

Advancing the discovery of predictive
biomarkers in drug high-throughput screens
and clinical trials for precision oncology



Alexander Joschua Ohnmacht

München 2024

Advancing the discovery of predictive biomarkers in drug high-throughput screens and clinical trials for precision oncology

Dissertation

der Fakultät für Biologie
der Ludwig-Maximilians-Universität München



Vorgelegt von

Alexander Joschua Ohnmacht

zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)

München, den 08. April 2024

Diese Dissertation wurde angefertigt
unter der Leitung von PD Dr. Michael P. Menden am
Helmholtz Zentrum München (Helmholtz Munich)
Deutsches Forschungszentrum für Gesundheit und Umwelt
Computational Health Center

1. Gutachter: PD Dr. Michael P. Menden
2. Gutachter: Prof. Dr. Wolfgang Enard

Tag der Abgabe: 08. April 2024

Tag der mündlichen Prüfung: 12. Dezember 2024

Eidesstattliche Erklärung:

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt ist. Außerdem erkläre ich hiermit, dass diese Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist und ich mich nicht anderweitig einer Doktorprüfung ohne Erfolg unterzogen habe.

23.04.2025, München

Alexander J. Ohnmacht

Abstract

Discovering a universal cure for all cancers, known as the one-drug-fits-them-all approach, is challenging due to the diverse and complex nature of the disease. Precision oncology is a paradigm in modern medicine which ought to overcome this approach by tailoring cancer treatments to tumour and patient characteristics for increased safety and efficacy. Carcinogenesis is driven by genetic alterations, which established themselves as suitable drug targets and predictive biomarkers in clinical practice. While these discoveries were previously limited to studying a few key cancer pathways or cancer genes, the contemporary accumulation of biomedical data, including molecularly profiled drug high-throughput screens and clinical trials, facilitates the discovery of biomarkers for predicting treatment success. Several computational models using data-driven methods were able to successfully predict responses to drugs in both preclinical and clinical settings based on molecular characteristics; however, the translation of the predicted biomarkers towards clinical utility has remained limited. In order to address this, this thesis presents a range of methods and analysis strategies that make sparse, interpretable and robust predictions of potential biomarkers for treatment efficacy.

The chapters of this thesis include (1) an integrative method for identifying DNA methylation biomarkers associated with drug susceptibility using drug high-throughput screens and multi-omics characterisations in cancer cell lines, (2) an assessment of the epithelial-mesenchymal transition in cancer cell lines and its causal impact on drug susceptibility and (3) a framework for the exploration and identification of the molecular and biomarker landscapes of randomised controlled clinical trials in oncology.

In summary, the presented work facilitates the discovery of predictive biomarkers by incorporating molecular data modalities into tailored modelling strategies to reflect cancer mechanisms in high-throughput screens and clinical trials. In future, these methods may become an indispensable part of a more integrated and data-driven drug discovery and development process to design more targeted and effective cancer treatment strategies.

Summary

Cancer is a leading burden for public health and thus urges the development of effective treatments. While scientific progress has been made in cancer biology and the development of targeted therapies in the past, many patients still do not respond to their treatments, suffer relapses or experience serious side effects. Tailoring treatments to predictive biomarkers in the form of tumour or patient characteristics has revolutionised cancer therapies, a concept known as precision oncology. For this, drug high-throughput screens and clinical trials provide pharmacological information to assess treatment benefits and therefore are key tools to advance the discovery of predictive biomarkers in cancer. Applying data-driven methods to these datasets complemented with molecular profiling allows the scalable discovery and evaluation of large numbers of therapies and potential molecular biomarkers, which can yield promising drug candidates and appropriate target populations for further development and validation.

This thesis proposes methods that use data from drug high-throughput screens and clinical trials for the discovery of predictive biomarkers in precision oncology.

First, the epigenetic component of drug sensitivity in cancer was elucidated by employing a sequential analysis design that identifies differentially methylated regions for drug responses in drug high-throughput screens (dDMRs) [1]. Subsequently, it integrates genomic and transcriptomic data modalities of cancer cell lines, matches the findings with human primary tumours, and proposes potential mechanisms on protein-protein interaction networks. The identified dDMRs were predominantly found in regulatory elements, particularly in promoters. For instance, a dDMR was found within the promoter region of *SLFN11*, a gene frequently associated with drug response to DNA-damaging agents. The consideration of the expression of genes proximal to all dDMRs in both cancer cell lines and primary human tumours prioritised tumour-generalisable dDMRs (tgdDMRs). For example, the expression of *SLFN11* was correlated to the methylation of its tgdDMR. Finally, the validation of tgdDMRs in another drug screen and DNA methylation profiling technology revealed high consistencies. Interestingly, DNA methylation was often accompanied by transcriptomic changes, but only modestly correlated with somatic genetic events. This suggests that it may function supplementary to gene expression, but rather complementarily to somatic genetic alterations for determining drug susceptibilities. In summary, this analysis offers a view of DNA methylation in the context of drug response by integrating and interpreting multiple data sources.

Secondly, the epithelial-to-mesenchymal (EMT) transition was systematically investigated as an intrinsic mechanism to determine drug responses in cancer [2]. For this, EMT was derived by molecular signatures quantified from gene expression and its effect on drug responses in drug high-throughput screens was estimated with predictive modelling by ablation and causal inference in the context of the tumour genetic background. Response to HSP90 inhibitors was robustly predicted by EMT signatures in melanoma, which was associated with the activity of the oncogenic transcription factor MITF. Finally, it was demonstrated that *in vitro* stimulation of EMT by TGF- β pretreatment can sensitise melanoma cell lines to the HSP90 inhibitor luminespib, which hints at a causal component of EMT.

Lastly, the Oncology Biomarker Discovery framework (OncoBird) for outlining the molecular and biomarker landscape of clinical trials for precision oncology is presented [3]. It enables explorative subgroup analysis in clinical trials by studying somatic alterations in tumour subtypes, mutually exclusive somatic alterations and their predictive components in tumour subtypes. It is showcased in the FIRE-3 phase III randomised controlled clinical trial in metastatic colorectal cancer. Among *KRAS* mutations, also chromosome arm 20q (chr20q) amplifications showed predictive benefits for cetuximab in the context of consensus molecular subtypes (CMS). It was also applied to the ADJUVANT clinical trial for non-small cell lung cancer, which yielded consistent predictive biomarkers for gefitinib benefits. Benchmarks with the current standard clinical practice and commonly employed methods for data-driven subgroup analysis showed consistencies between subgroups represented by sets of related biomarkers, and demonstrated that OncoBird can robustly identify biomarkers for smaller subgroups that predict higher treatment effects. For fostering further research and development, an OncoBird R package is available within an accessible dockerised Shiny application, which includes a reproducible demonstration of its presented analysis.

In summary, by allowing the systematic examination of different cancer hallmarks with the proposed analysis methods, this work contributes to the ongoing efforts to advance the data-driven discovery of predictive biomark-

ers for cancer treatments. Moreover, it provides a roadmap for this effort by highlighting key considerations and challenges for the accommodation of biomedical data and data-driven biomarker discovery methods for drug development in oncology.

Zusammenfassung

Krebs ist eine der enormsten Belastungen für das Gesundheitswesen und fordert daher dringlichst die Entwicklung wirksamer Behandlungen. Obwohl Fortschritte in der Forschung von Krebsbiologie und bei der Entwicklung zielgerichteter Therapien in der Vergangenheit erzielt wurden, sprechen dennoch viele Patienten entweder nicht auf ihre Behandlungen an, erleiden Rückfälle oder erfahren schwere Nebenwirkungen. Das Zuschneiden von Behandlungen gemäß prädiktiven Biomarkern in der Form von Tumor- oder Patientenmerkmalen hat Krebstherapien revolutioniert, ein Konzept das auch als Präzisionsonkologie bekannt ist. Dabei liefern Hochdurchsatz-Wirkstoffscreenings und klinische Studien pharmakologische Informationen für die Bewertung von Behandlungserfolgen und sind daher zentrale Unternehmungen für die weitere Entdeckung von prädiktiven Biomarkern für Krebs. Die Anwendung datengesteuerter Methoden auf diese mit molekularen Profilen ergänzte Datensätze ermöglicht die skalierbare Entdeckung und Bewertung einer großen Anzahl von Medikamenten und potenziellen molekularen Biomarkern, aus denen sich vielversprechende Arzneimittel und entsprechende Zielgruppen für die weitere Entwicklung und Validierung ergeben können.

In dieser Arbeit werden Methoden vorgeschlagen, die Daten aus Hochdurchsatz-Wirkstoffscreenings und klinischen Studien für die Entdeckung von prädiktiven Biomarkern in der Präzisionsonkologie nutzen.

Zunächst wurde die epigenetische Komponente der Krebswirkstoffsensitivität durch die Anwendung eines schichtweisen Analysedesigns aufgedeckt, welches differentiell methylierte Regionen (dDMRs) für das Wirkstoffansprechen in Hochdurchsatz-Wirkstoffscreening identifiziert [1]. Anschließend integriert diese Analyse genomische und transkriptomische Daten von Krebszelllinien, gleicht jene Erkenntnisse mit menschlichen Primärtumoren ab und schlägt mögliche Mechanismen in Protein-Protein-Interaktionsnetzwerken vor. Die identifizierten dDMRs befanden sich überwiegend in regulatorischen Elementen, insbesondere in Promotoren. So wurde beispielsweise eine dDMR in der Promotorregion von *SLFN11* gefunden, einem Gen, das häufig mit der Reaktion auf DNA-schädigende Wirkstoffe in Verbindung gebracht wird. Die nähere Betrachtung der Expression von Genen die proximal zu allen dDMRs sowohl in Krebszelllinien als auch in primären menschlichen Tumoren liegen führte zu der Priorisierung von tumorgeneralisierbaren dDMRs (tgdDMRs). So war beispielsweise die Expression von *SLFN11* mit der Methylierung seiner tgdDMR korreliert. Die Validierung der tgdDMRs in einem anderen Wirkstoffscreening und anderer DNA-Methylierung Profilierungstechnologie ergab eine hohe Übereinstimmung. Interessanterweise ging die DNA-Methylierung häufig mit transkriptomischen Veränderungen einher, korrelierte aber nur in geringem Maße mit somatischen genetischen Ereignissen. Dies deutet darauf hin, dass die DNA-Methylierung bei der Bestimmung der Wirkstoffsensitivität möglicherweise ergänzend zur Genexpression und eher komplementär zu somatischen genetischen Veränderungen fungiert. Zusammenfassend ermöglicht diese Analyse einen Blick auf die DNA-Methylierung im Zusammenhang mit Wirkstoffsensitivität durch die Integration und Interpretation mehrerer Datenquellen.

Zweitens wurde die Epithelial-mesenchymale Transition (EMT) als wesentlicher Mechanismus zur Bestimmung von Krebswirkstoffsensitivität systematisch untersucht [2]. Dafür wurde zunächst EMT aus molekularen Signaturen abgeleitet und durch Genexpression quantifiziert. Anschließend wurde seine Auswirkung auf Wirkstoffsensitivität in Hochdurchsatz-Wirkstoffscreens durch Ablation in prädiktiver Modellierung und kausale Inferenz im Kontext des genetischen Hintergrunds des Tumors abgeschätzt. Das Ansprechen auf HSP90-Inhibitoren wurde durch EMT-Signaturen in Melanomen, die mit der Aktivität des onkogenen Transkriptionsfaktors MITF verbunden waren, zuverlässig vorhergesagt. Schließlich wurde gezeigt, dass eine *in vitro* Stimulation von EMT durch eine TGF- β Vorbehandlung Melanom-Zelllinien für den HSP90-Inhibitor Luminespib sensibilisieren kann, was auf eine kausale Komponente von EMT hindeutet.

Schließlich wird das Onkologie Biomarker Entdeckungskonzept (OncoBird) für die Darlegung von molekularen Veränderungen und Biomarkern in klinischen Studien für die Präzisionsonkologie vorgestellt [3]. Diese ermöglicht explorative Subgruppenanalysen in klinischen Studien, indem sie somatische Veränderungen in Tumorsubtypen, sich gegenseitig ausschließende somatische Veränderungen und ihre prädiktiven Komponenten in Tumorsubtypen untersucht. Dieses Analysekonzept wird in der randomisierten, kontrollierten klinischen Phase-III-Studie FIRE-3 für die Begutachtung von metastasiertem Darmkrebs vorgestellt. Durch seine Anwendung in dieser Studie zeigten sich neben *KRAS* Mutationen auch Amplifikationen des Chromosom Arms 20q (chr20q)

einen prädiktiven Nutzen von Cetuximab im Kontext der Konsensus-Molekular-Subtypen (CMS). Dieses Konzept wurde auch auf die ADJUVANT klinische Studie zur Begutachtung von nicht-kleinzelligen Lungenkrebs angewendet, die konsistente prädiktive Biomarker für den Nutzen von Gefitinib ergab. Der Vergleich mit dem aktuellen klinischen Behandlungsstandard und gängigen Methoden zur datengesteuerten Subgruppenanalyse ergab Übereinstimmungen zwischen den von den einzelnen Methoden vorgeschlagenen Subgruppen, die durch Gruppen verwandter Biomarker repräsentiert wurden. Zudem zeigte dieser Vergleich, dass OncoBird robuste Biomarker für kleinere Subgruppen identifizieren kann die höhere Behandlungseffekte vorhersagen. Um weitere Forschung und Entwicklung zu fördern, ist das entwickelte OncoBird R Programm in einer Docker Shiny Anwendung verfügbar, die eine reproduzierbare Demonstration der vorgestellten Analyse enthält.

Zusammenfassend leistet diese Arbeit durch die systematische Untersuchung verschiedener Krebsmerkmale mit den hier vorgeschlagenen Analysemethoden einen Beitrag zu den laufenden Bemühungen der datengesteuerten Entdeckung von prädiktiven Biomarkern für Krebsbehandlungen. Darüber hinaus liefert sie einen Leitfaden für jene Bemühungen, indem sie die wichtigsten Überlegungen und Herausforderungen für die Nutzung biomedizinischer Daten und datengesteuerter Methoden für die Entdeckung von Biomarkern für die Arzneimittelentwicklung in der Onkologie aufzeigt.

Acknowledgements

I owe a debt of gratitude to numerous individuals who have supported and guided me throughout the journey of this work.

Foremost, I am grateful to my supervisor, PD Dr. Michael P. Menden, for his guidance, invaluable insights and constant encouragement. Thank you for the opportunity to shape my research with your experience, patience, and dedication.

I want to extend my sincere appreciation to my mentors and fellow researchers, Prof. Dr. Wolfgang Enard, Prof. Dr. Dieter Saur, Prof. Dr. Volker Heinemann, Prof. Dr. Emanuel Gonçalves, Prof. Dr. Annalisa Marsico, Prof. Dr. Sebastian Stintzing, Prof. Dr. Daniel Krappmann, Prof. Dr. Julien Gagneur, Prof. Dr. Jan Baumbach, Dr. Mathew Garnett, Dr. Benjamin Schubert, Dr. Matthias Heinig, Dr. Sebastian Vosberg, Dr. Ignacio Ibarra, Dr. Thomas O'Neill, Dr. Arndt Stahler and Dr. Julian Holch for their insightful feedback and constructive criticism. Their diverse perspectives in our joint discussions have shaped my ideas, expanded my horizon and enriched my work.


As I reflect on this journey, I would like to express gratitude to my colleagues and fellow researchers in my research group, including Ana Galhoz, Ginte Kutkaite, Phong Nguyen, Göksu Avar, Anantharamanan Rajamani, Dr. Ali Farnoud, Christina Hillig, Martin Meinel, Daniel Garger, Dr. Ines Assum, Dr. Diyuang Lu, Nikita Makarov, Maria Bordukova, Clara Meijs, Alina Arenth, Elisabeth Noheimer, Dr. Fabio Boniolo and many other colleagues from the institute. Our coffee breaks, discussions and shared experiences have made this journey enjoyable both intellectually and personally.

Alongside, I owe my dear friends Srikanth, Rayner and Cem for giving me a sense of normalcy during these times. I am deeply thankful to my brother David and my parents, Marlise and Heinrich, for their unconditional support and understanding. Lastly, to my partner Alissa, you have been my driving force, and your devotion has made it possible for me to pursue this endeavour.

Curriculum Vitæ

Personal Information

Name Alexander Joschua Ohnmacht

ORCID  <https://orcid.org/0000-0002-1481-9426>

Education

- 2019 – 2024 **Dr. rer. nat.**, *Faculty of Biology, Ludwig-Maximilians-Universität München*
Thesis title: Advancing the discovery of predictive biomarkers in drug high-throughput screens and clinical trials for precision oncology,
Computational Health Center, Helmholtz Munich
- 2016 – 2019 **M. Sc.**, *Faculty of Physics, Ludwig-Maximilians-Universität München*
Thesis title: Pattern formation in geometry and networks,
Department of Statistical and Biological Physics, Ludwig-Maximilians-Universität München
- 2012 – 2016 **B. Sc.**, *Department of Physics, Universität Konstanz*
Thesis title: Coarse-grained models for polymers,
Department of Chemistry and Biochemistry, University of Oregon
- 2004 – 2012 **Abitur**, *Fürstenberg Gymnasium, Donaueschingen*

Publications

The publications associated with this work are the following:

- Section 2.1 Ohnmacht, A. J. *et al.* The pharmacoeigenomic landscape of cancer cell lines reveals the epigenetic component of drug sensitivity. *Communications Biology* **6** (2023). URL <https://doi.org/10.1038/s42003-023-05198-y>
- Section 2.2 Ohnmacht, A. J. *et al.* The pharmacogenomic assessment of molecular epithelial-mesenchymal transition signatures reveals drug susceptibilities in cancer cell lines. *bioRxiv* (2024). URL <http://dx.doi.org/10.1101/2024.01.16.575190>
- Section 2.3 Ohnmacht, A. J. *et al.* The Oncology Biomarker Discovery framework reveals cetuximab and bevacizumab response patterns in metastatic colorectal cancer. *Nature Communications* **14** (2023). URL <https://doi.org/10.1038/s41467-023-41011-4>

Additionally, the following publications have been contributed to:

- Holch, J. W. *et al.* Refining first-line treatment decision in RAS wildtype (RAS-WT) metastatic colorectal cancer (mCRC) by combining clinical biomarkers: Results of the randomized phase 3 trial FIRE-3 (AIO KKR0306). *Journal of Clinical Oncology* **42**, 13–13 (2024). URL http://dx.doi.org/10.1200/JCO.2024.42.3_suppl.13
- Lu, D., Pamar, D. P., Ohnmacht, A. J., Kutkaite, G. & Menden, M. P. Enhancing gene expression representation and drug response prediction with data augmentation and gene emphasis. *bioRxiv* (2024). URL <http://dx.doi.org/10.1101/2024.05.15.592959>
- Vosberg, S. *et al.* DNA methylation profiling refines the prognostic classification of acute myeloid leukemia patients treated with intensive chemotherapy. *HemaSphere* **6**, 378–379 (2022). URL <https://doi.org/10.1097/01.hs9.0000844804.00811.c6>
- Boniolo, F. *et al.* Artificial intelligence in early drug discovery enabling precision medicine. *Expert Opinion on Drug Discovery* **16**, 991–1007 (2021). URL <https://doi.org/10.1080/17460441.2021.1918096>
- Nguyen, P. B. H., Ohnmacht, A. J., Sharifli, S., Garnett, M. J. & Menden, M. P. Inferred ancestral origin of cancer cell lines associates with differential drug response. *International Journal of Molecular Sciences* **22**, 10135 (2021). URL <https://doi.org/10.3390/ijms221810135>
- Farnoud, A., Ohnmacht, A. J., Meinel, M. & Menden, M. P. Can artificial intelligence accelerate preclinical drug discovery and precision medicine? *Expert Opinion on Drug Discovery* **17**, 661–665 (2022). URL <https://doi.org/10.1080/17460441.2022.2090540>

Declaration of contributions

Apart from Chapter 2, the remaining work in Chapters 1 and 3 was not previously presented and has not been published. All used sources are acknowledged as references. All co-authors and their contributions for Chapter 2 are stated below and are also documented in the respective publications. The shared-first authors are indicated with asterisks and have confirmed their contribution accordingly.

- Alexander J. Ohnmacht (A.J.O.; author of this thesis)
- Anantharamanan Rajamani (A.R.)
- Göksu Avar (G.A.)*
- Ginte Kutkaite (G.K.)
- Emanuel Gonçalves (E.G.)
- Dieter Saur (D.S.)
- Michael P. Menden (M.P.M.)
- Marisa K. Schübel (M.K.S.)*
- Thomas J. O’Neill (T.J.O.)
- Daniel Krappmann (D.K.)
- Arndt Stahler (A.S.)*
- Sebastian Stintzing (S.S.)*
- Dominik P. Modest (D.P.M.)
- Julian W. Holch (J.W.H.)
- Christoph B. Westphalen (C.B.W.)
- Linus Hölzel (L.H.)
- Ana Galhoz (A.G.)
- Ali Farnoud (A.F.)
- Minhaz Ud-dean (M.U.)
- Ursula Vehling-Kaiser (U.V.)
- Thomas Decker (T.D.)
- Markus Möhler (M.M.)
- Matthias Heinig (M.H.)
- Volker Heinemann (V.H.)

Section 2.1 Conceptualisation, A.J.O. and M.P.M.; Data curation, A.J.O. and A.R.; Analysis, A.J.O. and M.P.M.; Methodology, A.J.O. and M.P.M.; Supervision, M.P.M.; Visualisation, A.J.O.; Writing original draft, A.J.O. and M.P.M.; Writing, review and editing, A.J.O., A.R., G.A., G.K., E.G., D.S. and M.P.M.

Section 2.2 Conceptualisation, A.J.O. and M.P.M.; Data curation, A.J.O., G.A., M.K.S. and T.J.O.; Analysis, A.J.O., G.A. and M.K.S.; Methodology, A.J.O., G.A., M.K.S. and M.P.M.; Supervision, M.P.M. and D.K.; Visualisation, A.J.O. and G.A.; Writing original draft, A.J.O., M.K.S., G.A. and M.P.M.; Writing, review and editing, A.J.O., G.A., M.K.S., T.J.O., D.K. and M.P.M.

Section 2.3 Conceptualisation, M.P.M. and V.H.; Data curation, A.S., S.S., D.P.M., U.V., T.D., M.M. and A.J.O.; Formal analysis, A.J.O.; Methodology, A.J.O., A.S. and M.P.M.; Supervision, V.H. and M.P.M.; Visualisation, A.J.O. and L.H.; Writing original draft, A.J.O., A.S., V.H. and M.P.M.; Writing, review and editing, A.J.O., A.S., S.S., D.P.M., J.W.H., C.B.W., L.H., A.G., A.F., M.U., U.V., T.D., M.M., M.H. and M.P.M.

Contents

1	Introduction	1
1.1	Biomarkers in precision oncology	2
1.2	Cancer systems biology for precision oncology	2
1.2.1	Cancer genomics	2
1.2.1.1	DNA sequencing in cancer	3
1.2.1.2	Cancer drivers and mutational patterns	3
1.2.2	Cancer hallmarks	4
1.2.2.1	Cancer transcriptomics	5
1.2.2.2	Epithelial-mesenchymal transition	6
1.2.2.3	Epigenomics and DNA methylation in cancer	6
1.2.3	Cancer vulnerabilities, drug targets and discovered drugs	8
1.2.4	Pharmacogenomics for precision oncology	10
1.3	Principles of data-driven discovery of predictive biomarkers	11
1.3.1	Statistical learning	12
1.3.2	Hypothesis testing	13
1.3.3	Multiple testing	14
1.3.4	Central methods for machine learning	16
1.3.4.1	Model training	16
1.3.4.2	Model selection	16
1.3.4.3	Model validation	17
1.3.4.4	Model interpretation	17
1.3.5	A framework for data-driven biomarker discovery	18
1.4	Data-driven discovery of predictive biomarkers for clinical data in oncology	19
1.4.1	Estimating treatment effects in randomised controlled clinical trials	20
1.4.2	Survival analysis	21
1.4.3	Subgroup and biomarker discovery in clinical trials	22
1.4.4	Multiplicity adjustment for subgroup discovery in clinical trials	23
1.4.5	Estimating causal effects	23
1.5	Data sources for cancer research	24
1.5.1	Omics data repositories	25
1.5.2	Functional data	27
1.5.3	Databases for chemical compounds, drug targets, biological processes and cancer genomics	28
1.6	Methods for identifying predictive biomarkers	29
1.6.1	Methods for (epi)genome-wide association studies and differential gene expression	29
1.6.2	Methods for drug efficacy prediction	30
1.6.3	Methods for subgroup analysis	31
1.7	Regulatory considerations	33
1.8	Rethinking conventional drug discovery	34

1.9	Aims of the thesis	34
2	Results	37
2.1	The pharmacoepigenomic landscape of cancer cell lines reveals the epigenetic component of drug sensitivity in cancer, <i>Communications Biology</i> (2023)	38
2.2	The pharmacogenomic assessment of molecular epithelial-mesenchymal transition signatures reveals drug susceptibilities in cancer cell lines, <i>bioRxiv</i> (2024)	50
2.3	The Oncology Biomarker Discovery framework reveals cetuximab and bevacizumab response patterns in metastatic colorectal cancer, <i>Nature Communications</i> (2023)	70
3	Discussion	87
3.1	Conclusions	88
3.1.1	Biases in drug high-throughput screens and clinical trials and their analysis	88
3.1.2	Multifaceted biomarkers	90
3.1.3	Evaluation of causality	90
3.1.4	Challenges for the translation of biomarkers from preclinical to clinical studies	91
3.1.5	Establishing molecular profiling as predictive biomarkers in clinical studies	92
3.2	Outlook	93
3.2.1	Molecular profiling of tumour plasticity	93
3.2.2	Acquired drug resistance and drug combinations	93
3.2.3	Advances in modelling of response mechanisms in cancer	94
3.2.4	Advances in estimating treatment effects in cancer clinical studies	95
3.2.5	Enabling virtual drug discovery and treatment recommendations	96
3.3	Closing statement	96
	References	98
A	Abbreviations	121
A.1	List of acronyms	122
A.2	List of proteins	124
A.3	List of genes	125
B	Supplementary material	129
B.1	The pharmacoepigenomic landscape of cancer cell lines reveals the epigenetic component of drug sensitivity in cancer, <i>Communications Biology</i> (2023), supplementary information	130
B.2	The pharmacogenomic assessment of molecular epithelial-mesenchymal transition signatures reveals drug susceptibilities in cancer cell lines, <i>bioRxiv</i> (2024), supplementary information	139
B.3	The Oncology Biomarker Discovery framework reveals cetuximab and bevacizumab response patterns in metastatic colorectal cancer, <i>Nature Communications</i> (2023), supplementary information	147

Chapter 1

Introduction

Despite the improved outcomes of cancer patients over the last decades, cancer remains a substantial burden for humankind. Cancer is a disease in which cells proliferate uncontrollably in local and distant tissues, and is often fatal when cancerous cells impair the function of vital organs [10]. Due to the complexity and heterogeneity of the disease, possible avenues to tackle cancer include the development of precision therapies [11]. They rely on the fact that each cancer develops unique ways to control genes that determine its therapy response, and displays these characteristics representing vulnerabilities to be targeted by cancer treatments. In other words, precision oncology studies how patients and their tumours respond to therapies according to their response patterns. Identifying these patterns and their so-called biomarkers not only provides better treatment opportunities by tailoring the target group, but also may enhance our understanding of the disease and guide the discovery and development of compounds that exploit new cancer vulnerabilities. This chapter gives an introduction to this field. First, biomarkers are defined. Secondly, cancer is introduced in terms of systems biology, encompassing its development, hallmarks and possible targets. Third, general principles of data-driven biomarker discovery in preclinical studies for translational cancer research are presented and expanded to clinical cancer research. Fourth, appropriate data sources and methods for biomarker discovery are presented that take advantage of these concepts. Afterwards, regulatory aspects and the transformative potential of data-driven biomarker discovery for drug discovery efforts are discussed. Finally, within this scope, the general aims of this thesis are outlined.

1.1 Biomarkers in precision oncology

The high heterogeneity of tumours implies that clinical decision-making ought to be tailored to the characteristics of a patient and tumour. This is the dogma of precision oncology and relies on the discovery of biomarkers. In general, a biomarker is an objective, quantifiable characteristic of biological processes that can be utilised for diagnosis, prognosis and treatment decisions [12]. Accordingly, a biomarker can be diagnostic, prognostic or predictive [12]. Diagnostic biomarkers help determine relevant characteristics, prognostic biomarkers are surrogates for disease progression, and predictive biomarkers yield a measure for the treatment success probability of a particular treatment regimen accompanied by companion diagnostics for their diagnosis. Furthermore, a predictive biomarker can be called causal if an intervention to modulate the marker causes a change in the treatment success probability. In the clinical standard today, biomarkers in oncology do not usually go beyond genetic events. Specifically, from the 86 targeted therapies that require diagnostic biomarker testing approved between 1998 and 2022, 69 therapies (80%) had an associated genetic biomarker [13]. Their applications are attributed to their causal component and stability in the cancer disease aetiology.

1.2 Cancer systems biology for precision oncology

In 1914, Theodor Boveri first suggested that aberrant chromosomal changes may be the primary cause of cancer [14]. Accordingly, genetic changes in human cells confer the formation of a cancer as an evolutionary process that alters cellular function [15], a process which is termed *oncogenesis*. Genes that are causally linked to oncogenesis are called cancer genes. These are usually classified into *oncogenes* and *tumour suppressor genes*. When altered, the former promote oncogenesis, whereas the latter lose their ability to inhibit oncogenesis [16]. After having its roots in retroviral research, *RAS*, the first oncogene in the human genome, was discovered in 1982 [17]. Successively, the MAPK signalling pathway was depicted [18], which couples extracellular growth signals to intracellular signalling cascades to promote proliferation and survival pathways (Fig. 1.1a). Single biological mechanisms that are often found in cancer cells have been described as *cancer hallmarks* [19]. During the following centuries, new hallmarks were added that demonstrated the complexity of the disease [20, 21] and suggested that the only avenue was to treat cancer with a systems approach. The centrality of *RAS* and kinases of other mutated oncogenes, such as *BRAF* and *PIK3CA*, in many human cancers triggered a century-long search for treatments targeting them [22]. However, after the initial excitement of precision oncology in the clinical practice targeting these oncogenes [22], even after achieving the targeting of $\text{KRAS}^{\text{G12C}}$ and regulatory approvals [23], cancer remains a major obstacle. Modern technological innovations in sequencing and bioinformatics made the characterisations of tumours become progressively rich and refined [24]. Today, more than 3% of genes in the human genome are presumably involved in cancer [25]. About 80% of cancer patients have potentially targetable alterations with existing compounds [13, 26], and predictive biomarkers for standard treatments are now available for roughly 32% of cancer patients [13, 27]. This highlights the success of precision oncology and advocates expanding the efforts to identify actionable mutations or shifting the focus from the cause and exploring non-mutational cancer mechanisms to discover new cancer vulnerabilities and targeting opportunities, which can eventually result in suitable patient stratifications for clinical application [28].

1.2.1 Cancer genomics

Somatic mutations in the DNA are events that occur during the lifespan of an individual. They can be the result of internal or external mutagenic exposures that cause DNA defects, which are converted to mutations by erroneous DNA replication or DNA repair [29, 30]. Mutations which confer a fitness advantage of cancerous cells towards oncogenesis are called *driver mutations*. In contrast, cancer genomes can also include passenger mutations, which do not serve an evolutionary advantage and are not involved in oncogenesis [30]. Common types of alterations include base substitutions leading to missense or nonsense mutations in the encoded protein, structural variants such as insertions or deletions causing protein frameshifts, chromosomal translocations leading to gene deregulation or chimeric transcripts, and copy number alterations [31] (Fig. 1.1b). As an example, the $\text{KRAS}^{\text{G12D}}$ mutation is a

common missense mutation that produces the oncogenic protein KRAS^{G12D} by depriving it of its GTPase functionality and therefore locking it in its active GTP-bound state [16] (Fig. 1.1c,d). First efforts have assembled the Cancer Gene Census, which has constituted 291 cancer genes encompassing oncogenes, tumour suppressor genes and fusion partners [31], which is continuously expanding to 719 reported today [25]. These have been mostly neglecting mutations in non-coding regions, but efforts are taken to reveal their function [32, 33, 34].

1.2.1.1 DNA sequencing in cancer

The first wave of genomic discovery was based on previously laborious and expensive polymerase chain reaction (PCR) and exon-by-exon direct sequencing methods, which therefore focused on promising genes that encoded protein kinases for therapeutic targeting [16]. For example, BRAF^{V600E} mutations were discovered in cancer cell lines [35]. The resulting oncogenic BRAF plays a central role in the MAPK signalling pathway due to its upstream phosphorylation by RAS and downstream phosphorylation of MEK (Fig. 1.1e). Other examples include the discovery of activating mutations in *ERBB2* [36] or *PIK3CA* [37].

After the introduction of the first massively parallel high-throughput sequencing platform in 2005 [38], next-generation sequencing (NGS) has been the enabling technology towards the discovery of the cancer genome using whole exome sequencing (WES) [39] and whole genome sequencing (WGS) [40]. WES spans about 1% of the human genome, whereas WGS can capture about 99%. However, the amount of times a given genomic region is sequenced on average in one experiment, i.e. the sequencing depth, differs. Specifically, since WES usually operates on a sequencing depth of 100×, it is more accurate for mutation calling than WGS, which usually has a lower depth of 30×.

The massively parallel sequencing has since expanded to most protein-coding genes and, for instance, has led to the discovery of *IDH1* mutations in glioblastoma [41] as a crucial epigenetic regulator. Caveats for the computational analysis of cancer genomes include the detection of rare somatic events, the analysis of highly disarranged genomes and tumour heterogeneity [28]. Furthermore, the lack of driver mutations in non-coding regions of the genome may merely reflect the limited understanding of gene regulatory landscapes. Over the last decade, efforts have expanded towards the discovery of complex structural variations and rearrangements, mutational patterns in tumour evolution and heterogeneity or RNA alterations along with others [42, 43].

Targeted sequencing (TS) is less efficient for discovering these types of alterations. However, its great sequence depth makes it useful for analysing clinical samples, for which DNA quality or tumour contents can be low [44]. For example, the analytically and clinically validated companion diagnostic platform built by Foundation Medicine sequences more than 300 genes with a sequencing depth of 500×, which has been applied to clinical trial settings [45].

1.2.1.2 Cancer drivers and mutational patterns

The further elucidation of the cancer genome requires expanding the analysis of mutational patterns across the genome for detecting functional cancer mutations. In the current understanding of tumour progression, a single driver mutation is not sufficient to confer oncogenesis and multiple events are required, which are acquired successively and cooperate to drive oncogenesis [10]. In colorectal cancer for example, the first step of oncogenesis is often an inactivation of the tumour suppressor gene *APC* in about 60% of patient tumours, which is a critical negative regulator of cell growth [46].

Oncogenesis can arise through different evolutionary routes with the hijacking of different cancer genes in a stochastic manner [16]. However, the resulting tumour mutational patterns are not entirely random. For example, driver mutations are often mutually exclusive, such as *BRAF* and *KRAS* in colorectal cancer (COREAD), which are both key signalling proteins within the MAPK/ERK pathway [47] (Fig. 1.1f). The straightforward explanation for this pattern is the functional similarity of BRAF and KRAS, i.e. either mutation suffices to drive oncogenic signals through the activation of the MAPK signalling and no selective advantage exists for mutations in the counterpart [48]. Another part of the explanation could be the context-specificity of both mutations dependent on *APC* mutations. While *KRAS* and *APC* co-occur, *BRAF* mutations are found more frequently in *APC* wild

type mutations [49]. Accordingly, there may be more of a selective advantage to *KRAS* mutations in *APC* mutant tumours than to *BRAF* mutations. The conclusion is that driver mutations can be context-specific not only in terms of the tissue type but also in terms of their genetic background.

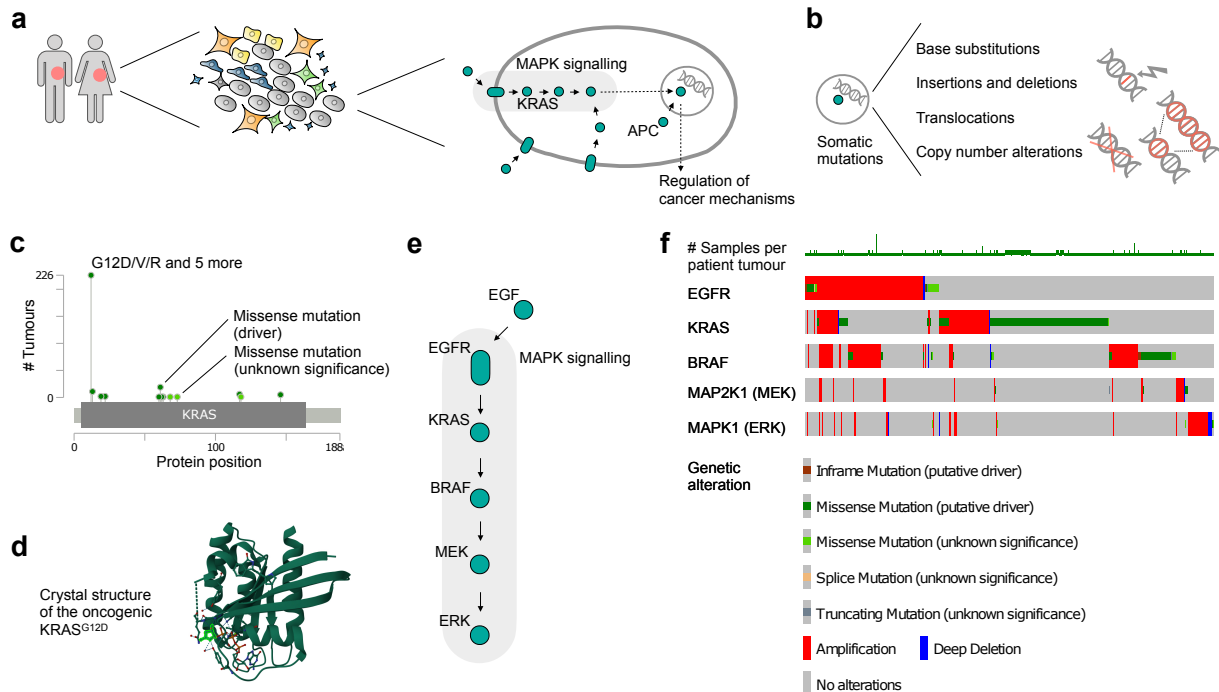


Figure 1.1: Cancer genomics as the foundation of precision oncology. **a** Simplified depiction of human cancers, the cellular composition of the dissected tissue and extra- and intra-cellular regulation of selected cancer mechanisms. In general, cancer cells exploit rewired cellular signalling pathways to enable cancer hallmarks. **b** Oncogenesis is initiated by somatic mutations and copy number alterations in DNA sequences. **c** Lollipop plot of the protein position of somatic missense mutations on *KRAS*, visualised with MutationMapper on cBioPortal [50] using whole genomes from the ICGC/TCGA [43], demonstrating the high-frequency mutations on the amino acid positions G12D/V/R. **d** Crystal structure of oncogenic *KRAS*^{G12D} visualised on the Protein Data Bank (PDB) with ID code 6GJ5 [51]. **e** Simplified depiction of the MAPK signalling with its central proteins and key downstream signalling events. **f** Oncoprint of mutually exclusive somatic mutations on proteins in the MAPK signalling pathway, visualised with OncoPrinter on cBioPortal using whole genomes from the ICGC/TCGA [43].

1.2.2 Cancer hallmarks

Cancer mutations in protein-coding sequences are responsible for altering protein function to dysregulate cell signalling to become cancerous. During the study of the human cancer genome, genetic alterations were found that can alter functions in epigenetics, chromatin modifications, cell signalling, metabolism and gene expression [52]. The characteristic cancer processes that enable cancer formation and progression have been continuously classified into so-called cancer hallmarks [19, 20, 21]. Today, cancer hallmarks encompass 14 processes that may alter cell function in terms of cell growth, growth control, immune evasion, inflammation, invasion, angiogenesis, genome instability, evading apoptosis, metabolic regulation, epigenetic reprogramming, polymorphic microbes, cell senescence, phenotypic plasticity and other emerging processes [21].

For example, a primary hallmark of cancer is its sustained proliferation [19]. In healthy cells, growth signals are required before a cell can proceed with its division. Mutated oncogenes can substitute this signal and promote

cancerous cell proliferation by modifying the extracellular, transcellular or intracellular processes [20]. Driver mutations in extracellular growth factors are rare, and even if the dependence of oncogenesis on those is often alleviated, they play a role in the stages of tumour progression and therapy resistance [53]. For example, the ERBB2 receptor overexpression leads to increased sensitivity to regular abundance of growth factors or to signal transduction independent of ligand binding due to its increased heterodimerisation with other members of the ERBB family of tyrosine receptor kinases [54]. Sustained proliferation signals can also stem from activating mutations in the cytoplasmic signalling circuits, such as the MAPK/ERK pathway. Its members KRAS and BRAF are only found mutated in about 30% and 8% of cancer types, respectively [26]. Thus, these pathways and molecular mechanisms are not the only way cancer can persist and progress. For example, the cytokine TGF- β usually stimulates apoptosis or differentiation through SMAD proteins [55], whereas in cancer it can induce the epithelial-mesenchymal transition for invasion and metastasis [56].

Research usually focuses on a handful of canonical cancer processes for inferring actionable cancer mechanisms that can enable cancer hallmarks [26, 57]. Thereby, the relevant processes within the scope of this work will be highlighted in the following sections.

1.2.2.1 Cancer transcriptomics

Cell transcripts in the form of messenger RNA (mRNA) contain transcribed DNA templates that are successively translated into proteins to engage in their appropriate cell functionality. This propagation of information from the genetic code to interacting proteins is known as the ‘central dogma’ of molecular biology (Fig. 1.2a). While mRNA is the most studied type of RNA, other RNA types are non-coding and can be responsible for regulating various cellular processes [16]. Compared to DNA sequences, transcriptomic profiles are highly dynamic and provide snapshots for the functional state, cell types and tissue-specific regulatory mechanisms of cells.

To monitor RNA production, hybridisation-based DNA microarrays have been a prominent profiling approach [16], while RNA sequencing (RNA-seq) is commonly used today due to less background noise, non-reliance on existing genome knowledge and ability to distinguish different isoforms and allelic expression [58]. This also allows researchers to study alternative splicing, post-transcriptional mechanisms and RNA editing [59].

Cancer is characterised by aberrant transcriptomic profiles that can reflect cancer hallmarks and tend to be highly conserved across tissues [60]. This is plausible when considering that many cancers converge to activate canonical cancer pathways through common oncogenes and tumour suppressors. However, transcriptional footprints of cancer signalling networks are highly context-specific. Therefore, it is common to define gene sets based on variations in gene expression to yield transcriptional signatures of cancer or differentiate cancer from each other. For example, different cancer types can be easily classified by their tissue gene expression levels (Fig. 1.2b). This started with an effort to distinguish acute lymphoblastic leukaemia (ALL) from acute myeloid leukaemia (AML), which was achieved with the expression levels of 50 genes that most correlated with these two conditions [61]. As an example shown in the uniform manifold approximation and projection (UMAP) of a harmonised transcriptomics dataset (Fig. 1.2b), the *IDH1* mutational status is a clear separator represented by the gene expression footprints of glioblastoma and low-grade glioma. Quantifying these types of variation can result in clinically impactful subtype classifications, such as the intrinsic breast cancer subtypes, which are defined by overexpression of *ERBB2*, estrogen or progesterone receptors [62].

Nowadays, subtypes are often based on transcriptional signatures encoded in machine learning models that are interpreted *post hoc* to reveal relevant tumour biology. For example, melanoma subtypes have been defined by transcriptomic signatures that revealed a dedifferentiation trajectory that reflects melanoma progression [63]. Similarly, the consensus molecular subtypes (CMS) in colorectal cancer revealed tumour subtypes that are enriched in distinct cancer hallmarks and somatic alterations [64]. Specifically, CMS2 was enriched in copy number alterations and left-sided tumours, whereas CMS1 was enriched in *BRAF* mutations and right-sided tumours. Their prognostic relevance in clinical trials was confirmed [65], however, it remained an open question if they can impact treatment decision-making.

1.2.2.2 Epithelial-mesenchymal transition

The epithelial-mesenchymal transition (EMT) is a program which is characterised by epithelial cells transitioning dynamically and reversibly to a mesenchymal phenotype in which they are deprived of their polarity and adhesive properties, restructure their cytoskeleton and extracellular matrix, and reprogram transcription to increase cell motility and their invasive capabilities [66].

It plays a critical role in the early stages of embryonic development, wound healing, fibrosis and cancer [10, 67]. In carcinomas, malignant cancer cells are confined by the baseline membrane and different types of cell-to-cell junctions and thus require a phenotypic switch to gain fibroblast-like morphology and mesenchymal characteristics for becoming motile and enable the invasion of surrounding tissue and metastasis [68], which is specified as a cancer hallmark [20] (Fig. 1.2c).

Its coordination does not require any DNA alterations. Instead, EMT is induced by factors such as TGF- β emitted from a reactive stroma that activate key transcription factors (TF) [69, 70] (Fig. 1.2c). During EMT, they pleiotropically downregulate common epithelial markers, which can be measured in gene expression or post-transcriptional changes. For example, E-cadherin is found on the cell membrane and mediates forming and maintaining cell-to-cell adherens junctions through its homotypic interactions between adjacent cells [71]. Its encoded gene *CDH1* is downregulated and replaced by upregulating N-cadherin, which only forms weak bonds and thus increases cell motility [10]. In addition, while cytokeratins make up the cytoskeleton of epithelial cells, mesenchymal cells contain vimentin to mediate spindle cell shape, which is reflected by molecular markers such as *VIM* expression.

Tumours have been shown to display a diverse spectrum of EMT programs [67] and high cellular plasticity is demonstrated through the fact that tumour subpopulations can show different and hybrid stages of EMT in the microenvironment [72]. This may contribute to the tumour's ability to adapt and anticipate external stresses. For example, tumours with mesenchymal features often show resistance to chemotherapeutic and immunotherapeutic treatment regimens [70]. The general resistance to common anticancer drugs has been attributed to the increased activity of anti-apoptotic pathways and slower proliferation rates [73].

Systematically targeting EMT by its prevention is infeasible since it encompasses a process necessary for homeostasis, and tumours may disseminate circulating tumour cells (CTC) before diagnosis. Furthermore, promoting mesenchymal-epithelial transitions (MET) is precarious since it may promote the formation of metastatic lesions from already present CTCs [74]. Therefore, selectively targeting transitioned cells may pose a viable option [73]. The essentiality of EMT for invasion and metastasis, as well as therapeutic response and resistance, advocates to expand efforts towards revealing its role as a predictive biomarker for therapeutic regimens and its associated mechanisms.

1.2.2.3 Epigenomics and DNA methylation in cancer

Epigenetics is particularly relevant for understanding differentiation in development because the whole developmental program must be written in the genome. Its concept was first introduced by Waddington in 1942 as the 'epigenotype' [75]. It led to the discovery of modifications that can overwrite genetic blueprints, regulate gene activity and continuously mediate interactions between the genome and cytoplasmic proteins that recognise these epigenetic marks [76]. These types of covalent modifications are reversible and dynamic and play major roles in gene regulation. In particular, the DNA is packaged and organised by histone complexes forming nucleosomes [16]. These proteins show modifications in the form of methylation or acetylation, for which particular combinations enable the binding of other regulatory proteins [16]. For example, transcription start sites are often free of nucleosomes that are not methylated if the nucleosomes they are flanked by are marked by di- or tri-methylation of histones H3 lysine 4 (H3K4) [77, 78] (Fig. 1.2d).

Apart from histone modifications, direct modifications of the DNA are another mechanism by which DNA sequences are marked for the recognition of DNA-binding proteins. In particular DNA methylation in the form of 5-methylcytosine (5mC) CpG dinucleotides, cytosine followed by guanine in 5' \rightarrow 3' direction, is frequently studied in the human genome [79, 80, 81].

Many methylation profiling techniques use sodium bisulfite treatment that measurably converts unmethylated cytosine to uracil [82]. These technologies fall into two categories, i.e. probe-based and sequencing-based. The former includes Infinium BeadChip arrays, which use probes that hybridise at the target CpG. Depending on the generation, it includes approximately 450,000 or 850,000 CpG sites [83]. An example of the latter is whole-genome bisulfite sequencing (WGBS), which in principle can screen all 28 million CpG sites in the human genome; however, it typically only has sufficient coverage for 15 million sites [84]. Alternatively, reduced representation bisulfite sequencing (RRBS) can only cover around 4 million CpG sites in the genome, but reduces the sequencing burden, allows higher throughput and increases confidence by allowing higher sequencing depth [85].

DNA methylation primarily shows in CpG islands that are usually around 1kb long, are enriched in CpG sequences and colocalise to approximately 70% of human proximal promoters [86] (Fig. 1.2d). Since 5mC can be converted to thymine via spontaneous deamination, it is thought that these regions only exist because they are usually hypomethylated in the germline [87]. However, about 70% of CpG sites in regions with low CpG density are actually methylated. For example, repetitive elements and transcribed regions of the gene body are usually methylated [88] and distal regulatory elements, such as enhancers, contain tissue-specific, highly variable and dynamic DNA methylation [89].

Inverse correlations between CpG island methylation in promoters and gene expression levels are observed frequently [79]. This led to the belief that methylation contributes to gene silencing by inhibiting the accessibility of the promoters for TFs (Fig. 1.2d). Accordingly, it was later found that about 22% of TFs showed decreased binding in their methylated motifs [90]. Indeed, DNA is also shown to ‘lock’ the repression of CpG island promoters after DNA has been encased by nucleosomes [80]. However, there are still discussions about silencing transcriptional initiation by methylation as the preferred regulatory mechanism to control expression levels, and studies rather propose that silenced genes precede methylation [80]. Rather than a lock, it is proposed that DNA methylation functions as a molecular mark for memorising and maintaining gene silencing [91]. Furthermore, for gene bodies and distal regions in particular, also positive correlations between methylation and expression levels can be observed [79], which suggest a context-specific function of DNA methylation.

Unlike somatic mutations, epigenetic mechanisms are dynamic and do not alter DNA base pair sequences. Thus, it has long been discussed if epigenomics in the form of DNA methylation plays a causal role in oncogenesis. As first evidence, gene body methylation was found to be a mutagen for classical tumour suppressor genes such as *TP53* [92], and in general, promoter hypermethylation of other tumour suppressors was linked to the silencing of key pathways for cancer progression [93]. For example, the promoter of the tumour suppressor P16, encoded by *CDKN2A*, was found to be hypermethylated and silenced in approximately 20% of cancers [94]. After the discovery of somatic alterations in genes that encode epigenomic regulators necessary for active methylation of the DNA, the discussion on its causal role was settled [95]. For example, the recurrent gain-of-function mutations of *IDH1* in gliomas and AML produce an onco-metabolite, 2-hydroxyglutarate, which interferes with the TET2 demethylating activity and causes hypermethylated DNA and subsequently altered regulatory interactions [96, 97]. Accordingly, the restoration of TET2 functionality can block leukaemia progression [98]. This type of non-mutational epigenetic reprogramming of cancer cells by epigenetic regulators has been recently added to cancer hallmarks [21]. Conversely, heterozygous R882H mutations in the DNA methyltransferase DNMT3A in AML reduce the methyltransferase activity by about 80% through disrupting tetramerisation that causes focal hypomethylation across the genome [99].

Cancer cells can be distinguished from healthy cells with their global methylation patterns created by these epigenetic regulators. For example, frequent hypermethylation of promoters has been defined as the CpG island methylator phenotype (CIMP) [100], which is discussed for a handful of cancer types [100]. To reiterate the previous example, the CIMP status reveals footprints in the glioma transcriptomes and mutations in *IDH1* seem to be sufficient to explain this phenotype [101] (Fig. 1.2b). This highlights the inherent dependencies between somatic mutations, DNA methylation and gene expression to determine cellular phenotypes.

Apart from global methylation patterns, local methylation patterns can reveal more actionable vulnerabilities. For example, the silencing of DNA repair gene *MGMT* is found with a hypermethylated promoter and precedes high rates of *TP53* and *KRAS* mutations at later tumour progression stages due to the predisposition of deficient cells to alkylation damage at guanines [102]. Thus, these epigenetic events can precede tumour initialising

mutations that are necessary for oncogenesis in the classical sense. That leads to the conclusion that genomics and epigenomics cooperate to unlock oncogenic potential, i.e. epigenetic changes can cause further mutations that can alter epigenetic regulators. Further basic research is required to reveal mechanistic interplays between chromatin remodelling, the hierarchy of gene silencing, the DNA methylation machinery, histone post-translational modifications and the DNA methylome in both healthy cells and cancer [103].

1.2.3 Cancer vulnerabilities, drug targets and discovered drugs

In the search for new cancer treatments, drug discovery traditionally starts with target identification and validation. The current cancer targets are diverse, which is attributed to the large space of opportunities to rewire pathways in cancer hallmarks during oncogenesis to arrive at a viable state to strive. The main principle of cancer therapies is to identify and selectively interfere with its essential processes, i.e. vulnerabilities, and sparing essential processes of healthy cells and tissues. Targetable entities are usually in the form of genes, proteins or other molecules that tumours depend on.

The first cancer therapy with chemical compounds was nitrogen mustard, which was originally developed for chemical warfare [105]. It is a cytotoxic alkylating agent that binds to DNA and forms cross-links that trigger apoptosis [105]. Its effectiveness in some cancer types resulted in the approval of mustard gas (mechlorethamine) by the Food and Drug Administration (FDA) in 1949 [106], the first chemotherapy. The earliest anti-cancer drugs are cytotoxic agents that target processes involving DNA and/or RNA, such as the inhibition of mitosis or induction of DNA damage. This class includes alkylating agents, antimetabolites, anthracyclines, topoisomerase inhibitors and anti-microtubule agents [107].

Their high toxicity in human patients has promoted the development of more selective agents with targeted therapies. The causal role of altered oncogenes and tumour suppressors in oncogenesis suggests targeting their activity. Indeed, it was shown that shutting down the signals stemming from oncogenes or restoring tumour suppressor activity in cancer cells unexpectedly inhibits their growth and induces apoptosis, termed oncogene addiction or tumour suppressor hypersensitivity [108]. Since cancer is thought of as a multifaceted disease, it is somehow surprising that the restoration of just a single altered protein function could have such tremendous effects. The first approved targeted treatment was tamoxifen, which inhibits the estrogen receptor and modulates its activity for the treatment of breast cancer [109].

The two main molecular targeting techniques are monoclonal antibodies and small-molecule inhibitors [110]. Within a cell, most key cancer signalling proteins are kinases, which are commonly used as targets by small molecules to inhibit their activity [110]. Conversely, proteins on the cell surface can be targeted through antibodies [110]. Some kinases, such as the ERBB receptor family, can be targeted by both techniques. Nevertheless, targeting EGFR with either the monoclonal antibody cetuximab or the small molecule gefitinib can still show different response patterns in cancer patients due to their differences in basic properties and mechanisms of action (MOA) [110]. Great successes were achieved by tyrosine kinase inhibitors (TKI). For example, imatinib was developed for chronic myelogenous leukaemia by targeting the BCR-ABL1 fusion [111] and gefitinib or erlotinib targeting EGFR in lung cancer [112].

Over the years, it became more apparent that therapeutic targets are not only confined to the list of classical mutated oncogenes or tumour suppressors, but ought to be expanded to non-oncogene dependencies [113]. As an example, the melanocyte master regulator MITF is often labelled an oncogene in melanomas and is proposed to act as a rheostat to regulate melanoma dedifferentiation and progression [63, 114]. Melanoma cancer cells in CRISPR screening experiments have been revealed to be self-addicted to MITF [115], which means that only melanoma cells with high *MITF* expression are addicted to it.

