Entwicklung eines Verfahrens zur fehlwerttoleranten Batcheffektkorrektur zwischen unabhängig generierten Proteomdatensätzen

Dissertation zum Erlangen der Würde des

Doktors der Naturwissenschaften
(Dr.rer. nat)

An der Fakultät für Mathematik, Informatik und Naturwisenschaften,
Fachbereich Chemie,
Universität Hamburg

Vorgelegt von

Hannah Voß

Aus Haltern am See

November 2022

1. Gutachter: Prof. Dr. rer. nat. Hartmut Schlüter
2. Gutachter: Priv. Dr. rer. nat. habil. Markus Perbandt

Datum der mündlichen Prüfung: 23.06.2023

Vorsitzender der Prüfungskomission: Prof. Dr. rer. nat. Hartmut Schlüter
2. Prüfer: Prof. Dr. rer. nat. Markus Fischer
3. Prüfung: Prof. Dr. rer. nat.Christian Betzel

Die Forschung für diese Arbeit wurde von Oktober 2019 bis November 2022 in der Arbeitsgruppe Massenspektrometrische Proteomik von Prof. Dr. Hartmut Schlüter am Universitätsklinikum Hamburg-Eppendorf durchgeführt.

Publikationsliste

Voß H , Wurlitzer M, Smit DJ, Ewald F, Alawi M, Spohn M, Indenbirken D, Omidi M, David K, Juhl H, Simon R, Sauter G, Fischer L, Izbicki JR, Molloy MP, Nashan B, Schlüter H, Jücker M. Differential regulation of extracellular matrix proteins in three recurrent liver metastases of a single patient with colorectal cancer. Clin Exp Metastasis. 2020 Dec;37(6):649-656. doi: 10.1007/s10585-020-10058-8. Epub 2020 Oct 24. PMID: 33099724; PMCID: PMC7666585.

Steffen P, Wu J, Hariharan S, **Voß H**, Raghunath V, Molloy MP, Schlüter H. OmixLitMiner: A Bioinformatics Tool for Prioritizing Biological Leads from 'Omics Data Using Literature Retrieval and Data Mining. Int J Mol Sci. 2020 Feb 19;21(4):1374. doi: 10.3390/ijms21041374. PMID: 32092871; PMCID: PMC7073124.

Kement D, Reumann R, Schostak K, **Voß H**, Douceau S, Dottermusch M, Schweizer M, Schlüter H, Vivien D, Glatzel M, Galliciotti G. Neuroserpin Is Strongly Expressed in the Developing and Adult Mouse Neocortex but Its Absence Does Not Perturb Cortical Lamination and Synaptic Proteome. Front Neuroanat. 2021 Feb 23;15:627896. doi: 10.3389/fnana.2021.627896. PMID: 33708076; PMCID: PMC7940840.

Guan Y, Zhang M, Gaikwad M, **Voß H** Fazel R, Ansari S, Shen H, Wang J, Schlüter H. An Integrated Strategy Reveals Complex Glycosylation of Erythropoietin Using Mass Spectrometry. J Proteome Res. 2021 Jul 2;20(7):3654-3663. doi: 10.1021/acs.jproteome.1c00221. Epub 2021 Jun 10. Erratum in: J Proteome Res. 2022 Oct 7;21(10):2552. PMID: 34110173; PMCID: PMC9472269.

Bartkowiak K, Heidrich I, Kwiatkowski M, Banys-Paluchowski M, Andreas A, Wurlitzer M, Geffken M, **Voß H**, Zeller T, Blankenberg S, Peine S, Joosse SA, Müller V, Schlüter H, Oliveira-Ferrer L, Pantel K. Circulating Cellular Communication Network Factor 1 Protein as a Sensitive Liquid Biopsy Marker for Early Detection of Breast Cancer. Clin Chem. 2022 Feb 1;68(2):344-353. doi: 10.1093/clinchem/hvab153. PMID: 34458901.

Hahn J, Moritz M, **Voß H**, Pelczar P, Huber S, Schlüter H. Tissue Sampling and Homogenization in the Sub-Microliter Scale with a Nanosecond Infrared Laser (NIRL) for Mass Spectrometric Proteomics. Int J Mol Sci. 2021 Oct 7;22(19):10833. doi: 10.3390/ijms221910833. PMID: 34639174; PMCID: PMC8509473.

Dottermusch M, Sumisławski P, Krevet J, Middelkamp M, **Voß H**, Siebels B, Bartsch H, Sotlar K, Meyer P, Frank S, Korshunov A, Glatzel M, Schüller U, Neumann JE. Co-activation of Sonic hedgehog and Wnt signaling in murine retinal precursor cells drives ocular lesions with features of intraocular medulloepithelioma. Oncogenesis. 2021 Nov 16;10(11):78. doi: 10.1038/s41389-021-00369-0. PMID: 34785636; PMCID: PMC8595639.

Krösser D, Dreyer B, Siebels B, **Voß H**, Krisp C, Schlüter H. Investigation of the Proteomes of the Truffles: Tuber albidum pico, T. aestivum, T. indicum, T. magnatum and T. melanosporum. Int J Mol Sci. 2021 Nov 30;22(23):12999. doi: 10.3390/ijms222312999. PMID: 34884803; PMCID: PMC8658033.

Voß H, Moritz M, Pelczar P, Gagliani N, Huber S, Nippert V, Schlüter H, Hahn J. Tissue Sampling and Homogenization with NIRL Enables Spatially Resolved Cell Layer Specific Proteomic Analysis of the Murine Intestine. Int J Mol Sci. 2022 May 30;23(11):6132. doi: 10.3390/ijms23116132. PMID: 35682811; PMCID: PMC9181169.

Voß H, Schlumbohm S, Barwikowski P, Wurlitzer M, Dottermusch M, Neumann P, Schlüter H, Neumann JE, Krisp C. HarmonizR enables data harmonization across independent proteomic datasets with appropriate handling of missing values. Nat Commun. 2022 Jun 20;13(1):3523. doi: 10.1038/s41467-022-31007-x. PMID: 35725563; PMCID: PMC9209422.
Teile dieser Arbeit wurden im Rahmen der genannten Publikation veröffentlicht

Nauth T, Bazgir F, **Voß H**, Brandenstein LI, Mosaddeghzadeh N, Rickassel V, Deden S, Gorzelanny C, Schlüter H, Ahmadian MR, Rosenberger G. Cutaneous manifestations in Costello syndrome: HRAS p.Gly12Ser affects RIN1-mediated integrin trafficking in immortalized epidermal keratinocytes. Hum Mol Genet. 2022 Aug 18:ddac188. doi: 10.1093/hmg/ddac188. Epub ahead of print. PMID: 35981076.

Guan Y, Zhang M, Gaikwad M,**Voß H**, Fazel R, Ansari S, Shen H, Wang J, Schlüter H. Correction to An Integrated Strategy Reveals Complex Glycosylation of Erythropoietin Using Mass Spectrometry. J Proteome Res. 2022 Oct 7;21(10):2552. doi: 10.1021/acs.jproteome.2c00509. Epub 2022 Sep 13. Erratum for: J Proteome Res. 2021 Jul 2;20(7):3654-3663. PMID: 36098618; PMCID: PMC9552228.

Wissenschaftliche Vorträge

2019 Poster:Quantitative proteomics of FFPE Medulloblastoma tissue reveals new molecular signatures for different cancer subtypes / Jahrestagung der deutschen Geselltschaft für Massenspektrometrie (DGMS) 2019/ Rostock, Deutschland

2019 Talk:Differential, Quantitative Tandem Mass Tag (TMT) based LC-MS/MS proteome analysis of FFPE Medulloblastoma tissue reveals new molecular signatures for different cancer subtypes/ UCCH- Cancer Retreat 2019/ Jesteburg, Deutschland

2020 Poster: Quantitative Tandem Mass Tag (TMT) based LC-MS/MS proteome analysis of FFPE Medulloblastoma tissue reveals new, orthogonally confirmable molecular signatures for different cancer types/ / Jahrestagung der deutschen Geselltschaft für Massenspektrometrie (DGMS) 2020/ Münster, Deutschland

2020 Poster: Quantitative, mass spectrometric proteomics of FFPE Medulloblastoma tissue reveals new, phenotypically relevant subtypes of Medulloblastoma/ UCCH- Cancer Retreat 2020 / Jesteburg, Deutschland

2020 Poster: Differential, Quantitative Tandem Mass Tag (TMT) based LC-MS/MS proteome analysis of FFPE Medulloblastoma tissue reveals new molecular signatures for different cancer subtypes / Kiel Mass Spec Forum 2020/ Kiel, Deutschland

2021 Poster: Sonic hedgehog (SHH)-medulloblastoma comprises two proteomic subtypes which are represented by Shh medulloblastoma mouse models/ UCCH- Cancer Retreat 2021/ Hamburg, Deutschland

2021 Poster: Infrared Laser based sampling for LC-MS/MS enables the guided, tissue depth resolved analysis of tissue specific proteomes / / Jahrestagung der amerikanischen Geselltschaft für Massenspektrometrie (ASMS) 2021/ Phildelphia, USA

2022 Talk: Missing-data tolerant integration of proteomic datasets enables the identification and characterization of disease subtypes / Jahrestagung der amerikanischen Geselltschaft für Massenspektrometrie (ASMS) 2022, Minneapolis, USA

2022 Poster:New developments in mass spectrometry-based proteome analysis enable the in-depth molecular characterization of cancer tissue/ UCCH- Cancer Retreat 2022/ Bad Bramsted, Deutschland

2022 Poster: LC-MS/MS based analysis of N-Glycans reveals specific glycan patterns for newly defined proteomic cancer subtypes /IMSC- 2022/ Maastricht, Neiederlande

2022 Poster: Multi-omic analyses of integrated cohorts reveal methylome, proteome and Nglycan signatures of brain cancer subtypes / Jahrestagung der deutschen Geseltschaft für Neurologie (DNG)/ Berlin, Deutschland

Wissenschaftliche Auszeichnungen

2020: Hubertus-Wald-Posterpreis in Würdigung einer hervorragenden Präsentation wissenschaftlicher Ergebnisse im Bereich der Krebsforschung und Krebstherapie.

2022: Posterpreis der Deutschen Gesekkschaft für Neurologie für das Poster Multi-omic analysis of integrated cohorts reveals methylome, proteome and N-glycane signatures of brain cancer subtypesïn der Kategorie Neuroonkologie".

Abstract

Investigating the proteome can add a significant layer of information to manifold existing methylation, mutation and transcriptome data, as proteins represent the pharmacologically addressable phenotype of a disease. Small sample cohorts limit the usability and validity of statistical methods, while variable technical setups, introducing phenotype suppressing batch effects to integrated datasets, make data integration from public sources challenging. Addressing this problem, here, the integrability of independently generated proteome datasets was investigated. Furthermore, the applicability of established bioinformatic strategies for batch effect reduction in other ”omics” entities to such proteome datasets was addressed. It was found that common methods for batch effect reduction in transcriptome and DNA methylation data, after experimental setup-specific preprocessing of proteomic data, can effectively reduce batch effects regardless of the used tissue conservation type, liquid chromatography- and mass spectrometry setup or quantification platform. As a limitation, existing, advanced strategies for batch effect removal cannot handle missing values of the ”missing not at random”(MNAR) type, frequently present in integrated proteome data sets. As a result, only 30-60% of identified proteins could be considered for batch effect correction for all datasets analyzed in this study. The application of established batch effect correction methods after machine learning-based imputation techniques, such as random forest imputation, were able to remduce batch effects, considering all proteins identified. As a limitation, data

imputation was found to distort non-imputed values during batch effect correction. In general, it was found that data imputation prior to batch effect correction can be considered as error prone, especially when MNAR and "missing at random" (MAR) type missing values are imputed at the same time. Targeting this problem, the matrix dissection framework for the missing value tolerant integration of independently generated data sets, without the need for data imputation, was developed in this study. By implementing several well-established algorithms, such as the empirical Bayesian framework of the ComBat algorithm and the linear regression model implemented in the Limma algorithm, the principle enables batch correction of Gaussian and non-Gaussian distributed proteomic data independently of the availability of spectral data. While the matrix dissection framework was established for proteomic data, the basic principle can be adapted for all "missing at random" and "missing completely at random" -type missing value tolerant batch effect correction strategies, data modalities and scientific questions.

Zusammenfassung

Die Untersuchung des Proteoms kann den vielfältigen vorhandenen DNA-Methylierungs-, Mutations- und Transkriptomdaten eine wichtige Informationsebene hinzufügen, da Proteine den Phänotyp biologischer Konditionen widerspiegeln, welcher häufig pharmakologisch adressiert werden kann. Kleine Kohorten schränken dabei die Validität statistischer Methoden zur Analyse hochdimensionaler Proteomdatensätze ein. Die Erweiterung eigener Datensätze mit Proteomdaten unabhängiger Studien, z.B. aus öffentlichen Datenbanken, hat das Potential, probenzahllimitierte Datensätze effizient zu erweitern. Eine solche Datenintegration ist allerdings durch die hohe technische Variabilität zwischen Proteomstudien limitiert, welche in integrierten Datensätzen Batcheffekte induziert, die biologisch relevante Unterschiede zwischen Phänotypen überlagern können. Im Rahmen der vorliegenden Arbeit wurde erstmalig die Integrierbarkeit von Proteomdaten aus unabhängig generierten Datensätzen, sowie die Anwendbarkeit etablierter bioinformatischer Verfahren zur Entfernung von Batcheffekten zwischen diesen, untersucht. Dabei konnte festgestellt werden, dass gängige Verfahren zum Entfernen von Batcheffekten in Transkriptom- und DNA-Methylierungsdaten, bei konfigurationsspezifischer Präprozessierung von Proteomdaten, Batcheffekte zwischen aus unterschiedlichen Gewebskonservierungstypen, Flüssigkeitschromatographie, Massenspektrometerkonfigurationen, Quantifizierungstechniken generierten Proteomdatensätzen erfolgreich reduzieren können. Bisher ist dabei limitierend, dass alle fort-

geschrittenen Verfahren zur Batcheffektreduktion keine Fehlwerte des "Missing not at random" (MNAR)-Typen tolerieren. Aus diesem Grund, reduzierte sich, für alle in dieser Studie untersuchten Datensätze, die Zahl der verwendbaren Datenpunkte nach Batcheffektkorrektur auf 30-60 % aller identifizierten Proteine. Derweil die Anwendung "Machine Learning"-basierter Imputationsverfahren (z. B. Random ForestImputation) eine Batcheffektkorrektur unter Berücksichtigung aller identifizierten Proteine ermög
-licht, konnte für alle getesteten Verfahren eine Datenverzerrung einzelner Proteine festgestellt werden. Die Anwendbarkeit von Imputationsverfahren vor der Batcheffektkorrektur zwischen Proteomdatensätzen ist dabei besonders durch die Notwendigkeit der gleichzeitigen Imputation von MNAR- und "Missing at random" (MAR)-Typ-Fehlwerten limitiert. Um dieses Problem zu umgehen, wurde im Rahmen dieser Studie das Matrix-Dissektionsverfahren zur fehlwerttoleranten Integration unabhängig generierter Datensätze, ohne die Notwendigkeit der Datenimputation, entwickelt. Das Prinzip ermöglicht durch die Implementierung verschiedener etablierter Algorithmen, wie dem empirischen Bayesian-Framework des ComBat-Algorithmus und dem linearen Regressionsmodell des Limma-Algorithmus, die Batcheffektkorrektur von normalverteilten und nicht
-normalverteilten Proteomdaten unabhängig der Verfügbarkeit von Spektraldaten. Derweil das Matrix-Dissektionsverfahren für Proteomdaten etabliert wurde, kann das grundlegende Prinzip für alle MAR-, MCAR-Typ-toleranten Batcheffektkorrekturstrategien, Datenmodalitäten und wissenschaftlichen Fragestellungen adaptiert werden.

Inhaltsverzeichnis

1

Einleitung

1.1 | Limitation der statistischen Validität von „Omics“-Studien durch kleine, unabhängige Kohorten

Groß angelegte ”Omics”-Experimente sollen ein vertieftes Verständnis molekularer Prozesse ermöglichen, einschließlich der Klassifizierung molekularer Subtypen, der Erkennung grundlegender molekularer Mechanismen und der Identifizierung neuer Biomarker oder Therapieziele. Aus ”Omics”-Studien resultieren Daten unterschied - licher Art. Darunter fallen z.B. DNA-Methylierungsprofile, Transkriptom- und Proteomdaten, für welche die Quantität mehrerer Tausend Faktoren über hohe Probenzahlen ermittelt wird. Bedingt durch die Hochdimensionalität solcher Daten werden statistische Verfahren benötigt, um relevante biologische Informationen aus ihnen zu extrahieren (42). Statistische Methoden zur Identifizierung krankheitsassoziierter biologischer Faktoren aus ”Omics”-Datensätzen sind, besonders für seltene Erkrankungen, häufig durch geringe Kohortengrößen limitiert. Die - durch den Publikationsdruck bedingte - verringerte Kooperationsbereitschaft einzelner Institutionen verstärkt diesen Effekt, sodass häufig mehrere größenlimitierte Einzeldatensätze zu einer Fragestellung existieren (11; 61).

Statistische Analysen kleiner Kohorten zeichnen sich durch eine geringe statistische Trennschärfe aus. Grund dafür ist, dass zufällige Schwankungen in kleinen Stichproben einen größeren Einfluss haben als in vergleichbaren Studien mit hohen Probenzahlen. Als Folge resultieren hohe Falsch-Negativ- und Falsch-Positiv-Raten. Die Wahrscheinlichkeit, dass aufgrund einer nicht repräsentativen Stichprobe verzerrte oder verfälschte Forschungsergebnisse resultieren, steigt (31).

Durch die Integration unabhängiger Datensätze, welche jedoch die gleiche Fragestellung betreffen - zum Beispiel aus öffentlichen Datenbanken - kann die Kohortengröße und statistische Validität stichprobenlimitierter Studien effizient erhöht werden. Die gemeinsame Analyse unabhängig generierter ”Omics”-Datensätze ist, aufgrund der An-

wendung unterschiedlicher technischer Konfigurationen, in einzelnen Studien, limitiert. Diese induzieren Batcheffekte in integrierten Datensätzen.

Batcheffekte bezeichnen nicht durch das biologische Experiment begründete Unterschiede zwischen mehreren Messungen (29). Solche technischen Varianzen können biologische Effekte maskieren, die Sensitivität biologischer Analysen reduzieren und in statistischen Artefakten resultieren (71).

1.2 | Varianztypen in Proteomdatensätzen

Häufig untersuchte Omics-Typen, wie Transkriptom- und DNA-Methylierungsdaten, werden mit einer begrenzten Anzahl technischer Konfigurationen erfasst und weisen eine relativ hohe Vergleichbarkeit und Datenvollständigkeit über Experimente hinweg auf. In diesen Bereichen ist die Reduzierung von Batcheffekten gut etabliert (39; 46). Im Gegensatz dazu leiden aufkommende Technologien – zu welchen auch die Proteomanalyse zählt – unter geringer Datenvollständigkeit und hohen experimentellen Varianzen (1; 8). Proteomanalysen nehmen eine Sonderstellung ein, da sie im Gegensatz zu DNA-Sequenzdaten, DNA-Methylierungsprofilen und Transkriptomdaten pharmakologisch adressierbare Phänotypen direkt widerspiegeln (14). Die Generierung globaler Proteomdatensätze mittels Flüssigkeitschromatographie-gekoppelter Tandem-Massenspektrometrie (LC-MS/MS) ermöglicht die gleichzeitige Quantifizierung von mehr als 1000 Proteinen. Dabei werden Proteine aus komplexen Proben extrahiert, denaturiert und proteolytisch zu Peptiden gespalten. Resultierende Peptide werden mittels Flüssigkeitschromatographie auf Basis ihrer chemischen Eigenschaften aufgetrennt. Zu jedem Retentionszeitpunkt der Chromatographie werden Peptide definierter chemischer Eigenschaften in ein massenspektrometrisches System überführt, ionisiert und analysiert. Dabei werden für jedes Peptid spezifische Fragmentmuster erzeugt, welche eine Identifikation der Peptide ermöglichen. Identifizierte Peptide werden anhand ihrer Aminosäuresequenz den ihnen zugrundeliegenden Proteinen zugeordnet und zur rela-

tiven Quantifizierung dieser innerhalb des Datensatzes herangezogen (69).
Technische Unterschiede in verschiedenen Schritten dieser Analyse induzieren Batcheffekte zwischen integrierten, unabhängig generierten Datensätzen.

1.2.1 | Technische Varianzen bei der Extraktion von Proteinen

Die erste Varianzquelle zwischen unabhängig generierten Proteomstudien stellt Extraktion von Proteinen aus biologischem Material wie Zellen oder Geweben dar. Dabei werden Detergenzien wie Natriumdodecylsulfat (SDS), Natriumdesoxycholat (SDC) oder Urea zur Solubilisierung von Zellen benötigt. Die Wahl des Puffersystems beeinflusst dabei die Effizienz von Proteindenaturierung, Löslichkeit und der folgenden proteolytischen Spaltung einzelner Proteine. So können beispielsweise im Vergleich zwischen verschiedenen Lysepuffern, bei identischen globalen Proteinidentifikationsraten und Signalwegabdeckungen, durchschnittlich 600 Proteine durch nur ein Puffersystem extrahiert werden (64).
Die meisten für die Proteinextraktion verwendeten Detergenzien und Puffersysteme sind nicht mit dem nachgeschalteten proteolytischen Verdau – der LC oder MS Analyse – kompatibel. Sie müssen durch die Verwendung von Protokollen zur Entfernung von Detergenzien und Salzen eliminiert werden. Solche Methoden resultieren häufig in Materialverlust und basieren auf unterschiedlichen chemischen Prinzipien, welche Batcheffekte in unabhängig generierten, integrierten Datensätzen induzieren können. So zeigten z.B. *Jehmlich et al. (2014)*, dass 4-5% aller identifizierten Proteine aus humanem Speichel nur nach Entsalzung mit einem von vier verglichenen Umkehrphasenchromatographie (”Reversed Phase”, RP)-basierten Verfahren mit unterschiedlicher Abundanz identifiziert werden (20).
In Bezug auf die Proteinextraktion nehmen Gewebeproben eine Sonderstellung ein. Die Fixierung mit Formalin und die Paraffineinbettung (FFPE) sind ein weltweiter Standard für die Konservierung, Lagerung und Aufbereitung von Geweben. FFPE-Material

stellt aufgrund der geringeren Proteinextraktionseffizienz, durch unvollständigen Umkehr von Methylenbrücken und der Induktion irreversibler chemischer Veränderungen, eine Herausforderung für die massenspektrometrische Analyse dar (33). Bedingt durch diese Herausforderung existieren verschiedene Protokolle zur Extraktion von Proteinen aus FFPE Gewebe, welche mit unterschiedlicher Proteinextraktionseffizienz und Identität extrahierter Proteine assoziiert sind (6). Zwischen FFPE- und FF-Gewebe derselben Probe kann eine hohe quantitative Korrelation aus beiden Gewebskonservierungstechniken identifizierter Proteine erwartet werden. Gleichzeitig zeigen FFPE-basierte Analysen deutlich geringere Proteinidentifikationsraten und Proteinabundanzen. So zeigten z.B. *Mantsiou et al.(2020)* (35), dass aus muriner FF-Niere über 500 Proteine mehr identifiziert wurden, als aus FFPE-Material der gleichen Tiere.

1.2.2 | Technische Varianzen im proteolytischen Verdau von extrahierten Proteinen

Neben der Proteinextraktion stellt der proteolytische Verdau von Proteinen eine Varianzquelle zwischen unabhängig generierten Datensätzen dar. So können z.B. verschiedene Proteasen mit unterschiedlicher Aminosäurespezifität verwendet werden. Am häufigsten wird die Serinprotease Trypsin verwendet, welche C-terminal hinter Lysin- und Argininresten schneidet und, zusammen mit der N-terminalen Aminogruppe, bei niedrigem pH Wert in mindestens zweifach geladenen Peptiden für die folgende LC-MS-Analyse resultiert (7). Für den tryptischen Verdau können z.B. das Magnetkugel-basierte "Single-pot, solid phase-enhanced sample preparation" (SP3)-Protokoll (17) oder das Filtermembran-basierte "Filter Aided Sample Preparation" (FASP)-Protokoll verwendet werden (65). Die Anwendung unterschiedlicher Protokolle für den tryptischen Verdau führt zur Identifikation unterschiedlicher Proteine. So zeigten z.B. *Sielaff et al. (2017)*, dass beim tryptischen Verdau von 20 Mikrogramm HeLa-Zellysat im Vergleich zwischen FASP, SP3 und "in-Stage-Tip digestion" zwischen 6 und 18% aller Proteine

nur nach tryptischem Verdau mit einem der verglichenen Verfahren identifiziert wurde. Dieser Effekt verstärkte sich bei der Verwendung geringerer Ausgangsmengen (52). Des Weiteren können verschiedene Inkubationszeiten, Temperaturen und die Konzentration von Zusatzstoffen wie Acetonitril (ACN) einen signifikanten Einfluss auf die Effizienz des tryptischen Verdaus sowie die Abundanz und Art identifizierter Proteine haben (23).

1.2.3 | Technische Varianzen durch verschiedene LC-MS- Konfigurationen

Aus dem Proteaseverdau resultierende Peptide werden durch eine Kopplung von Flüssigkeitschromatrographie (LC) und Massenspektrometrie-Systemen (MS) analysiert. Dabei können technische Varianzen durch Unterschiede in jedem Bestandteil der experimentellen Konfigurationen induziert werden. Ein Beispiel stellt die Wahl der stationären Phase zur chromatographischen Trennung der Peptide dar. So können z.B. Normalphasen-, Umkehrphasen-, hydrophile Interaktions- und Ionenaustauschchromatographie mit Massenpektrometern gekoppelt werden (68). Die Chromatographie wird verwendet, um die Komplexität biologischer Proben für die MS-Analyse zu reduzieren, da pro Scan nur eine Teilmenge aller injizierten Peptide analysiert werden kann. Welche Peptide pro MS-Scan analysiert werden, hängt mit der Peptidumgebung – bei einer definierten Pufferkomposition eluierender Peptide – zusammen (9), welche unmittelbar von der Wahl der stationären und mobilen Phase abhängt. Folglich werden bei der Verwendung unterschiedlicher stationärer und mobiler Phasen verschiedene Peptide identifiziert. Neben Unterschieden im chemischen Trennprinzip verschiedener chromatographischer Systeme beeinflusst die Wahl der Säulen- und Gradientenlängen die Zahl und Intensität identifizierter Peptide (9).

Zusätzlich kann die Konfiguration massenspektrometrischer Systeme technische Varianzen zwischen unabhängig generierten Proteomdatensätzen induzieren. Ein Beispiel

hierfür ist die Verwendung unterschiedlicher Massenanalysatoren. So werden beispielsweise bei Quadrupol-Orbitrap Hybrid Masssenspektrometern Vorläufer-/Produkt -Ionen in einer C-Trap akkumuliert und in eine Orbitrap-Ionenfalle injiziert. Die Massenanalyse erfolgt, indem die Frequenz der eliptischen, axialen Oszillation von Ionen um eine Zentralelektrode durch die Induktion eines elektrischen Stroms detektiert wird. Diese korreliert proportional mit dem Masse-zu-Ladungsverhältnis der Ionen. Mit Hilfe der Fourier-Transformation wird dieses Signal in ein Massenspektrum überführt, in welchem das Masse-zu-Ladungsverhältnis (m/z) von Ionen gegen ihre Intensität aufgetragen wird (15). Bedingt durch die Akkumulation der Ionen können Analyten mit einer hohen Sensitivität identifiziert werden. Im Vergleich zu "Time of Flight" (TOF)-Geräten, bei denen das Masse-zu-Ladungsverhältnis durch die mit ihrer korrelierender Flugzeit von Ionen in einer feldfreien Driftstrecke ohne Akkumulation von Ionen zusammenhängt (15), erscheint die Abundanz identifizierter Peptide/Proteine höher. Gleichzeitig wird durch die Akkumulation der Ionen die Scangeschwindigkeit Orbitrap-basierter Messungen verlangsamt. Dies resultiert in einer verringerten Proteinidentifikationsrate von Orbitrap-basierten Messungen im Vergleich zu TOF-Analysen (18).

1.2.4 | Technische Varianzen durch verschiedene Quantifizierungstechniken

Neben der LC-MS-Konfiguration zur Analyse proteolytischer Peptide kann die Verwendung unterschiedlicher Quantifizierungstechniken in verschiedenen Studien Batcheffekte in integrierten Datensätzen induzieren. Peptidabundanzen werden durch die Anwendung verschiedener Label-freier oder Label-basierter Verfahren ermittelt.
Die Label-freie Quantifizierung (LFQ) erfolgt dabei entweder auf Ebene der Vorläuferionen oder Fragmentionen. Es wird zwischen dem datenabhängigen- (”Data dependent acquisition“, DDA) und datenunabhängigen (”Data independent acquisition“, DIA) Aufnahmeprinzip unterschieden. In der Bottom-Up-Proteomanalyse wird zunächst das Masse-zu-Ladungsverhältnis unmarkierter, von der LC eluierender Peptide bestimmt (Massenspektrometrische Analyseebene 1, MS1). Analysierte Peptide (Vorläuferionen) werden im Folgenden fragmentiert. Die dabei entstehenden peptidspezifischen Fragmente werden massenspektrometrisch analysiert (Massenspektrometrische
Analyseebene 2,MS2) und - durch den Abgleich mit Datenbanken -theoretischen Peptidfragmentspektren zugeordnet. Durch die Zuordnung dieser Peptide zu Proteinen können Proteine identifiziert und quantifiziert werden.
Die datenbankgestützte Zuordnung von Peptiden zu theoretischen Datenbanken ist hinsichtlich der Komplexität zuordnungsbarer Spektren limitiert. Aus diesem Grund müssen Peptide vor der Fragmentierung elektrisch isoliert werden. Dadurch entstehen individuelle Fragmentspektren für einzelne Vorläuferionen. Bedingt durch den dadurch entstehenden hohen Zeitaufwand pro Peptid kann nur eine begrenzte Zahl von Peptiden pro chromatographischem Retentionszeitpunkt fragmentiert und identifiziert werden. Als Folge dieser Datenabhängigkeit im DDA-Modus werden häufig nur die abundantesten Peptide pro MS1-Scan berücksichtigt. Die Quantifizierung im Rahmen der markierungsfreien DDA-Messung erfolgt für alle identifizierten Peptide auf MS1-Ebene durch die Ermittlung der Intensität oder Fläche unter der Kurve der Vorläuferionen im

extrahierten Ionenchromatogramm (EIC) auf MS1-Ebene (16).

Im Gegensatz dazu werden bei DIA-Messungen alle Peptide innerhalb eines definierten m/z-Fensters pro chromatographischem Retentionszeitpunkt gleichzeitig fragmentiert. Die Identifikation von Peptiden aus solchen komplexen Spektren wird über die Verwendung von Spektralbibliotheken ermöglicht. Eine Spektralbibliothek bezeichnet dabei eine Datenbank, die massenspektrometrische und chromatographische Parameter wie Vorläufer- und Fragment-Masse-zu-Ladungsverhältnisse, Fragmenttyp, Ladung und Elutionszeit für potenziell in einer Probe vorhandene Peptide enthält. Diese Studien

-spezifischen Spektralbibliotheken werden üblicherweise durch eine DDA-basierte proteomische Charakterisierung derselben Proben, vor der Analyse mittels DIA-MS erstellt. Die Zuordnung von Peptiden aus DIA-Messungen ist deswegen, in der Regel, auf die Quantifizierung von Peptiden limitiert, welche in DDA Messungen der Probe eindeutig identifiziert werden können. Die Quantifizierung erfolgt dabei durch die Summe der Fläche unter der Kurve aller identifizierten Fragmente eines Vorläuferpeptides im fragmentspezifischen EIC auf MS2-Ebene (16; 32).

Label-basierte Quantifizierungsstategien beruhen auf dem DDA-Prinzip und quantifi -zieren Peptide unter Nutzung stabiler Isotopenlabel. Dabei wird zwischen metabolischen und chemischen Markierungsstrategien unterschieden. Bei der am häufigsten verwendeten metabolischen Quantifizierungsstategie „stable isotope labeling of amino acids in cell culture“ (SILAC) werden mit stabilen schweren Isotopen markierte Amino -säuren (in der Regel Lysin und/oder Arginin) im Kulturmedium wachsender Zellen zur Verfügung gestellt. Dabei wird jedes Vorhandene Protein in der Probe mit schweren Isotopen markiert. Aus diesen Zellen erzeugte Peptide werden gemeinsam mit unmarkierten Peptiden einer Referenzprobe vermessen. Peptidabundanzen werden auf MS1-Ebene ermittelt und als Verhältnis zwischen der zu analysierenden Probe und der Referenzprobe dargestellt (34).

Chemische Markierungsprinzipien wie die Tandem Mass Tag (TMT)-basierte Proteo-

manalyse koppeln mit stabilen Isotopen gelabelte Modifikationen nach dem proteolytischen Verdau an die entstandenen Peptide. Die TMT-Quantifizierung basiert auf dem TMT-Reagenz, welches aus drei Hauptbestandteilen besteht: einer N-Methyl-2,6-dimethylpiperidin
-Reportergruppe, dem N-Hydroxysuccinimid-Ester der Aminopropionsäure als reaktive Gruppe sowie eine Ausgleichsgruppe für die Massennormalisierung (4; 57; 66). Verschiedene TMT-Reagenzien weisen dabei dieselben chemischen Eigenschaften und Massen auf, haben aber unterschiedliche C13- und N15-Schwerisotopenverteilungen innerhalb ihrer Reporter- und Ausgleichsgruppe. Bedingt durch diese unterschiedliche Isotopenverteilung, lassen sich TMT-Label massenspektrometrisch unterscheiden. Für die TMT-Analyse werden einzelne Proben, die einen bestimmten Proteomzustand repräsentieren, mit verschiedenen TMT-Reagenzien markiert. Zu diesem Zweck reagiert der N-Hydroxysuccinimid-Ester der reaktiven Gruppe mit primären Aminen am N-Terminus der Peptidkette oder an Lysin-Seitenketten. Nach der Markierung werden die Peptide aus Proben mit verschiedenen Labeln kombiniert und gemeinsam analysiert. Bei der TMT-basierten LC-MS-Messung haben aus verschiedenen Proben extrahierte, identische Peptide den gleichen m/z-Wert auf MS1-Ebene, da die individuelle Isotopenzusammensetzung der TMT-Reportergruppe durch die Isotopenverteilung innerhalb der Ausgleichsgruppe ausgleichen wird. Bei der Isolation und Fragmentierung einzelner Peptide werden neben dem Peptidrückrad auch die TMT-Reagenzien fragmentiert. Dabei wird die TMT-Reportergruppe abgetrennt, wobei TMT-Reporterionen entstehen. Durch das Fehlen der Ausgleichsgruppe in TMT-Reporterionen können sie auf Grund ihrer individuellen Isotopenzusammensetzung unterschieden und einzelnen Proben zugeordnet werden. Die Intensität dieser Reporter in jedem Peptid-spezifischen MS2-Spektrum wird verwendet, um das Peptid in den Reporter-assoziierten Proben zu quantifizieren (4; 57; 66).

Die – durch die Verwendung von TMT Markierungen erreichte – "Multiplex"-Analyse ermöglicht eine erhöhte Reproduzierbarkeit, Vergleichbarkeit und Quantifizierungsge-

nauigkeit zwischen Proben, da bei der gemeinsamen Vermessung die Varianz von Umwelt -einflüssen verringert wird. Sie ist allerdings durch die maximale Verfügbarkeit der 16 verschiedenen Isotopenlabel limitiert (62). Werden im Rahmen größer angelegter Studien höhere Probenzahlen analysiert, so müssen mehrere TMT-Plexe vermessen werden. Die Integration verschiedener TMT-Plexe induzierte zusätzliche Batcheffekte in integrierten Datensätzen, welche vor der Datenanalyse eliminiert werden müssen (3). Neben den Varianzen zwischen verschiedenen Quantifizierungstechniken kann auch die Verwendung verschiedener Algorithmen zur Quantifizierung von Peptiden und Proteinen in Batcheffekten resultieren. So unterscheidet sich die Peptid- und Proteinquantifizierung unter Nutzung der etablierten bioinformatischen Plattformen MaxQuant (59) und Proteome Discoverer (25) in Ausbeute, Dynamikbereich, Reproduzierbarkeit, Spezialität und Sensitivität quantifizierter Peptide und Proteine für LFQ-DDA-Messungen. Einen weiteren Einflussfaktor bildet die Wahl der Normalisierungsstrategie innerhalb der verwendeten Plattform (40).

