 2, nchar(IND$REF[which(IND$InDType=="del")]))
IND$ALT2[which(IND$InDType=="ins")] <- substr(IND$ALT[which(IND$InDType=="ins")], 2, nchar(IND$ALT[which(IND$InDType=="ins")]))
IND$Start <- IND$POS +1 # see: https://www.ensembl.org/info/docs/tools/vep/vep_formats.html#input
IND$End <- ifelse(IND$InDType=="ins", IND$POS, IND$POS + (nchar(IND$REF)-1)) # - nchar(IND$ALT), IND$Start + (nchar(IND$REF)-1))
IND$Alleles <- ifelse(IND$InDType=="ins", paste("-",IND$ALT2,sep = "/"), paste(IND$REF2,"-",sep = "/"))
IND$Strand <- "*"
```

```
146 # Coordinates creation bam-readcount
147 IND$REF3 <- ifelse(IND$InDType=="ins", IND$REF, substr(IND$REF,2,2))
148 IND$ALT3 <- ifelse(IND$InDType=="ins", paste0("+",substr(IND$ALT, 2, nchar(IND$ALT))
        ), paste0("-", substr(IND$REF, 2, nchar(IND$REF))))
149 IND$POSRC <- ifelse(IND$InDType=="del", IND$POS+1, IND$POS)

151 # Index creation
152 IND$Position <- paste(IND$CHROM,IND$POSRC,sep = ":")
153 IND$Index <- paste(p, IND$Position, paste0(IND$REF3,"/", IND$ALT3), sep = ".")

155 # sort IND
156 IND$Chrom <- as.numeric(mapvalues(IND$CHROM, c("X", "Y"), 23:24))
157 IND <- IND[order(IND$POS),]
158 IND <- IND[order(IND$Chrom),]

160 # Variant caller Column
161 IND$VarScan <- ifelse(IND$Caller=="VarScan2", T, F)
162 IND$Strelka <- ifelse(IND$Caller=="Strelka", T, F)
163 IND$Seurat <- ifelse(IND$Caller=="Seurat", T, F)

165 # Mark variants found by more than one caller
166 ind.split <- split(IND[, "Caller"], IND$Index)
167 ind.split <- lapply(ind.split, function(x) paste(x, collapse = ", "))
168 IND$Caller2 <- NA
169 IND$Caller2 <- sapply(IND$Index, function(idx) x <- unlist(ind.split[[idx]]))
170 IND$Caller2 <- unlist(IND$Caller2)
171 IND$VarScan <- ifelse(grepl("VarScan", IND$Caller2),T, F)
172 IND$Strelka <- ifelse(grepl("Strelka", IND$Caller2),T, F)
173 IND$Seurat <- ifelse(grepl("Seurat", IND$Caller2),T, F)
174 IND$CallerCount <- ifelse(IND$Caller2 == "Strelka, Seurat" | IND$Caller2 == "
        VarScan2, Seurat" | IND$Caller2 == "VarScan2, Strelka", 2, ifelse(IND$Caller2 ==
         "VarScan2, Strelka, Seurat", 3, 1))
175 # Remove all variants found by more than one caller exept the first entry
176 IND.red <- IND[-which(duplicated(IND$Index)),]

178 # SNVs and InDels
179 allvars <- rbind(SNV.red,IND.red)
180 # use format() else R creates sometimes scientific numbers
181 # (e.g. 8000000 turns to 8e+06) -> problems with bam-readcount
182 allvars$POSRC <- format(allvars$POSRC, scientific = FALSE, trim = TRUE)

184 # write bed for bam-readcount ... for final Variant Allele Frequency
185 # 1-based
186 write.table(allvars[,c("CHROM", "POSRC", "POSRC")],
187             sep = "\t", quote = FALSE, row.names = FALSE, col.names = FALSE,
188             file = paste0("/Data/VariantCalls/bamreadcount_bed_input/", p, ".bed"))

190 # write table input for VarScan2 fpfilter
```

```
  # 1-based
  write.table(allvars[,c("CHROM", "POSRC", "REF3", "ALT3")],
              sep = "\t", quote = FALSE, row.names = FALSE, col.names = c("chrom",
                      "position", "ref", "var"),
              file = paste0("Data/VariantCalls/bamreadcount_bed_input/", p,
                            ".fpfilter"))