Another example is *IRF4*, an oncogene that is often overexpressed because of its chromosomal translocations in myeloma. Strikingly, cancer cells are addicted to aberrant IRF4 regulatory networks also if IRF4 is unaltered [116]. This demonstrates the concept of *synthetic lethality*, which states that non-oncogenic proteins can be essential for a tumour with a particular oncogenic alteration [117] (Fig. 1.2e). In general, since the loss-of-function of tumour suppressors is usually not directly targetable, the concept of synthetic lethality becomes relevant for drug target discovery. Among the not directly targetable oncogenic mutations are *BRCA1/2* alterations in breast cancer, which

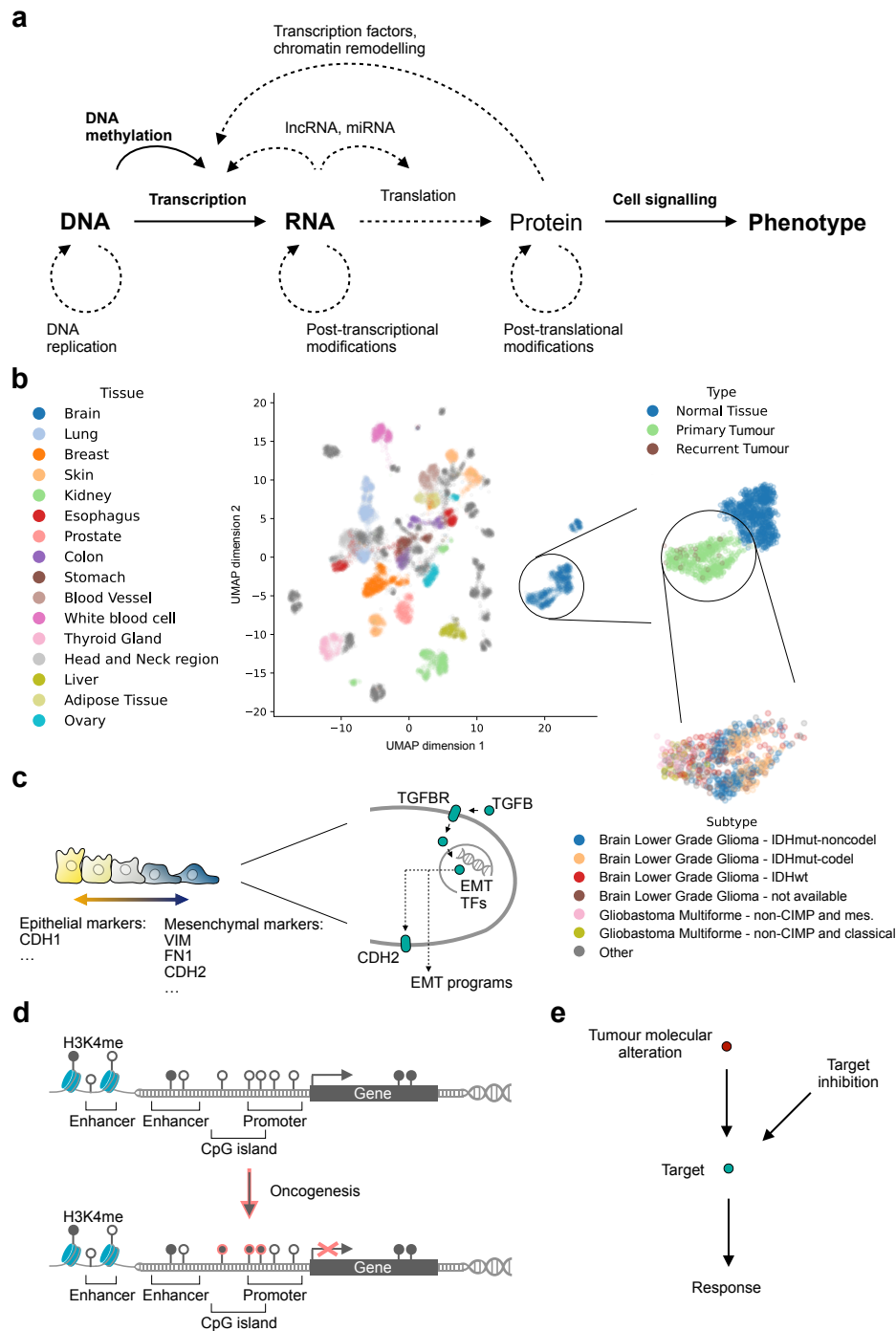


Figure 1.2: Tumour molecular profiling as the cornerstone for targeting cancer.

a Selected processes of gene regulation that map cancer genotypes to phenotypes, for which processes with bold text are addressed within this work. **b** UMAP representation of gene expression data of tissue samples obtained from harmonised GTEx, TCGA and the Therapeutically Applicable Research to Generate Effective Treatments initiative (TARGET) data [104], with subtypes of brain tumours as an example. **c** Simplified depiction of cancer cells undergoing EMT by down- and up-regulation of epithelial and mesenchymal markers, respectively. Extracellular stimuli promote the EMT program by activating its associated transcription factors. **d** Schematic visualisation of DNA methylation in histones, enhancers, CpG islands and promoters. As an example mechanism, promoters become hypermethylated and lead to the repression of target gene transcription during oncogenesis. **e** If the drug target (turquoise) is synthetically lethal with a tumour molecular alteration (red), the response to target inhibition depends on the presence of the alteration.

provide a successful proof of concept [118]: The inhibition of PARP leads to defective DNA repair of single-strand breaks and results in increased amounts of DNA double-strand breaks, from which only tumours with wild-type DNA repair protein *BRCA1/2* can recover [10]. Since mutations in tumour suppressors *BRCA1/2* impair the DNA damage repair of double-strand breaks, the synthetically lethal inhibition of PARP results in an accumulation of double-strand breaks and result in cell death [119]. Heat shock proteins (HSP) are another class of drug targets that are actively studied, but not categorised as cancer genes in a traditional sense. HSPs are responsible for the folding of their many client proteins to ensure their appropriate activity in both healthy and cancer cells. For example, while wild-type *BRAF* activity does not depend on HSPs, the oncogenic *BRAF*^{V600E} stability depends on HSP90 in melanoma cells [120].

The interplay between somatic cancer mutations and the cancer epigenome suggests targeting epigenetic cancer mechanisms. Epigenetic regulators such as DNMT proteins can be targeted by azacytidine or its derivative decitabine, a broad nucleoside DNMT inhibitor. Albeit its regulatory approval, their clinical utility is still in question due to elusive mechanisms because of the broad impact of DNMT on many genomic regions, limited activity in solid tumours, toxicity, poor pharmacokinetic properties and the lack of biomarkers to predict treatment efficacy [121]. Among the first-generation broad epigenetic modulators were also HDAC inhibitors blocking the activity of histone deacetylases to interfere with cancer-specific gene expression [122]. A more promising avenue may be targeted approaches, which include inhibitors of mutated *IDH1*, that have been shown to inhibit the growth of glioma cells, even though the inhibition of DNMT can be at least as effective [123]. Epigenetic synthetically lethal interactions are in discussion [124]; however, challenges associated with the targeted therapies exploiting epigenetic mechanisms are the recapitulation of epigenetic profiles in cultured cancer cells and the scarcity of available epigenetic data in the public domain [125].

In total, 332 compounds across 57 cancer targets gained regulatory approval between 2009 and 2020 [126]. New technologies enabling large-scale chemical or genetic screens facilitate the discovery of synthetically lethal targets. Emerging targets that gain attention are nodes and hubs in cancer-related protein-protein interaction networks, metabolic processes, master regulators, tumour plasticity, the tumour microenvironment and immune components [117]. In conclusion, the increasing universe of cancer targets and the larger amount of compounds currently in development paired with the discovery of synthetically lethal interactions for selective targeting will further contribute to optimising the efficacy of cancer treatments.

1.2.4 Pharmacogenomics for precision oncology

Pharmacogenomics aims to study the role of the genome in determining drug responses. Historically, it evolved from pharmacogenetics, which typically involved the study of a particular genetic polymorphism to alter drug effects [11], which was first introduced in 1959 [127]. Until the 1990s, most pharmacogenomic studies focused on inherited traits that alter enzymes involved in drug metabolism, disposition and transporters [128, 129]. For example, the 2 – 10% of individuals with homozygous non-functional *CYP2D6*, a drug-metabolising enzyme responsible for both detoxification and prodrug activation, are resistant to many opioid analgesics [130].

Upon the study of genetic polymorphisms in drug targets that alter drug efficacy, efforts in pharmacogenomics have been increasing [131]. In particular, the introduction and wide adoption of genome-wide sequencing technologies has accelerated the discovery of new variants in the human genome that could be subsequently associated with drug effects [11]. In the special case of cancer, strong pharmacogenomic interactions have been observed for somatic alterations [132]. An example is the fact that *KRAS* mutant *COREAD* does not respond to anti-EGFR therapies because the oncogenic signalling of *KRAS* is independent of upstream activation by EGFR (Fig. 1.1e). Other examples, such as mutations in *EGFR* determining response to gefitinib in non-small-cell lung carcinomas (NSCLC) or *ERBB2* overexpression as predictive biomarker for efficacy of treatment with trastuzumab in breast cancer [133, 134], has outlined the path for the future of pharmacogenomics to investigate drug response patterns in terms of molecular cancer data.

Drug development since then has led to a handful of success stories. For example, the *EML4-ALK* fusion oncogene was discovered in NSCLC patients, and the observation that cancer cell lines with this alteration showed higher sensitivity to ALK inhibitors promoted the use of crizotinib for this patient subgroups [135]. Recent ad-

vances in cancer immunotherapies also make use of response biomarkers, e.g. PD-1 inhibitors tend to show higher responses in tumours with high *PD-L1* expression, high tumour mutational burden (TMB) or mismatch repair deficiency (MMRd) across many cancer types [136]. Furthermore, the discovery of causal pharmacogenomic interactions between PARP and *BRCA1/2* mutated tumours led to the approval of the PARP inhibitor olaparib [119]. Today, there are 517 pharmacogenomic biomarkers involved in drug labelling, from which 221 are attributed to oncology [137]. From all 164 targeted therapies approved between 1998 and 2022, about half require diagnostic testing of their associated biomarkers [13].

Epigenetic alterations currently do not play a major role in precision oncology, but evidence for epigenetic biomarkers of drug response is expanding. Among the earliest and most established examples is the promoter methylation of *MGMT* [138, 139]. Its hypermethylation is associated with the downregulation of MGMT, which is a DNA damage repair protein for alkylation lesions by removing methyl groups from the O-6 position of guanine residues [16]. In glioblastoma, a hypermethylated *MGMT* promoter increases susceptibility to alkylating agents such as temozolomide [140]. It alkylates or methylates DNA at this guanine residue position, from which tumours with DNA repair impairment cannot recover. In accordance with the idea of pharmacogenomics, *pharmacoeigenomics* studies the impact of the epigenome in determining drug responses. Other than this example, there are only a handful of biomarkers used in oncology [141], leaving the predictive component of DNA methylation and the rest of the epigenome elusive.

A major obstacle in cancer is drug resistance; thus, identifying its biomarkers from drug response assays is crucial. Key differences between intrinsic and acquired resistance need to be considered. The former is characterised by an initial lack of response (short time scale), while the latter occurs after an initial response in a relapse (long time scale). Thereby, resistances can occur both through genetic mutations or non-mutational tumour plasticity [142]. For example, EMT was reported to be both an intrinsic and acquired resistance mechanism for *KRAS*^{G12C} inhibition in NSCLC [143]. In drug high-throughput screens (HTS), intrinsically resistant cell lines can be found among non-responding cell lines, which can share molecular characteristics with tumours from patients with acquired resistances [144].

These endeavours show that pharmacogenomics can accelerate cancer drug discovery and development two-fold through distinguishing drug responses. Firstly, molecular alterations that drive this stratification between responding and non-responding tumours may yield novel synthetic interactions and promising drug targets for designing new therapies. Secondly, it enables the retrospective analysis of toxicity or response patterns in clinical trials that could reveal patient subgroups tailored specifically to the compound of interest. An example for both cases are *KRAS* mutations, which are both an attractive drug target and a resistance marker for anti-EGFR therapies. Thus, pharmacogenomics is possible across different drug discovery and development stages and should cross-inform each other. It is evolving into a general term for leveraging sequencing technologies to enable the discovery and development of new treatments [145] by revealing drug response patterns, which can be facilitated by data-driven modelling using statistics, machine learning and artificial intelligence [9].

1.3 Principles of data-driven discovery of predictive biomarkers

With the previous sections, it became apparent that the discovery of predictive biomarkers is central to the deployment of precision oncology. Predictive biomarker discovery can be understood as a classification task to stratify cancers into subgroups based on their characteristics to predict drug responses. They can depend on simple biomarkers such as anatomical, histopathological or molecular features. As the simplest example and arguing in a data-driven and disease-agnostic manner, cancers from different tissues of origin require different therapies, which would render the cancer type a predictive biomarker for the success of these administered drugs.

Single genetic markers may suffice to predict some treatment responses and encompass most clinically approved examples [13]; however, many patients still do not respond well to their administered therapies. Already in the earlier days of pharmacogenomics, it was proposed that the variability of drug responses should be viewed through pathways of cooperating genes characterised by multi-omic data modalities [146]. This urges research to employ a more holistic approach to biomarker discovery, which relies on the precise measuring of high-

dimensional characterisations of tumour cells or patients that can yield biomarker signatures based on multiple genes and other data modalities.

Predictive biomarkers are subsets of these measurable disease characteristics that are able to predict treatment outcomes. For their discovery, statistical and machine learning approaches are commonly used modelling techniques fuelled by abundant biomedical data for characterising patients and their tumours. The following sections outline principles from statistical learning and hypothesis testing to arrive at a framework for the discovery of predictive biomarkers in cancer.

1.3.1 Statistical learning

Statistical learning aims to find models describing observed data. Within the scope of this work, suppose the data $D = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(N)}, y^{(N)})\}$ is observed for either N tumours or patients. Thereby, each $\mathbf{x}^{(i)}$ consists of p observed characteristics $x_1^{(i)}, \dots, x_p^{(i)}$ and each $y^{(i)}$ consists of a single continuous outcome measurement for the i -th observation. Models derived from data are inherently probabilistic because of uncertainties underlying the observed data stemming from missing information and noise. Thus, we assume that the observations in D can be modelled by a joint probability distribution $P(\mathbf{X}, Y)$ of p random variables $\mathbf{X} = X_1, \dots, X_p$ and a random variable Y . We are interested in the function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ that describes the relationship between the measurable observations of \mathbf{X} and Y , given by

$$Y = f(\mathbf{X}) + \varepsilon, \quad (1.1)$$

where ε is the noise term. The minimisation of the expected squared error $\text{Err} = E[Y - f(\mathbf{X})]^2$ in a point-wise manner for the measurable outcome ('features') of variables $\mathbf{X} = \mathbf{x}$ [147] yields the solution

$$f(\mathbf{x}) = E[Y | \mathbf{X} = \mathbf{x}]. \quad (1.2)$$

In other words, the expectation value of Y (output) conditioned on $\mathbf{X} = \mathbf{x}$ (inputs) gives their relationship. This representation is the discriminative setting, which allows us to reduce the probabilistic nature of the problem to a prediction task by approximating the function f [147, 148].

The function f is usually parameterised by a set of parameters θ . We are interested in parameters $\hat{\theta}$ such that the estimated f fits the data well, i.e.

$$f(\mathbf{x}^{(i)}, \hat{\theta}) \approx y^{(i)} \quad \forall i = 1, \dots, N. \quad (1.3)$$

Imposing the input-output relationship in equation 1.1 formulates a supervised learning task in which parameters θ are learned from the data D . For estimating parameters $\hat{\theta}$, *maximum likelihood estimation* is a commonly employed concept [147]. If we recall the probabilistic representation of our data, $P(D|\theta)$ is the likelihood of observing data D for fixed parameters θ . We intend to maximise the log-likelihood $\mathcal{L}(\theta) = \log P(D|\theta)$ [147]. For the additive and Gaussian error with zero mean, $\varepsilon = \mathcal{N}(0, \sigma^2)$ in equation 1.1, the likelihood for each observed sample $(\mathbf{x}^{(i)}, y^{(i)})$ is given by [147]

$$P(y^{(i)} | \mathbf{x}^{(i)}, \theta) = \mathcal{N}(f(\mathbf{x}^{(i)}, \theta), \sigma^2) \quad \forall i = 1, \dots, N. \quad (1.4)$$

For independent and identically distributed samples in data D (i.i.d.), the likelihood factorises, and the negative log-likelihood is given by [147]

$$\mathcal{L}(\theta) = -\frac{N}{2} \log(2\pi) - N \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^N (y^{(i)} - f(\mathbf{x}^{(i)}, \theta))^2. \quad (1.5)$$

If the predicted values are given by $\hat{y}^{(i)} = f(\mathbf{x}^{(i)}, \theta)$, this is equivalent to minimising the loss function $l(y^{(i)}, \hat{y}^{(i)})$,

which is commonly chosen to be the residual sum of squares [147]

$$\begin{aligned} l(y^{(i)}, \hat{y}^{(i)}) &= \sum_{i=1}^N (y^{(i)} - \hat{y}^{(i)})^2 \\ &= \sum_{i=1}^N (y^{(i)} - f(\mathbf{x}^{(i)}, \boldsymbol{\theta}))^2, \end{aligned} \quad (1.6)$$

with respect to parameters $\boldsymbol{\theta}$ for finding the optimal parameters $\hat{\boldsymbol{\theta}}$. This minimisation has infinitely many possible solutions [147]. Hence, the choice of parameterising and constraining $f(\mathbf{x}^{(i)}, \boldsymbol{\theta})$ will be crucial. Choosing $f(\mathbf{x})$ as a linear combination of its inputs $\mathbf{x} = x_1, \dots, x_p$ is a popular imposed model, which is often adequate and specified by

$$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^p \beta_j x_j, \quad (1.7)$$

where $\boldsymbol{\beta} = \beta_0, \dots, \beta_p$ are the parameters that are estimated using minimisation of the residual sum of squares

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^N (y^{(i)} - \boldsymbol{\beta} \mathbf{x}^{(i)})^2, \quad (1.8)$$

which yields the parameters $\hat{\boldsymbol{\beta}}$ for the fitted function $f(\mathbf{x}, \hat{\boldsymbol{\beta}}) = \hat{f}(\mathbf{x})$. For this simple case, there exists a closed-form solution [147]. The components of the parameter $\hat{\boldsymbol{\beta}}$ are the coefficients for each input \mathbf{x} and thus give an estimate of their effect for predicting the outcome.

1.3.2 Hypothesis testing

Statistical hypothesis tests are fundamental methods for statistical inference on parameters $\boldsymbol{\theta}$ of an observed dataset D consisting of N samples. Some of the basic principles are comprehensively laid out by Fay and Brittain (2022) [149]. Accordingly, when testing a single hypothesis regarding the unknown parameter $\boldsymbol{\theta}$, we test a null hypothesis H_0 with an alternative hypothesis H_1 while stating that H_0 is false.

After defining the hypotheses, a *test statistic* is defined, which is a function of the sample values. The test statistic quantifies the discrepancy between the sample observations and what to expect under H_0 , with functional values indicating the evidence against the null hypothesis. Subsequently, a decision rule is defined, which depends on the test statistic and the *significance level* α , which is designed such that the rate of type I errors is $\leq \alpha$. Thereby, falsely rejecting a true null hypothesis is a type I error, while falsely not rejecting a false null hypothesis is a type II error (Table 1.1).

	H_0	H_1
Accept	Correct decision	Type II error
Reject	Type I error	Correct decision

Table 1.1: Scenarios of statistical hypothesis testing. Depending on the true hypothesis H_0 or H_1 , either type I errors (false positive) or type II errors (false negative) can occur.

The significance level α is the probability of rejecting the null hypothesis if the null hypothesis is true. Often $\alpha = 0.05$ is chosen, but since this decision is rather arbitrary, we resort to defining the *p-value*, which is the smallest α for which H_0 is still rejected for all larger α .

Showcasing a simple example, suppose that the sample data D consists of measurements of a single variable, i.e., $D = \{x^{(1)}, \dots, x^{(N)}\}$. We are interested in the true mean $\mu = \theta$ of the distribution $P_\mu(X)$ and intend to test if this mean is greater than μ^* . Then, let H_0 state that $\mu = \mu^*$, whereas H_1 states $\mu > \mu^*$. Furthermore, let $\bar{x} = \frac{1}{N} \sum_{i=1}^N x^{(i)}$ be the sample mean of the sample data and $s^2 = \frac{1}{N-1} \sum_{i=1}^N (x^{(i)} - \bar{x})^2$ the unbiased sample variance. Testing this hypothesis required a *one sample one-sided t-test*. For that, the normality of the true distribution is assumed and a

test statistic T is employed with the form [149]

$$T = \frac{\mu^* - \bar{x}}{\frac{s}{\sqrt{N}}} \sim t_{N-1}, \quad (1.9)$$

which is distributed according to the *Student's t-distribution* t_{N-1} with $N - 1$ degrees of freedom (d.o.f.). The p -value is then given by [149]

$$p = 1 - F_{t,N-1}(T), \quad (1.10)$$

for which $F_{t,N-1}$ is the cumulative t -distribution with $N - 1$ d.o.f.'s. The null hypothesis H_0 is then rejected if $p < \alpha$. A two-sided test for the hypothesis with the alternative hypothesis $\mu \neq \mu^*$ is achieved if two one-sided tests are employed with the p -value given by the minimal one-sided p -value multiplied by 2 [149].

In an alternative formulation by inverting a series of hypothesis tests [149], the $100(1 - \alpha)\%$ *confidence interval* (CI) is the set of parameters for which we fail to reject the null hypothesis with a significance level α . Accordingly, the $100(1 - \alpha)\%$ CI is given by [149]

$$\text{CI} = \bar{x} \pm F_{t,N-1}^*(1 - \alpha/2) \frac{s}{\sqrt{N}}, \quad (1.11)$$

for which $F_{t,N-1}^*(q)$ is the q -th quantile of the t -distribution with $N - 1$ d.o.f.'s. Accordingly, the null hypothesis is rejected if $\mu^* \notin \text{CI}$.

After this introduction, we follow with a relevant example for this work, i.e. testing model parameters. From the linearity and normality of the error term of the specified model in 1.7 follows that [147]

$$\hat{\beta} \sim \mathcal{N}(\beta, \text{Var}(\beta)), \quad (1.12)$$

a multivariate normal distribution \mathcal{N} with the coefficients β as mean and covariance matrix $\text{Var}(\beta)$. For that, the null hypothesis is formulated that for a particular coefficient $\beta_j = 0$ with the alternative hypothesis $\beta_j \neq 0$. To test this hypothesis, the test statistic (Z-score) is employed with [147]

$$z_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}} \sim t_{N-p-1}, \quad (1.13)$$

which is distributed according to t_{N-p-1} . The statistical significance can then be evaluated analogously to equations 1.10 and 1.11. For all the statistical tests employed within Chapter 2, the outlined process of testing statistical significance will be analogous.

1.3.3 Multiple testing

When carrying out multiple statistical tests, the probability of making at least one type I error can increase dramatically. Consider testing a family of m independent hypotheses with significance level α . Then, the *family-wise error rate* (FWER), the probability of falsely rejecting at least one true null hypothesis, is $1 - (1 - \alpha)^m$, which is monotonically increasing with m [149]. Thus, *multiplicity adjustment* methods are required to control the overall type I error rates [149].

Taking into account the amount of statistical tests by adjusting the derived p -values is a convenient way to address multiplicity. The associated adjusted p -value p^{adj} describes the smallest α that still rejects its associated hypotheses in the family. Thus, rejecting all hypotheses with $p^{adj} < \alpha$ controls the FWER at significance level α . In the following sections, a selection of procedures that adjust raw unadjusted p -values are presented.

The simplest procedure that controls the FWER is the Bonferroni procedure [150], for which

$$p_j^{adj} = p_j m, \quad (1.14)$$

simply multiplying individual p -values p_j by the amount of tested hypotheses m . There are no assumptions on the dependence between individual tests, but at the cost of being conservative when m becomes large, which decreases power and increases false negatives. Another single-step procedure that adjusts p -values independently from each other is the Šidák procedure [151]. It is slightly less conservative but assumes statistical independence of the m hypothesis tests. Adjusted p -values are given by

$$p_j^{adj} = 1 - (1 - p_j)^m, \quad (1.15)$$

following the rationale of the probability for at least one false positive among m independent hypotheses.

A step-wise procedure that is universally more powerful than the Bonferroni procedure without further assumptions is the (step-down) Holm procedure [152] given by

$$p_i^{adj} = \max_{j \leq i} p_j(m + 1 - j) \quad \forall i = 1, \dots, m, \quad (1.16)$$

with all p_j sorted in ascending order and j as the rank of each value. Other procedures are the Hochberg (step-up) [153] and Hommel [154] (step-down) procedure, which both assume a certain positive dependence of test statistics [155].

Other procedures are based on resampling methods using permutations or bootstrapping [156], which come with a higher computational cost, but mostly require no assumptions and adjust for dependencies in between hypotheses m for controlling the FWER. For example, let the dataset D have labels $y^{(1)}, \dots, y^{(N)}$ randomly permuted B times and for m hypotheses and let P^* be a matrix with components $p_{i,j}^*$ across permuted sets $i = 1, \dots, B$ and hypotheses $j = 1, \dots, m$. The first iterative step in the free step-down procedure by Westfall and Young [156] is then given by

$$p_1^{adj} = \frac{1}{B} \sum_{i=1}^B \mathbb{1}(\max_{j \in \{1, \dots, m\}} p_{i,j}^* \geq p_1), \quad (1.17)$$

where $\mathbb{1}(\cdot)$ is the indicator function. The k -th step is then given by

$$p_k^{adj} = \frac{1}{B} \sum_{i=1}^B \mathbb{1}(\max_{j \in \{k, \dots, m\}} p_{i,j}^* \geq p_k) \quad \forall k = 2, \dots, m. \quad (1.18)$$

Experiments for biomarker discovery, such as the datasets in Section 1.5, often require testing many hypotheses. In these cases, controlling the FWER at a significance level $\alpha = 0.05$ can be too stringent. Thus, it is proposed to control the *false discovery rate* (FDR), which describes the expected rate of false positive rejections among the total set of rejected hypotheses [149]. Since generally the FDR is less conservative than FWER, it can lead to higher power, i.e. lower chances of false negatives. The utilisation of the appropriate multiplicity correction highly depends on the context and prior knowledge about the hypotheses in question. A popular method for controlling the FDR is the (step-up) Benjamini-Hochberg procedure [157] with adjusted p -values

$$p_i^{adj} = \min_{j \geq i} p_j \frac{m}{j} \quad \forall i = 1, \dots, m, \quad (1.19)$$

with all p_j sorted in descending order and j as the rank of each value. While this procedure is conservative for certain positive dependence structures, for arbitrary correlation structures an extension was given by Benjamini and Yekutieli [158].

In modern genomics and epigenomics, studies are often performed in a genome-wide fashion that require unbiased evaluations and interpretations across many statistical tests, which increases the burden of multiple testing [159], which can be termed the ‘curse of multiplicity’ [160]. A range of procedures based on the presented groundwork are introduced and used in Chapter 2.

1.3.4 Central methods for machine learning

Machine learning aims to develop methods for learning patterns in data to arrive at models that make reliable predictions [148]. While the prior sections predominantly focused on linear regression as an important example, a variety of machine learning methods with different characteristics have been developed in the past solving different types of problems [148], which will be introduced in Sections 1.6.2 and 1.6.3. Thereby, the process of building models generally involves four steps, i.e. (1) learning model parameters ('training'), (2) selecting appropriate models and hyperparameters ('model selection'), (3) predicting and evaluating the model ('validation') and (4) interpreting the model and its parameters ('interpretation'), which will be the subject of the four subsequent sections.

1.3.4.1 Model training

The training step typically is carried out by minimising the loss function, which commonly takes the form of the squared error loss in equation 1.6 for regression models and in equation 1.8 for linear models. Other loss functions can be used or combined to guide the model fitting [147, 148]. Simply minimising loss functions to fit a model with many parameters on the data can lead to *overfitting*, which is the case if a trained model is fit on the input-specific data characteristics and therefore cannot perform generalisable predictions on unseen data [148]. For example, the variance explained by highly correlated variables in standard regression models is shared among them. This can lead to multicollinearity issues, such as instability of coefficients, that result in unreliable predictions.

To alleviate this issue, regularisation can be used to control the model fitting and enforce sparsity by modifying the optimisation problem given in equation 1.6 [147]. A widely used regularisation strategy for regression is shrinkage, i.e. 'shrinking model coefficients' to trade bias (reduced model space) for decreased variance, for achieving a better fit and to receive more interpretable predictions. For example, the elastic net promotes the interpretability and stability of regression coefficients by promoting sparsity with the lasso component and redistribution of the magnitude of regression coefficients among correlated variables with the ridge component. With the elastic penalty term, the optimisation problem reads

$$\min_{\beta} \sum_{i=1}^N (y^{(i)} - \sum_{j=1}^p \beta_j x_j^{(i)})^2 + \lambda \sum_{j=1}^p (\alpha \beta_j^2 + (1 - \alpha) |\beta_j|), \quad (1.20)$$

including both a lasso and ridge penalty with hyperparameter λ and tuning each component with α , and yielding parameters of the elastic net $\hat{\beta}$ [161]. This and other types of regularisation techniques are employed for virtually all learning algorithms and depend on the chosen algorithm [147].

1.3.4.2 Model selection

Model selection involves choosing the best model from a set of candidate models [148]. For example, models can depend on *hyperparameters*, which have to be set prior to training in contrast to model parameters that are fitted during the training process [147]. Machine learning methods can have intrinsic hyperparameters and forms of regularisation can add further hyperparameters. Their selection is called *tuning*, which can be carried out by various algorithms. However, in the simplest case, a grid search is employed, i.e. different sets of hyperparameter combinations are tested and the hyperparameters with the best performance are selected.

Additionally, if performances upon employing regularisation and tuning are still unsatisfactory, especially in the case $N \ll p$, feature selection is a popular way to filter non-informative variables to avoid overfitting in order to decrease variance [147, 160]. This is particularly well-suited for biomarker discovery, since its final step is the selection of a biomarker set from relevant features of the model that should be understood in terms of model selection. However, this can result in model selection bias, which occurs if the model is selected based on the best-performing features and thus overestimates their true contribution [160]. This can be viewed as an alternative formulation of the 'curse of multiplicity' in Section 1.3.3. Namely, inflated type I error rates when testing multiple features hamper inferences with overfitting or model selection bias as its analogous concept [160]. For each of the

presented applications, the respective techniques are introduced and used in Chapter 2.

1.3.4.3 Model validation

Once the model is fitted, it can be used to predict unseen independent data from which a generalisation performance can be estimated. Importantly, if a dataset of sufficient size is available, the dataset can be split into a *training set*, *validation set* and *test set* before the training and model selection steps. Then, the training set can be used for training and model selection including feature selection and tuning; the validation set can be used for estimating prediction errors for selecting the final model; and the predictions in the test set can be used to estimate the performance of the selected final model, e.g. by quantifying the prediction error $\hat{\text{Err}}$.

Sometimes, a simple train-test split to assess performances suffices for model selection, but the estimated performance may depend on the exact train and validation split. Thus, the repeated use of one exact split for model selection of the model with the minimum validation prediction error may lead to model overfitting. To alleviate this, more commonly K -fold cross-validation (CV) is used. For this, the data D with the set of indices d is randomly partitioned into K roughly equally sized subsets D_k with their associated set of indices d_k . The error of the CV estimate is given by [147]

$$\hat{\text{Err}}_{\text{CV}} = \frac{1}{K} \sum_{k=1}^K \frac{1}{|d_k|} \sum_{i \in d_k} l(y^{(i)}, \hat{f}^{d \setminus d_k}(\mathbf{x}^{(i)})), \quad (1.21)$$

where $\hat{f}^{d \setminus d_k}$ is fitted with the subsets of data D with indices in $d \setminus d_k$. In other words, $K - 1$ folds are used for training, model selection and tuning, whereas the leftover fold is used for estimating the prediction error. Often, $K = 5$ or $K = 10$ is chosen as a good trade-off between bias and variance of the CV estimator.

Another popular method applicable to a wide range of estimation problems is bootstrapping [162]. Bootstrapped datasets are produced by randomly drawing samples with replacements from dataset D , which can be used to generate bootstrapped predictions [147]. The percentiles of the distribution of bootstrapped predictions account for the variability in the entire modelling strategy [147]. Thus, it can provide confidence intervals for estimation problems for which it is difficult to account for model selection using maximum likelihood [147]. Hence, it is used to correct biases of estimators and construct confidence intervals of estimated quantities in Section 2.3.

1.3.4.4 Model interpretation

Statistical learning for biomarker discovery can be assessed in terms of a prediction problem or testing of multiple hypotheses, but ultimately relies on the assessment of individual features. For predictive biomarkers to effectively be transferred from preclinical experiments to clinical practice, it is essential that they are robust and, ideally, reflect biologically plausible mechanisms.

If multivariate and nonlinear models are employed, the model parameters are difficult to interpret [147] and thus biological mechanisms are becoming increasingly elusive. Their interpretation would require either a *post hoc* assessments, i.e. identifying subsets of predictive features extracted from all features \mathbf{X} and an evaluation of their biological context [163], or a careful engineering of models to incorporate known contexts [164]. Conversely, linear models with fewer effective parameters suffice or can even outperform nonlinear alternatives in difficult scenarios such as small training data, low signal-to-noise ratio or sparsity [147]. Moreover, they provide high interpretability by specifying one model coefficient β_i for each X_i in equation 1.7. Additionally, if each feature X_i is used for building a set of univariate models, the problem can be formulated as testing multiple hypotheses for each associated single feature. Then, selecting the best models and their associated subsets of significant single features from all features \mathbf{X} allows for a straightforward interpretation of the predictive biomarkers by directly extracting model parameters β_i for the evaluation of their biological contexts.

Thus, the task of discovering predictive biomarkers can be approached in various ways depending on the data and use case. Further examples from the literature and the motivation of the modelling strategies in this thesis will be introduced in Section 1.6.

1.3.5 A framework for data-driven biomarker discovery

The general ideas in Sections 1.3.1-1.3.4 are the groundwork for building models for predictive biomarker discovery. To formulate this task, let all variables \mathbf{X} consist of three types of features $\mathbf{X} = (\mathbf{X}_d, \mathbf{X}_t, \mathbf{t})$, which are characterised by disease features \mathbf{X}_d (observed disease features $x_{d,1}, \dots, x_{d,q}$) in response to a certain treatment \mathbf{t} , a vector of binary treatment indicators for a combination of the n possible treatments $\mathbf{t} = t_1, \dots, t_n \in \{0, 1\}^n$, and by treatment features \mathbf{X}_t consisting of variables associated with the treatment (Fig. 1.3a).

The *outcome function* f predicts the measurable outcome Y of an entity of interest, such as tumours or patients. Then, a functional form for f is assumed in the form of [165, 166]

$$f(\mathbf{X}_d, \mathbf{X}_t, \mathbf{t}) = g\left(h(\mathbf{X}_d) + z(\mathbf{X}_d, \mathbf{X}_t, \mathbf{t})\right). \quad (1.22)$$

Thereby, h is a function that predicts the outcome independent of an applied treatment and refers to the *prognostic* component. Thus, it only depends on the disease features \mathbf{X}_d . In contrast, z is a *contrast function*, which refers to the *predictive* component. It depends on the treatment \mathbf{t} , its features \mathbf{X}_t and can also depend on disease features \mathbf{X}_d . g is a monotone function that takes the form of a link function [167].

For n possible single treatments and their pairwise combinations, z can be expressed in terms of the individual treatment contrast function $\mathbf{z}' = (z'_1, \dots, z'_n)$ expressing the predictive contribution of individual treatments t_1, \dots, t_n and the synergy contrast matrix \mathbf{Z}'' with components z''_{ij} expressing pairwise treatment interactions between t_i and t_j . With this, z can be written in terms of \mathbf{z}' and \mathbf{Z}'' as follows

$$\begin{aligned} z(\mathbf{t}) &= \mathbf{z}'^T \mathbf{t} + \mathbf{t}^T \mathbf{Z}'' \mathbf{t} \\ &= \sum_{k=1}^n z'_k t_k + \sum_{i=1}^n \sum_{j=1}^n z''_{ij} t_i t_j, \end{aligned} \quad (1.23)$$

where the dependencies on \mathbf{X}_d and \mathbf{X}_t are omitted for convenience. This work incorporates three different special cases of the problem formulation in 1.22 and 1.23¹. The considered features and treatments for each use case are introduced in the respective sections in Chapter 2. Thereby, Sections 2.1 and 2.2 focus on *in vitro* experiments in cancer cell lines, which offer both many treatment and disease features, which are decently sampled to allow models to potentially generalise for new tumours with features \mathbf{X}_d or new treatments when considering features \mathbf{X}_t ². In contrast, Section 2.3 contains clinical datasets that study only a handful of compounds. Thus, they only cover a small amount of treatment features but can contain and sample as many disease features, which allows models to generalise for new patients with features \mathbf{X}_d .

In Sections 2.1 and 2.2, the analysed experiments are designed such that the prognostic term h can be eliminated when quantifying drug responses³. Thus, modelling drug responses using the treatment contrast z suffices. Furthermore, only monotherapy treatments are measured, i.e. $\mathbf{Z}'' = 0$ and $|\mathbf{t}| = 1$, and each individual treatment contrast z'_k is estimated separately for each treatment t_k , so parameters are not shared across drugs and thus z_k are independent of the drug features \mathbf{X}_t . Assuming that z_k is linear in its disease features \mathbf{X}_d , the model for each drug k reads

$$\begin{aligned} z(\mathbf{X}_d, \mathbf{t}) &= z'_k(\mathbf{X}_d) \\ &= \beta_0 + \sum_{j=1}^q \beta_j x_{d,j}, \end{aligned} \quad (1.24)$$

with coefficients $\beta = \beta_0, \dots, \beta_q$ individually fitted for each index k . The coefficients β with high absolute values and significance assessed by hypothesis tests from Section 1.3.2 indicate predictive drug response biomarkers for

¹The matrix $\mathbf{Z}''(\mathbf{X}_d, \mathbf{X}_t)$ contains pairwise synergistic and antagonistic treatment components, which will be omitted within the scope of this work, i.e. $\mathbf{Z}'' = 0$ will be universally assumed.

²Within the scope of this work, the built models will focus on disease features \mathbf{X}_d rather than treatment features \mathbf{X}_t , which allows drawing conclusions only in the context of all the administered treatments.

³The analysed drug HTS introduced in Section 1.5.2 derive relative cell viabilities by quantifying differences to replicate untreated cells, and thus, the viability can be attributed to the predictive drug effect.

the treatment t_k . Thus, this approach models putative biomarkers explicitly, which allows for simple evaluations and interpretations. It may be insufficient because of its assumptions of linearity and constant variances, its sensitivity to outliers, multicollinearity of features and potential overfitting for $p \gg N$. However, many of these issues can be mitigated by the methods proposed in Chapter 2, rendering linear models such as equation 1.24 valuable for biomarker evaluations.

Sometimes, clinical data is analysed for only a single treatment with a clinical endpoint for quantifying an outcome, for which one can estimate the outcome function f . Therefore, since only a single treatment is observed, f is independent of \mathbf{X}_t and only one component of the individual treatment contrast \mathbf{z}' is non-zero, i.e.

$$\begin{aligned} f(\mathbf{X}_d, \mathbf{X}_t, \mathbf{t}) &= f(\mathbf{X}_d) \\ &= g\left(h(\mathbf{X}_d) + \mathbf{z}'(\mathbf{X}_d)\right). \end{aligned} \quad (1.25)$$

Since the prognostic and predictive components h and z cannot be distinguished, they are often modelled jointly using linear models with the form

$$f(\mathbf{X}_d) = g\left(\beta_0 + \sum_{j=1}^q \beta_j x_{d,j}\right), \quad (1.26)$$

with coefficients $\beta = \beta_0, \dots, \beta_q$ that contain both prognostic and predictive components.

Section 2.3 includes a framework for analysing clinical trials, i.e. there are two studied treatment regimes, and thus allows us to distinguish the prognostic and predictive components h and z . Since usually only two treatment regimens are observed in clinical trials, a single endogenous treatment indicator $t \in (\{0, 1\})$ is included, and f is independent of treatment features \mathbf{X}_t . Furthermore, linear models including treatment-covariate interaction terms are used for estimating both h and z , arriving at

$$\begin{aligned} f(\mathbf{X}_d, t) &= g\left(h(\mathbf{X}_d) + z(\mathbf{X}_d, t)\right) \\ &= g\left(h(\mathbf{X}_d) + \mathbf{z}'(\mathbf{X}_d)t\right) \\ &= g\left(\beta_0 + \sum_{j=1}^q \beta_j^{\text{prog}} x_{d,j} + \sum_{j=1}^q \beta_j^{\text{pred}} x_{d,j}t\right), \end{aligned} \quad (1.27)$$

containing two types of coefficients, i.e. prognostic β^{prog} and predictive β^{pred} components. Predictive components are interaction terms between disease features $x_{d,j}$ and the treatment indicator t , which can be directly estimated with the hypothesis tests from Section 1.3.2.

1.4 Data-driven discovery of predictive biomarkers for clinical data in oncology

The general term of subgroup analysis describes the effort of identifying patient subgroups distinguished by predictive biomarkers with superior (or inferior) responses to therapies. In the literature, this topic is most often discussed in the context of clinical trials and used in a retrospective and data-driven fashion, highlighting its inherent exploratory nature. Historically, scientists have been sceptical towards subgroup analysis [168], because an undisciplined analysis, so-called ‘data dredging’, can result in spurious false positive associations leading to both selecting the wrong subgroups and overestimating their effect size [169]. The statistical issues and often occurring malpractice [170] of conducted subgroup analysis have fuelled debates concerning their appropriateness in the past [171, 172], which has led to reporting standards in both the community and regulatory agencies [173, 174].

Some commonly discussed pitfalls for subgroup analysis are extreme effect estimates for small sample sizes, selection bias in its pre-planning, regression to the mean and occurrence of the Simpson’s paradox [169]. While

these issues can also occur in pre-specified subgroup analysis, issues for *post hoc* analyses include multiple testing [175] and the ‘cherry-picking’ of strong associations, i.e. selective inference [176]. Given the outlined issues and challenges, results from exploratory subgroup analysis must be interpreted with caution and cannot be utilised for confirmatory processes such as regulatory approval or drug label change [177].

However, in the era of precision medicine, in which the central idea is tailoring treatments to subgroup-specific characteristics, the analysis of subgroups becomes ever more relevant. As Alvin R. Feinstein stated, “statisticians are right in denouncing subgroups that are formed *post hoc* from exercises in pure data dredging. The clinicians are also right, however, in insisting that a subgroup is respectable and worthwhile when established *a priori* from pathophysiologic principles” (Alvin R. Feinstein, 1998, p. 299) [178]. Especially with the increasing relevance and abundance of biomedical and molecular data, subgroup discovery is becoming an essential aspect of modern drug development. Today, the main goal is to identify the right patient for the right treatment.

For the further development of subgroup analysis around these concerns, it is essential to clearly define its purpose in a given clinical trial. Thus, subgroup analysis is proposed to be divided into four categories [166]. First, *confirmatory subgroup analysis* in late-stage clinical trials is advised to only take place for a small number of subgroups defined prospectively [179]. For this, traditional multiplicity adjustment methods are essential for their evaluation to preserve the type I error rate [175]. In contrast, *exploratory subgroup evaluation* focuses on a smaller number (about 10) of pre-specified (or unexpected) subgroups and assesses the consistency of the overall treatment effects, typically via testing statistical interactions in equation 1.27 [177]. Similarly, *post hoc subgroup evaluation* is often employed for unexpected heterogeneity in treatment effects for regulatory or safety issues [177]. Lastly, *subgroup and biomarker discovery* is purely used for proposing candidate subgroups among an arbitrary number of subgroups in the available data. Typically, the algorithms designed for this purpose rely on principles from machine learning discussed in Sections 1.3, especially Section 1.3.4. By design, they require strategies for validation, preferably in an independent study cohort.

Subgroup analysis can be viewed as a special case of model selection [160, 169], i.e., finding the best subgroups from a candidate subgroup selection dependent on the parameters of the proposed procedure. The outcome function f in equation 1.22 can be directly estimated with the training data, and predictions of individual instances can be output by the final model. Then, these predictions can be directly tested with the test data using classic performance metrics since outcomes Y are observed. However, it is at the heart of precision medicine to estimate the predictive component z , which requires estimating the predictive terms in equation 1.27, which are often dominated by strong prognostic effects.

Having introduced some subgroup analysis terminology, the following sections will introduce heterogeneous treatment effects, censored variables, subgroup discovery, multiplicity adjustments and causal inference for guiding the assessment of predictive biomarkers in clinical trials.

1.4.1 Estimating treatment effects in randomised controlled clinical trials

The objective of exploratory subgroup discovery is often to identify subsets of patients with differential treatment effects by estimating parameters of the predictive components in equation 1.27. Many ideas in the literature extend into ideas from causal inference, which are commonly used for subgroup analysis today [180].

In a two-arm study, $E[Y|\mathbf{X}, t]$ is the expected outcome of a patient given the baseline disease features in \mathbf{X} treated with regimen t . A useful notion is the potential outcomes framework, for which the i -th patient can have multiple potential outcomes depending on the received treatment, i.e. $\tilde{Y}_i(t)$, of which all but the observed one are hypothetical [180]. This work focuses on randomised controlled clinical trials (RCT), for which the treatment assignment t is independent of the covariates \mathbf{X} used for subgroup analysis, i.e. $\{\tilde{Y}_i(t = 1), \tilde{Y}_i(t = 0)\} \perp\!\!\!\perp t_i$ [166, 180]. This is not necessarily the case in observational data, for which the distribution of patients in each treatment regimen may differ and therefore bias the inference. However, other types of confounding are possible if covariates in \mathbf{X} are correlated, e.g. highly positively correlated predictive biomarkers can be labelled predictive but are not causally linked. This is especially relevant for analysing genetic alterations, for which mutational patterns highlighted in Section 1.2.1.2 frequently occur.

Under the assumption of random treatment assignment of RCTs, the average treatment effect (ATE) is given

by [181]

$$\tau = E[\tilde{Y}_i(t=1) - \tilde{Y}_i(t=0)]. \quad (1.28)$$

In RCTs, a simple difference of observed outcomes is unbiased and consistent as an estimator [180]. For the assessment of heterogeneous treatment effects, the average treatment effect conditioned on disease features in \mathbf{X} is the conditional average treatment effect (CATE), given by [181, 182]

$$\tau(\mathbf{X}) = E[\tilde{Y}_i(t=1) - \tilde{Y}_i(t=0)|\mathbf{X}]. \quad (1.29)$$

The unconfoundedness assumption states that when conditioning on covariates \mathbf{X} , there are no further confounders that simultaneously have a direct effect on the outcomes $\tilde{Y}(t)$ and t that are present in \mathbf{X} , i.e. $\{\tilde{Y}_i(t=1), \tilde{Y}_i(t=0)\} \perp\!\!\!\perp t_i \mid \mathbf{X}$ [166, 180]. Under this assumption the ATE can be derived from the CATE with $\tau = E[\tau(\mathbf{X})]$, where the expectation value is taken over \mathbf{X} .

For estimating these quantities, the expected outcome can be formulated as [166]

$$Y_i = \tilde{Y}_i(t=1)t_i + \tilde{Y}_i(t=0)(1-t_i), \quad (1.30)$$

for which the expected values of both $\tilde{Y}_i(t=0)$ and $\tilde{Y}_i(t=1)$ can be estimated separately. Using a more heuristic approach, the expected outcome $E[Y|\mathbf{X}, t]$ can also be modelled with the outcome function f using its formulation in equation 1.27. The predictive components can then be used to form subgroups with differential treatment effects, after which one can estimate the average treatment effects in the subgroups [183]. In the estimated subgroups $\hat{A}(\mathbf{X})$ with outcome $Y(\hat{A}(\mathbf{X}))$, the expected value can be modelled by

$$E[Y(\hat{A}(\mathbf{X}))|\mathbf{X}, t] = \gamma_0 + \gamma_1 t, \quad (1.31)$$

where γ_1 is the CATE estimate. Since the CATE estimate in equation 1.31 was derived from the same dataset as subgroup \hat{A} , it suffers from bias due to selective inference. Thus, this estimate must be corrected with methods that are introduced and utilised in Section 2.3.

1.4.2 Survival analysis

While many datasets measure binary or continuous outcomes Y , censored outcomes consist of a time T^* and a binary event indicator C . The time T^* describes the actual survival time T or the time until censoring has occurred depending on the event indicator C . For example, since clinical studies are finite, a patient withdrawn from the study or otherwise lacking follow-ups will have an associated time T^* until censoring if the event has not occurred. Estimating mean survival times $E[T]$ is challenging since the actual survival times T are only known for uncensored events. Thus, censored outcomes are commonly modelled by estimating a survival function given by

$$S(\eta) = P(T > \eta), \quad (1.32)$$

which describes the probability of the observed survival time T being greater than η , from which $E[T] = \int_0^\infty S(\eta) d\eta$ [184]. If survival times $v_1 < v_2 < \dots$ from individuals are observed, the Kaplan-Meier estimator is used to estimate the survival function [184], which is given by

$$\hat{S}(\eta) = \prod_{j: v_j < \eta} \left(1 - \frac{c_j}{n_j}\right) \quad \text{for } \eta < \max(v_i), \quad (1.33)$$

where n_j and c_j are the number of subjects at risk and for which the event has occurred at time v_j , respectively. Thus, in the context of a clinical trial, this estimate can be used to evaluate treatment benefits using a log-rank test for testing the null hypothesis that there is no difference in the survival for the tested groups [185]. Since including additional features \mathbf{X} to adjust for is not possible for this estimator in a straightforward manner, other (semi-)parametric models, such as the Cox proportional hazards model, are often used. The hazard $H(\eta)$ describes the potential for an event occurrence at a given time unit given that the event did not yet happen, and can be

explicitly expressed by [167]

$$H(\eta) = -\frac{\frac{dS(\eta)}{d\eta}}{S(\eta)}. \quad (1.34)$$

In the presence of explanatory variables $\mathbf{X} = X_1, \dots, X_p$, the Cox proportional hazards model is specified by

$$H(\eta, \mathbf{X}) = H_0(\eta) \exp\left(\sum_{j=1}^p \beta_j x_j\right). \quad (1.35)$$

Note that the baseline hazard H_0 is independent of the characteristics of the subjects, which yields that the hazard ratio between two individuals is constant in time. Estimating hazard ratios and regression coefficients does not require estimating H_0 , which can be an advantage over parametric survival models, such as the Weibull model or exponential model, if the form of the hazard function is unknown [167]. This model can be fit through a partial maximum likelihood approach while only considering uncensored subjects. Minimising the negative partial log-likelihood similar to equation 1.5 is given by [186]

$$\min_{\boldsymbol{\beta}} \sum_{k=1}^N C_k \log \left(\sum_{i \in R_k} \exp \left(\left(\sum_{j=1}^p \beta_j x_j^{(i)} \right) - \left(\sum_{k=1}^p \beta_j x_j^{(k)} \right) \right) \right), \quad (1.36)$$

where C_k is the indicator for uncensored events and R_k the set of indices for subjects at risk from all subjects N . Analogously to the linear models in equation 1.24, the Cox model utilises a linear predictor to estimate regression coefficients and thus can be viewed as a generalised linear model tailored to the distributions of censored outcomes [167]. Thus, the presented ideas for standard outcomes can be analogously applied to censored outcomes while considering the necessary caveats.

1.4.3 Subgroup and biomarker discovery in clinical trials

The previously discussed scepticism regarding subgroup analysis has triggered the generation of checklists for best practices. Guidelines have been proposed before [187, 188, 189], and a list of recurring items has been collected [166]. Accordingly, subgroups should be pre-specified and biologically plausible, all tests must be adjusted for multiplicity, and all testing in subgroups should be only conducted if the tested coefficients for the predictive interaction terms in equation 1.27 is significant. However, some of these suggested guidelines contradict the nature of data-driven biomarker discovery for precision medicine. For example, the requirement of testing interactions for each pre-specified predictor variable ignores the possibility of exploratory and multivariate considerations. Having in mind that data-driven subgroup analysis can be viewed as a special case of model selection, i.e. forming subgroups from important features and estimate their treatment effects, it can be described as a search strategy for a final subgroup A . Formally, a predictive subgroup may be defined by [160]

$$z(\mathbf{X}) > \delta \implies \mathbf{X} \in A, \quad (1.37)$$

where A is the subgroups to be discovered and δ is a clinically relevant treatment effect, which can be selected as $\delta > 0$ indicating non-zero benefit or $\delta > \tau$ indicating benefit higher than the ATE [166]. Thus, the subgroup A is defined by subjects with features \mathbf{X} for which treatment contrasts $> \delta$ are found. Principles for data-driven subgroup analysis methods are derived from general principles of statistical learning (Section 1.3.4), multiple testing (Section 1.3.3) and causal inference (Section 1.4.5) [166, 177]: First, an assessment of the type I error rate should be given for the entire search strategy, which yields an estimate of how likely it is to find a treatment effect in the subgroup A by chance. Next, the strategy should incorporate controls for complexity and selection bias to prevent overfitting and provide an assessment for its reproducibility. Finally, it should provide an ‘honest’ estimate of treatment effects in the found subgroups [166, 190]. Based on these ideas, many data-driven subgroup analysis methods have been proposed that go beyond the classical approach to detect statistical treatment-covariate interactions in equation 1.27, which are introduced in Section 1.6.3. In Section 2.3, a method is presented that

builds on the classical ideas, uses resampling methods from Section 1.3.4 to fulfil the stated requirements, and includes benchmarks with modern methods to analyse clinical trials in oncology.

1.4.4 Multiplicity adjustment for subgroup discovery in clinical trials

Traditionally, multiplicity adjustments in confirmatory clinical trials control the FWER with the presented procedures in Section 1.3.3. Multiplicity arises from multiple endpoints, several drug dosages or several patient subpopulations [191]. Procedures for their adjustment in clinical trials with a single set of hypotheses and fixed design have been reviewed before [175]. For multiple sources of multiplicity, hypotheses can be either divided into separate families, or more complex procedures can be employed [192, 193].

In general, methods controlling the FDR are inappropriate for the confirmatory setting because of its non-stringent requirements, usually low number of hypotheses and its inability to support complex decision rules [192]. However, applications of the less conservative FDR are more common practice in genomics, because a low number of false discoveries is acceptable [159]. Thus, pharmacogenomic subgroup discovery for clinical trials with molecular tumour profiling can benefit from its adoption. For multiple types of families of hypotheses, controlling the FDR in each family can still conserve the overall error. This is rationalised by the correct scaling behaviour of rate with the number of tests m [194]. If a selection takes place for the tested families, adjustments due to the selective inference are required [195].

1.4.5 Estimating causal effects

The discovery of predictive biomarkers using hypothesis tests and predictive modelling relies on clever ways to design models, fitting procedures, multiple testing strategies and resampling procedures to yield valid inferences. Complementary, utilising methods from causal inference to learn predictive biomarkers seems to be a promising extension. In general, causal inference works with observational data, for which potential outcomes $\tilde{Y}_i(t)$ and treatment assignments t are not independent of the covariates \mathbf{X} . Thus, strategies are needed to account for different distributions of patients in treatment groups when estimating causal effects. For example, matching can be used to identify individuals with similar characteristics \mathbf{X} between treatment groups [196]. Alternatively, propensity score modelling, i.e. prediction of the treatment assignment t based on characteristics \mathbf{X} , can be used for matching by using inverse probability weighting [196].

Since estimating the CATE from equation 1.27 can suffer from model misspecification or low power for detecting interactions, several machine learning frameworks have been proposed for this purpose, from which we consider two categories, i.e. metalearners (S-/T-/X-learner) and double machine learning methods [182, 197, 198, 199, 200]. Both methodologies make use of baselearners, which can be arbitrary machine learning methods that function as backbones for inferring causal effects. As an example for the former, the simplest S-learner \mathcal{M}_S uses a single baselearner and is given by [182]

$$\mathcal{M}_S(\mathbf{X}, t) = E[Y|\mathbf{X}, t] = f(\mathbf{X}, t), \quad (1.38)$$

essentially globally fitting the outcome function f with an arbitrary machine learning model. From this, the estimated CATE is then given by

$$\hat{\tau}_S(\mathbf{X}) = \mathcal{M}_S(\mathbf{X}, t = 1) - \mathcal{M}_S(\mathbf{X}, t = 0). \quad (1.39)$$

Complementary, the T-learner \mathcal{M}_T fits one model per treatment in reference to equation 1.30, i.e. $\mathcal{M}_T^{\{0,1\}} = E[Y|\mathbf{X}, t = \{0, 1\}]$ [182]. From this, the estimated CATE is given by

$$\hat{\tau}_T(\mathbf{X}) = \mathcal{M}_T^1(\mathbf{X}, t = 1) - \mathcal{M}_T^0(\mathbf{X}, t = 0). \quad (1.40)$$

Metalearners combine predictions from baselearners to estimate the CATE, which enables them to be used in a flexible way. In contrast, the frameworks developed for double machine learning provide a more general theoretical

framework with new concepts for the estimation of heterogeneous treatment effects by partially linear regression models [198]. As an informal outline, it combines two baselearners for the propensity model and the outcome model with another final stage model to arrive at valid inferences of the treatment effects [198]. The estimates of the propensity model are used to orthogonalise the predicted treatment residuals, and the outcome model is used to remove the variance stemming solely from the features \mathbf{X} [199]. Finally, regressing the outcome residuals on the treatment residuals provides an estimate for the ATE, while the CATE is estimated by the third final stage machine learning model fitted on the features \mathbf{X} . If the structure of the CATE is generally unknown and many features are present in \mathbf{X} , a good choice for the final model are causal forests [200], since they provide honest treatment effect estimates and valid confidence intervals if their assumptions are met. All of these frameworks have shown promising results in simulations and real-world applications, even in the presence of complex data and confounding [182, 199, 200]. Thus, the double machine learning framework will be used to quantitatively estimate causal effects in Section 2.2.

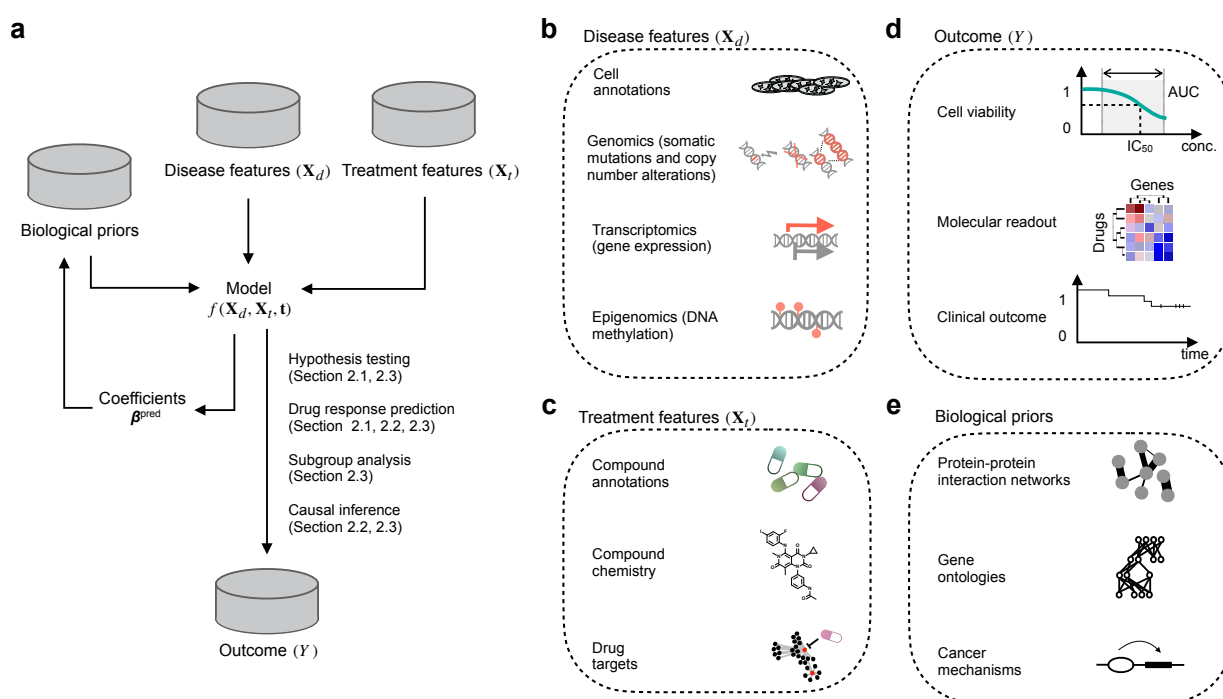


Figure 1.3: Data sources for cancer research enabling the development of statistical methods and machine learning for predictive biomarker discovery. **a** A standard workflow consisting of data inputs from disease features, treatment features and biological knowledge databases for building a model predicting the measured outcome from a functional screen using hypothesis testing (Section 2.1, 2.3), drug response prediction (Section 2.1, 2.2, 2.3), subgroup analysis (Section 2.3) and causal inference (Section 2.2, 2.3). **b** Data sources for disease features contain cell annotations and multi-omics characterisations. **c** Data sources for the treatment features contain compound annotations in the form of their chemistry or targets. **d** Measured functional outcomes are usually cell viability, molecular readouts or clinical endpoints. **e** Biological priors from knowledge databases can help to build or interpret models for predictive biomarker discovery.