1.3 | Eignung und Limitation etablierter Batcheffektkorrekturstrategien für Proteomdatensätze

Bislang gibt es nur wenige Proteomdaten-spezifische Tools, die eine Reduktion solcher Batcheffekte ermöglichen. Existierende Anwendungen, wie beispielsweise der ProNorM-Algorithmus, erfordern Negativkontrollen oder interne Standards um unerwünschte technische Varianzen zu reduzieren (45). Folglich sind sie nicht mit der Integration unabhängig generierter Proteomdatensätze kompatibel, welche über keine vergleichbaren internen Referenzen verfügen. Referenzprobenunabhängige Strategien zur Reduktion von Batcheffekten zwischen Proteomdatensätzen wie ”proBatch” implementieren grundlegende Techniken wie Quantil- und Median-Normalisierung. Diese Techniken sind allerdings – verglichen mit, für andere ”Omics”-Typen etablierte, Batcheffektkor-

rekturstrategien – wenig performant (71). Gleichzeitig stützen sich die meisten RNA-Sequenzierungs-/Transkriptom-basierten Strategien zur Reduktion von Batcheffekten, die mit der Struktur von Proteomdaten kompatibel sind, auf nicht überwachte, nicht-lineare Dimensionsreduktionsmethoden wie die Hauptkomponenten Analyse
(”Principal component analysis”, PCA) oder ”t-distributed stochastic neighbor embedding”
(t-SNE). Solche Verfahren sind nicht mit Fehlwerten kompatibel und erfordern eine vollständige Datenmatrix (Harmony (25), LIGER (58), deepMNN (70), MMD-resnet (70)).
Proteomdaten weisen hohe Raten fehlender Werte in Datensätzen auf, welche verschiedenen Fehlwerttypen zugeordnet werden können. Fehlwerte können in drei Klassen eingeordnet werden: ”missing completely at random” (MCAR), ”missing at random” (MAR) und ”missing not at random” (MNAR). Das Fehlen von MCAR-Typ-Fehlwerten ist zufällig verteilt und damit völlig unabhängig von allen anderen Variablen des Datensatzes, derweil das Fehlen von MAR-Typ-Fehlwerten auf beobachtbare Variablen des Datensatzes zurückgeführt werden kann (26). Besonders für im DDA-Modus aufgenommene Proteomdaten können 70-90% der identifizierten Proteinen mindestens in einer Probe einen fehlenden Wert des MCAR- oder MAR-Typen aufweisen (27). Durch die Bedingung, dass Matrix-Algebra-basierte Verfahren nur auf Datenmatritzen ohne Fehlwerte anwendbar sind, können solche Strategien nicht für die Batcheffektkorrektur an Proteomdaten genutzt werden. Modifizierte Versionen von t-SNE und PCA wie z. B. InDaPCa sind zwar anwendbar, aber derzeit nicht in der Struktur existierender Algorithmen zur Batcheffektreduktion implementiert (22).
Die bekanntesten, nicht Matrixalgebra-basierten Ansätze zur Batcheffektreduktion sind die Funktion „removeBatchEffect()“ des Limma-Algorithmus (47), welcher ein lineares Regressionsmodell implementiert und der ComBat-Algorithmus (21), der auf einem empirischen Bayesian-Framework beruht. Beide Strategien akzeptieren Fehlwerte des MAR- und MCAR-Typen. Die Anwendbarkeit solcher Verfahren zur Batcheffektkorrektur in Proteomdatensätzen wurde bereits mehrfach publiziert (37; 43; 54). Einschrän-

kend ist, dass Fehlwerte des "Missing not at Ramdom" (MNAR)-Typen, welche nicht zufällig sind, systematisch von den betrachteten Variablen eines Datensatzes abweichen und mit dem unbeobachteten Teil der Daten korrelieren (26), nicht toleriert werden. Das Fehlen von Proteinen in einzelnen Studien integrierter, unabhängig generierter Datensätze korreliert systematisch mit dem verwendeten experimentellen Setup und hängt dabei mit Faktoren zusammen, die von den untersuchten Variablen des biologischen Experimentes abweichen. Als Folge werden sie dem MNAR-Typen zugeordnet und müssen - vor der Batcheffektkorrektur - mit Limma (47) oder ComBat (21) eliminiert werden. Dies führt zu einer Reduktion der statistischen Validität und biologischen Aussagekraft integrierter Datensätze.

Als Alternative zur Datenreduktion können fehlende Werte imputiert werden, um vor der Batcheffektkorrektur artifiziell eine vollständige Datenmatrix zu erzeugen. Bei der einfachen Imputation werden fehlenden Werte durch einen nach einer bestimmten Regel definierten Wert ersetzt. Dabei können unterschiedlich komplexe Imputationsverfahren angewandt werden, welche auf verschiedenen Annahmen über die Natur fehlender Werte beruhen (19). Die in der verbreitetsten Statistiksoftware zur Analyse von Proteomdaten (Perseus) (59) implementierte Imputation über die Normalverteilung per Probe oder Matrix trifft dabei die Annahme, dass nicht identifizierte Proteine eine niedri -gere Abundanz aufweisen als der Durchschnitt der identifizierten Proteine pro Probe oder Matrix. Diese Annahme ist für MCAR- und MAR-Fehlwerte von LC-MS-Daten häufig unrealistisch und resultiert in verzerrten Daten (63). Die "Random Forest" (RF)-Imputation solcher Werte wird als deutlich performanter für LC-MS-Daten klassifiziert (24; 63). RF-Imputation beschreibt eine "Machine learning"-Strategie, welche ein nichtparametrisches Modell nutzt, das keine Annahmen über die Verteilung der Daten macht, komplexe Strukturen in den Daten lernen kann und sowohl für kontinuierliche als auch für kategoriale Variablen funktioniert (63).

Das Einführen artifizieller Werte ist vor allem für die Batcheffektkorrektur integrierter Datensätze kritisch, da Fehlwerte des MAR-, MCAR- und MNAR-Typs gleichzeitig

imputiert werden müssen, da für die valide Imputation verschiedener Fehlwerttypen unterschiedliche Voraussetzungen gegeben sind. So zeigten z.B. *Kokla et al. (2019)* (24), dass für die Imputationsmethode, für welche der geringste mittlere quadratische Fehler ("Mean Squarred Error" MSE) bei der Imputation von MAR- und MCAR-Fehlwerten aus quantitativen LC-MS-Daten identifiziert wurde, der höchste MSE-Wert für MNAR-Typ-Fehlwerte resultierte. Grund dafür ist, dass für MNAR-Typ-Fehlwerte eine geringe Abundanz des jeweiligen Faktors in einer Probe angenommen werden kann. Als Resultat nähert eine Imputation von MNAR-Typ-Fehlwerten durch den minimal im Datensatz identifizierten Wert den realen Zustand an. Durch ihre Korrelation mit den beobachtbaren Variablen des Datensatzes werden MAR- und MCAR-Typ-Fehlwerte hingegen am besten durch anhand der Datenstruktur selbst erzeugten Werte ersetzt, welche sich durch eine RF-Imputation erzeugen lassen.

Als Resultat können beide Fehlwerttypen nur durch die Nutzung multipler Imputationsmethoden effizient gleichzeitig imputiert werden. Multiple Imputationsverfahren werden allgemein als fehleranfällig eingestuft (19; 24; 41).

2

Zielsetzung

Die Untersuchung des Proteoms kann den vielfältigen vorhandenen Methylierungs-, Mutations- und Transkriptomdaten eine wichtige Informationsebene hinzufügen, da Proteine den pharmakologisch adressierbaren Phänotyp biologischer Konditionen widerspiegeln. Kleine Kohorten schränken dabei die Verwendbarkeit und Gültigkeit statistischer Methoden ein. Die Erweiterung eigener Datensätze mit Proteomdaten unabhängiger Studien, z.B. aus öffentlichen Datenbanken, hat das Potential, probenzahllimitierte Datensätze effizient zu erweitern. Sie ist aber durch die hohe technische Variabilität zwischen Proteomstudien limitiert, welche biologische Varianzen überlagert. Hohe Fehlwertraten verhindern dabei die effiziente Anwendbarkeit gängiger Verfahren zur Korrektur solcher Batcheffekte.

Um die gemeinsame statistische Analyse unabhängig generierter Proteomdatensätze zu ermöglichen, war das Ziel dieser Arbeit die Entwicklung eines fehlwerttoleranten Verfahrens zur Integration und Batcheffektkorrektur zwischen unabhängig generierten Proteomdatensätzen.

3

Ergebnisse

3.1 | Adressierte Varianztypen und verwendete Datensätze

Die Generierung von Proteomdatensätzen mittels Flüssigkeitschromatographie -gekoppelter Tandem-Massenspektrometrie (LC-MS/MS) unterscheidet sich zwischen verschiedenen Laboratorien besonders durch die Verwendung unterschiedlicher Chromatographiesysteme, Massenspektrometer-Aufbauten, Quantifizierungstechniken und, im Fall der Nutzung von Gewebeproben, Gewebskonservierungstechniken. Eine weitere Varianzquelle stellt die Probenanalyse zu unterschiedlichen Zeitpunkten, auch unter der Nutzung identischer experimenteller Aufbauten, dar. Die Integration unabhängig generierter Datensätze zur Erhöhung der Probenzahl erhöht die statistische Validität kleiner Kohorten, unterliegt diesen Varianzen. Um die Integrierbarkeit unabhängig generierter Proteomdatensätze unter Berücksichtigung verschiedener Varianztypen zu untersuchen, wurden in dieser Studie die in Abbildung 3.1. dargestellten Datensätze verwendet.

Zur Analyse der Integrierbarkeit mit unterschiedlichen LC-MS-Konfigurationen aufge -nommener Proteomdaten wurden zwei definierte Phänotypen durch die Kombination von Homo sapiens-, *E. coli*- und *Saccaromyces cerevisea* -Zell-Lysaten erzeugt. (Phänotyp 1: 80 % *Homo sapiens*, 15 % *E. coli*, 5 % *Saccaromyces cerevisea*; Phänotyp 2: 80 % *Homo sapiens*, 10 % *E. Coli*, 10 % *Saccaromyces cerevisea*, Abb. 3.1.a) und mit unterschiedlichen LC-MS-Setups (DDA-Messung auf QExactive-Massenspektrometer; DIA-Messung auf- QExactive-Massenspektrometer; und SWATH-Messung, TripleTOF 6600-Massen -spektrometer) vermessen. Insgesamt wurden 5543 Proteine identifiziert (Batch 1 (SWATH-TripleTOF 6600): 3151 Proteine; Batch 2 (DIA-QExactive): 2980 Proteine; Batch 3(DDA-QExactive): 4118 Proteine (Spike-in-Datensatz).

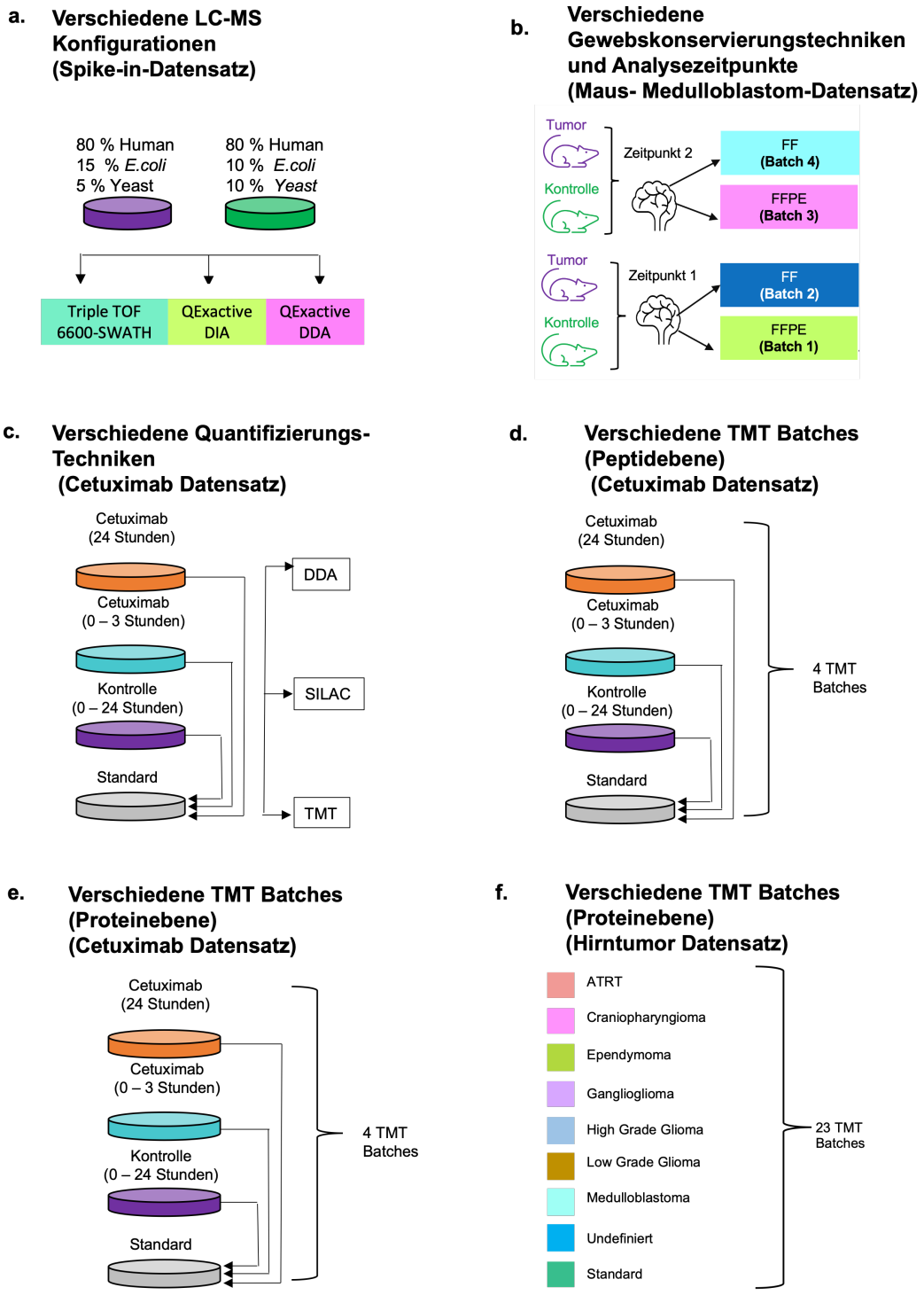

Abbildung 3.1: Schematische Darstellung der verwendeten Datensätze zur Analyse der Integrierbarkeit unabhängig generierter Proteomdatensätze.

Um Varianzen zwischen Proteomdaten verschiedener Gewebskonservierungstechniken und Analysezeitpunkten zu adressieren, wurden unterschiedlich aufbereitete Pro -ben eines etablierten Sonic Hedgehog(Shh)-Medulloblastom-Mausmodells ((50)) (hGFAP-cre::SmoM2Fl/+) analysiert. Im Einzelnen wurden Kleinhirntumore von hGFAP-cre::SmoM2Fl+ - Mäusen und Kontroll-Kleinhirn von SmoM2Fl+ - Wurfgeschwistern halbiert. Die eine Hälfte wurde bei -80 °C eingefroren (frisch gefrorener Zustand (fresh frozen, FF)). Die andere Hälfte wurde mit Formalin fixiert und in Paraffin eingebettet (FFPE). Die Experimente wurden wiederholt, um verschiedene Analysezeitpunkte zu simulieren. 3530 Proteine konnten zu mindestens einem Zeitpunkt aus mindestens einer Gewebs-Konservierungstechnik identifiziert werden (Batch 1 (Zeitpunkt1, FFPE):1613; Batch 2 (Zeitpunkt1, FF):2648; Batch 3 (Zeitpunkt2, FFPE):1609; Batch 4 (Zeitpunkt1, FF):2660)) (Abbildung 3.1.b)). Eine gemeinsame Prozessierung von LC-MS/MS-Rohdaten unter Nutzung des Minora-Algorithmus zur Reduktion von Batcheffekten und Erhöhung der Daten-Vollständigkeit auf Chromatogramm-Ebene((40)) ergab 4786 identifizierte Proteine. (Batch 1 (Zeitpunkt1, FFPE): 3938 Proteine; Batch 2 (Zeitpunkt1, FF): 4442 Proteine; Batch 3 (Zeitpunkt2, FFPE): 4544 Proteine; Batch 4 (Zeitpunkt1, FF): 4284 Proteine), Abbildung 3.1.b)).

Zur Analyse der Integrierbarkeit mit verschiedenen Quantifizierungstechniken aufge -nommener Proteomdaten wurden ein 2020 von Stepath et al. ((55), PDXD014565) veröffentlichter Datensatz verwendet. Der Datensatz enthält unnormalisierte Peptid- und Proteinabundanzen von Kolorektalkarzinom-Zellen (Zellinie DiFi) mit und ohne Cetuximab -Behandlung, nach 0, 3 und 24 Stunden Inkubation. Die Quantifizierung erfolgte mittels Stable Isotope Labling by Amino Acids in Cell Culture (SILAC), Tandem Mass Tag (TMT) oder Label-freier Quantifizierung (LFQ) im datenabhängigen Akquisitions-Modus (DDA). Insgesamt wurden 6754 Proteine quantifiziert. (TMT: 2579 Proteine; DDA: 6081 Proteine; SILAC: 4141 Proteine, (Abbildung 3.1.c)).

Die Verwendung von 8-Plex-TMT ermöglicht die Multiplex-Analyse von bis zu acht Proben. Wird diese Probenzahl überschritten, so müssen mehrere TMT-Batches vermessen werden. Resultierende Varianzen müssen nach Rohdatenprozessierung auf Peptid- oder Proteinebene eliminiert werden ((3)). Zur Adressierung dieser TMT -spezifischen Varianz, wurde der von *Stepath et al (2020)* publizierte ((55)) TMT-Datensatz separat auf Peptid und Proteinebene untersucht (Abbildung 3.1.d,e)). Insgesamt wurden 12615 Peptide (TMT-Batch 1: 8108 Peptide; TMT-Batch 2: 7747 Peptide; TMT-Batch 3: 8488 Peptide; Batch 4: 8200 Peptide) identifiziert. Diese wurden zur Quantifizierung von 2579 Proteinen (TMT-Batch 1: 1998 Proteine; TMT-Batch 2: 1986 Proteine; TMT-Batch 3: 1973 Proteine; TMT-Batch 4: 2072 Proteine), in mindestens einem TMT-Batch herangezogen.

Um TMT-spezifische Varianzen in einer größeren, biologischen Kohorte zu analysieren, wurde zusätzlich ein 2021 von *Petralia* et al. publizierter Datensatz verwendet((43)). Ziel der Studie war der Vergleich der Proteom-Profile von 8 verschiedenen kindlichen Hirntumor-Entitäten. Insgesamt konnten aus 23 TMT-11-Plex-Batches 9156 Proteine quantifiziert werden (Abbildung 3.1. f).

3.2 | Analyse der Anwendbarkeit eines empirischen Bayesian-Frameworks zur Batcheffektreduktion zwischen unabhängig generierten Proteomdaten-sätzen

Initial wurde die Integrierbarkeit von Proteomdatensätze welche über unterschiedliche LC-MS-Konfigurationen aufgenommen wurden geprüft. (Abbildung 3.2.a), 3.3.a)).

In Pearson-Korrelations-basiertem hierarchischem Clustering (HC) konnte ein Clustering in Abhängigkeit des LC-MS-Setup nach Datenintegration beobachtet werden. Mittels SWATH-TripleTOF 6600 vermessene Proben bildeten ein separates Cluster und zeig -ten deutlich geringere Proteinabundanzen im Vergleich zu DIA-QExactive- und DDA-QExactive-Messungen. Unterschiede zwischen Phänotyp 1 und Phänotyp 2 konnten nur innerhalb einzelner LC-MS-Konfiguration beobachtet werden. (Abbildung 3. ´2.b)).

Des Weiteren konnte für alle Konfigurationen innerhalb eines experimentellen Setups eine lineare Korrelation >99% zwischen Phänotyp 1-Proben beobachtet werden. Zwischen Proteinabundanzen TripleTOF 6600 (SWATH)- und QExactive (DIA,DDA)-basierter-Messungen wurde die geringste Korrelation festgestellt (SWATH-Triple TOF 6600 versus DIA-QExactive: 0.68; SWATH-TripleTOF 6600 versus DDA-QExactive: 0.66). Proteinabundanzen aus DIA-QExactive- und DDA-QExactive-Messungen zeigten eine Korrelation von 77%. (Abbildung 3.2.c, d).

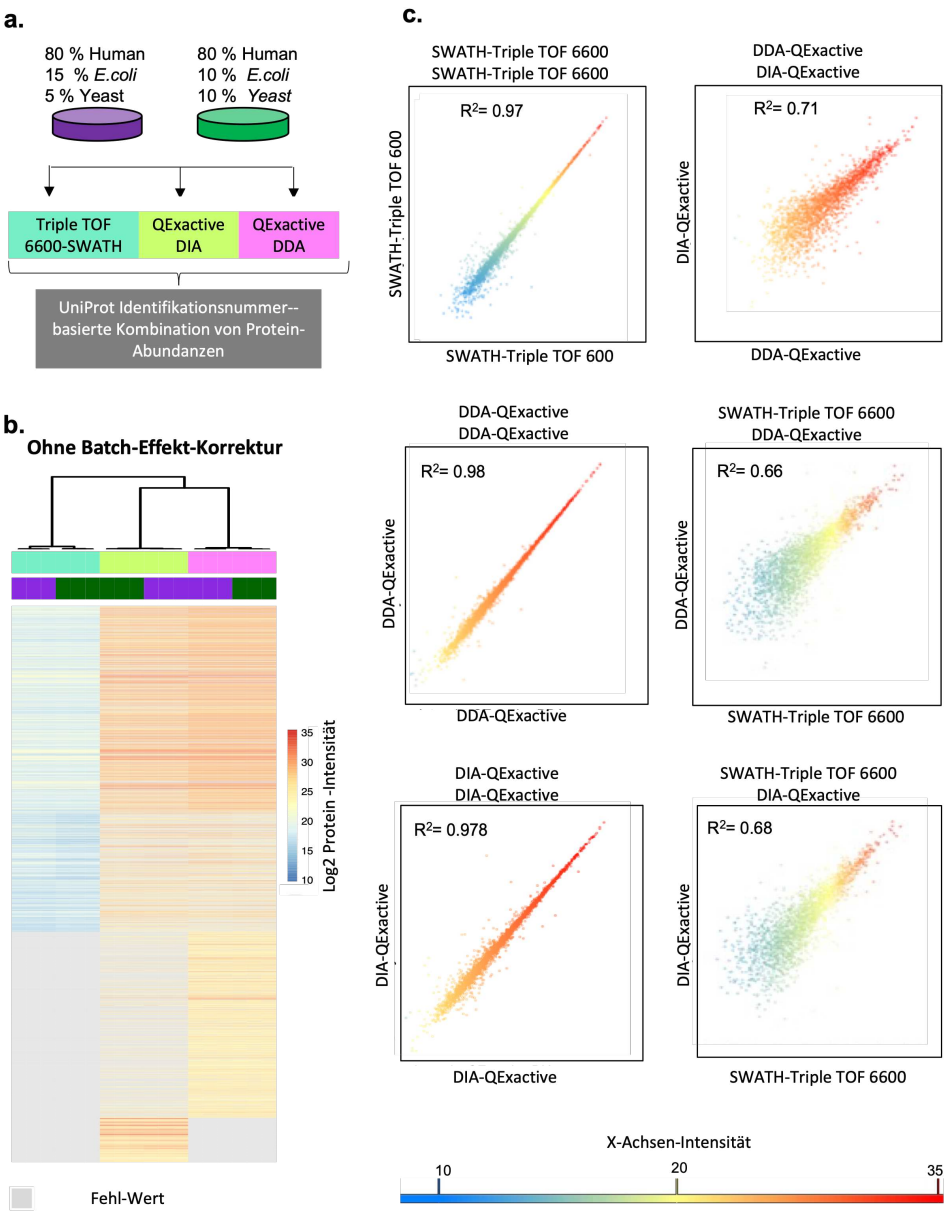

Abbildung 3.2: Einfluss der Integration unabhängig generierter Proteomdatensätze (unterschiedliche LC-MS-Konfigurationen) auf die Proteinabundanzverteilung zwischen definierten Phänotypen **a)** Schematische Darstellung des experimentellen Aufbaus.**b)** Heatmap-Visualisierung des Pearson-Korrelations-basierten hierarchischen Clustering nach UniProt-Identifikationsnummer-basierter Datenintegration, basierend auf 5543 identifizierten Proteinen.**c)**Streudiagramm-Visualisierung der Proteinabundanzen und linearer Korrelationskoeffizient von Phänotyp 1-Proben (lila), die mit identischen LC-MS-Konfigurationen gemessen wurden. **d)**Streudiagramm-Visualisierung der Proteinabundanzen und linearer Korrelationskoeffizient von Phänotyp 1-Proben die mit unterschiedlichen LC-MS-Konfigurationen gemessen wurden.

Durch nicht-biologische Faktoren induzierte Varianzen in biologischen Datensätzen werden als Batcheffekt bezeichnet und resultieren in der Verzerrung von Daten und Maskierung realer biologischer Effekte. Der auf einem empirischen Bayesian-Framework basierende ComBat-Algorithmus (21) sowie das lineare Regressionsmodell des Limma-Algorithmus (47) zur Entfernung von Batcheffekten sind mit der Datenstruktur von Proteomdatensätzen kompatibel. Im Gegensatz zu Limma integriert ComBat eine nicht-parametrische Methode zur Batcheffektreduktion , welche keine Gaußsche Normal - verteilung der Datenpunkte voraussetzt. Für den Spike-In-Datensatz (Abboldung 3.1. a) konnte eine trimodale Wahrscheinlichkeitsverteilung beobachtet werden. Aus diesem Grund wurde das nicht parametische Bayesian-Framework des ComBat-Algorithmus genutzt um die Anwendbarkeit etablierter Batcheffektkorrekturstrategien auf die Entfernung von Batcheffekten zwischen mit unterschiedlichen LC-MS-Konfigurationen aufgenommenen Proteomdaten zu testen. Bedingt durch die Inkompatibilität aller etablierten Batcheffektkorrekturstrategien mit Fehl-Werten des Missing not at Random (MNAR)-Typen, wurde die Daten-Matrix vor der Batcheffektkorrektur auf 1880 in allen experimentellen Konfigurationen identifizierte Proteine reduziert.

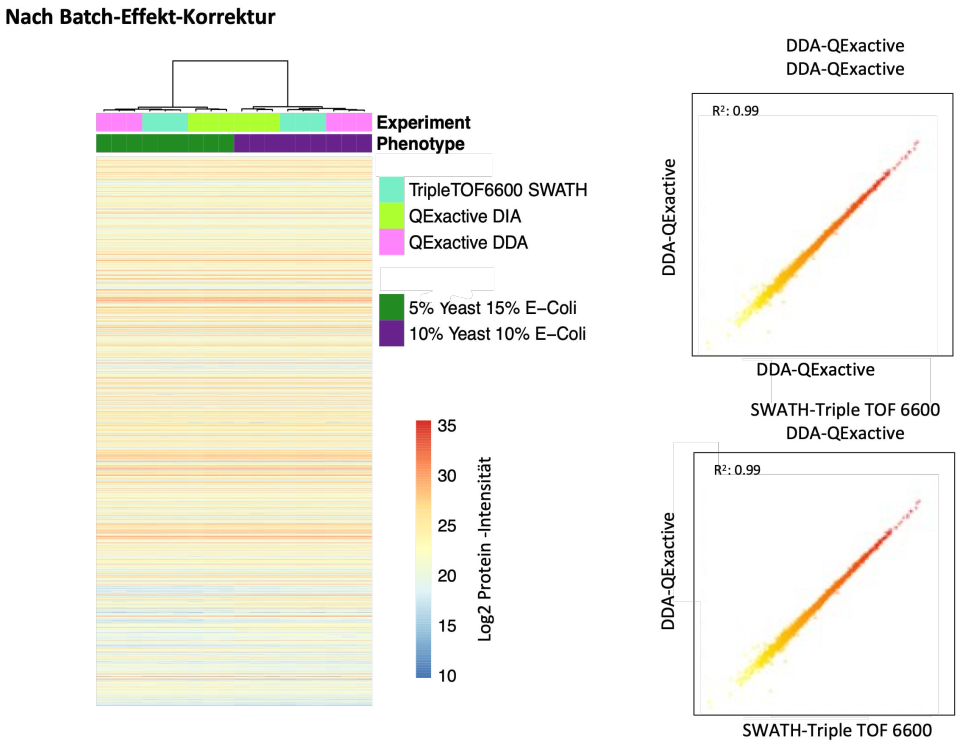

Abbildung 3.3: Anwendung des ComBat-Algorithmus zur Batcheffektreduktion zwischen technisch unabhängigen Proteomdatensätzen auf Basis von 1880 in allen Proben identifizierten Proteinen. Heatmap-Visualisierung des Pearson-Korrelations-basierten hierarchischen Clustering, nach UniProt-Identifikationsnummer-basierter Datenintegration; Streudiagramm-Visualisierung der Proteinabundanzen von Phänotyp 1-Proben (lila), identischer und unterschiedlicher LC-MS-Konfiguration mit annotiertem Pearson-Korrelationskoeffizienten nach Batcheffektkorrektur.

Nach der Batcheffektkorrektur clusterten Proben primär in Abhängigkeit des Phänotyps, derweil die LC-MS-Konfiguration das Clustering nur innerhalb eines Phänotyps beeinflusste. (Abbildung 3.3.) Konfigurationsgleiche Messungen von Phänotyp 1 zeigten eine konstante Korrelation von 99% (Abbildung 3.3.e) , derweil sich der Korrelationskoeffizient zwischen unterschiedlichen Setups auf 0.99 erhöhten (Abbildung 3.3.).

3.3 | Analyse gängiger Strategien zur Handhabung des Fehlwerttoleranz-Problems in der Batcheffektreduktion zwischen unabhängig generierten Proteomdatensätze

Die Anforderung, dass Datenmatrizen keine Fehlwerte des MNAR-Typs aufweisen dürfen, reduziert die Anzahl der verwendbaren Proteine auf 34% der ursprünglichen identifizierten 5543 Proteine für den Spike-in-Datensatz (Abbildung 3.1. a) Um die Auswirkung der Eliminierung von Proteinen, die in mindestens einem Batch nicht identifiziert wurden, auf unabhängig generierte, integrierte Datensätze näher zu untersuchen wurden die identifizierten Proteine pro Batch für alle verwendeten Datensätze (Abbildung 3.1.) verglichen (Abbildung 3.4.).

Für den Maus-Medulloblastom-Datensatz (Abbildung 3.1. b) wurde eine Überlappung von 28.4% zwischen allen betrachteten Gewebskonservierungstechniken und Analysezeitpunkten ermittelt (Abbildung 3. 4. a). Die gemeinsame Datenbanksuche von Rohdaten aller experimentellen Konfigurationen unter Nutzung des Minora-Algorithmus ((40) erhöhte die Übereinstimmung identifizierter Proteine zwischen allen Batches auf 75.8%.(Abbildung 3.4. b).

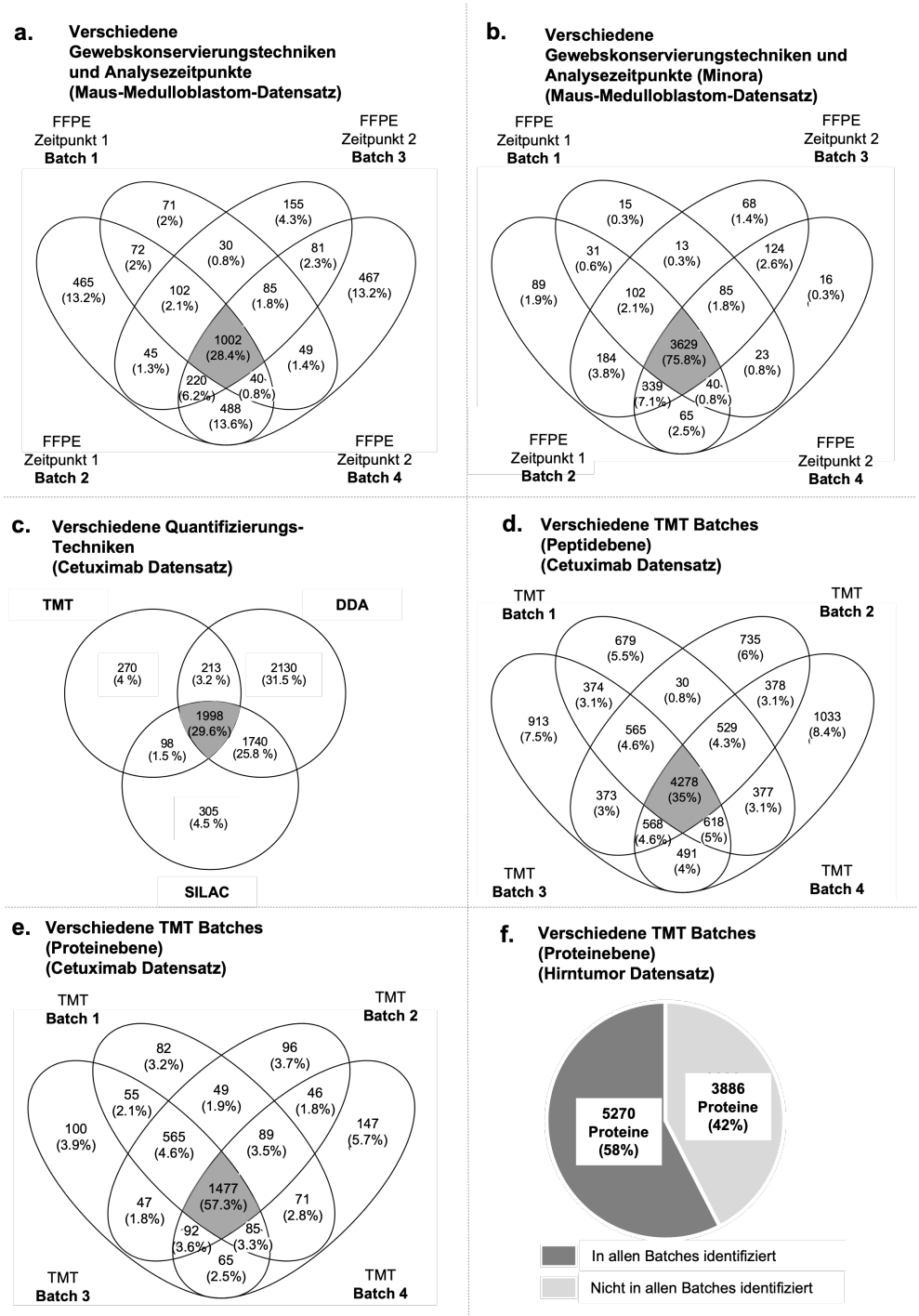

Abbildung 3.4: Venn-Diagramm-Visualisierung der Übereinstimmung identifizierter Proteine zwischen unabhängig generierten Proteomdatensätzen für **a)** Unterschiedliche Gewebskonservierungstechniken (FF; FFPE) und Analysezeitpunkte (Maus-Medulloblastom-Datensatzes auf Protein-ebene)**b)** Unterschiedliche Gewebskonservierungstechniken (FF; FFPE) und Analysezeitpunkte nach gemeinsamer Rohdatenprozessierung unter Verwendung des Minora-Algorithmus (Maus-Medulloblastom-Datensatz, Proteinebene)**c)** Unterschiedliche Quantifizierungstechniken (DDA, SILAC, TMT) (Cetuximab-Datensatz, Proteinebene) **d)** Unterschiedliche TMT-8-Plex-Batches (Cetuximab-Datensatz, Proteinebene) **e)** Unterschiedliche TMT-8-Plex-Batches (Cetuximab-Datensatz, Peptidebene) **f)**Unterschiedliche TMT-11-Plex-Batches (Hirntumor-Datensatz, Proteinebene.

Für den Cetuximab-Datensatz (Abbildung 3.1.c) wurden 29.5% aller Proteine mittels DDA-, TMT- und SILAC-Messungen identifiziert. Die größte Übereinstimmung wurde zwischen DDA- und SILAC- Messungen festgestellt. (55.4%). Den höchsten Anteil exklusiv identifizierter Proteine zeigten DDA-Messungen (31.5%) (Abbildung 3.4.c)). In Bezug auf TMT-Messungen des Cetuximab-Datensatzes (Abbildung 3.1.d,e) konnte auf Peptiebene eine Überlappung von 35% zwischen den identifizierten Proteinen aller TMT-8-Plex-Batches ermittelt werden (Abbildung 3. 4.d). Auf Proteinebene erhöhte sich dieser Anteil auf 57.3%(Abbilung 3.4.e)). Bei Analyse größerer TMT-Datensätze, wurden 58% aller identifizierten Proteine in allen 23 TMT-11-Plex-Batches des Hirntumor-Datensatzes (Abbildung 3.1.f) gefunden (Abbildung 3.4.f)).
Im Folgenden wurde die Auswirkung der Datenreduktion auf die Untersuchung biologisch relevanter Signalnetzwerke analysiert (Abbildung 3. 5.).