  # write table per patient for merging with readcounts and filter
  write.csv2(allvars, file = paste0("Data/VariantCalls/merged/", p, ".variants.csv"))
```

Code B.8: Merge VCF files

```
# ------------------------------------------------------------------------ #
# Counting Reads with bam-readcount                                        #
# Filter Variants with VarScan2 fpfilter                                   #
# ------------------------------------------------------------------------ #

# Counting reads with bam-readcount
# bam-readcount version: 0.8.0
# Filter variants with VarScan2 fpfilter
# Java version: 8
# VarScan2 version: 2.4.3
#
# bam-readcount options:
# -f Reference FASTA file
# -l BED file with variant positions
# -q Minimum read mapping quality
# -b Minimum base mapping quality
# -w Maximum number of warnings
# VarScan2 fpfilter options:
# --keep-failures writes all variants from input BED to output table
# --min-var-freq minimum variant allele frequency estimated from new read counts
# --min-var-count minimum number of reads supporting a variant



# readcounts DNA
bam-readcount -f $reference -l $bed -q 10 -b 13 -w 0 $tumorbam > $readcounts.DNA

# readcounts RNA
bam-readcount -f $reference -l $bed -q 1 -b 1 -w 0 $RNAtumorbam > $readcounts.RNA

java -jar VarScan2 fpfilter $fpfilter $readcounts.DNA --output-file $final.variants
    --keep-failures --min-var-freq 0.1

java -jar VarScan2 fpfilter $fpfilter $readcounts.RNA --output-file
    $final.variants.rna --keep-failures --min-var-freq 0 --min-var-count 2
```

Code B.9: Count reads and filter variants

```r
# ------------------------------------------------------------------------ #
# R Code                                                                   #
# Merge counts and filter results                                          #
# Create input table for Ensembl vep                                       #
# ------------------------------------------------------------------------ #

args <- commandArgs(TRUE)
p = args[1]

cn <- c("chrom", "position", "ref", "var", "ref_reads", "var_reads", "ref_strand", "
    var_strand", "ref_basequal", "var_basequal", "ref_readpos", "var_readpos", "ref_
    dist3", "var_dist3", "ref_mapqual", "var_mapqual", "mapqual_diff", "ref_mmqs", "
    var_mmqs", "mmqs_diff", "ref_avg_rl", "var_avg_rl", "avg_rl_diff", "filter_status"
    , "X")
filtered.VS <- read.table(paste0("Data/VariantCalls/filtered/final.", p), sep = "\t",
    header = TRUE, as.is = TRUE, fill = TRUE, col.names = cn)
filtered.VS$Index <- paste0(p, ".", filtered.VS$chrom, ":", filtered.VS$position, ".",
     filtered.VS$ref, "/", filtered.VS$var)
rna.VS <- read.table(paste0("Data/VariantCalls/filtered/final.rna.", p), sep = "\t",
    header = TRUE, as.is = TRUE, fill = TRUE, col.names = cn)
rna.VS$Index <- paste0(p, ".", rna.VS$chrom, ":", rna.VS$position, ".", rna.VS$ref, "/
    ", rna.VS$var)
vars <- read.csv2(paste0("Data/VariantCalls/merged/", p, ".variants.csv"), as.is =
    TRUE)
if(!all(table(vars$Index%in%filtered.VS$Index))){
  stop('Not all variants are in both tables run table(vars$Index%in%filtered.VS$Index)
      ')
  }

# add fpfilter status and readcounts
vars$filterstatus <- filtered.VS$filter_status[match(vars$Index, filtered.VS$Index)]
vars$REF.reads <- filtered.VS$ref_reads[match(vars$Index, filtered.VS$Index)]
vars$ALT.reads <- filtered.VS$var_reads[match(vars$Index, filtered.VS$Index)]
vars$REF.reads.rna <- rna.VS$ref_reads[match(vars$Index, rna.VS$Index)]
vars$ALT.reads.rna <- rna.VS$var_reads[match(vars$Index, rna.VS$Index)]