1.5 Data sources for cancer research

Data-driven assessments of drug response biomarkers depend on the appropriate datasets from which they can be derived, hypothesised and validated. The growing wealth of biomedical data in oncology increasingly allows

cancer research to leverage ‘big data’ techniques, which enables the scalable modelling across numerous genes or pathways [201]. Thereby, it is essential for the resulting models to be able to narrow down or project back to biomarkers encompassing individual genes or pathways to achieve high interpretability and yield biologically plausible insights.

For this task, four types of biomedical data laid the foundation of this work and will be described in the following sections (Fig. 1.3a). For each data type, the selected datasets relevant to the scope of this work are outlined in Table 1.2. Firstly, omics data can describe tumours based on their genetic, transcriptomic or epigenetic profiles covering disease features \mathbf{X}_d (Section 1.5.1, Fig. 1.3b). Secondly, compound databases contain treatment annotations such as compound chemistry and drug targets covering treatment features \mathbf{X}_t (Section 1.5.3, Fig. 1.3c). Thirdly, functional data contains screens of observed outcomes Y complemented with molecular profiling in cancer samples or patients, which allows to discover patterns between outcomes Y , disease features \mathbf{X}_d and treatment features \mathbf{X}_t . These include data from perturbation screens using chemical compounds, genetic engineering or other cancer therapies with functional readouts (Section 1.5.2, Fig. 1.3d). Lastly, knowledge databases include annotations accumulated from our current understanding of cancer biology. While these databases are not primarily tailored to oncology, they contain a large proportion of biological priors for cancer applications to guide or interpret models in terms of disease features \mathbf{X}_d and treatment features \mathbf{X}_t for predictive biomarker discovery (Section 1.5.3, Fig. 1.3e).

1.5.1 Omics data repositories

The first big consortium to attempt forming a comprehensive catalogue of cancer genomic alterations in human cancers was The Cancer Genome Atlas (TCGA), founded in 2006 as a pilot study in three prevalent cancer types. Shortly after, in 2008, the International Cancer Genome Consortium (ICGC) was launched to coordinate collaborative efforts toward this goal and jointly expanded efforts towards over 30 cancer types. These efforts elucidated the mutational landscape of 12 cancer types [202] by exome sequencing in human tumour samples with matched normal tissue. Today, the ICGC database contains over 20,000 samples with molecular profiles, including over 10,000 samples from TCGA [201]. Their resource and findings have advanced the current understanding of the cancer genome [203], which promoted further efforts into the investigations in whole genomes by the joint consortium Pan-Cancer Analysis of Whole Genomes (PCAWG) [43], which now encompasses over 2,600 tumour samples. Data from the TCGA is embedded within the Genomic Data Commons (GDC) data portal, including data from more than 20 other consortia and over 85,000 cancer samples [203]. Most cancer types only include under 1,000 samples each; however, this data is highly curated and standardised. As a result, more than 10,000 scientific articles have cited TCGA [201]. Further, molecular data repositories in public archives, such as the NCBI Genbank [204], the European Nucleotide Archive [205] and the Gene Expression Omnibus [206]), host large data collections of more than 1 million cancer samples [201], which are, however, hard to analyse jointly because of the lack of standardisation and integration.

While the primary focus of the ICGC and TCGA has been cancer genomics, the TCGA has expanded to profiling complementary molecular data such as epigenetics and transcriptomics in order to facilitate the integrative analysis of cancer beyond its genetic component. Namely, today more than 8,800 primary tumours have complete molecular data, i.e. whole exome sequencing, RNA-seq, and Illumina human 450k methylation array profiling. This data can be used to yield additional evidence for cancer mechanisms when matching it with molecular data of cancer cell lines that provide an expanded pharmacological view, which was carried out in Section 2.1.

Resource	Type	Data description	Number	Usage	Source
TCGA*/ PCAWG/ TARGET	omics, (functional)	whole genomes, exomes, gene expression, DNA methylation	10,000 tumours	discovery of can- cer mutations and multi-omic cancer mechanisms	[43, 202]
NCI-60	functional	whole exomes, gene expres- sion, DNA methylation, pro- teomics, metabolomics	60 cancer cell lines with 50,000 screened compounds	discovery of pre- dictive drug effi- cacy biomarkers	[207]
GDSC*/ CCLE*/ CTRP*	functional	whole genomes, exomes, gene expression, DNA methylation, proteomics, metabolomics	1,000 can- cer cell lines with 500 screened compounds	discovery of pre- dictive drug effi- cacy biomarkers	[132, 208, 209]
Score*	functional	whole genomes, exomes, gene expression, DNA methylation, proteomics, metabolomics	900 can- cer cell lines with CRISPR ge- netic screens	discovery of gene dependencies and synthetically lethal targets	[115]
LINCS*	functional	gene expression	70 cancer cell lines with 25,000 compounds	discovery of drug targets and mecha- nisms of action	[210]
FIRE-3*/ ADJU- VANT*	functional	targeted somatic mutations, gene expression	randomised controlled clinical trials with 300 tumours	explorative sub- group analysis for discovery of pre- dictive biomarkers	[211, 212]
ChEMBL/ PubChem/ DrugBank*	knowledge	biochemical assays, SMILES structures, drug targets	1,000,000 compounds	annotation for compounds and their targets	[213, 214, 215, 216]
KEGG/ GO/ OmniPath*	knowledge	pathways, gene ontologies, protein-protein interaction networks	-	annotation for molecular mecha- nisms	[217, 218, 219]
COSMIC*	knowledge	curated somatic mutations, non-coding mutations, gene fusions, genome rearrange- ments, copy number alter- ations, aberrant expression, DNA methylation	-	annotation for discovered cancer driver mutations and mechanisms	[220]
GTEx/ ENCODE*/ Roadmap	knowledge, (omics)	curated (expression) trait loci, tissue gene expression, DNA regulatory elements, DNA configurations, DNA and histone methylation	-	annotation for molecular mecha- nisms	[104, 221, 222, 223]

Table 1.2: Selected data sources for cancer research. The resource names, types classified by Fig. 1.3b-e, descriptions of provided data, approximate number of contained tumours or compounds, their usage in the scope of this work and their citations are shown. The indicated resources with the asterisks are used within this work for the purpose of predictive biomarker discovery. The other highlighted data sources can be used analogously or complementary. The dashes on the number of tumours or compounds indicate the focus on their provided metadata on biological processes.

1.5.2 Functional data

While molecular profiling in cancer can yield insights into its disease aetiology, functional screens can reveal biological mechanisms for a phenotype of interest to help identify potential therapeutic opportunities [207, 224]. Typically, this is achieved through a systematic perturbation inflicted on a cancer model system, such as compound HTS in cancer cell lines [207, 225].

In 1990, the NCI-60 human tumour cell lines HTS was the pioneering functional screen conducted in 59 immortalised human cancer cell lines [207, 226]. Immortalising and establishing a cancer cell line requires tumour material of a human cancer and the repeated culturing of a stable population of cells [227]. Initially, concerns were raised regarding their resemblance to the original tumour because of culturing artefacts such as the accumulation of passenger mutations or the missing microenvironment and immune component. However, they are shown to recapitulate meaningful tumour biology [228] and, therefore, serve as a suitable model system for pharmacogenomic studies with high-throughput experiments. For example, by evaluating the NCI-60 monotherapy drug screens, it was found that BRAF^{V600E} mutations in melanoma cell lines confer response to the MEK inhibitors hypohemycin and CI-1040 [229].

Other HTS efforts have since expanded on this concept. While many included cancer types in the NCI-60 only screened about 60 samples, the Genomics of Drug Sensitivity (GDSC) and Cancer Cell Line Encyclopedia (CCLE) / Cancer Therapeutic Response Portal (CTRP) project both expanded beyond 1,000 cancer cell lines from over 30 cancer types screened across approximately 500 compounds [132, 208, 230]. In contrast to the NCI-60, for which advanced sequencing technologies were only performed later, these screening efforts are complemented with multi-modal molecular characterisations, including genomics, epigenomics, transcriptomics, proteomics and metabolomics [132, 209, 231, 232, 233]. While these two screens are the main focus of this work, other monotherapy HTS with lower numbers of cell lines are available and have been integrated into public databases [234].

These datasets enabled pharmacogenomic assessments and a wide range of drug response prediction models, which are discussed in Section 1.6.2. They are phenotypic cell-based screens designed to assess drug responses by measuring biological activity upon drug perturbation, such as cell viability, gene expression or pathway modulation, without requiring prior knowledge about MOA compared to other assays [235]. Commonly used methods are assays quantifying the metabolic activity of ATP through luminescence as a proxy for cell viability [236], such as the CellTiter-Glo assay for the HTS performed by the GDSC [234]. Roughly, relative cell viability can be calculated from these assays by dividing intensities from drug-treated cell cultures by untreated controls for different drug doses to arrive at a dose-response curve [237]. Finally, summary metrics for these curves can be derived through curve-fitting sigmoid functions [237, 238], Gaussian processes [239] or hierarchical Bayesian models [240]. These metrics include the drug concentration at which cells experience a 50% decrease in viability, i.e. the half maximal inhibitory concentration (IC_{50}), and the area under the dose-response curve (AUC), which are popular outcomes Y for training *in silico* drug response prediction models. As opposed to the IC_{50} , the AUC metric depends on the used concentrations for the conducted experiments. However, many machine learning methods use the AUC metric because of its robustness [241].

Aside from drug treatments, other types of perturbations are possible. For example, CRISPR-Cas9 knockout screens measure the viability upon loss-of-function of the range of protein-coding genes, which can reveal gene dependencies and synthetically lethal targets [115, 242, 243]. Similarly, drug MOA can be revealed when comparing viabilities upon knockout of putative drug targets and drug treatments [244]. Other types of CRISPR (activation or inhibition) and RNA interference (RNAi) have been collected in numerous integrated databases [245, 246, 247]. Furthermore, other molecular readouts beyond cell viability are feasible. The NIH LINCS consortium is interested in human disease perturbations using various assays [248], for example, assessing transcriptional responses with gene expression profiling as readout in perturbed cancer cell lines [210]. Upon perturbation, up- or down-regulated genes are summarised into drug or gene signatures containing implicit information about the MOA or gene loss-of-function effects. Since transcriptional signatures upon chemical perturbations can contain information about potential drug MOAs and cancer response mechanisms, they were hypothesised to be able to ‘reverse’ transcriptional disease signatures, a concept called connectivity map (CMAP) [249]. Since then, many strategies have been proposed to exploit this concept for pharmacogenomics [250]. For example, a CMAP identified entinostat to

inhibit the maintenance of AML, which was subsequently validated *in vivo* [251].

All of the above screening efforts aim to identify treatment opportunities in human patients, which ultimately can only be validated with clinical observational data. Thus, clinical data are valuable resources for mining functional relationships between tumours and their response to therapy. Unfortunately, TCGA data contains only sparse or no information about administered treatments and outcomes for many of the profiled patient primary tumours [252], thus, drug response patterns can only be assessed in a limited number of tumours and compounds. Therefore, its ability to yield predictive drug response biomarkers is also limited. Additionally, observational data such as available in the TCGA can show selection biases and confounding that hamper drawing reliable conclusions [253]. Thus, randomised controlled clinical trials are preferred for outlining predictive and prognostic biomarkers by using different types of clinical endpoints that quantify favourable therapy outcomes Y . For instance, the ‘response evaluation criteria in solid tumours’ (RECIST) criteria [254] can be summarised into an objective response rate, which is a binary variable in order to quantify the response to therapy. Other types of clinical endpoints are censored outcomes, for which statistical details are given in Section 1.4.2 and include the time from diagnosis or treatment initiation until death (overall survival), disease progression (progression-free survival) or relapse (disease-free survival). This work uses data from the two clinical trials FIRE-3 [211] and ADJUVANT [212]. FIRE-3 included metastatic COREAD patients treated with either cetuximab or bevacizumab in combination with 5-fluorouracil, leucovorin and irinotecan (FOLFIRI) [211], whereas ADJUVANT included NSCLC patients treated with either vinorelbine plus cisplatin or gefitinib [212].

1.5.3 Databases for chemical compounds, drug targets, biological processes and cancer genomics

Compound and knowledge databases introduced in this section are often used for systems pharmacology approaches, however, they only have limited information on drug variabilities between tumours [255]. Nonetheless, these databases can help to contextualise the variability among individual samples or patients. To achieve this, they can be used either as priors to guide and constrain the data-driven modelling for reducing bias or in a *post hoc* analysis to accumulate evidence in terms of the biological interpretation of predictive biomarkers. To exemplify this, consider a study that investigates many different compounds in parallel. Mining annotations in the chemical compound databases to cover drug features \mathbf{X}_t can improve drug response prediction models by integrating them as features [256]. Conversely, drug response biomarkers can also be derived for each compound without considering drug features \mathbf{X}_t . Then, a subsequent assessment can reveal biomarkers that are consistently observed for all compounds with the same drug target, which is demonstrated in Section 2.2.

Compound annotations can be obtained from chemical databases such as ChEMBL [213], PubChem [214] or DrugBank [215], which include diverse pharmacological information on millions of compounds. For example, DrugBank is used to manually queue drug targets for compounds with putative drug response biomarkers in Section 2.1. Furthermore, refined drug annotations may be derived from biochemical assays for drug action or activity that encode information on drug MOAs. Additional compound annotations can be obtained through the simplified molecular-input line-entry system (SMILES) compound structures, which can be used to extract pharmacological information with tools such as RDkit [257] or molecular representation learning [258].

Another type of knowledge database contains biological processes associated with drug targets, which are often understood in terms of molecular cancer pathways. An example for such a database is the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [217], which also includes drug-target interactions additionally to annotated networks of molecular interactions. A similar concept is followed by the Gene Ontology (GO) [218], which provides hierarchically structured relationships between genes to gain functional insights into gene sets. For example, in Section 2.2, annotated GO terms are used to interpret drug MOAs from dysregulated gene sets obtained from the LINCS database. In some efforts, such as the OmniPath database curated from over 100 sources [219], signalling pathways are characterised by molecular interaction graphs that outline protein-protein interactions or gene regulatory networks. By mapping putative molecular drug targets and predictive biomarkers to OmniPath nodes and employing integer linear programming [259] or shortest paths [260] in Section 2.1, dysregulated molecular pathways can be contextualised. Furthermore, in Section 2.3, curated pathway networks are used to prioritise

mutually exclusive somatic alterations that are functionally related [261].

Complementary efforts such as the Human Genome Project [262] majorly powered knowledge databases for cancer genomics by providing a reference genome. For example, the Catalogue of Somatic Mutations in Cancer (COSMIC) [220] assembled curated somatic cancer mutations derived from cancer genomes contained in the omics data repositories outlined in Section 1.5.1, which constitute the current Cancer Gene Census [25]. Moreover, resources such as the Genotype Tissue Expression project (GTEx) [221], the Encyclopedia of DNA Elements (ENCODE) [222] or the NIH Roadmap Epigenomics Consortium [223] have provided functional annotations for the regulatory landscape in normal human epigenomes and transcriptomes that can be used to study aberrant gene regulations in cancer. Somatic alterations with functional relevance are often included in TS platforms analysed in Section 2.3. Furthermore, many somatic alterations that were used as mutational background of cancer cell lines for the models in Section 2.2 are annotated somatic driver mutations in the COSMIC database [132].

1.6 Methods for identifying predictive biomarkers

After the introduction of statistical and machine learning concepts along with the required cancer biology and its data sources, here, related works which proposed statistical and machine learning methods that can be utilised to discover predictive biomarkers in oncology are introduced, which served as the foundation for the modelling strategies chosen in this work. The first subsection discusses methods from genome-wide association studies (GWAS) and differential gene expression (DGE), which are both traditional examples of feature-wise linear modelling. Next, methods for drug efficacy prediction will be shown, for which a wide range of predictive modelling strategies using machine learning have been proposed before. Finally, previously proposed methods for subgroup analysis in clinical data are discussed, for which a comparably wide range of statistical and machine learning methods have been previously proposed.

1.6.1 Methods for (epi)genome-wide association studies and differential gene expression

For the systematic discovery of predictive biomarkers in cancer cell lines, associative pharmacogenomic studies using the linear models in equation 1.24 have primarily focused on somatic mutations [225, 263]. A common approach is to call likely somatic variants and copy number alterations from NGS data [132, 209] as a filtering step before the downstream analysis because of their causal component in the cancer disease aetiology. These variant calls are also used for the mutational background of cancer cell lines in Section 2.2. It is reported that consensus single-nucleotide variant (SNV) calls among different variant-calling software tools can achieve a sensitivity and precision of approximately 95%, respectively [43]. While the technical aspects of calling these variants are out-of-scope for this work⁴, it is noteworthy that typically tumours contain a few hundred coding SNVs [265].

In contrast, GWAS is a widely used methodology for identifying genetic variants associated with complex traits across the whole genome in an unbiased manner. It assesses many genetic variants across the genome that are tested for their association with the trait typically by using *t*-tests, analysis of variance (ANOVA) or linear models for each single nucleotide polymorphism (SNP) such as equation 1.24. The resulting summary statistics per SNP are then corrected for multiple hypothesis testing with methods introduced in Section 1.3.3. Typically the significant associated SNPs with high effect sizes⁵ are the selected biomarkers as candidate quantitative trait loci (QTL) for further validation and replication of the found effects. Thereby, statistical fine-mapping of the set of correlated SNPs due to linkage disequilibrium can help detect the causal variant [266]. Popular software tools for these tasks include *PLINK* [267] and *GEMMA* [268], and best practices and recommendations have been laid out [269].

For example, a GWAS to investigate the role of germline genetic variants in cancer drug responses has revealed that inherited variants can undeniably contribute to drug susceptibility in cancer cell lines [270]. For this,

⁴Calling somatic variants is a task for which different tools and best practices emerged over the last decade [264].

⁵The effect size represents the magnitude of the investigated effect and can be measured in different ways depending on the use case. Mean differences, fold changes, Cohen's *d*, correlation coefficients, odds ratios, hazard ratios and other model coefficients are valid choices depending on the type of trait and study.

the employed linear mixed model adjusted for fixed effects in the form of tissue type, and the inherited population structure was accounted for as a random effect. Alternatively, adjusting for principal components can also correct for stratification in GWAS [271]. An accumulation of evidence across multi-modal data can further reveal biological mechanisms. For example, expression quantitative trait loci (eQTL) are SNPs associated with the expression of proximal or distal genes [272]. Accordingly, the aforementioned study suggested that *NQO1* expression determines responses to the HSP90 inhibitor tanespimycin, which is modulated by the germline QTL and affects its activity in cancer cell lines [270].

For associative assessments of phenotypic traits regarding gene expression across all protein-coding genes, DGE analysis is used to identify gene transcripts that are differentially expressed in multiple sets of samples with different conditions, e.g. case and control samples or different drug treatments. In DGE analysis, after preprocessing the gene expression data (removing technical variation and normalisation depending on the sequencing technology and methodology) and quality assessment, similar (generalised) linear models are used, which are optimised and incorporated in R packages, such as *limma* [273] or *DESeq2* [274]. Similar methods are also implemented in *limma* for DNA methylation to detect differentially methylated probes (DMP) and differentially methylated regions (DMR)⁶. Software packages such as *minfi* [276] for DNA methylation array data and *methylKit* [277] for RRBS data facilitate the processing steps and calling of DMPs. Since DNA methylation is a highly dynamic process, instead of focusing on single sites, calling an extended region of CpG sites with differential methylation is favourable. Calling these DMRs can be performed with software such as *Bumphunter* [278], *DMRcate* [279], *Probe Lasso* [280] and *comb-p* [281], which have been previously benchmarked against each other [282]. As discussed before, DNA methylation in regulatory elements can regulate the expression levels of transcriptional targets. These types of regulation can be subsequently identified by workflows such as the *ELMER* (Enhancer Linking by Methylation/Expression Relationships) R package [283, 284]. Rather few efforts have been directed towards revealing epigenetic mechanisms for determining drug responses, even though it is reported to contribute to drug susceptibility in HTSs in a tissue-specific setting [132]. Therefore, a differential methylation analysis of drug responses coupled with an integrative study of genomic and transcriptomic data across GDSC, CCLE and TCGA datasets presented in Section 2.1 could reveal epigenetic drug response mechanisms to advance the field of pharmacogenomics.

1.6.2 Methods for drug efficacy prediction

Models for drug response prediction are usually trained with data obtained from HTS experiments in cell culture to model the contrast function $z(\mathbf{X}_d, \mathbf{X}_t, \mathbf{t})$ ⁷ and test its predictions in independent screening experiments. The first systematic benchmarking study for this task was conducted by the NCI DREAM challenge. It included 35 breast cancer cell lines treated across 28 drugs to train the models, which could be evaluated on 18 cell lines [285]. This challenge delivered several takeaways. As part of the challenge, the identity of the compounds was not known, which discouraged the utilisation of drug features \mathbf{X}_t , such as drug targets and chemical features. At the same time, the state of a cancer cell line was given by disease features \mathbf{X}_d from diverse molecular profiling technologies. Since the number of features significantly outweighed the sample size, the top-performing methods reduced the number of model parameters by employing kernel regression methods to measure the similarity between cell lines for each data modality [285]. In addition, the top-performing methods preferably used nonlinear modelling and exploited biological priors from knowledge databases such as biological pathways [285]. This effort spawned a multitude of drug response prediction models trained with the datasets introduced in Section 1.5.2 and utilising diverse methodologies, which have been comprehensively reviewed and categorised in recent years [241, 286]. These diverse models included regression with and without kernels, Bayesian inference, matrix factorisation or deep learning [286].

Data modalities sometimes contain redundant information, which makes it challenging to efficiently integrate

⁶Sometimes this type of study is also called epigenome-wide association study (EWAS), the GWAS analogue for epigenetic datasets [275].

⁷Drug response prediction is often conducted by predicting summary metrics such as IC_{50} or AUC , which eliminates the prognostic component h in equation 1.22. However, in theory, HTS experiments could also be used to predict absolute viabilities that take h as prognostic term into account.

them. For example, a regression model in COREAD can predict the *RAS* mutational status from expression profiles [287]. Thus, efficiently integrating different data modalities is a central effort to increase model performances. A common approach is to aggregate all available features from a dataset (early integration); however, it was shown to be a better approach to build one model per modality and aggregate each predictor (late integration) [288]. However, it remains in question if early integration has not yet been possible because of its inability to regularise to the model space appropriately.

Generally, gene expression is often reported to be the most informative modality [132, 285, 286, 288]. However, this depends on the tissue and compound. For example, Iorio *et al.* (2016) found that genetic alterations are generally more predictive than gene expression in the tissue-specific context [132]. Moreover, Aben *et al.* (2016) found that this also depends on the compound. Namely, somatic mutations predicted sensitivity to MEK inhibitors, but the response to DNA damaging agents was mainly driven by gene expression [289].

Since the functional relationships of drug response in cell lines are likely nonlinear and complicated, today, the focus of development lies on deep learning algorithms [290], which have shown the ability to outperform traditional methods [291]. However, these advances have been hampered by the need for interpretability, causal reasoning and limited capability to propose predictive biomarkers with clinical utility [290]. As a result, the most actionable drug response prediction models focus increasingly on the translatability of their predictions to clinical applications or the interpretability of drug response mechanisms [292]. For the former, the first efforts have used linear models for imputing drug response for clinical samples and applied univariate statistical tests for assessing the individual biomarkers [293, 294]. Similarly, few-shot learning is applied to pretrained models in a few samples of the testing cohort to improve generalisation for the remaining samples [295]. For the latter, model-agnostic approaches have used feature importance scores, including regression coefficients [289] or other feature attribution techniques [163]. However, while these methods allow arbitrary baselearners, they do not directly propose plausible mechanisms and rely on downstream analyses of the extracted gene sets. Another effort has been directed at engineering ‘visible neural networks’ that reflect molecular mechanisms propagating from somatic mutations [164], but this method does not yet consider non-mutational mechanisms. Altogether, there is no all-purpose solution for drug response prediction for biomarker discovery yet, and in many cases, linear regression models in conjunction with rigorous validation suffice to yield promising predictive biomarkers. Thus, this work includes regression models with appropriate regularisation that are used in conjunction with feature ablation to benchmark the contribution of the EMT score to drug responses beyond the mutational background in HTS experiments as presented in Section 2.2.

1.6.3 Methods for subgroup analysis

Exploratory subgroup analysis for predictive biomarker discovery is typically conducted in clinical datasets. A range of methods have been proposed and discussed previously [166, 183, 296]. Since subgroup analysis attempts to estimate the predictive components of the outcome function $f(\mathbf{X}, t)$ in equation 1.27, its goal shares substantial similarity to drug response prediction and thus shares some of the same methodologies.

First, the methodology for Section 2.3 is presented. A binary treatment indicator $\mathbf{t} = t \in \{0, 1\}$ will be considered, referring to two treatment arms and a set of disease candidate biomarkers in \mathbf{X} . Then, following equation 1.27, the classical approach for subgroup analysis is chosen. It refers to the fitting of individual univariate regression models to predict clinical outcomes with treatment interactions for each x_1, \dots, x_p representing somatic mutations, which have been the primary focus of clinical applications because of their causal role in cancer. Since the baseline features \mathbf{X} are binary, the final predictive subgroup $A(\mathbf{X})$ is then directly defined over the contrasts. This approach does not account for complex interactions between features in \mathbf{X} , and thus is prone to model misspecification. However, if predefined subtypes provide a first level of stratification which is of central interest or part of clinical standard practice, this approach can help to evaluate the predictive potential of a second layer of stratification.

Furthermore, many somatic variants are rare with relatively low mutational frequency, whereas only a few recurring SNVs have high mutational frequencies [43]. Since mutual exclusivity often occurs in cancer driver mutations due to selection pressure, it can help to group mutually exclusive somatic mutations to ‘gene modules’. This introduces interactions between features \mathbf{X} and can (i) increase the statistical power to detect low-frequency

pharmacogenomic variants and (ii) improve the interpretability when defining gene modules based on biological priors. For calling mutually exclusive features in \mathbf{X} , first efforts have used pairwise hypergeometric tests [297] or other analytic significance tests [298]. These purely statistical approaches can be complemented by data from protein-protein interaction networks introduced in Section 1.5.3, which impose a prior that reduces the search space to mutated genes in the same biological process, as implemented in the Mutual Exclusivity Modules in cancer (MEMo) [298] or Mutex algorithm [261]. The latter uses an iterative greedy one-sided hypergeometric test to gradually evaluate putative gene modules. It is reported to trade recall for precision found by a benchmarking study of the competing CoMEt method [299], which is a desirable property for the purpose of grouping putative biomarker candidates prior to downstream statistical modelling and therefore was used in Section 2.3.

This chosen methodology is compared with alternative methods in Section 2.3. For example, the regression framework in equation 1.27 can be extended to regularised regression models that globally estimate the outcome function $f(\mathbf{X}, t)$ across treatment arms. When using this linear model, predictive and prognostic terms are not distinguished in the penalty term. Since predictive contributions are usually weaker, the FindIt method [300] employs separate lasso penalties in their support vector machine classifier for predictive and prognostic components, from which predictive components are extracted from the regression coefficients. As an alternative to examining regression coefficients, the Virtual Twin method [190] fits a global outcome model using random forests and computes the counterfactual outcomes for each subject as a first step. In the second step, it fits a separate regression tree on the hypothetical counterfactual treatment effect difference to extract predictive biomarkers.

Instead of estimating the outcome function $f(\mathbf{X}, t)$, one can resort to directly estimating treatment contrasts $z(\mathbf{X}, t)$. For this task, tree-based methods have been popular because of their inherent interpretability. Instead of using the observed outcome in the splitting criterion for the growing regression trees, the criterium includes treatment interactions for each possible split, so-called interaction trees [301, 302]. This methodology stratifies subjects recursively into their leaf nodes with similar treatment effects, thus providing a piecewise constant treatment effect estimation. This methodology has been incorporated into a range of methods. For example, the GUIDE method [303] and model-based partitioning (MOB) [304] reduce selection bias of selecting features with many possible splits. The SIDES method [305] includes a splitting criterion that selects the split according to the maximal differential effect between candidate subgroups.

An alternative formulation of the problem is the search for optimal treatment regimens to arrive at an optimal treatment policy $d(\mathbf{X})$, which maps each subject to an available treatment regimen. Maximising expected values of potential outcomes under this policy $\tilde{Y}(d(\mathbf{X}))$ yields an optimal policy [166, 306], i.e. an individualised treatment rule for optimal treatment selection. This problem reduces to a weighted classification problem with the treatment assignment as predicted variable weighted by the outcome (‘outcome weighting’) [307].

More recently, machine learning methods for estimating causal effects introduced in Section 1.4.5 started to be used for subgroup analysis. Naturally, these methods provide a subject-level estimate of the treatment effects $z(\mathbf{X})$ using causal forests [200], double machine learning [198] or metalearners [182]. In contrast to the traditional methods described above, these methods are not specifically designed to identify interpretable subgroups and biomarkers and often do not support censored outcomes. For example, a T-learner using treatment-balanced deep neural networks as baselearner with qualitative *post hoc* assessments was used to identify potential biomarkers [308]. Indeed, metalearners and recent advances from double machine learning for causal effect estimation are becoming as accessible as traditional machine learning methods due to the range of software projects that actively support their implementations, such as DoWhy [309] or econML [310]. These methods are primarily used for clinical trials or observational data. Specifically, causal forests were used in Section 2.3. Additionally, this work assesses the utility of double machine learning for *in vitro* HTS experiments in Section 2.2. For this, the EMT state of cancer cell lines is considered as a continuous ‘treatment feature’ t . Then, the presented methods are used to give causal estimates of the EMT effect on drug responses (outcome Y) in the presence of confounders in the form of a mutational background \mathbf{X} .

Beyond the introduced ideas, hybrid methods are possible. For example, the PRISM method [311] first reduces the subgroup search space through regularised regression that filters important features for predicting the outcome Y . Then, it estimates subject-level treatment effects using a T-learner with random forests as baselearner, and uses the individualised effects to grow trees in which the tree leaves correspond to the proposed subgroups with their

estimated treatment effects in RCTs.

The performance of each method is highly dependent on the studied dataset, the functional forms of the treatment contrasts and the intended use case [166, 183, 296]. By interpreting produced trees or regression coefficients, these methods can be used to extract putative predictive biomarkers. Most of the presented methods have software implementation that can be used for censored outcomes, which are applied in Section 2.3 to benchmark their ability to discover predictive biomarkers.

1.7 Regulatory considerations

From all compounds in oncology that enter phase 1 clinical trials, only about 2.1% will receive the final approval for clinical use by regulatory agencies [312]. Strikingly, compounds that enter clinical development with an associated biomarker for patient stratification show improved approval rates at around 10.7% [312]. However, these rates assume that a biomarker is known, and unfortunately, compounds for which novel biomarkers are investigated show similarly low approval rates as compounds without biomarkers [312]. Thus, the European Medicines Agency (EMA) specifically highlighted their interest in developing biomarkers for precision medicine using omics technologies with early engagement with biomarker developers [313].

The discovery of predictive biomarkers in the preclinical setting using drug response prediction benefits from rich datasets, but its utility is ultimately limited by their ability to transfer conclusions to the clinical setting [7]. For example, in the literature, studies assessing the clinical evaluation of a candidate biomarker found from the drug response prediction models in Section 1.6.2 can often only compare biomarker-positive and biomarker-negative subpopulations in clinical observational datasets [293]. While this can yield additional evidence of their clinical utility, due to the lack of their randomised controlled designs, this effect could be attributed to prognostic effects, confounders or other biases.

The discovery of predictive biomarkers and treatment effects directly in clinical studies can be computationally ascertained with multiplicity adjustments and resampling methods presented in Section 2.3. Nonetheless, the approval of a proposed biomarker for treatment efficacy requires additional prospective studies for the assessment of the sensitivity, specificity, reproducibility, and clinical utility of the associated companion diagnostic tests [11]. Furthermore, the exploratory and retrospective nature of biomarker discovery methods and their *post hoc* evaluation in clinical studies raise statistical questions regarding their regulatory assessment, even if RCTs and validated molecular diagnostic tests are available.

Thereby, the distinction between confirmatory and exploratory testing strategies during the assessment of subgroups is crucial. Accordingly, subgroup analysis in confirmatory clinical studies should be conducted with the proposed stratification strategies according to the EMA guidelines for subgroup analysis [174], which promote assessments of treatment effects in well-defined subgroups for trial planning, analyses and inferences. Confirming proposed decision-making for the former scenario requires pre-planned subgroups and rigorous multiplicity adjustments to preserve the overall false positive rate [314]. In the exploratory scenario and an overall successful clinical trial, subgroup analysis can be conducted to test the consistency or heterogeneity of the treatment effects [174, 177]. In the special case of a formally failed clinical trial, additional validation for any *post hoc* subgroup analysis is required since after the primary null hypothesis cannot be rejected, in principle, no further confirmatory conclusions are reliable [315]. After all, “the best test of the validity of subgroup analyses is not significance but replication” (Peter M. Rothwell, 2005, p. 182) [188].

For example, cetuximab was initially approved in metastatic COREAD across all patients that express *EGFR* [316]. However, after evidence from several retrospective analyses [317], a prospective study found interactions with *KRAS* mutations [317], which resulted in a label change by the FDA upon this accumulating evidence and the biological plausibility that *KRAS* mutant tumours lack cetuximab benefits [318]. Moreover, a meta-analysis of tumour sidedness later revealed that right-sided metastatic COREAD patients do not respond well to cetuximab either [319]. Subsequently, this was adopted in the European Society for Medical Oncology (ESMO) treatment guidelines, which currently state the effectiveness of cetuximab for left-sided tumours, whereas suggesting bevacizumab for right-sided tumours due to the lack of cetuximab benefit [320].

1.8 Rethinking conventional drug discovery

The conventional drug development pipeline starts with the identification of a cancer target and hit compounds (hit series) with suitable characteristics to modulate target activity *in vitro* [321]. During the following steps, the hits are further optimised to arrive at a narrower set of compounds, for which *in vivo* assays will be carried out in order to select a candidate compound to be deployed in the clinical development phase [321]. Drug development may benefit from the assessment of predictive biomarkers across development stages. Pharmacogenomic approaches can be employed in both preclinical and clinical development stages, and this knowledge can be transferred and fed back across stages to drive the simultaneous development of a drug targeting specific responder patient subpopulations predicted to fall into this responder subgroup in order to maximise treatment efficacy. For example, the development of BRAF inhibitors for BRAF^{V600E} mutations benefited from the knowledge about the MAPK signalling pathway. Accordingly, the measurement of downstream protein abundance of MEK and ERK can be used for optimisation [28]. However, this knowledge is still incomplete. For example, observed toxicity in clinical studies promoted investigations that revealed interactions between BRAF and oncogenic RAS in melanoma, which demonstrated that BRAF inhibition can reactivate the MAPK pathway in RAS mutant melanomas [322, 323]. Furthermore, targeting BRAF^{V600E} mutations via vemurafenib in metastatic COREAD does not show any success, potentially due to the quick restoration of MAPK signalling via EGFR activation [324]. Thus, increased efforts in finding and understanding pharmacogenomic biomarkers are essential for further drug development efforts.

In vitro studies in cancer have pioneered precision medicine with the discovery of pharmacogenomic interactions, cancer vulnerabilities and appropriate target populations for clinical translation. Complementary, precision medicine in the clinical setting has been driven by subgroup analysis in order to prompt follow-up studies with *in vitro* assays and clinical trials for refined patient stratification. Coupled with the recent methodological developments, the growing amount of functional preclinical and clinical biomedical data accompanied by molecular profiling has allowed the advanced exploration of predictive biomarkers [7]. Thus, this work focuses on both the analysis of *in vitro* drug HTS experiments and clinical trials in oncology. Specifically, emerging cancer hallmarks derived from tumour transcriptomics or epigenomics profiling in the context of somatic alterations are employed to yield reproducible and transferable predictive biomarkers in order to facilitate the feedback between preclinical and clinical studies in oncology.

1.9 Aims of the thesis

Given the introduced scope of this work, the aims and objectives of the three sections in Chapter 2 are formulated here. They build upon the outlined methodologies for the discovery of predictive biomarkers in the data sources from Section 1.5. The introduction section of each of the three included articles or preprints in the sections of Chapter 2 assesses different datasets and aspects and thus contains a more encapsulated introduction and discussion to each objective, whereas here these aspects are summarised in more encompassing aims.

First, Section 2.1 focuses on DNA methylation as drug response biomarkers. Thereby, the designed analysis provides full resolution of CpG sites and integration of complementary molecular mechanisms and clinical datasets, which extends on previous efforts that did not pursue this holistic angle. The formulated hypothesis states that extended differentially methylated regions may be associated with proximal transcriptional target expression and somatic mutations in cancer cell lines and human tumours that may jointly determine drug responses. Since DNA methylation is difficult to interpret, a layer-wise data integration design with low model complexities maintains high interpretability and transferability, and in addition allows to retain intermediate results within the filtered hypotheses. This design yields sets of drugs with their target and associated putative biomarker, which can be contextualised by shortest paths on protein-protein interaction networks.

Secondly, in Section 2.2, the focus narrows on EMT since it is an attractive cancer target due to its dynamic and reversible nature. Since this study focuses on this isolated process, the focus of this study is the discovery and estimation of the quantitative effect of EMT on drug responses. Additionally, its strong signal in transcriptomic data in many cancer types suggests its association with genetic alteration and upstream TFs. Therefore, the employed

hypothesis for this study states that EMT may be causally involved in determining drug responses in cancer cell lines. To test this hypothesis, machine learning and causal inference methods are employed for the estimation of the predictive power and causal effects of EMT, while subsequent enrichment tests for TFs and biological processes attempt its mechanistic interpretation.

These two studies aimed at discovering and evaluating predictive biomarkers in preclinical drug HTS datasets. Instead of discovering drug response biomarkers in cancer cell lines, Section 2.3 employs the biomarker discovery efforts directly in clinical trials. Thus, for our analysed clinical trial in metastatic COREAD, it is hypothesised that the contribution of CMS as transcriptional tumour subtypes and low-frequency somatic mutations may enrich the current treatment guidelines, which currently propose decisions based on *RAS* mutations and tumour sidedness. Thereby, exploiting mutually exclusive somatic mutations and subtype-specific modelling may refine the target populations of cetuximab or bevacizumab. Finally, applying such a general framework to other independent clinical trials and benchmarking other methods for subgroup analysis evaluates the generalisability of this framework.

In summary, this thesis aims at advancing the discovery, evaluation and contextualisation of predictive biomarkers from emerging cancer hallmarks and associated non-mutational cancer mechanisms in the context of their genetic component to determine drug responses in preclinical and clinical studies, as well as the development of data-driven frameworks to achieve this.

Chapter 2

Results

This chapter contains the three research articles that constitute the main contribution of this work. Each article contains its specialised introduction, results and discussion. The articles in Sections 2.1 and 2.3 were peer-reviewed and published open-access in scientific journals. The article in Section 2.2 is a publicly available preprint prior to peer review.

2.1 The pharmacoepigenomic landscape of cancer cell lines reveals the epigenetic component of drug sensitivity in cancer, *Communications Biology* (2023)

This article was peer-reviewed and published open-access in *Communications Biology* [1] and is reproduced with permission from Springer Nature. It is publicly available at <https://doi.org/10.1038/s42003-023-05198-y>.



<https://doi.org/10.1038/s42003-023-05198-y>

OPEN

The pharmacoepigenomic landscape of cancer cell lines reveals the epigenetic component of drug sensitivity

Alexander Joschua Ohnmacht^{1,2}, Anantharamanan Rajamani^{3,4,5}, Göksu Avar^{1,2}, Ginte Kutkaite^{1,2}, Emanuel Gonçalves^{6,7}, Dieter Saur^{1,3,4,5} & Michael Patrick Menden^{1,2,8}✉

Aberrant DNA methylation accompanies genetic alterations during oncogenesis and tumour homeostasis and contributes to the transcriptional deregulation of key signalling pathways in cancer. Despite increasing efforts in DNA methylation profiling of cancer patients, there is still a lack of epigenetic biomarkers to predict treatment efficacy. To address this, we analyse 721 cancer cell lines across 22 cancer types treated with 453 anti-cancer compounds. We systematically detect the predictive component of DNA methylation in the context of transcriptional and mutational patterns, i.e., in total 19 DNA methylation biomarkers across 17 drugs and five cancer types. DNA methylation constitutes drug sensitivity biomarkers by mediating the expression of proximal genes, thereby enhancing biological signals across multi-omics data modalities. Our method reproduces anticipated associations, and in addition, we find that the *NEK9* promoter hypermethylation may confer sensitivity to the NEDD8-activating enzyme (NAE) inhibitor pevonedistat in melanoma through downregulation of *NEK9*. In summary, we envision that epigenomics will refine existing patient stratification, thus empowering the next generation of precision oncology.

¹Computational Health Center, Helmholtz Munich, 85764 Neuherberg, Germany. ²Department of Biology, Ludwig-Maximilians University Munich, 82152 Martinsried, Germany. ³Division of Translational Cancer Research, German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. ⁴Chair of Translational Cancer Research and Institute of Experimental Cancer Therapy, Klinikum rechts der Isar, School of Medicine, Technische Universität München, Ismaninger Str. 22, 81675 Munich, Germany. ⁵Center for Translational Cancer Research (TranslaTUM), School of Medicine, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany. ⁶Instituto Superior Técnico (IST), Universidade de Lisboa, 1049-001 Lisbon, Portugal. ⁷INESC-ID, 1000-029 Lisbon, Portugal. ⁸Department of Biochemistry and Pharmacology, University of Melbourne, Victoria, VIC 3010, Australia. ✉email: michael.menden@helmholtz-munich.de

Precision oncology adverts to stratifying patients based on tumour entities and their molecular profiles to enhance drug efficacy and reduce toxicity¹. The success rate of clinical trials without a molecular biomarker is estimated to be 1.6% and is increased to 10.7% when using an appropriate patient stratification². Accordingly, methods that identify biomarkers and thereby facilitate clinical translation are crucial for the rapid development of novel cancer treatments.

In human tumours, aberrant DNA methylation has been shown to deregulate oncogenic pathways³ and to contribute to the acquisition of drug resistance^{4,5}. For example, DNA methylation in promoter, enhancer and CpG island regions has revealed epigenetic mechanisms involved in the transcriptional activity of several key cancer genes^{3,6}. In particular, the downregulation of tumour suppressor genes by hypermethylation of CpG sites in gene promoters is a hallmark of many cancer types⁷. In contrast, the hypermethylation of CpG sites in gene bodies is often reported to be positively correlated with gene expression⁸.

Molecularly characterised cancer cell lines are a useful and scalable model system for drug discovery⁹. They have empowered large high-throughput drug screens (HTS)^{10–15}, which include cell line panels of >1000 cell lines and are aimed to characterise the biomarker landscape of cancer¹⁶. For example, skin cutaneous melanoma cell lines (SKCM) harbouring *BRAF* V600E mutations are vulnerable to *BRAF* kinase inhibitors, and furthermore, this in vitro observation generalises to in vivo models and melanoma patients¹⁷. Genetic alterations are the causally related disease aetiology of cancer. Thus, most molecular biomarker studies have focused on somatic mutations and copy number variations. However, despite the growing utility of epigenetic biomarkers in clinics and an increasing number of commercially available diagnostic tests involving DNA methylation¹⁸, prognostic and predictive epigenetic biomarkers are still sparse¹⁹.

Few efforts have been dedicated to identifying DNA methylation biomarkers of drug response. For example, DNA methylation has been used to identify the CpG island methylator phenotype (CIMP)²⁰. It has previously been suggested as a predictive biomarker²¹, however, its definition is still inconsistent²², challenging to mechanistically interpret and limited to a handful of cancer types^{20,23,24}. Furthermore, predictive DNA methylation biomarkers in HTS are commonly assessed by summarising CpG sites in promoters and CpG islands^{11,21}. For these summarised regions, machine learning models have been used to predict drug response^{25,26} of preselected genes involved in DNA methylation or demethylation²⁶. In summary, these methods either do not leverage the full epigenome on the CpG site resolution, build evidence in multi-omics data modalities across different datasets, or lack mechanistic interpretations.

In order to empower epigenetic response biomarkers, our objectives were: (1) Identify DNA methylation regions associated with drug response in HTS; (2) Integrate genetic, epigenetic and transcriptomic data modalities of cancer cell lines for increasing evidence and interpretability; (3) Verify these epigenetic regulations of gene expression in human primary tumours and thus enhancing clinical translatability; (4) Finally, map the epigenetically regulated genes onto protein-protein signalling networks, and link them to their respective drug targets, thereby obtaining interpretable, actionable and translatable mechanisms. Our systematic analysis of the pharmacoeigenomic landscape in HTS, accompanied by thorough filtering for layer-wise evidence, interpretability and translatability, may pave the way for epigenetic response biomarkers in cancer.

Results

For the discovery of DNA methylation biomarkers of drug response, we analysed methylation patterns of 721 cancer cell lines from 22 cancer types treated with 453 anti-cancer compounds. The data was derived from the Genomics of Drug Sensitivity in Cancer (GDSC; Fig. 1a) project¹¹, which has since expanded its set of screened compounds compared to the original publication^{27,28}. Drug responses of cancer cell lines were characterised by their area under the drug response curve (AUC; Fig. 1b), for which low AUC values convey high sensitivity to the respective compound.

We first systematically searched for methylation regions with differential drug response in cancer cell lines, i.e., drug differentially methylated regions (dDMRs) by adaptively grouping spatially correlated CpG sites contained in the Infinium HumanMethylation450 BeadChip array (Fig. 1c; Methods). Secondly, we filtered for dDMRs which may mediate proximal gene expression (Fig. 1d; Methods), which thereby increases evidence of functional epigenetic events impacting drug response (Fig. 1e). Subsequently, we filtered for concordantly observed epigenetic mechanisms in human primary tumour samples from The Cancer Genome Atlas (TCGA; Fig. 1f; Methods), which yielded a prioritisation list of tumour-generalisable dDMRs, (tgdDMRs). Lastly, we correlated tgdDMRs with somatic mutations in cancer genes (Fig. 1g) and used shortest path algorithms applied to protein-protein interaction networks (Fig. 1g, h; Methods) to derive relationships between drug targets and proximal tgdDMR genes encoding respective proteins to support tgdDMRs further. In total, we found 19 tgdDMRs, i.e., predictive epigenetic biomarkers of drug response.

Identification of epigenetic drug response biomarkers from high-throughput drug screens. Analysing the DNA methylation and gene expression profiles of cancer cell lines stemming from 22 cancer types highlighted that the variance within cancer types is lower compared to the variance between cancer types (Fig. 2a and Supplementary Fig. 1a). Hence, we stratified cell lines into cancer types for subsequent modelling. For each cancer type and screened compound, we employed linear models and called drug differentially methylated regions (dDMRs; Methods), i.e., regions for which the methylation in CpG sites associates with drug response quantified by AUC. In total, we identified 802 dDMRs for 186 drugs in 22 cancer types (dDMR calling, adj. $p < 10^{-6}$; Fig. 2b and Supplementary Fig. 1b). We observed a linear relationship between the amount of found dDMRs and the sample size of the investigated cancer type (Pearson's $r = 0.81$, $p = 5.1 \times 10^{-6}$, correlation test; Supplementary Fig. 1c).

The distribution of significant drugs across cancer types was heterogeneous, but we identified enrichments of drug classes between cancer types (one-sided hypergeometric test, FDR < 0.05; Supplementary Data 1): Drugs that target the ERK-MAPK signalling pathway (trametinib, PD0325901, ulixertinib, selumetinib, VX-11e and CI-1040) were enriched in colorectal cancer (COREAD, odds ratio = 6.3), drugs that target EGFR signalling (afatinib, sapitinib, AZD3759, erlotinib, gefitinib and pelitinib) were enriched in lung adenocarcinoma (LUAD, odds ratio = 15.0) and drugs that are involved in targeting mitosis (alisertib, vinblastine, vinorelbine, GSK1070916, epothilone B, docetaxel, ARRY-520, S-trityl-L-cysteine) were enriched in small-cell lung cancer (SCLC, odds ratio = 4.9).

The distribution of CpG site counts per dDMR had a median of seven sites per dDMR. Furthermore, 132/802 dDMRs comprised >10 CpG sites, whilst 147 dDMRs contained <5 sites (Supplementary Fig. 1d). dDMRs were enriched for DNase I hypersensitive sites (DHS, $p < 10^{-16}$, odds ratio = 3.32, one-sided

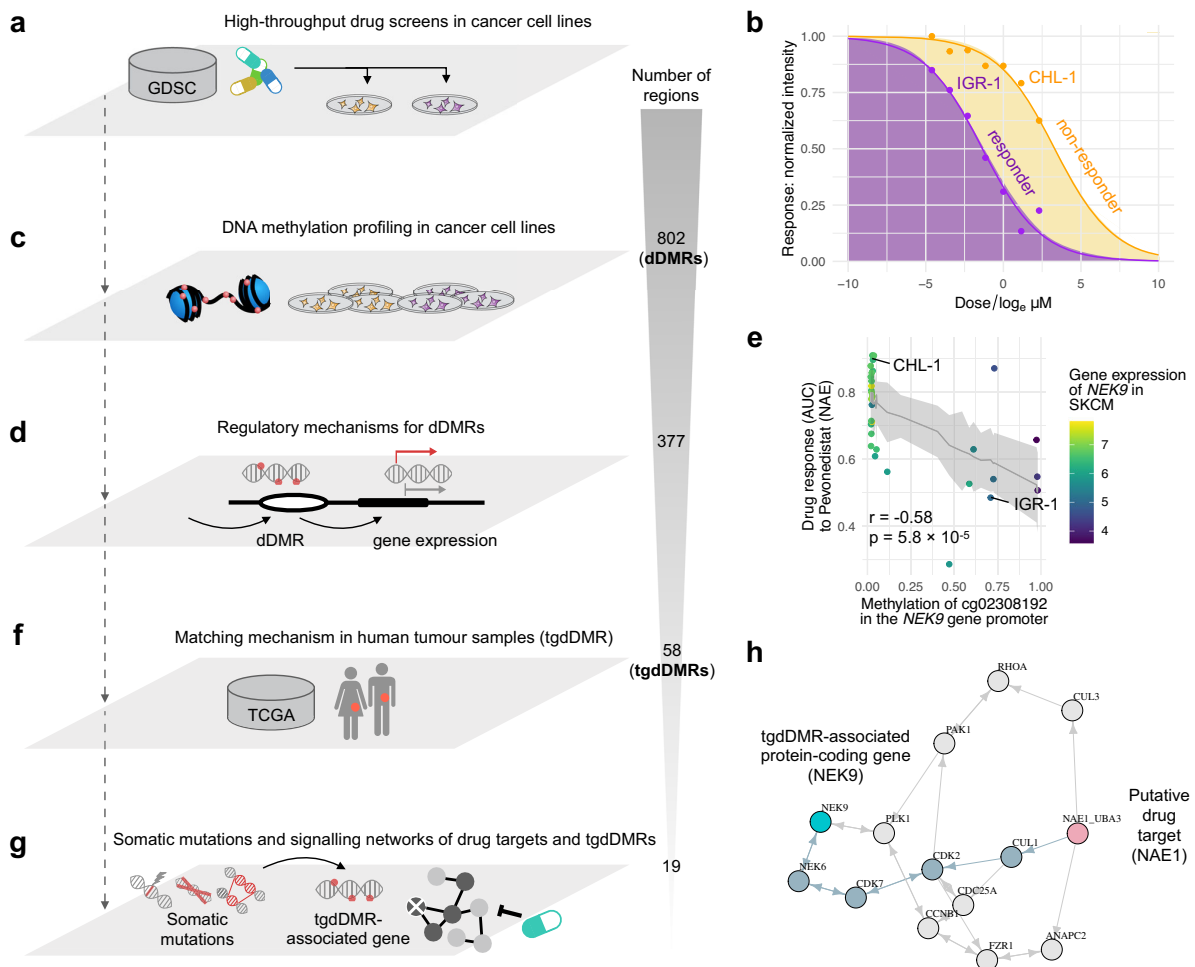


Fig. 1 Analysis workflow for the identification of epigenetic biomarkers of drug response. **a** The Genomics of Drug Sensitivity in Cancer (GDSC) project contains 721 cancer cell lines from 22 cancer types, which were epigenetically characterised and screened across 453 compounds. **b** The dose-response curves of a responder and non-responder melanoma cell line treated with pevonedistat. **c** We identified 802 drug differentially methylated regions (dDMRs). **d** The set of dDMRs is filtered for regulatory mechanisms, i.e., correlated gene expression of proximal genes, resulting in 377 functionally interpretable dDMRs. **e** For example, the dDMR in the *NEK9* promoter is associated with the expression of *NEK9* and is additionally correlated with drug response to pevonedistat. The error bars corresponding to 95% confidence intervals, the raw p -value (p) for the respective CpG site and the Pearson correlation coefficient (r) are displayed. **f** In total, the methylation of 58 epigenetic biomarkers of drug response were observed to be consistently correlated with the expression of their proximal gene in TCGA primary tumours. **g** The set of tgdDMRs was investigated for correlated somatic mutations in cancer cell lines. Additionally, for gaining further mechanistic insights, shortest-path algorithms traversed protein-protein signalling networks containing tgdDMR-associated genes as well as the respective drug targets and revealed additional evidence for 19 tgdDMRs. **h** The predictive biomarker *NEK9* (light blue) is connected within five steps to the drug target of pevonedistat, i.e., the NEDD8-activating enzyme NAE (pink). In the graph, nodes that are traversed with a shortest path are highlighted by the blue-grey colour among the alternative paths. The used human icons are from the AIGA symbol signs collection and are in the public domain.

Fisher's test; Fig. 2c, d) and sites in CpG islands ($p < 10^{-16}$, odds ratio = 3.13, one-sided Fisher's test; Fig. 2c, d). Furthermore, we investigated dDMRs in proximity of cancer genes based on annotations of the Network of Cancer Genes (NCG) project²⁹. DNA Methylation sites on the 450k microarrays have higher seeding density in the vicinity of cancer genes, i.e., 645/674 (96%) of cancer genes contained >10 profiled CpG sites compared to 16,213/20,557 (79%) of non-cancer genes. To alleviate this bias, we only tested genes with at least ten proximal CpG sites, which resulted in 16,858 background genes and 645/16,858 (3.8%) cancer genes. We observed 503 genes in proximity to identified dDMRs, of which 27 were cancer genes (5.4%; Supplementary Fig. 1e), thus cancer genes were significantly enriched ($p = 0.049$,

odds ratio = 1.44, one-sided Fisher's test). The most prevalent cancer genes were *APC* and *SKI* found across two cancer types. For reference, the most prevalent non-cancer genes were *PTPRN2* and *DDK1*, which were found in five and four cancer types, respectively (Supplementary Data 2).

Among the cancer genes associated with dDMRs, we found that *MGMT* dDMR methylation in low-grade glioma was associated with response to JQ1 (BET inhibitor, dDMR calling, adj. $p < 10^{-6}$; Supplementary Fig. 1f). The epigenetic silencing of *MGMT* is frequently debated as a clinical biomarker³⁰ and previous work revealed that JQ1 disturbs DNA damage responses by attenuating *MGMT* expression in glioblastoma cells³¹. While the different treatment responses are often attributed to somatic

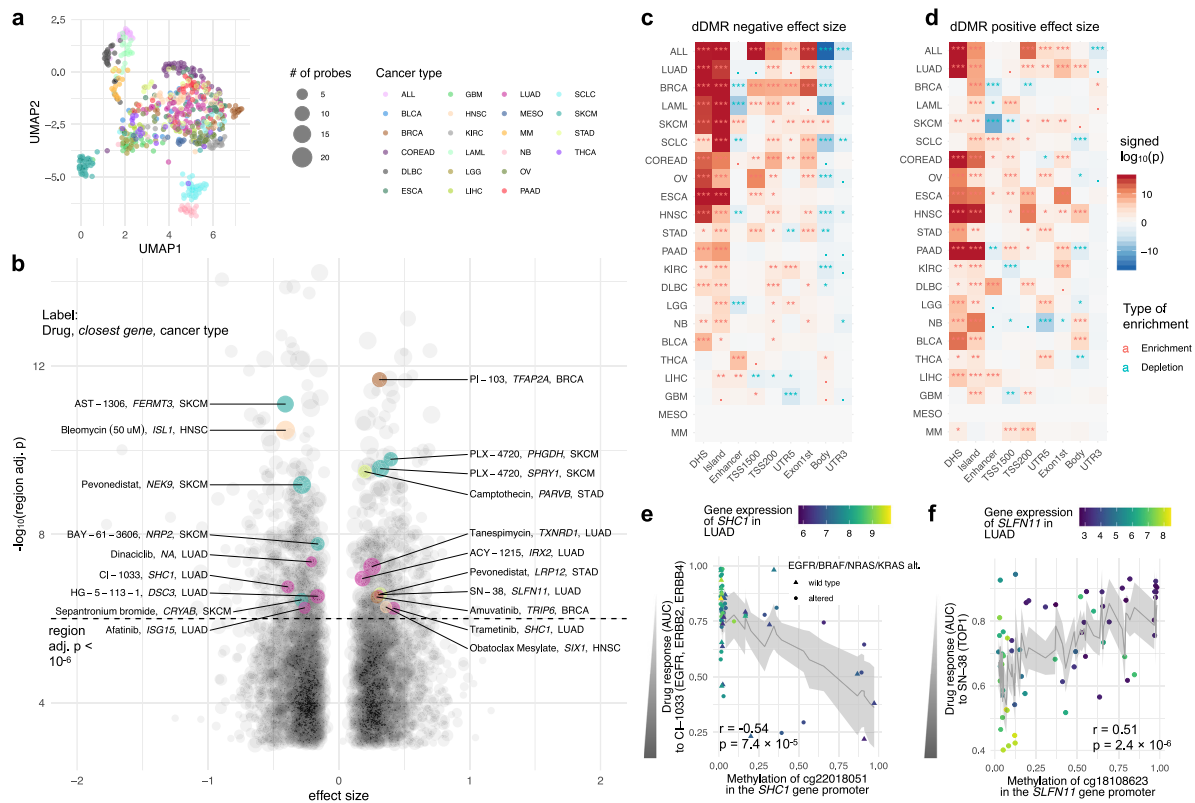


Fig. 2 Heterogeneity of epigenetic patterns across cancer types results in a rich resource of dDMRs. **a** Cancer type specific pattern of DNA methylation profiles of cancer cell lines in the GDSC. **b** Significant dDMRs across 22 cancer types and 186 drugs. The size of the data points indicates the amount of CpG sites in the identified dDMR. Genomic regions are labelled by the gene name in the closest vicinity. The enrichment of functional genomic regions in dDMRs is visualised in heatmaps for the scenario in which **c** hypermethylation confers drug sensitivity or **d** hypomethylation confers sensitivity. We tested enrichments for: genomic regions in the form of DNaseI hypersensitive sites (DHS), CpG sites within CpG islands, enhancer regions, regions within 200 and 1500 bases upstream of the transcriptional start site (TSS200 and TSS1500), the 5' untranslated region (UTR5), the 1st exon, gene body and 3' untranslated region (UTR3). **e** The association between *SHC1* promoter hypermethylation and CI-1033 response in LUAD; and **f** the association between *SLFN11* gene hypomethylation and response to SN-38. The error bars corresponding to 95% confidence intervals, the raw p-value (p) for the respective CpG site and the Pearson correlation coefficient (r) are displayed.

mutations in cancer genes, this suggests that DNA methylation can function as a complementary mechanism.