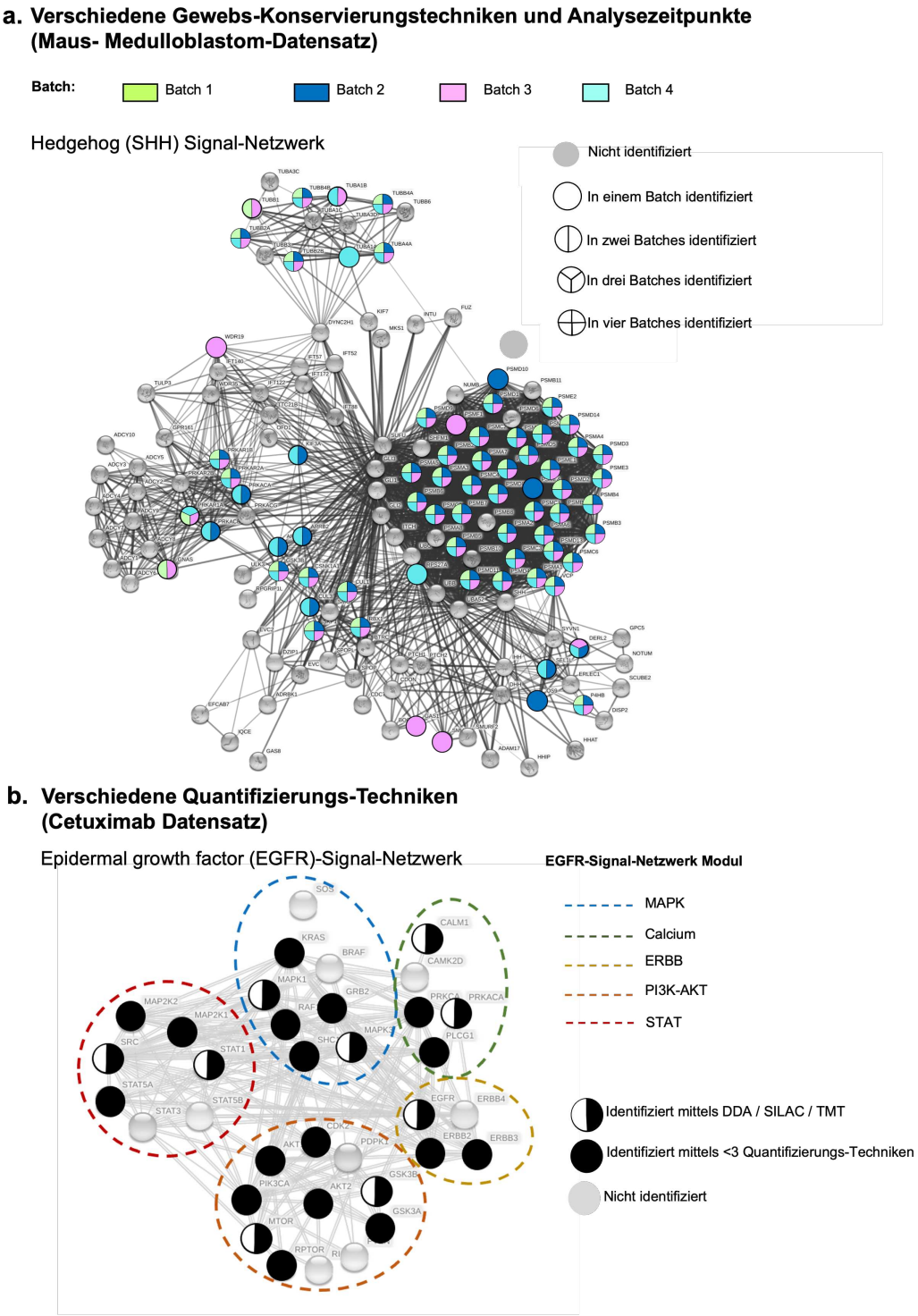

Abbildung 3.5: Auswirkung der Datenreduktion auf die Untersuchung biologisch relevanter Signalnetzwerke **a)** Signalnetzwerk-Abdeckung des SHH-Signalnetzwerkes anhand des Maus-Medulloblastom-Datensatzes. **b)** Signalnetzwerk-Abdeckung des EGFR-Signalnetzwerkes durch DDA-, TMT- und SILAC-Messungen des Cetuximab-Datensatzes.

Da eine Aktivierung des SHH-Signalweges das Wachstum von SHH-Typ -Medulloblastomen antreibt ((53), wurde die Abdeckung des SHH-Signalnetzwerks für den Maus-Medulloblastom-Datensatz vor und nach der Datenreduktion auf die in allen Batches identifizierte Proteine untersucht (Abbildung 3.6.a)). Nur 20 von 71 identifizierten Proteinen (28%) zeigten keine MNAR-Typ-Fehlwerte und konnten nach Datenreduktion zwecks Batcheffektkorrektur berücksichtigt werden.

Cetuximab bindet den Epidermal Growth Factor (EGFR) ((2). Aus diesem Grund wurde für die Messung des Cetuximab-Datensatzes mit verschiedenen Quantifizierungstechniken ein, von *Stepath et al. (2020)* definiertes , EGFR-Signalnetzwerk untersucht. Insgesamt konnten 26 von 34 (76%) Signalweg-assoziierten Faktoren mit mindestens einer Quantifizierungstechnik gefunden. Davon konnte 9 Proteine (35%) gleichzeitig in DDA- , Spike-in-SILAC- und TMT-Daten identifiziert und folglich in der Batcheffektkorrektur berücksichtigt werden.

Neben der Datenreduktion kann zur Lösung des Fehlwerttoleranz-Problems in der Batcheffektkorrektur integrierter, unabhängig generierter Datensätze die Ersetzung von fehlenden Werten durch künstliche Werte (Imputation) genutzt werden, um eine komplette Datenmatrix zu erzeugen. Um die Anwendbarkeit der Imputation fehlender Werte zur Lösung des Fehlwerttoleranz-Problems zu evaluieren, wurden verschiedene Imputationsstrategien für den Spike-In-Datensatz getestet: Die Imputation fehlender Werte aus der Normalverteilung (für jede individuelle Probe), Die Imputation fehlender Werte aus der Normalverteilung (für die gesamte Matrix) und die Random-Forest -Imputation. Nach Imputation wurden Batcheffekte zwischen unterschiedlichen LC-MS-Konfigurationen unter Nutzung des ComBat-Algorithmus (nicht-parametrisches Bayesian -Framework, L/S-Scaling) korrigiert (Abbildung 3.6.-9.).

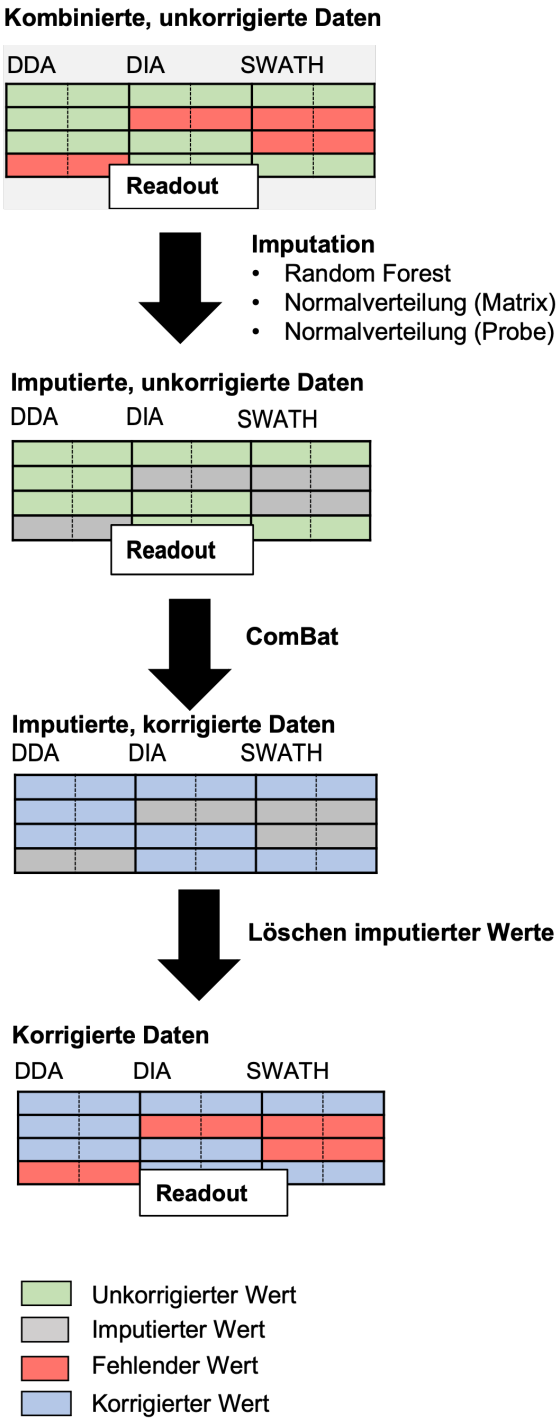

Abbildung 3.6: Schematische Darstellung der Analyse der Anwendbarkeit der Imputation fehlender Werte zur Lösung des Fehlwerttoleranz-Problems für die Batcheffektkorrektur zwischen unabhängig generierten Proteomdatensätzen.

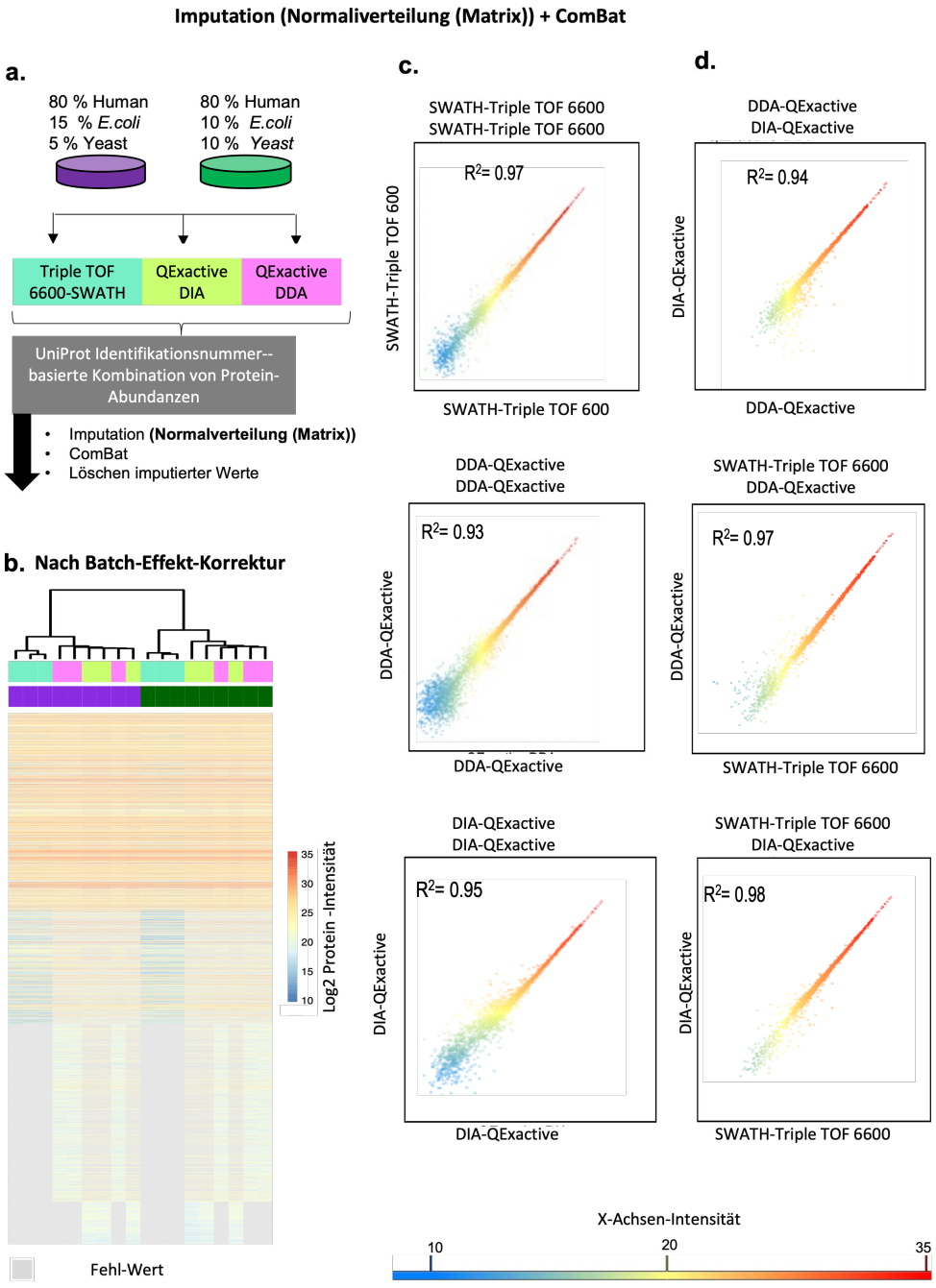

Abbildung 3.7: Evaluation des Einflusses der Matrix-Weise Imputation von Fehl-Werten über die Normalverteilung auf die Batcheffektkorrektur zwischen unabhängig generierten Proteomdatensätzen (Unterschiedliche LC-MS-Konfigurationen).**a)** Schematische Darstellung des experimentellen Aufbaus.**b)** Heatmap-Visualisierung des Pearson-Korrelations-basierten hierarchischen Clustering nach Batcheffektkorrektur. **c)**Streudiagramm-Visualisierung der Proteinabundanzen und linearer Korrelationskoeffizient von Phänotyp 1-Proben (lila), die mit identischen LC-MS- Konfigurationen gemessen wurden, nach Batcheffektkorrektur.**d)**Streudiagramm-Visualisierung der Proteinabundanzen und linearer Korrelationskoeffizient von Phänotyp 1 Proben die mit unterschiedlichen LC-MS-Konfigurationen gemessen wurden, nach Batcheffektkorrektur.

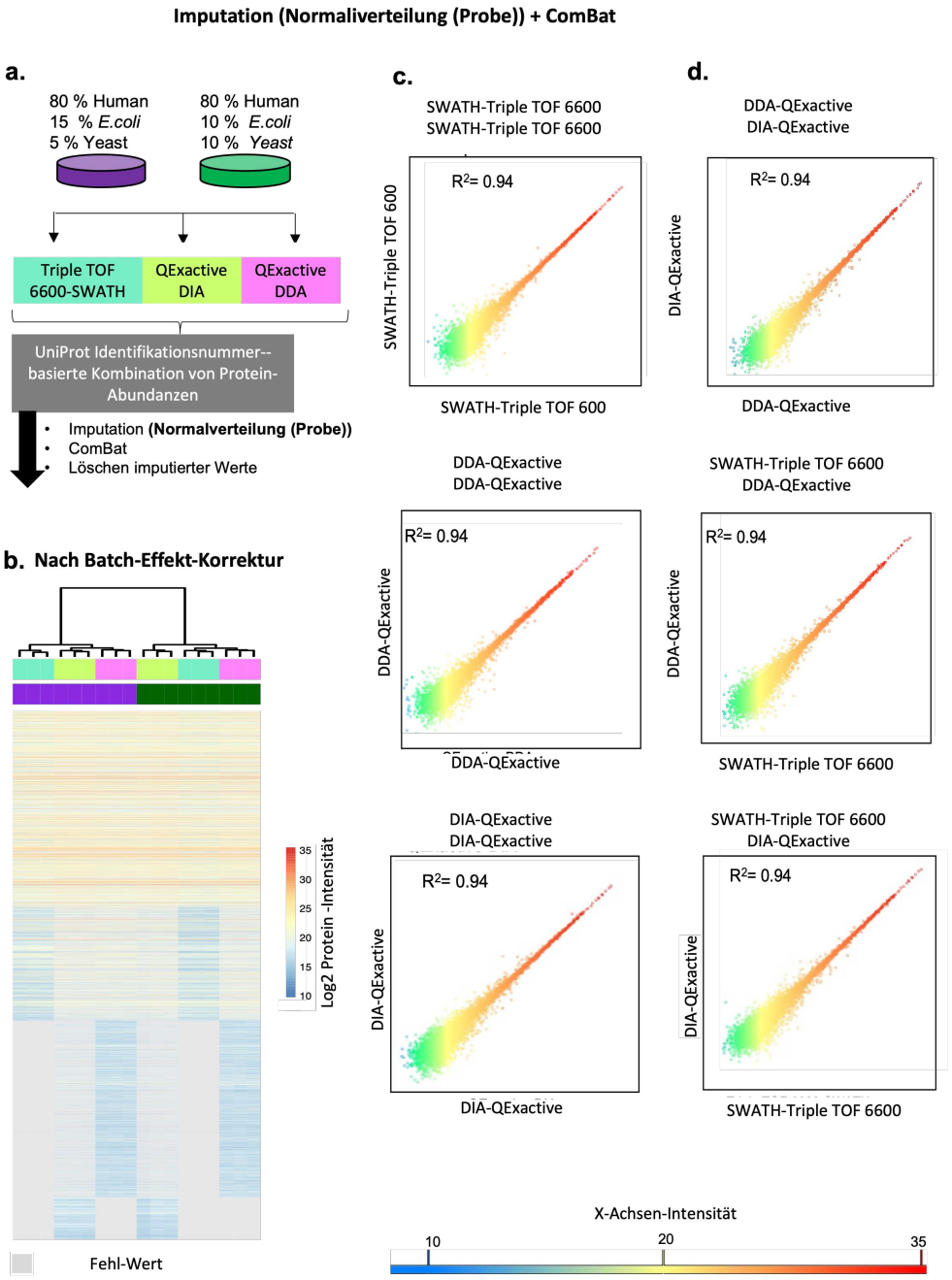

Abbildung 3.8: Evaluation des Einflusses der Proben-Weise Imputation von Fehl-Werten über die Normalverteilung auf die Batcheffektkorrektur zwischen unabhänig generierten Proteomdatensätzen (unterschiedliche LC-MS-Konfigurationen).**a)** Schematische Darstellung des experimentellen Aufbaus.**b)** Heatmap-Visualisierung des Pearson-Korrelations-basierten hierarchischen Clustering nach Batcheffektkorrektur. **c)**Streudiagramm-Visualisierung der Proteinabundanzen und linearer Korrelationskoeffizient von Phänotyp 1-Proben (lila), die mit identischen LC-MS-Konfigurationen gemessen wurden, nach Batcheffektkorrektur.**d)**Streudiagramm-Visualisierung der Proteinabundanzen und linearer Korrelationskoeffizient von Phänotyp 1 Proben die mit unterschiedlichen LC-MS-Konfigurationen gemessen wurden, nach Batcheffektkorrektur.

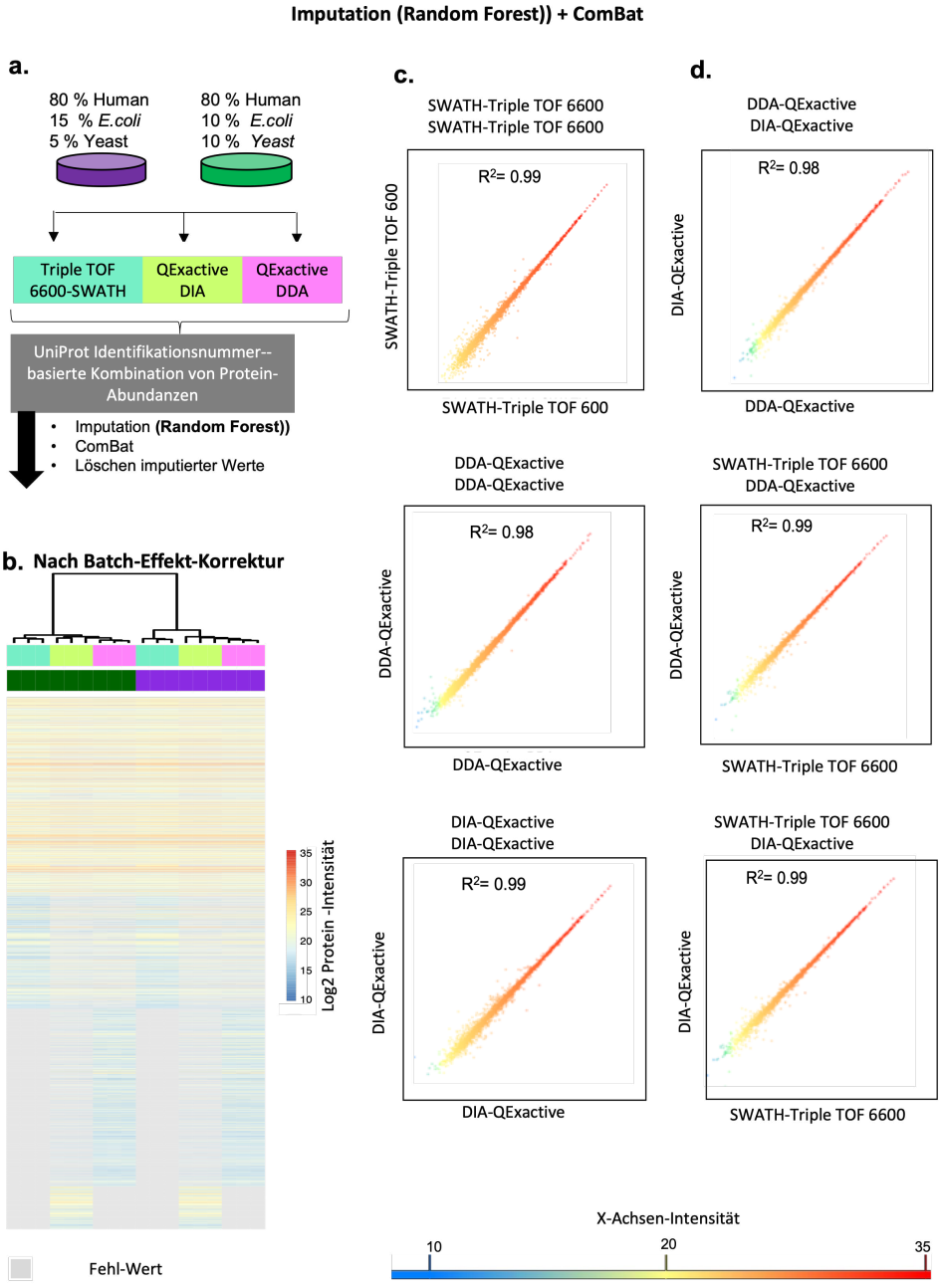

Abbildung 3.9: Evaluation des Einflusses der Random-Forest-Imputation von Fehlwerten auf die Batcheffektkorrektur vob unabhängig generierten Proteomdatensätzen (unterschiedliche LC-MS-Konfigurationen).**a)** Schematische Darstellung des experimentellen Aufbaus.**b)** Heatmap Visualisierung des Pearson-Korrelations-basierten hierarchischen Clustering nach Batcheffektkorrektur. **c)**Streudiagramm-Visualisierung der Proteinabundanzen und linearer Korrelationskoeffizient von Phänotyp 1-Proben (lila), die mit identischen LC-MS-Konfigurationen gemessen wurden, nach Batcheffektkorrektur.**d)**Streudiagramm-Visualisierung der Proteinabundanzen und des linearen Korrelationskoeffizienten von Phänotyp 1-Proben die mit unterschiedlichen LC-MS-Konfigurationen gemessen wurden, nach Batcheffektkorrektur.

Alle getesteten Methoden zeigten nach der Imputation fehlender Werte, gefolgt von einer Batcheffektkorrektur durch das empirische Byesian-Framework, ein Clustering in Abhängigkeit des Phänotypen (Abbildung 3.7-9.b). Nach der Imputation fehlender Werte aus der Normalverteilung (Matrix) erhöhte sich die Korrelation zwischen mit unterschiedlichen LC-MS-Konfigurationen vermessenen Proben des Phänotypen 1 auf 94-98 % (Abbildung 3.7.d)). Gleichzeitig reduzierte sich die Korrelation zwischen QExactive-Messungen identischer experimenteller Setups um 2-5%.(Abbildung 3.7.c)). Für Proben, die via SWATH-Messungen auf dem TripleTOF 6600 gemessen wurden konnte keine Änderung des Korrelationskoeffizienten zwischen unkorrigierten (Abbildung 3.1.c)) und Batcheffekt korrigierten Daten festgestellt werden (Abbildung 3.7.c)). Die Imputation fehlender Werte aus der Normalverteilung (Probe) (Abbildung 3.9.a) resultierte in einer konstanten Korrelation von 94 % innerhalb und zwischen allen LC-MS -Konfigurationen (Abbildung 3.7.c, d). Für die Random-Forest-Imputation konnte innerhalb und zwischen allen LC-MS-Konfigurationen ein Korrelationskoeffizient zwischen 0.98 und 0.99 ermittelt werden (Abbildung 3.9.c, d).

3.4 | Matrix-Dissektionsverfahren als Alternative zur Handhabung des Fehlwerttoleranz-Problems in der Batcheffektkorrektur zwischen integrierten, unabhängig generierten Proteomdatensätzen.

Als eine von der Datenreduktion und dem Einführen artifizieller Werte unabhängige Lösung des Fehlwerttoleranz-Problems wurde in dieser Arbeit das Matrix-Dissektions -verfahren zur Batcheffektkorrektur zwischen integrierten, unabhängige generierten Proteomdatensätzen, entwickelt. Eine exemplarische, algorithmische Implementierung dieses Prinzips nutzt das Matrix-Dissektionsverfahren als Framework für die Fehlwert-tolerante Nutzung des ComBat-Algorithmus und Limma-Algorithmus zur Reduktion von Batcheffekten. (Abbildung 3.10.). Eine integrierte Eingabematrix wird durch UniProt-Identifik -ationsnummer-basierte Kombination log2-transformierter, probenspezifischer Protein -abundanzen aus unabhängig generierten Datensätzen erzeugt (integrierte Datenmatrix). Zu Beginn durchsucht der Algorithmus die Eingabematrix nach fehlenden Werten. Werden für ein Protein Werte <2 in einem Batch identifiziert, so wird es als MNAR-Typ Fehlwert in dem betreffenden Batch klassifiziert. Im Folgenden werden auf Grundlage der Batch-Verteilung einzelner Proteine, Untermatrizen generiert, welche keine MNAR-Typ Fehlwerte enthalten. Missing at Random (MAR)-Typ Fehlwerte werden toleriert. Basierend auf der Wahrscheinlichkeitsverteilung und Varianz (probenspezifischer Mittelwert, probenspezifischer Varianzkoeffizient) der Daten kann der Benutzer zwischen dem linearen Regressionsmodell (Funktion *RemoveBatcheffects()* -in Limma oder dem empirischen Bayesian-Framework (ComBat) wählen. Das empirische Bayesian-Frame -work kann auf parametrischer und nicht-parametrischer Grundlage genutzt werden. Beide Implementationen unterstützen die Nutzung des Model-basierten „location and scale adjustment“ (L/S-scaling), sowie einer rein Mittelwert-basierten

Batcheffektkorrektur (21).

Das gewählte Batcheffektkorrektur-Verfahren wird individuell auf jede Untermatrix angewandt. In nur einem Batch identifizierte Proteine werden keiner Korrektur unterzogen. Korrigierte Untermatrizen sowie unkorrigierte Untermatrizen für batchspezifische Proteine werden rekombiniert. Es entsteht eine Batcheffekt-korrigierte Ausgabematrix, welche für die weitere Datenanalyse genutzt werden kann. MAR- und MNAR-Typ Fehlwerte bleiben erhalten und werden nicht verändert.

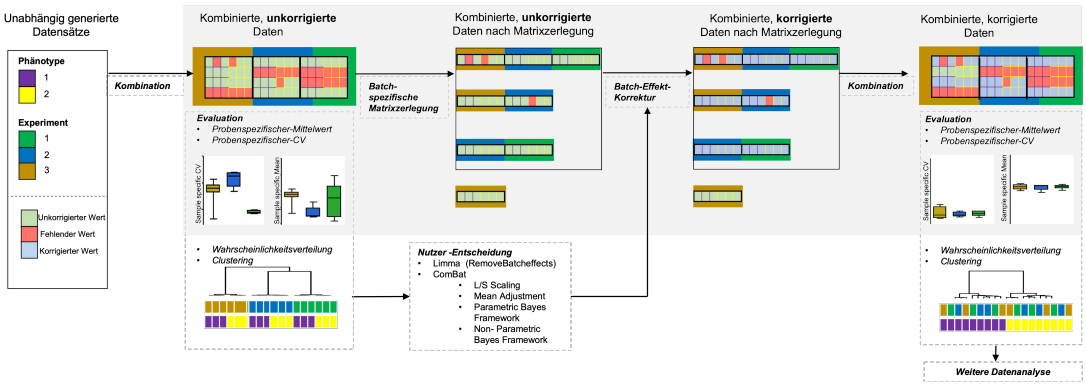

Abbildung 3.10: Schematische Darstellung des Matrix-Dissektionsverfahrens zur Lösung des Fehlwerttoleranz-Problems in der Batcheffektkorrektur zwischen unabhängig generierten Proteomdatensätzen für die Anwendung des ComBat-Algorithmus und der *RemoveBatcheffects()*-Fuktion in Limma.

Um die Eignung des Matrix-Dissektionsverfahrens für Batcheffektkorrektur zwischen unabhängigen Proteomstudien zu evaluieren, wurde es exemplarisch für den Spike-in-Datensatz (Abbildung 3.1.a, 3.12.a) unter Verwendung des ComBat-Algorithmus (L/S-Scaling, Nicht-parametrisches Bayesian-Framework) getestet. Viele Matrixalgebra -basierte Verfahren zur statistischen Folgeanalyse wie die lineare Hauptkomponenten-Analyse (PCA) benötigen komplette Daten-Matrizen ohne Fehlwerte (49). Aus diesem Grund wurde im Folgenden die Auswirkung einer Random-Forest-Imputation fehlender Werte auf den mittels Matrix-Dissektionsverfahren Batcheffekt-korrigierten Daten-

satz analysiert (Abbildung 3.11-3.13).

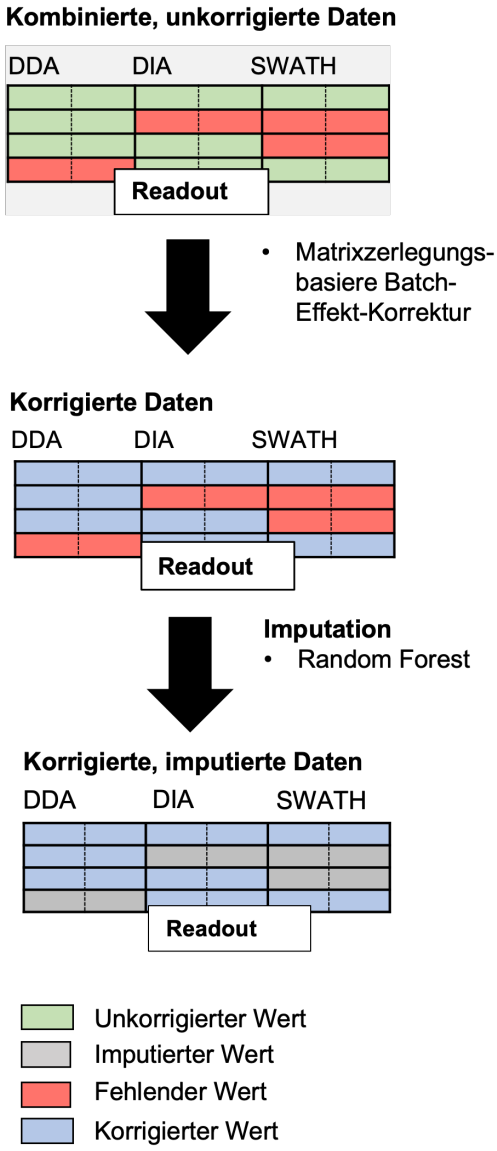

Abbildung 3.11: Schematische Darstellung der Analyse der Eignung des Matrix-Dissektionsverfahrens zur Handhabung des Fehlwerttoleranz-Problems in der Batcheffektkorrektur zwischen unabhängigen Proteomstudien.

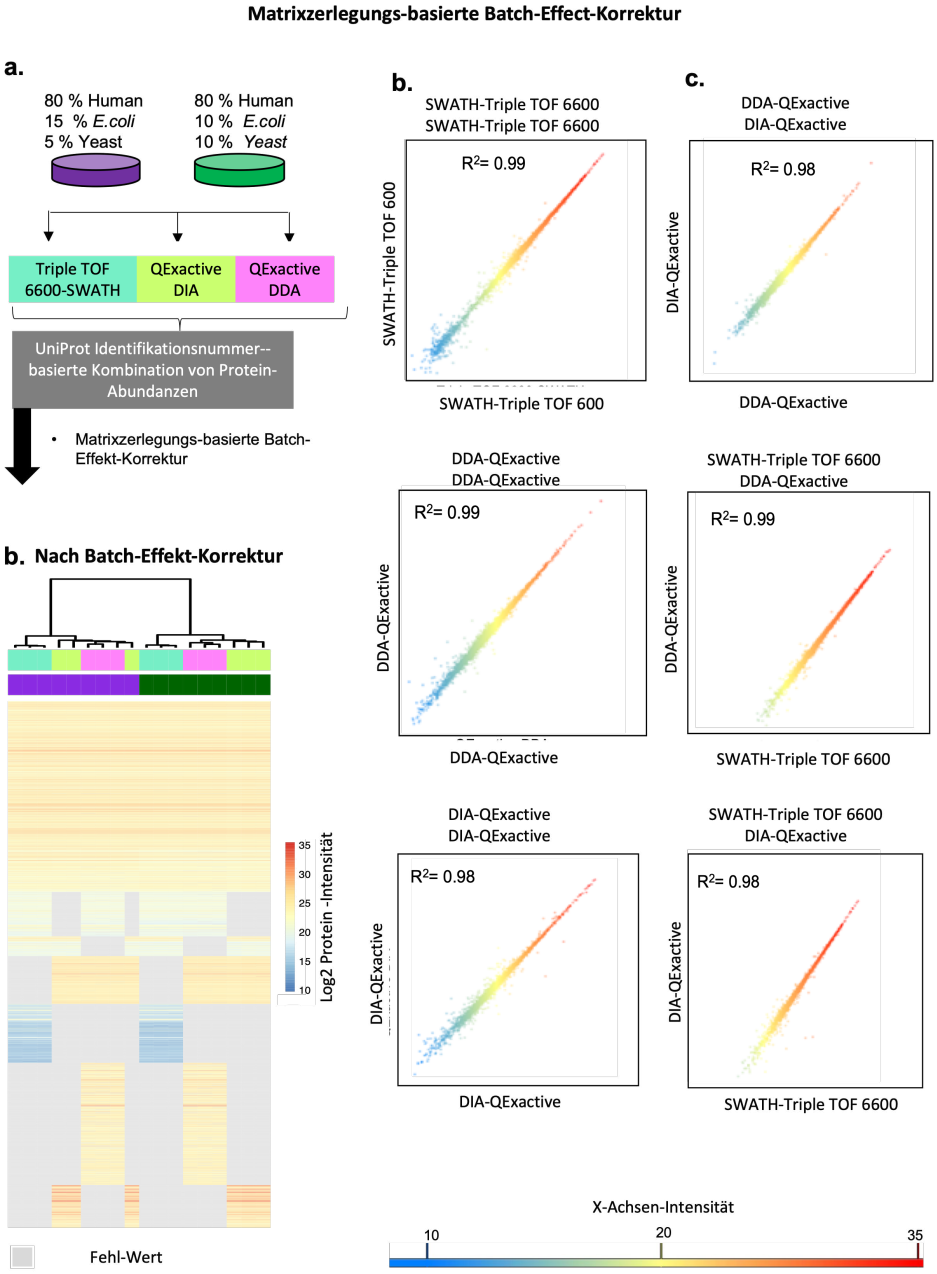

Abbildung 3.12: Evaluation des Matrix-Dissektionsverfahren als Alternative zur Handhabung des Fehlwerttoleranz-Problems in der Batcheffektkorrektur zwischen integrierten, unabhängig generierten Proteomdatensätzen (unterschiedliche LC-MS-Konfigurationen).**a)** Schematische Darstellung des experimentellen Aufbaus.**b)** Heatmap-Visualisierung des Pearson-Korrelations-basierten hierarchischen Clustering nach Batcheffektkorrektur durch das Matrix-Dissektionsverfahren, auf Grundlage aller 5543 identifizierten Proteinen **c)**Streudiagramm-Visualisierung der Proteinabundanzen und linearer Korrelationskoeffizient von Phänotyp 1-Proben (lila), die mit identischen LC-MS/MS-Konfigurationen gemessen wurden. **d)**Streudiagramm-Visualisierung der Proteinabundanzen und linearer Korrelationskoeffizient von Phänotyp 1-Proben die mit unterschiedlichen LC-MS-Konfigurationen gemessen wurden.