# calculate VAF and PRESENT status
vars$VAF <- vars$ALT.reads/(vars$REF.reads+vars$ALT.reads) *100
vars$VAF[vars$REF.reads+vars$ALT.reads == 0] <- 0
vars$VAF.RNA <- vars$ALT.reads.rna/(vars$REF.reads.rna+vars$ALT.reads.rna) *100
vars$VAF.RNA[vars$REF.reads.rna+vars$ALT.reads.rna == 0] <- 0
vars$DNA.present <- ifelse(vars$VAF>=0.1 & vars$ALT.reads > 1, T, F)
vars$RNA.present <- ifelse(vars$ALT.reads.rna > 1, "Expressed", ifelse(vars$REF.reads.
    rna > 0 & vars$ALT.reads.rna == 0, "Variant not expressed", "Gene not expressed"))

# only variants which passed filters and are called by more than one caller
vars.red <- vars[which(vars$filterstatus=="PASS"),]
vars.red <- vars.red[which(vars.red$CallerCount>1),]
```

```
# write table for Annotation with Ensembl vep
# 1-based
# SNV Startpos == Endpos e.g. 12600 12600 C/A * Index
# Insertion Startpos > Endpos e.g. 12601 12600 -/ACC * Index (Startpos = Endpos + 1)
# Deletion Startpos < Endpos e.g. 12600 12602 CGT/- * Index

write.table(vars.red[,c("CHROM", "Start", "End", "Alleles", "Strand", "Index")],
            sep = " ", quote = FALSE, row.names = FALSE, col.names = FALSE,
            file = paste0("VariantCalls/vep_files/", p,"_SNVs.VEP.txt"))

# save variant table in R workspace
save(vars.red, vars, file = paste0("VariantCalls/variants_raw/", p, "_raw.variants.
    RData"))
```

Code B.10: Merge counts and filter results of variants

```
# ------------------------------------------------------------------------ #
# Annotation with Ensembl vep                                              #
# ------------------------------------------------------------------------ #

# Annotation with Ensembl variant effect predictor
# Human Genome Annotations Ensembl version 94
# vep version: 0.19.4
#
# options:
# -i input variants
# -o annotated variants
# --fasta reference FASTA file
# --symbol HGNC gene symbol
# --fork fork input for parallel computation
# --cache use cached reference files for faster and offline computations
# --tab output in table format
# --sift SIFT predictions and/or score
# --polyphen PolyPHEN predictions and/or score
# --check_existing check existence of known variants that are co-located
#   with the input
# --af global allelic frequency from 1000 Genomes Phase 2 data for any
#   co-located variant
# --max_af highest allele frequency observed in any population from
#   1000 Genomes, ESP or gnomAD
# -v verbose
# --offline compute annotations offline

./vep -i $input -o $output --fasta $reference.fa --symbol --fork 12 --cache --tab
                --sift b --polyphen b --check_existing --af --max_af -v --offline
```

Code B.11: Annotate variants

```
# ------------------------------------------------------------------------ #
# R Code                                                                   #
```

```
# Combine annotation and variant table                                    #
# ------------------------------------------------------------------------ #

args <- commandArgs(TRUE)
p = args[1]

library(stringr)

# load variant table
load(paste0("VariantCalls/variants_raw/",p,"_raw.variants.RData"))

# load Ensembl vep results
vep <- read.table(paste0("/VariantCalls/vep_files/", p, "SNVs.VEP.out"), as.is = TRUE)

colnames(vep) <- c("Index", "Location", "Allele",  "Gene", "Feature",  "Feature_type"
    , "Consequence", "cDNA_position", "CDS_position", "Protein_position", "Amino_
    acids",  "Codons",  "Existing_variation", "IMPACT", "DISTANCE", "STRAND",  "
    FLAGS",  "SYMBOL",  "SYMBOL_SOURCE",  "HGNC_ID", "GIVEN_REF", "USED_REF", "
    BAM_EDIT", "SIFT", "PolyPhen", "AF",  "MAX_AF",  "MAX_AF_POPS",  "CLIN_SIG", "
    SOMATIC",  "PHENO")

vepFull <- vep
allvars <- vars.red
# vep <- vepFull

# Variant consequences ordered by severity as estimated by Sequence Ontology
consorder <- c("splice_acceptor_variant", "splice_donor_variant", "stop_gained", "
    frameshift_variant", "stop_lost", "start_lost", "inframe_insertion", "inframe_
    deletion", "missense_variant", "splice_region_variant", "start_retained_variant",
    "stop_retained_variant", "synonymous_variant")