A negative effect size of a dDMR indicates that drug-sensitive cell lines are hypermethylated. Here, this is exemplified by the methylation status of *SHC1*, which was found to be associated with the EGFR, ERBB2 and ERBB4 inhibitor CI-1003 in LUAD (Fig. 2e). We observed that LUAD cell lines with a hypermethylated promoter of *SHC1* showed lower expression levels and were more sensitive to CI-1003 (Fig. 2e). Indeed, the adaptor protein SHC1 is involved in promoting the downstream signalling of ERK through EGFR³². No correlations between *SHC1* methylation and alterations in the ERK signalling pathway such as *EGFR*, *BRAF*, *NRAS* or *KRAS* mutations or amplifications were found. Clinical trials have reported benefits for non-small cell lung cancer patients with EGFR mutant tumours treated with the pan-ERBB inhibitor dacomitinib^{33,34}. Thus, *SHC1* silencing through DNA hypermethylation may be a sufficient but not necessary condition for sensitivity to ERBB inhibitors.

Overall, CpG sites in gene promoters were particularly enriched in dDMRs with a negative effect size ($p < 10^{-15}$, one-sided Fisher's test; Fig. 2c). For dDMRs with a negative effect size, methylation sites were usually hypomethylated across cancer cell lines, with a few treatment-sensitive cell lines that were hypermethylated (Supplementary Fig. 2).

In contrast to above, dDMRs with positive effect size contained methylated CpG sites that were mostly distributed across diverse genomic locations (Fig. 2d) and their hypomethylation was associated with drug sensitivity (Supplementary Fig. 2). Furthermore, we found enrichments of dDMRs with positive effect size within 200 bases upstream of the transcriptional start site (TSS200) for 11/22 cancer types ($p < 0.001$, one-sided Fisher's test; Fig. 2d). Exemplifying a dDMR with positive effect size, the hypomethylation of the *SLFN11* promoter was significantly associated with sensitivity to SN-38 in LUAD (Fig. 2f). The topoisomerase I inhibitor SN-38, the active metabolite of irinotecan, inhibits DNA replication through binding to the topoisomerase I-DNA complex and thus promotes DNA double-strand breaks. *SLFN11* is a putative DNA/RNA helicase that sensitises cancer cells to DNA damaging agents by killing cells with defective DNA repair³⁵. Its expression has been discussed extensively as a predictive biomarker for compounds targeting the DNA damage response^{36,37}. Here, we show that cells with hypomethylated *SLFN11* show high *SLFN11* expression and sensitivity to SN-38.

For validating dDMRs, we retrieved independent drug response data from the Cancer Therapeutics Response Portal (CTRP; Methods). We found that 236/802 dDMRs (29.4%) had overlapping data on cancer cell lines and drugs between GDSC

and CTRP. Among these, 193/236 (81.8%) had consistent effect size (Supplementary Fig. 3a), with an overall correlation of Pearson's $r = 0.46$ ($p = 9.7 \times 10^{-14}$, correlation test; Supplementary Fig. 3b). Furthermore, we validated our dDMRs with independent methylation data, i.e., reduced representation bisulfite sequencing for DNA methylation profiling (RRBS; Methods) extracted from the Cancer Cell Line Encyclopedia (CCLE). This only reduced the overlapping data of dDMRs slightly to 227/802 (28.3%), and 164/227 (72.2%) of these dDMRs displayed consistent effect size (Supplementary Fig. 3a), with a correlation of Pearson's $r = 0.43$ ($p = 1.2 \times 10^{-11}$, correlation test; Supplementary Fig. 3c), highlighting the ability of our method to yield reproducible results for independent drug screenings and DNA methylation experiments.

Epigenetic biomarkers interpreted through gene regulatory mechanisms. Hypermethylation of promoter regions is an established mechanism to reduce sufficient transcription factor binding and regulate gene expression accordingly³⁸. Thus, most methylation biomarker discovery efforts focus on gene promoter regions and neglect other regulatory mechanisms^{11,21,25,26}. For example, the deregulation of methylation patterns in gene bodies was also reported to alter gene expression profiles⁸. In order to address this, we generalised our working hypothesis and explored the DNA methylation of any dDMR that may mediate gene expression of proximal genes (Methods).

Upon systematic analysis with the Enhancer Linking by Methylation/Expression Relationships (ELMER) method³⁹, we observed that 377/802 dDMRs (47.0%) showed at least one significantly associated gene in the proximity of its genomic region (emp. adj. $p < 0.001$; Methods). In total, 576 genes were associated with these 377 dDMRs. For each gene associated with a dDMR, we independently correlated its expression and drug response with a linear model fit (Fig. 3a–d). In summary, we observed four distinct mechanisms which may drive drug sensitivity, i.e., hypermethylation with either downregulated gene expression (Case 1, $n = 216$; Fig. 3a) or upregulated gene expression (Case 2, $n = 110$; Fig. 3b), and hypomethylation with either upregulated gene expression (Case 3, $n = 162$; Fig. 3c) or downregulated gene expression (Case 4, $n = 88$; Fig. 3d). We exemplified each case in cancer cell lines and their mechanistic consistency in primary tumours (Fig. 3e–l).

For both Cases 1 and 2, hypermethylated dDMRs were associated with drug sensitivity (negative effect size in Fig. 2b). The majority of dDMRs belonged to Case 1, which was distinguished by promoter regions (Fig. 3a). It resembles the canonical mechanism in which hypermethylation of promoter regions downregulates the expression of their associated proximal gene and thereby confers drug sensitivity. This behaviour is exemplified by the methylation of the *SHCI* promoter and its gene expression in LUAD cell lines (Fig. 3e). Additionally, we verified the association of the epigenetic status and gene expression in LUAD human tumour samples (Fig. 3f).

For Case 2, hypermethylation of dDMRs correlated with higher expression of proximal genes (Fig. 3g, h). This is a less frequent epigenetic regulation mechanism, however, it is consistent with previous studies reporting both behaviours^{8,40–42}. As an example, the hypermethylation of the *OPLAH* dDMR was associated with the upregulation of *OPLAH* expression in SKCM cancer cell lines and HG-6-64-1 drug sensitivity (Fig. 3g). In addition, this epigenetic regulation of *OPLAH* expression was also demonstrated in primary tumour samples (Fig. 3h).

Cases 3 and 4 were characterised by hypomethylated dDMRs that were associated with drug sensitivity (positive effect size in Fig. 2b), which could also be distinct by negative or positive

correlations of dDMRs with gene expression for Case 3 and Case 4, respectively. For example, we found that the hypomethylation of the *SLFN11* dDMR in LUAD was associated with higher *SLFN11* expression (Fig. 3i), which was further verified in human tumour samples (Fig. 3j). In contrast, the hypomethylation of *PITX2* dDMR was linked to teniposide drug sensitivity, however, the hypermethylation of *PITX2* dDMR was positively associated with *PITX2* expression in cancer cell lines and human tumour samples (Fig. 3k, l).

In summary, drug sensitivity in cancer cell lines may be driven by either hypermethylation (Cases 1 and 2) or hypomethylation (Cases 3 and 4) of dDMRs and can either present negatively correlated gene expression (Cases 1 and 3) or positively correlated gene expression (Cases 2 and 4). Case 1 has been the focus of most epigenetic biomarker studies, whilst we systematically investigated all 4 cases (Supplementary Data 2) and therefore can provide broader mechanistic insights.

Epigenetic and transcriptional mechanisms in primary tumours increase evidence of drug response biomarkers. In the section above, we highlighted four distinct epigenetic mechanisms that may drive drug response, i.e., Case 1–4. Each of them was exemplified in cancer cell lines (Fig. 3e, g, i, k), and consecutively, further supported by concordant methylation and proximal gene expression patterns in tumours (Fig. 3f, h, j, l). Here, we systematically assessed all 377 short-listed dDMRs from above, to investigate concordant epigenetic regulation patterns in primary tumours leveraging ELMER³⁹ also in TCGA tumour samples⁴³ (Methods). In total, we investigated a subset of 241/377 dDMRs for which the associated cancer type data was available in TCGA. We observed that 58/241 (24.1%) of dDMRs showed a significant association with their proximal genes in tumours (ELMER, emp. adj. $p < 0.001$; Methods). We called this selection of epigenetic biomarkers tumour-generalisable dDMRs (tgdDMRs). For the final selection, we found 19/58 tgdDMRs for which the protein encoded by the associated gene was connected to the corresponding drug targets in the protein-protein signalling network OmniPath⁴⁴ (Methods). These 19 tgdDMRs (Supplementary Data 2) contained proposed biomarkers for 17 anti-cancer drugs across five cancer types (Fig. 4a), i.e., LUAD $n = 7$ (Supplementary Fig. 4), SKCM $n = 6$ (Supplementary Fig. 5), breast cancer (BRCA) $n = 2$ (Supplementary Fig. 6), head and neck cancer (HNSC) $n = 2$ (Supplementary Fig. 6), and stomach adenocarcinoma (STAD) $n = 2$ (Supplementary Fig. 6).

We found that the majority of tgdDMRs (15/19) were in promoter regions, which is concordant with previous computational strategies that focused solely on promoters to identify epigenetic response biomarkers. However, the remaining 4 tgdDMRs, which constitute >20% of our identified lead biomarkers, had distinctly different epigenetic regulation mechanisms, i.e., were located in either the gene body or distal regions (Fig. 4b). In addition, we found that all tgdDMRs had negative correlations with a proximal gene, which correspond to mechanism Case 1 or Case 3 (Fig. 3a, c). Furthermore, for 10/19 tgdDMRs the expression of proximal genes in cell lines itself was independently associated with drug response in cancer cell lines ($p < 0.05$, linear model fit; Methods), thus having a functional interpretation across two molecular layers.

For additional evidence of tgdDMRs, we again leveraged the CTRP and CCLE datasets as validation cohorts. For the tgdDMRs that had overlapping drug response data, we found that 7/9 tgdDMRs showed consistent effect sizes in the CTRP screen, with an increased correlation of Pearson's $r = 0.75$ ($p = 0.02$, correlation test; Fig. 4c) compared to unfiltered dDMRs in the previous section. Additionally, 5/7 of the tgdDMRs overlapping with the

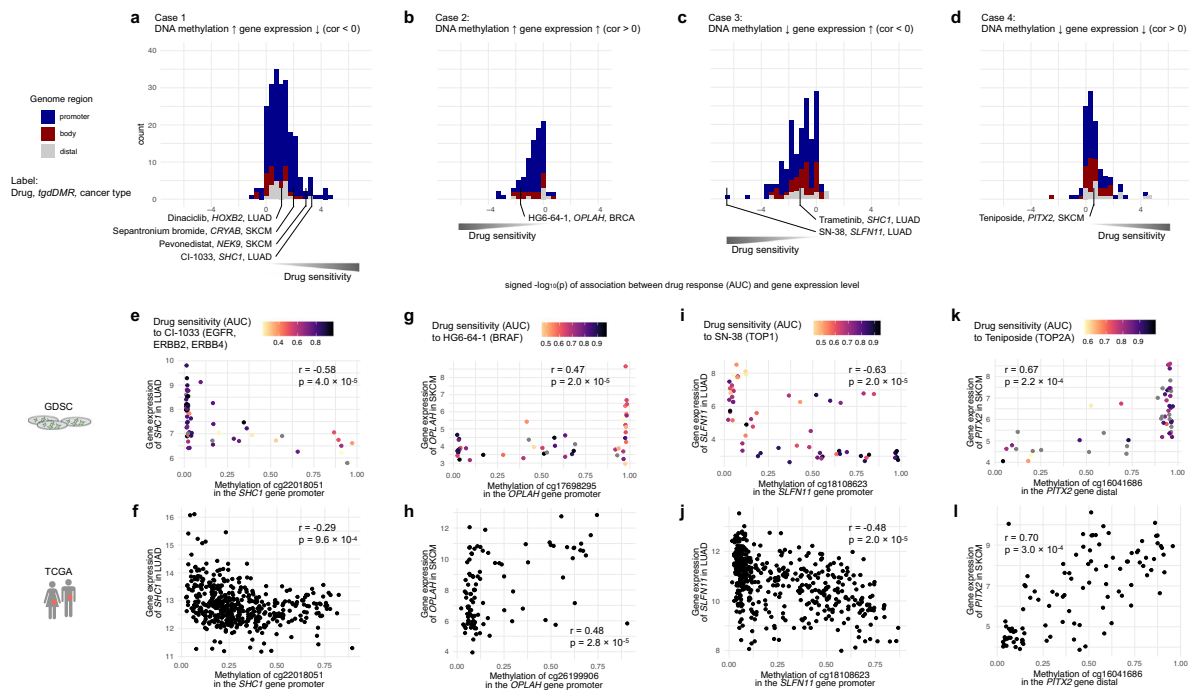


Fig. 3 Epigenetic drug response biomarkers are empowered by studying DNA methylation and gene expression patterns. This analysis revealed four distinct mechanisms observed across 377 dDMRs, i.e., Case 1-4: Cases 1 and 2 included dDMRs for which hypermethylation was associated with drug sensitivity and either **a** negative or **b** positive correlation with gene expression. For Cases 3 and 4 hypomethylated dDMRs were associated with drug sensitivity and either **c** negative or **d** positive correlation with gene expression. The x-axis shows the signed $-\log_{10}(p\text{-value})$ derived from a t -test of the coefficient of a linear model fit explaining drug response by proximal gene expression (Methods). Case 1 is exemplified by **e** the hypermethylation of the dDMR in the *SHC1* promoter regulating the expression in LUAD cancer cell lines, **f** which was validated in human tumour samples. In contrast, for Case 2 **g** hypermethylation in the *OPLAH* promoter promoted its expression in SKCM cell lines, and **h** tumour samples. For Case 3, **i** the hypermethylation of the *SLFN11* gene promoter downregulated the expression of *SFN11* in cancer cell lines, and **j** tumour samples. In Case 4, **k** positive correlations could be observed in the *PITX2* promoter and its expression in cell lines, and **l** tumour samples. The empirical adjusted p -value (p) for the respective CpG site and the Pearson correlation coefficient (r) are displayed. The used human icons are from the AIGA symbol signs collection and are in the public domain.

CCLE RRBS methylation data showed consistent effect sizes with an increased correlation of Pearson's $r = 0.85$ ($p = 0.01$, correlation test; Fig. 4c) compared to unfiltered dDMRs in the previous section. This highlights that reproducibility across independent drug screens and methylation datasets increased when focusing on tgdDMRs.

Currently, the majority of biomarkers for patient stratification are genetic alterations, thus, we investigated if genetic mutations and copy number alterations may reflect the methylation of tgdDMRs. We tested for associations between somatic mutations and tgdDMRs using linear models (Methods). We only observed weak correlations between somatic mutations and tgdDMRs (FDR < 0.1; Supplementary Fig. 7a; Methods).

While most tgdDMRs are found in gene promoters or bodies, we observed a distal region in a CpG island in the vicinity of the *HOXB2* gene that marked favourable drug responses for treatment with dinaciclib (CDK inhibitor), if the *HOXB2* tgdDMR was hypermethylated (dDMR calling, adj. $p < 10^{-6}$; Fig. 4d). Furthermore, the methylation status was correlated with *HOXB2* expression in cell lines (ELMER, emp. adj. $p < 0.001$; Fig. 4e) and primary tumours (ELMER, emp. adj. $p < 0.001$; Fig. 4f). Additionally, DNA repair enzyme encoding gene *APEX1* essentiality obtained from CRISPR knockout screens was significantly higher, if the tgdDMR was hypermethylated (FDR < 0.2; Supplementary Fig. 7d; Methods). HOX genes are a family of transcription factors that are frequently associated with cancer⁴⁵. Their expression is reported to be regulated by DNA

methylation⁴⁶, however, the mechanisms by which they affect responses to dinaciclib remain elusive. Notably, we were able to validate this association in the independent CTRP drug screen (Pearson's $r = -0.59$, $p = 0.02$, correlation test; Supplementary Fig. 7b) and additionally observed consistent trends with an alternative methylation profiling based on RRBS in the CCLE (Pearson's $r = -0.48$, $p = 0.10$, correlation test; Supplementary Fig. 7c).

Next, we highlight further associations included in the identified tgdDMRs. For instance, hypermethylation of the tgdDMR in the *NEK9* promoter conferred sensitivity to NAE inhibition with pevonedistat in cell lines (dDMR calling, adj. $p < 10^{-6}$; Fig. 4g). In particular, we observed that tumours with hypermethylated tgdDMR in the *NEK9* promoter showed low *NEK9* expression in both cell lines (ELMER, emp. adj. $p < 0.001$; Fig. 4h) and patient tumours (ELMER, emp. adj. $p < 0.001$; Fig. 4i). *NEK9* has been previously reported to participate in G1/S phase transition and progression and to regulate the kinase activity of CHK1 upon replication stress⁴⁷. Examining the neighbourhood of signalling networks, the inhibition of NAE by pevonedistat leads to the inactivation of cullin-RING ligases⁴⁸, which target key proteins during the cell cycle progression such as CDK2 and CDC25A (Fig. 1h)⁴⁹. This is supported by the Library of Integrated Network-Based Cellular Signatures (LINCS) database, which revealed the transcriptional dysregulation of *CUL3*, *CDC25A*, *CCNB1* and *PLK1* in SKCM cell lines upon treatment with pevonedistat (FDR < 0.1; Supplementary Fig. 7e;

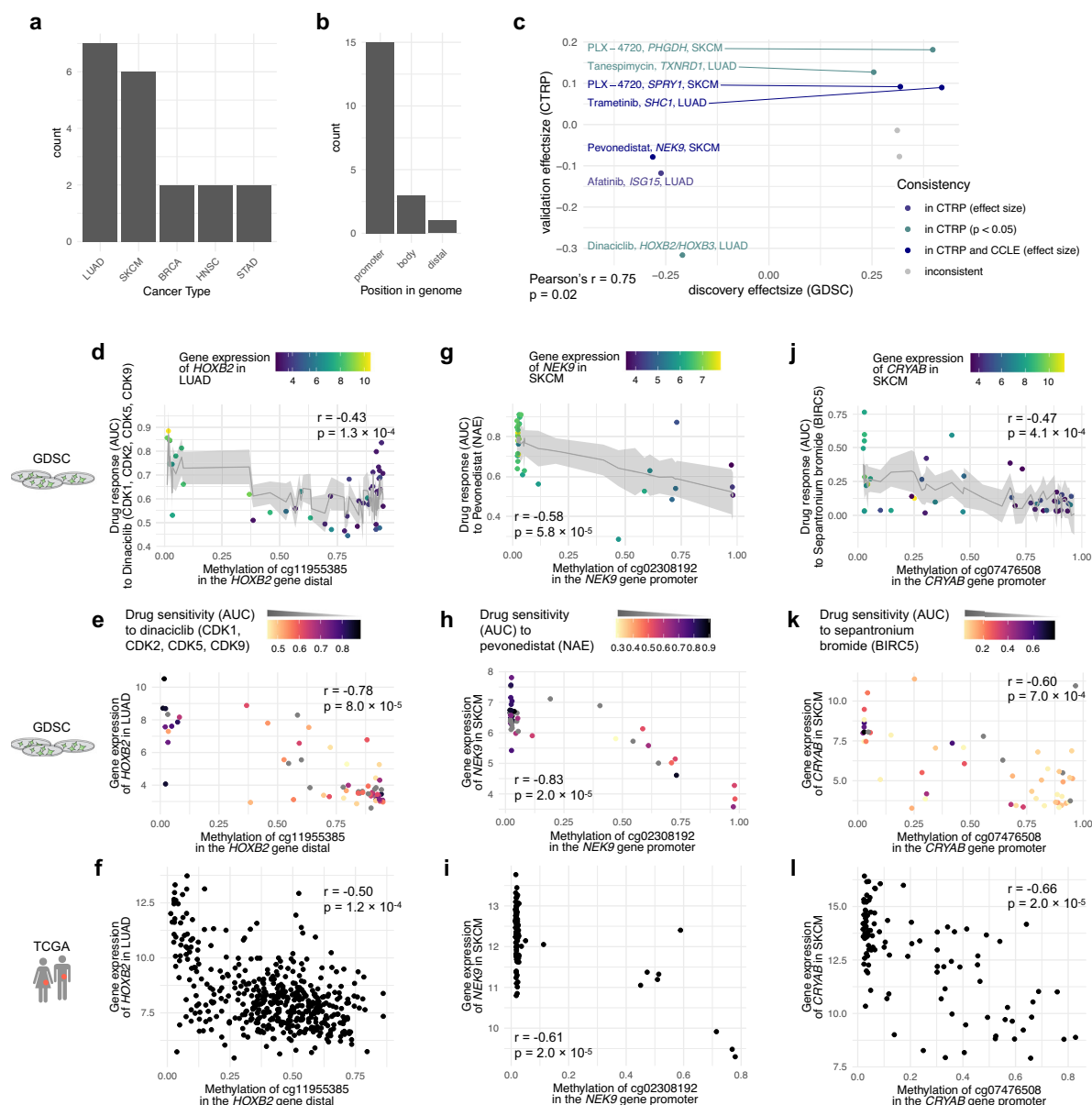


Fig. 4 tgdDMRs in the context of lung cancer and melanoma. **a** A histogram of tgdDMR in cancer types and **b** tgdDMR functional positions. **c** Scatter plot for validation of tgdDMRs, showing consistent effect sizes for CTRP and CCLE cohorts. DNA methylation of the distal dDMR in the vicinity of *HOXB2* is correlated with **d** response to dinaciclib and the expression of *HOXB2* in **e** cancer cell lines and **f** primary tumour samples. **g** Scatter plots show that the hypermethylated *NEK9* promoter confers sensitivity to pevonedistat in SKCM, the correlation between methylation in the *NEK9* promoter and its gene expression in SKCM in **h** cancer cell lines and **i** primary tumours. **j** Furthermore, scatter plots show that SKCM cell lines with hypomethylated *CRYAB* promoter do not respond to the apoptotic agent sepantronium bromide. Aberrant *CRYAB* expression with a hypermethylated promoter of *CRYAB* in **k** cell lines and **l** in tumour samples. For analysing DNA methylation and drug response, the error bars corresponding to 95% confidence intervals, the raw p -value (p) for each CpG site and the Pearson correlation coefficient (r) are reported. For analysing DNA methylation and gene expression, the empirical adjusted p -value (p) and the Pearson correlation coefficient (r) are reported. The used human icons are from the AIGA symbol signs collection and are in the public domain.

Methods). Concordantly, pevonedistat has been shown to induce DNA damage and cell cycle arrest^{50,51}, from which the cells with downregulated *NEK9* may not be able to recover.

A second tgdDMR in SKCM was identified, which involved a pro-apoptotic agent. Specifically, hypermethylation of the *CRYAB* promoter was associated with drug sensitivity to the BIRC5 inhibitor sepantronium bromide (dDMR calling, adj. $p < 10^{-6}$;

Fig. 4j) and aberrant *CRYAB* expression (ELMER, emp. adj. $p < 0.001$; Fig. 4k, l). Sepantronium bromide functions as a pro-apoptotic agent by inhibiting BIRC5, a member of the inhibitor of apoptosis (IAP) family⁵². The signalling network neighbourhood of the *CRYAB* tgdDMR shows interactions with CASP3 and P53 (Supplementary Fig. 5), which have been previously reported to show anti-apoptotic activity through *CRYAB*^{53,54}. This

observation suggests that activated CRYAB may protect from apoptosis induced by sepantronium bromide, however, the exact nature of this relationship remains elusive. Nevertheless, the signalling network neighbourhoods of tgdDMRs offer interpretable indications about putative drug response mechanisms associated with tgdDMRs.

Discussion

For advancing predictive epigenetic biomarkers in cancer, we presented an epigenome-wide multi-omic analysis for identifying interpretable and actionable epigenetic drug sensitivity biomarkers in HTS. In total, we identified 802 dDMRs demonstrating the epigenetic component of drug sensitivity in human cancer cell lines. Furthermore, we guided our method by the functional relationship that DNA methylation can mediate proximal gene expression, which resulted in a filtered set of 377 dDMRs that showed explainable regulation of transcriptional activity in human cancer cell lines. Furthermore, identifying consistency between cancer cell lines and primary tumours yielded evidence across epigenomic and transcriptomic data modalities and overcame limitations imposed by cell line artefacts⁵⁵. This step prioritised 58 tgdDMRs of which 19 were further supported by protein-protein interaction networks. This thorough filtering was necessary because direct evidence of epigenetic biomarkers is lacking and validation was only possible for a limited number of dDMRs.

We observed an enrichment of cancer genes in the proximity of dDMRs, however, many established cancer genes lacked dDMRs, which suggests that only a minority of cancer genes may be epigenetically regulated. Furthermore, the modest correlations with somatic mutations suggest that DNA methylation may function complementary to genetic alterations for determining cancer drug susceptibilities. In contrast, DNA methylation was often accompanied by transcriptomic changes; however, it was not able to substitute DNA methylation pattern of dDMRs, i.e., more than half of dDMRs did not reveal regulations of a proximal gene. This suggests that tgdDMR methylation may either assist cancer cells in rewiring key signalling pathways through altering transcriptional signals or accompany other more elusive epigenetic mechanisms. This notion advocates our study design that first focuses on differentially methylated regions and consecutive integration of genetic and transcriptomic data. The layer-wise filtering starting with DNA methylation allowed us to evaluate intermediate results on all separate analysis steps and provide a comprehensive resource of epigenetic biomarkers (Supplementary Data 2).

Within this study, we focused on cancer type specific dDMRs and observed strong epigenetically diverse patterns across cancer types. Since the amount of found dDMRs was directly related to the studied sample size, we anticipate that forthcoming large-scale screening efforts can increase the power to detect dDMRs focusing on tumour subtypes, e.g., in BRCA⁵⁶ or COREAD⁵⁷. Since DNA methylation can correlate with tumour subtypes, our analysis of dDMRs corrects for global methylation patterns through its principal components, which increases the ability to capture local mechanisms.

We showed consistency of tgdDMRs with an independent HTS and a different methylation profiling technology. Furthermore, we highlighted concordant epigenetic regulation of gene expression in human tumour samples, however, matched drug response readouts in human tumours are lacking. Nonetheless, our mechanisms may be validated in retrospective analyses of previously conducted molecularly characterised clinical trials for exploratory biomarker discovery. Although the signalling network neighbourhoods give insights into the potential mechanisms

for causal relationships or synthetically lethal interactions between drug targets and tgdDMRs-associated genes, tgdDMRs as predictive biomarkers remain to be further evaluated. In particular, melanoma patient subpopulations with promoter hypermethylation of tgdDMRs in the *NEK9* or *CRYAB* promoters could reveal benefits if treated with pevonedistat or pro-apoptotic agents such as sepantronium bromide, respectively.

We confirmed that DNA methylation in promoters is the major regulatory mechanism, and only sparse evidence supports mechanisms in gene bodies or distal regions. Thus, the role of methylation in cancer beyond its relevance in tumorigenesis and potential epigenetic vulnerabilities remains elusive. Upcoming technologies may enable the investigation of alternative epigenetic mechanisms in mediating drug responses beyond DNA methylation. For example, another class of epigenetic modifications, histone acetylation and histone methylation, are commonly associated with tumorigenesis and transcriptional regulations in cancer⁵⁸. Furthermore, sequencing technologies beyond the traditional epigenome, e.g., ATAC-seq chromatin accessibility and Hi-C chromosome conformation, can yield further regulatory insights.

In essence, epigenetic data has the potential to yield the next generation of predictive biomarkers for precision medicine. The results of our analysis show that DNA methylation complemented with multi-omic data integration can reveal interpretable biomarkers for expanding the limited number of epigenetic biomarkers in clinical use. Our analysis for pharmacogenomics can be applied to any drug screening effort with complementary multi-omics characterisation. Therefore, it may refine existing patient stratification and enhance the development of personalised cancer therapies in future.

Methods

Cancer cell lines and primary tumours. We leveraged cancer cell lines from the Genomics of Drug Sensitivity in Cancer (GDSC) project¹⁰ and the Cancer Cell Line Encyclopedia (CCLE) project¹² as discovery and validation cohort, respectively. Both databases have been extensively characterised and curated⁵⁹. The primary tumour samples are included in The Cancer Genome Atlas (TCGA), which aims to adhere to established guidelines and regulations regarding the use of human data⁶⁰. Ethics and policies regarding the TCGA study are available at <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/policies>. Additional demographic characteristics of TCGA are available under <https://portal.gdc.cancer.gov/> and have been reported previously⁶¹.

DNA methylation. The raw methylation profiling data from GDSC, generated with the Infinium HumanMethylation450 BeadChip array, were downloaded from the Gene Expression Omnibus (GEO: accession number GSE68379 <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68379>). The data was processed with the R Bioconductor package Minfi⁶², performing the noob background subtraction with dye-bias normalisation. After that, we filtered cross-reactive probes⁶³ and probes falling on sex chromosomes. The methylation beta-values were extracted and normalised by using the BMIQ method implemented in the R Bioconductor package ChAMP⁶⁴. The probe annotations were obtained from the package IlluminaHumanMethylation450kanno.ilmn12.hg19⁶⁵.

The raw methylation profiling data from CCLE, generated with the reduced representation bisulfite sequencing (RRBS) methylation profiling technology, were downloaded in the form of fastq files from the Sequence Read Archive (SRA: accession number PRJNA523380 <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA523380/>) using the SRA toolkit. We found 651 cell lines in our selected cancer types and performed quality control analysis and adaptor trimming using FastQC and TrimGalore⁶⁶, respectively. Subsequently, methylation percentage calls were retrieved from Bismark⁶⁷ using methylKit⁶⁸.

For the human primary tumours in TCGA, the preprocessed beta-values from the Infinium HumanMethylation450 BeadChip were downloaded from the GDC data portal (<https://portal.gdc.cancer.gov/>), accessed on the 18th October 2019. They were downloaded and processed with the R package TCGAbiolinks⁶⁹, using the ChAMP preprocessing pipeline consisting of filtering, imputation and normalisation methods with default parameters. Cancer types that either lacked DNA methylation or gene expression data, or had low sample size ($n < 8$), were excluded from further analysis, i.e., LAML, ALL, SCLC, NB, MM and OV.

Gene expression. For the cell lines in the GDSC project, we downloaded the RMA-processed Affymetrix array data from their website <http://www.cancerrxgene.org/gdsc1000/>, accessed on the 8th August 2019. For the human tumours, we downloaded the Hi-Seq count data from the RNAseq experiments in the TCGA database <https://portal.gdc.cancer.gov/>, accessed on the 18th October 2019. For the subsequent analysis, we performed variance stabilising transformation (VST) on the transcript count matrix.

High-throughput drug response screens. For the discovery cohort, we leveraged the HTS from the GDSC project http://www.cancerrxgene.org/downloads/bulk_download release 8.0. We limited the analysis to the 22 cancer types that had >15 fully treated and molecularly characterised cancer cell lines. Drug response was quantified by using the area-under-the-curve (AUC). A drug was required to display partial drug response across at least three cell lines, i.e., $AUC \leq 0.7$. For the independent validation cohort, we used the Cancer Therapeutics Response Portal (CTRP) project <https://portals.broadinstitute.org/ctrp.v2.1>.

Linear models and spatially correlated methylation sites for the identification of differentially methylated regions (dDMR calling). We employed a two-step analysis method to identify the differentially methylated regions of drug response (dDMRs). First, we identified differentially methylated sites in cancer cell lines. For that, we built linear models which fit the drug response denoted as y by the methylation beta-value denoted as m for each CpG site and drug in all cancer types, while correcting for the screening medium (c_1), growth properties (c_2), micro-satellite instabilities (c_3) and the first two principal components (pc_1 , pc_2) to correct for global methylation patterns. Thus, the linear model was defined by

$$y = \beta_0 + \beta_1 m + \beta_2 c_1 + \beta_3 c_2 + \beta_4 c_3 + \beta_5 pc_1 + \beta_6 pc_2, \quad (1)$$

where β_0, \dots, β_6 are the regression coefficients. The analysis was performed for each cancer type separately. The p -values were derived from the significance of the regression coefficient β_1 using a t -test for the respective CpG site. For the extraction of differentially methylated regions of drug response (dDMRs), we employed the software Comb-p^{70,71} with default parameters. We first calculated the auto-correlation (ACF) between sites and the Stouffer-Liptak-Kechris correction of ACFs, followed by subsequent extraction of regions based on the Šidák-adjusted p -values (adj. p) while merging peaks within 1000 bases. dDMRs were called with a cutoff of adj. $p < 10^{-6}$. For the post-processing, the extracted regions were filtered such that there existed more than three cell lines that were aberrantly methylated for each dDMR. For this, we counted the number of cell lines which showed a methylation beta-value < 0.3 and beta-value > 0.7. Furthermore, we filtered regions for which the contained CpG sites did not meet the threshold for the raw $p < 0.01$. The identified region is labelled a dDMR, if both criteria were fulfilled. This subsequently yielded 802 drug differentially methylated regions (dDMRs) for 186 drugs. The effect size for each dDMR was defined as the mean of the regression coefficients β_1 across all CpG sites contained in the called region. The raw p -value (p) for each CpG site and the Pearson correlation coefficient (r) are reported for statistical tests analysing DNA methylation and drug response in the manuscript scatter plots.

Inference of gene regulatory mechanisms as potential drug response biomarkers in cancer cell lines and human tumour samples. To identify the proximal genes that were associated with aberrant methylation, we used the R package ELMER³⁹. We focused on either promoter or distal regions within each cancer type⁴³. For each dDMR, we tested the association between the methylation status and the gene expression with a Mann-Whitney U test according to the default parameters of ELMER³⁹. We corrected for multiple hypothesis testing using a permutative approach with permutation size = 50000, raw p -value threshold = 0.05 and empirical adjusted p -value (emp. adj. p) threshold = 0.001. The empirical adjusted p -value (p) and the Pearson correlation coefficient (r) are reported for statistical tests analysing DNA methylation and gene expression in the manuscript scatter plots. In addition, for cancer cell lines, we tested if the proximal gene expression was associated with drug response independently of its dDMR. For this, we used linear models which fit the drug response to the respective proximal gene expression accordingly with the analogous linear models built using the methylation data.

Protein-protein interaction networks between dDMR proximal genes and drug targets. We identified protein-protein interaction networks in the neighbourhood of tgdDMR-associated genes and drug targets based on the OmniPath database⁴⁴. For each of the 58 tgdDMRs, we extracted the correlated proximal gene and identified the ten shortest paths to each putative drug target using Yen's algorithm⁷². If no path from a gene to a drug target was found in the directed network, we identified paths traversing from the drug target to the tgdDMR gene. In summary, we were able to display protein-protein interaction networks with their shortest paths for 19/58 tgdDMRs, thus enhancing the mechanistic understanding of tgdDMRs.

Somatic variants and their association with tgdDMRs. The GDSC project has compiled a selection of somatic variants and copy number alterations¹¹, which are available at Cell Model Passports (<https://cellmodelpassports.sanger.ac.uk/downloads>). Only somatic mutations in coding regions were considered, which were binarised to represent the mutant and wild type status. Similarly, we binarised amplifications and deletions of gene-level copy number alterations. For both we only considered alterations which showed >3 altered cell lines. For assessing the correlation between genetic alterations and tgdDMRs, we used univariate linear models explaining tgdDMR methylation by the mutational status of each alteration. The p -values were derived from the significance of the regression coefficients and were multiplicity-adjusted by using the Benjamini-Hochberg method.

CRISPR screens and their association with tgdDMRs. CRISPR knockout data and associated gene effects on viability were downloaded from the DepMap Public 22Q4 primary files (<https://depmap.org/portal/download/all/>)^{28,73}. Univariate linear models assessed associations between CRISPR knockouts for each gene in signalling network neighbourhoods of all tgdDMRs. The p -values were derived from the significance of the regression coefficients and were multiple hypothesis-adjusted by the Benjamini-Hochberg correction.

LINCS drug transcriptomic signatures and their association with tgdDMRs. We used the CLUE knowledge base (<https://clue.io/lincs/>)⁷⁴ and its provided API to retrieve transcriptomic gene signatures from the overlapping compounds with matching tissue. Next, we tested for enrichments of each tgdDMR-associated gene and the corresponding genes in the signalling network neighbourhood in the set of gene signatures using a binomial test. The resulting p -values were adjusted using the Benjamini-Hochberg method.

Statistics and reproducibility. The sample sizes of the GDSC, CCLE/CTRP and TCGA data were predetermined by their data availability. We selected cancer types with >15 distinct molecularly characterised cell lines in the GDSC dataset. Cancer cell lines in the GDSC were parallelly treated according to the previously published study protocol¹¹. For the matching cancer types, all distinct primary tumour samples with both available DNA methylation and gene expression data in the CCLE and TCGA data were selected. For all datasets, this resulted in 22 cancer types: small-cell lung cancer (SCLC; $n_{GDSC} = 63$; $n_{CCLE} = 36$; $n_{TCGA} = 0$), lung adenocarcinoma (LUAD; $n_{GDSC} = 63$; $n_{CCLE} = 87$; $n_{TCGA} = 484$), skin cutaneous melanoma (SKCM; $n_{GDSC} = 52$; $n_{CCLE} = 50$; $n_{TCGA} = 104$), breast invasive carcinoma (BRCA; $n_{GDSC} = 49$; $n_{CCLE} = 39$; $n_{TCGA} = 861$), colorectal adenocarcinoma (COREAD; $n_{GDSC} = 46$; $n_{CCLE} = 47$; $n_{TCGA} = 325$), head and neck squamous cell carcinoma (HNSC; $n_{GDSC} = 40$; $n_{CCLE} = 29$; $n_{TCGA} = 520$), glioblastoma (GBM; $n_{GDSC} = 35$; $n_{CCLE} = 37$; $n_{TCGA} = 51$), esophageal carcinoma (ESCA; $n_{GDSC} = 35$; $n_{CCLE} = 14$; $n_{TCGA} = 170$), ovarian serous cystadenocarcinoma (OV; $n_{GDSC} = 34$; $n_{CCLE} = 30$; $n_{TCGA} = 7$), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC; $n_{GDSC} = 33$; $n_{CCLE} = 28$; $n_{TCGA} = 48$), neuroblastoma (NB; $n_{GDSC} = 32$; $n_{CCLE} = 14$; $n_{TCGA} = 0$), kidney renal clear cell carcinoma (KIRC; $n_{GDSC} = 30$; $n_{CCLE} = 21$; $n_{TCGA} = 344$), pancreatic adenocarcinoma (PAAD; $n_{GDSC} = 29$; $n_{CCLE} = 38$; $n_{TCGA} = 181$), acute myeloid leukemia (LAML; $n_{GDSC} = 25$; $n_{CCLE} = 29$; $n_{TCGA} = 0$), acute lymphocytic leukemia (ALL; $n_{GDSC} = 25$; $n_{CCLE} = 24$; $n_{TCGA} = 0$), stomach adenocarcinoma (STAD; $n_{GDSC} = 23$; $n_{CCLE} = 29$; $n_{TCGA} = 338$), mesothelioma (MESO; $n_{GDSC} = 21$; $n_{CCLE} = 8$; $n_{TCGA} = 86$), bladder urothelial carcinoma (BLCA; $n_{GDSC} = 19$; $n_{CCLE} = 24$; $n_{TCGA} = 428$), multiple myeloma (MM; $n_{GDSC} = 17$; $n_{CCLE} = 24$; $n_{TCGA} = 0$), liver hepatocellular carcinoma (LIHC; $n_{GDSC} = 17$; $n_{CCLE} = 20$; $n_{TCGA} = 412$), brain low-grade glioma (LGG; $n_{GDSC} = 17$; $n_{CCLE} = 15$; $n_{TCGA} = 511$) and thyroid carcinoma (THCA; $n_{GDSC} = 16$; $n_{CCLE} = 10$; $n_{TCGA} = 551$). The reproducibility of biomarkers was assessed by the overlapping CCLE/CTRP DNA methylation and drug response data as independent validation cohort. Discrepancies between drug response biomarkers in CCLE/CTRP may arise due to technical noise or differences in drug screening assays, but showed high consistency as reported.

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets that were analysed in this study are publicly available within the outlined repositories. Specifically, the GDSC and CCLE DNA methylation data are available on Gene Expression Omnibus (GEO; accession number [GSE68379](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68379)) and Sequence Read Archive (SRA; accession number [PRJNA523380](https://www.ncbi.nlm.nih.gov/sra/PRJNA523380)), respectively. The TCGA DNA methylation data is available on the GDC data portal <https://portal.gdc.cancer.gov/>. The GDSC and CCLE drug response data are available on http://www.cancerrxgene.org/downloads/bulk_download release 8.0 and the Cancer Therapeutics Response Portal <https://portals.broadinstitute.org/ctrp.v2.1>, respectively. The GDSC and TCGA gene expression data are available on <http://www.cancerrxgene.org/gdsc1000/> and the GDC data portal <https://portal.gdc.cancer.gov/>, respectively. The GDSC somatic variants and copy number alterations are available at Cell Model Passports <https://cellmodelpassports.sanger.ac.uk/downloads>. The CRISPR screens are available on DepMap <https://depmap.org/>.

[org/portal/download/all/](https://portal/download/all/) and the LINCS data is available on CLUE <https://clue.io/lincs>. The processed datasets are publicly available on Zenodo⁷⁵. Source data for the figure panels are provided in Supplementary Data 3.

Code availability

The source code for the presented analysis is available at <https://github.com/MendenLab/pheb> v0.1.0. It refers to a runnable docker image that contains all used software for data analysis. The statistical analysis can be reproduced with the source code and datasets provided on Zenodo⁷⁵.

Received: 14 January 2023; Accepted: 1 August 2023;

Published online: 09 August 2023

References

- Shameer, K., Readhead, B. & Dudley, J. T. Computational and experimental advances in drug repositioning for accelerated therapeutic stratification. *Curr. Top. Med. Chem.* **15**, 5–20 (2015).
- Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2019).
- Saghafinia, S., Mina, M., Riggi, N., Hanahan, D. & Ciriello, G. Pan-cancer landscape of aberrant DNA methylation across human tumors. *Cell Rep.* **25**, 1066–1080.e8 (2018).
- Nyce, J., Leonard, S., Canupp, D., Schulz, S. & Wong, S. Epigenetic mechanisms of drug resistance: drug-induced DNA hypermethylation and drug resistance. *Proc. Natl Acad. Sci. USA* **90**, 2960–2964 (1993).
- Wilting, R. H. & Dannenberg, J.-H. Epigenetic mechanisms in tumorigenesis, tumor cell heterogeneity and drug resistance. *Drug Resist. Updat.* **15**, 21–38 (2012).
- Nishiyama, A. & Nakanishi, M. Navigating the DNA methylation landscape of cancer. *Trends Genet.* **37**, 1012–1027 (2021).
- Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
- Yang, X. et al. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26**, 577–590 (2014).
- Masters, J. R. W. Human cancer cell lines: fact and fantasy. *Nat. Rev. Mol. Cell Biol.* **1**, 233–236 (2000).
- Garnett, M. J. et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
- Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
- Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
- Seashore-Ludlow, B. et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* **5**, 1210–1223 (2015).
- Basu, A. et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151–1161 (2013).
- Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
- Menden, M. P. et al. The germline genetic component of drug sensitivity in cancer cell lines. *Nat. Commun.* **9**, 3385 (2018).
- Kim, G. et al. FDA approval summary: vemurafenib for treatment of unresectable or metastatic melanoma with the BRAFV600E mutation. *Clin. Cancer Res.* **20**, 4994–5000 (2014).
- Kronfol, M. M., Dozmorov, M. G., Huang, R., Slattum, P. W. & McClay, J. L. The role of epigenomics in personalized medicine. *Expert Rev. Precis Med. Drug Dev.* **2**, 33–45 (2017).
- Kamińska, K. et al. Prognostic and predictive epigenetic biomarkers in oncology. *Mol. Diagn. Ther.* **23**, 83–95 (2019).
- Issa, J.-P. CpG island methylator phenotype in cancer. *Nat. Rev. Cancer* **4**, 988–993 (2004).
- Lv, W. et al. Exploration of drug-response mechanism by integrating genetics and epigenetics across cancers. *Epigenomics* **10**, 993–1010 (2018).
- Jia, M., Gao, X., Zhang, Y., Hoffmeister, M. & Brenner, H. Different definitions of CpG island methylator phenotype and outcomes of colorectal cancer: a systematic review. *Clin. Epigenetics* **8**, 25 (2016).
- Noushmehr, H. et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17**, 510–522 (2010).
- Kelly, A. D. et al. A CpG island methylator phenotype in acute myeloid leukemia independent of IDH mutations and associated with a favorable outcome. *Leukemia* **31**, 2011–2019 (2017).
- Yuan, R., Chen, S. & Wang, Y. Computational prediction of drug responses in cancer cell lines from cancer omics and detection of drug effectiveness related methylation sites. *Front. Genet.* **11**, 917 (2020).
- Vural, S. et al. Association of expression of epigenetic molecular factors with DNA methylation and sensitivity to chemotherapeutic agents in cancer cell lines. *Clin. Epigenetics* **13**, 49 (2021).
- Picco, G. et al. Functional linkage of gene fusions to cancer cell fitness assessed by pharmacological and CRISPR-Cas9 screening. *Nat. Commun.* **10**, 2198 (2019).
- Gonçalves, E. et al. Drug mechanism-of-action discovery through the integration of pharmacological and CRISPR screens. *Mol. Syst. Biol.* **16**, e9405 (2020).
- Repina, D. et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol.* **20**, 1 (2019).
- Butler, M. et al. MGMT status as a clinical biomarker in glioblastoma. *Trends Cancer Res.* **6**, 380–391 (2020).
- Tancredi, A. et al. BET protein inhibition sensitizes glioblastoma cells to temozolomide treatment by attenuating MGMT expression. *Cell Death Dis.* **13**, 1037 (2022).
- Zheng, Y. et al. Temporal regulation of EGF signalling networks by the scaffold protein Shc1. *Nature* **499**, 166–171 (2013).
- Ramalingam, S. S. et al. Randomized phase II study of dacomitinib (PF-00299804), an irreversible pan-human epidermal growth factor receptor inhibitor, versus erlotinib in patients with advanced non-small-cell lung cancer. *J. Clin. Oncol.* **30**, 3337–3344 (2012).
- Wu, Y.-L. et al. Dacomitinib versus gefitinib as first-line treatment for patients with EGFR-mutation-positive non-small-cell lung cancer (ARCHER 1050): a randomised, open-label, phase 3 trial. *Lancet Oncol.* **18**, 1454–1466 (2017).
- Zoppoli, G. et al. Putative DNA/RNA helicase Schlafen-11 (SLFN11) sensitizes cancer cells to DNA-damaging agents. *Proc. Natl Acad. Sci. USA* **109**, 15030–15035 (2012).
- Coleman, N., Zhang, B., Byers, L. A. & Yap, T. A. The role of Schlafen 11 (SLFN11) as a predictive biomarker for targeting the DNA damage response. *Br. J. Cancer* **124**, 857–859 (2021).
- Winkler, C. et al. SLFN11 informs on standard of care and novel treatments in a wide range of cancer models. *Br. J. Cancer* **124**, 951–962 (2021).
- Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23–38 (2013).
- Silva, T. C. et al. ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. *Bioinformatics* **35**, 1974–1977 (2019).
- Smith, J., Sen, S., Weeks, R. J., Eccles, M. R. & Chatterjee, A. Promoter DNA hypermethylation and paradoxical gene activation. *Trends Cancer Res.* **6**, 392–406 (2020).
- Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V. & Jordan, I. K. On the presence and role of human gene-body DNA methylation. *Oncotarget* **3**, 462–474 (2012).
- Spainhour, J. C., Lim, H. S., Yi, S. V. & Qiu, P. Correlation patterns between DNA methylation and gene expression in the cancer genome atlas. *Cancer Inform.* **18**, 1176935119828776 (2019).
- Weinstein, J. N. et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113 (2013).
- Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
- Shah, N. & Sukumar, S. The Hox genes and their roles in oncogenesis. *Nat. Rev. Cancer* **10**, 361–371 (2010).
- Flagiello, D., Poupon, M. F., Cillo, C., Dutrillaux, B. & Malfoy, B. Relationship between DNA methylation and gene expression of the HOXB gene cluster in small cell lung cancers. *FEBS Lett.* **380**, 103–107 (1996).
- Smith, S. C. et al. A gemcitabine sensitivity screen identifies a role for NEK9 in the replication stress response. *Nucleic Acids Res.* **42**, 11517–11527 (2014).
- Wolinski, F. S. et al. The NAE inhibitor pevonedistat (MLN4924) synergizes with TNF- α to activate apoptosis. *Cell Death Discov.* **1**, 15034 (2015).
- Jang, S.-M., Redon, C. E., Thakur, B. L., Bahta, M. K. & Aladjem, M. I. Regulation of cell cycle drivers by Cullin-RING ubiquitin ligases. *Exp. Mol. Med.* **52**, 1637–1651 (2020).
- Paiva, C., Godbersen, J. C., Berger, A., Brown, J. R. & Danilov, A. V. Targeting neddylation induces DNA damage and checkpoint activation and sensitizes chronic lymphocytic leukemia B cells to alkylating agents. *Cell Death Dis.* **6**, e1807 (2015).
- Michelen, J. et al. Analysis of PARP inhibitor toxicity by multidimensional fluorescence microscopy reveals mechanisms of sensitivity and resistance. *Nat. Commun.* **9**, 2678 (2018).
- Mazzio, E. A., Lewis, C. A., Elhag, R. & Soliman, K. F. Effects of sepantronium bromide (YM-155) on the whole transcriptome of MDA-MB-231 cells: highlight on impaired ATR/ATM fanconi anemia DNA damage response. *Cancer Genomics Proteom.* **15**, 249–264 (2018).
- Hu, W.-F. et al. α A- and α B-crystallins interact with caspase-3 and Bax to guard mouse lens development. *Curr. Mol. Med.* **12**, 177–187 (2012).

54. Liu, S. et al. As a novel p53 direct target, bidirectional gene HspB2/aB-crystallin regulates the ROS level and Warburg effect. *Biochim. Biophys. Acta* **1839**, 592–603 (2014).
55. Mirabelli, P., Coppola, L. & Salvatore, M. Cancer cell lines are useful model systems for medical research. *Cancers* **11**, 1098 (2019).
56. Reis-Filho, J. S. & Pusztai, L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet* **378**, 1812–1823 (2011).
57. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
58. Cheng, Y. et al. Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. *Signal Transduct. Target Ther.* **4**, 62 (2019).
59. van der Meer, D. et al. Cell Model Passports—a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic Acids Res.* **47**, D923–D929 (2019).
60. Siu, L. L. et al. Facilitating a culture of responsible and effective sharing of cancer genome data. *Nat. Med.* **22**, 464–471 (2016).
61. Wang, X. et al. Characteristics of The Cancer Genome Atlas cases relative to U.S. general population cancer cases. *Br. J. Cancer* **119**, 885–892 (2018).
62. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
63. Chen, Y.-A. et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).
64. Morris, T. J. et al. ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics* **30**, 428–430 (2014).
65. Kd, H. IlluminaHumanMethylation450kanno. ilmn12. hg19: annotation for illumina's 450k methylation arrays. *R package version 0.2.1*, (2016).
66. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
67. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
68. Akalin, A. et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87 (2012).
69. Colaprico, A. et al. TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71 (2016).
70. Pedersen, B. S., Schwartz, D. A., Yang, I. V. & Kechris, K. J. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics* **28**, 2986–2988 (2012).
71. Mallik, S. et al. An evaluation of supervised methods for identifying differentially methylated regions in Illumina methylation arrays. *Brief. Bioinform.* **20**, 2224–2235 (2019).
72. Yen, J. Y. An algorithm for finding shortest routes from all source nodes to a given destination in general networks. *Quart. Appl. Math.* **27**, 526–530 (1970).
73. Behan, F. M. et al. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature* **568**, 511–516 (2019).
74. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17 (2017).
75. Ohnmacht, A. J. et al. The pharmacoeigenomic landscape of cancer cell lines reveals the epigenetic component of drug sensitivity. <https://doi.org/10.5281/ZENODO.8161472> (Zenodo, 2023).

Acknowledgements

This project was supported by the European Union's Horizon 2020 Research and Innovation Programme (Grant agreement No. 950293-COMBAT-RES).

Author contributions

Conceptualization: A.J.O. and M.P.M.; Data curation: A.J.O. and A.R.; Analysis, A.J.O. and M.P.M.; Methodology: A.J.O. and M.P.M.; Supervision: M.P.M.; Visualisation: A.J.O.; Writing original draft: A.J.O. and M.P.M.; Writing, review and editing: A.J.O., A.R., G.A., G.K., E.G., D.S. and M.P.M.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

M.P.M. collaborates with GSK, Roche and AstraZeneca, and receives funding from Roche and GSK. M.P.M. is a former employee at AstraZeneca. The remaining authors declare no competing interest.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-023-05198-y>.

Correspondence and requests for materials should be addressed to Michael Patrick Menden.

Peer review information *Communications Biology* thanks Yuan Liu, Christian Bergsland and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editors: Silvia Belluti and George Inglis. A peer review file is available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

2.2 The pharmacogenomic assessment of molecular epithelial-mesenchymal transition signatures reveals drug susceptibilities in cancer cell lines, *bioRxiv* (2024)

This article is an open-access preprint in *bioRxiv* prior to peer review [2]. It is publicly available at <https://doi.org/10.1101/2024.01.16.575190>.

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The pharmacogenomic assessment of molecular epithelial-mesenchymal transition signatures reveals drug susceptibilities in cancer cell lines

Alexander J. Ohnmacht ^{1,2,*}, Göksu Avar ^{1,2,*}, Marisa K. Schübel ^{1,2,*}, Thomas J. O'Neill ³, Daniel Krappmann ^{2,3,#}, Michael P. Menden ^{1,4,#}

¹ Computational Health Center, Helmholtz Munich, Neuherberg, Germany

² Department of Biology, Ludwig-Maximilians University Munich, Martinsried, Germany

³ Research Unit Signaling and Translation, Group Signaling and Immunity, Helmholtz Munich, Neuherberg, Germany

⁴ Department of Biochemistry and Pharmacology, University of Melbourne, Victoria, Australia

* equal contribution

corresponding authors: daniel.krappmann@helmholtz-munich.de; michael.menden@unimelb.edu.au

Abstract

The epithelial-mesenchymal transition (EMT) is characterised by the loss of cell-cell adhesion and cell polarity, which is often exploited by cancer cells to adopt a motile, invasive and metastatic phenotype. Whilst EMT is often linked with cancer progression and therapy resistance, strategies for its selective targeting remain limited. In order to address this, we infer EMT states of cancer cell lines from their molecular signatures and use predictive and causal modelling to estimate the effect of EMT on drug susceptibility in high-throughput drug screens. For example, we show that EMT signatures in melanoma cells can predict favourable responses to the HSP90 inhibitor luminespib and demonstrate that epithelial-like melanoma cells can be sensitised to luminespib upon stimulation of EMT by TGF- β . Thus, we provide an analysis that systematically yields a set of potent drugs by exploiting vulnerabilities of cancer cells undergoing EMT, which may pave the way for therapies to target these cells.

Introduction

The epithelial-mesenchymal transition (EMT) is a cellular process that allows cells to transition between different phenotypic states ¹. Rather than a switch between two distinct phenotypes, the EMT program describes a dynamic spectrum of phenotypes between epithelial and mesenchymal cells ranging from apical-basal polarity and strong cell-cell contacts to motile and spindle-like characteristics ². EMT is an essential mechanism for embryonic development, wound healing and tumour plasticity, and has been regarded as a hallmark of cancer ^{1,3–5}. The invasion of the extracellular matrix by cancer cells undergoing EMT prior to metastasis is accompanied by the loss of the adherens junction protein E-cadherin and upregulation of N-cadherin, vimentin and fibronectin^{1,6}. Scores derived from gene expression signatures of these molecular markers are typically used for assessing EMT and its associated cellular processes in cancers ^{2,7–10}. Some of these processes can be used to externally stimulate cells to undergo EMT. For example, TGF- β signalling is an established mechanism for inducing EMT ¹¹ and thus TGF- β treatment is widely used for external EMT induction *in vitro* ^{12–17}.

Sparse findings in cancer cell lines and human tumours have reported EMT as a putative drug response biomarker ^{7–10}. For example, acquired resistance through EMT has been reported for commonly

employed chemotherapeutic agents, e.g. cisplatin and doxorubicin¹⁸, and targeted therapies, e.g. EGFR or PI3K inhibitors¹⁹. Furthermore, EMT was found to cause intrinsic resistance to KRAS inhibitors in lung cancer²⁰. Although the genetic background has been shown to play an important role in enabling EMT in cancer progression¹⁰, it is still unclear to what extent initial cancer drug responses can be attributed to EMT. Thus, we hypothesised that predictive and causal modelling of EMT scores in drug high-throughput screens can assess the role of EMT in cancer drug sensitivity, which may lead to strategies that systematically exploit EMT as a cancer vulnerability.

Here, we first estimated continuous EMT scores based on gene expression profiles of 790 cancer cell lines from 31 cancer types using four different methods^{7,8,10,21}. Consecutively, we benchmarked the contribution of EMT in drug response prediction models and quantitatively estimated the EMT effect with causal inference. For example, we revealed that EMT and its related processes in melanoma robustly predict sensitivity to HSP90 inhibition with luminespib and other HSP90 inhibitors. Indeed, we experimentally demonstrated that stimulating EMT with TGF- β pretreatment can sensitise epithelial melanoma cell lines to luminespib.

Results

We leveraged a high-throughput drug screen (HTS; **Fig. 1a**) of 790 cancer cell lines across 31 cancer types, which were treated with 544 unique compounds to obtain dose-response curves (**Fig. 1b**)^{22–24}. This was complemented with molecular profiling of cancer cell lines, i.e. somatic mutations, copy number alterations and gene expression (**Fig. 1c**)^{22–24}. For estimating EMT, we derived EMT scores from four established methods that leverage molecular signatures to infer EMT on a continuous spectrum using gene expression data (**Fig. 1c**); these were: Mak *et al.*⁸, gene set variation analysis²¹, Tan *et al.*⁷ and Tagliazucchi and Wiecek *et al.*¹⁰ (**Methods**), abbreviated as MAK, GSVA, TAN and TW, respectively (**Supplementary Data 1**). Then, we systematically benchmarked the EMT scores for predicting drug responses across all compounds and cancer types using (1) ablation of the EMT score and (2) causal inference of the EMT effect (**Fig. 1c; Methods**). Thereby, the cancer somatic alterations served as background predictors for assessing the EMT-specific component.