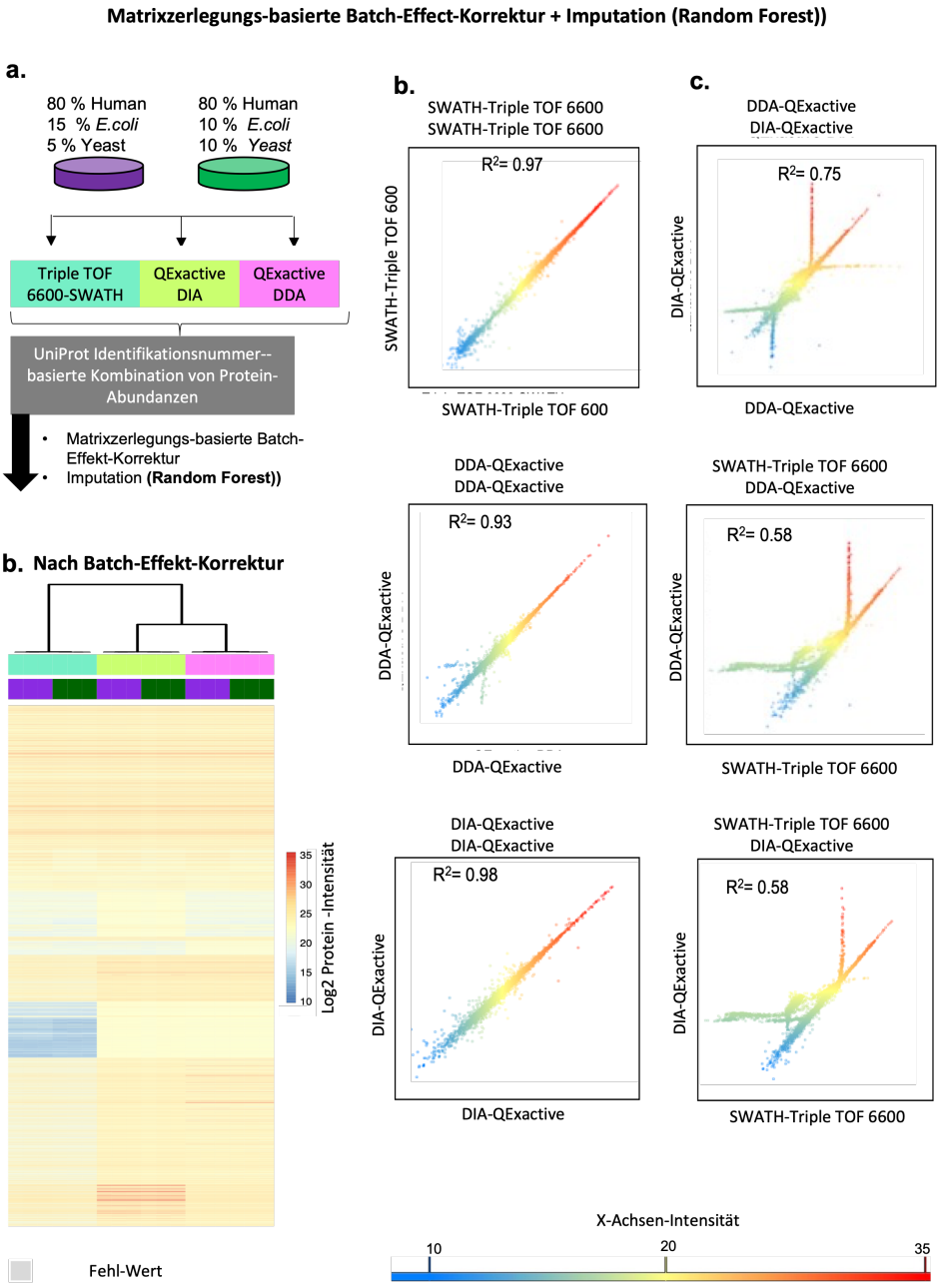

Abbildung 3.13: Evaluation des Einflusses der Random-Forest-Imputation auf integrierte, mittels Matrix-Dissektionsverfahren batcheffektkorrigierte, Proteomdatensätze (Unterschiedliche LC-MS-Konfigurationen).**a)** Schematische Darstellung des experimentellen Aufbaus.**b)** Heatmap-Visualisierung des Pearson-Korrelations-basierten hierarchischen Clustering nach Batcheffektkorrektur durch das Matrix-Dissektionsverfahren auf Grundlage aller 5543 identifizierten Proteinen und folgender Random-Forest-Imputation fehlender Werte. **c)**Streudiagramm-Visualisierung der Proteinabundanzen und linearer Korrelationskoeffizient von Phänotyp 1-Proben (lila), die mit identischen LC-MS-Konfigurationen gemessen wurden. **d)**Streudiagramm-Visualisierung der Proteinabundanzen und linearer Korrelationskoeffizient von Phänotyp 1-Proben die mit unterschiedlichen LC-MS-Konfigurationen gemessen wurden.

Nach Anwendung der Matrix-Dissektions-basierten Batcheffektkorrektur zeigte sich ein Phänotypen-basiertes Clustering. Ein Clustering in Abhängigkeit der LC-MS -Konfiguration wurde sekundär innerhalb phänotypenspezifischer Cluster beobachtet. (Abbildung 3. 13.b). Zwischen Phänotyp 1-Proben aller LC-MS-Konfigurationen wurde nach Batcheffektkorrektur eine Korrelation von 98-99% ermittelt (Abbildung 3.12.c,d). Nach zusätzlicher Random-Forest-Imputation zur Generierung einer kompletten Matrix, für batcheffektkorrigierte Daten, clusterten Proben in Abhängigkeit des LC-MS-Setups. Phänotypenspezifische Unterschiede wurde sekundär innerhalb konfigurations -spezifischer Cluster beobachtet (Abbildung 3.13.a). Bei Gegenüberstellung einzelner Phänotyp 1-Proben konnte nach der Random-Forest-Imputation kein lineares Verhalten zwischen Proteinabundanzverteilungen unterschiedlicher LC-MS -Konfigurationen mehr beobachtet werden. Der Korrelationskoeffizient zwischen Proben identischen LC-MS-Setups reduzierte sich auf 0.93-0.98, derweil zwischen verschiedenen LC-MS-Konfigurationen die Pearson Korrelation auf 58-75% sank.
Um das Matrix-Dissektionsverfahren näher mit der Imputation fehlender Werte vor Batcheffektkorrektur zu vergleichen, wurden statistisch signifikant differentiell abundante Proteine zwischen Phänotyp 1 und 2 des Spike-in-Datensatzes ermittelt. Für jede individuelle LC-MS-Konfiguration sowie für den integrierten Datensatz vor und nach der Batcheffektkorrektur wurden Proteine als signifikant verändert klassifiziert, wenn sie einen P-Wert < 0.05 zwischen beiden Konditionen aufwiesen (Abbildung 3.14.-3.15.).
Für alle individuellen LC-MS-Konfigurationen konnte vor der Batcheffektkorrektur eine höhere Abundanz von *Saccaromyces cerevisea* Proteinen in Phänotyp 1-Proben ermittelt werden. Phänotyp 2 zeigte eine höhere Abundanz von *E. Coli* Proteinen. Für jede LC-MS-Konfiguration wurde ein negativer dekadischer Logarithmus des P-Wertes von maximal 7 ermittelt (Abbildung 3.14. a). Für DDA-QExactive-Messungen wurden 20% (823 Proteine) aller identifizierten Proteine als signifikant differentiell abundant zwischen Phänotyp 1 und 2 klassifiziert. Signifikante SWATH-TripleTOF 6600-Messungen zeigten den höchsten Anteil differentieller Proteine (798 Proteine, 25.3%). Für DIA

-QExactive-Messungen wurden 349 Proteine als statistisch signifikant verändert klassifiziert (11.7%).

Trotz des Anstiegs der berücksichtigten Proteine auf 5543 Kandidaten, nach Datenintegration, reduzierte sich die Zahl signifikanter Proteine ohne Batcheffektkorrektur auf 369 Kandidaten (6.6%, Abbildung 3.15. a). Nach der Batcheffektkorrektur erhöhte sich die Zahl signifikant differentiell abundanter Proteine für alle getesteten Verfahren auf 19.3% -19.8%. Nach Anwendung Matrix-Dissektionsverfahren zur Batcheffektkorrektur wurde 1094 Proteine als differentiell klassifiziert (19.7%) (Abbildung 3.15. a)).

Bei Anwendung der Random-Forest-Imputation nach Matrix-Dissektions-basierter Batcheffektkorrektur reduzierte sich die Zahl differentieller Proteine, im Vergleich zur alleinigen Anwendung des Matrix-Dissektions-Verfahrens um 10.4% auf 515 Kandidaten (9.3%) (Abbildung 3.14. c, Abbildung 3.15. a). In Bezug auf den negativen dekadischen Logarithmus des P-Wertes konnte für alle getesteten Verfahren eine Erhöhung im Vergleich zu individuellen LC-MS-Setups und integrierten, unkorrigierten Daten beobachtet werden.

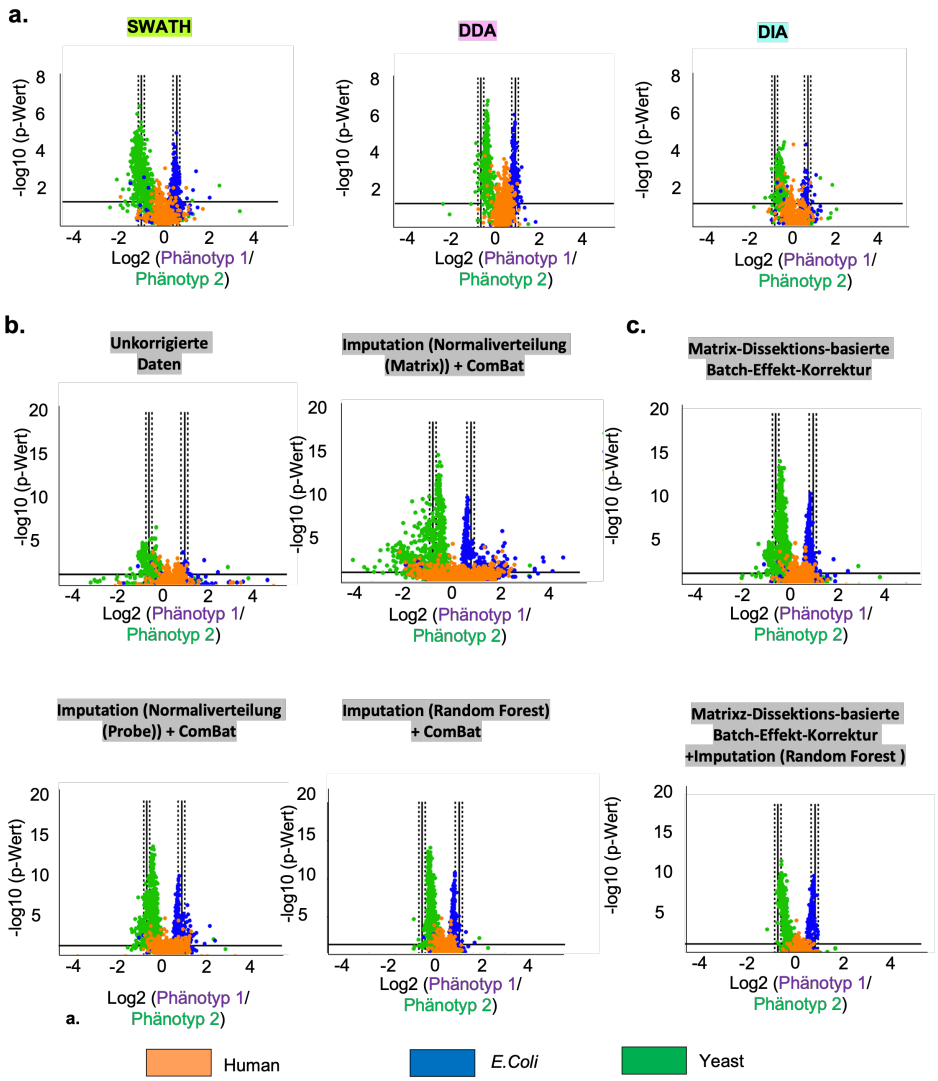

Abbildung 3.14: Visualisierung T-test-signifikanter Proteine (P-Wert < 0.05) zwischen Phänotyp 1 (80 % *Homo sapiens*, 15 % *E. coli*, 5 % *Saccaromyces cerevisea*; und Phänotyp 2 (80 % *Homo sapiens*, 10 % *E. coli*, 10 % *Saccaromyces cerevisea*, für den Spike-in-Datensatz.**a)**Für alle innerhalb einer LC-MS/MS-Konfiguration identifizierten Proteine vor der Batcheffektkorrektur (SWATH-TripleTOF 6600, QExactive-DIA, QExactive-DDA) **v)**Für integrierte Daten vor der Batcheffektkorrektur und nach Imputations-basierter Batcheffektkorrektur **c)**Für integrierte Daten nach der Batcheffektkorrektur durch das Matrix-Dissektions-Verfahren vor und nach Random-Forest Imputation.

a.

	Unkorrigierte Daten	Matrix-Dissektions-basierte Batch-Effekt-Korrektur	Imputation (Normaliverteilung (Matrix)) + ComBat	Imputation (Normaliverteilung (Probe)) + ComBat	Imputation (Random Forest) + ComBat	Matrixz-Dissektions-basierte Batch-Effekt-Korrektur +Imputation (Random Forest
Kombiniert	369	1094	1076	1101	1069	515
SWATH	798	798	698	709	700	1615
DIA	349	349	307	307	307	350
DDA	823	822	601	611	603	808

b.

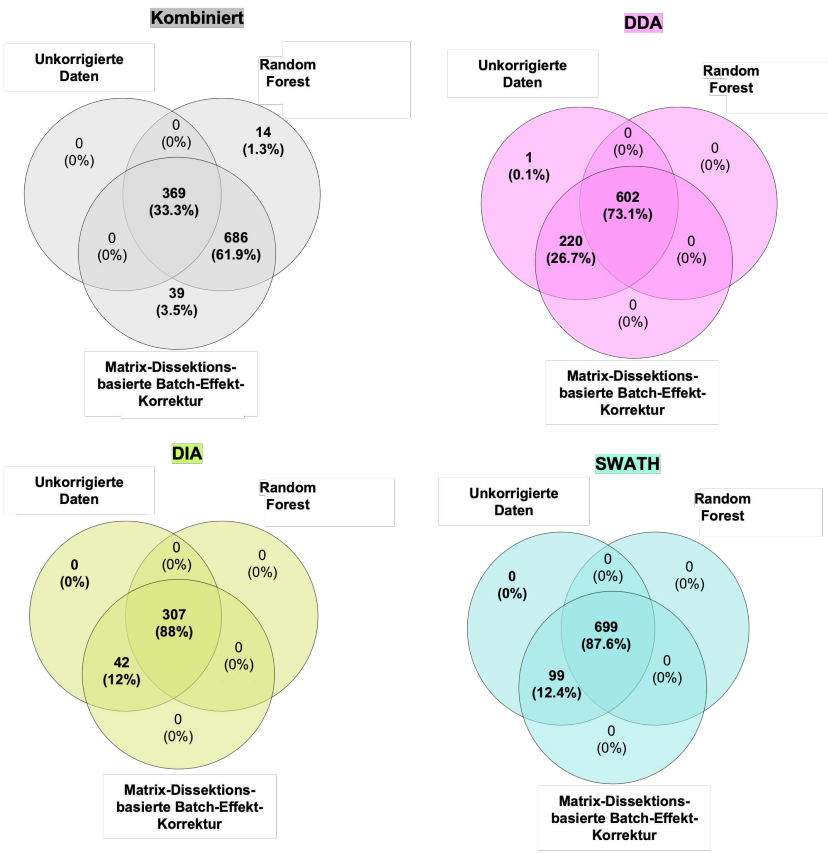

Abbildung 3.15: **a)**Zahl T-Test-signifikanter Proteine (P-Wert <0.05) zwischen Phänotyp 1 (80 % *Homo sapiens*, 15 % *E. Coli*, 5 % *Saccaromyces cerevisea* ; und Phänotyp 2 (80 % *Homo sapiens*, 10 % *E. Coli*, 10 % *Saccaromyces cerevisea* für den Spike-In-Datensatz, vor und nach der Anwendung aller getesteten Batcheffektkorrektur-Verfahren für den integrierten Datensatz und innerhalb jeder LC-MS-Konfiguration.**b)** Vergleich T-Test signifikanter Proteine zwischen Phänotyp 1 und 2 vor und nach der Batcheffektkorrektur mittels Matrix-Dissektionsverfahren und nach Random-Forest-Imputation, für den integrierten Datensatz und innerhalb jeder LC-MS-Konfiguration (SWATH-TripleTOF 6600, QExactive-DIA, QExactive-DDA)

Die Zahl T-Test signifikanter Proteine (P-Wert < 0.05) zwischen Phänotyp 1 und 2 innerhalb einzelner LC-MS-Konfigurationen blieb nur für das Matrix-Dissektionsverfahren vor und nach der Batcheffektkorrektur konstant. Für alle imputationsbasierten Verfah-

ren verringerte sich die Zahl signifikanter Proteine pro Setup (Abbildung 3.15. a) nach Batcheffektkorrektur. Für die Random-Forest-Imputation nach Matrix-Dissektions -basierter Batcheffektkorrektur erhöhte sich die Zahl signifikant differentiell abundanter Proteine für SWATH-TripleTOF 6600- und DIA-QExactive-Messungen im Vergleich zu unkorrigierten Daten auf 1615 (SWATH) und 350 (DIA). Für DDA-QExactive-Messungen reduzierte sich die Zahl signifikanter Proteine auf 808.

Da für die Random-Forest-Imputation vor Batcheffektkorrektur sowie das Matrix -Dissektionsverfahren die höchste Korrelation zwischen Phänotyp 1-Proben nach der Batcheffektkorrektur beobachtet werden konnte (Abbildung 3.9., 3.12.), wurden die statistisch signifikant differentiell abundanten Proteine für beide Verfahren vor und nach der Batcheffektkorrektur im Folgenden näher untersucht. Für den integrierten Datensatz konnte nach Batcheffektreduktion zwischen beiden Verfahren eine Überschneidung von 95.2% festgestellt werden. Alle im unkorrigierten Datensatz als P-Wert signifikant (P-Wert < 0.05) identifizierten Proteine wurden nach Batcheffektkorrektur unter Verwendung beider Verfahren als signifikant klassifiziert. 39 Proteine wurden ausschließlich nach Anwendung des Matrix-Dissektionsverfahrens als statistisch signifikant differentiell abundant eingeordnet.
Bei einzelner Betrachtung jeder LC-MS-Konfiguration konnte für alle Setups eine 100% ige Überschneidung zwischen unkorrigierten Daten und mittels Matrix-Dissektions -verfahren korrigierten Werten ermittelt werden. Nach Random-Forest-basierter Anwendung des ComBat-Algorithmus wurden 12-26% der zuvor als differentiell klassifizierten Proteine innerhalb einzelner LC-MS-Konfigurationen kein signifikanter P-Wert < 0.05 mehr zugeordnet (Abbidung 3.16. b).

3.5 | Matrix-Dissektionsverfahren zur Korrektur von Batcheffekten zwischen verschiedenen Gewebskonservierungstechniken und Analysezeitpunkten

Formalin-Fixiertes-Parrafin eingebettetes (FFPE-) Gewebe ist ein weltweiter Standard für die Konservierung und Lagerung von Geweben, stellt jedoch aufgrund der unvollständigen Reversion von Methylenbrücken und der Induktion irreversibler chemischer Modifikationen eine Herausforderung für die massenspektrometrische Analyse dar. Frischgewebe (FF) zeigt eine höhere Kompatibilität mit LC-MS-Analysen. Die Analyse größerer, statistisch valider Kohorten ist jedoch durch die geringe Verfügbarkeit und Haltbarkeit von FF-Proben limitiert (33). Die Verwendung unterschiedlicher Gewebskonservierungstypen in unabhängig generierten Proteomstudien resultiert nach Datenintegration in der Induktion von Batcheffekten.

Um die Anwendbarkeit des Matrix-Dissektionsverfahrens auf die Reduktion von Batcheffekten zwischen verschiedenen Gewebskonservierungtechniken zu evaluieren, wurde der Maus-Medulloblastom-Datensatz (Abbildung 3.1. b) verwendet. Im Framework des Matrix-Dissektionsverfahrens, wurde zu diesem Zweck der der ComBat-Algorithmus (L/S-Scaling, parametrisches Bayesian-Framework) und das lineare Regressionsmodell der *RemoveBatcheffects()*-Funktion des Limma-Algorithmus verglichen. (Abbildung 3.17.-3.19.). Des Weiteren wurden anhand des Datensatzes die Matrix-Dissektionsver
-fahren-bedingte Batcheffektkorrektur zwischen unterschiedlichen Aanlysezeitpunkten bei identischer Probenvorbereitung und LC-MS/MS-Messung evaluiert.

Vor der Batcheffektkorrektur konnte ein klares Clustering in Abhängigkeit der verwendeten Gewebskonservierungstypen beobachtet werden. Innerhalb eines jeden Batches clusterten Proben in Abhängigkeit des Analysezeitpunktes. Nur innerhalb jedes Ge-

webstypen und Analysezeitpunktes konnte ein Clustering in Abhängigkeit des Phänotyps (Tumor, Kontrolle) beobachtet werden. Für FFPE-basierte Messungen konnte eine höhere Variabilität zwischen probenspezifischen Varianzkoeffizienten im Vergleich zu FF-Proben ermittelt werden. Trotz identischer Injektionsmenge (1 Mikrogram Probe/Messung), zeigten FF-basierte Messungen einen höheren probenspezifischen Mittelwert (Abbildung 3.17. a).

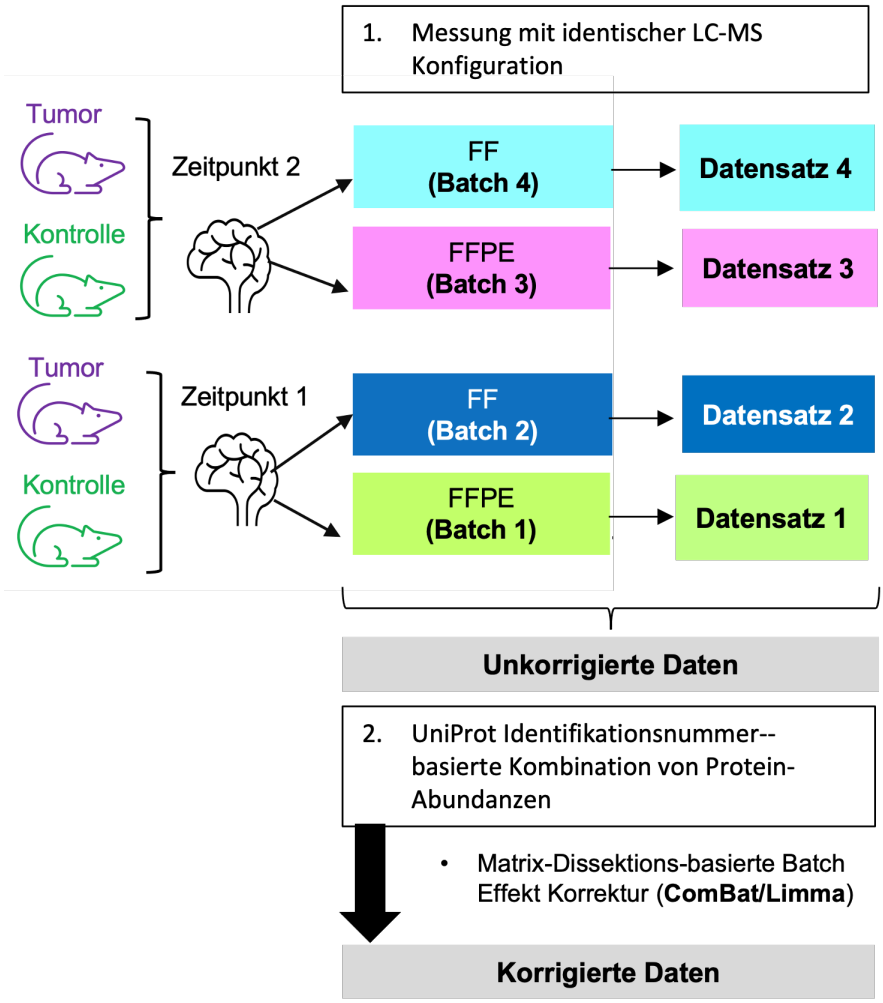

Abbildung 3.16: Schematische Darstellung der Evaluation der Eignung des Matrix-Dissektionsverfahrens zur Korrektur von Batcheffekten zwischen verschiedenen Gewebs-Konservierungstechniken und Analysezeitpunkten unter Verwendung des ComBat-Algorithmus und der *RemoveBatcheffects()*-Funktion des Limma-Algorithmus.

Nach Anwendung des Matrix-Dissektionsverfahrens konnte unter Nutzung beider Algorithmen ein dominantes Clustering in Abhängigkeit der Phänotypen beobachtet werden. Des Weiteren zeigte sich keine Clustertendenz in Abhängigkeit des Gewebskonservierungstyps oder Analysezeitpunktes. Für alle Batches wurde nach Batcheffektkorrektur ein Angleichen des probenspezifischen Mittelwertes festgestellt. Des Weiteren erhöhte sich die Variabilität Proben-spezifischer Varianzen für FF-Batches, derweil diese sich für FFPE-Batches reduzierten (Abbildung 3.17 b),c)). Im Hinblick auf die Zahl regulierter Proteine zwischen Tumor und Kontrolle wurden vor der Batcheffektkorrektur im integrierten Datensatz 713 t-Test signifikante Proteine identifiziert (P-Wert < 0.05). Innerhalb individueller Batches wurden je 567 (FF, Zeitpunkt 1), 449 (FFPE, Zeitpunkt 1), 547 (FF, Zeitpunkt 2) und 328 (FFPE, Zeitpunkt 2) differentielle Proteine ermittelt. Nach Batcheffektkorrektur konnten mit Limma und ComBat 1555 Proteine als differentiell klassifiziert werden. Zwischen Limma und ComBat konnte eine 93.2% iger Überschneidung zwischen signifikant differentiell abundanten Proteinen ermittelt werden. 36 Kandidaten (2.1 % aller signifikant differentiell abundanten Proteine) wurden nur im unkorrigierten Datensatz als differentiell klassifiziert. Je 3.9 % (ComBat, 67 Proteine) und 2.1 % (Limma, 36 Proteine) wurden nur nach Batcheffektkorrektur mit einer Methode als reguliert zwischen Tumor und Kontrolle eingeordnet.

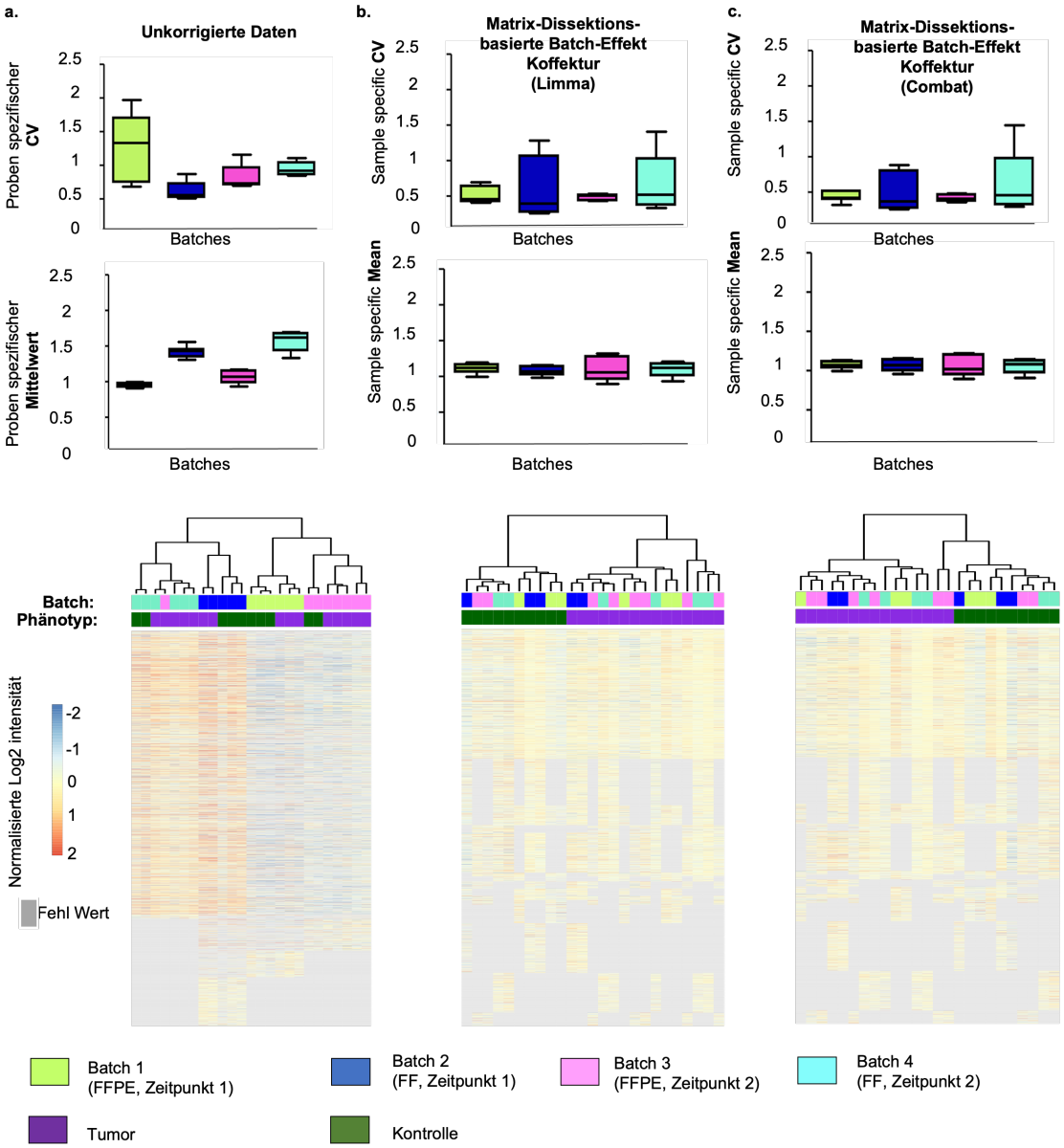

Abbildung 3.17: Limma- und ComBat-basierte Anwendung des Matrix-Dissektionsverfahrens zur Batcheffektreduktion zwischen unterschiedlichen Gewebskonservierungs-Techniken und Analysezeitpunkten am Beispiel des Maus-Medulloblastom-Datensatzes.**(oben)**Heatmap-Visualisierung des auf der Pearson-Korrelation basierenden hierarchischen Clustering mit Ward D Linkage für unkorrigierte integrierte Daten **(a)** und nach Limma- **(b)**- und ComBat**(c)**-basierter (L/S-Scaling, nicht-parametrisches Bayesian-Framework) Batcheffektkorrektur im Matrix-Dissektions-Framework.**(unten)**probenspezifischer Varianzkoeffizient (CV) und der Mittelwert für unkorrigierte integrierte Daten und nach ComBat- und Limma-basierter Batcheffektkorrektur im Matrix-Dissektions-Framework.

Um die Auswirkung der Batcheffektkorrektur auf ein einzelnes Protein zu untersuchen, wurde die Abundanzverteilung des etablierten SHH-Medulloblastom-Biomarkers Filamin A (FLNA) (10) zwischen Tumor und Kontrollproben vor und nach Batcheffektkorrektur untersucht (Abbildung 18).

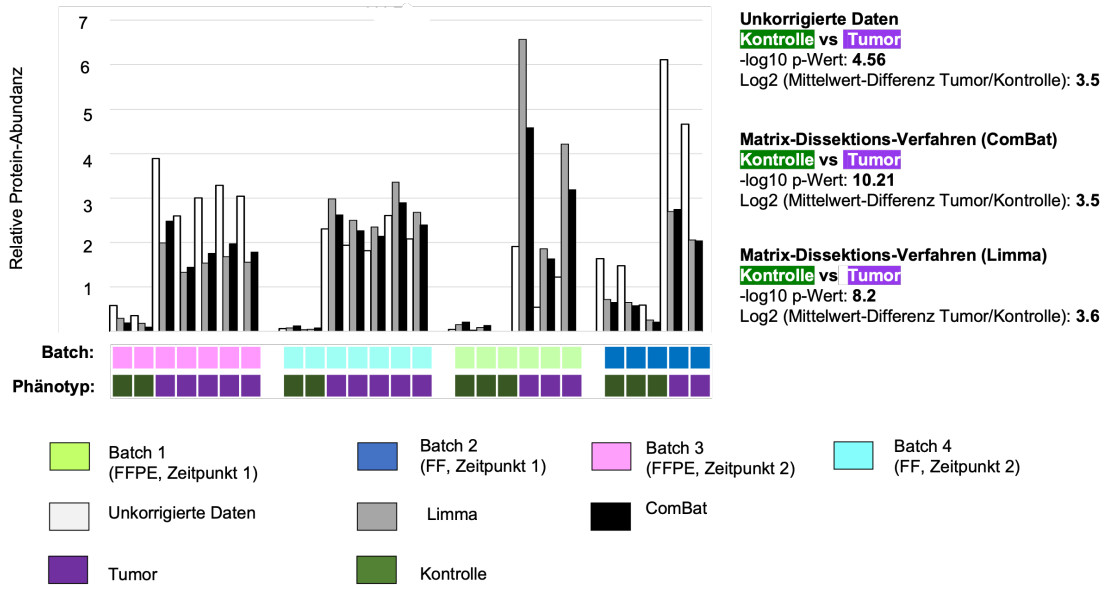

Abbildung 3.18: Abundanz-Verteilung des Sonic-Hedgehog-Medulloblastom-Markers Filamin A in integrierten Daten (unterschiedliche Gewebskonservierungs-Techniken und Analysezeitpunkte) vor und nach der Batcheffektkorrektur mittels Matrix-Dissektionsverfahren unter Verwendung der *RemoveBatcheffects()*-Funktion in Limma und des ComBat Algorithmus (L/S-Scaling, nicht-parametrisches Bayesian-Framework), am Beispiel des Maus-Medulloblastom-Datensatzes (Abbildung 3.1. b)

Derweil im Hinblick auf die Mittelwertdifferenz zwischen Tumor und Kontrolle im integrierten Datensatz vor und nach der Batcheffektkorrektur kein Unterschied festgestellt werden konnte, erhöhte sich die T-Test-Signifikanz zwischen Tumor und Kontrolle auf einen negativen dekadischen Logarithmus des P-Wertes von 10.2 für die Anwendung des ComBat-Algorithmus und 8.2 nach Nutzung des linearen Regressionsmodells der *RemoveBatcheffects()*-Funktion des Limma Algorithmus. Die Abundanz -verteilung von Proteinen innerhalb einzelner Batches blieb vor und nach der Batcheffektkorrektur mit beiden verwendeten Algorithmen konstant.

3.6 | Matrix-Dissektionsverfahren zur Reduktion von Batcheffekten zwischen verschiedenen Tandem Mass Tag-Batches

Markierungsbasierte Quantifizierungstechniken wie die Tandem Mass Tag (TMT)- Quantifizierung bieten die Möglichkeit der Multiplex-Analyse. Hierbei können bis zu 16 Proben, durch den Einsatz probenspezifischer, isobarer Markierungen gleichzeitig vermessen werden. Als Folge werden kleinere Probenmengen einer Einzelprobe benötigt, da die Proben in der LC-MS-Messung kombiniert werden können. Multiplexing führt des Weiteren zu einer höheren Vergleichbarkeit zwischen Proben, da die Varianz der Umwelteinflüsse bei der LC-MS-Analyse deutlich reduziert werden kann.

Die Vergleichbarkeit zwischen TMT-Messungen ist limitiert, wenn die Probenzahl die Zahl verfügbarer isobarer Markierungen übersteigt (36). Als Folge resultiert die Induktion von Batcheffekten bei der Vermessung mehrerer TMT-Batches für große Probenkohorten. Nach aktuellem Stand der Technik werden TMT-Batcheffekte durch die Nutzung interner Referenzstandards (iRS) ausgeglichen, wobei die Abundanz jedes Peptids/
Proteins einer Probe auf die mittlere Abundanz des Peptids/Proteins in der Referenz-Probe des jeweiligen Batches normalisiert wird (3).

Da iRS von der Verwendung interner Referenzen abhängt, können unabhängige generierte TMT-Batches ohne identische interne Standards nicht der gemeinsamen Batcheffektkorrektur unterzogen und folglich nicht integriert werden. Des Weiteren ist die Verwendung einzelner Proben für die Anpassung der Batcheffekte sehr anfällig für durch experimentelle Ungenauigkeiten induzierte Fehler. Aus diesem Grund wurde im Rahmen dieser Arbeit die Anwendbarkeit des Matrix-Dissektionsverfahrens für vier TMT-

8-Plex-Batches des Cetuximab-Datensatz (Abbildung 3.1. a) evaluiert und mit der iRS-Normalisierung auf Peptid- und Proteinebene verglichen. Da auf Peptid- und Proteinebene eine Gaußsche Wahrscheinlichkeitsverteilung und divergierende probenspezifische Mittelwerte und CVs zu beobachten waren, wurde die L/S-Skalierung innerhalb des parametrischen Bayes-Frameworks von ComBat für den Vergleich mit iRS verwendet (Abbildung 3.19.-3.21.).

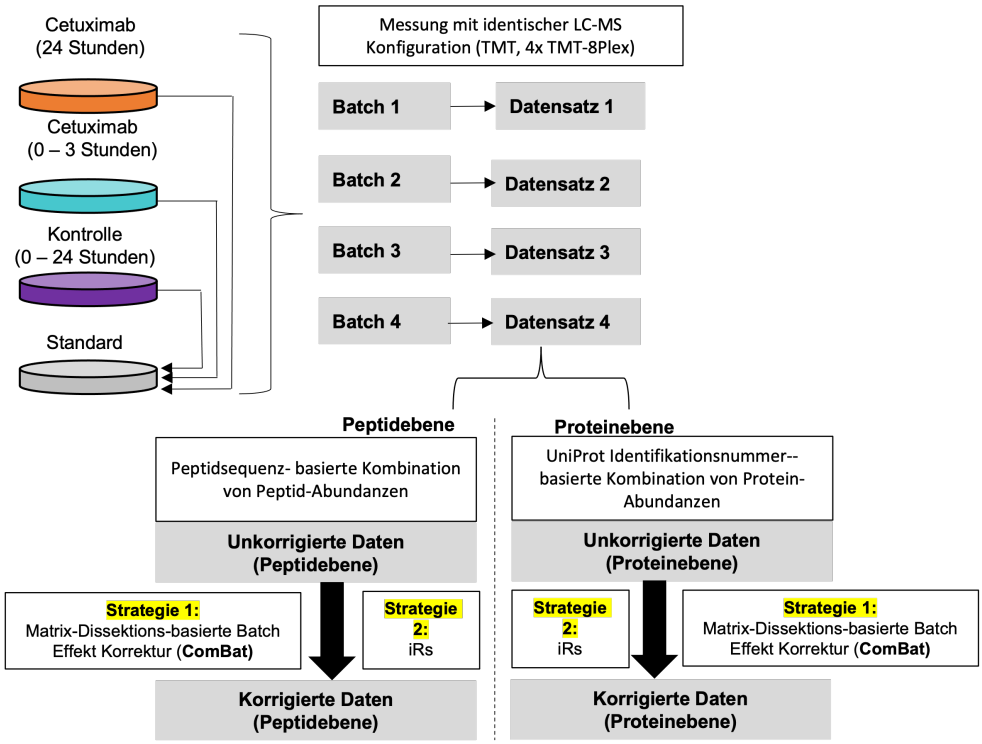

Abbildung 3.19: Schematische Darstellung des Vergleichs von Matrix-Dissektionsverfahrens und Nutzung interner Referenzstandards zur Reduktion von Batcheffekten zwischen Tandem Mass Tag (TMT)-Batches am Beispiel des Cetuximab-Datensatzes auf Protein- und Peptidebene.