# Remove unwanted consequences
vep <- subset(vep, SYMBOL_SOURCE == "HGNC")
vep <- subset(vep, !(Consequence %in% c("upstream_gene_variant", "downstream_gene_
    variant")))
vep <- vep[grep("synon|frame|stop|miss|start|splice", vep$Consequence),]
vep$Consequence2 <-  unlist(lapply(str_split(vep$Consequence, ","), function(x) x[1]))
vep$Consequence2 <- factor(vep$Consequence2, levels = consorder)

# Sort consequence in order of severity
vep <- vep[order(vep$Consequence2),]

# Split by variant index
vepl <- split(vep, vep$Index)
vepFullConswanted <- lapply(split(vep$Consequence2, vep$Index), function (x) paste(
    unique(unlist(str_split(x, ","))), collapse = ","))
vepFullConsall <- lapply(split(vep$Consequence, vep$Index), function (x) paste(unique(
    unlist(str_split(x, ","))), collapse = ","))
```

```
# reduced vep table ... one entry per variant
vep2 <- do.call("rbind", lapply(vepl, function(x) x <- x[1,]))
vep2$ConsequenceFullwanted <- unlist(vepFullConswanted[vep2$Index])
vep2$ConsequenceFullall <- unlist(vepFullConsall[vep2$Index])

# Annotation to variant table
allvars[,c("Symbol", "Consequence", "ConsequenceFullwanted", "ConsequenceFullall", "
    Impact", "Existing_variation", "Codons", "Amino_acids", "Protein_position", "
    EnsemblGeneID", "MAX_AF", "MAX_AF_POP", "CLIN_SIG")] <- vep2[match(allvars$Index,
    vep2$Index), c("SYMBOL", "Consequence2", "ConsequenceFullwanted", "
    ConsequenceFullall", "IMPACT", "Existing_variation", "Codons", "Amino_acids", "
    Protein_position", "Gene", "MAX_AF", "MAX_AF_POPS", "CLIN_SIG")]

allvars$PolyPhen <- str_remove_all(vep2$PolyPhen, "\\(|\\)|[0-9]|\\.")[match(allvars$
    Index, vep2$Index)]
allvars$SIFT <- str_remove_all(vep2$SIFT, "\\(|\\)|[0-9]|\\.")[match(allvars$Index,
    vep2$Index)]
allvars$isCoding <- ifelse(allvars$Codons!="-", TRUE, FALSE)
allvars$isSynonymous <- ifelse(grepl("synon",allvars$Consequence), TRUE, FALSE)
allvars$inPopulation <- ifelse(allvars$MAX_AF_POP!="-", TRUE, FALSE)
allvars$Protein_Change <- ifelse(allvars$Consequence=="missense_variant", paste0("p.",
     substr(allvars$Amino_acids,1,1), allvars$Protein_position, substr(allvars$Amino_
    acids,3,3)), NA)

allvars$Probe <- substr(allvars$Index, 1, 5)

# Add expression values from cpm and rpkm normalization
(load(".../RS_norm.all_G180726_181022.Rdata"))
(load(".../RS_rpkm.norm.all_G180726_190207.Rdata"))
nc <- log2(nclist$nc[, which(colnames(nclist$nc) %in% unique(allvars$Probe))]+1)
tp <- log2(rpkm[, which(colnames(rpkm) %in% unique(allvars$Probe))]+1)
allvars$Expression <- NA
allvars$Expression.rpkm <- NA
for (i in 1:nrow(allvars)){
  allvars$Expression[i] <- ifelse(!any(rownames(nc) %in% allvars$EnsemblGeneID[i]), NA
      , nc[which(rownames(nc) %in% allvars$EnsemblGeneID[i]),which(colnames(nc) %in%
      allvars$Probe[i])])
  allvars$Expression.rpkm[i] <- ifelse(!any(rownames(tp) %in% allvars$EnsemblGeneID[i
      ]), NA, tp[which(rownames(tp) %in% allvars$EnsemblGeneID[i]),which(colnames(tp)
      %in% allvars$Probe[i])])
}
allvars$Expression[which(allvars$Expression<1)] <- NA
allvars$Expression.rpkm[which(allvars$Expression.rpkm<=0)] <- NA