Exemplifying our method, we leveraged the MAK EMT scores and drug responses quantified by IC₅₀ values in skin cutaneous melanoma (SKCM) and identified four inhibitory compounds, for which the full model including EMT significantly outperformed the baseline model (**Fig. 1d; Methods**). For example, response to luminespib in SKCM was predicted well by the full model, i.e. leveraging EMT scores and the mutational background with Pearson's $r = 0.50$ between actual and predicted IC₅₀ values. However, the performance of actual versus predicted IC₅₀ dropped to Pearson's $r = 0.02$ upon exclusion of the EMT score ($\Delta r = 0.47$, $p = 5.0 \times 10^{-4}$, t -test for resampled performance metrics; **Supplementary Fig. 1a**). For the identified compounds, we applied double machine learning in conjunction with causal random forests to estimate the EMT-specific effect on drug responses with a 95% confidence interval (**Fig. 1e; Methods**)^{25–28}. Compounds with significantly increased performance and high inferred effect size for multiple EMT scores and both IC₅₀ and area under the drug response curve (AUC) suggested

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a causal component of EMT for determining drug responses (**Supplementary Fig. 1b-f**). This hypothesis was systematically dissected across the remaining cancer types, EMT scores and drug response readouts in the next section.

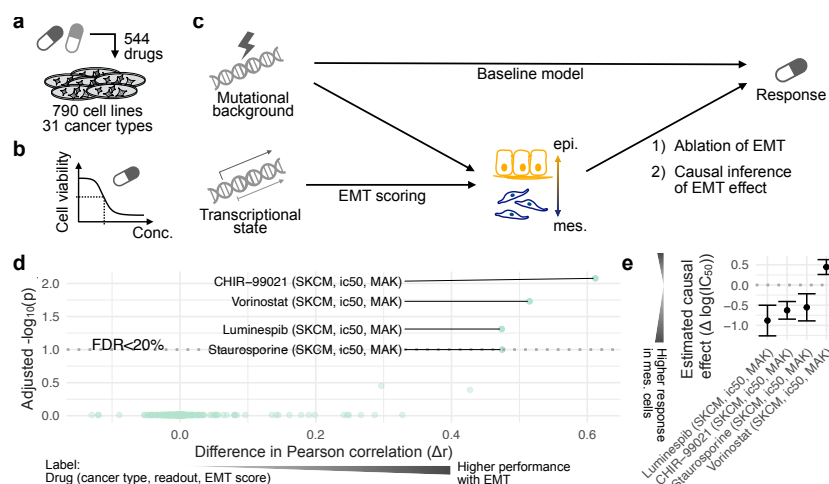


Figure 1: Modelling drug susceptibility in cancer cell lines in the context of EMT. (a) The drug high-throughput screen contained 790 cancer cell lines treated with 544 compounds and (b) their dose-response curves. (c) Molecular profiling of cancer cell lines quantifies their mutational background and the transcriptional state. Shown in the schematic workflow for predictive and causal modelling, the mutational background are baseline features and EMT scores are derived from gene expression data. First, a regression-based ablation study assessed the predictive performances of the drug response prediction model upon excluding EMT from the model. Secondly, the EMT-specific effect on drug susceptibility was estimated by causal inference methods (**Methods**). (d) The systematic ablation study in SKCM yielded a set of compounds for which EMT improved the response predictions, showing the adjusted p -values of a t -test for performance metrics and the difference in Pearson's correlation Δr . (e) The inferred EMT effects on responses to the identified set of compounds with the 95% confidence interval in SKCM is shown.

Systematic analysis of EMT and its regulators as biomarkers of cancer drug sensitivity

The distributions of MAK, GSVA, TAN and TW scores were predominantly cancer type specific (**Fig. 2a-d**). For example, SKCM cell lines showed a more mesenchymal MAK EMT score, whilst breast cancer (BRCA) and colorectal cancer (COREAD) cell lines displayed rather epithelial MAK EMT scores (**Fig. 2a**), which highlighted the high tissue-specificity of EMT molecular signatures. MAK, TAN and TW scores showed high overall correlations (Pearson's $r > 0.87$; **Supplementary Fig. 2a**), which were consistently high within cancer types. GSVA showed lower overall correlations with these scores (Pearson's $r < 0.39$; **Supplementary Fig. 2a**) due to normalised scores (**Supplementary Fig. 2b**), but displayed consistently high correlations within cancer types as well (**Supplementary Fig. 2a**).

We conducted the benchmark with the outlined modelling strategies (**Fig. 1c**; **Methods**), and recorded its results across all included cancer types, EMT scores, compounds and IC_{50} or AUC (**Supplementary Data 2**; **Methods**). Six cancer types showed at least one significant compound with $FDR < 0.2$ (**Fig. 2e**; **Methods**). We estimated the EMT effects and confidence intervals for all compounds (**Supplementary Fig. 3a,b**; **Methods**), and further focused on five compounds in three cancer types

2 Results

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that consistently showed significant performance gains in at least three models (labelled in **Fig. 2e,f**; **Table 1**; **Methods**). For all of these compounds, mesenchymal cells showed higher drug responses than epithelial cells (**Fig. 2f**).

Drug	Target	Cancer type	Readout	EMT Score	Δr	Raw p -value	Adj. p -value	Effect and interval	TF	Responder GO	Downregulated GO	Upregulated GO
Luminespib	HSP90	SKCM	ic50	MAK	0.47	0.0005	0.049	-0.881±0.380	MITF	GO:0016241 GO:1902600	Epi: GO:0042127 ¹ GO:0045595 ² Mes: GO:0007178 GO:0006355 GO:0045893 GO:1903844 ⁴ GO:0017015 ⁵ GO:0045596 ⁶	Epi: GO:0036503 GO:0006986 ³ Mes: GO:0034976 GO:0006986 ³
			auc	MAK	0.51	0.0002	0.050	-0.084±0.035	MITF	GO:0007035 GO:1902600 GO:0035751		
			auc	GSVA	0.39	0.0008	0.156	-0.076±0.020				
			ic50	MAK	0.61	2.8×10^{-5}	0.008	-0.628±0.218	MITF	GO:0032438 GO:0045333	Mes: GO:1901203 GO:0007178 GO:0007179 ⁷	
			auc	MAK	0.60	0.0005	0.066	-0.037±0.012	MITF	GO:0019646 GO:0042775 GO:0045333		
			auc	GSVA	0.54	0.0012	0.169	-0.031±0.015				
CHIR-99021	GSK3A/B	SKCM	ic50	MAK	0.61	2.8×10^{-5}	0.008	-0.628±0.218	MITF	GO:0032438 GO:0045333	Mes: GO:1901203 GO:0007178 GO:0007179 ⁷	
			auc	MAK	0.60	0.0005	0.066	-0.037±0.012	MITF	GO:0019646 GO:0042775 GO:0045333		
			auc	GSVA	0.54	0.0012	0.169	-0.031±0.015				
			ic50	GSVA	0.54	0.0001	0.042	-0.538±0.259				
			auc	TW	0.59	7.4×10^{-7}	0.0003	-0.038±0.014				
			auc	MAK	0.47	0.0013	0.100	-0.553±0.336	MITF	GO:0051452 GO:0032438		
Staurosporine	broad multi-kinase	SKCM	ic50	MAK	0.47	0.0013	0.100	-0.553±0.336	MITF	GO:0051452 GO:0032438		
			auc	MAK	0.43	0.0010	0.091	-0.065±0.036	MITF	GO:0007032 GO:0051452 GO:1902600		
			auc	GSVA	0.61	0.0003	0.146	-0.068±0.039				
			ic50	GSVA	0.63	0.0005	0.074	-0.618±0.379				
GSK269962A	ROCK1/2	LUAD	ic50	MAK	0.51	0.0001	0.032	-0.666±0.236	SOX2	GO:0018212 GO:0010632 ⁸		
			ic50	TAN	0.46	0.0006	0.157	-0.604±0.288				
			ic50	TW	0.58	5.4×10^{-5}	0.014	-0.517±0.310				
AZD7762	CHEK1/2	BRCA	auc	MAK	0.59	2.3×10^{-5}	0.013	-0.046±0.069	ESR1	GO:0010256 GO:0072659 GO:0006892 GO:1990778		
			auc	GSVA	0.63	1.5×10^{-7}	8.0×10^{-5}	-0.069±0.042				
			auc	GSVA	0.56	0.0014	0.183	-0.118±0.054				

¹ Regulation of cell population proliferation

² Regulation of cell differentiation

³ Response to unfolded protein

⁴ Regulation of cellular response to transforming growth factor beta stimulus

⁵ Regulation of transforming growth factor beta receptor signaling pathway

⁶ Negative regulation of cell differentiation

⁷ Transforming growth factor beta receptor signaling pathway

⁸ Regulation of epithelial cell migration

Table 1: The five top-ranked EMT-dependent compounds. Each of the five compounds is characterised by its name, target and cancer type for which the association was found. The statistics for the ablation study (Δr and (adjusted) p -value) and causal inference (effect size plus interval in terms of $\Delta \log(\text{IC}_{50})$ or ΔAUC) are given for each response readout and EMT score. Furthermore, the enriched TFs and GO terms for the responding cell lines and the enriched GO terms in transcriptional signatures for the compounds in SKCM are shown. Selected GO terms are annotated in the footnotes.

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To interpret the five top-ranked EMT-dependent compounds, we employed gene set enrichment analysis of the set of differentially expressed genes between higher and lower responding cell lines leveraging the entire transcriptome (**Supplementary Data 3; Methods**). The enrichment analysis found upstream transcription factors (TF) acting as master regulators that can affect both molecular EMT markers and the set of differentially expressed genes. For example, we found that responses to the CHK1 inhibitor AZD7762 in BRCA were successfully predicted by the MAK and GSVA EMT scores (**Fig. 2e; Supplementary Fig. 3c; Table 1**), and identified that the TF target genes of ESR1 were enriched (adj. $p = 4.3 \times 10^{-28}$, odds ratio = 2.44, Fisher's exact test; **Supplementary Data 3; Table 1**). Furthermore, lower response to AZD7762 was associated with higher *ESR1* expression (**Supplementary Fig. 4a**).

ESR1 expression is associated with clinical BRCA subtypes (PAM50), especially the invasive basal BRCA subtype is characterised by low *ESR1* expression²⁹. Accordingly, we confirmed that cell lines derived from the more invasive basal-like BRCA displayed higher MAK EMT scores resembling the mesenchymal phenotype ($p = 0.002$, ANOVA F-test; **Fig. 2g**). Therefore, we added the PAM50 subtype to the EMT score and *ESR1* expression as fixed effects in a regression model predicting AZD7762 response and found that it did not further improve our model ($p = 0.64$, ANOVA F-test for multiple regression coefficient; **Fig. 2g**). Similar to *ESR1* expression, it is established that BRCA1/2 regulates the cell cycle by activating CHK1 in response to DNA damage and its mutations are associated with oncogenesis³⁰. Thus, we repeated the same analysis by excluding cell lines that carry *BRCA1/2* mutations ($p = 0.73$, ANOVA F-test for multiple regression coefficient; **Fig. 2g**), which also did not further improve our model. Concordantly, EMT regulators were previously shown to underlie DNA damage responses through their interaction with CHK1/2 (target of AZD7762) in BRCA cells³¹. In summary, EMT as a predictive biomarker for AZD7762 response in BRCA reflected but was not further enhanced by BRCA subtypes and somatic mutations in *BRCA1/2*.

Furthermore, we observed performance gains for the ROCK1 (Rho kinase 1) inhibitor GSK269962A, to which lung adenocarcinoma mesenchymal-like cell lines with a higher MAK, TAN and TW EMT score were more responsive (**Supplementary Fig. 4b; Table 1**). We identified an associated TF SOX2 (adj. $p = 1.5 \times 10^{-13}$, odds ratio = 8.95, Fisher's exact test; **Supplementary Data 3; Table 1**), which was previously found to be associated with EMT and metastasis in multiple cancer types, including lung cancer³². We expanded the enrichment analysis of the set of differentially expressed genes for responder cell lines to Gene Ontology (GO) biological processes and found that upregulated genes in LUAD cell lines responding to GSK269962A were enriched in genes involved in the regulation of epithelial cell migration (adj. $p = 0.0004$, odds ratio = 35.07, Fisher's exact test; **Supplementary Data 4; Table 1; Methods**), which is orchestrated by ROCK1³³.

In summary, our proposed method was able to robustly identify compounds in HTS that demonstrated distinct drug responses depending on EMT, its upstream regulators and related processes in several

2 Results

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cancer types. In the next section, we focused on elucidating further mechanisms on the compounds identified in SKCM.

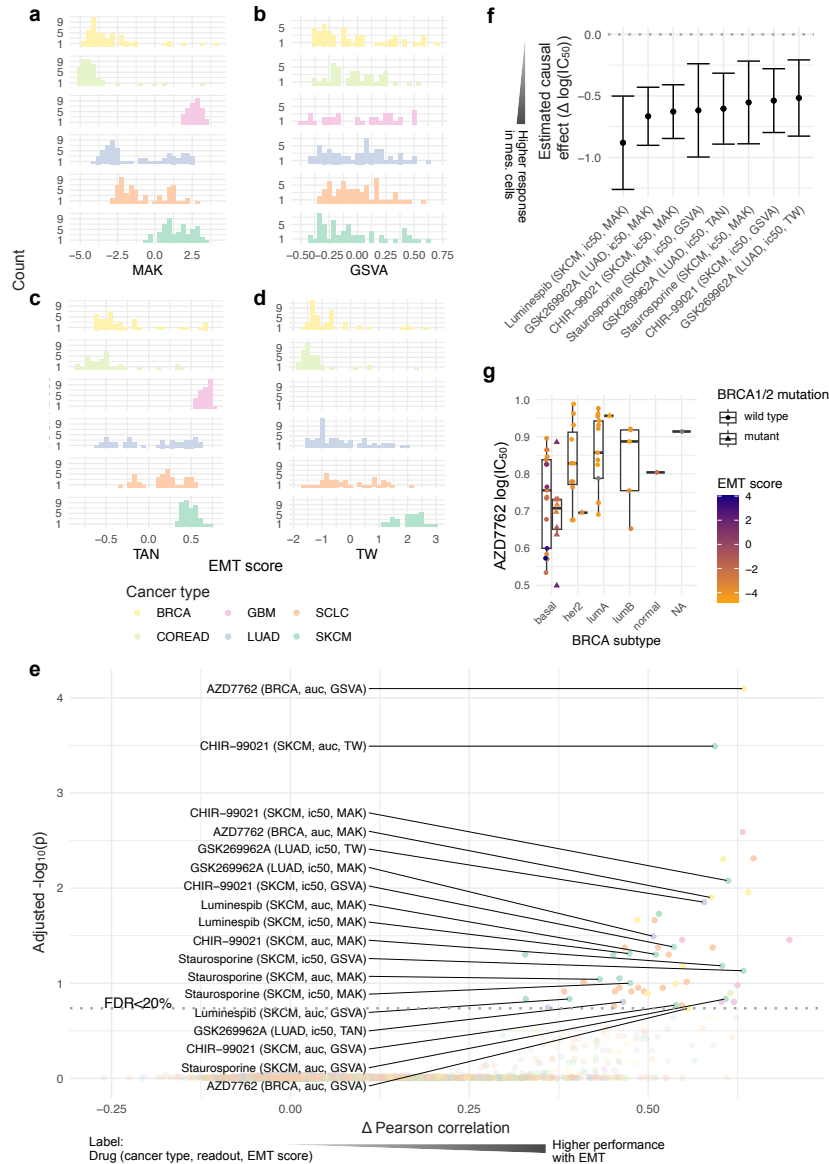


Figure 2: EMT as a predictive component of drug sensitivity. Distributions of (a) MAK, (b) GSVA, (c) TAN and (d) TW EMT scores depending on the cancer type are displayed. (e) The systematic ablation study demonstrates EMT as a predictor of drug sensitivity in cancer cell lines for four different EMT scores and two response readouts, i.e. IC_{50} and AUC, with six cancer types that showed at least one significant compound (FDR < 0.2). The compounds and cancer types that showed at least 3 significant performance changes are labelled. The horizontal axis represents the difference in mean Pearson's correlation Δr between predicted and actual IC_{50} or AUC values of the models, including and excluding EMT, whereas the vertical axis measures the significant improvement in the performance over the baseline model using a t -test for resampled performance metrics. (f) For the IC_{50} prediction models, the estimated EMT effects plus 95% confidence intervals are displayed. (g) The boxplot shows the response to CHK1/2 inhibitor AZD7762 depending on BRCA PAM50 subtypes, the MAK EMT score and mutations in *BRCA1/2*. The centre on the boxplot represents the median, while the box illustrates the interquartile range (IQR). The whiskers show a range that is 1.5 times the size of the IQR.

Potential regulators and drug response mechanisms of EMT in melanoma cell lines

We focused on the three remaining compounds in SKCM, i.e. CHIR-99021, luminespib and staurosporine, for which mesenchymal-like SKCM cell lines showed higher sensitivity consistently across at least three models (**Fig 2e,f; Table 1**). In the TF enrichment analysis (**Methods**), we found that MITF was enriched in the set of differentially expressed genes for these compounds, i.e. CHIR-99021 (adj. $p = 3.7 \times 10^{-36}$, odds ratio = 3.44), luminespib (adj. $p = 6.8 \times 10^{-7}$, odds ratio = 7.54) and staurosporine (adj. $p = 5.3 \times 10^{-69}$, odds ratio = 4.47; Fisher's exact test; **Supplementary Data 3; Table 1**) and showed responses associated with MITF expression (**Fig. 3a**). For luminespib, 39 genes were significantly downregulated in responding mesenchymal-like cells, from which 23 were putative MITF target genes (**Fig. 3b**). MITF is a melanocyte master regulator and is often described as an oncogene in melanoma³⁴. It was proposed to act as a phenotype-switching regulator in melanoma, for which cells with trace MITF levels show senescent properties characterised by cell cycle arrest and cell motility, low-to-intermediate MITF levels display proliferative properties, and higher MITF levels can drive cell differentiation^{35–38}.

The MAK EMT score in SKCM was associated with previously proposed SKCM subtypes³⁹ ($p = 5.6 \times 10^{-8}$, ANOVA F-test; **Fig. S4c-e**), i.e. melanocytic cell lines characterised by high MITF expression showed low EMT scores (**Fig. S4c-e**). To quantify their impact on responses to the three compounds, we added these SKCM subtypes to the MAK EMT score and MITF expression as fixed effects in a regression model predicting IC₅₀ values. Modelling subtypes improved predictions for staurosporine ($p = 0.0008$, ANOVA F-test for multiple regression coefficient; **Fig. S4c**), whilst we did not observe improvements for luminespib ($p = 0.31$, ANOVA F-test for multiple regression coefficient; **Fig. S4d**) or CHIR-99021 ($p = 0.24$, ANOVA F-test for multiple regression coefficient; **Fig. S4e**), thus highlighting the predictive capability of EMT in SKCM.

To gain further insights into the mechanisms of luminespib, CHIR-99021 and staurosporine, we extracted transcriptional signatures from the Library of Integrated Network-Based Cellular Signatures (LINCS)⁴⁰. We retrieved luminespib signatures of mesenchymal-like A375 and epithelial-like SK-MEL-28 SKCM cell lines and tested the 100 up- and down-regulated genes for enrichment in Gene Ontology (GO) biological processes (**Supplementary Data 5; Table 1; Methods**). The top process for both cells was the upregulation of genes involved in the response to unfolded proteins (A375: adj. $p = 4.9 \times 10^{-19}$, odds ratio = 130.34; SK-MEL-28: adj. $p = 6.0 \times 10^{-11}$, odds ratio = 68.09, Fisher's exact test; **Supplementary Data 5; Table 1**). Notably, genes involved in the regulation of TGF- β receptor signalling, such as SMAD3, were significantly downregulated among the top two enriched processes (A375: adj. $p = 0.0004$, odds ratio = 17.55, Fisher's exact test; SK-MEL-28: adj. $p = 0.02$, odds ratio = 8.81, Fisher's exact test; **Supplementary Data 5; Table 1**), suggesting that luminespib response may depend on TGF- β signalling components. Commonly downregulated genes of the CHIR-99021 signature included SMAD3 and PXN, which were also enriched in TGF- β receptor signalling (A375: adj. $p = 0.0006$, odds ratio = 123.73, Fisher's exact test; **Supplementary Data 5; Table 1**), whereas for the

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

staurosporine signature, TGF- β receptor signalling showed trends of enrichment (A375: adj. $p = 0.06$, odds ratio = 23.7, Fisher's exact test; **Supplementary Data 5**).

In summary, these results demonstrated that response to the HSP90 inhibitor luminespib in the GDSC HTS may depend on EMT components, their regulators MITF and TGF- β , and their associated subtypes. Therefore, the next section assessed the generalisation of EMT-dependent drug responses to other HSP90 inhibitors and their validation in independent HTS experiments.

EMT is associated with drug sensitivity to HSP90 inhibition in melanoma cell lines

For evaluating the robustness of EMT as a drug sensitivity biomarker to HSP90 inhibitors in SKCM, we tested for correlations between EMT scores and sensitivity to five HSP90 inhibitors across two high-throughput drug screens (**Methods; Supplementary Fig 5**). First, we assessed the IC₅₀ values of HSP90 inhibitors in the GDSC, here exemplified with tanespimycin ($r = -0.40$, $p = 0.036$, correlation test; cell lines with higher than mean *NQO1* expression⁴¹; **Fig. 3c**), elesclomol ($r = -0.34$, $p = 0.015$, correlation test; **Fig. 3d**), a luminespib replicate screened in both GDSC1 and GDSC2 ($r = -0.33$, $p = 0.021$, correlation test; **Supplementary Fig. 5a**) and SNX 2112 ($r = -0.21$, $p = 0.14$, correlation test; **Supplementary Fig. 5c**). Furthermore, consistent correlations were observed for AUC values across these HSP90 inhibitors in the GDSC (**Supplementary Fig. 5a-h**), thus highlighting the robustness of the association between EMT and responses to HSP90 inhibition regardless of the drug response readout.

To gain further evidence across independent datasets, we calculated the MAK EMT score based on gene expression data obtained from the Cancer Cell Line Encyclopedia (CCLE)⁴² and analysed the HTS of the Cancer Therapeutics Response Portal (CTRP)⁴³ (**Supplementary Data 6; Supplementary Fig 5i-m**). The AUC values of the screened HSP90 inhibitors SNX 2112 ($r = -0.44$, $p = 0.002$, correlation test; **Supplementary Fig. 5j**) and tanespimycin ($r = -0.47$, $p = 0.001$, correlation test; **Supplementary Fig. 5l**) were significantly associated with the EMT score in this independent HTS, and AT13387 (onalespib) displayed consistent trends ($r = -0.24$, $p = 0.221$, correlation test; **Supplementary Fig. 5m**).

In essence, EMT scores were consistently associated with drug sensitivity to HSP90 inhibition in SKCM cell lines across independent drug HTS and transcriptomic profiles (**Fig. 3e; Supplementary Fig. 5i-m**). For the next section, luminespib was selected as the lead compound for further experimental validation of our method, since it showed significant performance gains with the highest estimated EMT effects (**Table 1**).

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

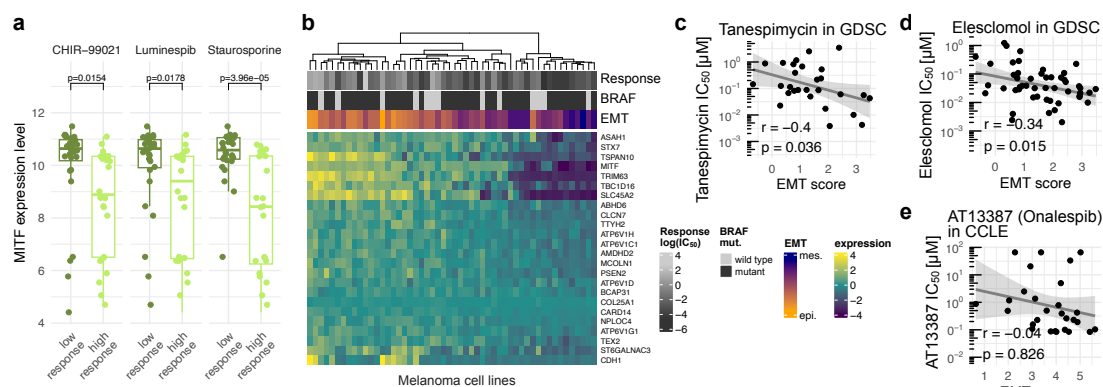


Figure 3: EMT is associated with transcription factors and susceptibility to HSP90 inhibitors. (a) Boxplots of binarised drug response (i.e. discretisation by median IC_{50} values) of CHIR-99021, luminespib and staurosporine show associations with *MITF* expression levels quantified by a two-sided *t*-test and its derived *p*-value. The centre on the boxplot represents the median, while the box illustrates the interquartile range (IQR). The whiskers show a range that is 1.5 times the size of the IQR. (b) A heatmap shows differentially expressed genes of luminespib drug response quantified by $\log(IC_{50})$ values in SKCM cell lines (FDR < 0.1) for the subset of *MITF* targets. Scatter plots show drug susceptibility of SKCM cell lines to other HSP90 inhibitors such as (c) tanespimycin, (d) elesclomol and (e) AT13387 (onalespib) in an independent dataset. The solid line depicts a fitted ordinary least squares regression model with its 95% confidence interval. The Pearson's correlation coefficient (*r*) and the associated *p*-value of the correlation test (*p*) are displayed.

TGF- β modulates the response to HSP90 inhibition with luminespib in epithelial-like melanoma cell lines

We conducted experiments on whether drug response to luminespib could be modulated by EMT induction. TGF- β is a known inducer of EMT¹⁷, which was also suggested by the upregulation of *TGFB1* expression in many mesenchymal-like SKCM cell lines (**Supplementary Fig. 6a-d**). Thus, we chose two epithelial-like cell lines (IGR-37, SK-MEL-5; **Fig. 4a**) and two mesenchymal-like cell lines (RPMI-7951, A375; **Fig. 4a**), which showed different levels of sensitivity to our lead compound luminespib in the GDSC, respectively. Following a 7-day pretreatment with TGF- β 1, we treated the cells with different concentrations of luminespib (**Supplementary Fig. 6e; Methods**) and fitted dose-response curves for each experiment to obtain IC_{50} and AUC values (**Supplementary Data 7; Methods**).

While the mesenchymal RPMI-7951 and A375 showed no distinguishable change in luminespib response upon TGF- β 1 treatment (**Fig. 4b,c**), the epithelial cell lines IGR-37 and SK-MEL-5 displayed increased luminespib sensitivity (**Fig. 4d,e**). To quantify this effect, we calculated the difference in $\log(IC_{50})$ values, i.e. $\Delta\log(IC_{50})$, for the screened cell lines and compared it to the 95% CI of the predicted causal effect upon change in the EMT score (**Methods**). Accordingly, the epithelial cell lines IGR-37 and SK-MEL-5 showed decreased IC_{50} within this CI (**Fig. 4f**). Analogously, we compared differences in AUC values (ΔAUC), which showed consistency within the CI of the predicted causal effect (**Supplementary Fig. 6f**). In summary, this highlights that EMT can modulate HSP90 inhibitor response in epithelial-like SKCM cell lines.

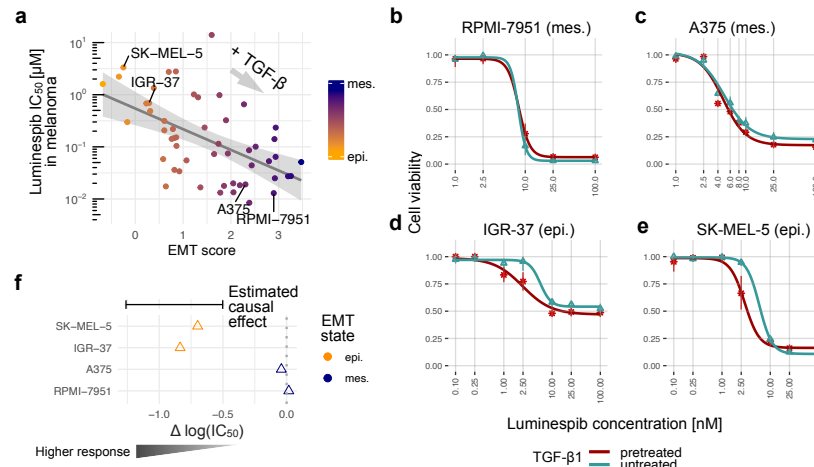


Figure 4: TGF- β sensitises melanoma cell lines to luminespib. (a) The scatter plot shows luminespib drug response stratified by EMT scores in SKCM cell lines and highlighted selected cell lines. The mesenchymal cell lines (b) RPMI-7951 and (c) A375 showed indistinguishable luminespib response upon TGF- β pretreatment. In contrast, the two epithelial cell lines (d) IGR-37 and (e) SK-MEL-5 responded stronger after pretreatment with TGF- β . Shown are the derived cell viabilities averaged across replicates and the fitted dose-response curves (**Methods**). (f) TGF- β treated epithelial cell lines demonstrate higher responses (decreased IC_{50}) to luminespib within the predicted causal effect 95% CI.

Discussion

We presented a drug response analysis encompassing the causal exploration of EMT in the context of mutational backgrounds and their upstream regulators and processes. We quantified EMT based on molecular biomarkers from gene expression profiles, thus offering a continuous score that accounts for the spectrum of intermediate and hybrid EMT states. By combining predictive and causal modelling, we identified five compounds across three cancer types with robust associations across different EMT scoring methods and drug response readouts (**Table 1**). Exemplifying our approach, we found that mesenchymal-like cell lines showed increased sensitivity to HSP90 inhibitors, particularly luminespib, which we experimentally validated.

Our pharmacogenomic modelling approaches corrected for confounders from the mutational background. Therefore, the estimated EMT effects from the causal modelling approach assumed no hidden confounders in the gene expression data. In order to address this, we performed *post hoc* differential gene expression analyses considering all genes to identify transcription factors as upstream regulators and GO biological processes. Furthermore, we mined drug transcriptional signatures to identify transcriptional confounders. Our analysis pursued the contribution of EMT on drug responses, however our systematic and causal modelling framework is generalisable to any putative drug response biomarker and its mechanisms.

We showed that epithelial-like cell lines can become more responsive to luminespib upon TGF- β treatment, whereas mesenchymal-like cell lines displayed no distinguishable change in their response.

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In addition, drug responses to luminespib quantified by IC₅₀ values in non-TGF- β treated cells were comparable to values observed in the GDSC. Therefore, the sensitivity of some melanoma cell lines to luminespib may indeed be induced by a phenotypic conversion of cell lines. Although molecular markers do not fully capture all intricacies of the EMT program, the sensitisation of epithelial-like cell lines upon TGF- β treatment to luminespib suggests that EMT markers with their regulator MITF in melanoma may constitute a promising biomarker for selectively targeting epithelial-mesenchymal transitioning cells.

HSP90 is an ATP-dependent molecular chaperone necessary for protein folding and stabilisation of oncogenic proteins including BRAF and TGF- β receptors^{44,45}. In melanoma, HSP90 levels have been found to correlate with melanoma progression metrics such as Breslow's depth and Clark level⁴⁶. The effect of HSP90 inhibition on cell viability seems to depend on MITF, which is a master regulator in melanoma cells that allow phenotype switching between distinct states ranging from arrested to highly invasive or highly proliferative phenotypes^{34,37,38}. TGF- β induces EMT across many cancer entities⁴⁷, and has inhibitory downstream effects on *MITF* expression^{48,49}. The sensitisation of epithelial cells via pretreatment with TGF- β suggests that TGF- β might regulate MITF in epithelial cells to allow switching to an invasive state, thereby rendering them more vulnerable to luminespib.

The exact mechanisms through which mesenchymal-like melanoma cell lines respond better to HSP90 inhibition remain elusive. They may be revealed by considering common mechanisms between the two compounds that were identified by our framework in conjunction with luminespib, namely the GSK3 β inhibitor CHIR-99021 and secondly, the non-selective multi-kinase inhibitor staurosporine. Potentially, the downregulation of TGF- β signalling might be the common link between these inhibitors.

In conclusion, we demonstrated that the pharmacogenomic assessment of EMT markers with predictive and causal modelling can predict drug susceptibilities and reveal relevant tumour biology in cancer cell lines. We anticipate that considering additional parameters of EMT-like phenotype transitions, such as cell morphology and proteomics, will increase mechanistic insights to EMT and its impact on drug responses. These and other types of follow-up studies may ultimately enable the selective targeting of transitioned cancer cells from the primary tumour or circulating tumour cells to prevent dissemination and metastasis.

Methods

Drug response data

The drug response data from the Genomics of Drug Sensitivity in Cancer (GDSC) was obtained from its release 8.4 under <https://ftp.sanger.ac.uk/project/cancerrxgene/releases/>. Both GDSC1 and GDSC2 datasets were used in this analysis, using the half maximal inhibitory concentration log(IC₅₀) and area under the curve (AUC) as metrics for quantifying drug responses. This resulted in 700 drug response profiles from 544 unique compounds. The Cancer Therapeutic Response Portal (CTRP) drug response data was downloaded from DepMap (<https://depmap.org/portal/>) contained in the file

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

'CTRPv2.0_2015_ctd2_ExpandedDataset.zip', which included 545 drug response profiles from 496 unique screened compounds.

Somatic mutations and copy number alterations

The GDSC project has previously compiled a selection of high-confidence cancer driver genes, including somatic mutations and copy number alterations, available under http://www.cancerrxgene.org/downloads/bulk_download. These binary matrices comprised the somatic mutational status for each identified genetic event for all cancer cell lines, thus characterising their genetic landscape. They contained the status of somatic mutations from 218 cancer genes and 802 copy number segments of 775 cancer cell lines across 31 cancer types.

Gene expression profiling and cancer subtypes

The GDSC RMA-processed Affymetrix array gene expression data was downloaded from https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Data/preprocessed/Cell_line_RMA_proc_basalExp.txt.zip, containing 781 cell lines for our investigated cancer types. The CCLE log₂ transformed RNA-seq gene expression data was downloaded from DepMap (<https://depmap.org/portal/>) contained in the file 'OmicsExpressionProteinCodingGenesTPMLogp1.csv' (22Q4). The BRCA (PAM50) ²⁹ and SKCM ³⁹ subtype annotations were obtained from the supplementary material of Jaaks *et al.* (Table S2) ⁵⁰ and Warren *et al.* ⁵¹, respectively.

EMT scores

We quantified EMT in 27 cancer types that had > 5 cancer lines available using four established methods, i.e. Mak *et al.* ⁸ (MAK), gene set variation analysis ²¹ (GSVA), Tan *et al.* ⁷ (TAN) and Tagliacruzchi and Wiecek *et al.* ¹⁰ (TW). For the MAK EMT score, we ranked genes based on their Pearson's correlation to four EMT marker genes, i.e. *CDH1*, *CDH2*, *VIM* and *FN1*. The genes were then ordered by their respective correlation coefficients and the top 25 genes highly correlated to *CDH1* expression were selected as 'epithelial' marker genes, whereas the top 25 genes that were highly correlated to each respective mesenchymal gene were grouped as 'mesenchymal' markers, resulting in EMT gene signatures comprised of all unique genes from the 25 epithelial and 75 mesenchymal markers for each cancer type. For each cell line, the EMT score was then calculated by the difference in mean expression levels of mesenchymal and epithelial markers.

For the GSVA EMT score, the 'msgdbr' R package was used to queue gene sets for the subsequent gene set variation analysis using the 'GSVA' R package, which yielded gene set enrichment scores from the EMT gene set for each cell line ⁹. For the TAN EMT score, we downloaded the provided tables in their supplementary material (Table S4C ⁷) and extracted the scores from their set of cancer cell lines. Similarly, for the TW EMT scores, we used the provided supplementary tables in their supplementary material to extract scores (Source data ¹⁰). All EMT scores are supplied in **Supplementary Data 1** and **Supplementary Data 6**.

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Predictive modelling and ablation study

Drug responses denoted by y and quantified by $\log(\text{IC}_{50})$ and AUC values were modelled by the EMT score s and somatic alterations $x = \{x_1, \dots, x_m\}$, consisting of m binary encoded genetic alterations. The model was specified by

$$y = \alpha + \beta x + \gamma s + e, \quad (1)$$

with intercept α , confounder coefficients β , EMT coefficient γ and the error term e . The python package 'sklearn' was used to fit the regression model with lasso penalty and cross-validation for choosing the optimal penalty hyperparameter for each compound, cancer type and EMT score separately. For benchmarking the model performance, we performed 5-fold cross-validation with 5 repetitions. The Pearson's correlation (r) between predicted and ground truth response was calculated on the test set to quantify model performance for each of the five folds and five initialisations.

An ablation study was conducted to prioritise drugs and cancer types for which EMT is suggested to contribute to the drug response phenotype. It was performed by refitting the models with removed EMT score s and recording its performance with the same splits. To compare the model performances of the full versus the model with ablation of s , a t -test for resampled performance metrics was used for assessing significant decreases of Pearson's r across all the performances from the 25 models⁵². The resulting p -values were corrected for multiplicity using the Benjamini-Hochberg false discovery rate (FDR) method⁵³ for each cancer type and EMT score separately. We found 32 compounds with FDR < 0.2 across six cancer types. In the main manuscript, we focused on five compounds that showed robustly significant performance differences in at least three out of eight possible models (4 EMT scores \times 2 response readouts).

We only performed modelling if at least 25 cell lines for a given cancer type and drug were observed in the screening experiment. Furthermore, for modelling IC_{50} values, we did not consider models for which $> 70\%$ of IC_{50} values for a given drug and cancer type were extrapolated considerably beyond the maximum tested concentration c_{\max} , i.e. $\text{IC}_{50} > 2c_{\max}$. The full results are supplied in **Supplementary Data 2**.

Causal modelling

Double machine learning (DML) is often used for estimating treatment effects on observed outcomes. It consists of two stages, (1) learning the propensity and outcome models as nuisance functions to extract their residuals, and (2) regressing outcome residuals on treatment residuals to obtain valid treatment effects and confidence intervals (CI)^{25,26}. Accordingly, we estimated the causal component of EMT by fitting a causal forest²⁷ in conjunction with DML for each drug, cancer type and EMT score, implemented in the CausalForestDML method within the python package 'econml'²⁸. The two nuisance functions were fitted using the same lasso regression model as used above. Thereby, we modelled the drug responses as outcome y to estimate the effect of the EMT score s as a continuous variable in the presence of the mutational background as confounders x . The estimated effect (EMT effect) then assesses the impact of undergoing EMT via non-mutational tumour plasticity on drug response. Since

the EMT scores are continuous, the effect was given per unit of EMT change, i.e. for the interval of one standard deviation from the distribution of EMT scores for each cancer type. This effect and its 95% CI was compared with the validation experiments. The full results are supplied in **Supplementary Data 2**.

Transcription factor and gene ontology enrichments

We sought to identify enrichments of genes correlated to drug responses in transcription factor (TF) targets and Gene Ontology (GO) biological processes from the transcriptional background of cancer cell lines. For a given drug response and transcriptomic profile within a given cancer type, we performed differential gene expression between continuous drug responses using linear models implemented in the 'limma' R package. The differentially expressed genes (FDR < 0.1) were then used as query genes for a gene set enrichment analysis with the 'enrichR' R package, for which we tested gene sets consisting of curated TF target genes⁵⁴ as potential upstream regulators of EMT and biological processes in the GO knowledge base⁵⁵. We only considered the gene set positively correlated with drug response and its top enriched TF and the top two enriched GO terms by their adjusted *p*-values including ties in **Table 1**, while the full results are supplied in **Supplementary Data 3** for TFs and **Supplementary Data 4** for GO terms.

LINCS transcriptional signatures

The transcriptional signatures of the Library of Integrated Network-Based Cellular Signatures (LINCS) program contain sets of genes with up- and down-regulated gene expression levels upon chemical or genetic perturbations⁴⁰. Using the CLUE knowledge base (<https://clue.io/lincs>) and its provided API, we retrieved the signatures of luminespib for two SKCM cell lines, i.e. mesenchymal-like A375 and epithelial-like SK-MEL-28. We aggregated the 100 up- and down-regulated genes from all available signatures for cell lines. Then, we used these genes as a query for a gene set enrichment analysis with the 'enrichR' R package for each cell line to test for enrichments of GO biological processes. For staurosporine and CHIR-99021, only mesenchymal-like A375 cells were available. We used the overlapping signature genes of the transcriptional signatures of luminespib in A375 cells as a query for the same enrichment analysis in order to check for common mechanisms between the three compounds. We only considered the top two enriched GO terms by their adjusted *p*-values including ties in **Table 1**, while the full results are supplied in **Supplementary Data 5**.

Cell culture

SK-MEL-5 (source: ATCC), A375 (source: ATCC), RPMI-7951 (source: DSMZ) were cultured in Gibco Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% Fetal Bovine Serum (FBS) and 1% Penicillin-Streptomycin (Pen-strep) (10000 U/mL). IGR-37 (source: DSMZ) was cultured in Gibco DMEM supplemented with 15% FBS 1% Pen-strep. To induce EMT based on previous literature¹⁷, the media were supplemented with 5 ng/mL TGF- β 1 (R&D Systems 7754-BH/CF) for 7 days.

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Luminespib treatment

10,000 cells in 100 µL medium per well were seeded in 96-well opaque, white, flat-bottom plates. After the cells were allowed to attach at 37 °C, luminespib (Selleck-Chem: S1069) dissolved in DMSO was added into the wells at the indicated concentrations and 0.5% DMSO. The negative control wells were treated with 0.5% DMSO alone, whereas the blank wells contained only the media. The plates were incubated for 72h before the CellTiter-Glo® 2.0 Cell Viability Assay (Promega: G924A) was performed following the manufacturer's protocol. Luminescence was measured using the Perkin Elmer EnVision 2104 Multilabel Plate Reader.

Dose-response analysis

Cell viability (v) (capped between 0 and 1) was calculated with intensities from blank (I_B), negative control (I_{NC}) and luminespib treatment (I_L) wells with

$$v = \frac{I_L - I_B}{I_{NC} - I_B}. \quad (2)$$

Dose-response curves were fitted and IC₅₀ values were calculated using the four-parameter log-logistic (LL.4) model in the R package 'drc' ⁵⁶ and AUC values were calculated using the R package 'PharmacoGx' ⁵⁷. The results are supplied in **Supplementary Data 7**.

Code accessibility

The source code for the presented analysis is available at <https://github.com/mendenlab/emtpb>.

Author Contributions

Conceptualization, A.J.O. and M.P.M.; Data curation, A.J.O., G.A., M.K.S. and T.J.O.; Analysis, A.J.O., G.A. and M.K.S.; Methodology, A.J.O., G.A., M.K.S. and M.P.M.; Supervision, D.K. and M.P.M.; Visualisation, A.J.O. and G.A.; Writing original draft, A.J.O., M.K.S., G.A. and M.P.M.; Writing, review and editing, all authors.

Competing Interests

M.P.M. collaborates with GSK, Roche and AstraZeneca, and receives funding from Roche and GSK. M.P.M. is a former employee at AstraZeneca. The remaining authors declare no competing interest.

References

1. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* **119**, 1420–1428 (2009).
2. Yang, J. *et al.* Guidelines and definitions for research on epithelial–mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* **21**, 341–352 (2020).
3. Hüsemann, Y. *et al.* Systemic spread is an early step in breast cancer. *Cancer Cell* **13**, 58–68 (2008).

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4. Thiery, J. P., Acloque, H., Huang, R. Y. J. & Nieto, M. A. Epithelial-mesenchymal transitions in development and disease. *Cell* **139**, 871–890 (2009).
5. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).
6. Park, J. & Schwarzbauer, J. E. Mammary epithelial cell interactions with fibronectin stimulate epithelial-mesenchymal transition. *Oncogene* **33**, 1649–1657 (2014).
7. Tan, T. Z. *et al.* Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.* **6**, 1279–1293 (2014).
8. Mak, M. P. *et al.* A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. *Clin. Cancer Res.* **22**, 609–620 (2016).
9. Vasaikar, S. V. *et al.* EMTome: a resource for pan-cancer analysis of epithelial-mesenchymal transition genes and signatures. *Br. J. Cancer* **124**, 259–269 (2021).
10. Malagoli Tagliazucchi, G., Wiecek, A. J., Withnell, E. & Secrier, M. Genomic and microenvironmental heterogeneity shaping epithelial-to-mesenchymal trajectories in cancer. *Nat. Commun.* **14**, 789 (2023).
11. Miettinen, P. J., Ebner, R., Lopez, A. R. & Derynck, R. TGF- β induced transdifferentiation of mammary epithelial cells to mesenchymal cells: involvement of type I receptors. *J. Cell Biol.* **127**, 2021–2036 (1994).
12. Bhowmick, N. A. *et al.* Transforming growth factor- β 1 mediates epithelial to mesenchymal transdifferentiation through a RhoA-dependent mechanism. *Mol. Biol. Cell* **12**, 27–36 (2001).
13. Kasai, H., Allen, J. T., Mason, R. M., Kamimura, T. & Zhang, Z. TGF- β 1 induces human alveolar epithelial to mesenchymal cell transition (EMT). *Respir. Res.* **6**, 56 (2005).
14. Gao, J., Zhu, Y., Nilsson, M. & Sundfeldt, K. TGF- β isoforms induce EMT independent migration of ovarian cancer cells. *Cancer Cell Int.* **14**, 72 (2014).
15. Li, S. *et al.* Transcriptional regulation of autophagy-lysosomal function in BRAF-driven melanoma progression and chemoresistance. *Nat. Commun.* **10**, 1693 (2019).
16. Kim, B. N. *et al.* TGF- β induced EMT and stemness characteristics are associated with epigenetic regulation in lung cancer. *Sci. Rep.* **10**, 10597 (2020).
17. Schlegel, N. C., von Planta, A., Widmer, D. S., Dummer, R. & Christofori, G. PI3K signalling is

- required for a TGF β -induced epithelial-mesenchymal-like transition (EMT-like) in human melanoma cells. *Exp. Dermatol.* **24**, 22–28 (2015).
18. Shibue, T. & Weinberg, R. A. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. *Nat. Rev. Clin. Oncol.* **14**, 611–629 (2017).
 19. Byers, L. A. *et al.* An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clin. Cancer Res.* **19**, 279–290 (2013).
 20. Adachi, Y. *et al.* Epithelial-to-Mesenchymal Transition is a Cause of Both Intrinsic and Acquired Resistance to KRAS G12C Inhibitor in KRAS G12C-Mutant Non-Small Cell Lung Cancer. *Clin. Cancer Res.* **26**, 5962–5973 (2020).
 21. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
 22. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
 23. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).
 24. Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–61 (2013).
 25. Chernozhukov, V. *et al.* Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21**, C1–C68 (2018).
 26. Foster, D. J. & Syrgkanis, V. Orthogonal statistical learning. *aos* **51**, 879–908 (2023).
 27. Athey, S., Tibshirani, J. & Wager, S. Generalized random forests. *aos* **47**, 1148–1178 (2019).
 28. Battocchi, K., Dillon, E., Hei, M., Lewis, G. & Oka, P. EconML: A Python package for ML-Based heterogeneous treatment effects estimation. *Version 0. x*.
 29. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
 30. Yarden, R. I., Pardo-Reoyo, S., Sgagias, M., Cowan, K. H. & Brody, L. C. BRCA1 regulates the G2/M checkpoint by activating Chk1 kinase upon DNA damage. *Nat. Genet.* **30**, 285–289 (2002).
 31. Zhang, P. *et al.* ATM-mediated stabilization of ZEB1 promotes DNA damage response and radioresistance through CHK1. *Nat. Cell Biol.* **16**, 864–875 (2014).

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

32. Mamun, M. A., Mannoor, K., Cao, J., Qadri, F. & Song, X. SOX2 in cancer stemness: tumor malignancy and therapeutic potentials. *J. Mol. Cell Biol.* **12**, 85–98 (2018).
33. Riento, K. & Ridley, A. J. ROCKs: multifunctional kinases in cell behaviour. *Nat. Rev. Mol. Cell Biol.* **4**, 446–456 (2003).
34. Garraway, L. A. *et al.* Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117–122 (2005).
35. Goding, C. R. Commentary. A picture of Mitf in melanoma immortality. *Oncogene* vol. 30 2304–2306 (2011).
36. Carreira, S. *et al.* Mitf regulation of Dia1 controls melanoma proliferation and invasiveness. *Genes Dev.* **20**, 3426–3439 (2006).
37. Kaur, A., Webster, M. R. & Weeraratna, A. T. In the Wnt-er of life: Wnt signalling in melanoma and ageing. *Br. J. Cancer* **115**, 1273–1279 (2016).
38. Arozarena, I. & Wellbrock, C. Phenotype plasticity as enabler of melanoma progression and therapy resistance. *Nat. Rev. Cancer* **19**, 377–391 (2019).
39. Tsoi, J. *et al.* Multi-stage Differentiation Defines Melanoma Subtypes with Differential Vulnerability to Drug-Induced Iron-Dependent Oxidative Stress. *Cancer Cell* **33**, 890–904.e5 (2018).
40. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437–1452.e17 (2017).
41. Menden, M. P. *et al.* The germline genetic component of drug sensitivity in cancer cell lines. *Nat. Commun.* **9**, 3385 (2018).
42. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
43. Seashore-Ludlow, B. *et al.* Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov.* **5**, 1210–1223 (2015).
44. da Rocha Dias, S. *et al.* Activated B-Raf is an Hsp90 client protein that is targeted by the anticancer drug 17-allylamino-17-demethoxygeldanamycin. *Cancer Res.* **65**, 10686–10691 (2005).
45. Taipale, M. *et al.* Quantitative analysis of HSP90-client interactions reveals principles of substrate recognition. *Cell* **150**, 987–1001 (2012).

bioRxiv preprint doi: <https://doi.org/10.1101/2024.01.16.575190>; this version posted January 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

46. McCarthy, M. M. *et al.* HSP90 as a marker of progression in melanoma. *Ann. Oncol.* **19**, 590–594 (2008).
47. Xu, J., Lamouille, S. & Derynck, R. TGF-beta-induced epithelial to mesenchymal transition. *Cell Res.* **19**, 156–172 (2009).
48. Kim, D.-S., Park, S.-H. & Park, K.-C. Transforming growth factor-beta1 decreases melanin synthesis via delayed extracellular signal-regulated kinase activation. *Int. J. Biochem. Cell Biol.* **36**, 1482–1491 (2004).
49. Wellbrock, C. & Arozarena, I. Microphthalmia-associated transcription factor in melanoma development and MAP-kinase pathway targeted therapy. *Pigment Cell Melanoma Res.* **28**, 390–406 (2015).
50. Jaaks, P. *et al.* Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature* **603**, 166–173 (2022).
51. Warren, A. *et al.* Global computational alignment of tumor and cell line transcriptional profiles. *Nat. Commun.* **12**, 22 (2021).
52. Bouckaert, R. R. & Frank, E. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. in *Advances in Knowledge Discovery and Data Mining* 3–12 (Springer Berlin Heidelberg, 2004).
53. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
54. Keenan, A. B. *et al.* ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res.* **47**, W212–W224 (2019).
55. Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, (2023).
56. Ritz, C., Baty, F., Streibig, J. C. & Gerhard, D. Dose-Response Analysis Using R. *PLoS One* **10**, e0146021 (2015).
57. Smirnov, P. *et al.* PharmacGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics* **32**, 1244–1246 (2016).

2.3 The Oncology Biomarker Discovery framework reveals cetuximab and bevacizumab response patterns in metastatic colorectal cancer, *Nature Communications* (2023)

This article was peer-reviewed and published open-access in *Nature Communications* [3] and is reproduced with permission from Springer Nature. It is publicly available at <https://doi.org/10.1038/s41467-023-41011-4>.

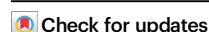


The Oncology Biomarker Discovery framework reveals cetuximab and bevacizumab response patterns in metastatic colorectal cancer

Received: 4 March 2022

Accepted: 17 August 2023

Published online: 04 September 2023



Alexander J. Ohnmacht^{1,2,10}, Arndt Stahler^{3,10}, Sebastian Stintzing^{1,3,4,10}, Dominik P. Modest^{1,3}, Julian W. Holch^{4,5}, C. Benedikt Westphalen⁵, Linus Hölzel¹, Marisa K. Schübel^{1,2}, Ana Galhoz^{1,2}, Ali Farnoud¹, Minhaz Ud-Dean¹, Ursula Vehling-Kaiser⁶, Thomas Decker⁷, Markus Moehler⁸, Matthias Heinig¹, Volker Heinemann⁵ ✉ & Michael P. Menden^{1,2,9} ✉

Precision medicine has revolutionised cancer treatments; however, actionable biomarkers remain scarce. To address this, we develop the Oncology Biomarker Discovery (OncoBird) framework for analysing the molecular and biomarker landscape of randomised controlled clinical trials. OncoBird identifies biomarkers based on single genes or mutually exclusive genetic alterations in isolation or in the context of tumour subtypes, and finally, assesses predictive components by their treatment interactions. Here, we utilise the open-label, randomised phase III trial (FIRE-3, AIO KRK-0306) in metastatic colorectal carcinoma patients, who received either cetuximab or bevacizumab in combination with 5-fluorouracil, folinic acid and irinotecan (FOLFIRI). We systematically identify five biomarkers with predictive components, e.g., patients with tumours that carry chr20q amplifications or lack mutually exclusive ERK signalling mutations benefited from cetuximab compared to bevacizumab. In summary, OncoBird characterises the molecular landscape and outlines actionable biomarkers, which generalises to any molecularly characterised randomised controlled trial.

Precision medicine aims to tailor therapeutic interventions to specific patient subgroups defined by predictive biomarkers detected in tumours. Accordingly, strategies are required to identify such patient subgroups systematically¹. For performing subgroup analysis and

exploratory biomarker discovery, the European Medicines Agency (EMA) has provided specific guidelines². According to these, biological knowledge should underpin subgroup definitions, and subgroup-specific effects in late-stage clinical trials should still be interpreted

¹Computational Health Center, Helmholtz Munich, 85764 Neuherberg, Germany. ²Department of Biology, Ludwig-Maximilians University Munich, 82152 Martinsried, Germany. ³Charité Universitätsmedizin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Hematology, Oncology, and Cancer Immunology, Charitéplatz 1, 10117 Berlin, Germany. ⁴German Cancer Consortium (DKTK), partner sites Berlin and Munich, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. ⁵Department of Medicine III and Comprehensive Cancer Center Munich, University Hospital, Ludwig-Maximilians University Munich, 81377 Munich, Germany. ⁶Oncological Practice, 84028 Landshut, Germany. ⁷Oncological Practice, 88212 Ravensburg, Germany. ⁸Department of Medicine I and Research Center for Immunotherapy (FZI), Johannes Gutenberg-University Clinic, 55131 Mainz, Germany. ⁹Department of Biochemistry and Pharmacology, University of Melbourne, Victoria 3010, Australia. ¹⁰These authors contributed equally: Alexander J. Ohnmacht, Arndt Stahler, Sebastian Stintzing. ✉ e-mail: volker.heinemann@med.uni-muenchen.de; michael.menden@helmholtz-munich.de

with caution owing to the exploratory and retrospective nature of the analyses. For this purpose, a large number of computational methods have been proposed and discussed^{3–5}, e.g., tree-based methods using recursive partitioning^{6–8}, virtual twins⁹, outcome weighted methods^{10,11}, causal forests¹² and metalearners for estimating heterogeneous treatment effects¹³. However, most of these computational methods neglect cancer biology, i.e., exploiting the molecular landscape of a clinical trial and customising models to cancer subtypes and mutational patterns.

Clinical outcomes of patients with metastatic colorectal cancer (mCRC) significantly improved upon the introduction of targeted treatments, including anti-EGFR and anti-VEGF directed monoclonal antibodies such as cetuximab and bevacizumab, respectively¹⁴. Tumours of colorectal cancer patients were shown to exhibit, for instance, either *KRAS* or *NRAS* mutations (referred to as *RAS* mutations) with a rate of about 50%, which tend to occur mutually exclusive^{15,16}. These *RAS* mutations are clinically approved predictive biomarkers of resistance against anti-EGFR directed monoclonal antibodies such as cetuximab¹⁷. Bevacizumab has been reported to improve progression-free survival in first-line mCRC trials¹⁸; however, no comparable biomarker has been depicted yet.

In this study, we focused on the open-label randomised phase III clinical trial FIRE-3. Here, patients with *KRAS* exon 2 wild-type mCRC were randomised to receive either cetuximab or bevacizumab in combination with 5-fluorouracil, leucovorin and irinotecan (FOLFIRI) as a first-line regimen. Several retrospective subgroup analyses revealed potential prognostic and predictive biomarkers based on tumour DNA and clinical characteristics, such as the relevance of the molecular status, i.e., alterations other than *KRAS* exon 2, such as *KRAS* exon 3–4, *NRAS* exon 2–4 and BRAF V600E, or primary tumour sidedness^{19–23}. For example, targeting EGFR in *RAS* wild-type mCRC tumours located in the left hemicolon (left-sided) was shown to be beneficial, whilst *RAS* wild-type tumours located in the right colon (right-sided) were less likely to respond²⁴. Additionally, in the more recent FIRE-4.5 study, it was demonstrated that patients with BRAF V600E mutant tumours may benefit from the treatment with 5-fluorouracil, oxaliplatin, leucovorin and irinotecan (FOLFOXIRI) backbone plus bevacizumab²⁵, whereas in contrast, these patients lacked benefits from cetuximab^{26,27}. This hints towards tumour subtype-specific interactions and alternative mechanisms to acquire EGFR inhibitor resistance²⁸.

Previously proposed tumour subtypes in colorectal adenocarcinoma are based on the gene expression-derived consensus molecular subtypes (CMS) and could identify subtypes that reflected distinct tumour biology²⁹. Recently, the prognostic value of CMS has been confirmed in the FIRE-3, CALGB/SWOG 80405 and AGITG MAX clinical trials for FOLFIRI combined with either cetuximab or bevacizumab^{21,30,31}. In particular, CMS4 patients with *RAS* wild-type have shown a significantly longer overall survival when treated with cetuximab compared to bevacizumab in metastatic disease²¹. However, the clinical translation of the CMS classification of colorectal cancer is still in its infancy and is further investigated in multiple clinical trials³². These sparse results have illustrated that modelling interactions between somatic alterations and tumour subtypes can yield insights into complex biomarkers and highlight the urgent need for computational frameworks to systematically decipher the molecular landscape, tumour subtypes and biomarkers. Thus, we hypothesised that predictive response biomarkers may be revealed by systematically deconvoluting cancer genetic events and tumour subtypes within a clinical trial.