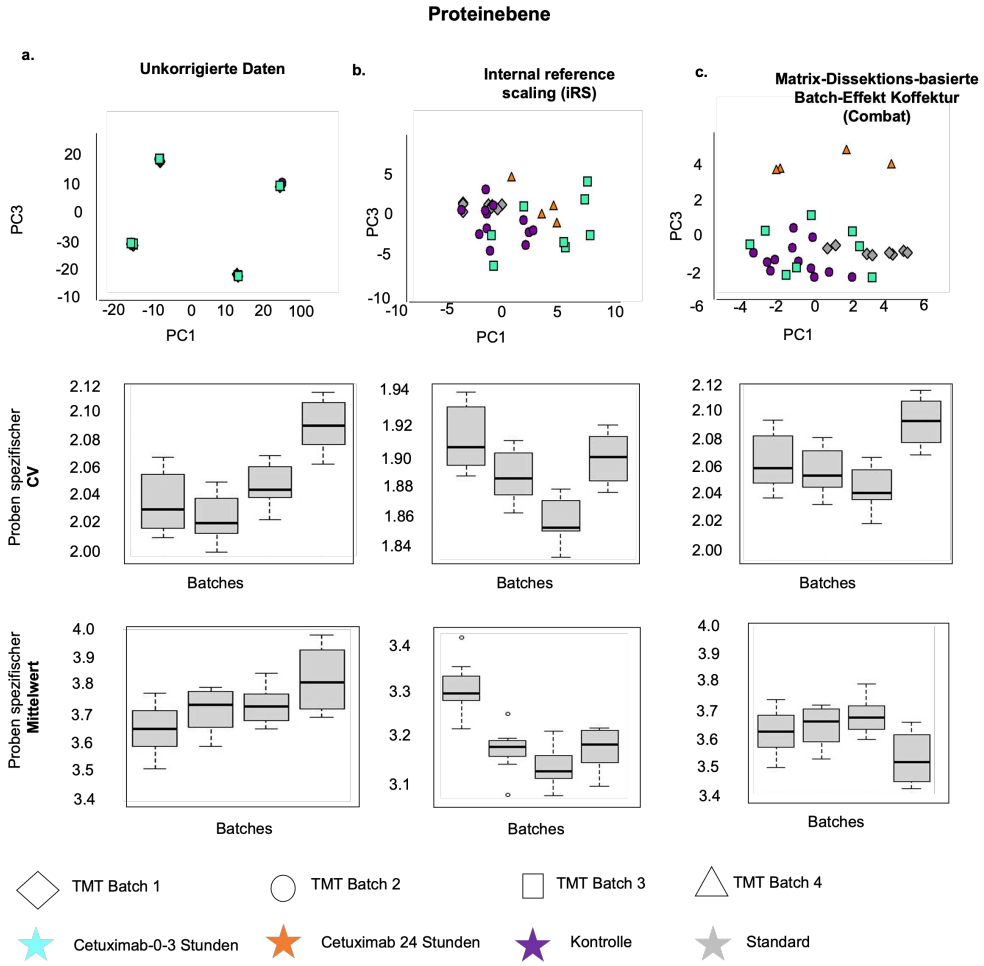

Abbildung 3.20: Vergleich der Anwendung des Matrix-Dissektionsverfahrens und Nutzung interner Referenzstandards zur Reduktion von Batcheffekten zwischen Tandem Mass Tag (TMT)-Batches am Beispiel des Cetuximab-Datensatzes auf Proteinebene.**(a)**Streudiagramm der Verteilung der Proben über Hauptkomponente 1 und 3 in NIPALS-PCA, basierend auf 2152 Proteinen, die in 50% aller Proben gefunden wurden sowie probenspezifische CV/Mittelwerte für jedes TMT-Batch für unkorrigierte Daten (alle Batches repräsentieren n = 8 biologisch unabhängige Proben. (Kontrolle: n = 2; Cetuximab 24 h: n = 1; Cetuximab 0-3 h: n = 3; interne Referenz: n = 2). **(b)**Streudiagramm der Verteilung der Proben über Hauptkomponente 1 und 3 in NIPALS-PCA, basierend auf 2152 Proteinen, die in 50% aller Proben gefunden wurden sowie probenspezifische CV/Mittelwerte für jedes TMT-Batch nach iRS-basierter Batcheffektkorrektur.**(c)**Streudiagramm der Verteilung der Proben über Hauptkomponente 1 und 3 in NIPALS-PCA, basierend auf 2152 Proteinen, die in 50% aller Proben gefunden wurden sowie probenspezifische CV/Mittelwerte für jedes TMT Batch nach Matrix-Dissektionsverfahren-basierter Batcheffektkorrektur.

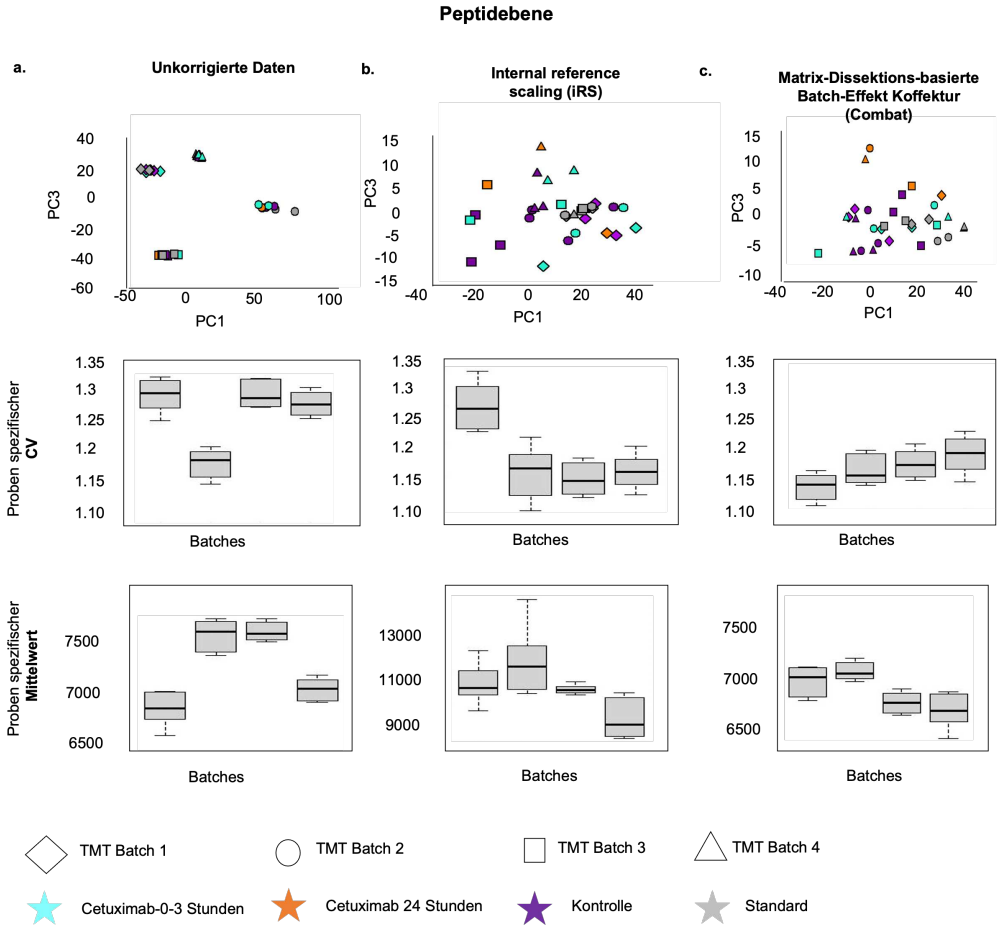

Abbildung 3.21: Vergleich der Anwendung des Matrix-Dissektionsverfahrens und Nutzung interner Referenzstandards zur Reduktion von Batcheffekten zwischen Tandem Mass Tag (TMT)-Batches am Beispiel des Cetuximab-Datensatzes auf Peptidebene.**(a)**Streudiagramm der Verteilung der Proben über Hauptkomponente 1 und 3 in NIPALS-PCA, basierend auf 8877 Peptiden, die in 50% aller Proben gefunden wurden sowie probenspezifische CV/Mittelwerte für jedes TMT-Batch für unkorrigierte Daten (Alle Batches repräsentieren n = 8 biologisch unabhängige Proben. (Kontrolle: n = 2; Cetuximab 24 h: n = 1; Cetuximab 0-3 h: n = 3; interne Referenz: n = 2) **(b)**Streudiagramm der Verteilung der Proben über Hauptkomponente 1 und 3 in NIPALS-PCA, basierend auf 8877 Peptiden, die in 50% aller Proben gefunden wurden sowie probenspezifische CV/Mittelwerte für jedes TMT-Batch nach iRS-basierter Batcheffektkorrektur.**(c)**Streudiagramm der Verteilung der Proben über Hauptkomponente 1 und 3 in NIPALS-PCA, basierend auf 8877 Peptiden, die in 50% aller Proben gefunden wurden sowie probenspezifische CV/Mittelwerte für jedes TMT-Batch nach Matrix-Dissektionsverfahren-basierter Batcheffektkorrektur.

Die fehlwerttolerante "Nonlinear Iterative vertical Least Squares (NIPALS)"-PCA wurde genutzt, um die Ähnlichkeit von Proben in integrierten, unkorrigierten Datensätzen, sowie nach iRS und Anwendung des Matrix-Dissektionsverfahrens zur Reduk-

tion von TMT-Batcheffekten, unter Berücksichtigung mehrerer Dimensionen, zu analy -sieren. Vor der Batcheffekt-Reduktion ordneten sich die Proben auf Peptid- und Proteinebene in Abhängigkeit ihrer jeweiligen TMT-Batches an (Abbildung 3.20.a), 3.21. a)). Nach iRS-Korrektur konnten Batcheffekte weiterhin für Hauptkomponenten 1 bis 3 beobachtet werden. Nach Anwendung des Matrix-Dissektionsverfahrens wurde keine batchabhängige Anordnung der Proben festgestellt.

Nach *Stepath et al. (2020)*, ist für die Cetuximab Behandlung der Kolorektalkarzinom-Zellinie DiFi nach 24 Stunden eine klare Unterscheidbarkeit zu erwarten. Dies konnte im Rahmen der Originalpublikation für SILAC- und DDA-Daten nachgewiesen werden. Nach Anwendung von iRS zur Korrektur von TMT-Batcheffekten, zeigte sich für den TMT-Datensatz keine klare Abgrenzung des mit Cetuximab behandelten Phänotyps nach 24 Stunden (55). In Übereinstimmung mit diesen Ergebnissen (55), konnte im Rahmen dieser Arbeit keine Abgrenzung mit Cetuximab behandelter Zellen nach 24h in Folge der iRS-Korrektur auf Protein- und Peptidebene nachgewiesen werden (Abbildung 3.20. b, 3. 21.b). Nach Anwendung des Matrix-Dissektionsverfahrens, bildeten 24h mit Cetuximab inkubierte Proben separate Cluster auf Peptid- und Proteinebene (Abbildung 3. 20.c, 3.21.c). Des Weiteren konnte auf beiden Ebenen die bestmögliche Annäherung der probenspezifischen Mittelwerte und CVs nach der Verwendung des Matrix-Dissektionsverfahrens beobachtet werden (Abbildung 3.20, 3.21).

Um die Auswirkung der iRS-Normalisierung- und Matrix-Dissektionsverfahren -basierten Batcheffektkorrektur auf ein einzelnes Protein zu bewerten, wurde der gesam -te und batchspezifische Varianzkoeffizient für das etablierte Housekeeping-Protein Nuxid Hydrolase 21 (NUDT21) berechnet. Für NUDT21 wird gewebsübergreifend ein Varianzkoeffizient von maximal 4.9% erwartet (28). Auf Peptidebene wurde der Gesamt-CV beispielhaft für drei einzigartige tryptische Peptide von NUDT21 ermittelt. Für alle Peptide reduzierte die Anwendung des Matrix-Dissektionsverfahrens den CV von 23.3-25.5% auf 6-7%, während für iRS-normalisierte Peptide CV-Werte zwischen 12 und 14.5% berechnet wurden (Abbildung 3. 22). Auf Proteinebene wurde vor der Reduk-

tion des Batcheffekts ein Gesamt-CV von 6% beobachtet. Nach iRS sank der Batch-übergreifende CV auf 4.8% während die Matrix-Dissektionsverfahren-basierte Batch-effektreduktion mit 4.1% im niedrigsten Varianzkoeffizienten für NUDT21 resultierte.

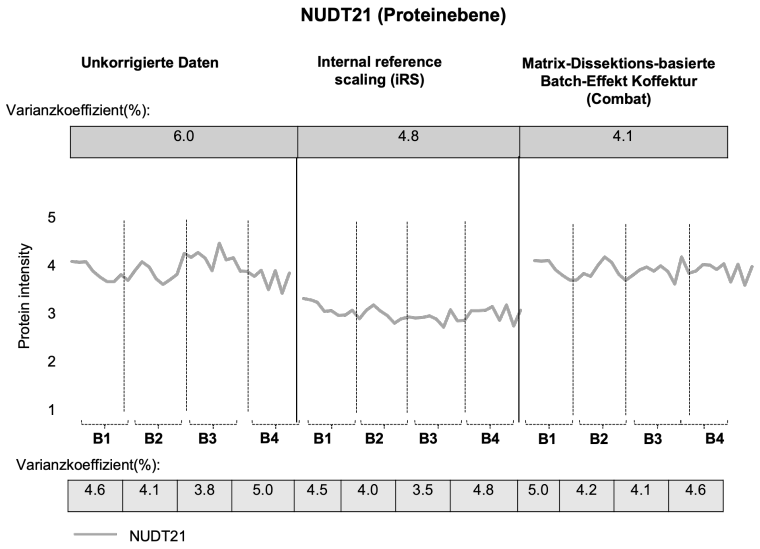

Abbildung 3.22: Batchübergreifender und batchspezifischer Varianzkoeffizient für das Housekeeping-Protein NUDT21 für unkorrigierte Daten nach iRS-Normalisierung und nach Matrix-Dissektionsverfahren-basierter Batcheffektkorrektur auf Proteinebene.

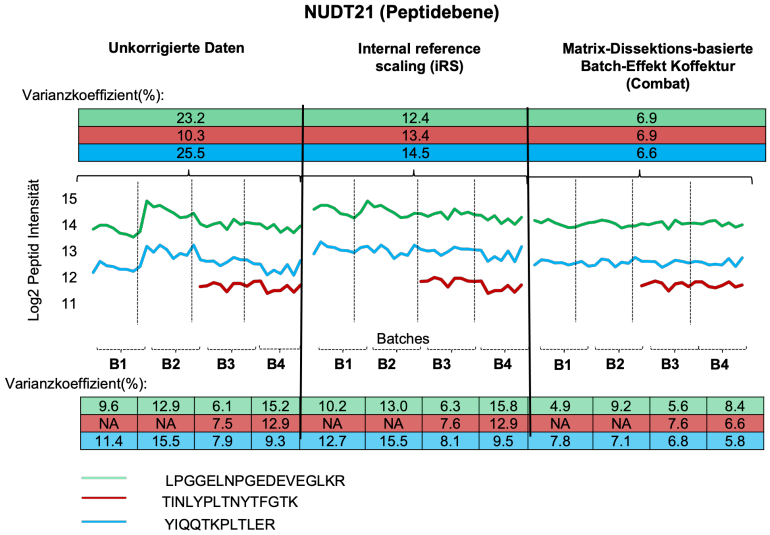

Abbildung 3.23: Batchübergreifender und batchspezifischer Varianzkoeffizient für das Housekeeping-Protein NUDT21 für unkorrigierte Daten nach iRS-Normalisierung und nach Matrix-Dissektionsverfahren-basierter Batcheffektkorrektur auf Peptidebene.

Um die Anwendbarkeit des Matrix-Dissektionsverfahrens auf die Batcheffektkorrektur zwischen TMT-Batches für größere Datensätze zu evaluieren wurde der von *Petrialia et al. (2021)* - (43) veröffentlichte Hirntumor-Datensatz vor und nach der Batcheffektkorrektur untersucht (Abbildung 3.24.-3.26.).

In der Original-Publikation (43) wurden TMT-Batcheffekte unter Verwendung des ComBat -Algorithmus (parametrisches Bayesian-Framework, L/S-Scaling) nach Random-Forest-basierter Imputation von Fehlwerten korrigiert. Im Rahmen der Studie wurden proteomische Profile von sieben verschiedenen pädiatrischen Hirntumorentitäten verglichen. (ATRT, Eraniphagyngioma, Ependymoma, Ganglioglioma, High Grade Glioma, Low Grade Glioma, Medulloblastoma). Unter Anderem beobachteten *Pertalia et al. (2021)* - klar definierte Cluster für Medulloblastome und Ependymome. Niedriggradige Gliome (LGG), hochgradige Gliome (HGG) und Gangliogliome bildeten gemischte Untercluster auf Proteomebene.

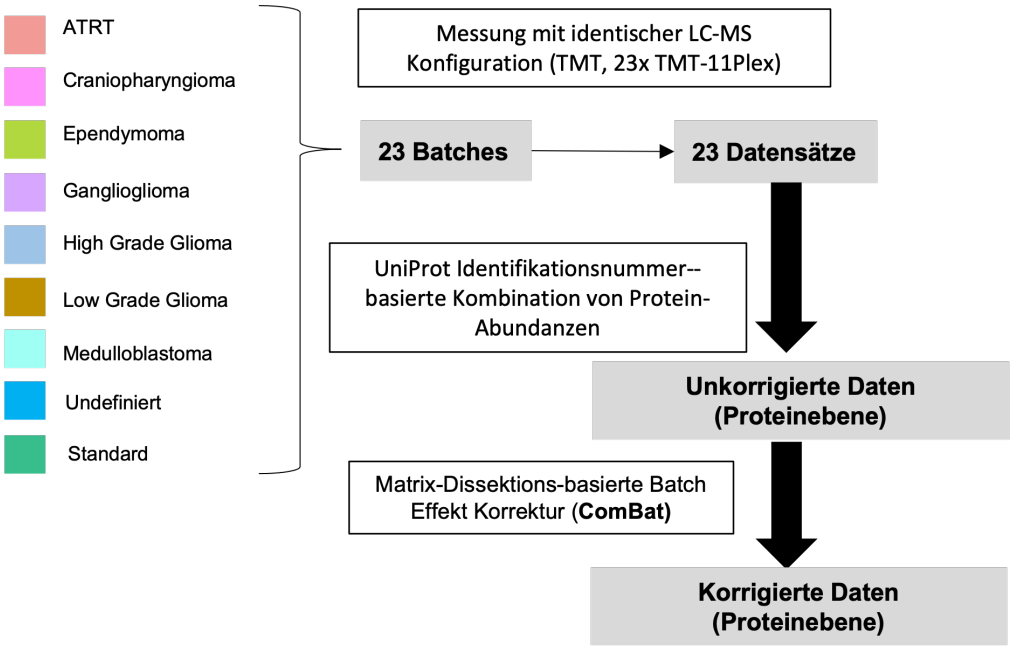

Abbildung 3.24: Schematische Darstellung des Vergleichs der Evaluation der Anwendbarkeit des Matrix-Dissektionsverfahrens zur Reduktion von Batcheffekten zwischen Tandem Mass Tag (TMT)-Batches in großen TMT-Datensätzen am Beispiel des humanen Hirntumor-Datensatzes auf Proteinebene.

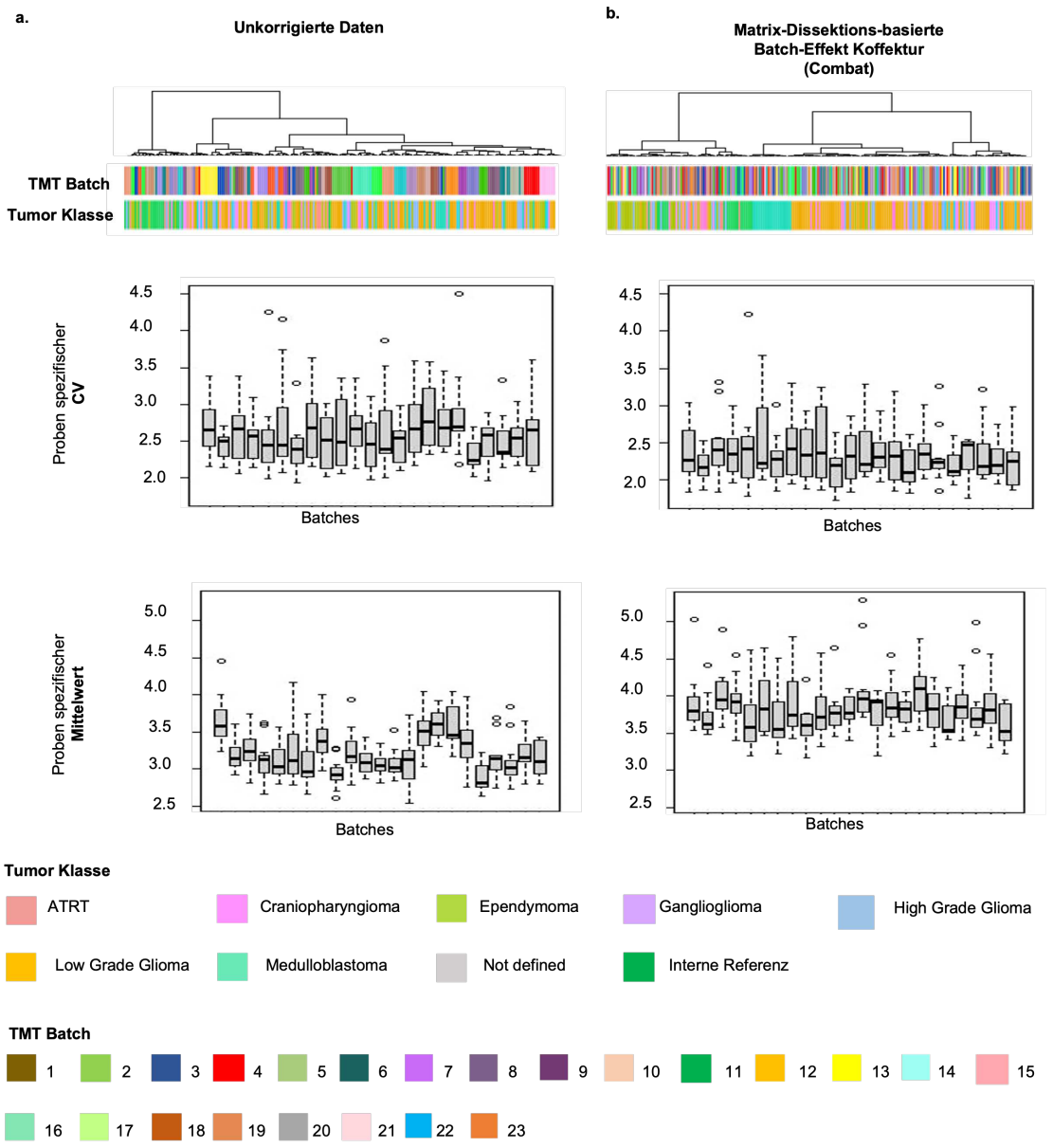

Abbildung 3.25: Vergleich der Evaluation der Anwendbarkeit des Matrix-Dissektionsverfahrens zur Reduktion von Batcheffekten zwischen Tandem Mass Tag (TMT)-Batches in großen TMT-Datensätzen am Beispiel des humanen Hirntumor-Datensatzes auf Proteinebene **(a)**Heatmap Visualisierung des Pearson-Korrelations-basierten hierarchischen Clustering nach UniProt-Identifikationsnummer-basierter Datenintegration, basierend auf 9156 identifizierten Proteinen sowie probenspezifische CV/Mittelwerte für jedes TMT-Batch für unkorrigierte Daten. **(b)**Heatmap-Visualisierung des Pearson-Korrelations-basierten hierarchischen Clustering nach UniProt-Identifikationsnummer-basierter Datenintegration, basierend auf 9156 identifizierten Proteinen, sowie probenspezifische CV/Mittelwerte für jedes TMT-Batch nach Matrix-Dissektionsverfahren-basierter Batcheffektkorrektur (ComBat,-parametrisches Bayesian-Framework, L/S-Scaling).

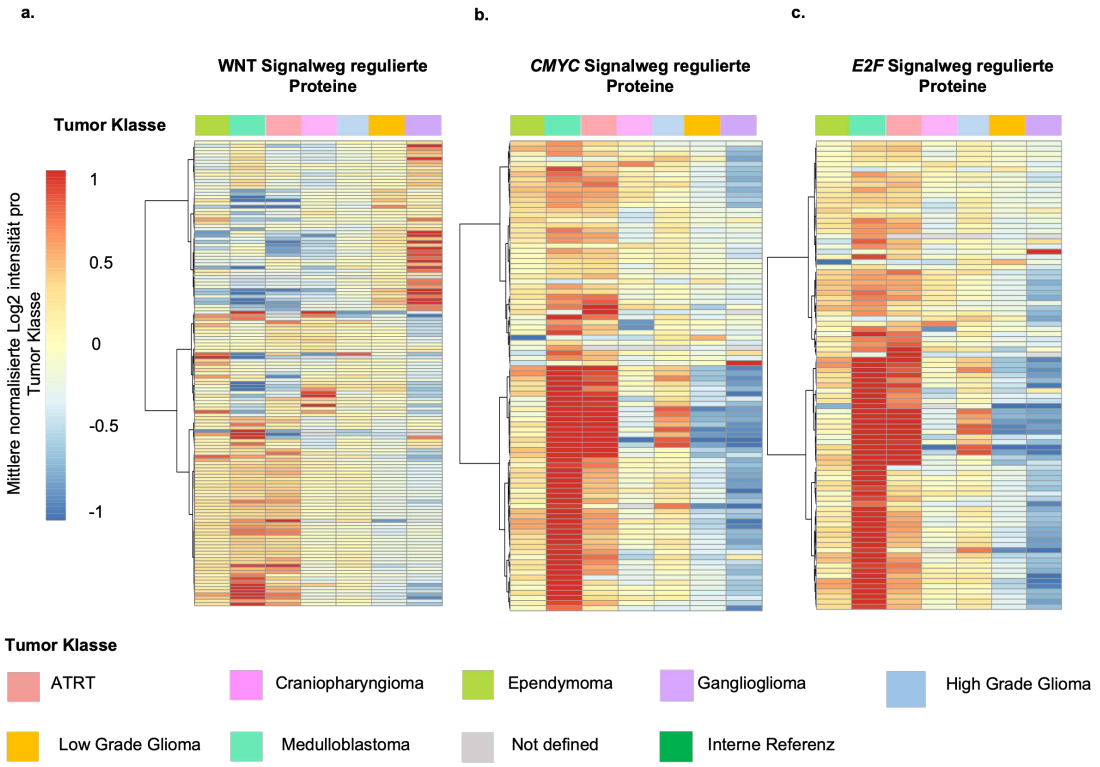

Abbildung 3.26: Heatmap-Visualisierung der Tumor-typspezifischen Häufigkeitsverteilung von Proteinen, die mit den Tumor-relevanten Gensets "Hallmark-MYC Targets; Hallmark-E2F-Targets und REACTOME-Signaling by WNT" assoziiert sind nach Matrix-Dissektionsverfahren-basierter Batcheffektkorrektur (ComBat, parametrisches Bayesian-Framework, L/S-Scaling).

Auf Signalweg-Ebene wurden für die vorherrschenden Krebssignalwege MYC, E2F und WNT untersucht. Insbesondere MYC- und E2F-assoziierte Proteine wiesen eine signifikant höhere Abundanz bei ATRT und Medulloblastom im Vergleich zu allen anderen untersuchten Tumorentitäten auf. Für den WNT-Signalweg wurde eine tendenziell niedrigere Abundanz in niedriggradige Gliome (LGG) und hochgradige Gliome (HGG) sowie Gangliogliome beschrieben. (43)

Vor der Batcheffektkorrektur konnte ein TMT-Batch-abhängiges Clustering beobachtet werden, derweil im hierarchischen Clustering nur für klar abgrenzbare Tumorentitäten (Medulloblastom, Ependymom) Batch-unabhängige Cluster identifiziert wurden. Probenspezifische Mittelwerte und probenspezifische Varianzkoeffizienten zeigten klare Unterschiede zwischen Batches. (Abbildung 3.25. a) Nach der Matrix-Dissektions-

basierten Batcheffektreduktion zeigte sich eine klare Unterscheidbarkeit von Medulloblastomen und Ependymomen, während mehrere gemischte Untercluster für niedriggradige LGG, HGG und Gangliogliome identifiziert wurden. Des Weiteren konnte kein Batch-basiertes Clustering beobachtet werden. (Abbildung 3.25. b). Der probenspezifische Varianzkoeffizient und Mittelwert zeigten eine geringere Fluktuation zwischen Batches im Vergleich zum unkorrigierten Datensatz.

Nach Batcheffektkorrektur wiesen insbesondere MYC- und E2F-assoziierte Proteine eine höhere Abundanz in ATRTs und Medulloblastomen auf. Für den WNT-Signalweg konnte eine deutlich geringere Abundanz aller assoziierten Faktoren in LGG und HGG beobachtet werden (Abbildung 3.26.). Gangliogliome, Medulloblastome, Ependymome und Craniopharyngnome zeigte eine erhöhte Abundanz einzelner WNT-assoziierter Proteine. Dabei konnten für Medulloblastome, Ependymome und Craniopharyngnome ähnliche Proteinabundanz-Verteilungen beobachtet werden. Gangliogliome zeigten eine erhöhte Abundanz von Proteinen, welche in allen anderen Entitäten eine geringe Abundanz aufwiesen.

Die Matrixzerlegung von 23 Batches kann potenziell zu 4194281 Kombinationen von Submatrizen für die Batcheffektkorrektur führen. Real wurden jedoch nur Batches 3654 Submatrizen für 23 TMT-Batches des humanen Hirntumor-Datensatzes im Matrix -Dissektionsverfahren beobachtet.

3.7 | Matrix-Dissektionsverfahren zur Reduktion von Batcheffekten zwischen verschiedenen Quantifizierungstechniken

Für die relative Quantifizierung von Proteinen können verschiedene Label-basierte und Label-freie Strategien angewandt werden. Um die Anwendbarkeit des Matrix -Dissektionsverfahrens auf die Reduktion von Batcheffekten zwischen Lable-freien und Label-basierten Quantifizierungstechniken zu evaluieren, wurde die Integration von DDA-LFQ-, SILAC- und TMT-Daten des Cetuximab-Datensatzes untersucht (Abbil -dung 28-30). In Bezug auf den TMT-Datensatz wurden mittels Matrix-Dissektions -verfahren angeglichene TMT-8-Plex-Batches verwendet, da im Vergleich zur iRS-Normali -sierung ein niedrigerer Varianzkoeffizient für das Housekeeping-Protein NUDT21 ermittelt wurde und für die 24h-Cetuximab-Kondition ein klar definiertes Cluster nach Batcheffektkorrektur ermittelt wurde.

Für SILAC-Daten werden die relativen Proteinabundanzen als Verhältnisse zwischen jeder einzelnen Probe und einer mit schweren Isotopen markierten Referenz dargestellt (SILAC-Ratios). Für große Kohorten wird hierzu meist ein interner Standard als Referenz gewählt, welcher die mittlere Abundanz der Proteine über alle Phänotypen repräsentieren soll (Spike in-SILAC/Super-SILAC) (13).

Um aus TMT- und DDA-Messungen resultierende Daten an die SILAC-Daten anzugleichen, wurde für jedes Protein das Verhältnis zwischen der probenspezifischen Abundanz der mittleren Häufigkeit in allen Proben berechnet. Im Folgenden wurde Batcheffekte zwischen SILAC-, TMT- und DDA-Daten unter Anwendung des ComBat-Algorith -mus (L/S-Scaling, parametisches Bayesian-Framework) im Rahmen des Matrix -Dissektionsverfahrens korrigiert.

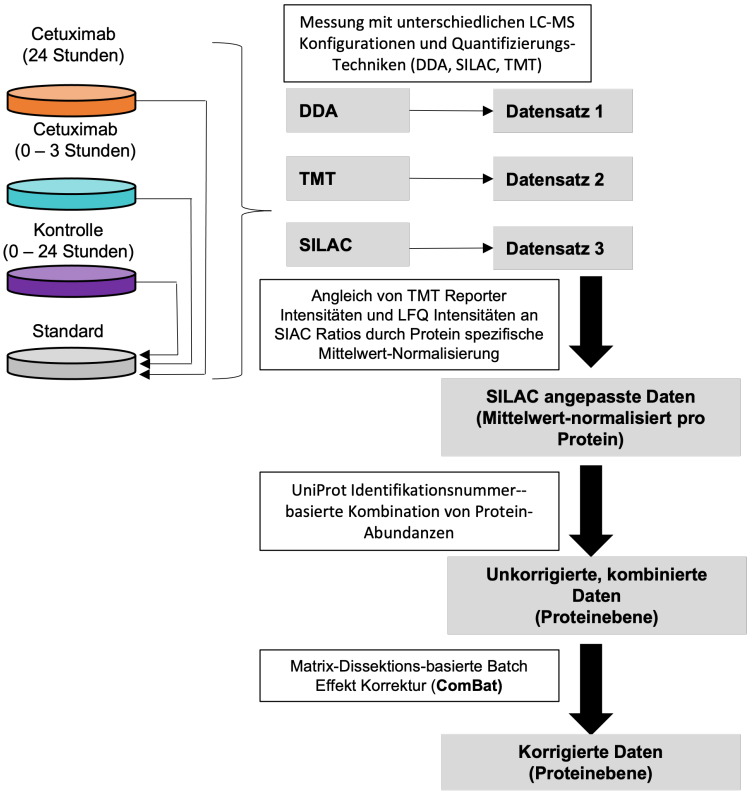

Abbildung 3.27: Schematische Darstellung der Evaluation der Anwendbarkeit des Matrix-Dissektionsverfahrens zur Reduktion von Batcheffekten zwischen verschiedenen Quantifizierungstechniken.

Vor der Batcheffektkorrektur konnte in der NIPALS-PCA auf Basis der ersten zwei Hauptkomponenten ein Clustering in Abhängigkeit der verwendeten Quantifizierungstechnik beobachtet werden. (67 % der beobachteten Varianz wurden durch Hauptkomponente 1 erklärt). Phänotypische Unterschiede wurden bis Hauptkomponente 3 nicht repräsentiert. Nach dem Angleich von DDA-LFQ- und TMT-basierten Proteinabundanzen an SILAC-Ratios konnte eine allgemeine Unterscheidbarkeit der 24 Stunden mit Cetuximab behandelten Proben beobachtet werden. Gleichzeitig konnte weiterhin eine Quantifizierungstechnik-bedingte Anordnung der Proben identifiziert werden, wobei im Vergleich zu unkorrigierten Daten eine geringere Batch-abhängige Ähnlichkeit zwischen Proben vorlag. Nach zusätzlicher Anwendung des Matrix-Dissektionsverfahrens wurde der technische Batcheffekt effizient reduziert, während 24 h mit Cetuximab be-

handelte Proben ein eindeutiges Cluster bildeten (Abbildung 3.28.a)-c)).