# Add mapping with CoMMpass data
CoMMpass <- read.table("/Data/COMMPASS/MMRF_CoMMpass_IA10c_All_Canonical_NS_Variants_
    lfthg38.txt", header = TRUE, as.is = TRUE)
allvars$GeneinCoMMpass <- allvars$Symbol %in% CoMMpass$ANN....GENE
allvars$VarinCoMMpass <- allvars$Existing_variation %in% CoMMpass$ID
```

```
77 # Subset
78 variants <- allvars[which(allvars$isCoding & !allvars$isSynonymous & allvars$
       CallerCount > 1),]
79 variants <- variants[order(variants$POS),]
80 variants <- variants[order(variants$Chrom),]
81 variants <- variants[order(variants$Probe),]

83 # Save
84 assign(paste0(p, "_variants"), variants)
85 assign(paste0(p, "_allvars"), allvars)
86 assign(paste0(p, "_vep"), vep)
87 assign(paste0(p, "_vep2"), vep2)
88 assign(paste0(p, "_vepFull"), vepFull)
89 save(list = ls(pattern = paste0(p, "_")), file =  paste0("VariantCalls/VariantTables/"
       , p, "_all_variants.processed.RDa"))
```

Code B.12: Create final table of variants per patient

```
1  # ------------------------------------------------------------------------ #
2  # R Code                                                                   #
3  # Merge variant tables                                                     #
4  # ------------------------------------------------------------------------#

6  patients <- substr(list.files("Data/VariantCalls/Strelka/"), 9, 13)

8  allvars <- NULL
9  variants <- NULL
10 vep <- NULL
11 vepFull <- NULL

13 for (p in patients){
14   load(paste0("VariantCalls/VariantTables/", p, "_all_variants.processed.RDa"))
15   allvars <- rbind(allvars, get(ls(pattern = paste0(p, "_allvars"))))
16   variants <- rbind(variants, get(ls(pattern = paste0(p, "_variants"))))
17   vep <- rbind(vep, get(ls(pattern = paste0(p, "_vep"))))
18   vepFull <-  rbind(vepFull, get(ls(pattern = paste0(p, "_vepFull"))))
19 }

21 save(allvars, variants, vep, vepFull,
22      file = "VariantCalls/all_variants_processed.RData")
```

Code B.13: Merge final variant tables

```
1  # ------------------------------------------------------------------------ #
2  # Copy number Calling with VarScan2                                        #
3  # ------------------------------------------------------------------------ #

5  # Copy number Calling with VarScan2
6  # samtools version: 1.5 using htslib 1.5
```

```
# Java version: 8
# VarScan2 version: 2.4.3
#
# samtools options:
# -C adjust mapping quality; recommended: 50
# -B disable BAQ (per-Base Alignment Quality)
# -q skip alignments with mapping quality smaller than
# -Q skip alignments with base quality smaller than
# -f fasta indexed reference sequence file
# VarScan2 options:
# -mpileup Input is samtools mpileup
# --min-coverage Minimum coverage in normal and tumor to call variant
# --min-segment-size Minimum number of bases in a copy number altered segment
# --max-segment-size Maximum number of bases in a copy number altered segment



samtools mpileup \
-C50 -B -q 10 -Q 13 -f $reference \
$normalbam $tumorbam>$pileup

java -jar VarScan2 copynumber \
$pileup $cnvdir/$patient \
-mpileup 1 \
--min-coverage 20 \
--min-segment-size 100 \
--max-segment-size 1000

java -jar VarScan2 copyCaller \
$cnvdir/$patient.copynumber \
--output-file $cnvdir/$patient.cnv.called
```

Code B.14: Copy number call with VarScan2