Here, we present the Oncology Biomarker Discovery (OncoBird) framework, which empowers the systematic identification of actionable biomarkers for clinical trials in oncology. OncoBird is publicly available as a software package at <https://github.com/MendenLab/OncoBird> and a demo run is available at <https://codeocean.com/capsule/9911222/tree/v1>. Furthermore, users can run a graphical user interface within a docker container (Supplementary Fig. 1).

The OncoBird workflow is divided into five distinct steps: it systematically (1) investigates the molecular landscape of a clinical trial, i.e., copy number alterations, somatic mutations, mutually exclusive patterns and predefined tumour subtypes; (2) identifies biomarkers within a treatment arm based on genetic alterations, and (3) in relation to the predefined tumour subtypes; consecutively, (4) evaluates their predictive component across treatment arms; and finally, (5) it comprehensively corrects for multiple hypothesis testing and adjusts treatment effects of biomarkers based on resampling methods. To enhance the biological signal, this analysis integrates the molecular and biomarker landscape of cancer clinical trials by customising models to established cancer subtypes and mutational patterns. In essence, OncoBird yields subtype-specific biomarkers with treatment benefits in an interpretable and transparent manner and therefore operates complementary to existing methods. The utility of OncoBird is exemplified by the application to the FIRE-3 clinical trial, generalises to the ADJUVANT clinical trial^{33–35}, and in fact, would generalise to any molecularly characterised randomised controlled trial (RCT) in oncology.

Results

OncoBird is applicable to RCTs accompanied with molecular characteristics, including genetic sequencing panels which yield copy number alterations and somatic driver mutations (Fig. 1a, b). In addition, a second layer of stratification can be supplied in the form of predefined tumour subtypes (Fig. 1a). Then, OncoBird systematically assesses the genetic landscape in the context of tumour subtypes (Fig. 1c) and outlines the biomarker landscape across multiple clinical responses (Fig. 1d), i.e., time-to-event data (overall or progression-free survival; “Methods”), and binary variables capturing treatment success (objective response rate; “Methods”).

Here, we leveraged the FIRE-3 RCT, including 752 mCRC patients who have been treated with FOLFIRI and either cetuximab or bevacizumab. We defined tumour subtypes based on CMS²¹, and tumour sidedness, i.e., left- or right-sided mCRC. In addition, 373 tumours were genetically characterised, i.e., the mutational status of 277 frequently altered cancer genes. To reveal the biomarker landscape, we employed the following stratification and modelling strategies (Supplementary Data 1; “Methods”): We first investigated each alteration for stratifying patients by their prognosis within each treatment arm (Fig. 1e). Consecutively, we inspected alterations in tumour subtypes (Fig. 1f), revealing subtype-specific biomarkers. Finally, we tested for treatment interactions to reveal biomarkers with predictive effects (Fig. 1g). Importantly, subtypes and genetic alterations ought to be independent of the treatment assignment. The molecular landscape and individual treatment arm analysis could be applied to any trial design without limitations.

Exemplified with a well-established biomarker of cetuximab response¹⁷, *RAS* wild-type mCRC patients showed longer overall survival (Fig. 1h; $p = 0.0002$, HR = 0.53 [0.38–0.73]). Consistent with a previous study³⁶ and more recently defined treatment guidelines for mCRC³⁷, the cetuximab overall survival (OS) benefit for patients with *RAS* wild-type tumours was conserved in left-sided tumours (Fig. 1i; $p = 7.6 \times 10^{-5}$, HR = 0.44 [0.29–0.66]). Furthermore, we observed interactions between *RAS* mutations and the treatment arm in left-sided tumours ($p_{\text{int}} = 0.07$): Cetuximab remained superior to bevacizumab in *RAS* wild type and left-sided tumours (Fig. 1j; $p = 0.05$, HR = 0.73 [0.52–1.00]) in terms of OS, whilst bevacizumab and cetuximab achieved comparable OS for patients with *RAS* mutant and left-sided tumours (Supplementary Fig. 2; $p = 0.32$, HR = 1.22 [0.85–1.75]).

Whilst we particularly focused on the FIRE-3 trial in colorectal cancer, we also demonstrate the generalisability of OncoBird by applying it with the same default biomarker thresholds to the

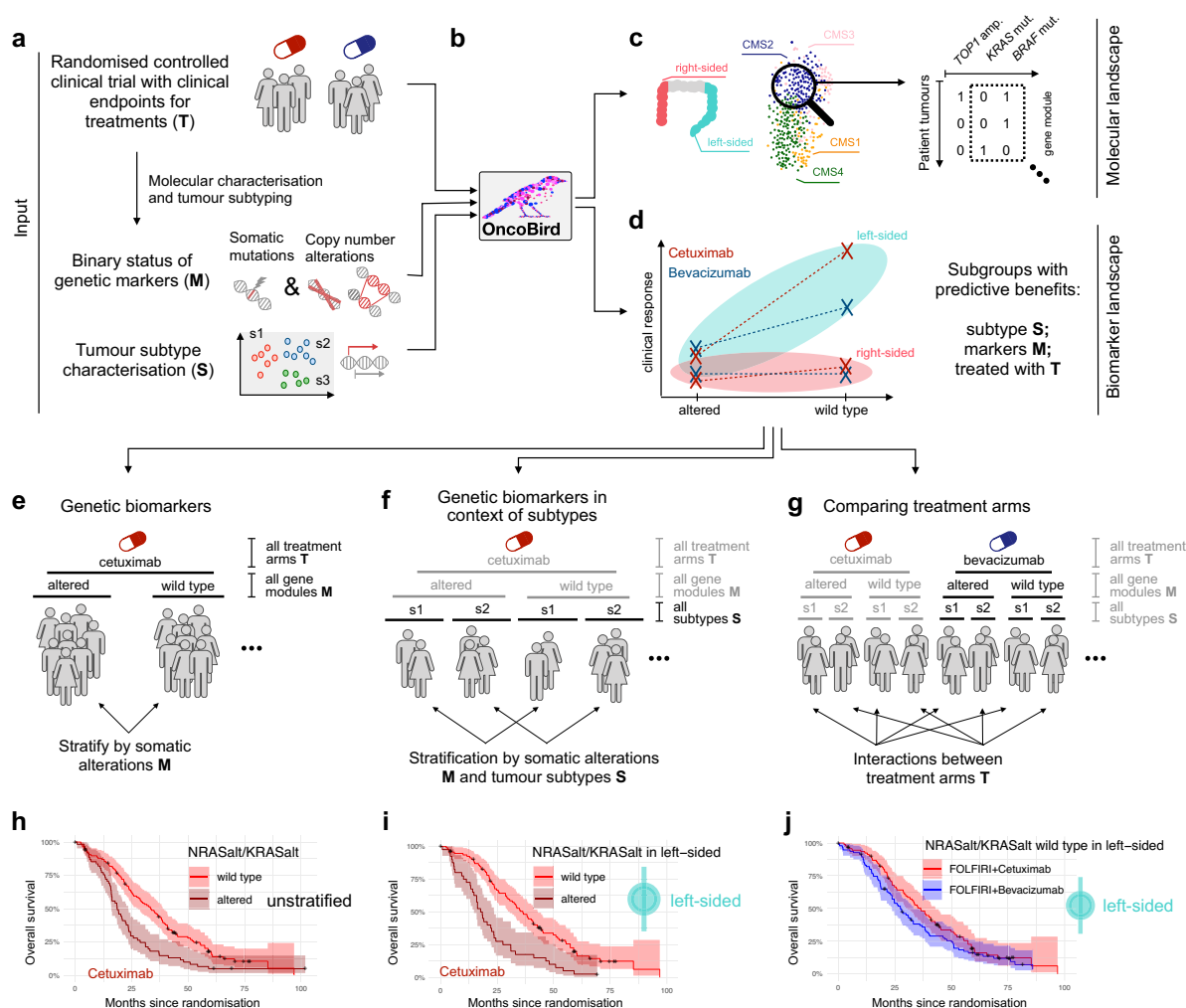


Fig. 1 | The Oncology Biomarker Discovery (OncoBird) workflow. **a** Patients in clinical trials were treated with (T) two treatment regimens with measured clinical endpoints. Subsequently, their tumours are characterised according to (M) tumour genetic alterations (somatic mutations and copy number alterations) and (S) tumour subtypes. **b** With this input, OncoBird outlines **c** the molecular landscape and **d** the biomarker landscape. For the latter, **e** somatic alterations are explored for a differential patient prognosis for each treatment arm. **f** Consecutively, for each treatment arm, subtype-specific biomarkers are derived. **g** Finally, interactions

between treatment arms are examined. The grey shadings indicate the data included in the previous analysis step. Here, this is exemplified in the FIRE-3 clinical trial using Kaplan–Meier plots, including 95% confidence intervals (CI) and summary statistics of the Cox regression models. **h** RAS mutations are established biomarkers of cetuximab resistance. **i** Patients with RAS wild-type tumours showed a better prognosis when treated with cetuximab within left-sided tumours compared to right-sided tumours. In addition, **j** the RAS wild-type subpopulation in left-sided tumours showed benefits when treated with cetuximab compared to bevacizumab.

ADJUVANT clinical trial (“Methods”), which explored gefitinib in non-small cell lung cancer (NSCLC)^{33–35}. The ADJUVANT study reported predictive components of five alterations, i.e., *TP53* mutations, *RB1* alterations and copy number amplifications of *NKX2-1*, *CDK4* and *MYC*³⁵. Four out of five biomarkers were concordantly identified for disease-free survival with OncoBird ($FDR_{int} < 0.2$; Supplementary Data 2; Supplementary Fig. 3–6). In addition, OncoBird suggests that the mutual exclusivity patterns play a role in the biomarker landscape of NSCLC (Supplementary Fig. 3c, d). In more detail, we observed gefitinib benefits in tumours that were characterised by mutations in either *TP53*, *SMAD4* or *CDK4* amplifications ($p = 0.0002$, $HR = 0.37$ [0.21–0.63]; Supplementary Data 2; Supplementary Figs. 5c and 6a), for which the resampling-based adjustment of the conditional average treatment effect yielded $p_{adj} = 0.001$ with $HR = 0.32$ [0.14–0.86] (Supplementary Data 2; “Methods”). These findings highlight the accessibility, reproducibility and interoperability of OncoBird.

The molecular landscape of the FIRE-3 clinical trial

Leveraging OncoBird, we assessed the genetic landscape of patient tumours in the FIRE-3 clinical trial. In total, 373 tumours were genetically characterised, including 31 frequently altered cancer genes observed in at least 12 patients (Fig. 2a). We observed amplifications in chromosome arm 20q (chr20q) in 74/373 tumours (19.8%), which includes *SRC*, *TOP1*, *BCL2L1*, *ZNF217*, *AURKA*, *GNAS* and *ARFRP1* (Fig. 2a). Indeed, chr20q amplifications have been reported to define a distinct subtype of left-sided colon cancers³⁸. In addition, we identified 39 mutually exclusive somatic alterations (gene modules) using the Mutex algorithm (Fig. 2b; “Methods”) ³⁹, thus grouping low frequent but functionally similar somatic events within a signalling pathway. We could confirm that chr20q amplifications were mutually exclusive to somatic mutations in the ERK signalling pathway (*KRAS*, *NRAS* or *BRAF*; $p = 0.0002$, Fisher’s exact test).

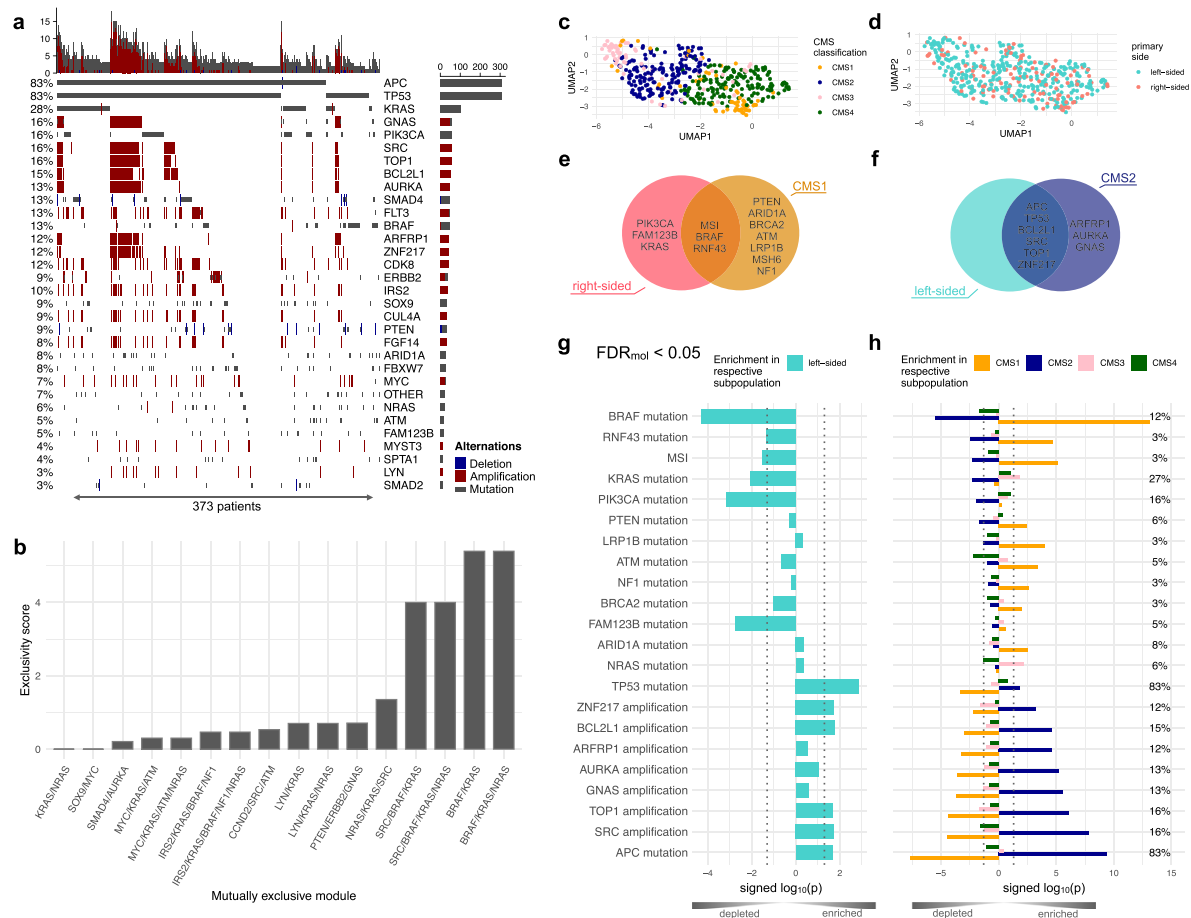


Fig. 2 | Molecular landscape of the FIRE-3 clinical trial. a Oncoprint of 373 mCRC tumours, including mutations and copy number alterations detected in more than 12 tumours. **b** The mutually exclusive alteration patterns were derived with the Mutex algorithm. Gene expression profiles of 451 mCRC tumours are annotated by **c** the consensus molecular subtypes (CMS) and **d** the primary tumour side. **e** Venn

diagram showing all enriched somatic alterations for CMS1 and right-sided tumours, and **f** enriched somatic alterations for CMS2 and left-sided tumours. **g** Frequently altered cancer genes tested for enrichment in left- or right-sided tumours, and **h** tested against CMS subtypes using one-sided hypergeometric tests. Source data for the figure panels are provided as Source Data file.

In addition, we analysed 451 gene expression profiles and showed consistency with their derived CMS subtypes (Fig. 2c), whilst the primary tumour side displayed a heterogeneous gene expression pattern (Fig. 2d). Right-sided tumours were particularly enriched in CMS1 tumours ($p = 0.009$, hypergeometric test; Supplementary Fig. 7) and depleted in CMS2 tumours ($p = 0.007$, hypergeometric test; Supplementary Fig. 7).

The concordance between right-sided tumours and CMS1 (Fig. 2e) was reflected by genetic alterations that were enriched in both tumour subtypes. Microsatellite instabilities (MSI) and somatic mutations in *BRAF* and *RNF43* were enriched in both CMS1 and right-sided tumours ($FDR_{mol} < 0.05$, hypergeometric test). Additionally, mutations in *PIK3CA*, *FAM123B* and *KRAS* were only associated with right-sided tumours (Fig. 2e; $FDR_{mol} < 0.05$, hypergeometric test). In contrast, the similarity of left-sided tumours and CMS2 (Fig. 2f) was characterised by mutations in *APC*, *TP53* and chr20q amplifications (*SRC*, *TOP1*, *BCL2L1*, *ZNF217*), which were all significantly enriched in both left-sided and CMS2 tumours (Fig. 2g, h; $FDR_{mol} < 0.05$, hypergeometric test). Somatic mutations in *PTEN*, *ARID1A*, *ATM*, *LRP1B*, *BRCA2* and *NF1* did not show a preference for a particular primary tumour side, but were enriched in CMS1 tumours (Fig. 2h), and were associated with an increased tumour mutational burden ($p = 0.008$, $p = 0.002$, $p = 0.017$, $p = 0.0001$, $p = 0.010$ and $p = 0.051$, respectively, Fisher's exact test).

In summary, leveraging OncoBird and investigating patterns of genetic events in tumour subtypes revealed meaningful tumour biology. For example, mutations of either *BRAF* or *KRAS* promote ERK signalling and therefore occur mutually exclusive. *BRAF* mutations were predominantly found in CMS1, but nevertheless, 27 out of 53 *BRAF* mutant tumours were distributed among CMS2-4. Therefore, it is of utmost importance to gain an enhanced understanding of the molecular landscape of mCRC prior to the interpretation of biomarkers, which is further empowered by OncoBird.

Genetic biomarkers of cetuximab

First, independent of tumour subtypes, we assessed single genes and mutually exclusive gene modules (Fig. 2a, b) as biomarkers for cetuximab. For this, we leveraged Cox proportional hazards regression and logistic regression models ("Methods"), considering overall survival (OS; Fig. 3a–h), progression-free survival (PFS; Supplementary Fig. 8) and the objective response rate (ORR; Supplementary Fig. 9). We quantified effect sizes by hazard ratios (HR) for survival data and odds ratios (OR) for binary data including 95% confidence intervals (Supplementary Data 3).

The clinically established resistance biomarkers of cetuximab were recovered, i.e., mutations in *RAS* (either *KRAS* or *NRAS*) referred to a poorer OS in the cetuximab treatment arm (Fig. 1h; $p = 0.0002$,

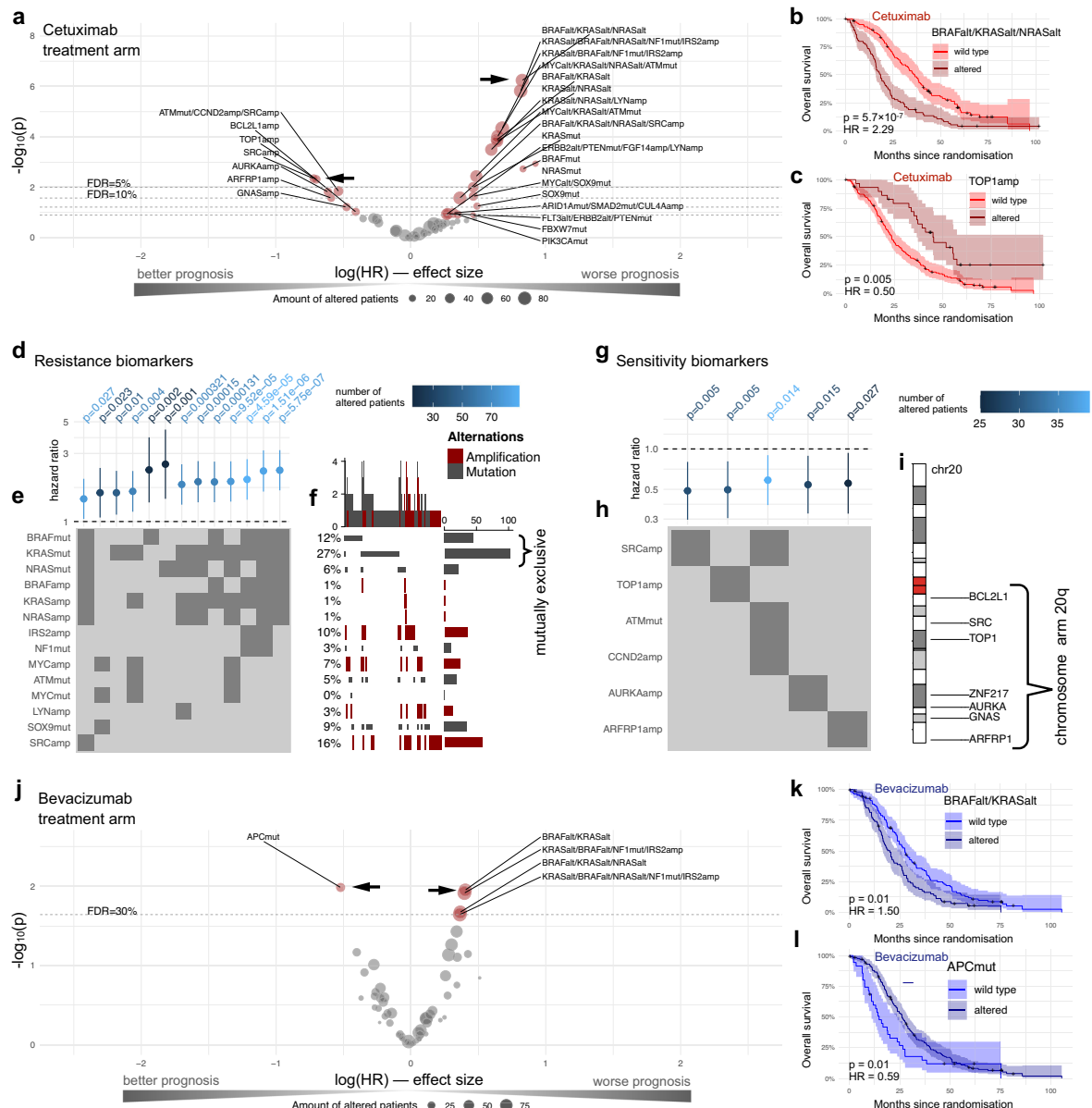


Fig. 3 | Identification of genetic biomarkers for FOLFIRI plus cetuximab or bevacizumab. **a** Volcano plot for genetic biomarkers of cetuximab in the form of mutually exclusive gene modules or single gene mutations. Each point shows the effect of a particular group of alterations summarised by its hazard ratio derived by the Cox regression models and its raw p -value derived by a Wald test. Exemplifying the most significant associations, Kaplan–Meier plots, including 95% confidence intervals (CI) and summary statistics of the Cox regression models, are shown for **b** the mutually exclusive module consisting of *RAS* and *BRAF* mutations, and **c** the amplification of *TOP1* treated with cetuximab. For investigating the biomarker composition, we focus on **d** resistance biomarkers of FOLFIRI plus cetuximab with $FDR_{cet} < 0.1$, showing their hazard ratios and 95% CIs. For these, **e** the composition of mutually exclusive genes is indicated by dark grey colour, and **f** an oncoprint

highlighting mutational frequencies of biomarker combinations is shown. In like manner, **g** cetuximab sensitivity biomarkers, shown by their hazard ratios and 95% CIs, and **h** their composition are summarised. **i** Karyoplot showing transcription start sites of co-amplified genes on chromosome 20q. **j** Volcano plot of the genetic biomarkers of bevacizumab with $FDR_{bev} < 0.3$, shown in brown colour by their hazard ratios derived by the Cox regression models and their raw p -values derived by a Wald test. Kaplan–Meier plot including 95% CIs and summary statistics of the Cox regression models of **k** mutations in *KRAS* or *BRAF* and **l** *APC* mutations treated with bevacizumab. The compositions of bevacizumab biomarkers are shown in Supplementary Fig. 10a, b. A Source Data file is provided, which contains the source data for the figure panels and the sample sizes of the conducted statistical tests.

$HR = 1.90$ [1.36–2.65], $FDR_{cet} < 0.1$). In addition, we confirmed that *BRAF* mutations are mutually exclusive to *RAS* mutations (Fig. 2b; $p = 0.0008$, Fisher's exact test), and both contributed to a poor OS when treated with cetuximab (Fig. 3b; $p = 5.7 \times 10^{-7}$, $HR = 2.29$ [1.65–3.16], $FDR_{cet} < 0.1$), which has been consistently observed in an independent cohort⁴⁰.

Most resistance biomarker modules grouped mutations in *KRAS* and *BRAF* ($FDR_{cet} < 0.1$). In addition, we found a gene module including mutations in *SOX9* and *MYC* amplifications, for which mutant tumours displayed a worse prognosis based on OS (Fig. 3d, e; $p = 0.02$, $HR = 1.50$ [1.07–2.37], $FDR_{cet} < 0.1$). By inspecting their oncoprint (Fig. 3f), 27/59 tumours harboured mutations in either *SOX9* or *MYC* and were wild-

type in either *BRAF*, *KRAS* or *NRAS*, hinting towards an alternative cetuximab resistance mechanism.

In addition, we found *TOP1* amplifications to be a strong predictor of a prolonged OS for treatment with cetuximab (Fig. 3c; $p = 0.005$, HR = 0.50 [0.30–0.81], FDR_{cet} < 0.1). In fact, we could identify multiple co-amplifications that showed prognostic value for the cetuximab treatment arm, which are located on chromosome 20q. Among the most predictive amplifications for a longer OS were *SRC*, *TOP1*, *AURKA* and *ARFRP1* (Fig. 3g–i; Supplementary Data 3). Consistent trends were observed with *SRC* amplifications in PFS ($p = 0.10$, HR = 0.69 [0.44–1.07], median PFS wild-type tumours 9.6 months vs mutants 11.1 months) and ORR ($p = 0.18$, OR = 0.45 [0.14–1.45], ratio ORR wild-type 0.66 vs mutant tumours 0.83).

Genetic biomarkers of bevacizumab

Analogously to the cetuximab biomarker analysis, for the bevacizumab treatment arm, we also built Cox proportional hazards regression models (“Methods”) applied to OS (Fig. 3j–l; Supplementary Fig. 10) and PFS (Supplementary Fig. 11), and logistic regression models for ORR (Supplementary Fig. 12). For exploring bevacizumab biomarker trends, we employed a lenient threshold of FDR_{bev} < 0.3, which deviates from the default setting (“Methods”). The mutually exclusive module of *KRAS* and *BRAF* mutations showed lower OS (Fig. 3j, k; $p = 0.01$, HR = 1.50 [1.10–2.04], FDR_{bev} < 0.3), which is consistent with literature reports^{41,42}. A better predictor for poor OS was the *APC* wild-

type status for tumours treated with FOLFIRI plus bevacizumab (Fig. 3j; $p = 0.01$, HR = 1.69 [1.14–2.50], FDR_{bev} < 0.3).

Subtype-specific biomarkers of cetuximab and bevacizumab

The previous analyses focused on genetic biomarkers in isolation, whilst here, we investigated them within the context of tumour subtypes (“Methods”). In FIRE-3, tumour subtypes are defined as either left- or right-sided tumours, or alternatively, classified according to the consensus molecular subtypes, i.e., CMS1–4 (“Methods”) ²⁹. Here, we tested stratifications based on each single gene or gene module within tumour subtypes for OS (Fig. 4a, b), PFS (Supplementary Fig. 13) and ORR (Supplementary Fig. 14).

In total, we found 38 subtype-specific biomarkers of cetuximab for OS (FDR_{cet} < 0.1; “Methods”). In particular, we recovered favourable OS of CMS2 patients treated with cetuximab (Fig. 4a), if their tumours additionally carried chr20q amplifications, i.e., *ARFRP1* (Fig. 4c; $p = 0.01$, HR = 0.32 [0.13–0.77], FDR_{cet} < 0.1), *TOP1* (Supplementary Fig. 15a; $p = 0.01$, HR = 0.34 [0.15–0.74], FDR_{cet} < 0.1) and *SRC* (Supplementary Fig. 15b; $p = 0.01$, HR = 0.37 [0.17–0.78], FDR_{cet} < 0.1). Additionally, CMS4 *KRAS* mutant tumours treated with cetuximab showed worse OS (Fig. 4d; $p = 0.002$, HR = 2.60 [1.44–4.70], FDR_{cet} < 0.1) and PFS (Supplementary Fig. 13a, c).

For reporting bevacizumab biomarker trends, we employed a lenient false discovery rate (FDR_{bev} < 0.3), which deviates from the conservative OncoBird default setting (“Methods”). Tumours with

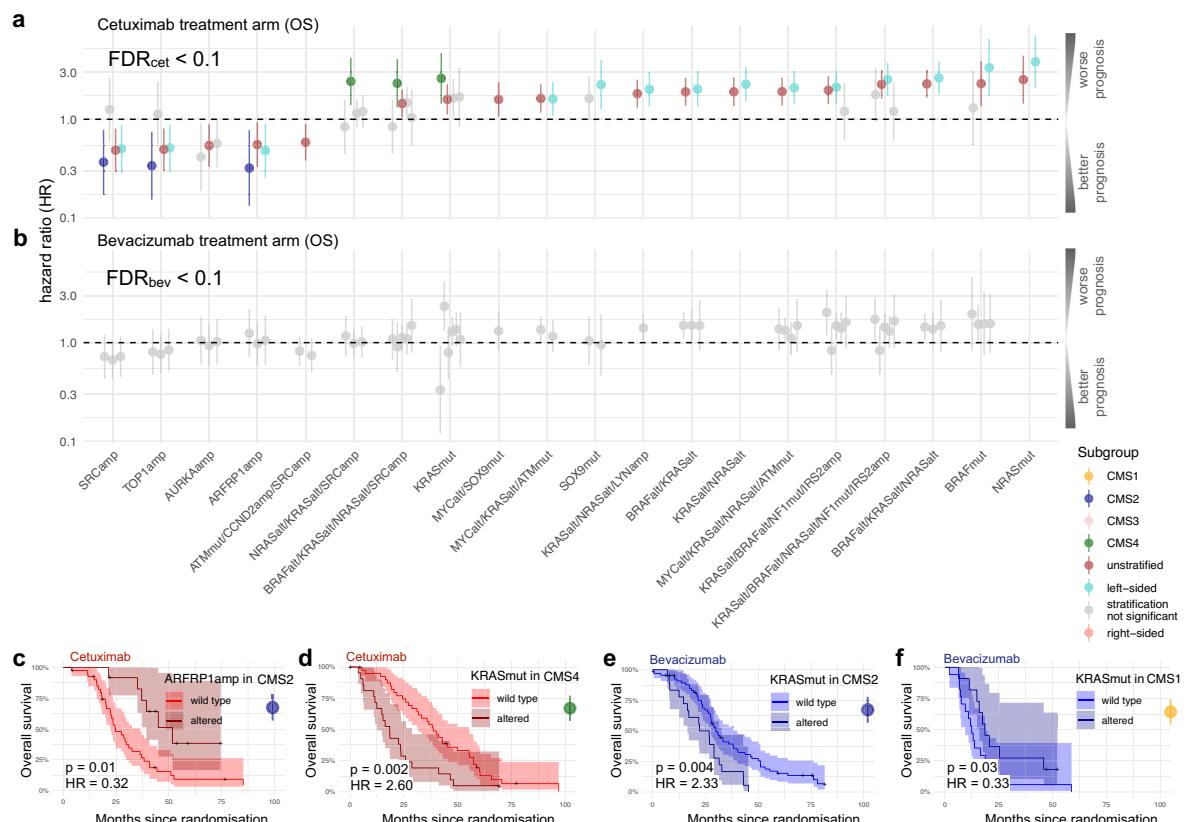


Fig. 4 | Identification of subtype-specific genetic biomarkers for FOLFIRI plus cetuximab or bevacizumab. Subtype-specific genetic biomarkers for OS of **a** cetuximab and **b** bevacizumab using hazard ratios including 95% confidence intervals (CI) derived from single Cox regression models. Subtypes are defined by either the primary tumour side, CMS or unstratified (reference model). Kaplan–Meier plots including 95% CIs, hazard ratios and raw p -values derived by

Wald tests from the Cox regression models of subtype-specific genetic biomarkers for **c** *ARFRP1* in CMS2, **d** *KRAS* in CMS4, **e** *KRAS* in CMS2 and **f** *KRAS* in CMS1 in either the cetuximab or bevacizumab treatment arm. A Source Data file is provided, which contains the source data for the figure panels and the sample sizes of the conducted statistical tests.

KRAS mutations classified as CMS2 tended to show worse OS when treated with bevacizumab (Fig. 4e; $p = 0.004$, HR = 2.33 [1.31–4.15], $\text{FDR}_{\text{bev}} < 0.3$). In contrast, *KRAS* mutated tumours classified as CMS1 tended to show a longer OS compared to wild-type tumours when treated with bevacizumab (Fig. 4f; $p = 0.03$, HR = 0.33 [0.12–0.93], $\text{FDR}_{\text{bev}} < 0.3$).

Predictive components of biomarkers

For assessing the predictive component of response biomarkers, here, we compared the cetuximab and bevacizumab treatment arms against each other by focusing on interactions between genetic alterations in the context of tumour subtypes (“Methods”). Subsequently, we compared the prognosis of both inhibitors for each subgroup according to the interaction biomarkers, thus assessing potential treatment benefits. In addition, we corrected the conditional average treatment effects in the identified subgroups using resampling methods to obtain multiplicity-adjusted p -values and bias-corrected confidence intervals (“Methods”). The results were summarised for OS (Fig. 5a, b) and PFS (Supplementary Fig. 16), whereas no significant interactions were detected for ORR. In total, we found five putative interactions (Supplementary Data 4; $\text{FDR}_{\text{int}} < 0.2$; “Methods”). For reporting other biomarker trends, we also included summary statistics of 57 subgroups with a lenient threshold of $\text{FDR}_{\text{int}} < 0.6$, which deviates from the default setting (Supplementary Data 3).

For example, we found predictive value of chr20q amplifications in CMS2 tumours treated with FOLFIRI plus cetuximab (Fig. 5a, b), which is evident by the significant interactions of *TOP1* ($p_{\text{int}} = 0.07$, $\text{FDR}_{\text{int}} < 0.2$) and *ARFRP1* ($p_{\text{int}} = 0.01$, $\text{FDR}_{\text{int}} < 0.2$). *ARFRP1* amplifications showed the largest predictive component among the chr20q

amplifications. Accordingly, we observed longer OS in the cetuximab treatment arm compared to bevacizumab in CMS2 (Fig. 5a, c; *ARFRP1*: $p = 0.003$, HR = 0.21 [0.07–0.59], $\text{FDR}_{\text{int}} < 0.2$; Supplementary Data 3). The resampling-based adjusted treatment effect confirmed this observation and yielded a hazard ratio in this subgroup of HR = 0.21 [0.09–0.54] with $p_{\text{adj}} = 0.04$ (Fig. 5a, c). Previous reports have indicated a prognostic value of chr20q amplifications in colorectal cancer patients^{38,43}, whilst OncoBird yielded additional evidence that they harbour a predictive component.

Another interaction example was tumours with *KRAS* mutations that showed CMS-specific responses. In CMS4, patients with *KRAS* wild-type tumours responded better to cetuximab compared to patients treated with bevacizumab (Fig. 5b, d; *KRAS* wild types: $p = 0.02$, HR = 0.57 [0.35–0.93]; $p_{\text{int}} = 0.02$, $\text{FDR}_{\text{int}} < 0.2$), for which the resampling-based adjusted treatment effect yielded HR = 0.70 [0.25–2.35] with $p_{\text{adj}} = 0.14$ (Fig. 5b, d). Our results suggest a predictive role of *KRAS* mutations in CMS4 for cetuximab, which we also identified for PFS (Supplementary Fig. 16c, d). Notably, modules containing alterations in *NRAS*, *BRAF* and *SRC* showed similar statistics since only four, eight and twelve mutant tumours were present in CMS4. Insignificant but numerically longer OS was observed for patients with *KRAS* mutated tumours classified as CMS4 treated with bevacizumab (Fig. 5e, *KRAS* mutants: $p = 0.24$, HR = 0.66 [0.33–1.31]), with a median OS 28.3 months compared to 18.4 months when treated with cetuximab.

In order to assess the ability of OncoBird to discover the same biomarkers for different datasets, we applied 5-fold cross-validation repeated five times and extracted the ten most significant biomarkers for OS across each of the 25 models (Fig. 6a). Consistent with our

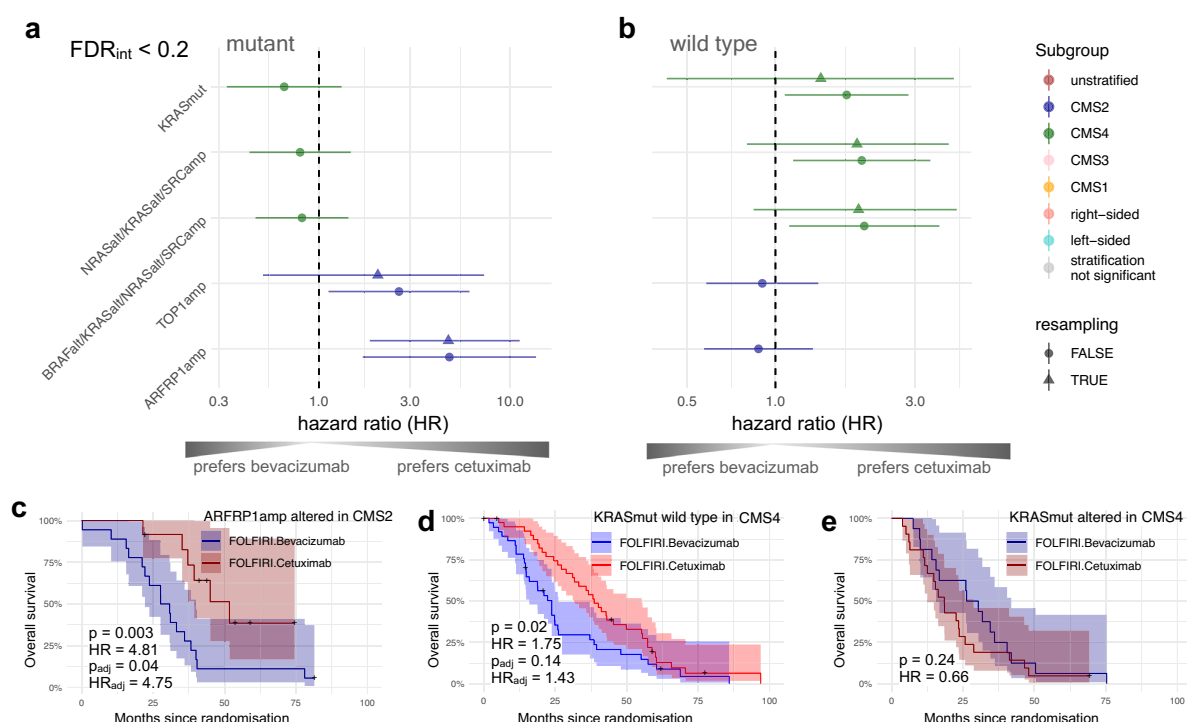


Fig. 5 | Predictive biomarkers in the context of tumour subtypes. Overview of interaction biomarkers ($\text{FDR}_{\text{int}} < 0.2$) focusing on **a** mutant and **b** wild-type tumours when comparing cetuximab and bevacizumab treatment, using hazard ratios including 95% confidence intervals (CI) derived from single Cox regression models fitted on OS. Triangle points and confidence intervals were obtained from the bootstrap-based bias-correction of treatment effects. For the conducted statistical

tests, the sample sizes are given in Supplementary Data 3. Here exemplified, Kaplan–Meier plots including 95% CIs, hazard ratios and raw p -values derived by Wald tests from the Cox regression models compare treatments in subgroups for **c** *ARFRP1* amplifications in CMS2 and **d, e** *KRAS* mutations in CMS4. Source data for the figure panels are provided as Source Data file.

previous findings, gene modules containing *KRAS* mutations for CMS4 were found in 21/25 training sets and chr20q amplifications in CMS2 were reproduced in 22/25 training sets (Fig. 6a).

Benchmarking of methods for subgroup analysis

For benchmarking OncoBird, we compared it to alternative methods that can be used to investigate predictive biomarkers based on overall survival. Together with OncoBird, eight algorithms and implementations were used in order to identify subgroups with differential treatment effects, i.e., virtual twins (VT)⁹, model-based partitioning (MOB)⁸, an outcome-weighting method (OWE)¹¹, causal random forests (CRF)¹², policy learning (POL)⁴⁴, GUIDE⁴⁵ and PRISM⁴⁶ (Supplementary Table 1; “Methods”; Fig. 6b).

For the evaluation, we first derived hazard ratios for cetuximab benefit based on OS in the subgroups according to the predicted biomarkers for all methods across five times 5-fold cross-validation (“Methods”). We also focused on the current treatment guidelines for mCRC, according to which patients should receive cetuximab if their tumours are *RAS* wild-type and left-sided (std; Fig. 6b)³⁷. While the treatment benefit was not significant for the std-positive subgroup

(Fig. 6b, median HR = 0.78, $p_{cv} = 0.129$), the methods that found the highest significant benefits were OncoBird (median HR = 0.74, $p_{cv} = 0.046$), POL (median HR = 0.81, $p_{cv} = 0.048$), MOB (median HR = 0.83, $p_{cv} = 0.048$) and OWE (median HR = 0.84, $p_{cv} = 0.049$) ordered by the magnitude of the hazard ratio (Fig. 6b).

Next, we leveraged the whole dataset to identify cetuximab sensitivity biomarkers with each method and compared them to the treatment guidelines. On average, 73% of methods identified cetuximab benefit for a patient in the std-positive subgroup, whereas only 33% of methods detected further benefits in the std-negative subgroup (Fig. 6c). 7/8 (88%) methods found mutually exclusive alterations in *KRAS*, *NRAS* or *BRAF* as a predictive biomarker, from which one, two and four methods proposed this marker in conjunction with tumour sidedness, CMS and across all patients, respectively (Supplementary Table 1). Only 2/8 (25%) methods highlighted *TOP1* amplifications as a potential biomarker (Supplementary Table 1). This highlights that current subgroup analysis methods mostly recover standard clinical practice, whilst sparsely identifying complementary predictive subgroups, thus highlighting the unmet need for cancer biology-driven frameworks such as OncoBird.

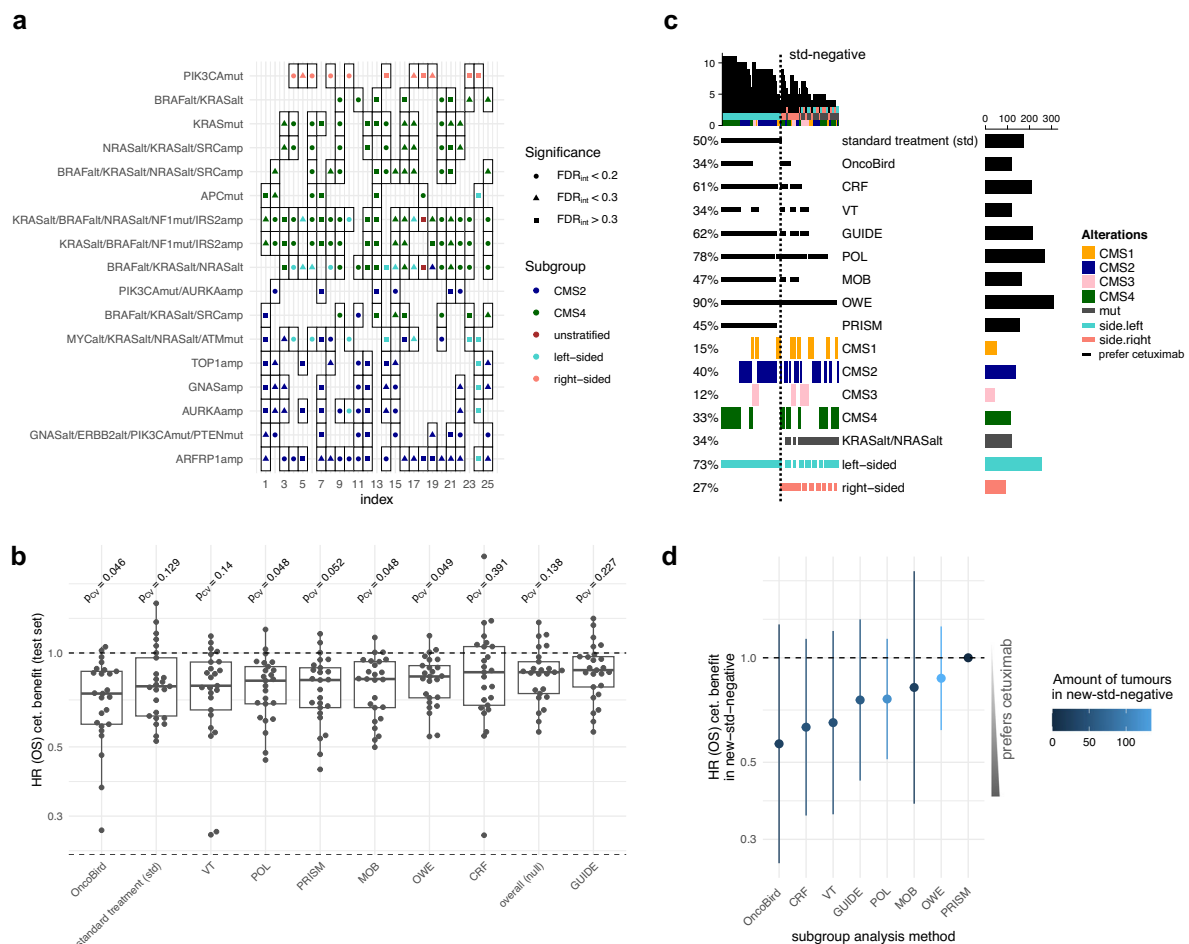


Fig. 6 | Stability analysis and benchmark with other methods. a The ten most significant biomarkers across 25 models of five times repeated 5-fold cross-validation. **b** Boxplots of treatment effects in terms of hazard ratios for the predicted subgroups in the 25 test sets for the benchmarked methods, including standard treatment guidelines (std) and overall across all patients (null). The centre line depicts the median; the box represents the inter-quartile range (IQR) and the whiskers the interval 1.5 times the IQR. **c** Oncoprint showing identified subgroups

for the benchmarked methods, including std, CMS subtypes, tumour sidedness and mutations in *KRAS* and *NRAS*. **d** Forest plot showing hazard ratios including 95% confidence intervals (CI) and amount of patients in the subgroup for which standard treatment is not recommended and which was found by subgroup analysis methods (new-std-negative). A Source Data file is provided, which contains the source data for the figure panels and the sample sizes of the conducted statistical tests.

Ideally, subgroup analysis should reveal subgroups with high treatment effects for refining treatment strategies and recover subgroups in the standard treatment strategy. Therefore, we evaluated the newly proposed subgroup for which standard treatment is not recommended (new-std-negative) for each method. We derived the hazard ratios for cetuximab benefit based on OS for all methods in the new-std-negative subgroups (Fig. 6d). Lower hazard ratios in new-std-negative patients indicate the discovery of off-label subgroups for which cetuximab is currently not recommended (Fig. 6d). Accordingly, OncoBird showed the numerically lowest hazard ratio $HR = 0.57$ ($p = 0.16$, $N = 29$) for the new-std-negative subgroup compared to all other methods (Supplementary Table 1; Fig. 6d).

In summary, many of the computational methods reproduced the clinically established biomarkers, whilst OncoBird empowers advanced biomarker identification by thoroughly integrating biological priors in the form of tumour subtypes. The simplicity of statistical models leveraged in OncoBird further increases interpretability and transparency.

Discussion

We demonstrated that OncoBird has the capabilities to characterise the molecular and biomarker landscape of RCTs. Here exemplified, we captured the established clinical biomarkers of FIRE-3, and proposed five predictive biomarker hypotheses ($FDR_{int} < 0.2$). The biomarkers were based on either individual cancer genes or mutually exclusive patterns and exploited these genetic events in the context of well-characterised cancer subtypes. In addition, OncoBird thoroughly corrects for multiple hypothesis testing and includes resampling-based adjustments of treatment effects. In essence, OncoBird systematically investigated the molecular landscape of the FIRE-3 clinical trial, suggested biomarkers based on genetic alterations, performed a data-driven subgroup analysis, and finally, presented the results in an interpretable and intuitive way.

The statistical power of detecting biomarkers depends on the amount of screened genes and subtypes, sample sizes and magnitude of treatment effects. For example, subtype-specific analyses reduce patient subgroup sizes, thus limiting the power for detecting interactions. In order to gain statistical power to detect genetic biomarkers with low mutational frequency, OncoBird exploits mutually exclusive modules ("Methods"). Despite the use of resampling-based treatment effect estimation in the found subgroups, hypotheses generated by exploratory tools such as OncoBird ought to be replicated in independent clinical trials. Nonetheless, OncoBird identified promising patient subpopulations within the FIRE-3 and ADJUVANT clinical trials with supported biological interpretation, which indicated refined predictive benefits in cancer subtypes.

A limitation of data-driven subgroup analysis is that these may produce spurious results if not biologically interpretable⁴⁷. To mitigate this risk, we used established tumour subtypes with distinct tumour biology in mCRC, i.e., here, the consensus molecular subtypes (CMS)²⁹ and primary tumour sidedness²¹. Furthermore, the grouping of functionally similar mutually exclusive somatic mutations in the cancer gene sequencing panel reinforced the identification of biological signals.

Somatic mutations may drive tumour subtypes, therefore, we systematically investigated mutational patterns within CMS1-4 and tumour sidedness. We found the majority of *BRAF* mutations in CMS1 and observed a co-occurrence between CMS2, left-sided tumours and amplifications in chr20q. In particular, CMS2 is characterised by a *MYC* signalling activation²⁹, which may be co-regulated by activation of the co-amplified *AURKA*⁴⁸. While we predominantly identified CMS-specific biomarkers, our results suggest that both primary tumour side and CMS subtypes play a major role in the landscape of predictive biomarkers. This highlights the need for OncoBird, an integrated biomarker discovery framework, which integrates the molecular landscape of RCTs with its biomarkers.

Several genes were co-amplified in chr20q, i.e., *ARFRP1*, *TOP1*, and *SRC*, thus determining the drivers among these biomarker candidates is challenging. Among the prominent chr20q amplifications, *TOP1* was previously proposed as a biomarker for irinotecan efficacy in metastatic colorectal cancer^{49,50}, which is part of the chemotherapeutic backbone of the FIRE-3 trial. Literature suggests that *TOP1* abundance is essential for irinotecan-induced DNA double-strand breaks during DNA replication⁵¹. Additionally, *TOP1* was identified to regulate EGFR through an endogenous interaction with the transcription factor c-Jun⁵², which supports the hypothesis that *TOP1* amplifications may be the actionable biomarker. *SRC* has been reported to play a role in cancer progression^{53,54}, whereas for *ARFRP1*, no functional evidence has been presented yet.

The resulting co-amplifications between these cancer genes complicate the determination of the genetic driver in chr20q. To understand the causality of cancer aetiologies, further efforts require additional treatment regimes. Alternative clinical trials for metastatic colorectal cancer often involve different chemotherapy backbones, i.e., fluorouracil, leucovorin, and oxaliplatin (FOLFOX) or fluorouracil, leucovorin, and irinotecan (FOLFIRI)³⁰. The use of other therapy backbones may unravel the role of *ARFRP1*, *TOP1* and *SRC* amplifications regarding better efficacy for patients treated with cetuximab. However, discrepancies may arise due to the synergism and antagonism of the different chemotherapy backbones and targeted treatments⁵⁵.

The prognostic potential of *APC* wild-type tumours for bevacizumab has been previously reported⁵⁶, whereas OncoBird did not yield enough evidence to support this. Indeed, a confounding factor is the enrichment of *BRAF* mutations in the *APC* wild-type tumours ($p = 1.4 \times 10^{-10}$, Fisher's exact test). This is, 48% of *APC* wild-type tumours were *BRAF* mutated in the bevacizumab treatment arm, whereas in the cetuximab treatment arm, only 29% were *BRAF* mutated ($p = 0.13$, Fisher's exact test). Nevertheless, independently a correlation between *VEGFA* expression and the mutational status of *APC* has been previously observed in primary colorectal tumour samples⁵⁷, suggesting that within *APC* mutated tumours, anti-VEGF treatment may indeed be beneficial.

Furthermore, *RAS/BRAF* mutations are known to harbour prognostic value in terms of overall survival^{38,43}. Furthermore, we observed that *KRAS* mutations showed highly CMS-specific responses. In particular, treatment response differed for tumours classified as CMS4 by *KRAS* status, showing better response for cetuximab in *KRAS* wild-type and for bevacizumab in *KRAS* mutated tumours, respectively. CMS4 has been reported to be associated with VEGF pathway activation and is thus associated with angiogenesis²⁹. Thus, patients with tumours resistant towards anti-EGFR treatment may benefit from VEGF inhibition. Further exclusion of *BRAF* mutations did not elevate the predictive potential of *KRAS* mutations in CMS4. However, the statistical power is limited by the fact that only six tumours harboured the prognostically unfavourable *BRAF* V600E mutation in CMS4²⁰.

In summary, OncoBird reproduced clinically established biomarkers and derived five hypotheses of biomarkers with predictive roles for FOLFIRI plus either cetuximab or bevacizumab. Highlighted examples include chr20q amplifications in CMS2 and *KRAS* mutations in CMS4, which may optimise patient stratification for metastatic colorectal cancer. Leveraging OncoBird for molecular profiling in the FIRE-3 clinical trial offered an expanded perspective on the molecular and biomarker landscape of these patients.

In the future, we anticipate that the analysis of clinical trials will progressively demand molecular patient tumour data, including predefined subtypes, highlighting the urgent need for integrative analysis tools such as OncoBird. Notably, OncoBird was developed for RCT designs and is generalisable to any trial designs for which the intention-to-treat population was defined before the treatment randomisation, i.e., the treatment assignment is independent of patient characteristics. According to this, OncoBird is applicable to modern clinical trial

designs based on master protocols⁵⁸, i.e., basket, umbrella, and platform trials if control arms are included. In an emerging landscape of predictive molecular biomarkers in cancer, OncoBird may untangle complex dependencies between somatic alterations and tumour subtypes in RCTs. Furthermore, OncoBird is generalisable to any cancer entity, thus ultimately paving the way for the next generation of precision oncology therapies.

Methods

Clinical data of the FIRE-3 clinical trial

FIRE-3 is an open-label, randomised phase III trial to compare first-line treatment in *KRAS* exon 2 wild-type metastatic colorectal cancer patients (mCRC) with either cetuximab or bevacizumab in combination with 5-fluorouracil, leucovorin and irinotecan (FOLFIRI). The protocol and rules of conduct were previously published^{23,59} (NCT00433927). The trial was conducted in accordance with the declaration of Helsinki (1996). All translational analyses were approved by the local ethics committee (University of Munich, registry no. 186-15). All patients included in this analysis provided written informed consent. 24% and 34% of the patients had female sex in the FOLFIRI plus cetuximab and bevacizumab arm, respectively. The sex is reported according to the study protocol^{23,59}, and gender cannot be distinguished retrospectively. The biological sex of patients (i.e., male or female) was assigned by the study doctor of the respective trial centre and reported to the clinical research organisation (CRO). The original intention-to-treat population consisted of 752 patients in total. Primary and secondary endpoints of the FIRE-3 trial, including the median overall survival (OS) and progression-free survival (PFS), were expressed as months and defined as stated in the respective articles^{23,59}. The objective response rate (ORR) was evaluated by the RECIST 1.0 criteria^{23,59}.

Next-generation sequencing and genetic alterations in FIRE-3

Primary tumour tissues from 373 patients have been molecularly characterised by next-generation sequencing (NGS) with the FoundationOne® panel (Foundation Medicine, Inc., MA, USA; catalogue number not available), which identified somatic mutations and copy number alterations, i.e., deletions and amplifications, of 277 key cancer genes, microsatellite instability (MSI) and tumour mutational burden²⁰. Somatic alterations were delivered in the form of binary matrices, that reflect the mutant or wild-type status of a given gene based on single nucleotide variants (SV), copy number amplifications (AMP) and deletions (DEL). MSI is an important prognostic predictor and enriched in CMSI⁶⁰, which is observed in our study, with 8 of 10 MSI-H tumours being classified as CMSI. However, MSI-H tumours are less prevalent in metastatic disease (~5%)⁶⁰. Furthermore, only six and four MSI-H tumours were treated with bevacizumab and cetuximab, respectively.

Gene expression profiling in FIRE-3

The genetic characterisation is complemented with gene expression profiles from Xcel® microarrays (Almac Ltd, Belfast, UK; catalogue number: 902016) in a subset of 451 patients. The clinical data and the layers of molecular characterisation led to 163 and 186 patients, which are fully characterised in the cetuximab and bevacizumab treatment arms, respectively.

Tumour subtypes in FIRE-3

A clinically established subtype for mCRC is its primary tumour sidedness. Left-sided tumours were located in the left hemicolon, e.g., splenic flexure to the rectum. In contrast, right-sided tumours were located in the right colon, e.g., caecum to the transverse colon. In addition, annotations for molecular subtypes of mCRC were obtained from transcriptome data that has been previously used to classify

patients into their closest consensus molecular subtype (CMS)^{21,29} using the *cmsclassifier* package with the SSP predictor. Thereby, 24 of out 373 patient tumours were not allocated to any CMS because of missing transcriptomics data and were left out of the CMS-specific analysis. The CMS classification was used as a complementary alternative to the primary tumour side and is currently discussed in multiple clinical settings⁶¹.

Oncology Biomarker Discovery workflow

The Oncology Biomarker Discovery (OncoBird) framework applies to RCTs for which patients received either treatment $t \in \{0,1\}$ according to the treatment indicator T , had an associated outcome Y and can be classified into q subtypes $\{s_1, \dots, s_q\}$ according to the subtype variable S (clinical data). Additionally, patient tumours are characterised by m candidate genetic biomarkers $\mathbf{X} = X_1, \dots, X_m$ with the observed biomarkers for patients $\mathbf{x} = x_1, \dots, x_m$ (genetic data). The genetic data can be used to group functionally similar genes that can be added to the set of candidate biomarkers. Furthermore, it is possible to add additional binary features to \mathbf{X} such as binarised copy number alterations with appropriate cutoffs or the MSI status of a tumour. Both genetic data (*MUT*) and clinical data (*CLIN*) are required inputs to the OncoBird workflow (Supplementary Data 1), which is described in the following sections. All implemented thresholds of OncoBird can be adjusted by the user, thus empowering more lenient or stringent analyses.