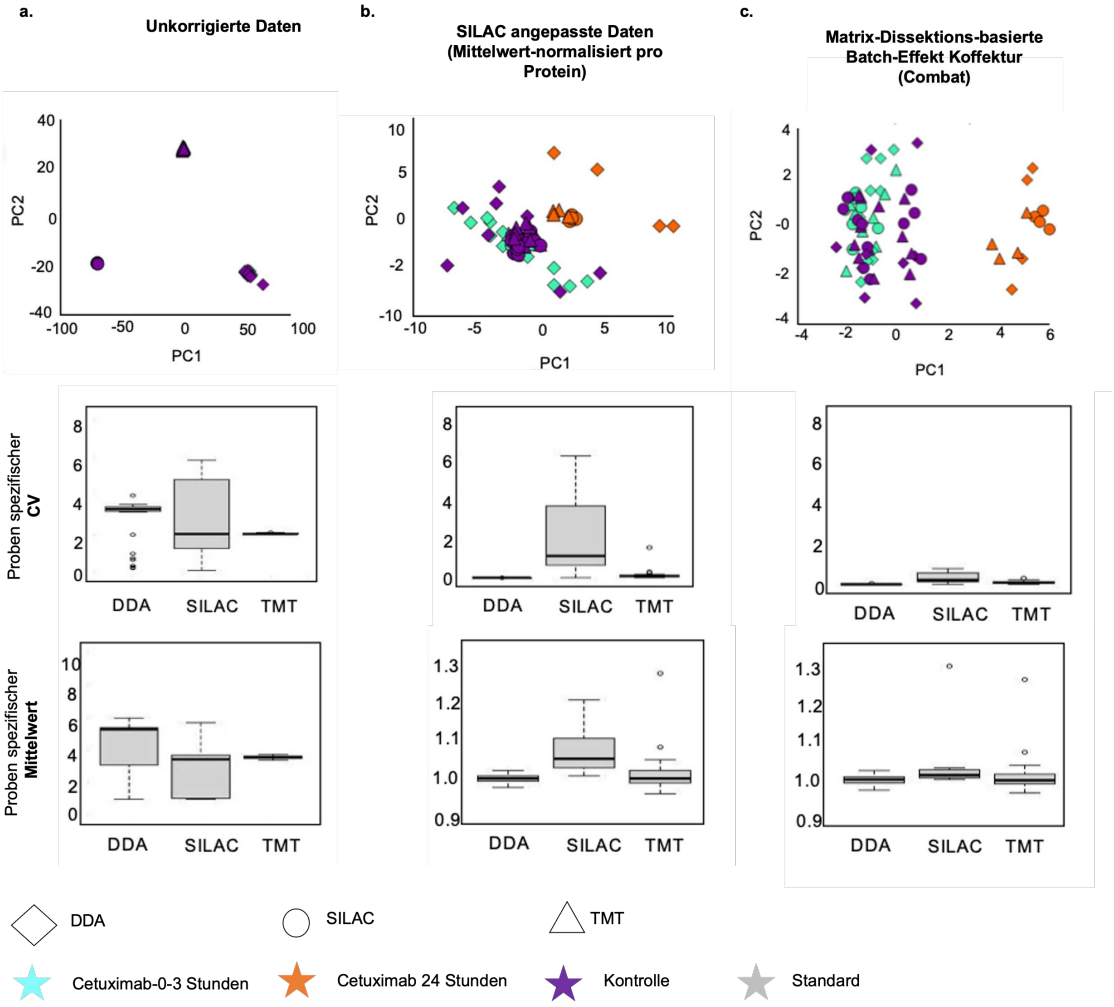

Abbildung 3.28: Evaluation der Anwendbarkeit des Matrix-Dissektionsverfahrens zur Reduktion von Batcheffekten zwischen SILAC-, TMT- und DDA-LFQ-Daten am Beispiel des Cetuximab-Datensatzes **(a)**Streudiagramm-Visualisierung der Verteilung der Proben über Hauptkomponente 1 und 2 in NIPALS-PCA, basierend auf 2368 Proteinen, die in 50% aller Proben gefunden wurden sowie probenspezifische CV/Mittelwerte für jedes TMT-Batch für unkorrigierte Daten (alle Batches repräsentieren n = 24 biologisch unabhängige Proben. (Kontrolle: n =8; Cetuximab 24 h: n = 4; Cetuximab 0-3 h: n = 12) **(b)**Streudiagramm der Verteilung der Proben über Hauptkomponente 1 und 2 in NIPALS-PCA, basierend auf 2368 Proteinen, die in 50% aller Proben gefunden wurden sowie probenspezifische CV/Mittelwerte für jedes TMT-Batch nach iRS-basierter Batcheffektkorrektur.**(c)**Streudiagramm der Verteilung der Proben über Hauptkomponente 1 und 2 in NIPALS-PCA, basierend auf 2368 Proteinen, die in 50% aller Proben gefunden wurden sowie probenspezifische CV/Mittelwerte für jedes TMT-Batch nach Matrix-Dissektionsverfahren-basierter Batcheffektkorrektur.

Vor der Datenharmonisierung wurden für jede Quantifizierungsplattform unterschiedliche probenspezifische Mittelwerte und Varienzkoeffizienten beobachtet (Abbildung 3.28. a-c). SILAC-Proben zeigten den niedrigsten mittleren probenspezifischen CV. Gleichzeitig konnte die höchste Varianz probenspezifischer CVs für den SILAC-Datensatz beobachtet werden.

Nach der Anpassung von DDA-LFQ- und TMT-Daten an SILAC-Ratios wurden probenspezifische CVs und Mittelwert für TMT- und DDA-LFQ-Daten deutlich reduziert. Für den SILAC-Datensatz wurde ebenfalls eine deutliche Verschiebung nach unten beobach -tet. Jedoch blieb die hohe Varianz probenspezifischer CVs für SILAC-Daten konstant. Nach Anwendung des Matrix-Dissektionsverfahrens, wurden vergleichbaren CVs und Mittelwerte für alle Quantifizierungstechniken beobachtet (Abbildung 3.28.. a-c).

Darüber hinaus wurden der Varianzkoeffizient für das Housekeeping-Protein NUDT21 berechnet für jede und zwischen allen Quantifizierungsplattformen ermittelt. Vor der Batcheffektkorrektur betrug der Gesamt-CV 50.5%. Über alle SILAC-Proben wurde ein CV von 4.7% ermittelt. Zwischen TMT-Proben wurde die geringste NUDT21-Varianz beobachtet (4.1%), derweil DDA-LFQ-Proben, mit 10.3% den höchsten CV zeigten.

Die SILAC-Ratio-Anpassung von TMT- und DDA-LFQ-Daten verringerte den Gesamtvarianzkoeffizienten auf 7.5% während der Variationskoeffizienten für jede individuelle Quantifizierungsplattform erhalten blieben. Die Matrix-Dissektionsverfahren-abhängige Batcheffektkorrektur reduzierte den Quantifizierungstechnik-übergreifenden CV zusätzlich auf 7.0%. Gleichzeitig erhöhte sich der SILAC- und TMT-spezifische CVs auf 6.9% und 7.0%. Für DDA-LFQ-Daten wurde eine Reduktion auf 7.1%, nach Anwendung des Matrix-Dissektionsverfahrens beobachtet (Abbildung 3.29.).

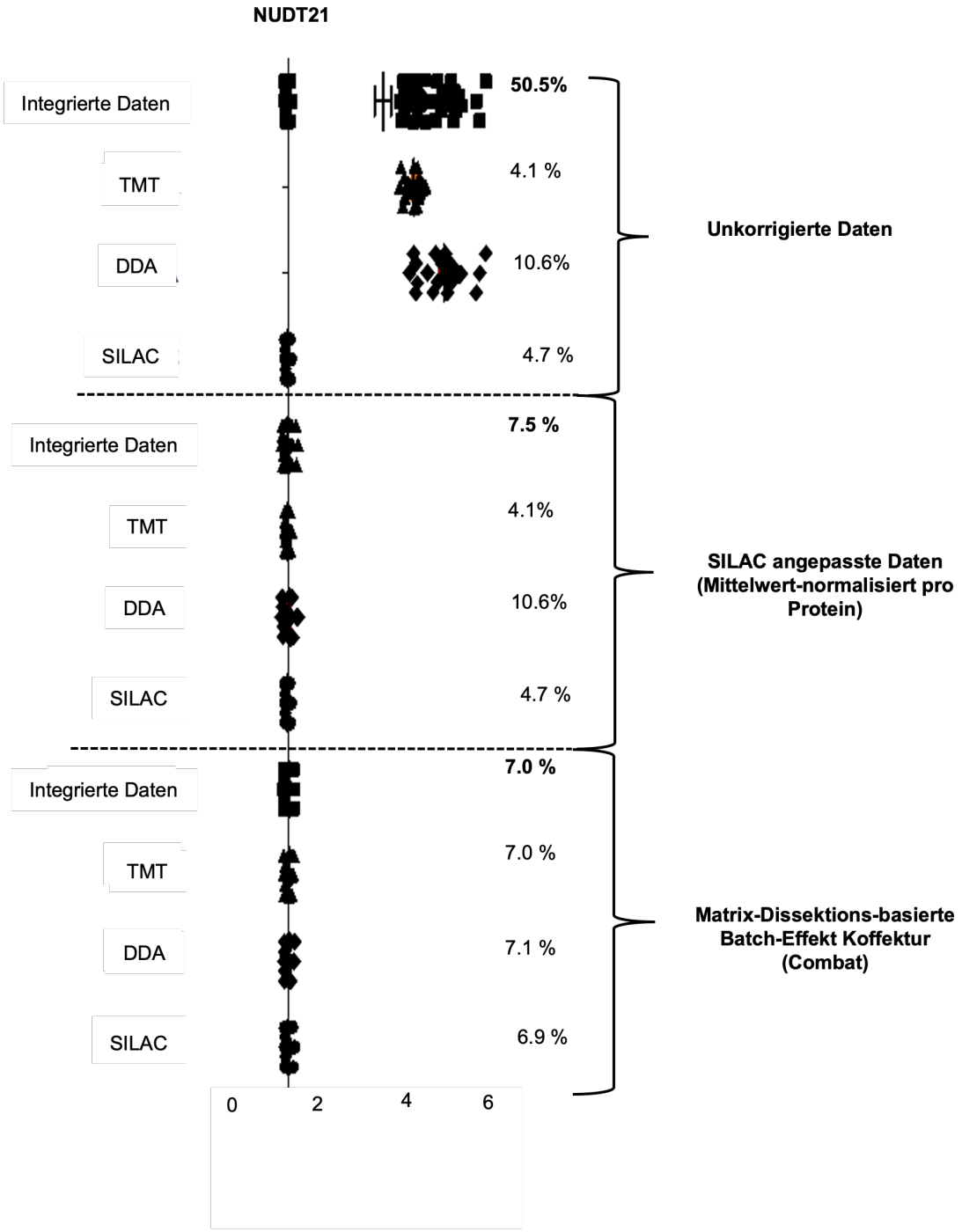

Abbildung 3.29: Batchübergreifender und batchspezifischer Varianzkoeffizient für das Housekeeping-Protein NUDT21 für unkorrigierte Daten nach Angleich von TMT- und DDA-Daten an SILAC-Verhältnisse und nach Matrix-Dissektionsverfahren-basierter Batcheffektkorrektur auf Proteinebene

4

Diskussion

Ziel dieser Arbeit war die Entwicklung einer Fehlwert-toleranten Pipeline zur Integration unabhängig generierter Proteomdatensätze, welchen unterschiedliche experimentelle Konfigurationen zugrunde liegen.
”Omics”-Analysen charakterisieren und quantifizieren eine große Anzahl von Bio -molekülen mit dem Ziel biologische Mechanismen zu verstehen und neue Biomarker und therapeutische Targets zu identifizieren (42). Proteomanalysen nehmen hierbei eine Sonderstellung ein, da sie im Gegensatz zu DNA-Sequenzdaten, DNA-Methylierungs -profilen und Transkriptomdaten pharmakologisch adressierbare Phänotypen direkt widerspiegeln (14). ”Omics”-Studien erzeugen eine große Menge an Daten. Als Folge daraus werden statistische Methoden benötigt, um biologisch relevante Informationen aus den hochdimensaionalen Datensätzen zu extrahieren (67). Unabhängig von der Art der ”Omics”-Daten, ist die statistische Aussagekraft von einzelnen Studien oft aufgrund relativ geringer Kohortengrößen limitiert (38). Die Integration von Daten aus unabhängig durchgeführten Studien zu gleichen Fragestellungen kann die Kohortengröße von ”Omics”-Datensätzen effizient erhöhen und, besonders für seltene Erkrankungen, die statistische Validität von ”Omics”-Analysen signifikant steigern.
Die gemeinsame Analyse unabhängig generierter ”Omics”-Datensätze ist durch Batcheffekte limitiert, welche bei der Integration mit unterschiedlicher technischer Konfiguration aufgenommener Einzeldatensätze induziert werden. Für Proteomdatensätze können diese Varianzen vor allem auf die Nutzung unterschiedlicher LC-MS-Systeme, Quantifizierungstechniken und auf unterschiedliche Analysezeitpunkte zurückgeführt werden. Im Rahmen dieser Arbeit zum konnte beispielsweise festgestellt werden, dass technische Varianzen, welche durch die Verwendung unterschiedlicher LC-MS -Konfigurationen (Quadrupol-Orbitrap-Hybrid-Massenspektrometer mit einer Messung im „Data dependent acquisition“ (DDA)-Modus, Quadrupol-Orbitrap-Hybrid-Massen -spektrometer unter Verwendung des „Data independent acquisition“ DIA-Modus, TripleTOF-Massenspektrometer unter Nutzung des „Sequential window acquisition of all theorethical mass spectra“ (SWATH)-Modus induziert werden, phänotypische

Differenzen zwischen Zellysaten aus 80 % *Homo sapiens*, 15 % *E. Coli*, 5 % *Saccaromyces cerevisea*und 80 % *Homo sapiens*, 10 % textitE. Coli, 10 % *Saccaromyces cerevisea* überlagern . Aus DIA-und DDA-Messungen an Quadrupol-Orbitrap-Hybrid-Massenspektrometern konnten deutlich weniger Proteine identifiziert werden als in vergleichbaren SWATH-Messungen an einem TripleTOF-Massenspektrometer. Gleichzeitig zeigten alle quantifizierten Proteine eine geringe Proteinabundanzen in TOF-Messungen (Abbildung 3.2). Dies kann auf die technische Konfiguration beider Gerätesysteme zurückgeführt werden. Bei Quadrupol-Orbitrap-Hybrid Masssenspektrometern werden Vorläufer- und Produktionen in einer C-Trap akkumuliert und in die Orbitrap injiziert. Bedingt durch diese Akkumulation wird die Sensitivität des Systems erhöht. Im Vergleich zu TOF-Geräten erscheint die Abundanz identifizierter Proteine höher. Gleichzeitig wird, durch die Akkumulation der Ionen, die Scangeschwindigkeit verlangsamt. Dies resultiert in einer verringerten Proteinidentifikationsrate von Orbitrap-basierten Messungen im Vergleich zu TOF-Analysen (18).

Bei der Proteomanalyse von Geweben können zusätzlich durch die verwendete Gewebekonservierungstechnik vor der Proteinextraktion Batcheffekte induziert werden. Die Verfügbarkeit von Frischgeweben (FF) ist, durch ihre zeitlich begrenzte und kostenintensive Lagerung, limitiert. Die Fixierung mit Formalin und die Paraffineinbettung (FFPE) ist ein weltweiter Standard für die kostengünstige Konservierung, Lagerung und Aufbereitung von Gewebe für die Histologie. Allgemein ist eine effiziente Proteomanalyse aus FFPE- und FF-Material möglich. Durch die, im Gegensatz zu Frischgewebe, verringerte Effizienz der Proteinextraktion auf FFPE-Gewebe und die Induktion irreversibler chemischer Modifikationen an Makromolekülen während der Formalin-Fix -ierung stellt FFPE Material allerdings eine Herausforderung für die massenspektro -metrische Analyse dar (33). Im Rahmen dieser Arbeit wurden zum Beispiel aus FFPE-Kleinhirntumoren von hGFAP-cre::SmoM2Fl/+-Mäusen und Kontroll-Kleinhirnen von SmoM2Fl+-Wurfgeschwistern durchschnittlich 40% weniger Proteine identifiziert als bei vergleichbaren Messungen aus Frischgewebe (Abschnitt 3.3). Gleichzeitig konnte

eine klare Unterscheidbarkeit zwischen Tumoren und Kontrollen auf Basis extrahierter Proteine aus beiden Gewebekonservierungstechniken festgestellt werden (Abbildung 3.17 a)). Nach Integration von Proteinabundanzen aus FFPE- und FF-Material zeigte sich ein deutlicher, die phänotypischen Differenzen zwischen Tumor und Kontrolle supprimierender, Batcheffekt. Des Weiteren konnte für den Mausmedulloblastom-Daten -satz auch innerhalb einer Gewebskonservierungstechnik ein klarer Unterschied in Abhängigkeit der Analysezeitpunktes der Proben beobachtet werden. Zusammenfassend konnten im Rahmen dieser Arbeit, bei der Integration unabhängig generierter Proteomdaten für alle analysierten Datensätze signifikante Batcheffekte beobachtet werden. Als Folge wird eine Batcheffektkorrektur zwischen unabhängig generierten Proteomdatensätzen für die valide statistische Analyse integrierter Daten dringend empfohlen.

Im Gegensatz zu anderen "Omics"-Typen weisen Proteomdaten hohe Raten fehlender Werte auf. Diese können verschiedenen Fehlwertklassen zugeordnet werden („Missing at random" (MAR), „Missing not at ranom" (MNAR) oder „missing completely at random" (MCAR)). Die bekanntesten Ansätze zur Batcheffektreduktion sind die Funktion *RemoveBatcheffects()*- des Limma-Algorithmus (47) und der Combat-Algorithmus (21). Beide Strategien akzeptieren Fehlwerte des MAR- und MCAR-Typen. Die Anwendbarkeit solcher Verfahren zur Batcheffektkorrektur in Proteomdatensätzen wurde bereits mehrfach für die Entfernung von Batcheffekten zwischen verschiedenen Analysezeitpunkten publiziert (37; 43; 54) und konnte im Rahmen dieser Studie für die Batcheffektkorrektur zwischen verschiedenen LC-MS-Konfigurationen (Abbildung 3.3.), Quantifizierungstechniken, Analysezeitpunkten und Gewebskonservierungstechniken bestätigt werden (ergänzende Abbildungen 6.1.-6.6). Einschränkend ist, dass Fehlwerte des MNAR-Typen nicht toleriert werden und vor der Batcheffektkorrektur eliminiert werden müssen.

Bei den vier in dieser Studie untersuchten Datensätzen konnte festgestellt werden, dass zwischen 42.7 und 71.6 % aller identifizierten Peptide/Proteine in mindestens einem der einbezogenen Batches fehlten (Abbildung 3.4.) und als MNAR-Typ-Fehlwerte klassifiziert werden können. Die Reduktion dieser Datensätze auf Proteine, die mit bestehenden Strategien zur Reduzierung von Batcheffekten (Combat, Limma) kompatibel sind, ist daher mit einem erheblichen Verlust an relevanten biologischen Informationen verbunden. Ein Beispiel dafür, stellt die Datenintegration von FFPE- und FF-Gewebeproben zu unterschiedlichen Analysezeitpunkten für den Mausmedulloblastom-Datensatzes dar.
Hedgehoge (SHH)-Typ Medulloblastome, werden durch Mutationen im SHH Signalweg induziert. Das im Rahmen dieser Studie untersuchte hGFAP-cre::SmoM2Fl/+
-Mausmodell bildet in Folge einer Mutation des SHH-Signalwegeffektors Smoothened zerebellare Tumore aus (50). Eine molekulare Analyse des SHH-Signaltransduktions
-netzwerkes ist in diesem Kontext biologisch relevant. Nach der Integration von Daten aus unterschiedlichen Gewebskonservierungstechniken und Analysezeitpunkten, zeigten nur 28% aller Proteine des SHH-Signaltransdukitionsnetzwerkes keine Fehlwerte des MNAR-Typen. Folglich reduzierte sich die Zahl berücksichtigbarer Faktoren des SHH-Signalweges nach der Batcheffektkorrektur um >70%, derweil die statistische Validität als differentiell abundant klassifizierter Proteine zwischen Tumor und Kontrolle durch die Batcheffektkorrektur signifikant stieg.
Ein weiteres Exempel statuiert die Datenintegration von „Lable free quantification“ (LFQ), DDA, „Tandem Mass Tag“ (TMT) und „Stable isotope labling by amino acids in cell culture“ (SILAC) (Abbildung 3.5 b)). (55) Messungen an Cetuximab-stimulierten Zellen der Kolorektalkarzinomzellinie DiFi. Hier reduzierte sich die Abdeckung des durch Cetuximab inhibierten EGFR-Signalnetzwerkes durch die mangelnde MNAR-Typ-Fehlwerttoleranz etablierter Batcheffektkorrekturverfahren von 76% auf 35% aller, einem EGFR-Signalnetzwerk assoziierten Faktoren.

Zu beachten ist, dass die Vorverarbeitung der Proteomdaten einen signifikanten Einfluss auf die Zahl der identifizierten Proteine ohne MNAR-Typ-Fehlwerte und folglich die Zahl der Proteine, welche nach Anwendung von etablierten Batcheffektkorrekturverfahren für Folgeanalysen verwendet werden können hat. So erhöhte die gemeinsame Prozessierung von LC-MS/MS-Rohdaten, unter Verwendung des Minora-Algorithmus (40), die Zahl der in allen Batches identifizierter Proteine des Mausmedulloblastom-Datensatzes von 28.4% auf 75.6% der insgesamt identifizierten Proteine (Abbildung 3.5.). Eine solche gemeinsame Datenprozessierung ist allerdings auf Daten limitiert, welche mit identischen LC-MS-Konfigurationen und Qualifizierungstechniken aufgenommen wurden.

Zusätzlich konnte im Rahmen dieser Arbeit festgestellt werden, dass auf Peptidebene eine erwartete, deutlich geringere Überschneidung identifizierter Faktoren zwischen Batches festgestellt werden als auf Proteinebene (Abbildung 3.4.). Allgemein wird die Korrektur von Batcheffekten auf Peptidebene empfohlen. Grund dafür ist, dass Peptid-Rohdaten stärkere Batcheffekte zeigen als gefilterte Proteindaten. Die Quantifizierung von Proteinen in unterschiedlichen Proben kann auf unterschiedliche Peptide zurückzuführen sein, die demselben Protein zugeordnet werden können. Sind Batcheffekt-Informationen nicht konsistent über die Peptide für jedes Protein verteilt, so kann dies zu späteren Fehlern bei der Batcheffektkorrektur auf Proteinebene führen (44).

Als Alternative zur Datenreduktion können fehlende Werte imputiert werden, um vor der Batcheffektkorrektur eine komplette Datenmatrix zu erzeugen. Bei der Datenimputation werden fehlende Werte durch einen nach einer bestimmten Regel definierten Wert ersetzt. Im Rahmen dieser Studie konnten sowohl nach Imputationen über die Normalverteilung als auch bei der Verwendung der „Random Forest“ (RF)-Imputation, vor der Batcheffektkorrektur Batcheffekte zwischen mit unterschiedlichen LC-MS -Konfigurationen vermessenen definierten Phänotypen (Phänotyp 1: 80 % *Homo sapi-*

ens, 15 % *E. Coli*, Phänotyp 2:5 % *Saccaromyces cerevisea* und 80 % *Homo sapiens*, 10 % *E. Coli*, Phänotyp 2:10 % *Saccaromyces cerevisea*), effizient reduziert werden. Allerdings konnte nur nach der Anwendung der RF-Imputation vor der Batcheffektkorrektur ein Korrelationskoeffizient > 0.98 zwischen mit unterschiedlichen LC-MS-Setups vermessenen Phänotypreplikaten erzielt werden, derweil die Korrelation zwischen Phänotypreplikaten innerhalb eines Batchs konstant blieb (Abbildung 3.8 - 3.10). Für alle verwendeten Imputationsstrategien konnte nach der Batcheffektkorrektur eine deutliche Erhöhung p-Wert-signifikanter Proteine zwischen Phänotyp 1 und Phänotyp 2 festgestellt werden (Abbildung 3.14.). Als negativ zu bewerten ist die Reduktion Students-t-Test-signifikanter Proteine zwischen Phänotyp 1 und Phänotyp 2 nach der Batcheffektkorrektur für einzelne Technische Konfigurationen (Batches), welche in Folge aller verwendeten Imputationsverfahren nach der Batcheffektkorrektur beobachtet werden konnte. So wurden nach RF-basierter Batcheffektkorrektur innerhalb einzelner Batches zwischen 12 und 27% weniger Proteine als signifikant differentiell abundant zwischen Phänotyp 1 und Phänotyp 2 klassifiziert (Abbildung 3.15.) als in der Analyse unkorrigierter Datensätze. Die Veränderung batchinterner Metriken indiziert eine imputationsbedingte Verzerrung einzelner Datenpunkte während der Batcheffektkorrektur.

Allgemein ist die Imputation fehlender Werte vor der Batcheffektkorrektur als kritisch zu bewerten. Vor allem, da im Fall integrierter, unabhängig generierter Proteomdatensätze MAR-, MCAR- und MNAR-Typ-Fehlwerte gleichzeitig imputiert werden müssen. Eine effiziente, gleichzeitige Imputation beider Fehlwerttypen kann nur durch die Nutzung multipler Imputationsmethoden erfolgen. Die Verwendung multipler Imputationstechniken wird als fehleranfällig eingestuft (19; 24), besonders da die meisten verfügbaren multiplen Imputationsverfahren der MAR-Hypothese folgen (41). So zeigten z.B. Gardner et al. (2017) in einer Simulation, dass die multiple Imputation vor allem dann fehleranfällig ist, wenn hohe Anteile an MNAR-Typ-Fehlwerten vorliegen. Im Rahmen der von Gardner et al. (2017) vorgestellten Arbeit, konnten nach der ar-

tifizieller Induktion von 8.5% MNAR-Typ-Fehlwerten in einem Proteomdatensatz aus Triple-negativem Brustkrebs nur 40.1% der im Originaldatensatz als statistisch signifikant differentiell abundanten Proteine als p-Wert-signifikant eingeordnet werden (12).

Als eine von der Datenreduktion und dem Einführen artifizieller Werte unabhängige Lösung des Fehlwerttoleranz-Problems wurde in dieser Arbeit das Matrix-Dissektions -verfahren zur Batcheffektkorrektur zwischen integrierten, unabhängig generierten, Proteomdatensätzen vorgestellt. Dabei wird eine aus integrierten Datensätzen erzeugte Eingabematrix in Abhängigkeit der Batchverteilung von Proteinen/Peptiden in Untermatrizen zerlegt, sodass in keiner Untermatrix MNAR-Typ-Fehlwerte resultieren. Für jede Untermatrix wird eine unabhängige Batcheffektkorrektur mit etablierten Algorithmen durchgeführt. Korrigierte Untermatrizen werden rekombiniert, sodass eine Ausgabematrix mit batchkorrigierten Werten und unveränderten Fehlwerten entsteht (Abbildung 3.11.). In das Framework können prinzipiell alle MAR- und MCAR-Typfehlwerttoleranten Batcheffektkorrekturalgorithmen eingebunden werden.

Die Implementierung von ComBat und der *RemoveBatcheffects()*- Funktion des Limma-Algorithmus ermöglicht das Anpassen von normalverteilten Daten, durch ein lineares Regressionsmodell (47) oder ein parametrisches, empirisches Bayesian-Framework (21). ComBat ermöglicht zusätzlich die Korrektur von nicht-normalverteilten Daten über die Integration eines nicht-parametrischen empirischen Bayesian-Frameworks (21). Für die ComBat-basierte Datenkorrektur kann entweder eine modellbasierte Lage-/Skalen -anpassungen (location and scale adjustment, L/S) oder eine rein Mittelwert-basierte Datenanpassung vorgenommen werden. Dabei ist das verwendete L/S-Modell auch mit kleinen Batchgrößen kompatibel und kann folglich zur Anpassung unabhängiger Proteomdatensätze mit kleinen Stichprobengrößen verwendet werden (21).

Das Matrix-Dissektionsverfahren reduzierte effizient Batcheffekte zwischen verschiedenen LC-MS-Setups. Wie für die Batcheffektkorrektur nach RF-Imputation, konnte nach Anwendung des Matrix-Dissektionsverfahrens ein Korrelationskoeffizient > 0.98 zwischen mit unterschiedlichen LC-MS-Setups vermessenen Phänotypreplikaten erzielt werden, derweil die Korrelation zwischen Phänotypreplikaten innerhalb eines Batchs konstant blieb (Abbildung 3.12). Im Gegensatz zum RF-basierten Ansatz blieb die Zahl differentiell abundanter Proteine zwischen den betrachteten Phänotypen innerhalb einzelner Batches, allerdings vor und nach der Korrektur, konstant (Abbildung 3.14.; Abbildung 3.15.) Da das Matrix-Dissektionsverfahren effizient Batcheffekte reduziert, statistische Metriken zwischen Phänotypen integrierter Datensätze verbessert und gleichzeitig die Datenstruktur innerhalb einzelner Batches nicht verzerrt, kann es als valide Technik zur fehlwerttoleranten Entfernung von Batcheffekten eingeordnet werden. Des Weiteren konnte gezeigt werden, dass die Nutzung des Matrix-Dissektionsverfahrens Imputations-basierten Strategien zur Korrektur von Batcheffekten überlegen ist.

Dies ist nur dann einzuschränken, wenn eine vollständige Datenmatrix für die nachfolgenden Schritte der statistischen Datenanalyse zwingend erforderlich ist. So konnte festgestellt werden, dass eine RF-Imputation nach Matrix-Dissektionsverfahren-basierter Batcheffektkorrektur, Batcheffekte reimplementiert und den linearen Zusammenhang der Proteinabundanzverteilung zwischen unterschiedlichen Proben verzerrt (Abbildung 3.13.) In solchen Fällen ist die RF-basierte Imputation mit folgender Batcheffektkorrektur dem Matrix-Dissektionsverfahren vorzuziehen.

Neben Batcheffekten zwischen unterschiedlichen LC-MS-Setups konnten auch technische Varianzen zwischen verschiedenen Quantifizierungsansätzen, verschiedenen Gewebskonservierungstechniken und Analysezeitpunkten erfolgreich durch das Matrix -Dissektionsverfahren angeglichen werden. In allen Fällen konnte im korrigierten Datensatz ein Phänotyp-basiertes Clustering beobachtet werden (Abbildung 3.13.; Abbil-

dung 3.17.; Abbildung 3.20.; Abbildung 3.21.; Abbildung 3.25.). Erzielte Ergebnisse zeigten eine Vergleichbarkeit mit der Anwendung des unmodifizierten ComBat-Algorithmus (Abbildung 3.4.; Ergänzende Abbildung 6.1-6.6), berücksichtigten jedoch eine deutlich höhere Anzahl von Proteinen. Darüber hinaus wurden erwartete biologische Eigenschaften einzelner Phänotypen nach der Batcheffektkorrektur biologisch korrekt widergespiegelt. Dies zeigt sich zum Beispiel an einem 2021 von Petrialia et al. veröffentlichten Proteomdatensatz an verschiedenen humanen Hirntumor-Entitäten, nach der Matrix-Dissektionsverfahrens-basierten Batcheffektkorrektur zwischen 23 Messbatches (43). Wie in der Originalpublikation bei der Untersuchung der Krebssignalwege MYC, EF2 und WNT beschrieben, konnte nach Batcheffektkorrektur eine erhöhte Abundanz MYC-und E2F-assoziierter Proteine in Medulloblastomen und atypischen teratoiden/rhabdoiden Tumoren (ATRT) festgestellt werden. Für den WNT-Signalweg wurde die beschriebene niedrigere Abundanz in Niedriggradigen Gliomen (LGG) und Hochgradigen Gliome (HGG), sowie Gangliogliomen beobachtet (Abbildung 3.26.). Des Weiteren zeigte sich nach der Batcheffektkorrektur zwischen unterschiedlichen Gewebskonservierungstechniken und Analysezeitpunkten eine deutliche Erhöhung der statistischen Signifikanz für den Hedgehog (SHH)-Medulloblastom-Biomarker Filamin A (10) zwischen Kleinhirntumoren von hGFAP-cre::SmoM2Fl/+-Mäusen und Kontroll-Kleinhirnen von
SmoM2Fl+-Wurfgeschwistern, im Vergleich zu integrierten, nicht batcheffektkorrigierten Daten (Abbildung 3.18.).

Für die Datenintegration von auf Basis unterschiedlicher Gewebskonservierungstechniken identifizierten Proteinen des Mausmedulloblastom-Datensatzes zu unterschied
-lichen Analysezeitpunkten konnte zusätzlich festgestellt werden, dass für normalver
-teilte Daten das parametrische, empirische Baesian-Framework des ComBat
-Algorithmus und das lineare Regressionsmodell des Limma-Algorithmus ähnliche Resultate erzielten. So verdoppelte sich in beiden Fällen die Zahl signifikanter Protei-

ne nach Batcheffektkorrektur. 95 % aller als signifikant (p-Wert < 0.05) klassifizierten Proteine wurden nach Anwendung beider Algorithmen im Framework des Matrix-Dissektionsverfahrens als reguliert zwischen Kleinhirntumoren von hGFAP-cre::-SmoM2Fl/+-Mäusen und Kontroll-Kleinhirnen von SmoM2Fl+-Wurfgeschwistern eingestuft.

Eine besondere Herausforderung bei der Integration unabhängig generierter Proteomdatensätze stellt die Verwendung unterschiedlicher Quantifizierungstechniken dar. So konnte hier beispielsweise gezeigt werden, dass „Spike-in"-SILAC-Daten nur dann mit TMT- und LFQ-quantifizierten Daten integriert werden können, wenn die quantitativen Werte für jedes Protein aus TMT- und LFQ-Daten auf ihre mittlere Häufigkeit, über alle Proben hinweg, normalisiert wurden. Im Rahmen der Spike-in- oder super-SILAC-Quantifizierung werden Proteinabundanzen als Verhältnis zwischen einer Probe und einem schwer-isotopenmarkierten internen Standard dargestellt, welcher die mittlere Proteinabundanz über alle Proben repräsentieren soll (13; 51). In Ziellinien-basierten Experimenten wird hierzu meist eine schwer markierte Mischung aller vermessenen Phänotypen verwendet (13). Im Rahmen dieser Studie konnte festgestellt werden, dass durch die Mittelwertnormalisierung in TMT- und LFQ-Messungen dieses Verhältnis vor der Batcheffektkorrektur imitiert werden kann.

Nach Stepath et al. (2020) (55) ist nach 24h Cetuximab-Behandlung der Kolorektalkarzinomzelllinie DiFi eine deutliche Unterscheidbarkeit zu unbehandelten Kontrollen und 0 bis 3h behandelten Proben zu erwarten. Dies konnte individuell für LFQ- und SILAC-Messungen bestätigt werden (55). Für die Integration von Proteomdaten aus diesen Messungen und TMT-basierten Messungen derselben Proben, konnte nach Mittelwertangleich von TMT- und LFQ-Daten an SILAC-Verhältnisse mit anschließender Batcheffektkorrektur durch das Matrix-Dissektionsverfahren eine klare Abgrenzung des 24h Cetuximab-behandelten Zellen beobachtet werden (Abbildung 3.28). Darüber hinaus

unterstrich die deutliche Verringerung des Gesamt-Varianzkoeffizienten (Coeficcient of Variation, CV) für das Housekeeping-Protein Nuxid Hydrolase 21 (NUDT21) (28) auf einen Erwartungswert <10% (Abbildung 3.28) eine erfolgreiche Batcheffektkorrektur durch das Matrix-Dissektionsverfahren. Daraus kann geschlussfolgert werden, dass bei Beachtung der korrekten Normalisierung von Einzeldatensätzen vor der Batcheffektkorrektur eine Integration von Proteomdaten aus verschiedenen Quantifizierungsansätzen möglich ist.

Des Weiteren konnte das Matrix-Dissektionsverfahren als wirksames Verfahren zur Reduktion von Batcheffekten zwischen verschiedenen TMT-Plexen etabliert werden. Die TMT-basierte Quantifizierung ermöglicht die gleichzeitige Messung von bis zu 16 Proben durch die Kopplung von probenspezifischen schweren Isotopenmarkierungen an N-terminale primäre Amine von Peptiden sowie Lysinresten. Diese "Multiplex"-Analyse ermöglicht eine erhöhte Reproduzierbarkeit, Vergleichbarkeit und Quantifizierungsgenauigkeit zwischen Proben. Grund dafür ist, dass auf gemeinsam vermessene Proben identische Umwelteinflüsse wirken. Es entsteht folglich keine, durch diese bedingte, technische Varianz zwischen Einzelproben. Darüber hinaus ermöglicht die gleichzeitige Messung mehrerer Proben, die Verwendung geringerer Protein/Peptidmengen pro Probe (36). Die TMT-basierte Quantifizierung von Peptiden/Proteinen ist durch die maximale Verfügbarkeit von 16 verschiedenen Isotopenmarkierungen limitiert. Werden im Rahmen größer angelegter Studien höhere Probenzahlen analysiert, so müssen mehrere TMT-Plexe vermessen werden. Die Integration verschiedener TMT-Batches induzierte Batcheffekte in integrierten Datensätzen, welche eliminiert werden müssen. Dabei sind Multibatch-TMT Messungen, neben einer hohen Falsch-Positiv-Rate besonders durch die hohe Zahl induzierter MNAR-Typ-Fehlwerte limitiert. So konnten Brenes et al. (2019) (3) anhand von induzierten, pluripotenten Stammzellen zeigen, dass die Integration von fünf TMT-Plexen im Mittel auf Peptidebene 40% Fehlwerte und auf Proteinebene >10% Fehlwerte induziert. Im Rahmen dieser Studie zeigten zum Beispiel für

den Cetuximab-Datensatz, 75% aller Peptide mindestens einen MNAR-Typ-Fehlwert in integrierten Daten, derweil auf Proteinebene 42% aller Proteine nicht in allen vier Batches identifiziert wurden (55) (Abbildung 3.4.). Für den humanen Hirntumor-Daten -satz (43) konnten 58% aller identifizierten Proteine in allen 23 Batches gefunden werden (Abbildung 3.4.). Da die Zahl an MNAR-Typ- Fehlwerten linear mit der Anzahl integrierter Batches korreliert, sollte vor allem für groß angelegte TMT-Studien eine fehlwerttolerante Methode zur Batcheffektreduktion angewandt werden.