```
# ------------------------------------------------------------------------ #
# R Code                                                                   #
# Copy number data analysis                                                #
# ------------------------------------------------------------------------#

library(DNAcopy)
# read in copy number calling files created with VarScan2
# for every patient a copy number calling file is saved to $pathtofile
cn <- read.table($pathtofile, as.is=T, header=T)
cn$adjlr <- as.numeric(gsub(",", ".", cn$adjusted_log_ratio))
cn2 <- cn[which(cn$chrom%in%1:22),]
cn2 <- cn2[order(cn2$chrom),]
cn2$adjlr <- cn2$adjlr-mean(cn2$adjlr, na.rm=T)
CNA.subset <- CNA(genomdat = cn2$adjlr,
                  chrom = cn2$chrom,
                  maploc = cn2$chr_start,
```

```
                    data.type = 'logratio',
                    sampleid = sampleid)
CNA.smoothed <- smooth.CNA(CNA.subset)
segs <- segment(CNA.smoothed, verbose=0, min.width=2)
save($segs, file = "segs$patient.RData"))
segsAL <- list()
segsAL <- c(segsAL, list($segs))

# save all segments together in one file for GISTIC2.0
segs.gistic <- NULL
for (i in names(segsAL)){
   segments <- segs$output
   segs.gistic <- rbind(segs.gistic,segments)
}
write.table(segs.gistic, file = "Segments_gistic2",
            sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
```

Code B.15: Copy number segmentation with DNAcopy

```
# ------------------------------------------------------------------------ #
# Copy number analysis with GISTIC 2.0                                     #
# ------------------------------------------------------------------------ #

# Copy number analysis with GISTIC 2.0
# GISTIC version: 2.0.23
# Matlab Compiler Runtime (MCR) Enviroment version: 8
#
# GISTIC options:
# -b directory for results
# -seg input segment data
# -refgene gistic2 reference genome
# -genegistic 1 = use the gene GISTIC algorithm
# -smallmem 1 = compress memory
# -broad additional broad-level analysis should be performed 1:yes
# -brlen broad from focal events in units of fraction of chromosome arm
# -conf confidence level for calculating region containing a driver
# -armpeel Flag set to enable arm-level peel-off of events during peak definition.
# -savegene 1 = save segmented input data
# -gcm Method for reducing marker-level copy number data to the gene-level copy number
     data

# output directory
echo --- creating output directory ---
basedir=~/Gistic
mkdir -p $basedir

# input file definitions
segfile=Segments_gistic2
refgenefile=$gisticdir/refgenefiles/hg38.UCSC.add_miR.160920.refgene.mat
```

```
cd $gisticdir
# call script that sets MCR environment and calls GISTIC executable
./gistic2 -b $basedir -seg $segfile -refgene $refgenefile \
-genegistic 1 -smallmem 1 \
-broad 1 -brlen 0.5 \
-conf 0.90 -armpeel 1 \
-savegene 1 -gcm extreme
```

Code B.16: Copy number alterations with GISTIC2

Acknowledgements

Here I want to express my gratitude to all who supported me during writing this thesis. First of all I would like to thank PD Dr. Dr. Dirk Hose for the supervision and for the permanent support while writing this thesis. Also I would like to thank him and his deputy head PD Dr. Anja Seckinger for offering me the opportunity to work in their group the Multiple Myeloma Research Laboratory and for the conception of this thesis.

Special thanks to Martina Emde, for contributions to RNA sequencing analysis and HDAL score creation and also for all the time she spent on the discussions with me, the critical reading of my thesis while writing her own and for being the best colleague anyone could imagine.

Many thanks to all technicians who worked with me in the past years in the Multiple Myeloma Research Laboratory, Maria Dörner, Birgit Schneiders, Tomi Bähr-Ivacevic, Ewelina Nickel, Malu Brygider and Rike Seidt for the many performed plasma cell purifications and subsequent wet lab analyses, which made the study in this thesis possible.

I am also grateful to all who read this thesis for useful comments and remarks.

Finally, I would like to thank my husband, my family and friends for the encouragement and patience.

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema

   "**Molecular pathogenesis and prognosis of light chain amyloidosis**"

   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