Characterising the molecular landscape in clinical trials

OncoBird first examines genetic features \mathbf{X} in tumour subtypes $\{s_1, \dots, s_q\}$ independent of the treatment and patient response (function *GET-MUTATIONS-IN-SUBTYPES* in Supplementary Data 1). For examining enrichment or depletion of each genetic feature in tumour subtypes, one-sided hypergeometric tests are performed using the 'phyper' R function. Consecutively, the resulting p -values are corrected for multiple hypothesis testing with the Benjamini–Hochberg (BH) method⁶². The FDR cutoff for this analysis step is denoted by FDR_{mol} and controlled at $\text{FDR}_{\text{mol}} = 0.05$ as our default setting. Our method generalises to any binary tumour characterisation, e.g., the MSI status in FIRE-3. As a default setting, we test genetic features that were mutated in at least ten tumours ($n = 10$).

Identifying mutual exclusivity

For the identification of mutually exclusive modules, we used the Mutex algorithm³⁹ (function *GET-MUTATIONS-MODULES* in Supplementary Data 1). It leverages a signalling network⁶³ collecting interactions from Pathway Common⁶⁴, SPIKE⁶⁵ and SignalLink⁶⁶ in order to scan for common downstream effects of combinations of somatic alterations \mathbf{X} . The default setting only uses somatic variants that were altered in at least ten tumours ($n = 10$).

Genetic and subtype-specific biomarkers

OncoBird tests single somatic alterations and previously derived mutually exclusive somatic alterations for differential prognosis in each treatment arm separately (function *GET-TREATMENT-SPECIFIC-BIOMARKERS* in Supplementary Data 1). The patient outcome $Y(T = t, S = s_k)$ for the treatment arm $T = t$ in subtype $S = s_k$ with $k = 1, \dots, q$ may be defined by survival data (OS or PFS) or a binary variable measuring the objective response rate (ORR). Depending on the type of outcome, this is modelled with either Cox proportional hazards regression models or logistic regression models expressed by their linear predictor function $f(\mathbf{x}, t)$. Using this classical approach for subgroup analysis, the treatment-specific regression models in subtypes take the form

$$f(\mathbf{x}, t) = \alpha_{0j} + \alpha_{1j}x_j + \sum_l C_l \quad (1)$$

Cox proportional hazards regression models for survival end-points were implemented with the ‘coxph’ function from the *survival* R package or logistic regression models for binary response variables were implemented using the ‘glm’ function. We test each $\mathbf{x} = x_1, \dots, x_m$ first across all tumours, and subsequently in tumour subtypes $\{s_1, \dots, s_q\}$, i.e., CMS or primary sidedness. α_{ij} is the coefficient estimating the contribution of candidate biomarker $j = 1, \dots, m$ for patient outcomes in the context of each treatment arm $T = t$ in the subtype $S = s_k$. The predictors C_1, \dots, C_l include additional prognostic covariates and their coefficients.

The p -value $p_{\alpha_{ij}}$ derived by a Wald test from the coefficient α_{ij} is multiplicity-adjusted for each treatment arm t and across all biomarkers x_j with $j = 1, \dots, m$ for either all patients or across subtypes s_k with $k = 1, \dots, q$ and yields adjusted p -values $\tilde{p}_{\alpha_{ij}}$ using the Benjamini–Hochberg (BH) method⁶². The default false discovery rates (FDR) are controlled at $\text{FDR}_\alpha = 0.1$ for either treatment-specific component α_{ij} .

The adjustable default setting of OncoBird is to only perform statistical tests if, for a given candidate biomarker x_j and tumour subtype s_k , at least $n = 10$ samples were present in each mutant and wild-type population. Additionally, OncoBird only tested alterations for which its corresponding gene module had at least n tumours redistributed compared to the single gene alteration.

Predictive components of biomarkers

For the subsequent comparison of treatment arms, OncoBird tests for significant statistical interactions between treatment arms and genetic alterations in tumour subtypes (function `GET-PREDICTIVE-BIOMARKERS` in Supplementary Data 1). For that, we modelled the outcome $Y(S = s_k)$ in subtype $S = s_k$ with $k = 1, \dots, q$ using regression models with interactions between T and X_j which take the form

$$f(\mathbf{x}, t) = \beta_{0j} + \beta_{1j}x_j + \beta_{2j}x_jt + \sum_l C_l, \quad (2)$$

where the coefficients β_{1j} and β_{2j} estimate the prognostic and predictive component of biomarker x_j in subtype s_k , respectively. The p -value $p_{\beta_{2j}}$ derived with a Wald test from the coefficient β_{2j} is multiplicity-adjusted across all m biomarkers for either all patients or across subtypes s_k with $k = 1, \dots, q$ and yields BH adjusted p -values $\tilde{p}_{\beta_{2j}}$. The default FDR is controlled at $\text{FDR}_\beta = 0.2$ for predictive components. The biomarker x_j in subtype s_k is a putatively predictive biomarker if $\tilde{p}_{\alpha_{ij}} < \text{FDR}_\alpha$ for either t and $\tilde{p}_{\beta_{2j}} < \text{FDR}_\beta$.

Furthermore, OncoBird only performs statistical tests if for a given genetic alteration X_j and tumour subtype s_k , at least $n = 10$ samples were present in each mutant and wild-type population for each treatment arm as default setting.

Resampling for correction of conditional average treatment effects

Lastly, we estimate the conditional average treatment effect (CATE) for the found biomarkers (function `GET-PREDICTIVE-BIOMARKERS` in Supplementary Data 1). For each significant X_j in s_k , there is one CATE estimate in each found subpopulation with a positive (mutant) biomarker $x_j = 1$ and negative (wild type) biomarker $x_j = 0$. In each population, we estimate the CATE by modelling the outcome Y by

$$f(\mathbf{x}, t) = \gamma_0 + \gamma_1t + \sum_l C_l, \quad (3)$$

where γ_1 estimates the (biased) CATE in terms of either hazard ratios or odds ratios dependent on outcome type in the subgroup defined by biomarker x_j and subtype s_k . The population with the larger absolute estimate γ_1 is used to estimate the subgroups A_{x_j, s_k} .

For each found subgroup A , we assess the significance to the associated CATE estimate γ_1 and derive the p -value p_{γ_1} using a Wald test. Furthermore, we perform a multiplicity-adjustment of p_{γ_1} and derive honest estimates of the CATE.

The p -values are adjusted for multiplicity using a permutation-based approach that takes into account the entire subgroup search strategy³. For that, we permuted the treatment labels $U = 1000$ times to obtain null datasets without any differential treatment effects. Next, for each null dataset, we select significant subgroups $A^{(u)}$ for the same thresholds and record the treatment effect p -value of the best subgroup $p^{(u)}$ with $u = 1, \dots, U$. The adjusted p -values are then given by

$$\tilde{p}_{\gamma_1} = \frac{1}{U} \sum_{u=1}^U I_{\{p^{(u)} \leq p_{\gamma_1}\}}(p^{(u)}), \quad (4)$$

the fraction of p -values $p^{(u)}$ that are smaller or equal than p_{γ_1} with the indicator function I . Furthermore, we derive an honest estimate of the treatment effect γ_1 . Since subgroups A are derived from the same data as the treatment effect estimates, the estimates from the resubstitution $\gamma_1(A_{x_j, s_k})$ will be biased. In order to derive a bias-corrected estimate $\tilde{\gamma}_1$, we use a previously proposed non-parametric bootstrap approach⁹. For that, we generated $B = 500$ bootstrapped datasets. For each resampled dataset $b = 1, \dots, B$ we estimate subgroups $\hat{A}_{x_j, s_k}^{(b)}$. The treatment effects can then be either estimated on the b -th resampled dataset $\gamma_1^{(b)}(A^{(b)})$ or on the original dataset $\gamma_1(A^{(b)})$. The bias-corrected CATE estimate is then given by

$$\tilde{\gamma}_1 = \frac{1}{B} \sum_{b=1}^B (\gamma_1(A) + \gamma_1(A^{(b)}) - \gamma_1^{(b)}(A^{(b)})). \quad (5)$$

The 95% confidence intervals are constructed by the 0.025 and 0.975 quantiles of the bootstrapped distribution.

OncoBird parameterisation for FIRE-3

We used the function `GET-MUTATIONS-IN-SUBTYPES` to evaluate the primary tumour side and CMS as tumour subtypes with the default setting $\text{FDR}_{\text{mol}} < 0.05$. In total, we performed 156 and 312 statistical tests for the primary tumour sidedness and CMS, respectively. Using the `GET-MUTATIONS-MODULES` function with default settings, we analysed 42 genes which yielded 29 mutually exclusive modules. Mutations in *KRAS* or *NRAS* are the established clinical biomarkers for anti-EGFR treatment, thus we jointly modelled *KRAS* and *NRAS* as *RAS* mutations resulting in 10 additional modules.

The `GET-TREATMENT-SPECIFIC-BIOMARKERS` function was used with the number of metastatic sites and the information about a prior tumour resection as added covariates C_1, C_2 . With the OncoBird default setting, we performed 816 statistical tests across all readouts Y (OS, PFS and ORR), the cetuximab and bevacizumab treatment arm and tumour subtypes, i.e., CMS1-4, left- and right-sided and across all tumours. FDR cutoffs are employed for each treatment arm separately and are denoted FDR_{cet} and FDR_{bev} for the analysis in the cetuximab and bevacizumab treatment arms, respectively. In total, we found 92 significant associations with the default setting $\text{FDR}_{\text{cet/bev}} < 0.1$. The criteria $\text{HR} < 1$ and $\text{OR} < 1$ corresponded to a better prognosis for the mutant tumours compared to the wild-type tumours and vice versa. To consistently report $\text{HR} < 1$ and $\text{OR} < 1$ as beneficial risk reduction, reciprocal values of HRs and ORs were used if wild-type tumours displayed a better prognosis. We represent p -values, hazard/odds ratios with the 95% confidence intervals (CI) in square brackets and the associated FDRs.

In FIRE-3, the `GET-PREDICTIVE-BIOMARKERS` function with default settings resulted in a total amount of 396 statistical tests across the readouts Y (OS, PFS and ORR) and the tumour subtypes s_k . FDR cutoffs for the interaction tests across both treatment arms are denoted by FDR_{int} . We explored 57 associations with $FDR_{int} < 0.6$ and $FDR_{cet/bev} < 0.1$ (Supplementary Data 3) and further focused on a subset of five biomarkers with default setting $FDR_{int} < 0.2$ for OS, i.e., two gene modules and three single genes (Supplementary Data 4). For the cross-validation analysis, a more lenient $FDR_{int} < 0.3$ was employed, which deviated from default setting to account for reduced sample sizes in the training and testing splits. HRs and ORs > 1 and < 1 corresponded to benefit with cetuximab and bevacizumab, respectively. To report the benefits of cetuximab treatment, the reciprocal values of HRs and ORs were used in the manuscript in order to display treatment benefits consistently with $HR < 1$ and $OR < 1$. We reported p -values and hazard/odds ratios with the 95% CIs for the treatment comparison and the p -values and associated FDRs for the interaction tests.

OncoBird parameterisation for ADJUVANT

The ADJUVANT clinical trial in *EGFR* mutant non-small cell lung cancer (NSCLC) aimed to assess the efficacy of gefitinib versus chemotherapy with vinorelbine and cisplatin (NCT01405079)³⁴. The trial was previously approved by the research ethics boards of Guangdong Provincial People's Hospital and all other participating hospitals³⁵. Of note, 58% and 59% of patients had female sex in the gefitinib and chemotherapy arm, respectively. The sex was reported according to the study protocol³⁴, and gender cannot be distinguished retrospectively. We used the *EGFR* subtype, i.e., exon 19 deletion or exon 21 Leu858Arg, and the smoking history as putative tumour subtypes and clinical endpoints were disease-free survival (DFS) and overall survival (OS). We analysed 22 somatic alterations in 171 patients, from which 76 patients were treated with chemotherapy alone, and 95 were treated with gefitinib. For the subsequent analysis, we used the OncoBird default settings. The obtained results (Supplementary Data 2; Supplementary Figs. 3–6) and an associated extensive report can be reproduced in a runnable demo on Code Ocean (<https://codeocean.com/capsule/9911222/tree/v1>).

Benchmarking of alternative methods with FIRE-3

For benchmarking the biomarker identification, we compared OncoBird to seven competing subgroup analysis algorithms leveraging the overall survival of FIRE-3 (Supplementary Table 1)^{8,9,11,12,44–46}. We formed predictors by concatenating clinical annotations, including information about tumour resection, number of metastatic sites, age, gender, MSI and lung metastatic status. We added single genetic alterations and mutually exclusive modules observed across at least ten patients and in both investigated tumour subtypes, thus mirroring the OncoBird default settings. Furthermore, we investigated interactions between genetic alterations and tumour primary sidedness or CMS as predictors. Subgroups for the method evaluation were formed as the union of the subgroups showing cetuximab benefit according to the identified biomarkers (Supplementary Table 1).

All benchmarked models were 5-fold cross-validated with five repetitions. A univariate Cox proportional hazards model assessed performances leveraging the treatment effect based on OS in the subgroups with predicted benefits according to the found biomarkers. This included the treatment effect across the whole test set and in the subgroup defined by the current treatment guidelines, i.e., left-sided and *RAS* wild-type tumours³⁷. The significance of the treatment effect in the subgroups of the test set was assessed using a modified t -test for resampled performance metrics⁶⁷, denoted by p_{cv} .

For comparing computational methods and their predicted biomarkers, the models were fitted on the whole dataset. The parameterisation of these methods was followed according to the suggested default settings unless in conflict with the above outlined

use case. For example, for tree-based methods, the features contained in the resulting tree were used as biomarkers with tree depths = 2, with a minimum subgroup size of $n = 10$. For the implementation of the virtual twins method (VT)⁹, we used the R package *randomForestSRC* with default parameters and averaged predictions over 10 times repeated 10-fold cross-validation. Subsequently, a regression tree was fitted to the original data. In order to perform model-based recursive partitioning⁸, we used the R package *model4you*⁶⁸ using an exponential model with default conditional inference tree control parameters. The PRISM method⁴⁶ was implemented in the R package *StratifiedMedicine*, for which we used Cox proportional hazards regression. We used the implementation of causal survival forests⁶⁹ (CRF) in the R package *grf*⁷⁰ for estimating conditional treatment effects. The propensity scores were set as constant and the target estimand was set to restricted mean survival time (RMST) with horizon = 100. After model fitting, variable importance scores were extracted, and biomarkers were selected according to predictors with significant linear projections of the conditional average treatment effects ($p < 0.05$). Next, we employed policy learning (POL)⁴⁴ to find optimal treatment regimens using the R package *policytree*⁷¹. We used the 50 most important predictors according to the CRF causal survival forest model variable importance scores and their treatment effect estimates to produce a decision tree.

The remaining methods were not based on trees. For the outcome weighted method (OWE)¹¹, implemented in the R package *personalized*⁷², we used a constant propensity score, lasso loss and 10-fold cross-validation. The GUIDE method⁴⁵ was available as a binary executable under <https://pages.cs.wisc.edu/~loh/guide.html>. We used Cox proportional hazards regression with interactions tests and mean-based trees with pruning. For the SIDES method (R package *SIDES*)⁷, we used `level_control=0` and `alpha=0.05`.

Statistics and reproducibility

The investigators were not blinded to the randomised treatment allocation during the data collection and outcome assessment. Since the conducted subgroup analysis is retrospective, the sample sizes were not predetermined. No data were excluded from the analysis. Details of the conducted statistical tests are provided in the figure captions, Supplementary Data 2–4 and Source Data. The results of the statistical analysis of the ADJUVANT clinical trial are reproducible from a demo run on Code Ocean (<https://codeocean.com/capsule/9911222/tree/v1>).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The clinical data summary from the FIRE-3 clinical trial analysed in this study has been deposited in the Pharmnet.bund online platform of the German Federal Ministry of Health (https://portal.dimdi.de/data/ctr/O-0329_OI-2-1-B80630-20190731152224.pdf) and was published before¹⁹. The clinical and molecular data is available under restricted access due to data privacy laws. The raw and processed data can be obtained through the corresponding author at volker.heinemann@med.uni-muenchen.de. The data from the results of OncoBird v0.1.0 executed on the FIRE-3 trial are available in Supplementary Data 3 and Source Data. The processed data from the ADJUVANT clinical trial is available on Zenodo^{33,35}. The data from the results of OncoBird v0.1.0 executed on the ADJUVANT trial are available in Supplementary Data 2, Source Data and on Code Ocean (<https://codeocean.com/capsule/9911222/tree/v1>). Source data are provided with this paper.

Code availability

Oncology Biomarker Discovery (OncoBird) is publicly available at <https://github.com/MendenLab/OncoBird>. The repository contains an

R package as well as a Shiny application with a graphical user interface in a local docker container (Supplementary Fig. 1). Additionally, a demo run of OncoBird v0.1.0 used for analysis is available on Code Ocean (<https://codeocean.com/capsule/9911222/tree/v1>).

References

1. Ting, N., Cappelleri, J. C., Ho, S. & Chen, D.-G. (eds) *Design and Analysis of Subgroups with Biopharmaceutical Applications* (Springer, 2020).
2. European Medicines Agency. *Guideline on the Investigation of Subgroups in Confirmatory Clinical Trials*. Draft. European Medicines Agency/Committee for Medicinal Products for Human Use. EMA/CHMP/539146/2013 (EMA, 2014).
3. Lipkovich, I., Dmitrienko, A. & D'Agostino Sr, B. R. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat. Med.* **36**, 136–196 (2017).
4. Zhang, Z., Seibold, H., Vettore, M. V., Song, W.-J. & François, V. Subgroup identification in clinical trials: an overview of available methods and their implementations with R. *Ann. Transl. Med.* **6**, 122 (2018).
5. Loh, W., Cao, L. & Zhou, P. Subgroup identification for precision medicine: a comparative review of 13 methods. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **9**, e1326 (2019).
6. Lipkovich, I., Dmitrienko, A., Denne, J. & Enas, G. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat. Med.* **30**, 2601–2621 (2011).
7. Lipkovich, I. & Dmitrienko, A. Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. *J. Biopharm. Stat.* **24**, 130–153 (2014).
8. Seibold, H., Zeileis, A. & Hothorn, T. Model-based recursive partitioning for subgroup analyses. *Int. J. Biostat.* **12**, 45–63 (2016).
9. Foster, J. C., Taylor, J. M. G. & Ruberg, S. J. Subgroup identification from randomized clinical trial data. *Stat. Med.* **30**, 2867–2880 (2011).
10. Xu, Y. et al. Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics* **71**, 645–653 (2015).
11. Chen, S., Tian, L., Cai, T. & Yu, M. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* **73**, 1199–1209 (2017).
12. Wager, S. & Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* **113**, 1228–1242 (2018).
13. Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl Acad. Sci. USA* **116**, 4156–4165 (2019).
14. Cremolini, C. et al. First-line chemotherapy for mCRC—a review and evidence-based algorithm. *Nat. Rev. Clin. Oncol.* **12**, 607–619 (2015).
15. Thomas, R. K. et al. High-throughput oncogene mutation profiling in human cancer. *Nat. Genet.* **39**, 347–351 (2007).
16. Kawazoe, A. et al. A retrospective observational study of clinicopathological features of KRAS, NRAS, BRAF and PIK3CA mutations in Japanese patients with metastatic colorectal cancer. *BMC Cancer* **15**, 258 (2015).
17. Van Cutsem, E. et al. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *N. Engl. J. Med.* **360**, 1408–1417 (2009).
18. Saltz, L. B. et al. Bevacizumab in combination with oxaliplatin-based chemotherapy as first-line therapy in metastatic colorectal cancer: a randomized phase III study. *J. Clin. Oncol.* **26**, 2013–2019 (2008).
19. Heinemann, V. et al. FOLFIRI plus cetuximab or bevacizumab for advanced colorectal cancer: final survival and per-protocol analysis of FIRE-3, a randomised clinical trial. *Br. J. Cancer* **124**, 587–594 (2021).
20. Stahler, A. et al. Single-nucleotide variants, tumour mutational burden and microsatellite instability in patients with metastatic colorectal cancer: next-generation sequencing results of the FIRE-3 trial. *Eur. J. Cancer* **137**, 250–259 (2020).
21. Stintzing, S. et al. Consensus molecular subgroups (CMS) of colorectal cancer (CRC) and first-line efficacy of FOLFIRI plus cetuximab or bevacizumab in the FIRE3 (AIO KRK-0306) trial. *J. Clin. Orthod.* **35**, 3510–3510 (2017).
22. Laurent-Puig, P. et al. MiR-31-3p is a predictive biomarker of cetuximab response in FIRE3 clinical trial. *Ann. Oncol.* **27**, vi151 (2016).
23. Heinemann, V. et al. FOLFIRI plus cetuximab versus FOLFIRI plus bevacizumab as first-line treatment for patients with metastatic colorectal cancer (FIRE-3): a randomised, open-label, phase 3 trial. *Lancet Oncol.* **15**, 1065–1075 (2014).
24. Duarte, S. et al. Right vs left-sided RAS wild-type metastatic colorectal cancer treated with EGFR inhibitors: prognostic differences. *Ann. Oncol.* **30**, iv53 (2019).
25. Stintzing, S. et al. Randomized study to investigate FOLFOXIRI plus either bevacizumab or cetuximab as first-line treatment of BRAF V600E-mutant mCRC: the phase-II FIRE-4.5 study (AIO KRK-0116). *J. Clin. Orthod.* **39**, 3502–3502 (2021).
26. Peeters, M. et al. Massively parallel tumor multigene sequencing to evaluate response to panitumumab in a randomized phase III study of metastatic colorectal cancer. *Clin. Cancer Res.* **19**, 1902–1912 (2013).
27. Seymour, M. T. et al. Panitumumab and irinotecan versus irinotecan alone for patients with KRAS wild-type, fluorouracil-resistant advanced colorectal cancer (PICCOLO): a prospectively stratified randomised trial. *Lancet Oncol.* **14**, 749–759 (2013).
28. Dienstmann, R., Salazar, R. & Tabernero, J. Overcoming resistance to anti-EGFR therapy in colorectal cancer. *Am. Soc. Clin. Oncol. Educ. Book* **35**, e149–e156 (2015).
29. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
30. Lenz, H.-J. et al. Impact of consensus molecular subtype on survival in patients with metastatic colorectal cancer: results from CALGB/SWOG 80405 (Alliance). *J. Clin. Oncol.* **37**, 1876–1885 (2019).
31. Mooi, J. K. et al. The prognostic impact of consensus molecular subtypes (CMS) and its predictive effects for bevacizumab benefit in metastatic colorectal cancer: molecular analysis of the AGITG MAX clinical trial. *Ann. Oncol.* **29**, 2240–2246 (2018).
32. Sveen, A., Kopetz, S. & Lothe, R. A. Biomarker-guided therapy for colorectal cancer: strength in complexity. *Nat. Rev. Clin. Oncol.* **17**, 11–32 (2020).
33. cancer-oncogenomics. *cancer-oncogenomics/minerva-adjuvant-nsccl: adjuvant minerva study v1.0.0*. Zenodo <https://doi.org/10.5281/zenodo.5242512> (2021).
34. Zhong, W.-Z. et al. Gefitinib versus vinorelbine plus cisplatin as adjuvant treatment for stage II-III (N1-N2) EGFR-mutant NSCLC (ADJUVANT/CTONG1104): a randomised, open-label, phase 3 study. *Lancet Oncol.* **19**, 139–148 (2018).
35. Liu, S.-Y. et al. Genomic signatures define three subtypes of EGFR-mutant stage II-III non-small-cell lung cancer with distinct adjuvant therapy outcomes. *Nat. Commun.* **12**, 6450 (2021).
36. Holch, J. W., Ricard, I., Stintzing, S., Modest, D. P. & Heinemann, V. The relevance of primary tumour location in patients with metastatic colorectal cancer: a meta-analysis of first-line clinical trials. *Eur. J. Cancer* **70**, 87–98 (2017).
37. Chiorean, E. G. et al. Treatment of patients with late-stage colorectal cancer: ASCO Resource-Stratified Guideline. *JCO Glob. Oncol.* **6**, 414–438 (2020).
38. Ptashkin, R. N. et al. Chromosome 20q amplification defines a subtype of microsatellite stable, left-sided colon cancers with wild-

- type RAS/RAF and better overall survival. *Mol. Cancer Res.* **15**, 708–713 (2017).
39. Babur, Ö. et al. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biol.* **16**, 45 (2015).
 40. Hsu, H.-C. et al. Mutations of KRAS/NRAS/BRAF predict cetuximab resistance in metastatic colorectal cancer patients. *Oncotarget* **7**, 22257–22270 (2016).
 41. Díaz-Rubio, E. et al. Role of Kras status in patients with metastatic colorectal cancer receiving first-line chemotherapy plus bevacizumab: a TTD group cooperative study. *PLoS ONE* **7**, e47345 (2012).
 42. Modest, D. P. et al. Outcome according to KRAS-, NRAS- and BRAF-mutation as well as KRAS mutation variants: pooled analysis of five randomized trials in metastatic colorectal cancer by the AIO colorectal cancer study group. *Ann. Oncol.* **27**, 1746–1753 (2016).
 43. Zhang, B., Yao, K., Zhou, E., Zhang, L. & Cheng, C. Chr20q amplification defines a distinct molecular subtype of microsatellite stable colorectal cancer. *Cancer Res.* **81**, 1977–1987 (2021).
 44. Athey, S. & Wager, S. Policy learning with observational data. *Econometrica* **89**, 133–161 (2021).
 45. Loh, W.-Y. & Zhou, P. The GUIDE approach to subgroup identification. *Design and Analysis of Subgroups with Biopharmaceutical Applications* (eds Ting, N. et al.) 147–165 (Springer, 2020).
 46. Jemielita, T. O. & Mehrotra, D. V. PRISM: patient response identifiers for stratified medicine. Preprint at <https://arxiv.org/abs/1912.03337> (2019).
 47. Dmitrienko, A., Muysers, C., Fritsch, A. & Lipkovich, I. General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. *J. Biopharm. Stat.* **26**, 71–98 (2016).
 48. Takahashi, Y. et al. The AURKA/TPX2 axis drives colon tumorigenesis cooperatively with MYC. *Ann. Oncol.* **26**, 935–942 (2015).
 49. Nygård, S. B. et al. DNA topoisomerase I gene copy number and mRNA expression assessed as predictive biomarkers for adjuvant irinotecan in stage II/III colon cancer. *Clin. Cancer Res.* **22**, 1621–1631 (2016).
 50. Palshof, J. A. et al. Topoisomerase I copy number alterations as biomarker for irinotecan efficacy in metastatic colorectal cancer. *BMC Cancer* **17**, 48 (2017).
 51. Xu, Y. & Her, C. Inhibition of topoisomerase (DNA) I (TOP1): DNA damage repair and anticancer therapy. *Biomolecules* **5**, 1652–1670 (2015).
 52. Mialon, A. et al. DNA topoisomerase I is a cofactor for c-Jun in the regulation of epidermal growth factor receptor expression and cancer cell proliferation. *Mol. Cell. Biol.* **25**, 5040–5051 (2005).
 53. Chen, J., Elfiky, A., Han, M., Chen, C. & Saif, M. W. The role of Src in colon cancer and its therapeutic implications. *Clin. Colorectal Cancer* **13**, 5–13 (2014).
 54. Koh, H. M. et al. Aurora kinase A is a prognostic marker in colorectal adenocarcinoma. *J. Pathol. Transl. Med.* **51**, 32–39 (2017).
 55. Aderka, D., Stintzing, S. & Heinemann, V. Explaining the unexplainable: discrepancies in results from the CALGB/SWOG 80405 and FIRE-3 studies. *Lancet Oncol.* **20**, e274–e283 (2019).
 56. Wang, C., Ouyang, C., Sandhu, J. S., Kahn, M. & Fakih, M. Wild-type APC and prognosis in metastatic colorectal cancer. *J. Clin. Orthod.* **38**, 223–223 (2020).
 57. Easwaran, V. et al. beta-Catenin regulates vascular endothelial growth factor expression in colon cancer. *Cancer Res.* **63**, 3145–3153 (2003).
 58. Meyer, E. L. et al. The evolution of master protocol clinical trial designs: a systematic literature review. *Clin. Ther.* **42**, 1330–1360 (2020).
 59. Stintzing, S. et al. FOLFIRI plus cetuximab versus FOLFIRI plus bevacizumab for metastatic colorectal cancer (FIRE-3): a post-hoc analysis of tumour dynamics in the final RAS wild-type subgroup of this randomised open-label phase 3 trial. *Lancet Oncol.* **17**, 1426–1434 (2016).
 60. Battaglin, F., Naseem, M., Lenz, H.-J. & Salem, M. E. Microsatellite instability in colorectal cancer: overview of its clinical significance and novel perspectives. *Clin. Adv. Hematol. Oncol.* **16**, 735–745 (2018).
 61. Fontana, E., Eason, K., Cervantes, A., Salazar, R. & Sadanandam, A. Context matters-consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials. *Ann. Oncol.* **30**, 520–527 (2019).
 62. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Stat. Methodol.* **57**, 289–300 (1995).
 63. Babur, Ö. et al. Pattern search in BioPAX models. *Bioinformatics* **30**, 139–140 (2014).
 64. Cerami, E. G. et al. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685–D690 (2011).
 65. Paz, A. et al. SPIKE: a database of highly curated human signaling pathways. *Nucleic Acids Res.* **39**, D793–D799 (2011).
 66. Fazekas, D., Koltai, M. & Türei, D. Signalink 2—a signaling pathway resource with multi-layered regulatory networks. *BMC Syst. Biol.* **7**, 7 (2013).
 67. Bouckaert, R. R. & Frank, E. Evaluating the replicability of significance tests for comparing learning algorithms. *Advances in Knowledge Discovery and Data Mining*, 3–12 (Springer, 2004).
 68. Seibold, H., Zeileis, A. & Hothorn, T. Model4you: an R package for personalised treatment effect estimation. *J. Open Res. Softw.* **7**, 17 (2019).
 69. Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S. & Zhu, R. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *J. R. Stat. Soc. Series B Stat. Methodol.* **85**, 179–211 (2023).
 70. Athey, S., Tibshirani, J. & Wager, S. Generalized random forests. *AOS* **47**, 1148–1178 (2019).
 71. Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S. & Wager, S. policytree: policy learning via doubly robust empirical welfare maximization over trees. *J. Open Source Softw.* **5**, 2232 (2020).
 72. Huling, J. D. & Yu, M. Subgroup identification using the personalized package. *J. Stat. Softw.* **98**, 1–60 (2018).

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 950293, M.P.M.). The clinical study received industrial funding from Merck KGaA, Darmstadt, Germany and Pfizer GmbH, Germany. The transcriptome-based microarray for gene expression using Xcel® Array received funding from Almac Ltd, Belfast, UK. The FoundationOne® based sequencing analysis (MSI) received funding from Roche Pharma AG, Grenzach, Germany (grant numbers: n/a, V.H., S.S.).

Author contributions

Conceptualisation, M.P.M. and V.H.; Data curation, A.S., S.S., D.P.M., U.V., T.D., M.M. and A.J.O.; Formal analysis, A.J.O.; Methodology, A.J.O., A.S. and M.P.M.; Supervision, V.H. and M.P.M.; Visualisation, A.J.O. and L.H.; Writing original draft, A.J.O., A.S., V.H. and M.P.M.; Writing, review and editing, all authors.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

A.S. served on advisory boards for BMS and Novocure, received honoraria for talks by Roche, Servier and Taiho Pharmaceuticals and

received reimbursement for travel by Roche, Merck KGaA, MSD Sharp & Dohme, Pfizer, Lilly Oncology, and Amgen. V.H., S.S. and D.P.M. received honoraria for talks, advisory boards and travel expenses by Merck KGaA, Amgen, Roche, Pfizer, BMS, MSD, AstraZeneca, Novartis, Terumo, Oncosil, Nordic, Seagen, GSK, Takeda, Servier, Pierre Fabre, Taiho, Lilly Oncology, Servier, Sanofi and Bayer Pharmaceuticals. M.P.M. is a former employee at AstraZeneca, academically collaborates with AstraZeneca, GSK and Roche, and receives funding from GSK and Roche. J.W.H. served on an advisory board for Roche, has received honoraria from Roche, and travel support from Novartis. M.M. received honoraria for advisory boards or talks by Amgen, BMS, Roche, Merck KGaA, MSD Sharp & Dohme, Lilly Oncology, Servier, Pierre Fabre, Taiho Sanofi and Bayer Pharmaceuticals and serves as officer for the European Organisation on Research and Treatment of Cancer (EORTC), and Arbeitsgemeinschaft internistische Onkologie (AIO). C.B.W. has received honoraria from Amgen, Bayer, Chugai, Celgene, GSK, MSD, Merck, Janssen, Ipsen, Roche, Servier, SIRTEx, Taiho; served on advisory boards for Bayer, BMS, Celgene, Servier, Shire/Baxalta, Rafael Pharmaceuticals, RedHill, Roche, has received travel support by Bayer, Celgene, RedHill, Roche, Servier, Taiho and research grants (institutional) by Roche. C.B.W. serves as an officer for the European Society of Medical Oncology (ESMO), Deutsche Krebshilfe (DKH) and AIO. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-023-41011-4>.

Correspondence and requests for materials should be addressed to Volker Heinemann or Michael P. Menden.

Peer review information *Nature Communications* thanks Saskia Wilting and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

Chapter 3

Discussion

This chapter delves into a more holistic discussion of the presented results in Chapter 2 compared to the discussion sections from each of the included articles or preprints. Thereby, Section 3.1 starts with the reiteration of the aims and findings in Chapter 2, and subsequently provides an integrated contextualisation of the conclusions, properties and limitations of this work that are essential for interpreting it. Drawing from the gained insights from this work, Section 3.2 describes the shifting landscape of predictive biomarker discovery, outlines its current discourse in the scientific literature and lays the foundation for the evolution of this field. Finally, this thesis ends with a brief closing statement.

3.1 Conclusions

In this work, analysis frameworks using statistical and machine learning methods were designed and applied to discover predictive biomarkers for cancer treatments leveraging molecular profiling in various preclinical and clinical datasets.

First, it became apparent that DNA methylation may serve as viable predictive drug response biomarkers. To demonstrate this, Section 2.1 characterised the epigenetic component of drug response and mapped out the pharmacoeigenomic landscape of cancer. More specifically, it delivered a multi-omics integration framework that leveraged HTS in cancer cell lines, the profiling of DNA methylation, gene expression and somatic driver mutations in both cell lines and primary tumours for systematically building evidence of epigenetic drug response biomarkers. Thereby, it arrived at a resource of putative DNA methylation biomarkers, reproduced anticipated clinical associations and generated new hypotheses.

Secondly, the presented results suggested that EMT may be exploited as a putative cancer vulnerability because of its role in determining drug responses. For assessing this, Section 2.2 delivered a methodology for the causal assessment of EMT as a predictive biomarker in cancer cell lines. It yielded a set of compounds and the HSP90 inhibitor luminespib as the lead compound, which showed robust predictive power and high causal estimates across different EMT scoring methods and response readouts. Enrichments of TF target genes and GO biological processes in both responder cell lines and transcriptional responses provided mechanistic evidence for this association. In addition, this prediction was validated by demonstrating that EMT induction can sensitise cell lines to luminespib, thereby providing evidence for a causal component.

Lastly, the systemic discovery of predictive biomarkers from molecular profiling in RCTs can lead to refined patient stratifications for achieving higher drug efficacies. For this, Section 2.3 presented the OncoBird framework for outlining the molecular and biomarker landscape in oncology RCTs. It prioritises actionable predictive biomarkers through revealing molecular patterns and systematically screening for statistical interactions between mutually exclusive somatic mutations and treatment regimes, while considering exploratory tumour CMS subtypes or tumour sidedness as clinical standard. Its utility was assessed by benchmarking it with other commonly employed subgroup analysis methods and current treatment guidelines in metastatic colorectal cancer, for which it prioritised smaller subgroups with consistently higher treatment effects. Its generalisability was demonstrated in a second clinical trial, in which it reproduced known associations and proposed new stratification strategies. Thereby, the OncoBird R package, its dockerised Shiny application and its reproducible demo application promote its reusability and interoperability for subgroup analyses in other clinical trials.

Despite the demonstrated effectiveness of the three proposed analysis frameworks, the presented datasets, methodologies and results are subject to their associated assumptions and caveats that constitute their properties and limitations. While these considerations have been laid out in each section of Chapter 2, more encompassing properties and limitations are discussed in this section. In particular, this section focuses on discussions regarding biases, complexity, causality, translatability and feasibility of the used datasets, the presented methodologies and ultimately their proposed predictive biomarkers.

3.1.1 Biases in drug high-throughput screens and clinical trials and their analysis

Data-driven biomarker discovery relies on efforts to design unbiased experiments. However, inherent biases in HTSs and RCTs can inadvertently lead to inaccurate representations of the studied phenomena and therefore false conclusions.

For example, the set of included cancer cell lines and compounds in the HTSs used in Sections 2.1 and 2.2 can introduce a selection bias, e.g., commercially available and well-established cell lines or drug libraries can fail to catch the full diversity of cancers and drug targets. Additionally, possible biases in producing HTSs are varying experimental protocols or culturing. For instance, the GDSC1 data release used Syto60 cell viability assays, whereas the GDSC2 release used CellTiter-Glo [234]. Furthermore, cell lines can have different growth properties and require different screening mediums, which were included in the built models as potential confounding variables. These caveats become more apparent when increasing throughput, since experimental parameters cannot

be optimised specifically for each cell line or compound anymore. Finally, during the downstream analysis, batch effects and inconsistent data processing procedures can introduce biases. For example, curve fitting methods for IC_{50} and AUC as summary metrics can vary depending on the used curve fitting method [237, 238, 239, 240]. Inconsistencies that can arise in large HTS efforts have been debated before [325, 326]. Therefore, testing the derived biomarkers in experiments with independent cell lines, alternative molecular profiling and HTSs with different protocols can help assess biases and ensure the robustness of the given conclusions. An example of a study that focused on robust and translatable drug response biomarkers is a meta-analysis of seven large HTS datasets [327].

For the RCTs in Section 2.3, assessing potential biases is necessary for ensuring valid conclusions from retrospective biomarker discovery efforts. Especially the selection bias when estimating treatment effect in subgroups was approached by resampling strategies. However, selection bias can still occur if trial participants are not representative of the subgroup populations, to which the yielded results may then not generalise well. While no validation of these results in independent clinical trials was performed, resampling strategies in the designed subgroup analysis methodology in Section 2.3 were used to assess the confidence and stability of the yielded results and cross-validation was used to evaluate the found subgroups.

Complementary to HTSs and RCTs, tumour molecular profiling was primarily used for assessing predictive biomarkers. Both the analysis of HTSs and RCTs focused on somatic mutations and gene expression, while Section 2.1 expanded this to DNA methylation¹. Somatic mutations in RCTs are often given by clinically validated targeted sequencing, which limits the discovery of biomarkers on the selected cancer genes. In contrast, WES or WGS efforts in cancer cell lines allow a more unbiased assessment of somatic cancer mutations. Especially in Section 2.1, DNA methylation and its association with drug responses was systematically assessed at the resolution of single CpG sites. However, while the Infinium BeadChip arrays contained about 450,000 CpG sites, the whole human genome contains roughly 29 million CpGs [328]. A few biological signals could be highlighted by integrating transcriptomic data, however, the effect sizes were rather small for the remaining methylation sites. To alleviate this, molecular profiling is commonly summarised into stronger biological signals, i.e. EMT signatures and CMS in Sections 2.2 and 2.3, respectively. This choice can bias the analysis by loss of information, but the introduced bias can reduce model variance by increasing model stability, reducing dimensionality and handling variable multicollinearity.

In all sections of Chapter 2, cancer-specific modelling was performed, motivated by the context-specificity of many cancer mechanisms and the fact that they showed distinct data distributions. This stratification may lead to biased estimates either if cancer (sub-)types are not indicative of the underlying population differences or if their strata show imbalances. As an example for the latter, the resulting reduced sample sizes for each studied cancer (sub-)type required higher effect sizes to yield significant biomarkers due to reduced power, which leads to the inability to find significant biomarkers for small cancer (sub-)types. In general, small sample sizes often observed in subgroup analysis are a major bottleneck in estimating heterogeneous effects in clinical trials and observational studies [329]. To achieve the same power for detecting an overall treatment effect as interactions with the same magnitude, the sample size should be increased 4-fold [330]. For example, in Section 2.3, the proposed OncoBird model tended to choose small subgroups with higher treatment effects compared to other models, which can increase the risk of overfitting. Thus, it was necessary to implement a rigorous model complexity control with its limitations discussed in the following section.

Understanding and mitigating the highlighted biases is crucial for ensuring the accuracy, reproducibility, robustness, generalisability and overall reliability of the reported findings and their interpretation before their implementation for therapeutic applications. Since these biases are highly dependent on the study, the choices in designing the presented methods remain highly dependent on the available data.

¹When characterising cancer samples with molecular profiling, various technology-dependent technical limitations and biases in sequencing, sample preparation and computational processing steps emerge that are out of scope to this work.

3.1.2 Multifaceted biomarkers

Complex diseases exhibit complex biomarker landscapes [331], which renders the prediction of drug responses a challenging task [332]. Predictive modelling for drug response in preclinical and clinical datasets often benefits from employing large global and nonlinear models. However, with a large amount of often correlated features in a small-sample regime $p \gg N$, allowing for interactions between features is usually infeasible. For example, if penalised regression including all CpG sites were used in Section 2.1, issues regarding multicollinearity, dimensionality and stability would have occurred. Similarly, employing nonlinear baseline models for Section 2.2 was unlikely to yield increased performances. In Section 2.3, the models were guided by prior derivation of mutually exclusive features, which introduced regularised nonlinearities to the statistical models. Thus, multiple statistical models, each including a single biomarker as predictor, were considered to reduce the complexity. This choice also facilitated the interpretation of the models in terms of their candidate biomarkers, since the interpretation is facilitated by this design. For example, the interpretation of the coefficients for a model built in the context of a particular drug or cancer type is directly interpretable without requiring feature importance methods.

As mentioned in Section 3.1.1, the choice of lower model complexity was beneficial for the small sample size regimen. While this approach can be also prone to model misspecification and bias the model coefficients, it did not increase variance and outperformed models with higher complexities. This approach resulted in employing individual linear regression model fits, e.g. single linear models for each CpG and drug in Section 2.1, single models for each drug and EMT signature in Section 2.2, and single models for each somatic mutation in Section 2.3.

In summary, the chosen designs facilitated the discovery of single predictive features in the studied datasets. Due to disease complexity, it is highly likely that the proposed univariate biomarkers underestimate the true complexity of molecular response mechanisms in the investigated contexts. While it would be possible to attempt building larger and more complex models that may further optimise the performance of the associated prediction tasks, these models are not optimal for the applications of this work due to their inefficiency in extracting context-dependent biomarkers and *de novo* mechanisms that are biologically meaningful.

3.1.3 Evaluation of causality

If the discovery of causal biological mechanisms for drug response is put as the central interest, it seems natural to formulate predictive biomarker discovery as a causal discovery problem instead of a prediction task. While recent methodological advances have already started expanding on this, e.g. causal feature selection [333], many study designs and datasets can only reliably yield correlative conclusions. Thus, this work has focused on either qualitative discussions regarding the causal component of the found relationships or quantitative estimations and validations of their causal contributions in Chapter 2. When assessing these causal relationships, it is crucial to carefully evaluate potential confounders and biases in the data and assumptions of the used methods to avoid concluding spurious relationships and false positive findings.

For example, since CpG sites showed correlated DNA methylation among neighbouring sites in Section 2.1, the spatial correlations of CpG methylation for calling differentially methylated regions were taken into account by the employed calling algorithm. Additionally, the confounding bias by global methylation patterns likely stemming from DNA active (de-)methylation processes was considered by adjusting linear models for principal components. For the called regions, consistency in an independent drug screen and also with another DNA methylation profiling technology was achieved. However, the statistical power for validating the effects in independent cell lines was lacking due to drastically reduced sample sizes. Moreover, since the resulting DNA methylation regions were challenging to interpret mechanistically, correlations between DNA methylation, gene expression and genetic alterations were assessed. Concordances between DNA methylation and gene expression of proximal genes were found robust in independent primary tumour samples. Additionally, plausible causal biomarkers were assessed by interpreting protein-protein interaction networks between the putative drug targets and the methylation-associated genes, with additional evidence given by CRISPR knockouts and drug transcriptomic signatures of genes in these networks. Even if the theory on the causality between DNA methylation and gene expression is still unclear, as

stated in Section 1.2.2.3, the results suggest that both data modalities should be jointly evaluated, which suggests considering DNA methylation jointly with *SLFN11* expression as predictive biomarker [334] and supports debates for popular DNA methylation biomarkers such as *MGMT* [335]. In contrast, only weak correlations between DNA methylation and somatic alteration were observed. A lack of an underlying causal genetic component to determine drug response cannot be claimed; however, since only protein-coding regions were considered, somatic mutations in non-coding regions, which can have higher mutation rates, or the germline component may impact DNA methylation in the found regions. While the levels of correlative evidence for the found biomarkers suggest their causal components, it does not provide a quantitative methodology for inferring them. Despite the presented efforts to maximise interpretability, many biological mechanisms remain elusive. For example, in Section 2.1, *NEK9* was consistently hypermethylated and downregulated in melanoma cells responding to the NEDD8 inhibitor pevonedistat. The signalling network neighbourhoods produced from shortest paths between the potential biomarker *NEK9* and drug target NAE revealed their common involvement with proteins regulating the cell cycle, which does not provide insights into the exact mechanisms but a short-listed gene set which can be subject to future studies with narrowed scopes.

For clinical trials in Section 2.3, the randomised controlled designs allowed to draw average treatment effects in the investigated treatment contexts. However, inferring conditional average treatment effects relies on further assumptions such as unconfoundedness, i.e. subgroups stratified by covariates fulfil RCT properties. Additionally, the investigated somatic alterations are usually not independent. While the grouping of functionally similar mutually exclusive alterations in the context of tumour subtypes improved the interpretability of the proposed biomarkers, the co-occurring alterations were harder to interpret because the causal component could not be resolved without any functional validation. Even if higher confidence in the derived biomarkers can be generated through multiplicity adjustments, resampling methods and the application and benchmarking of different statistical methods, the validation of the results suffers from the fundamental problem of causal inference, i.e. the inability to observe counterfactual ground truth outcomes. Therefore, any data-driven biomarker discovery effort in clinical studies ideally requires follow-up confirmatory RCTs for their approval.

In Section 2.2, the used methods for estimating the contribution of EMT to drug response phenotypes only yield valid causal relationships if all confounders are known and prognostic and propensity models are not misspecified and sufficiently well-estimated. For example, other confounders beyond the investigated genetic alterations could be present since EMT is a highly dynamic effect that can be triggered by other components, for example, by upstream regulators such as MITF or signals from unknown variables such as the tumour microenvironment. The significant associations of multiple HSP90 inhibitors, consistency with independent drug screens and validation experiments showing that TGF- β sensitises melanoma cells demonstrate substantial evidence towards a causal component of EMT and its regulators such as MITF². HSP90 has hundreds of protein clients which renders the pinpointing of the exact causal protein difficult, however, transcriptional responses in SKCM cancer cell lines revealed the downregulation of TGF- β signalling components upon treatment with luminespib. However, the mechanisms behind the found sensitisation that connects these components to MITF remain subject to speculation.

Altogether, for the presented work, the identification and rigorous validation of supporting biologically plausible mechanisms of the proposed predictive biomarkers was of pivotal interest. For this, the results in Section 2.2 demonstrated that causal modelling for predictive biomarker discovery is still rather impractical because of disease complexity, often high dimensions p compared to limited sample sizes N and unknown confounders. Thus, the majority of this work resorted to associative methods that have been more suitable for an exploratory hypothesis generation.

3.1.4 Challenges for the translation of biomarkers from preclinical to clinical studies

Translating proposed biomarkers derived from preclinical cancer models, such as cancer cell lines, to clinical studies is challenging. This domain shift is the main bottleneck for predictive biomarker discovery and can introduce

²Note that the classic EMT is not present in melanoma, since melanoma cells stem from melanocytes that are not epithelial cells [63]. Nonetheless, melanomas utilise EMT TFs to regulate their phenotype plasticity [336], and thus were labelled either epithelial-like or mesenchymal-like.

substantial bias, which, as stated in Section 3.1.1, needs to be addressed prior to clinical studies. For example, in Section 2.1, the domain shifts due to artefacts inherent in cancer cell lines were mitigated by matching them to primary tumour samples, whereby consistent correlations between DNA methylation and gene expression were subsequently filtered. As stated in Sections 3.1.3 and 3.1.2, it would be desirable to employ causal nonlinear methods integrating multiple data types and sources in order to yield better response predictions and make claims about nonlinear biomarker relationships across studies to increase the likelihood of finding translatable mechanisms. However, since modelling drug response mechanisms in preclinical models itself already suffers from challenges in revealing complex and causal drug response biomarkers, this complicates matching data distributions between preclinical and clinical domains, which may require even better domain adaptation strategies and validation. Moreover, since cancer data is more readily accessible without functional readouts, leveraging chemical or genetic perturbations from *in vitro* cancer models can provide an enriched understanding of cancer mechanisms and vulnerabilities. In turn, this approach generates biological priors that can enhance the predictability of clinical drug responses.

Furthermore, the aims of predictive biomarker discovery of preclinical and clinical studies conceptually differ, as has been previously argued [166]. Accordingly, preclinical studies for biomarker discovery aim at proposing patients for a given treatment determined by the biomarker. This approach is interested in finding subgroups with high treatment effects for an otherwise poorly responding population. For instance, in the analysed HTS in Sections 2.1 and 2.2, for which the large bulk of cancer cell lines were non-responders with only a few responders. In contrast, clinical studies for biomarker discovery rather aim at the opposite, i.e. selecting a treatment for a given patient. While both of these aims seem consistent, biomarkers can differ depending on what question is asked. In order to make this clear, consider the biomarkers reported for RCTs in Section 2.3. While small subgroups with high treatment effects have been discovered by some methods, other benchmarked methods for subgroup analysis yielded larger subgroups with lower but still significant treatment effects characterised by other biomarkers. While the former proposes a biomarker-positive subgroup with high treatment effects and does not employ further considerations about the biomarker-negative subgroup, the latter proposes a treatment policy for a given biomarker-positive or biomarker-negative patient considering the available treatment options.

3.1.5 Establishing molecular profiling as predictive biomarkers in clinical studies

The current ESMO treatment guidelines suggest cetuximab for left-sided *RAS* wild-type tumours [320] and did not produce subgroups with significantly enhanced treatment effects in the cross-validation benchmarks, whereas significant treatment effects were found with OncoBird, MOB, policy learning (POL) and outcome-weighting (OWE). Thereby, OncoBird recovered the smallest subgroups with the highest treatment effects, whereas POL and OWE still recovered a smaller effect but larger subgroups.

Therefore, also the discovered biomarkers substantially differ and the contextualisation is still challenging even with maximised interpretability. For example, MOB selected the simplest model, i.e. selecting cetuximab treatment for wild-type tumours in either *RAS*, *BRAF*, *IRS2* or *NF1*. Thereby, mutations in *NF1* can mimic oncogenic *RAS* mutations and therefore are a biologically plausible predictive biomarker in metastatic COREAD [10]. MOB supported this hypothesis and the stability assessment of OncoBird showed that this mutually exclusive gene module was frequently identified. However, the stability assessment showed that other modules including *KRAS* and not including *NF1* were also frequently identified. Furthermore, comparing the employed subgroup discovery methods also only showed limited consistency beyond mutations in *RAS* or *BRAF*. For example, chr20q amplifications such as *TOP1* were only consistently identified by OncoBird since they are enriched in CMS2 and left-sided tumours, which favours the subtype-specific modelling strategy. However, since chr20q amplifications were observed to be mutually exclusive to *RAS* or *BRAF* mutations, these subgroups had considerable overlap. Thus, the different subgroup discovery methods tended to produce consistent subgroups. Therefore, treatment decision-making likely benefits from molecular profiling in the form of somatic alterations and CMS for this studied clinical trial, however, the exact selection of biomarkers requires further considerations.

Furthermore, these derived predictive biomarkers are only useful in the context of the two treatment options that were studied, e.g. cetuximab and bevacizumab and the chemotherapy backbones in Section 2.3. Since no

control arm without a targeted therapy was included, the predictive effects of cetuximab and bevacizumab cannot be distinguished, and their biomarkers may differ from single-drug RCTs. Furthermore, since both treatment arms contain the same chemotherapy backbone, its predictive component is independent of the administered treatment arm t and cannot be distinguished from other prognostic effects, rendering the predictive effects of the backbone prognostic in the context of the outcome function $f(\mathbf{X}, t)$ ³. Similarly, if $f(\mathbf{X}, t)$ is estimated for each treatment arm separately, the prognostic and predictive effects cannot be distinguished.

3.2 Outlook

It has become apparent that biomarker discovery is not primarily interested in predicting observable patterns but rather in understanding the relevant latent structures of the data, i.e. explanatory instead of predictive. Building on the insights and conclusions from this work, promising current and future developments are proposed and discussed here. These include the continued discovery of cancer vulnerabilities, synthetically lethal interactions, drug combinations, drug resistances and new targets from cancer hallmarks, facilitated by advancements in molecular profiling technologies and machine learning for predicting and explaining tasks in drug discovery and development.

3.2.1 Molecular profiling of tumour plasticity

While it is well-established that genetic components are the leading cause of cancer, this work supports the increasing appreciation of non-mutational cancer mechanisms and tumour plasticity that affects tumour responses, i.e. epigenetic drug response biomarkers in Section 2.1, EMT as a causal component for drug responses in the genetic background in Section 2.2, and somatic mutations in the context of transcriptional tumour subtypes in Section 2.3. Thus, the future study of tumour responses should encompass the integration of the molecular profiling technologies that capture dynamic cancer mechanisms and tumour plasticity.

For example, cells undergoing EMT can adopt hybrid states between the mesenchymal and epithelial state through diverse molecular programs in a bidirectional manner [337]. Instead of resorting to a one-dimensional molecular EMT score, it remains in question if sensitivity profiles are dependent on other dimensions, e.g. specific epithelial-mesenchymal plasticity mechanisms, EMT TF activities or metastable hybrid EMT states. Additionally, mechanisms for epigenetic plasticity, such as DNA methylation, assist tumours in acquiring hallmark capabilities [338]. Spatial and temporal modelling strategies based on specific molecular mechanisms will be increasingly required to interpret these dynamic regulations. For example, modelling EMT trajectories dependent on genetic and microenvironmental backgrounds can help discover interpretable cancer mechanisms [339].

While this work focused only on somatic mutations, gene expression and DNA methylation in bulk tumour cells, other data types are currently emerging that require further considerations. In particular, proteomic and metabolomic profiling of cancer cell lines can reveal meaningful cancer biology [231, 232]. Furthermore, molecular resolutions of cancer samples on the single-cell level and their spatial or temporal component can reveal refined contexts for molecular cancer mechanisms [201]. These data types were not used within the scope of this work but are exciting data sources for refined discoveries of potential cancer vulnerabilities within these dimensions.

3.2.2 Acquired drug resistance and drug combinations

The efforts towards modelling the response mechanisms of single compounds presented in this work extend into the study of drug resistance and drug combinations. Acquired drug resistance is thought to be caused by rare events in tumour evolution, for example, secondary mutations [340] or non-genetic tumour plasticity, the ability of a tumour to give rise to transient cell-to-cell heterogeneity in gene expression. Plasticity gives rise to a ‘primed’ cellular state that is resistant to the administered therapy [341]. For example, therapy resistance to the BRAF inhibitor vemurafenib is driven by rare pre-existing and pre-resistant cells [342].

³This is only the case if the predictive component of the chemotherapy backbone does not have synergistic or antagonistic interactions with the targeted treatments in the respective treatment arms.

A typical resistance mechanism is oncogenic bypassing, in which the initial target remains inhibited, but an alternative kinase adaptively activates to sustain oncogenic signals [343]. Accordingly, patients with *KRAS* wild-type metastatic COREAD treated with the EGFR inhibitor cetuximab often acquire secondary *KRAS*^{G13D} or *KRAS*^{G12R} mutations [344]. Another prominent example is the success of the BRAF inhibitor vemurafenib with high initial responses in *BRAF*^{V600E} mutant melanoma, for which common resistance mechanisms are acquiring mutations in *KRAS*, *NRAS* or *MAP2K1* [340]. Many resistance mechanisms were already elucidated for EGFR inhibition with gefitinib or erlotinib in lung cancers [345]. For example, the ERBB3-mediated PI3K activation via amplifications of *MET* occurs in about 20% of *EGFR*-mutant lung cancer patients [346]. While these discussed resistance mechanisms underlie a genetic cause, efforts are increasing in studying non-genetic tumour plasticity as resistance mechanisms that enable lineage switching or immune surveillance. Staying with this example, the phenotypic changes with EMT are also frequently observed concordant with resistance to EGFR inhibition [347]. As another example, sensitivity to treatment immune checkpoint inhibitors in NSCLC can be robustly predicted by high tumour mutational burden and high PD-L1 expression as a simple and the best-performing model in a recent community DREAM challenge [348]; however, lung cancers treated with immune checkpoint inhibitors can show adaptive resistance by upregulating alternative immune checkpoints [349].

Modelling acquired resistances requires longitudinal molecular profiling to characterise tumour plasticity and acquired mutations. For the former, epigenetics may provide refined insights into epigenetic reprogramming for studying altered tumour plasticity upon single compound treatments [350]. For example, the LINCS consortium has pioneered the study of transcriptional responses [210]; however, it does not sufficiently sample the heterogeneity of disease features \mathbf{X}_d due to its low number of studied cell lines and only covers short time scales to study intrinsic response mechanisms.

Tumours can acquire resistance through various ways, and hence most chemotherapies and targeted treatments are not curative when administered alone [351, 352]. This prompted the use of drug combinations guided by a handful of proposed principles, e.g. individual sufficient efficacy, differential MOAs to anticipate resistance or mutual exclusive toxicities [353, 354]. For instance, to anticipate acquired resistance of *BRAF*^{V600E} mutant melanoma, combination therapies of BRAF inhibitors with MEK inhibitors such as cobimetinib showed promising results in increasing progression-free survival [355]. For *BRAF*^{V600E} mutant in metastatic COREAD, combination therapies of vemurafenib and EGFR inhibitors such as cetuximab were proposed, which showed success in a recent phase II trial [356]. Many drug combinations in use today are effective because of independent drug action, i.e., they target different tumour subpopulations and, therefore, expand the responding cohort to the union of those populations [357]. In contrast, drug synergy, which is the case if a drug pair shows higher efficacy than their additive or independent effects, may provide deeper and more robust responses. However, drug synergies or antagonisms in drug combination HTS are rare, highly context-dependent and only provide insights about intrinsic responses or resistances on shorter time scales [358]. Thus, the quest of overcoming resistance with drug combinations is still holding back immense successes of targeted treatments, which is dependent on longitudinal molecular profiling and the exploitation of drug combinations.