Aktuell wird zur Batcheffektreduktion zwischen unterschiedlichen TMT-Plexen vor allem die interne Referenzskalierung ("internal referece scaling", iRS) verwendet. Dabei wird in jedem TMT-Batch ein interner Standard integriert. Die Batcheffektkorrektur erfolgt durch die Normalisierung jedes Peptid/Proteins einer Probe auf die mittlere Abundanz des Peptids/Proteins in der Referenz-Probe des jeweiligen Batches (3). Mit dem Matrix-Dissektionsverfahren, unter Einbindung des ComBat-Algorithmus, konnte im Vergleich zu iRS eine effizientere Entfernung von Batcheffekten erzielt werden.

Nach Stepath et al. (2019) konnte für 24h-Cetuximab-inkubierte Zellen der Kolorektalkarzinomzellinie DiFi, in TMT-basierten Messungen mit iRS-basierter Batcheffektkorrektur, kein erwarteter Unterschied zu unbehandelten Zellen (55) nachgewiesen werden. Dieses Ergebnis ließ sich im Rahmen dieser Studie auf Peptid- und Proteinebene reproduzieren. Im Vergleich dazu führte die Anwendung des Matrix-Dissektionsverfahrens zu einer klaren Differenzierung des Proteomprofils 24h-Cetuximab-inkubierter Zellen von Kontrollproben und 0-3h mit Cetuximab behandelten Zellen (55) (Abbildung 3.20.; Abiildung 3.21). Des Weiteren konnten für NUDT21 auf Peptid- und Proteinebene deutlich geringere CV-Werte nach Verwendung des Matrix-Dissektionsverfahrens festgestellt werden als nach iRS-basierter Datenkorrektur (Abbildung 3.22.; Abildung 3.23). Die ComBat-/Limma-basierte Batcheffektreduktion für TMT-Daten zieht im Vergleich zu iRS mehr Datenpunkte pro Batch heran. Dadurch ist sie weniger anfällig für ex-

perimentelle Fehler und Varianzen in internen Referenzen. Zusätzlich werden im Vergleich zu iRS durch die Verwendung der L/S-Skalierung nicht nur Mittelwerte, sondern auch Varianzen zwischen Batches in der ComBat-basierten Batcheffektkorrektur berücksichtigt. Ein weiterer Vorteil ist, dass die mangelnde Notwendigkeit identischer Referenzstandards in allen Batches die Versuchsplanung erleichtert und die nachträgliche Integration unabhängig erzeugter TMT-Batches ermöglicht. Während ComBat bereits in früheren Studien zur Kompensation von TMT-Batcheffekten verwendet wurde (37; 43; 54), ermöglicht das Matrix-Dissektionsverfahren die Integration von Multibatch-TMT-Daten, ohne die Notwendigkeit der Imputation oder Datenreduktion.

Zusammenfassend ermöglicht das Matrix-Dissektionsverfahren die fehlwerttolerante Harmonisierung von Daten über verschiedene massenspektrometrische Setups, Quantifizierungsplattformen, Analysezeitpunkte und Gewebskonservierungstechniken.
Dadurch können unabhängig voneinander generierte Proteomdatensätze integriert und gemeinsam analysiert werden, was die statistische Validität assoziierter Hypothesen, besonders für seltene Erkrankungen, signifikant erhöht. Allgemein ermöglicht das Framework durch die Implementierung verschiedener, etablierter Algorithmen die Batcheffektkorrektur von Gauß- und nicht Gauß-verteilten Proteomdaten, unabhängig von der Verfügbarkeit von Spektraldaten. Derweil das Matrix-Dissektionsverfahren für Proteomdaten etabliert wurde, kann das grundlegende Prinzip für alle MAR- und MCAR-Typ-fehlwerttoleranten Batcheffektkorrekturstrategien, Datenmodalitäten und wissen -schaftlichen Fragestellungen adaptiert werden.

5

Material und Methoden

5.1 | Methoden

5.1.1 | Öffentlich verfügbare Datensätze

Um die Anwendbarkeit des Matrix-Dissektionsverfahrens für die Reduzierung des Batcheffekts zwischen verschiedenen Quantifizierungstechniken zu testen, wurde ein von *Stepath et al. (2020)* (55) veröffentlichter Datensatz verwendet. Der Datensatz enthält unnormalisierte Peptid- und Proteinabundanzen von der Kolorektalkarzinom-Zelllinie DiFi mit und ohne Cetuximab-Behandlung nach 0, 3 und 24 Stunden Inkubation. Die Quantifizierung erfolgte mittels Stable Isotope Labling by Amino Acids in Cell Culture (SILAC), Tandem Mass Tag (TMT) oder Label-freier Quantifizierung (LFQ) im datenabhängigen Akquisitions-Modus (DDA) unter Verwendung eines Quadrupol-Orbitrap-Hybrid-Massenspektrometers (QExactive, Thermo Fisher Scientific, Bremen, Deutschland). Der Datensatz kann über das PRIDE-Archiv (PXD014565) abgerufen werden. Um TMT-Batch-spezifische Varianzen in einer größeren biologischen Kohorte zu analysieren, wurde zusätzlich ein von *Petralia et al. (2021)* ((43)) publizierter Datensatz verwendet. Ziel der Studie war der Vergleich der Proteom-Profile von 8 verschiedenen kindlichen Hirntumor-Entitäten (Hirntumor-Datensatz). Insgesamt wurden 23 TMT-11-Plex Batches mit einem Orbitrap-Iontrap-Quadrupol-Tribrid-Massenspektrometer (Orbitrap Fusion, Thermo Fisher Scientific, Bremen, Deutschland) vermessen. Der Datensatz ist über das Clinical Proteomic Tumor Analysis Consortium Data Portal (https://cptac-data-portal.georgetown.edu/cptacPublic/) und die Proteomics Data Commons (https://pdc.cancer.gov/pdc/) abrufbar.

5.1.2 | Proteinextraktion und LC-MS/MS-Analyse zur Generierung hauseigener Daten

5.1.2.1 | Spike-in Datensatz

Zur Analyse der Integrierbarkeit mit unterschiedlichen LC-MS/MS-Konfigurationen aufgenommener Proteomdaten wurden zwei definierte Phänotypen durch die Kombination von homo sapiens, *E. Coli* und *Saccaromyces cerevisea*-Zelllysaten erzeugt. (Phänotyp 1: 80 % *Homo sapiens*, 15 % *E. Coli*, 5 % *Saccaromyces cerevisea*; Phänotyp 2: 80 % *Homo sapiens*,10 % *E. Coli*, 10 % *Saccaromyces cerevisea*)und mit unterschiedlichen LC-MS-Setups vermessen (Label-freie Quantifizierung (LFQ) im datenabhängigen Akquisitions-Modus (DDA) auf einem Quadrupol-Orbitrap-Hybrid-Massenspektrometer (QExactive, Thermo Fisher Scientific), Label-freie Quantifizierung (LFQ) im datenunabhängigen Akquisitions-Modus (DIA) auf einem Quadrupol-Orbitrap-Hybrid-Massenspektrometer (QExactive, Thermo Fisher Scientific, Bremen, Deutschland) und LFQ im Sequential Window Acquisition of All Theoretical Mass Spectra-Modus (SWATH) auf einem Triple TOF Massenspektrometer (TripleTOF 6600, Sciex, Farmingham, USA). Weitere Details zur LC-MS/MS-Konfiguration, individueller Setups können über das PRIDE-Archiv (PXD027467) abgerufen werden.

5.1.2.2 | Maus-Medulloblastom-Datensatz

Um Varianzen zwischen Proteomdaten verschiedener Gewebskonservierungstechniken und Analysezeitpunkten zu adressieren, wurden unterschiedlich aufbereitete Proben eines etablierten Sonic Hedgehog (Shh)-Medulloblastom-Mausmodells ((50)) und (hGFAP-cre::SmoM2Fl/+)-Zerebellen von SmoM2Fl+-urfgeschwistern am postnatalen Tag 13 analy -siert. Es wurden sowohl männliche als auch weibliche Mäuse verwendet. hGFAP-cre-Mäuse und SmoM2Fl/Fl-Mäuse wurden von The Jackson Laboratories (Bar Harbor, ME, USA) erworben. Die transgenen Mäuse wurden mit einem C57BL/6J-

Hintergrund gekreuzt. Alle Experimente mit Tieren wurden von der örtlichen Tierschutzkommission genehmigt (Behörde für Justiz und Verbraucherschutz in Hamburg, TVA N99/2019) genehmigt und der Umgang mit den Tieren wurde in Übereinstimmung mit den örtlichen behördlichen und institutionellen Tierschutzbestimmungen durchgeführt. Kleinhirntumore von hGFAP-cre::SmoM2Fl/+-Mäusen und Kleinhirne von SmoM2Fll+-Wurfgeschwistern wurden halbiert. Eine Hälfte wurde eingefroren und bis zur weiteren Verarbeitung bei -80 °C gelagert (frisch gefrorener Zustand (FF)). Die andere Hälfte wurde über Nacht in 4% Paraformaldehyd/PBS bei Raumtemperatur fixiert. Das Gewebe für paraffineingebettete Schnitte wurde dehydriert, eingebettet und bei 4 m nach Standardprotokollen geschnitten (FFPE). Die Histomorphologie des Tumor- oder Kleinhirngewebe wurde durch HE-Färbung überprüft. Insgesamt wurden 2 Würfe verwendet und separat gehandhabt, um zwei verschiedene Analysezeitpunkte bei identischem experimentellem Setup zu simulieren.

Für FF-Proben wurden Proteine über 10 Minuten in 200 μl 0.1M TEAB-Puffer mit 1% SDC bei 99°C bei 1400 rpm in einem ThermoMixer® C (Eppendorf, Hamburg, Deutschland) extrahiert und denaturiert. FFPE-Schnitte wurden mit 500 μl N-Heptan über einen Zeitraum von 10 Minuten deparaffinisiert. Die Proben wurden 10 Minuten lang bei 14000 x g zentrifugiert und der Überstand entfernt. Anschließend wurden Formalin-fixierte Proben in 200 μl 0.1M TEAB-Puffer mit 1% Natrium-Desoxycholat (SDC) resuspendiert. Formalinfixierungen wurden mittels Hitze über einen Zeitraum von 60 Minuten bei 99°C bei 1400 rpm in einem ThermoMixer® C (Eppendorf, Hamburg, Deutschland) revertiert.

Für FFPE- und FF-Proben wurde nach Proteinextraktion die Proteinkonzentration unter Verwendung eines Pierce TM BCA Protein assay kits nach Herstellerangaben bestimmt. Für den tryptischen Verdau wurden je 50g Protein verwendet. Disulfid Brücken wurden mit 10mM Dithrothreitol für 30 Minuten bei 60°C und 1400 rpm in einem ThermoMi-

xer® C (Eppendorf, Hamburg, Deutschland) reduziert. Freie Thiolgruppen an Cysteinen wurden anschließend mit 20mM Iodacetamid (IAA) für 30 Minuten bei 37°C und 1400 rpm in einem ThermoMixer® C (Eppendorf, Hamburg, Deutschland) im Dunkeln alkyliert. Der tryptische Verdau wurde bei einem Trypsin/Protein-Ratio von 1:100 bei 37°C und 1400 rpm in einem ThermoMixer® C (Eppendorf, Hamburg, Deutschland) über Nacht durchgeführt. 1% Ameisensäure (FA) wurde verwendet, um SDC auszufällen und den tryptischen Verdau zu stoppen. Die Proben wurden 10 Minuten lang bei 14000 x g zentrifugiert. Der Überstand wurde in ein neues Eppendorfgefäß überführt und in einer Vakuumzentrifuge (SpeedVac SC110 Savant, Thermo Fisher Scientific, Bremen, Deutschland) getrocknet.

Unmittelbar vor der LC-MS/MS-Messung wurden die lyophilisierten tryptischen Peptide in 0.1% FA bis zu einer finalen Konzentration von 1mg/ml resuspendiert. 1g Peptide wurden in eine Glasviole überführt, in einen Autosampler transferriert und in ein Dionex Ultimate 3000 nano-UPLC-System (Thermo Fisher Scientific, Bremen, Deutschland) injiziert. Die Peptide wurden mit einer Acclaim PepMap 100 C18-Vorsäule (100 m x 2 cm, 100 Å Porengröße,5 m Partikelgröße, Thermo Fisher Scientific, Bremen, Deutschland) aufgereinigt und entsalzt. Die chromatographische Trennung erfolgte über eine Acclaim PepMap 100 C18-analytische Säule (75 m x 50 cm, 100 Å pore size, 2 m Partikelgröße, Thermo Fisher Scientific, Bremen, Deutschland). Peptide wurden über einen linearen Gradienten, gebildet aus definierten Mengen von Puffer A (0.1% FA in H2O) und Puffer B (0.1% FA in Acetonitril) eluiert. Die Elution erfolgte über einen Zeitraum von 60 Minuten, in welchem die Konzentration von Puffer B von 3-35% lerhöht wurde. Eluierende Peptide wurden über eine nano-Elektrosprayionisations (ESI)-Ionenquelle bei einer Spannung von 1800V in ein Quadrupole-Orbitrap-Iontrap-Tribrid-Massenspektrometer überführt (Orbitrap Fusion, Thermo Fisher Scientific, Bremen, Deutschland). Vorläuferionen wurden für maximal 50 Millisekunden oder bis zum Erreichen einer maximalen Ladungsdichte (AGC Target) von 2x105 Ionen akkumuliert und in einer

Orbitrap-Ionenfalle über einen Massenbereich von m/z 400-1200 mit einer Auflösung von 120000 bei m/z 200 analysiert. Isolierte Vorläuferionen wurden in einer linearen Ionenfalle unter Nutzung der kollisionsinduzierten Dissoziation (CID) bei einer normalisierten Kollisionsenergie von 35 fragmentiert. Für jeden MS2-Scan wurden die Ionen für eine maximale Füllzeit der Ionenfalle von 50 Millisekunden oder bis zu einem AGC-Target von 5x104 Ionen akkumuliert und über einen Massenbereich von m/z 400-1200 analysiert. Die m/z-Bereiche bereits fragmentierter Vorläuferionen wurden für 30 Sekunden nach ihrer Analyse von der Fragmentierung ausgeschlossen (Dynamic Exclusion). Für jedes Vorläuferspektrum wurde eine Zykluszeit von 3 Sekunden erlaubt.

5.1.3 | Rohdatenprozessierung

5.1.3.1 | LFQ-DDA Daten

DDA-Rohspektren des Spike-In- und des Mausmedulloblastom-Datensatzes wurden mit dem in MaxQuant (Max-Planck-Institut für Biochemie, Version 1.6.2.10) integrierten Andromeda-Algorithmus prozessiert. Alle Batches wurden separat durchsucht, um unabhängig voneinander erzeugte Datensätze zu simulieren. Für den Spike-In-Datensatz wurden eine menschlichen FASTA-Datenbank (heruntergeladen von Uniprot Dezember 2017, 26559 Einträge), eine Hefe-Datenbank (*Saccaromyces cerevisea*Stamm ATCC 204508, heruntergeladen von Uniprot Dezember 2017, 6721 Einträge), und eine *coli*-Datenbank (Stamm K12, heruntergeladen von Uniprot Dezember 2017, 31400 Einträge) verwendet. Gewebeproben von Mäusen wurden gegen eine Mausdatenbank abgeglichen (heruntergeladen von Uniprot Dezember 2020, 17015 Einträge). Trypsin wurde als proteolytisches Enzym gewählt, wobei maximal 2 verpasste Schnittstellen zulässig waren. Eine minimale Peptidlänge von 6 Aminosäuren und eine maximale Peptidmasse von 6000 Da wurden festgelegt. Die Oxidation von Methionin, die Acetylierung des Protein-N-Terminus und die Umwandlung von Glutamin in Pyro-Glutaminsäure wurden als variable Modifikationen erlaubt. Die Carbamidomethylierung von Cysteinen wurde als

statische Modifikation gesetzt. In Anlehnung an das gewählte LC-MS-Setup wurde für den Spike-In-Datensatz eine Massentoleranz von 20 ppm auf Vorläuferionen-Ebene festgelegt. Für Mausproben wurden 10 ppm verwendet. Für Fragmentspektren wurde für den Spike-in-Datensatz eine Massentoleranz von 0.02 Da verwendet. Für den Maus -medulloblastom-Datensatz wurde eine Massentoleranz von 0.6 Da festgesetzt. Für die Peptididentifizierung wurde eine Falscherkennungsrate (”False discovery rate”, FDR) von <0.01 festgelegt, wobei eine umgekehrte Decoy-Peptiddatenbank zur FDR -Berechnung verwendet wurde. Auf Proteinebene wurden nur Proteine, welche mit einer FDR <0.01 identifiziert wurden zugelassen. Die Label-freie Quantifizierung wurde mit einem LFQ-Mindestverhältnis von 1 durchgeführt. Um die Vollständigkeit der Daten zu erhöhen, wurden unidentifizierte Vorläuferionen mit per MS2 identifizierten Vorläuferionen anderer Rohspektren verglichen (”Matching Between Runs”, MBR).

Um zusätzlich den Effekt einer gemeinsamen Datenbanksuche von Rohspektren unabhängig generierter Proteomdatensätze zu evaluieren, wurde der Minora-Algorithmus zur Reduktion von Batcheffekten und Erhöhung der Datenvollständigkeit in Proteomdiscoverer 2.4 (Thermo Fisher Scientific, Bremen) verwendet. Zusätzlich zu den beschriebenen Parametern für MaxQuant wurde ein Retentionszeitalignment durchgeführt. Ein maximaler Retentionszeitshift von 5 Minuten wurde erlaubt. Für Peptide, welche über MBR identifiziert wurden, wurde ein minimales Signal-zu-Rauschverhältnis von 5 vorausgesetzt.

5.1.3.2 | LFQ-DIA/SWATH Daten

Für DIA-Daten wurden DDA-LC-MS/MS-Daten des Spike-in-Datensatzes sowie einzelne DDA-Messungen tryptischer Peptide aus humanen Hefe- und *E.coli*-Lysaten verwendet, um eine Referenz-Peptidspektrenbibliothek für die Datenextraktion aus DIA /SWATH-Daten zu erzeugen. Zur Erstellung der Peptidspektrenbibliothek wurden Roh-

spektren gegen eine menschlichen FASTA-Datenbank (heruntergeladen von Uniprot Dezember 2017, 26559 Einträge), eine Hefe-Datenbank (Saccharomyces cerevisiae Stamm ATCC 204508, heruntergeladen von Uniprot Dezember 2017, 6721 Einträge), und eine *Escherichia coli*-Datenbank (Stamm K12, heruntergeladen von von Uniprot Dezember 2017, 31400 Einträge) in Proteome Discoverer 2.0 gesucht. Die Massentoleranzen für die Vorläufer wurden auf 10 ppm und für die Fragmentionen auf 0.02 Da festgelegt. Die Carbamido -methylierung wurde als feste Modifikation für Cysteinreste festgelegt. Die Oxidation von Methionin, die Pyroglutamatbildung an Glutaminresten sowie die Acetylierung des Protein-N-Terminus, der Methioninverlust am Protein-N-Terminus und die Acety -lierung nach Methioninverlust am Protein-N-Terminus wurden als variable Modifikationen zugelassen. Für die Peptididentifizierung wurde eine Falscherkennungsrate (FDR) von <0.01 festgelegt, wobei eine umgekehrte Decoy-Peptiddatenbank zur FDR

-Berechnung verwendet wurde. Die Proteome Discoverer-Suchergebnisse wurden in die Skyline-Software (Version 4.2, MacCossLab Software, Washington, USA) importiert wobei nur Peptide mit mehr als 4 Fragmentionen zugelassen wurden. Für die Datenextraktion aus den DIA-Rohspektren für Peptide wurden maximal 5 Fragmentionen pro Peptid verwendet. Für die weitere Analyse wurden nur Peptide mit einem ”dot product” >0.85 zugelassen. Relative Proteinabundanzen wurden durch die Summe der ächen unter der Kurve („Area under the curve,“ AUC) im extrahierten Ionen

-Chromatogramm (XIC) aller identifizierten Peptide eines Proteins in einer Probe ermittelt.

5.1.4 | Normalisierung und Integration individueller Datensätze

Zwecks Normalisierung wurden unnormalisierte Proteinabundanzen individueller Datensätze in Perseus (Version 1.5.8.5, Max-Planck-Institut für Biochemie, München, Deutschland) geladen. Jedes Experiment und jeder Batch wurde separat gehandhabt.

Im Folgenden wurden Proteinabundanzen normalisierter Datensätze anhand ihrer UniProt-Identifikationsnummer kombiniert.

5.1.4.1 | Spike-in Datensatz

Für den Spike-in-Datensatz wurden vor der Batcheffekt-Reduktion zwischen SWATH-, DIA- und DDA-Daten relative Proteinabundanzen log2-transformiert. Aufgrund der vorliegenden trimodalen Wahrscheinlichkeitsverteilung wurde keine weitere Normalisierung durchgeführt.

5.1.4.2 | Maus-Medulloblastom-Datensatz

Für den Maus-MedulloblastomDatensatz wurden die relativen Proteinabundanzen log2-transformiert und für jede Einzelprobe über den probenspezifischen Median normalisiert.

5.1.4.3 | Cetuximab-Datensatz

Für mit Cetuximab behandelten Kolorektalkarzinom Zellen (Zelllinie DiFi) (55)) wurde zunächst der Batcheffekt zwischen TMT-8-Plex-Batches reduziert. TMT-Reporter -intensitäten wurden log2-transformiert und für jede Einzelprobe über den probenspezifischen Median normalisiert. (55) Zur Batcheffekt-Reduktion wurde „internal reference scaling“ (iRS) verwendet, wobei durch Division der Reporterionenintensitäten durch das arithmetische Mittel der Referenzproben jedes Batches eine Reduktion des Batcheffektes erzielt wurde. Als alternatives Verfahren wurde das Matrix-Dissektionsverfahren zur Batcheffektreduktion verwendet. (Siehe Abschnitt 5.1.5.)
Für die weitere Analyse wurden ausschließlich mittels Matrix-Dissektionsverfahren korrigierte TMT-Daten verwendet. Proteinabundanzen aus DDA- und SILAC-Messungen wurden log2-transformiert und für jede Einzelprobe über den probenspezifischen Median normalisiert. DDA- und TMT-basierte Proteinabundanzen wurden zusätzlich, zur

Nachahmung der Spike-in-SILAC-Verhältnisse, vor der Datenintegration über den Mittelwert jedes Proteins zentriert.
Auf Peptidebene wurden Peptidabundanzen unabhängig generierter TMT-8-Plex-Batches anhand ihrer zugeordneten Peptidsequenz integriert.

5.1.4.4 | Hirntumor Datensatz

Für die Hirntumor wurden TMT-Reporterintensitäten log2-transformiert und für jede Einzelprobe über den probenspezifischen Median normalisiert.

5.1.5 | Anwendung des Matrix-Dissektionsverfahrens zur Batcheffektkorrektur zwischen unabhängig generierten Proteomdatensätzen

Das Matrix-Dissektionsverfahren implementiert die *RemoveBatcheffects()*-Funktion des Limma-Packages sowie den ComBat-Algorithmus zur Reduktion von Batcheffekten (21) (47). Normalisierte, über UniProt-Identifikationsnummern kombinierte Datenmatritzen wurden für die Batcheffekt-Reduktion verwendet. Im Rahmen des Matrix-Dissektions -verfahrens wurden auf der Grundlage der Anzahl der vorhandenen numerischen Werte innerhalb jedes Batches Untermatrizen erzeugt. Ein Protein wurde für einen bestimmten Batch verworfen, wenn <2 Werte identifiziert wurden. Die Batcheffektreduktion erfolgt für jede Untermatrix separat, unter Anwendung des durch den Nutzer gewählten Algorithmus zur Batcheffekt-Korrektur. Die batcheffektrreduzierten Untermatritzen werden anschließend wieder zusammengefügt und eine Ausgabematrix wird erstellt. Proteine, welche nur in einem Batch identifiziert werden, werden keiner Korrektur unterzogen und der Ausgabematrix angehangen. Fehlende Werte werden nicht korrigiert oder verändert.
Alle Datensätze mit Ausnahme des Spike-in-Datensatzes wurden im Rahmen des Matrix-Dissektionsverfahrens unter Verwendung des parametrischen Bayesian-Frameworks mit

L/S-Scaling verarbeitet, welches im ComBat-Algorithmus(21) integriert ist. Zu Vergleichszwecken wurde zusätzlich das lineare Regressionsmodell(47) der *RemoveBatcheffects()*-Funktion des Limma-Algorithmus für den Maus-Medulloblastom-Datensatz angewandt. Dabei wurde für alle verwendeten Datensätze eine Gaußsche Normalverteilung vorausgesetzt.
Für den Spike-in-Datensatz war eine trimodale Wahrscheinlichkeitsverteilung gegeben. Daher wurde der nichtparametrische empirische Bayes-Rahmen mit L/S-Scaling im Rahmen des ComBat-Algorithmus zur Batcheffekt-Reduktion verwendet.

5.1.6 | Datenimputation für den Spike-in-Datensatz

Für den Vergleich des Matrix-Dissektionsverfahrens mit etablierten Strategien zur Batcheffektreduktion wurden drei verschiedene Imputationsstrategien auf den Spike-in-Daten
-satz vor der Batcheffektreduktion angewendet. 1. Matrix- und 2. Spaltenweise Imputation über die Normalverteilung: Die Imputation wurde in Perseus (Max-Planck-Institut für Biochemie, Version 1.5.8.5) (59) mit einer Breite von 0.3 und einer Wertreduktion um 1.8 durchgeführt. 3. Die Random-Forest-Imputation wurde mit dem in der R-Softwareumgebung implementierten Paket "RandomForest" (5) durchgeführt. Die Imputation wurde mit 1000 Bäumen in 10 Iterationen durchgeführt.
Um die Imputation fehlender Werte nach Anwendung des Matrix-Dissektionsverfahres zu testen, wurden identische Parameter für die Random-Forest-Imputation gewählt.
Zur Reduktion des Batcheffektes zwischen unabhängig generierten Proteomdatensätzen wurden für imputierte Daten in 5.1.5 beschriebene Parameter für den ComBat-Algo
-rithmus (21) und die *RemoveBatcheffects()*-Funktion des Limma-Algorithmus (47) gewählt.

5.1.7 | Statistische Analyse und Datenvisualisierung

Alle in dieser Studie integrierten Students t-Tests wurden mit der Perseus-Software durchgeführt (59). Proteine, die mit einem p-Wert < 0.05 zwischen den verglichenen Phänotypen identifiziert wurden, wurden als statistisch signifikant differentiell abundant klassifiziert. Für hierarchisches Clustering wurde die Pearson-Korrelation als Korrelationsmetrik genutzt. WardD-Linkage wurde als Verknüpfungsmethode gewählt. Heatmaps wurden mit dem mit dem ”pheatmap”-Paket (Version 1.0.12) in der R-Software -umgebung (Version 4.0.4) erstellt. Die fehlwerttolerante ”non linear itterative squares” (NIPALS) Hauptkomponentenanalyse (PCA) wurde unter Nutzung der im ”mixomics”-Algorithmus (Version 6.20.0) integrierten NIPALS-PCA-Funktion (48) durchgeführt. Die Streudiagrammverteilungen der Proben über verschiedene Hauptkomponenten wurden in PRISM (GraphPad, Version 5) visualisiert. Venn-Diagramme wurden mit Venny (BioinfoGP, Version 2.1.0) erstellt. Die Häufigkeitsverteilungen einzelner Proteine wurden mit Microsoft Excel (Version 16.5) visualisiert. Die Genesets ”REACTOME - SHH Signaling, ”REACTOME -WNT Targets Hallmark -MYC Targets” und ”Hallmark -E2F Targets” wurden der Molecular Signature-Database (https://www.gsea-msigdb.org/gsea/msigdb/) entnommen (30). Genset-assoziierte Proteine wurden mit einem hauseigenen Skript in der R-Softwareumgebung aus der kombinierten, mittels Matrix-Dissektionsverfahren batcheffektreduzierten Matrix des Hirntumor-Datensatzes extrahiert. Zur Unterstützung der Visualisierung des EGFR- und SHH Signaltransduktionsnetzwerks wurde die STRING-Protein-Protein -Interaktionsdatenbank verwendet (https://string-db.org) (56). Boxplots wurden mit der in R integrierten Funktion boxplot() erstellt.

5.1.8 | Datenverfügbarkeit

Alle verwendeten Datensätze dieser Studie wurden in öffentlich verfügbaren Datenbanken hinterlegt. Der Cetuximab-Datensatz kann über das PRIDE-Archiv (PXD014565) ab-

gerufen werden. Der Hirntumor-Datensatz ist über das Clinical Proteomic Tumor Analysis Consortium Data Portal (https://cptac-data-portal.georgetown.edu/cptacPublic/) und die Proteomics Data Commons (https://pdc.cancer.gov/pdc/) abrufbar. Hauseigene Daten (Spike-in-Datensatz; Maus-Medulloblastom-Datensatz) können über das PRIDE-Archiv (PXD027467) abgerufen werden.
Zentrale Inhalte dieser Arbeit wurden 2022 in *Nature Communications* veröffentlicht. (60)

5.1.9 | Codeverfügbarkeit

Auf Basis des Quellcodes des Matrix-Dissektionsverfahrens wurde in enger Zusammenarbeit mit Simon Schlumbohm das R-Package HarmonizR, entwickelt welches unter https://github.com/SimonSchlumbohm/HarmonizR
(https://doi.org/10.5281/zenodo.6553171 abgerufen werden kann.

5.2 | Material

5.2.1 | Verbrauchsmaterialien und Geräte

Verbrauchsmaterialien und Geräte	Hersteller	Katalognummer
ThermoMix®C	Eppendorf, Germany	Cat#5382000015
Probe Sonicator	N/A	N/A
SpeedVac SC110 Savant	Thermo Fisher Scientific (Bremen, Deutschland)	N/A
Dionex Ultimate 3000 UPLC	Thermo Fisher Scientific (Bremen, Deutschland)	https://www.thermofisher.com/de/de/home/industrial/chromatography/liquid-chromatography-lc/hplc-uhplc-systems
Acclaim PepMap 100 C18 analytische Säule (75 µm x 50 cm, 100 Å pore size, 2 µm Partikelgröße)	Thermo Fisher Scientific (Bremen, Deutschland)	Cat#164564
Acclaim PepMap 100 C18 Vorsäule (100 µm x 2 cm, 100 Å Porengröße,5 µm Partikelgröße)	Thermo Fisher Scientific (Bremen, Deutschland)	Cat#164946
Orbitrap Fusion Quadrupole Orbitrap Iontrap Tribrid Massenspektrometer	Thermo Fisher Scientific (Bremen, Deutschland)	https://www.thermofisher.com/de/de/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-systems
QExactive Quadrupole Orbitrap Hybrid Massenspektrometer	Thermo Fisher Scientific (Bremen, Deutschland)	https://www.thermofisher.com/de/de/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-systems
Triple TOF 6600 Massenspektrometer	Sciex (Farmingham, USA)	https://sciex.com/products/mass-spectrometers/qtof-systems/tripletof-systems/tripletof-6600plus-system

Abbildung 5.1: Auflistung verwendeter Verbrauchsmaterialien und Geräte

5.2.2 | Chemikalien

Chemicalien	Hersteller	Katalognummer
Triethylammoniumbromid (TEAB)	Sigma-Aldrich (St.Louis, USA)	Cat#241059
Natrium-Desoxycholat(SDC)	Sigma-Aldrich (St.Louis, USA)	Cat#30970
Dithiothreitol (DTT)	Sigma-Aldrich -Merck (Darmstadt, Deutschland)	Cat#DTT-RO
Iodoacetamid (IAA)	Sigma-Aldrich (St.Louis, USA)	Cat#I1149
Ameisensäure(FA)	Promega (Fitchbrug, USA)	Cat# 88328
Acetonitrile (ACN)	Sigma-Aldrich-Merck(Darmstadt, Deutschland)	Cat#271004-100ML
N-Heptan	Sigma-Aldrich-Merck(Darmstadt, Deutschland)	Cat#142825
BCA Protein Assay	Thermo Fisher Scientific (Waltham, USA)	Cat#23227

Abbildung 5.2: Auflistung verwendeter Chemikalien

5.2.3 | Biomaterialen

Biomaterial	Hersteller	Katalognummer
Bovines Serumalbumin	Thermo Fisher Scientific (Waltham, USA)	Cat#23209
Trypsin (Prokine)	Promega (Fitchbrug, USA)	Cat#V5111
hGFAP-cre::SmoM2Fl/+ Mäuse und SmoM2Fl+ Wurfgewschwister (Frischgewebe)	Universitätsklinikum Hamburg Eppendorf (Hamburg, Deutschland)	-
hGFAP-cre::SmoM2Fl/+ Mäuse und SmoM2Fl+ Wurfgewschwister (Formalin-fixiert)	Universitätsklinikum Hamburg Eppendorf (Hamburg, Deutschland)	-
***E. Coli* Stamm K12**	Sigma-Aldrich (St.Louis, USA)	Cat#EC1-5G

Abbildung 5.3: Auflistung verwendeter Biomaterialien

5.2.4 | Software und Datenbanken

Software und Datenbanken	Hersteller	Katalognummer
MaxQuant Version 1.6.2.10	Max Planck Institut für Biochemie (Martinsried, Deutschland)	https://maxquant.net
Perseus Version 1.5.8.5.	Max Planck Institut für Biochemie (Martinsried, Deutschland)	https://maxquant.net
Prism Version 5	GraphPad Software (San Diego, USA)	https://www.graphpad.com/scientific-software/prism/
Proteome Discoverer Version 2.4.	Thermo Fisher Scientific (Bremen, Deutschland)	https://www.thermofisher.com/de/de/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-software/multi-omics-data-analysis/proteome-discoverer-software.html
Proteome Discoverer Version 2.1.	Thermo Fisher Scientific (Bremen, Deutschland)	https://www.thermofisher.com/de/de/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-software/multi-omics-data-analysis/proteome-discoverer-software.html
MSigDB Version 2022.1	Board Institute, University of California (Berkley, USA)	https://www.gsea-msigdb.org/gsea/msigdb
R Software Umgebung Version 4.0.4	R Core Team, 2022	https://cran.r-project.org/
pheatmap (CRAN Package) Version 1.0.12.	Kolde et al. 2019	https://cran.r-project.org/web/packages/pheatmap/index.html
mixomics (Bioconductor Package) Version 6.20.0	Rohart et al. 2017	http://www.bioconductor.org/packages/release/bioc/html/mixOmics.html
Venny Version 2.1.0	BioinfoGP (Spanien)	https://bioinfogp.cnb.csic.es/
Github June 2022	Github inc. (San Francisco, USA)	https://github.com/
humane FASTA Datenbank (Dezember 2017, 26559 Einträge)	UniProt (EMBL, Schweiz)	https://www.uniprot.org/
Saccharomyces cerevisiae Stamm ATCC 204508 FASTA (Dezember 2017, 6721 EinträgeEinträge)	UniProt (EMBL, Schweiz)	https://www.uniprot.org/
E.Coli FASTA Datenbank (Dezember 2017, 31400 Einträge)	UniProt (EMBL, Schweiz)	https://www.uniprot.org/
Mus Musculus FASTA Datenbank (Dezember 2020, 17015 Einträge)	UniProt (EMBL, Schweiz)	https://www.uniprot.org/

Abbildung 5.4: Auflistung verwendeter Software und Datenbanken

6

Anhang

6.1 | Ergänzende Abbildungen

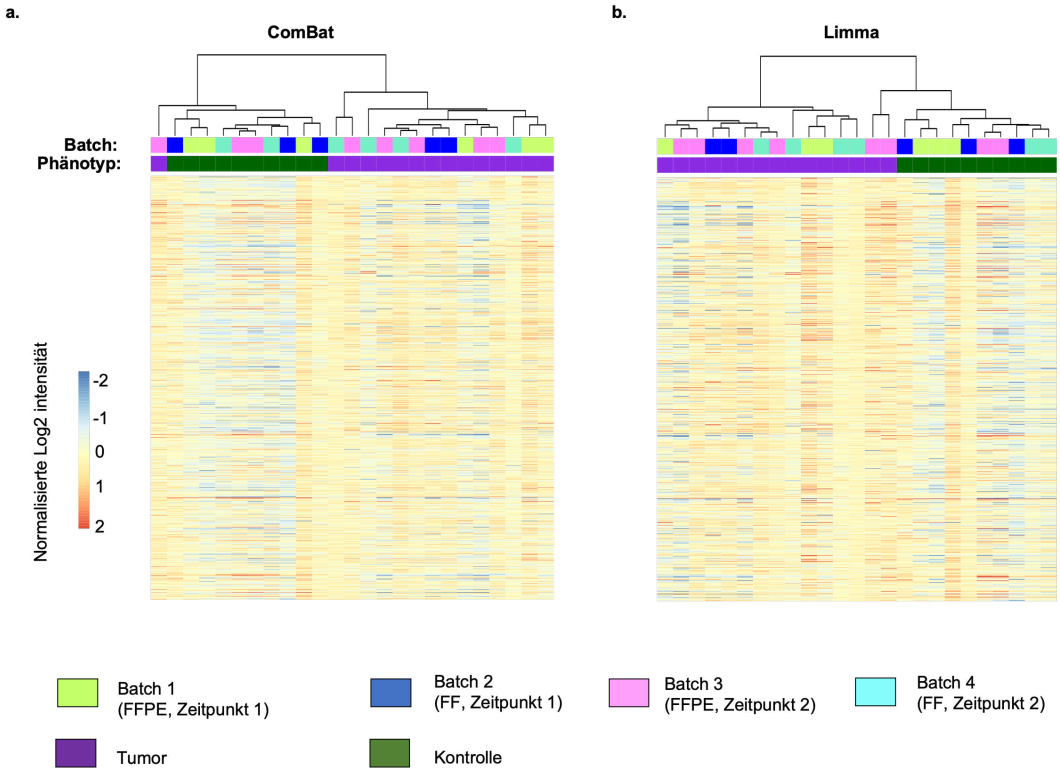

Abbildung 6.1: Limma- und ComBat-basierte Batcheffektreduktion zwischen unterschiedlichen Gewebskonservierungs-Techniken und Analysezeitpunkten am Beispiel des Maus-Medulloblastom-Datensatzes. Heatmap-Visualisierung des auf der Pearson-Korrelation basierenden hierarchischen Clustering mit Ward D Linkage nach Limma- **(a)**- und ComBat**(b)** (L/S-Scaling, nicht-parametrisches Bayesian-Framework)-basierter Batcheffektkorrektur. Da von Limma und ComBat keine MNAR-Typ Fehlwerte tolerieren, wurde die Datenmatrix auf 1002, in allen Batches identifizierte Proteine reduziert.

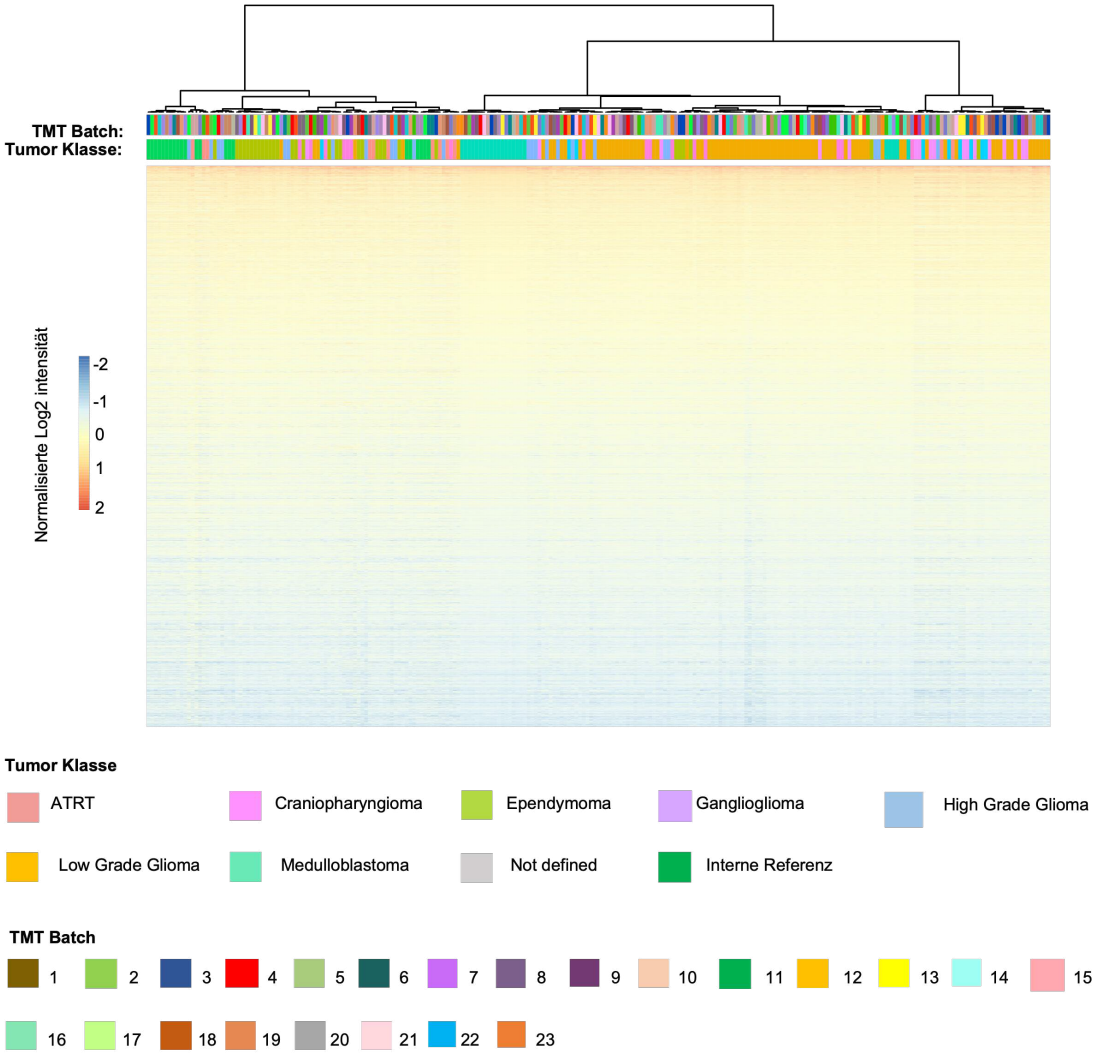

Abbildung 6.2: ComBat (L/S-Scaling, nicht-parametrisches Bayesian-Framework)-basierte Batcheffektkorrektur zwischen allen 23 für den humanen Hirntumor-Datensatz vermessenen TMT-11-Plex-Batches . Da L ComBat keine MNAR-Typ Fehlwerte toleriert, wurde die Datenmatrix auf 5270, in allen Batches identifizierte Proteine reduziert.

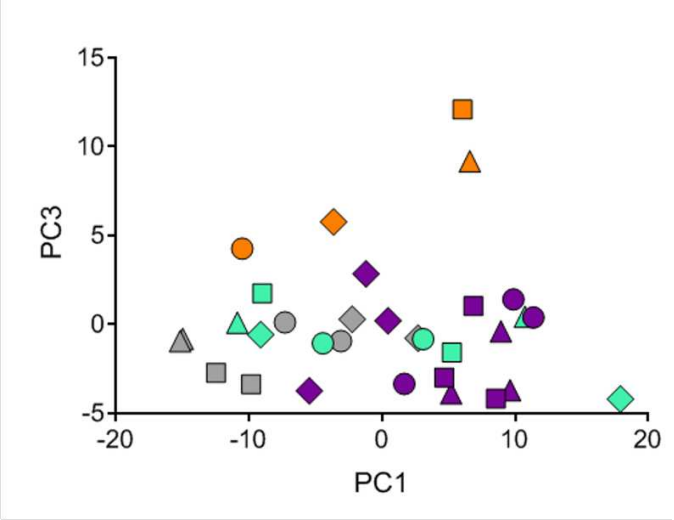

Abbildung 6.3: Evaluation der Anwendbarkeit von ComBat (L/S-Scaling, parametrisches Bayesian-Framework) zur Reduktion von Batcheffekten zwischen TMT-Batches auf Peptidebene am Beispiel des Cetuximab-Datensatzes. Streudiagramm der Verteilung der Proben über Hauptkomponente 1 und 3, ermittelt mittels linearer-PCA, basierend auf 4278 Peptiden, die in 100% aller Proben gefunden wurden.

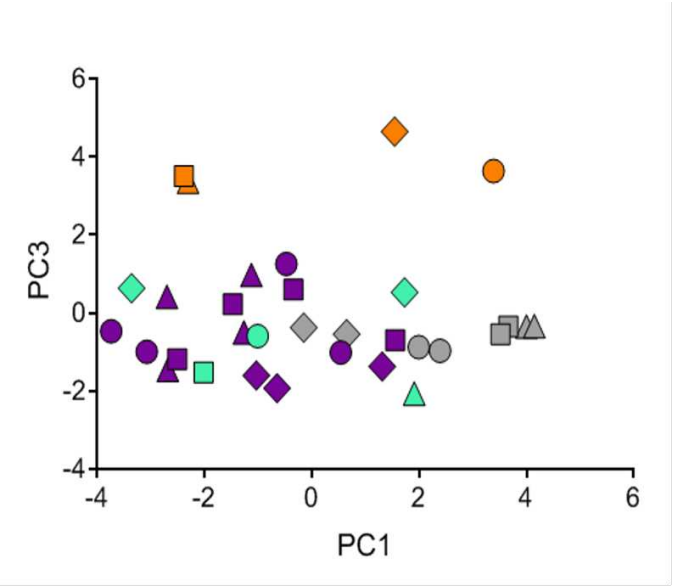

Abbildung 6.4: Evaluation der Anwendbarkeit von ComBat (L/S-Scaling, parametrisches Bayesian-Framework) zur Reduktion von Batcheffekten zwischen TMT-Batches auf Proteinebene am Beispiel des Cetuximab-Datensatzes. Streudiagramm der Verteilung der Proben über Hauptkomponente 1 und 3, ermittelt mittels linearer-PCA, basierend auf 1477 Proteinen, die in 100% aller Proben gefunden wurden.

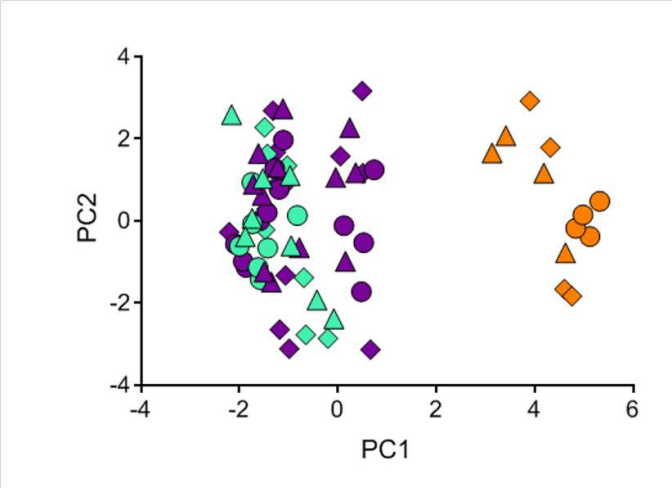

Abbildung 6.5: Evaluation der Anwendbarkeit von ComBat (L/S-Scaling, parametrisches Bayesian-Framework) zur Reduktion von Batcheffekten zwischen SILAC-, TMT- und DDA-LFQ-Daten am Beispiel des Cetuximab-Datensatzes. Streudiagramm der Verteilung der Proben über Hauptkomponente 1 und 2, ermittelt mittels linearer PCA, basierend auf 1036 Proteinen, die in 100% aller Proben gefunden wurden.

6.2 | Abkürzungsverzeichnis

ACN	Acetonitril
CV	Varianzkoeffizient (engl.: coefficient of variation)
DDA	Datenunabhängiger Aufnahmemodus (engl.: data independant acquisition)
DIA	Datenabhängiger Aufnahmemodus (engl.: data dependant acquisition)
EGFR	"Epidermal growth factor"
EIC	Extrahiertes Ionenchromatogram
FASP	"Filter aided sample preparation"
FF	Frischgewebe (engl.: Fresh Forzen)
FFPE	Formalin fixiert Parafin eingebettet (engl.: formalin fixated parafine embedded)
FLNA	"Filamin A"
HGG	Hochgradiges Gliom (engl.: high grade glioma)
iRS	Interne Referenz Skalierung (engl.: internal reference scaling)
iST	"in stage Tip Digestion"
LC	Flüssigkeitschromatographie (engl.: Liquid chromatography)
LFQ	Markierungsfreie Quantifizierung (engl.: Lable free quantification)
LGG	Niedriggradiges Gliom (engl.: low grade glioma)
MAR	"Missing at random"
MCAR	"Missing completly at random"
MCAR	"Missing not at random"
MS	Massenspektrometrie
MS/MS	Tandem Massenspektrometrie
MS1	Massenspektrometrische Analyseebene 1
MS2	Massenspektrometrische Analyseebene 2
MSE	Mittlerer quandratischer Fehler (engl.: Mean Squarred Error)
PCA	Hauptkomponentenanalyse (engl.: Principal component analysis)
RF	Zufallsbaum (engl.: Random Forest)
RP	Umkehrphase (engl.: Reversed Phase)
SHH	Biochemischer Signalweg "Sonic Hedgehog"
SILAC	"Stable isotope labling by amino acids in cell culture"
SP3	"Single-pot, solid phase-enhanced sample preparation"
t-SNE	"T-Distributed stochastic neighbor embedding"
TMT	"Tandem Mass Tag"
TOF	Flugzeitmassenanalysator (engl.: Time of flight)

6.3 | Auflistung der verwendeten Gefahrenstoffe nach GHS

Chemikalie	GHS Symbol	P-Sätze	H-Sätze
Acetonitril		P210, P280, P305+P351+P338 , P403+P235	H225, H302+H312+ H332, H319
Iodacetamid		P261, P264, P273, P301+P310, P302, P352	H301, H317, H334, H413
Dithiotreitol		P280, P302+P352, P305+P351+P338, P308+P311	H302, H315, H319, H335
Ameisensäure		P280,P301 + P330 + P331,P305 + P351 + P338, P310, P303 + P361 + P353, P210	H226, H302, H314
Pierce BCA Protein Assay, Reagent B		P273 , P391 , P501	H400, H411
Natrium -desoxycholat		P264, P270, P301+P312, P501	H302
N-Heptan		P210, P233, P73,P301+P310, P303+P361+P353, P331	H225, H304,H315, H336,H410

6.4 | KMR-Stoffe

In dieser Arbeit wurden keine krebserzeugenden, erbgutveränderten oder fortpflanzungsgefährdenden Stoffe (KMR-Stoffe) der Kategorie GHS 1A und 1B verwendet.

6.5 | Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere außerdem, dass diese Arbeit in gleicher oder ähnlicher Form noch nie einer Prüfungsstelle zur Erlangung des Doktorgrades vorgelegt wurde. Es wurden keine früheren Promotionsversuche meinerseits unternommen.

,Hamburg den 24.11.2021

Hannah Voß

6.6 | Danksagung

Ich Danke von Herzen allen Menschen, die mich auf diesem langen Weg unterstützt haben.

Ganz besonders Danke ich Prof. Hartmut Schlüter. Vielen Dank für die stets motivierende, freundliche und individuelle Betreuung meiner Dissertation. Danke für alles, was mir im Rahmen meiner Doktoranden-Zeit ermöglicht wurde. Sowohl im Bezug auf wissenschaftliche Ratschläge, spannende Themen und vielseitige fachliche und persönliche Diskussionen, als auch im Bezug auf die Möglichkeiten mich frei als Wissenschaftler zu entfalten und meine Arbeit international vorzustellen, hätte ich mir keine bessere Betreuung wünschen können.

Des Weiteren Danke ich Priv. Dr. rer. nat. habil. Markus Perbandt herzlich für die Bereitsschaft zum Erstellen meines Zweitgutachtens.

Sehr dankbar bin ich für die Hilfe von Anna Beckmann, Philip Barwikowski und Maria Riedner! Danke für die detaillierte Durchsicht und Korrektur meiner Arbeit. Ich bin unglaublich dankbar, Menschen zu haben, die so viel ihrer wertvollen Zeit opfern um mir zur Seite zu stehen und weiß, dass das alles andere als selbstverständlich ist. Danke!

Des Weiteren Danke ich meiner Arbeitsgruppe: AG Schlüter, Massenspektrometrische Proteomanalytik, am Universitätsklinikum Hamburg , für viele nette Kaffeepausen, viel Verständnis und wissenschaftlichen Input. Mein besonderer Dank gilt hier Bente Siebels und Maria Riedner, die in der stressigen Hochphase meiner Doktorarbeit viele Aufgaben außerhalb der Doktorarbeit für mich leichter gemacht haben.

Ebenfalls Danke ich Julia Neumann, welche das Thema dieser Arbeit mitdefiniert hat und mir im Rahmen des Verfassens der zugehörigen Publikation betreuend zur Seite stand. Danke auch an Philipp Neumann und Simon Schlumbohm für die informatische Unterstützung in diesem Projekt.

Mein aller größter Dank gilt meinen Freunden und meiner Familie für ganz viel Verständnis, Liebe und Support. Ganz lieben Dank, besonders an Philip Barwikowski für den besten täglichen Support den man sich hätte wünschen können, über das letzte Jahr hinweg. Danke für viele gekochte Kaffees und Abendessen, angeschaltete Waschmaschinen und ganz viel emotionale Unterstützung, als ich im Schreibprozess wenig Zeit für anderes hatte. Danke, dass Du mit mir die kleinen und großen Meilensteine gefeiert und die schlechten Tage getragen hast. Ohne deine Motivation hätte ich die letzten Meter nicht so schnell gehen können. Du bist ein wundervoller Mensch, für den ich jeden Tag dankbar bin.

Am allermeisten Danken möchte ich meinen Eltern Udo und Edeltraud Voß. Mama und Papa, ihr seid meine großen Vorbilder. Danke dass ihr mich die 7 Jahre von meiner ersten Bachelorvorlesung bis zur Abgabe meiner Dissertation in allem unterstützt habt. Es gibt keine besseren Eltern als euch. Ihr seit unglaublich liebevolle Menschen und dank eures Verständnisses, eurer liebevollen Art und eurer Ehrlichkeit, waren selbst die schwierigsten Tage erträglich und die schönen Tage und Erfolge noch schöner. Danke! Ich liebe Euch von ganzem Herzen. Ohne Euch würde ich hier heute nicht stehen und würde nicht jeden Tag mehr zu dem Menschen finden, der ich sein möchte.

Literaturverzeichnis

[1] B. Aslam, M. Basit, M. A. Nisar, M. Khurshid, and M. H. Rasool. Proteomics: Technologies and Their Applications. *J Chromatogr Sci*, 55(2):182–196, 02 2017.

[2] J. Baselga. The EGFR as a target for anticancer therapy–focus on cetuximab. *Eur J Cancer*, 37 Suppl 4:16–22, Sep 2001.

[3] A. Brenes, J. Hukelmann, D. Bensaddek, and A. I. Lamond. Multibatch TMT Reveals False Positives, Batch Effects and Missing Values. *Mol Cell Proteomics*, 18(10):1967–1980, 10 2019.

[4] R. Bąchor, M. Waliczek, P. Stefanowicz, and Z. Szewczuk. Trends in the Design of New Isobaric Labeling Reagents for Quantitative Proteomics. *Molecules*, 24(4), Feb 2019.

[5] R. Couronné, P. Probst, and A. L. Boulesteix. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 19(1):270, 07 2018.

[6] I. Dapic, N. Uwugiaren, J. Kers, Y. Mohammed, D. R. Goodlett, and G. Corthals. Evaluation of Fast and Sensitive Proteome Profiling of FF and FFPE Kidney Patient Tissues. *Molecules*, 27(3), Feb 2022.

[7] T. Dau, G. Bartolomucci, and J. Rappsilber. Proteomics Using Protease Alternatives to Trypsin Benefits from Sequential Digestion with Trypsin. *Anal Chem*, 92(14):9523–9527, 07 2020.

[8] E. W. Deutsch, S. Orchard, P. A. Binz, W. Bittremieux, M. Eisenacher, H. Hermjakob, S. Kawano, H. Lam, G. Mayer, G. Menschaert, Y. Perez-Riverol, R. M. Salek, D. L. Tabb, S. Tenzer, J. A. Vizcaíno, M. Walzer, and A. R. Jones. Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. *J Proteome Res*, 16(12):4288–4298, 12 2017.

[9] E. J. Dupree, M. Jayathirtha, H. Yorkey, M. Mihasan, B. A. Petre, and C. C. Darie. A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of this Field. *Proteomes*, 8(3), Jul 2020.

[10] D. W. Ellison, J. Dalton, M. Kocak, S. L. Nicholson, C. Fraga, G. Neale, A. M. Kenney, D. J. Brat, A. Perry, W. H. Yong, R. E. Taylor, S. Bailey, S. C. Clifford, and R. J. Gilbertson. Medulloblastoma: clinicopathological correlates of SHH, WNT, and non-SHH/WNT molecular subgroups. *Acta Neuropathol*, 121(3):381–396, Mar 2011.

[11] D. Fanelli. Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PLoS One*, 5(4):e10271, Apr 2010.

[12] M. L. Gardner and M. A. Freitas. Multiple Imputation Approaches Applied to the Missing Value Problem in Bottom-Up Proteomics. *Int J Mol Sci*, 22(17), Sep 2021.

[13] T. Geiger, J. R. Wisniewski, J. Cox, S. Zanivan, M. Kruger, Y. Ishihama, and M. Mann. Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. *Nat Protoc*, 6(2):147–157, Feb 2011.

[14] P. R. Graves and T. A. Haystead. Molecular biologist's guide to proteomics. *Microbiol Mol Biol Rev*, 66(1):39–63, Mar 2002.

[15] A. M. Haag. Mass Analyzers and Mass Spectrometers. *Adv Exp Med Biol*, 919:157–169, 2016.

[16] A. Hu, W. S. Noble, and A. Wolf-Yadlin. Technical advances in proteomics: new developments in data-independent acquisition. *F1000Res*, 5, 2016.

[17] C. S. Hughes, S. Moggridge, T. Müller, P. H. Sorensen, G. B. Morin, and J. Krijgsveld. Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat Protoc*, 14(1):68–85, 01 2019.

[18] M. Ishikawa, R. Konno, D. Nakajima, M. Gotoh, K. Fukasawa, H. Sato, R. Nakamura, O. Ohara, and Y. Kawashima. Optimization of Ultrafast Proteomics Using an LC-Quadrupole-Orbitrap Mass Spectrometer with Data-Independent Acquisition. *J Proteome Res*, 21(9):2085–2093, 09 2022.

[19] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel. When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. *BMC Med Res Methodol*, 17(1):162, Dec 2017.

[20] N. Jehmlich, C. Golatowski, A. Murr, G. Salazar, V. M. Dhople, E. Hammer, and U. Völker. Comparative evaluation of peptide desalting methods for salivary proteome analysis. *Clin Chim Acta*, 434:16–20, Jul 2014.

[21] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 01 2007.

[22] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci*, 374(2065):20150202, Apr 2016.

[23] Y. Kim, D. Lee, and J. Kim. Effects of incubation temperature and acetonitrile amount on microwave-assisted tryptic digestion of proteins. *Anal Biochem*, 569:31–38, 03 2019.

[24] M. Kokla, J. Virtanen, M. Kolehmainen, J. Paananen, and K. Hanhineva. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. *BMC Bioinformatics*, 20(1):492, Oct 2019.

[25] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. R. Loh, and S. Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*, 16(12):1289–1296, 12 2019.

[26] S. K. Kwak and J. H. Kim. Statistical data preparation: management of missing values and outliers. *Korean J Anesthesiol*, 70(4):407–411, Aug 2017.

[27] C. Lazar, L. Gatto, M. Ferro, C. Bruley, and T. Burger. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J Proteome Res*, 15(4):1116–1125, Apr 2016.

[28] H. G. Lee, J. Jo, H. H. Hong, K. K. Kim, J. K. Park, S. J. Cho, and C. Park. State-of-the-art housekeeping proteins for quantitative western blotting: Revisiting the first draft of the human proteome. *Proteomics*, 16(13):1863–1867, Jul 2016.

[29] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, 11(10):733–739, 10 2010.

[30] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, Jun 2011.

[31] L. Lin. Bias caused by sampling error in meta-analysis with small sample sizes. *PLoS One*, 13(9):e0204056, 2018.

[32] C. Ludwig, L. Gillet, G. Rosenberger, S. Amon, B. C. Collins, and R. Aebersold. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol Syst Biol*, 14(8):e8126, 08 2018.

[33] S. Magdeldin and T. Yamamoto. Toward deciphering proteomes of formalin-fixed paraffin-embedded (FFPE) tissues. *Proteomics*, 12(7):1045–1058, Apr 2012.

[34] M. Mann. Functional and quantitative proteomics using SILAC. *Nat Rev Mol Cell Biol*, 7(12):952–958, 12 2006.

[35] A. Mantsiou, M. Makridakis, K. Fasoulakis, I. Katafigiotis, C. A. Constantinides, J. Zoidakis, M. G. Roubelakis, A. Vlahou, and V. Lygirou. Proteomics Analysis of Formalin Fixed Paraffin Embedded Tissues in the Investigation of Prostate Cancer. *J Proteome Res*, 19(7):2631–2642, 07 2020.

[36] D. A. Megger, L. L. Pott, M. Ahrens, J. Padden, T. Bracht, K. Kuhlmann, M. Eisenacher, H. E. Meyer, and B. Sitek. Comparison of label-free and label-based strategies for proteome analysis of hepatoma cell lines. *Biochim Biophys Acta*, 1844(5):967–976, May 2014.

[37] J. Mergner, M. Frejno, M. Messerer, D. Lang, P. Samaras, M. Wilhelm, K. F. X. Mayer, C. Schwechheimer, and B. Kuster. Proteomic and transcriptomic profiling of aerial organ development in Arabidopsis. *Sci Data*, 7(1):334, 10 2020.

[38] A. A. Mitani and S. Haneuse. Small Data Challenges of Studying Rare Diseases. *JAMA Netw Open*, 3(3):e201965, 03 2020.

[39] C. Müller, A. Schillert, C. Röthemeier, D. A. Trégouët, C. Proust, H. Binder, N. Pfeiffer, M. Beutel, K. J. Lackner, R. B. Schnabel, L. Tiret, P. S. Wild, S. Blankenberg, T. Zeller, and A. Ziegler. Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data. *PLoS One*, 11(6):e0156594, 2016.

[40] A. Palomba, M. Abbondio, G. Fiorito, S. Uzzau, D. Pagnozzi, and A. Tanca. Comparative Evaluation of MaxQuant and Proteome Discoverer MS1-Based Protein Quantification Tools. *J Proteome Res*, 20(7):3497–3507, 07 2021.

[41] A. B. Pedersen, E. M. Mikkelsen, D. Cronin-Fenton, N. R. Kristensen, T. M. Pham, L. Pedersen, and I. Petersen. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol*, 9:157–166, 2017.

[42] N. Perakakis, A. Yazdani, G. E. Karniadakis, and C. Mantzoros. Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. *Metabolism*, 87:A1–A9, 10 2018.

[43] F. Petralia, N. Tignor, B. Reva, M. Koptyra, S. Chowdhury, D. Rykunov, A. Krek, W. Ma, Y. Zhu, J. Ji, A. Calinawan, J. R. Whiteaker, A. Colaprico, V. Stathias, T. Omelchenko, X. Song, P. Raman, Y. Guo, M. A. Brown, R. G. Ivey, J. Szpyt, S. Guha Thakurta, M. A. Gritsenko, K. K. Weitz, G. Lopez, S. Kalayci, Z. H. Gümüş, S. Yoo, F. da Veiga Leprevost, H. Y. Chang, K. Krug, L. Katsnelson, Y. Wang, J. J. Kennedy, U. J. Voytovich, L. Zhao, K. S. Gaonkar, B. M. Ennis, B. Zhang, V. Baubet, L. Tauhid, J. V. Lilly, J. L. Mason, B. Farrow, N. Young, S. Leary, J. Moon, V. A. Petyuk, J. Nazarian, N. D. Adappa, J. N. Palmer, R. M. Lober, S. Rivero-Hinojosa, L. B. Wang, J. M. Wang, M. Broberg, R. K. Chu, R. J. Moore, M. E. Monroe, R. Zhao, R. D. Smith, J. Zhu, A. I. Robles, M. Mesri, E. Boja, T. Hiltke, H. Rodriguez, B. Zhang, E. E. Schadt, D. R. Mani, L. Ding, A. Iavarone, M. Wiznerowicz, S. Schürer, X. S. Chen, A. P. Heath, J. L. Rokita, A. I. Nesvizhskii, D. Fenyö, K. D. Rodland, T. Liu, S. P. Gygi, A. G. Paulovich, A. C. Resnick, P. B. Storm, B. R. Rood, P. Wang, A. Francis, A. M. Morgan, A. J. Waanders, A. N. Viaene, A. M. Buccoliero, A. M. Chinnaiyan, C. A. Leonard, C. N. Kline, C. Caporalini, C. R. Kinsinger, C. Li, D. E. Kram, D. Hanson, E. Appert, E. A. Kawaler, E. H. Raabe, E. M. Jackson, J. P. Greenfield, G. S. Stone, G. Getz, G. Grant, G. C. Teo, I. F. Pollack, J. E. Cain, J. B. Foster, J. J. Phillips, J. E. Palma, K. A. Ketchum, K. V. Ruggles, L. Blumenberg, M. Cornwell, M. Sarmady, M. J. Domagalski, M. P. Cieślik, M. Santi, M. M. Li, M. J. Ellis, M. A. Wyczalkowski, M. Connors, M. Scagnet, N. Gupta, N. J. Edwards, N. A. Vitanza, O. M. Vaske, O. Becher, P. B. McGarvey, R. Firestein, S. Mueller, S. G. Winebrake, S. M. Dhanasekaran, S. Cai, S. Partap, T. Patton, T. Le, T. D. Lorentzen, W. Liu, and W. E. Bocik. Integrated Proteogenomic Characterization across Major Histological Types of Pediatric Brain Cancer. *Cell*, 183(7):1962–1985, 12 2020.

[44] S. X. Phua, K. P. Lim, and W. W. Goh. Perspectives for better batch effect correction in mass-spectrometry-based proteomics. *Comput Struct Biotechnol J*, 20:4369–4375, 2022.

[45] R. C. Poulos, P. G. Hains, R. Shah, N. Lucas, D. Xavier, S. S. Manda, A. Anees, J. M. S. Koh, S. Mahboob, M. Wittman, S. G. Williams, E. K. Sykes, M. Hecker, M. Dausmann, M. A. Wouters, K. Ashman, J. Yang, P. J. Wild, A. deFazio, R. L. Balleine, B. Tully, R. Aebersold, T. P. Speed, Y. Liu, R. R. Reddel, P. J. Robinson, and Q. Zhong. Strategies to enable large-scale proteomics for reproducible research. *Nat Commun*, 11(1):3793, 07 2020.

[46] E. M. Price and W. P. Robinson. Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned. *Front Genet*, 9:83, 2018.

[47] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, 43(7):e47, Apr 2015.

[48] F. Rohart, B. Gautier, A. Singh, and K. A. Lê Cao. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol*, 13(11):e1005752, Nov 2017.

[49] M. Scholz, F. Kaplan, C. L. Guy, J. Kopka, and J. Selbig. Non-linear PCA: a missing data approach. *Bioinformatics*, 21(20):3887–3895, Oct 2005.

[50] U. Schüller, V. M. Heine, J. Mao, A. T. Kho, A. K. Dillon, Y. G. Han, E. Huillard, T. Sun, A. H. Ligon, Y. Qian, Q. Ma, A. Alvarez-Buylla, A. P. McMahon, D. H. Rowitch, and K. L. Ligon. Acquisition of granule neuron precursor identity is a critical determinant of progenitor cell competence to form Shh-induced medulloblastoma. *Cancer Cell*, 14(2):123–134, Aug 2008.

[51] A. Shenoy and T. Geiger. Super-SILAC: current trends and future perspectives. *Expert Rev Proteomics*, 12(1):13–19, Feb 2015.

[52] M. Sielaff, J. Kuharev, T. Bohn, J. Hahlbrock, T. Bopp, S. Tenzer, and U. Distler. Evaluation of FASP, SP3, and iST Protocols for Proteomic Sample Preparation in the Low Microgram Range. *J Proteome Res*, 16(11):4060–4072, 11 2017.

[53] P. Skowron, H. Farooq, F. M. G. Cavalli, A. S. Morrissy, M. Ly, L. D. Hendrikse, E. Y. Wang, H. Djambazian, H. Zhu, K. L. Mungall, Q. M. Trinh, T. Zheng, S. Dai, A. S. G. Stucklin, M. C. Vladoiu, V. Fong, B. L. Holgado, C. Nor, X. Wu, D. Abd-Rabbo, P. Bérubé, Y. C. Wang, B. Luu, R. A. Suarez, A. Rastan, A. H. Gillmor, J. J. Y. Lee, X. Y. Zhang, C. Daniels, P. Dirks, D. Malkin, E. Bouffet, U. Tabori, J. Loukides, F. P. Doz, F. Bourdeaut, O. O. Delattre, J. Masliah-Planchon, O. Ayrault, S. K. Kim, D. Meyronet, W. A. Grajkowska, C. G. Carlotti, C. de Torres, J. Mora, C. G. Eberhart, E. G. Van Meir, T. Kumabe, P. J. French, J. M. Kros, N. Jabado, B. Lach, I. F. Pollack, R. L. Hamilton, A. A. N. Rao, C. Giannini, J. M. Olson, L. Bognár, A. Klekner, K. Zitterbart, J. J. Phillips, R. C. Thompson, M. K. Cooper, J. B. Rubin, L. M. Liau, M. Garami, P. Hauser, K. K. W. Li, H. K. Ng, W. S. Poon, G. Yancey Gillespie, J. A. Chan, S. Jung, R. E. McLendon, E. M. Thompson, D. Zagzag, R. Vibhakar, Y. S. Ra, M. L. Garre, U. Schüller, T. Shofuda, C. C. Faria, E. López-Aguilar, G. Zadeh, C. C. Hui, V. Ramaswamy, S. D. Bailey, S. J. Jones, A. J. Mungall, R. A. Moore, J. A. Calarco, L. D. Stein, G. D. Bader, J. Reimand, J. Ragoussis, W. A. Weiss, M. A. Marra, H. Suzuki, and M. D. Taylor. The transcriptional landscape of Shh medulloblastoma. *Nat Commun*, 12(1):1749, 03 2021.

[54] J. A. Staal, Y. Pei, and B. R. Rood. A Proteogenomic Approach to Understanding MYC Function in Metastatic Medulloblastoma Tumors. *Int J Mol Sci*, 17(10), Oct 2016.

[55] M. Stepath, B. Zülch, A. Maghnouj, K. Schork, M. Turewicz, M. Eisenacher, S. Hahn, B. Sitek, and T. Bracht. Systematic Comparison of Label-Free, SILAC, and TMT Techniques to Study Early Ad-

aption toward Inhibition of EGFR Signaling in the Colorectal Cancer Cell Line DiFi. *J Proteome Res*, 19(2):926–937, 02 2020.

[56] D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, L. J. Jensen, and C. von Mering. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*, 49(D1):D605–D612, 01 2021.

[57] A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, R. Johnstone, A. K. Mohammed, and C. Hamon. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem*, 75(8):1895–1904, Apr 2003.

[58] H. T. N. Tran, K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, and J. Chen. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol*, 21(1):12, 01 2020.

[59] S. Tyanova and J. Cox. Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. *Methods Mol Biol*, 1711:133–148, 2018.

[60] H. Voß, S. Schlumbohm, P. Barwikowski, M. Wurlitzer, M. Dottermusch, P. Neumann, H. Schlüter, J. E. Neumann, and C. Krisp. HarmonizR enables data harmonization across independent proteomic datasets with appropriate handling of missing values. *Nat Commun*, 13(1):3523, 06 2022.

[61] G. Wang and Q. Kong. The Dilemmas of Scientific Research Cooperation and Their Resolution From the Perspective of Evolutionary Psychology. *Front Psychol*, 10:2561, 2019.

[62] Z. Wang, K. Kavdia, K. K. Dey, V. R. Pagala, K. Kodali, D. Liu, D. G. Lee, H. Sun, S. R. Chepyala, J. H. Cho, M. Niu, A. A. High, and J. Peng. High-throughput and Deep-proteome Profiling by 16-plex Tandem Mass Tag Labeling Coupled with Two-dimensional Chromatography and Mass Spectrometry. *J Vis Exp*, (162), 08 2020.

[63] R. Wei, J. Wang, M. Su, E. Jia, S. Chen, T. Chen, and Y. Ni. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci Rep*, 8(1):663, 01 2018.

[64] L. A. Weston, K. M. Bauer, and A. B. Hummon. Comparison of bottom-up proteomic approaches for LC-MS analysis of complex proteomes. *Anal Methods*, 5(18), Sep 2013.

[65] J. R. Wiśniewski. Filter-Aided Sample Preparation for Proteome Analysis. *Methods Mol Biol*, 1841:3–10, 2018.

[66] M. Wühr, W. Haas, G. C. McAlister, L. Peshkin, R. Rad, M. W. Kirschner, and S. P. Gygi. Accurate multiplexed proteomics at the MS2 level using the complement reporter ion cluster. *Anal Chem*, 84(21):9214–9221, Nov 2012.

[67] R. Yamada, D. Okada, J. Wang, T. Basak, and S. Koyama. Interpretation of omics data analyses. *J Hum Genet*, 66(1):93–102, Jan 2021.

[68] X. Zhang, A. Fang, C. P. Riley, M. Wang, F. E. Regnier, and C. Buck. Multi-dimensional liquid chromatography in proteomics–a review. *Anal Chim Acta*, 664(2):101–113, Apr 2010.

[69] Y. Zhang, B. R. Fonslow, B. Shan, M. C. Baek, and J. R. Yates. Protein analysis by shotgun/bottom-up proteomics. *Chem Rev*, 113(4):2343–2394, Apr 2013.

[70] B. Zou, T. Zhang, R. Zhou, X. Jiang, H. Yang, X. Jin, and Y. Bai. deepMNN: Deep Learning-Based Single-Cell RNA Sequencing Data Batch Correction Using Mutual Nearest Neighbors. *Front Genet*, 12:708981, 2021.

[71] J. Čuklina, P. G. A. Pedrioli, and R. Aebersold. Review of Batch Effects Prevention, Diagnostics, and Correction Approaches. *Methods Mol Biol*, 2051:373–387, 2020.