3.2.3 Advances in modelling of response mechanisms in cancer

Models for understanding drug responses are increasingly using strategies that can process and integrate multiple data sources and modalities. Instead of posing the problem as a single drug response prediction task, it can be advantageous first to consider more fundamental tasks, i.e. learning latent representations of cells, patients, compounds and treatments as an initial step and make use of various techniques to interpret and understand these representations when leveraging these representations for downstream tasks such as drug response prediction.

For example, generative modelling with variational autoencoders (VAE) enables domain adaptations between molecular data on human tumours (TCGA) and cancer cell lines (GDSC/CCLE), either through pretraining [5, 359, 360] and few-shot learning [295], disentanglement between both representations using an adversarial loss [361], domain adaptation with an alignment and consistency loss [362] or data augmentation [5]. Furthermore, considering transcriptional responses in cancer cell lines can improve drug response prediction [363, 364], as it encodes clues for drug MOAs and causal response mechanisms. Additionally, integrating molecular fingerprints

or SMILES chemical structures can improve performances for predicting cell viability [256, 365] and transcriptional responses [366] of unseen compounds, or can be used to propose new chemical structures with suitable pharmacological properties using reinforcement learning [367].

In order to extract context-dependent mechanisms from these models, local feature importance scores need to be derived sample-wise, which can be computationally expensive when using Shapley values [163], but can yield interpretable insight into the biomarker landscape. Complementary, structured and interpretable latent representations can advance the discovery of biomarkers. This can be achieved through visible neural networks [368], which hierarchically include network layers that represent biological processes such as downstream effects of somatic mutations [164], cancer pathways from other molecular features [369], or through other ways such as attention modules [370]. Furthermore, an interpretable conditional VAE enables constraining the network architecture by known gene programs, which can be used for single-cell reference mapping while allowing the learning of *de novo* programs [371].

Modelling biological systems requires prior domain knowledge to overcome limitations in small, sparse and noisy biological data when using data-driven modelling approaches, preferably through ‘differentiable programs’ that are general-purpose trainable models tailored to a particular domain using appropriate priors [372]. Biological priors often can be formulated through graphs that represent different types of interactions between biological entities derived from knowledge databases mentioned in Section 1.5.3. The manual curation of these knowledge graphs takes enormous effort, and thus, recently, work has been directed towards facilitating the generation and integration of knowledge graphs [373]. For example, adopting heterogeneous graphs with graph entities representing drug and disease annotations enabled the discovery of drug-disease mechanisms [374]. Moreover, graph neural networks have been used for the discovery of new drug indications (drug repurposing) [375] or drug response prediction [376, 377].

Complementary to monotherapy screens, drug combination screens typically involve cell viability screening of drug pairs. There are many combination screens within only a handful of cell lines [378], with a few recent larger datasets [358], which enabled the discovery of synergistic drug combinations and their biomarkers using machine learning [379]. Other types of functional screens, such as CRISPR technologies, can provide refined insights into biological mechanisms by providing richer readouts [224], especially when combined with single-cell assays such as Drug-seq [380] or Perturb-seq [381]. The modelling of these chemical or genetic perturbations using a compositional perturbation autoencoder [382], graph neural networks [383] and transfer learning from single-cell data to bulk HTSs [384] can yield viable cancer targets and biomarkers.

Finally, fully data-driven methods can be combined with traditional mechanistic models. In contrast to machine learning methods, they are often used in systems pharmacology and provide a natural way to model and interpret causal disease and drug response mechanisms representing biochemical mechanisms and their kinetic parameters by ordinary differential equations [255]. They have been used before to make predictions using a mechanistic model of selected canonical cancer pathways and a handful of compounds for predicting cell viability [385] or transcriptional responses [386]. Especially for understanding resistance mechanisms of single drugs, mechanistic models can yield interpretable insights, e.g. for cetuximab in gastric cancer [387] or vemurafenib in BRAF^{V600E} mutant melanoma [388]. Similarly, a quantitative systems pharmacology model informed by transcriptomic data yielded refined insights into predictive biomarkers for immunotherapies [389]. In future, mixing and matching different model components with other theoretical foundations to the existing statistical and machine learning methods has the potential to advance the limited interpretability of current methods for predictive biomarker discovery.

3.2.4 Advances in estimating treatment effects in cancer clinical studies

Subgroup analysis methods, such as those benchmarked in Section 2.3, have been increasingly incorporating machine learning and causal inference methods to estimate heterogeneous treatment effects in clinical trials or observational studies. Thereby, the literature has started to propose modelling strategies for specialised problems and usually focuses on estimating individual treatment effects.

For example, the Bayesian additive tree (BART) methodology can consider heterogeneity and uncertainty

across all provided covariates [390] as well as implicitly account for multiplicity by correcting for variable selection with an appropriate choice of model prior probabilities [391]. Furthermore, treatment-balanced generative adversarial neural networks were proposed for estimating individual treatment effects [392]. However, only recently, related methods were proposed for censored outcomes. For example, treatment-balanced Cox regression models that replace the linear predictor function in equation 1.36 by an arbitrary neural network (BITES) [308] or a neural network that estimates time-discrete and treatment-specific conditional hazard functions (SurvITE) [393]. As an example of another specialised problem, instead of modelling drug response by a summary statistic as in Section 2.2, adding drug dosage as a parameter enables the estimation of individual dose-response curves [394].

While these methods focus on patient outcomes Y in the presence of baseline disease features \mathbf{X}_d , other datasets could expand on predicting longitudinal components. Electronic health records (EHR) that record longitudinal hospital visits, diagnoses and prescriptions have especially received increasing attention. In conjunction with long short-term memory (LSTM) models with inverse probability treatment weighting, EHRs were used for drug repurposing [395]. For longitudinal data, the bias from time-varying confounders can be alleviated by balancing time-dependent representation of patient histories [396]. SyncTwin is a recurrent neural network (RNN) which learns time-invariant representation from the observed pretreatment outcomes in order to make counterfactual predictions for a single-time binary treatment [397]. A causal transformer combines these methods using three transformer subnetworks for each time-varying covariates, previous treatments, and previous outcomes [398].

Despite these advances, the discovery of predictive biomarkers in clinical studies requires moving from estimations of heterogeneous treatment effects to interpretable and actionable insights. The highlighted methods are not specifically designed for this purpose, and thus, additional considerations for the derivation of putative predictive biomarkers are required. In principle, this can be achieved through using *post hoc* feature importance methods [399] or tree-based policy learning for estimating optimal treatment regimens [400, 401]. Since the presented framework in Section 2.3 is inherently interpretable, it could be combined with some of the mentioned methods, either as a preselection of likely predictive variables, or as a feature interpreter when applied to the predictions of a backbone treatment effect estimation model.

3.2.5 Enabling virtual drug discovery and treatment recommendations

A natural extension to this work is combining molecular response modelling *in vitro* and treatment effect estimation in clinical data for a more integrated drug discovery and development compared to the traditional drug discovery pipelines [7]. Comprehensive sets of response biomarkers have been included in precision oncology applications for oncogenomic reporting and interpretation [402]. While platforms such as this can successfully give treatment recommendations, usually no predictive modelling is performed. Coupled with generative chemistry and library design principles, these models can enable virtual biomarker screening and *de novo* generation of interesting small molecules [403]. On the other hand, treatment recommendation systems essentially predict the treatment effects of a given set of treatments for a given query tumour instance. Combining these two approaches, generated compounds could be automatically queued into a treatment recommendation system to receive clues for their efficacy and associated biomarkers in clinical practice. In order to enable these advanced treatment recommendation systems using predictive modelling, they need to be trained with substantial coverage of the disease features \mathbf{X}_d , drug (treatment) features \mathbf{X}_t and other priors of known relationships such as knowledge databases containing clinically used pharmacogenomic variants [404]. Furthermore, since counterfactual outcomes are unknown, any treatment recommendation system will require a confirmatory clinical trial that is able to confirm the efficacy of its proposed treatments.

3.3 Closing statement

Predictive biomarkers define tailored subgroups in which they predict treatment successes for a given treatment. Their discovery is fuelled by datasets that provide information on tumours and treatments that allow drawing conclusions about their relationships. This thesis demonstrated that biomarker discovery integrates into both pre-

clinical and clinical stages of drug development and promotes the understanding of drug response patterns and their associated mechanisms. The growing ability of models for biological, biochemical and biomedical entities and processes to make accurate and interpretable predictions enables biomarkers to become more informative and nuanced and therefore more suitable for developing curative therapies.

References

- [1] Ohnmacht, A. J. *et al.* The pharmacopigenomic landscape of cancer cell lines reveals the epigenetic component of drug sensitivity. *Communications Biology* **6** (2023). URL <https://doi.org/10.1038/s42003-023-05198-y>.
- [2] Ohnmacht, A. J. *et al.* The pharmacogenomic assessment of molecular epithelial-mesenchymal transition signatures reveals drug susceptibilities in cancer cell lines. *bioRxiv* (2024). URL <http://dx.doi.org/10.1101/2024.01.16.575190>.
- [3] Ohnmacht, A. J. *et al.* The Oncology Biomarker Discovery framework reveals cetuximab and bevacizumab response patterns in metastatic colorectal cancer. *Nature Communications* **14** (2023). URL <https://doi.org/10.1038/s41467-023-41011-4>.
- [4] Holch, J. W. *et al.* Refining first-line treatment decision in RAS wildtype (RAS-WT) metastatic colorectal cancer (mCRC) by combining clinical biomarkers: Results of the randomized phase 3 trial FIRE-3 (AIO KRK0306). *Journal of Clinical Oncology* **42**, 13–13 (2024). URL http://dx.doi.org/10.1200/JCO.2024.42.3_suppl.13.
- [5] Lu, D., Pamar, D. P., Ohnmacht, A. J., Kutkaite, G. & Menden, M. P. Enhancing gene expression representation and drug response prediction with data augmentation and gene emphasis. *bioRxiv* (2024). URL <http://dx.doi.org/10.1101/2024.05.15.592959>.
- [6] Vosberg, S. *et al.* DNA methylation profiling refines the prognostic classification of acute myeloid leukemia patients treated with intensive chemotherapy. *HemaSphere* **6**, 378–379 (2022). URL <https://doi.org/10.1097/01.hs9.0000844804.00811.c6>.
- [7] Boniolo, F. *et al.* Artificial intelligence in early drug discovery enabling precision medicine. *Expert Opinion on Drug Discovery* **16**, 991–1007 (2021). URL <https://doi.org/10.1080/17460441.2021.1918096>.
- [8] Nguyen, P. B. H., Ohnmacht, A. J., Sharifli, S., Garnett, M. J. & Menden, M. P. Inferred ancestral origin of cancer cell lines associates with differential drug response. *International Journal of Molecular Sciences* **22**, 10135 (2021). URL <https://doi.org/10.3390/ijms221810135>.
- [9] Farnoud, A., Ohnmacht, A. J., Meinel, M. & Menden, M. P. Can artificial intelligence accelerate preclinical drug discovery and precision medicine? *Expert Opinion on Drug Discovery* **17**, 661–665 (2022). URL <https://doi.org/10.1080/17460441.2022.2090540>.
- [10] Weinberg, R. A. *The Biology of Cancer* (W.W. Norton & Company, 2013).
- [11] Relling, M. V. & Evans, W. E. Pharmacogenomics in the clinic. *Nature* **526**, 343–350 (2015).
- [12] Strimbu, K. & Tavel, J. A. What are biomarkers? *Current Opinion in HIV and AIDS* **5**, 463–466 (2010).
- [13] Suehnholz, S. P. *et al.* Quantifying the expanding landscape of clinical actionability for patients with cancer. *Cancer Discovery* **14**, 49–65 (2023).

- [14] Boveri, T. *Zur Frage der Entstehung maligner Tumoren* (Fischer, 1914).
- [15] Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
- [16] Alberts, B. *et al. Molecular Biology of the Cell* (W.W. Norton & Company, 2017).
- [17] Parada, L. F., Tabin, C. J., Shih, C. & Weinberg, R. A. Human EJ bladder carcinoma oncogene is homologue of harvey sarcoma virus ras gene. *Nature* **297**, 474–478 (1982).
- [18] Seger, R. & Krebs, E. G. The MAPK signaling cascade. *The FASEB Journal* **9**, 726–735 (1995).
- [19] Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
- [20] Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).
- [21] Hanahan, D. Hallmarks of cancer: New dimensions. *Cancer Discovery* **12**, 31–46 (2022).
- [22] Cohen, P., Cross, D. & Jänne, P. A. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nature Reviews Drug Discovery* **20**, 551–569 (2021).
- [23] Puneekar, S. R., Velcheti, V., Neel, B. G. & Wong, K.-K. The current state of the art and future trends in RAS-targeted cancer therapies. *Nature Reviews Clinical Oncology* **19**, 637–655 (2022).
- [24] Mardis, E. R. & Wilson, R. K. Cancer genome sequencing: a review. *Human Molecular Genetics* **18**, R163–R168 (2009).
- [25] Sondka, Z. *et al.* The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer* **18**, 696–705 (2018).
- [26] Sanchez-Vega, F. *et al.* Oncogenic signaling pathways in the cancer genome atlas. *Cell* **173**, 321–337.e10 (2018).
- [27] Chakravarty, D. *et al.* Oncokb: A precision oncology knowledge base. *JCO Precision Oncology* 1–16 (2017).
- [28] Chin, L., Andersen, J. N. & Futreal, P. A. Cancer genomics: from discovery science to personalized medicine. *Nature Medicine* **17**, 297–303 (2011).
- [29] Hoeijmakers, J. H. Dna damage, aging, and cancer. *New England Journal of Medicine* **361**, 1475–1485 (2009).
- [30] Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
- [31] Futreal, P. A. *et al.* A census of human cancer genes. *Nature Reviews Cancer* **4**, 177–183 (2004).
- [32] Elliott, K. & Larsson, E. Non-coding driver mutations in human cancer. *Nature Reviews Cancer* **21**, 500–509 (2021).
- [33] Liu, E. M., Martinez-Fundichely, A., Bollapragada, R., Spiewack, M. & Khurana, E. CNCDatabase: a database of non-coding cancer drivers. *Nucleic Acids Research* **49**, D1094–D1101 (2020).
- [34] Zhang, X. & Meyerson, M. Illuminating the noncoding genome in cancer. *Nature Cancer* **1**, 864–872 (2020).
- [35] Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002).
- [36] Stephens, P. *et al.* Intragenic ERBB2 kinase mutations in tumours. *Nature* **431**, 525–526 (2004).
- [37] Samuels, Y. *et al.* High frequency of mutations of the PIK3ca gene in human cancers. *Science* **304**, 554–554 (2004).

-
- [38] Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
- [39] Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
- [40] Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics* **11**, 685–696 (2010).
- [41] Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
- [42] Cortés-Ciriano, I., Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M. & Park, P. J. Computational analysis of cancer genome sequencing data. *Nature Reviews Genetics* **23**, 298–314 (2021).
- [43] ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578** (2020).
- [44] Bewicke-Copley, F., Kumar, E. A., Palladino, G., Korfi, K. & Wang, J. Applications and analysis of targeted genomic sequencing in cancer studies. *Computational and Structural Biotechnology Journal* **17**, 1348–1359 (2019).
- [45] Stahler, A. *et al.* Single-nucleotide variants, tumour mutational burden and microsatellite instability in patients with metastatic colorectal cancer: Next-generation sequencing results of the FIRE-3 trial. *European Journal of Cancer* **137**, 250–259 (2020).
- [46] Powell, S. M. *et al.* Apc mutations occur early during colorectal tumorigenesis. *Nature* **359**, 235–237 (1992).
- [47] Rajagopalan, H. *et al.* RAF/RAS oncogenes and mismatch-repair status. *Nature* **418**, 934–934 (2002).
- [48] Morkel, M., Riemer, P., Bläker, H. & Sers, C. Similar but different: distinct roles for KRAS and BRAF oncogenes in colorectal cancer development and therapy resistance. *Oncotarget* **6**, 20785–20800 (2015).
- [49] Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
- [50] Cerami, E. *et al.* The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* **2**, 401–404 (2012).
- [51] Kessler, D. *et al.* Drugging an undruggable pocket on KRAS. *Proceedings of the National Academy of Sciences* **116**, 15823–15829 (2019).
- [52] Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
- [53] Witsch, E., Sela, M. & Yarden, Y. Roles for growth factors in cancer progression. *Physiology* **25**, 85–101 (2010).
- [54] Rubin, I. & Yarden, Y. The basic biology of her2. *Annals of Oncology* **12**, S3–S8 (2001).
- [55] Derynck, R., Zhang, Y. & Feng, X.-H. Transcriptional activators of TGF- β responses: Smads. *Cell* **95**, 737–740 (1998).
- [56] Moustakas, A. & Heldin, C.-H. Non-smad TGF- β signals. *Journal of Cell Science* **118**, 3573–3584 (2005).
- [57] Sever, R. & Brugge, J. S. Signal transduction in cancer. *Cold Spring Harbor Perspectives in Medicine* **5**, a006098–a006098 (2015).

REFERENCES

- [58] Wang, Z., Gerstein, M. & Snyder, M. RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57–63 (2009).
- [59] Cieřlik, M. & Chinnaiyan, A. M. Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics* **19**, 93–109 (2017).
- [60] Dudley, J. T., Tibshirani, R., Deshpande, T. & Butte, A. J. Disease signatures are robust across tissues and experiments. *Molecular Systems Biology* **5**, 307 (2009).
- [61] Golub, T. R. *et al.* Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
- [62] Sørli, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* **98**, 10869–10874 (2001).
- [63] Tsoi, J. *et al.* Multi-stage differentiation defines melanoma subtypes with differential vulnerability to drug-induced iron-dependent oxidative stress. *Cancer Cell* **33**, 890–904.e5 (2018).
- [64] Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nature Medicine* **21**, 1350–1356 (2015).
- [65] Stintzing, S. *et al.* Consensus molecular subgroups (cms) of colorectal cancer (crc) and first-line efficacy of foliri plus cetuximab or bevacizumab in the fire3 (aio krk-0306) trial. *Annals of Oncology* **30**, 1796–1803 (2019).
- [66] Thiery, J. P., Acloque, H., Huang, R. Y. & Nieto, M. A. Epithelial-mesenchymal transitions in development and disease. *Cell* **139**, 871–890 (2009).
- [67] Yang, J. *et al.* Guidelines and definitions for research on epithelial–mesenchymal transition. *Nature Reviews Molecular Cell Biology* **21**, 341–352 (2020).
- [68] Nieto, M. A., Huang, R. Y.-J., Jackson, R. A. & Thiery, J. P. EMT: 2016. *Cell* **166**, 21–45 (2016).
- [69] Sánchez-Tilló, E. *et al.* ZEB1 represses e-cadherin and induces an EMT by recruiting the SWI/SNF chromatin-remodeling protein BRG1. *Oncogene* **29**, 3490–3500 (2010).
- [70] Dongre, A. & Weinberg, R. A. New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer. *Nature Reviews Molecular Cell Biology* **20**, 69–84 (2018).
- [71] Pećina-Šlaus, N. Tumor suppressor gene e-cadherin and its role in normal and malignant cells. *Cancer Cell International* **3**, 17 (2003).
- [72] Pastushenko, I. *et al.* Identification of the tumour transition states occurring during EMT. *Nature* **556**, 463–468 (2018).
- [73] Shibue, T. & Weinberg, R. A. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. *Nature Reviews Clinical Oncology* **14**, 611–629 (2017).
- [74] Visal, T. H., den Hollander, P., Cristofanilli, M. & Mani, S. A. Circulating tumour cells in the -omics era: how far are we from achieving the ‘singularity’? *British Journal of Cancer* **127**, 173–184 (2022).
- [75] Waddington, C. H. The epigenotype. *International Journal of Epidemiology* **41**, 10–13 (2011).
- [76] Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187**, 226–232 (1975).
- [77] Ooi, S. K. T. *et al.* DNMT3l connects unmethylated lysine 4 of histone h3 to de novo methylation of DNA. *Nature* **448**, 714–717 (2007).

-
- [78] Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics* **39**, 457–466 (2007).
- [79] Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics* **37**, 853–862 (2005).
- [80] Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* **13**, 484–492 (2012).
- [81] Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23–38 (2012).
- [82] Carmona, J. J. *et al.* Empirical comparison of reduced representation bisulfite sequencing and Infinium BeadChip reproducibility and coverage of DNA methylation in humans. *npj Genomic Medicine* **2** (2017).
- [83] Bibikova, M. *et al.* Genome-wide DNA methylation profiling using Infinium assay. *Epigenomics* **1**, 177–200 (2009).
- [84] Sun, Z., Cunningham, J., Slager, S. & Kocher, J.-P. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics* **7**, 813–828 (2015).
- [85] Beck, D., Maamar, M. B. & Skinner, M. K. Genome-wide CpG density and DNA methylation analysis method (MeDIP, RRBS, and WGBS) comparisons. *Epigenetics* **17**, 518–530 (2021).
- [86] Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes & Development* **25**, 1010–1022 (2011).
- [87] Smallwood, S. A. *et al.* Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nature Genetics* **43**, 811–814 (2011).
- [88] Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916–919 (2010).
- [89] Parry, A., Rulands, S. & Reik, W. Active turnover of DNA methylation during cell fate decisions. *Nature Reviews Genetics* **22**, 59–66 (2020).
- [90] Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356** (2017).
- [91] Raynal, N. J.-M. *et al.* DNA methylation does not stably lock gene expression but instead serves as a molecular mark for gene silencing memory. *Cancer Research* **72**, 1170–1181 (2012).
- [92] Rideout, W. M., Coetzee, G. A., Olumi, A. F. & Jones, P. A. 5-methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science* **249**, 1288–1290 (1990).
- [93] Jones, P. A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics* **3**, 415–428 (2002).
- [94] Merlo, A. *et al.* 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nature Medicine* **1**, 686–692 (1995).
- [95] Dawson, M. A. & Kouzarides, T. Cancer epigenetics: From mechanism to therapy. *Cell* **150**, 12–27 (2012).
- [96] Figueroa, M. E. *et al.* Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell* **18**, 553–567 (2010).

REFERENCES

- [97] Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2015).
- [98] Cimmino, L. *et al.* Restoration of TET2 function blocks aberrant self-renewal and leukemia progression. *Cell* **170**, 1079–1095.e20 (2017).
- [99] Russler-Germain, D. A. *et al.* The r882h DNMT3a mutation associated with AML dominantly inhibits wild-type DNMT3a by blocking its ability to form active tetramers. *Cancer Cell* **25**, 442–454 (2014).
- [100] Issa, J.-P. CpG island methylator phenotype in cancer. *Nature Reviews Cancer* **4**, 988–993 (2004).
- [101] Turcan, S. *et al.* IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* **483**, 479–483 (2012).
- [102] Hawkins, N. J. *et al.* MGMT methylation is associated primarily with the germline c>t SNP (rs16906252) in colorectal cancer and normal colonic mucosa. *Modern Pathology* **22**, 1588–1599 (2009).
- [103] Baylin, S. B. & Jones, P. A. Epigenetic determinants of cancer. *Cold Spring Harbor Perspectives in Biology* **8**, a019505 (2016).
- [104] Vivian, J. *et al.* Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology* **35**, 314–316 (2017).
- [105] Tew, K., Colvin, O. M. & Chabner, B. *Cancer chemotherapy and biotherapy: principles and practice* (Lippincott-Raven: Philadelphia, PA, 1996).
- [106] Kinch, M. S. An analysis of FDA-approved drugs for oncology. *Drug Discovery Today* **19**, 1831–1835 (2014).
- [107] Chabner, B. A. & Roberts, T. G. Chemotherapy and the war on cancer. *Nature Reviews Cancer* **5**, 65–72 (2005).
- [108] Weinstein, I. B. & Joe, A. Oncogene addiction. *Cancer Research* **68**, 3077–3080 (2008).
- [109] Mouridsen, H., Palshof, T., Patterson, J. & Battersby, L. Tamoxifen in advanced breast cancer. *Cancer Treatment Reviews* **5**, 131–141 (1978).
- [110] Imai, K. & Takaoka, A. Comparing antibody and small-molecule therapies for cancer. *Nature Reviews Cancer* **6**, 714–727 (2006).
- [111] Druker, B. J. *et al.* Effects of a selective inhibitor of the abl tyrosine kinase on the growth of bcr–abl positive cells. *Nature Medicine* **2**, 561–566 (1996).
- [112] Pao, W. *et al.* EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proceedings of the National Academy of Sciences* **101**, 13306–13311 (2004).
- [113] Luo, J., Solimini, N. L. & Elledge, S. J. Principles of cancer therapy: Oncogene and non-oncogene addiction. *Cell* **136**, 823–837 (2009).
- [114] Goding, C. R. A picture of mitf in melanoma immortality. *Oncogene* **30**, 2304–2306 (2011).
- [115] Pacini, C. *et al.* A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer Cell* (2024).
- [116] Shaffer, A. L. *et al.* IRF4 addiction in multiple myeloma. *Nature* **454**, 226–231 (2008).
- [117] Hahn, W. C. *et al.* An expanded universe of cancer targets. *Cell* **184**, 1142–1155 (2021).

- [118] Lord, C. J. & Ashworth, A. PARP inhibitors: Synthetic lethality in the clinic. *Science* **355**, 1152–1158 (2017).
- [119] Dziadkowiec, K. N., Gasiorowska, E., Nowak-Markwitz, E. & Jankowska, A. PARP inhibitors: review of mechanisms of action and BRCA1/2 mutation targeting. *Menopausal Review* **4**, 215–219 (2016).
- [120] Grbovic, O. M. *et al.* V600e b-raf requires the hsp90 chaperone for stability and is degraded in response to hsp90 inhibitors. *Proceedings of the National Academy of Sciences* **103**, 57–62 (2005).
- [121] Linnekamp, J., Butter, R., Spijker, R., Medema, J. & van Laarhoven, H. Clinical and biological effects of demethylating agents on solid tumours – a systematic review. *Cancer Treatment Reviews* **54**, 10–23 (2017).
- [122] Jones, P. A., Issa, J.-P. J. & Baylin, S. Targeting the cancer epigenome for therapy. *Nature Reviews Genetics* **17**, 630–641 (2016).
- [123] Turcan, S. *et al.* Efficient induction of differentiation and growth inhibition in IDH1 mutant glioma cells by the DNMT inhibitor decitabine. *Oncotarget* **4**, 1729–1736 (2013).
- [124] Pfister, S. X. & Ashworth, A. Marked for death: targeting epigenetic changes in cancer. *Nature Reviews Drug Discovery* **16**, 241–263 (2017).
- [125] Mohammad, H. P., Barbash, O. & Creasy, C. L. Targeting epigenetic modifications in cancer therapy: erasing the roadmap to cancer. *Nature Medicine* **25**, 403–418 (2019).
- [126] Olivier, T., Haslam, A. & Prasad, V. Anticancer drugs approved by the US food and drug administration from 2009 to 2020 according to their mechanism of action. *JAMA Network Open* **4**, e2138793 (2021).
- [127] Heilmeyer, L., Schoen, R. & Rudder, B. (eds.) *Ergebnisse der Inneren Medizin und Kinderheilkunde* (Springer Berlin Heidelberg, 1959).
- [128] McLeod, H. L. & Evans, W. E. Pharmacogenomics: Unlocking the human genome for better drug therapy. *Annual Review of Pharmacology and Toxicology* **41**, 101–121 (2001).
- [129] Evans, W. E. & Relling, M. V. Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science* **286**, 487–491 (1999).
- [130] Desmeules, J., Gascon, M.-P., Dayer, P. & Magistris, M. Impact of environmental and genetic factors on codeine analgesia. *European Journal of Clinical Pharmacology* **41**, 23–26 (1991).
- [131] Evans, W. E. & McLeod, H. L. Pharmacogenomics — drug disposition, drug targets, and side effects. *New England Journal of Medicine* **348**, 538–549 (2003).
- [132] Iorio, F. *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
- [133] Slamon, D. J. *et al.* Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New England Journal of Medicine* **344**, 783–792 (2001).
- [134] Slamon, D. *et al.* Adjuvant trastuzumab in HER2-positive breast cancer. *New England Journal of Medicine* **365**, 1273–1283 (2011).
- [135] McDermott, U. *et al.* Genomic alterations of anaplastic lymphoma kinase may sensitize tumors to anaplastic lymphoma kinase inhibitors. *Cancer Research* **68**, 3389–3395 (2008).
- [136] Li, H., van der Merwe, P. A. & Sivakumar, S. Biomarkers of response to PD-1 pathway blockade. *British Journal of Cancer* **126**, 1663–1675 (2022).

REFERENCES

- [137] Table of pharmacogenomic biomarkers in drug labeling. U.S. Food and Drug Administration (FDA) Website. Accessed: June 14, 2023.
- [138] Esteller, M. *et al.* Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *New England Journal of Medicine* **343**, 1350–1354 (2000).
- [139] Oldrini, B. *et al.* MGMT genomic rearrangements contribute to chemotherapy resistance in gliomas. *Nature Communications* **11** (2020).
- [140] Hegi, M. E. *et al.* MGMT gene silencing and benefit from temozolomide in glioblastoma. *New England Journal of Medicine* **352**, 997–1003 (2005).
- [141] Kamińska, K. *et al.* Prognostic and predictive epigenetic biomarkers in oncology. *Molecular Diagnosis & Therapy* **23**, 83–95 (2018).
- [142] Boumahdi, S. & de Sauvage, F. J. The great escape: tumour cell plasticity in resistance to targeted therapy. *Nature Reviews Drug Discovery* **19**, 39–56 (2019).
- [143] Adachi, Y. *et al.* Epithelial-to-mesenchymal transition is a cause of both intrinsic and acquired resistance to KRAS g12c inhibitor in KRAS g12c–mutant non–small cell lung cancer. *Clinical Cancer Research* **26**, 5962–5973 (2020).
- [144] Ayestaran, I. *et al.* Identification of intrinsic drug resistance and its biomarkers in high-throughput pharmacogenomic and CRISPR screens. *Patterns* **1**, 100065 (2020).
- [145] Pirmohamed, M. Pharmacogenomics: current status and future perspectives. *Nature Reviews Genetics* (2023).
- [146] Evans, W. E. & Relling, M. V. Moving towards individualized medicine with pharmacogenomics. *Nature* **429**, 464–468 (2004).
- [147] Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer New York, 2009).
- [148] Murphy, K. P. *Machine Learning: A Probabilistic Perspective* (The MIT Press, 2012).
- [149] Fay, M. P. & Brittain, E. H. *Statistical Hypothesis Testing in Context* (Cambridge University Press, 2022).
- [150] Miller, R. G. *Simultaneous Statistical Inference* (Springer New York, 1981).
- [151] Šidák, Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* **62**, 626–633 (1967).
- [152] Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70 (1979).
- [153] Hochberg, Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802 (1988).
- [154] Hommel, G. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika* **75**, 383–386 (1988).
- [155] Tamhane, A. C. & Gou, J. Advances in p-value based multiple test procedures. *Journal of Biopharmaceutical Statistics* **28**, 10–27 (2017).
- [156] Westfall, P. & Young, S. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley Series in Probability and Statistics (Wiley, 1993).
- [157] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

-
- [158] Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29** (2001).
 - [159] Goeman, J. J. & Solari, A. Multiple hypothesis testing in genomics. *Statistics in Medicine* **33**, 1946–1978 (2014).
 - [160] Lipkovich, I., Dmitrienko, A., Muysers, C. & Ratitch, B. Multiplicity issues in exploratory subgroup analysis. *Journal of Biopharmaceutical Statistics* **28**, 63–81 (2017).
 - [161] Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**, 301–320 (2005).
 - [162] Efron, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7** (1979).
 - [163] Janizek, J. D. *et al.* Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models. *Nature Biomedical Engineering* **7**, 811–829 (2023).
 - [164] Kuenzi, B. M. *et al.* Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* **38**, 672–684.e6 (2020).
 - [165] Xu, Y. *et al.* Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics* **71**, 645–653 (2015).
 - [166] Lipkovich, I., Dmitrienko, A. & D’Agostino, R. B. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine* **36**, 136–196 (2016).
 - [167] McCullagh, P. & Nelder, J. *Generalized Linear Models* (Routledge, 2019).
 - [168] Yusuf, S. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA: The Journal of the American Medical Association* **266**, 93 (1991).
 - [169] Rosenkranz, G. *Exploratory Subgroup Analyses in Clinical Research* (Wiley, 2019).
 - [170] Assmann, S. F., Pocock, S. J., Enos, L. E. & Kasten, L. E. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet* **355**, 1064–1069 (2000).
 - [171] Feinstein, A. R. & Horwitz, R. I. Problems in the “evidence” of “evidence-based medicine”. *The American Journal of Medicine* **103**, 529–535 (1997).
 - [172] Hasford, J. *et al.* Inconsistent trial assessments by the national institute for health and clinical excellence and IQWiG: standards for the performance and interpretation of subgroup analyses are needed. *Journal of Clinical Epidemiology* **63**, 1298–1304 (2010).
 - [173] Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J. & Drazen, J. M. Statistics in medicine — reporting of subgroup analyses in clinical trials. *New England Journal of Medicine* **357**, 2189–2194 (2007).
 - [174] European Medicines Agency. Guideline on the investigation of subgroups in confirmatory clinical trials (2019).
 - [175] Dmitrienko, A. & D’Agostino, R. Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine* **32**, 5172–5218 (2013).
 - [176] Taylor, J. & Tibshirani, R. J. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences* **112**, 7629–7634 (2015).
 - [177] Ting, N., Cappelleri, J. C., Ho, S. & Chen, D. D.-G. (eds.) *Design and Analysis of Subgroups with Biopharmaceutical Applications* (Springer International Publishing, 2020).

REFERENCES

- [178] Feinstein, A. R. The problem of cogent subgroups: A clinicostatistical tragedy. *Journal of Clinical Epidemiology* **51**, 297–299 (1998).
- [179] Mayer, C., Lipkovich, I. & Dmitrienko, A. Survey results on industry practices and challenges in subgroup analysis in clinical trials. *Statistics in Biopharmaceutical Research* **7**, 272–282 (2015).
- [180] Imbens, G. W. & Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences* (Cambridge University Press, 2015).
- [181] Wager, S. Stats 361: Causal inference. *Stanford University Lecture Notes* (2020).
- [182] Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* **116**, 4156–4165 (2019).
- [183] Loh, W.-Y., Cao, L. & Zhou, P. Subgroup identification for precision medicine: A comparative review of 13 methods. *WIREs Data Mining and Knowledge Discovery* **9** (2019).
- [184] Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481 (1958).
- [185] Marchenko, O. V. & Katenka, N. V. (eds.) *Quantitative Methods in Pharmaceutical Research and Development* (Springer International Publishing, 2020).
- [186] Kvamme, H., Borgan, O. & Scheel, I. Time-to-event prediction with neural networks and cox regression. *arXiv* (2019).
- [187] Brookes, S. T. *et al.* Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technology Assessment* **5** (2001).
- [188] Rothwell, P. M. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet* **365**, 176–186 (2005).
- [189] Sun, X., Briel, M., Walter, S. D. & Guyatt, G. H. Is a subgroup effect believable? updating criteria to evaluate the credibility of subgroup analyses. *BMJ* **340**, c117–c117 (2010).
- [190] Foster, J. C., Taylor, J. M. & Ruberg, S. J. Subgroup identification from randomized clinical trial data. *Statistics in Medicine* **30**, 2867–2880 (2011).
- [191] Dmitrienko, A. & D’Agostino, R. B. Multiplicity considerations in clinical trials. *New England Journal of Medicine* **378**, 2115–2122 (2018).
- [192] Dmitrienko, A., Tamhane, A. C. & Bretz, F. (eds.) *Multiple Testing Problems in Pharmaceutical Statistics* (Chapman and Hall/CRC, 2009).
- [193] Dmitrienko, A., D’Agostino, R. B. & Huque, M. F. Key multiplicity issues in clinical drug development. *Statistics in Medicine* **32**, 1079–1111 (2012).
- [194] Efron, B. Simultaneous inference: When should hypothesis testing problems be combined? *The Annals of Applied Statistics* **2** (2008).
- [195] Benjamini, Y. & Bogomolov, M. Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 297–318 (2013).
- [196] Austin, P. C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* **46**, 399–424 (2011).
- [197] Lipkovich, I., Svensson, D., Ratitch, B. & Dmitrienko, A. Modern approaches for evaluating treatment effect heterogeneity from clinical trials and observational data. *arXiv* (2023).

- [198] Chernozhukov, V. *et al.* Double/debiased machine learning for treatment and causal parameters. *arXiv* (2016).
- [199] Foster, D. J. & Syrgkanis, V. Orthogonal statistical learning. *arXiv* (2019).
- [200] Athey, S., Tibshirani, J. & Wager, S. Generalized random forests. *The Annals of Statistics* **47** (2019).
- [201] Jiang, P. *et al.* Big data in basic and translational cancer research. *Nature Reviews Cancer* **22**, 625–639 (2022).
- [202] Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
- [203] Grossman, R. L. *et al.* Toward a shared vision for cancer genomic data. *New England Journal of Medicine* **375**, 1109–1112 (2016).
- [204] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Research* **37**, D26–D31 (2009).
- [205] Burgin, J. *et al.* The european nucleotide archive in 2022. *Nucleic Acids Research* **51**, D121–D125 (2022).
- [206] Edgar, R. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**, 207–210 (2002).
- [207] Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer* **6**, 813–823 (2006).
- [208] Seashore-Ludlow, B. *et al.* Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discovery* **5**, 1210–1223 (2015).
- [209] Ghandi, M. *et al.* Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 (2019).
- [210] Subramanian, A. *et al.* A next generation connectivity map: L1000 platform and the first 1, 000, 000 profiles. *Cell* **171**, 1437–1452.e17 (2017).
- [211] Heinemann, V. *et al.* FOLFIRI plus cetuximab versus FOLFIRI plus bevacizumab as first-line treatment for patients with metastatic colorectal cancer (FIRE-3): a randomised, open-label, phase 3 trial. *The Lancet Oncology* **15**, 1065–1075 (2014).
- [212] Zhong, W.-Z. *et al.* Gefitinib versus vinorelbine plus cisplatin as adjuvant treatment for stage II–IIIA (n1–n2) EGFR-mutant NSCLC (ADJUVANT/CTONG1104): a randomised, open-label, phase 3 study. *The Lancet Oncology* **19**, 139–148 (2018).
- [213] Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40**, D1100–D1107 (2011).
- [214] Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research* **49**, D1388–D1395 (2020).
- [215] Wishart, D. S. *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research* **36**, D901–D906 (2007).
- [216] Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J. & Bork, P. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Research* **36**, D684–D688 (2007).
- [217] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45**, D353–D361 (2016).

REFERENCES

- [218] The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Research* **47**, D330–D338 (2018).
- [219] Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods* **13**, 966–967 (2016).
- [220] Tate, J. G. *et al.* COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research* **47**, D941–D947 (2018).
- [221] Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nature Genetics* **45**, 580–585 (2013).
- [222] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- [223] Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
- [224] Przybyla, L. & Gilbert, L. A. A new era in functional genomics screens. *Nature Reviews Genetics* **23**, 89–103 (2021).
- [225] Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
- [226] Stinson, S. F. *et al.* Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. *Anticancer research* **12**, 1035–53 (1992).
- [227] Dove, A. The art of culture: Developing cell lines. *Science* **346**, 1013–1015 (2014).
- [228] Najgebauer, H. *et al.* CELLector: Genomics-guided selection of cancer in vitro models. *Cell Systems* **10**, 424–432.e6 (2020).
- [229] Solit, D. B. *et al.* BRAF mutation predicts sensitivity to MEK inhibition. *Nature* **439**, 358–362 (2005).
- [230] Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528**, 84–87 (2015).
- [231] Gonçalves, E. *et al.* Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell* **40**, 835–849.e8 (2022).
- [232] Li, H. *et al.* The landscape of cancer cell line metabolism. *Nature Medicine* **25**, 850–860 (2019).
- [233] Nusinow, D. P. *et al.* Quantitative proteomics of the cancer cell line encyclopedia. *Cell* **180**, 387–402.e16 (2020).
- [234] Feizi, N. *et al.* PharmacDB 2.0: improving scalability and transparency of in vitro pharmacogenomics analysis. *Nucleic Acids Research* **50**, D1348–D1357 (2021).
- [235] Macarron, R. *et al.* Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery* **10**, 188–195 (2011).
- [236] Riss, T. L., Moravec, R. A., Niles, A. L. & *et al.* Cell viability assays. In *Assay Guidance Manual [Internet]* (Eli Lilly & Company and the National Center for Advancing Translational Sciences, Bethesda (MD), 2013), updated 2016 jul 1 edn.
- [237] Vis, D. J. *et al.* Multilevel models improve precision and speed of IC50 estimates. *Pharmacogenomics* **17**, 691–700 (2016).
- [238] Hafner, M., Niepel, M., Chung, M. & Sorger, P. K. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nature Methods* **13**, 521–527 (2016).

- [239] Wang, D. *et al.* A statistical framework for assessing pharmacological responses and biomarkers using uncertainty estimates. *eLife* **9** (2020).
- [240] Tansey, W. *et al.* Dose–response modeling in high-throughput cancer drug screenings: an end-to-end approach. *Biostatistics* **23**, 643–665 (2021).
- [241] Firoozbakht, F., Yousefi, B. & Schwikowski, B. An overview of machine learning methods for monotherapy drug response prediction. *Briefings in Bioinformatics* **23** (2021).
- [242] Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR–cas9 essentiality screens in cancer cells. *Nature Genetics* **49**, 1779–1784 (2017).
- [243] Behan, F. M. *et al.* Prioritization of cancer therapeutic targets using CRISPR–cas9 screens. *Nature* **568**, 511–516 (2019).
- [244] Gonçalves, E. *et al.* Drug mechanism-of-action discovery through the integration of pharmacological and CRISPR screens. *Molecular Systems Biology* **16** (2020).
- [245] Rauscher, B., Heigwer, F., Breinig, M., Winter, J. & Boutros, M. GenomeCRISPR - a database for high-throughput CRISPR/cas9 screens. *Nucleic Acids Research* **45**, D679–D686 (2016).
- [246] Cui, Y. *et al.* CRISP-view: a database of functional genetic screens spanning multiple phenotypes. *Nucleic Acids Research* **49**, D848–D854 (2020).
- [247] Choi, A. *et al.* iCSDB: an integrated database of CRISPR screens. *Nucleic Acids Research* **49**, D956–D961 (2020).
- [248] Keenan, A. B. *et al.* The library of integrated network-based cellular signatures NIH program: System-level cataloging of human cells response to perturbations. *Cell Systems* **6**, 13–24 (2018).
- [249] Lamb, J. The connectivity map: a new tool for biomedical research. *Nature Reviews Cancer* **7**, 54–60 (2007).
- [250] Musa, A. *et al.* A review of connectivity map and computational approaches in pharmacogenomics. *Briefings in Bioinformatics* bbw112 (2017).
- [251] Ramsey, J. M. *et al.* Entinostat prevents leukemia maintenance in a collaborating oncogene-dependent model of cytogenetically normal acute myeloid leukemia. *Stem Cells* **31**, 1434–1445 (2013).
- [252] Moiso, E. Manual curation of TCGA treatment data and identification of potential markers of therapy response. *medRxiv* (2021).
- [253] Visvanathan, K. *et al.* Untapped potential of observational research to inform clinical decision making: American society of clinical oncology research statement. *Journal of Clinical Oncology* **35**, 1845–1854 (2017).
- [254] Eisenhauer, E. *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer* **45**, 228–247 (2009).
- [255] Turner, R. M., Park, B. K. & Pirmohamed, M. Parsing interindividual drug variability: an emerging role for systems pharmacology. *WIREs Systems Biology and Medicine* **7**, 221–241 (2015).
- [256] Menden, M. P. *et al.* Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS ONE* **8**, e61318 (2013).
- [257] Landrum, G. Rdkit: Open-source cheminformatics. <http://www.rdkit.org>.

REFERENCES

- [258] Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *Proceedings of the 35th International Conference on Machine Learning* **80**, 2323–2332 (2018).
- [259] Liu, A. *et al.* From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *npj Systems Biology and Applications* **5** (2019).
- [260] Bradley, G. & Barrett, S. J. CausalR: extracting mechanistic sense from genome scale data. *Bioinformatics* **33**, 3670–3672 (2017).
- [261] Özgün Babur *et al.* Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biology* **16** (2015).
- [262] Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- [263] Barretina, J. *et al.* The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
- [264] Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Medicine* **12** (2020).
- [265] Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
- [266] Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics* **19**, 491–504 (2018).
- [267] Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).
- [268] Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821–824 (2012).
- [269] Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1** (2021).
- [270] Menden, M. P. *et al.* The germline genetic component of drug sensitivity in cancer cell lines. *Nature Communications* **9** (2018).
- [271] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909 (2006).
- [272] Vösa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics* **53**, 1300–1310 (2021).
- [273] Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47–e47 (2015).
- [274] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15** (2014).
- [275] Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics* **12**, 529–541 (2011).
- [276] Aryee, M. J. *et al.* Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
- [277] Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology* **13**, R87 (2012).
- [278] Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology* **41**, 200–209 (2012).

-
- [279] Peters, T. J. *et al.* De novo identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin* **8** (2015).
- [280] Butcher, L. M. & Beck, S. Probe lasso: A novel method to rope in differentially methylated regions with 450k DNA methylation data. *Methods* **72**, 21–28 (2015).
- [281] Pedersen, B. S., Schwartz, D. A., Yang, I. V. & Kechris, K. J. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated *p*-values. *Bioinformatics* **28**, 2986–2988 (2012).
- [282] Mallik, S. *et al.* An evaluation of supervised methods for identifying differentially methylated regions in illumina methylation arrays. *Briefings in Bioinformatics* **20**, 2224–2235 (2018).
- [283] Yao, L., Shen, H., Laird, P. W., Farnham, P. J. & Berman, B. P. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biology* **16** (2015).
- [284] Silva, T. C. *et al.* ELMER v.2: an r/bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. *Bioinformatics* **35**, 1974–1977 (2018).
- [285] Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology* **32**, 1202–1212 (2014).
- [286] Chen, J. & Zhang, L. A survey and systematic assessment of computational methods for drug response prediction. *Briefings in Bioinformatics* **22**, 232–246 (2020).
- [287] Guinney, J. *et al.* Modeling RAS phenotype in colorectal cancer uncovers novel molecular traits of RAS dependency and improves prediction of response to targeted agents in patients. *Clinical Cancer Research* **20**, 265–272 (2014).
- [288] Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C. & Ester, M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* **35**, i501–i509 (2019).
- [289] Aben, N., Vis, D. J., Michaut, M. & Wessels, L. F. TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics* **32**, i413–i420 (2016).
- [290] Partin, A. *et al.* Deep learning methods for drug response prediction in cancer: Predominant and emerging trends. *Frontiers in Medicine* **10** (2023).
- [291] Baptista, D., Ferreira, P. G. & Rocha, M. Deep learning for drug response prediction in cancer. *Briefings in Bioinformatics* **22**, 360–379 (2020).
- [292] Azuaje, F. Computational models for predicting drug responses in cancer research. *Briefings in Bioinformatics* bbw065 (2016).
- [293] Geeleher, P., Cox, N. J. & Huang, R. S. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biology* **15** (2014).
- [294] Geeleher, P. *et al.* Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. *Genome Research* **27**, 1743–1751 (2017).
- [295] Ma, J. *et al.* Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nature Cancer* **2**, 233–244 (2021).
- [296] Ondra, T. *et al.* Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics* **26**, 99–119 (2015).
- [297] Yeang, C.-H., McCormick, F. & Levine, A. Combinatorial patterns of somatic gene mutations in cancer. *The FASEB Journal* **22**, 2605–2622 (2008).

REFERENCES

- [298] Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research* **22**, 398–406 (2011).
- [299] Leiserson, M. D., Wu, H.-T., Vandin, F. & Raphael, B. J. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biology* **16** (2015).
- [300] Imai, K. & Ratkovic, M. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* **7** (2013).
- [301] Su, X., Zhou, T., Yan, X., Fan, J. & Yang, S. Interaction trees with censored survival data. *The International Journal of Biostatistics* **4** (2008).
- [302] Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M. & Li, B. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* **10**, 141–158 (2009).
- [303] Loh, W.-Y., He, X. & Man, M. A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine* **34**, 1818–1833 (2015).
- [304] Zeileis, A., Hothorn, T. & Hornik, K. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics* **17**, 492–514 (2008).
- [305] Lipkovich, I., Dmitrienko, A., Denne, J. & Enas, G. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine* **30**, 2601–2621 (2011).
- [306] Qian, M. & Murphy, S. A. Performance guarantees for individualized treatment rules. *The Annals of Statistics* **39** (2011).
- [307] Zhao, Y., Zeng, D., Rush, A. J. & Kosorok, M. R. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107**, 1106–1118 (2012).
- [308] Schrod, S. *et al.* BITES: balanced individual treatment effect for survival data. *Bioinformatics* **38**, i60–i67 (2022).
- [309] Sharma, A. & Kiciman, E. Dowhy: An end-to-end library for causal inference. *arXiv* (2020).
- [310] Battocchi, K. *et al.* EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation (2019). Version 0.x.
- [311] Jemielita, T. O. & Mehrotra, D. V. Prism: Patient response identifiers for stratified medicine. *arXiv* (2019).
- [312] Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2018).
- [313] European Medicines Agency. EMA regulatory science to 2025: strategic reflection (2020).
- [314] European Medicines Agency. Guideline on multiplicity issues in clinical trials (2017).
- [315] Pocock, S. J. & Stone, G. W. The primary outcome fails — what next? *New England Journal of Medicine* **375**, 861–870 (2016).
- [316] Jonker, D. J. *et al.* Cetuximab for the treatment of colorectal cancer. *New England Journal of Medicine* **357**, 2040–2048 (2007).
- [317] Karapetis, C. S. *et al.* K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *New England Journal of Medicine* **359**, 1757–1765 (2008).
- [318] U.S. Food and Drug Administration. Label change of cetuximab. https://www.accessdata.fda.gov/drugsatfda_docs/label/2009/125084s167lbl.pdf (2009).

-
- [319] Holch, J. W., Ricard, I., Stintzing, S., Modest, D. P. & Heinemann, V. The relevance of primary tumour location in patients with metastatic colorectal cancer: A meta-analysis of first-line clinical trials. *European Journal of Cancer* **70**, 87–98 (2017).
 - [320] Cervantes, A. *et al.* Metastatic colorectal cancer: Esmo clinical practice guideline for diagnosis, treatment and follow-up. *Annals of Oncology* **34**, 10–32 (2023).
 - [321] Hughes, J., Rees, S., Kalindjian, S. & Philpott, K. Principles of early drug discovery. *British Journal of Pharmacology* **162**, 1239–1249 (2011).
 - [322] Heidorn, S. J. *et al.* Kinase-dead BRAF and oncogenic RAS cooperate to drive tumor progression through CRAF. *Cell* **140**, 209–221 (2010).
 - [323] Amaral, T. *et al.* Mapk pathway in melanoma part ii—secondary and adaptive resistance mechanisms to braf inhibition. *European Journal of Cancer* **73**, 93–101 (2017).
 - [324] Prahallad, A. *et al.* Unresponsiveness of colon cancer to BRAF(v600e) inhibition through feedback activation of EGFR. *Nature* **483**, 100–103 (2012).
 - [325] Haibe-Kains, B. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389–393 (2013).
 - [326] Geeleher, P., Gamazon, E. R., Seoighe, C., Cox, N. J. & Huang, R. S. Consistency in large pharmacogenomic studies. *Nature* **540**, E1–E2 (2016).
 - [327] Smirnov, P. *et al.* Meta-analysis of preclinical pharmacogenomic studies to discover robust and translatable biomarkers of drug response. *bioRxiv* (2022).
 - [328] Fouse, S. D., Nagarajan, R. P. & Costello, J. F. Genome-scale DNA methylation analysis. *Epigenomics* **2**, 105–117 (2010).
 - [329] Yadowlowsky, S., Pellegrini, F., Lionetto, F., Braune, S. & Tian, L. Estimation and validation of ratio-based conditional average treatment effects using observational data. *Journal of the American Statistical Association* **116**, 335–352 (2020).
 - [330] Brookes, S. T. *et al.* Subgroup analyses in randomized trials: risks of subgroup-specific analyses;. *Journal of Clinical Epidemiology* **57**, 229–236 (2004).
 - [331] Sveen, A., Kopetz, S. & Lothe, R. A. Biomarker-guided therapy for colorectal cancer: strength in complexity. *Nature Reviews Clinical Oncology* **17**, 11–32 (2019).
 - [332] Adam, G. *et al.* Machine learning approaches to drug response prediction: challenges and recent progress. *npj Precision Oncology* **4** (2020).
 - [333] Soleymani, A., Raj, A., Bauer, S., Schölkopf, B. & Besserve, M. Causal feature selection via orthogonal search. *arXiv* (2020).
 - [334] Mandrekas, S. J. & Sargent, D. J. Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. *Journal of Clinical Oncology* **27**, 4027–4034 (2009).
 - [335] Butler, M. *et al.* MGMT status as a clinical biomarker in glioblastoma. *Trends in Cancer* **6**, 380–391 (2020).
 - [336] Arozarena, I. & Wellbrock, C. Phenotype plasticity as enabler of melanoma progression and therapy resistance. *Nature Reviews Cancer* **19**, 377–391 (2019).
 - [337] Yuan, S., Norgard, R. J. & Stanger, B. Z. Cellular plasticity in cancer. *Cancer Discovery* **9**, 837–851 (2019).

REFERENCES

- [338] Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357** (2017).
- [339] Tagliazucchi, G. M., Wiecek, A. J., Withnell, E. & Secrier, M. Genomic and microenvironmental heterogeneity shaping epithelial-to-mesenchymal trajectories in cancer. *Nature Communications* **14** (2023).
- [340] Nazarian, R. *et al.* Melanomas acquire resistance to b-RAF(v600e) inhibition by RTK or n-RAS upregulation. *Nature* **468**, 973–977 (2010).
- [341] Emert, B. L. *et al.* Variability within rare cell states enables multiple paths toward drug resistance. *Nature Biotechnology* **39**, 865–876 (2021).
- [342] Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).
- [343] Holohan, C., Schaeybroeck, S. V., Longley, D. B. & Johnston, P. G. Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer* **13**, 714–726 (2013).
- [344] Misale, S. *et al.* Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* **486**, 532–536 (2012).
- [345] Tetsu, O., Hangauer, M. J., Phuchareon, J., Eisele, D. W. & McCormick, F. Drug resistance to EGFR inhibitors in lung cancer. *Chemotherapy* **61**, 223–235 (2016).
- [346] Engelman, J. A. *et al.* Met amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. *Science* **316**, 1039–1043 (2007).
- [347] Uramoto, H., Shimokawa, H., Hanagiri, T., Kuwano, M. & Ono, M. Expression of selected gene for acquired drug resistance to EGFR-TKI in lung adenocarcinoma. *Lung Cancer* **73**, 361–365 (2011).
- [348] Mason, M. *et al.* A community challenge to predict clinical outcomes after immune checkpoint blockade in non-small cell lung cancer. *Journal of Translational Medicine* **22** (2024).
- [349] Koyama, S. *et al.* Adaptive resistance to therapeutic PD-1 blockade is associated with upregulation of alternative immune checkpoints. *Nature Communications* **7** (2016).
- [350] Marine, J.-C., Dawson, S.-J. & Dawson, M. A. Non-genetic mechanisms of therapeutic resistance in cancer. *Nature Reviews Cancer* **20**, 743–756 (2020).
- [351] Sartorelli, A. C. Some approaches to the therapeutic exploitation of metabolic sites of vulnerability of neoplastic cells. *Cancer Research* **29**, 2292–2299 (1969).
- [352] Kwak, E. L., Clark, J. W. & Chabner, B. Targeted agents: The rules of combination. *Clinical Cancer Research* **13**, 5232–5237 (2007).
- [353] DeVita, V. T. & Schein, P. S. The use of drugs in combination for the treatment of cancer. *New England Journal of Medicine* **288**, 998–1006 (1973).
- [354] Kummar, S. *et al.* Utilizing targeted cancer therapeutic agents in combination: novel approaches and urgent requirements. *Nature Reviews Drug Discovery* **9**, 843–856 (2010).
- [355] Larkin, J. *et al.* Combined vemurafenib and cobimetinib in BRAF-mutated melanoma. *New England Journal of Medicine* **371**, 1867–1876 (2014).
- [356] Kopetz, S. *et al.* Randomized trial of irinotecan and cetuximab with or without vemurafenib in BRAF-mutant metastatic colorectal cancer (SWOG s1406). *Journal of Clinical Oncology* **39**, 285–294 (2021).

- [357] Palmer, A. C. & Sorger, P. K. Combination cancer therapy can confer benefit via patient-to-patient variability without drug additivity or synergy. *Cell* **171**, 1678–1691.e13 (2017).
- [358] Jaaks, P. *et al.* Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature* **603**, 166–173 (2022).
- [359] Chiu, Y.-C. *et al.* Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Medical Genomics* **12** (2019).
- [360] Jia, P. *et al.* Deep generative neural network for accurate drug response imputation. *Nature Communications* **12** (2021).
- [361] He, D., Liu, Q., Wu, Y. & Xie, L. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. *Nature Machine Intelligence* **4**, 879–892 (2022).
- [362] Sharifi-Noghabi, H., Harjandi, P. A., Zolotareva, O., Collins, C. C. & Ester, M. Out-of-distribution generalization from labelled and unlabelled gene expression data for drug response prediction. *Nature Machine Intelligence* **3**, 962–972 (2021).
- [363] Rampášek, L., Hidru, D., Smirnov, P., Haibe-Kains, B. & Goldenberg, A. Dr.VAE: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics* **35**, 3743–3751 (2019).
- [364] Lotfollahi, M. *et al.* Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology* **19** (2023).
- [365] Manica, M. *et al.* Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular Pharmaceutics* **16**, 4797–4806 (2019).
- [366] Hetzel, L. *et al.* Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Advances in Neural Information Processing Systems* (2022).
- [367] Born, J. *et al.* PaccMannRL: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *iScience* **24**, 102269 (2021).
- [368] Yu, M. K. *et al.* Visible machine learning for biomedicine. *Cell* **173**, 1562–1565 (2018).
- [369] Elmarakeby, H. A. *et al.* Biologically informed deep neural network for prostate cancer discovery. *Nature* **598**, 348–352 (2021).
- [370] Li, Y., Hostallero, D. E. & Emad, A. Interpretable deep learning architectures for improving drug response prediction performance: myth or reality? *Bioinformatics* **39** (2023).
- [371] Lotfollahi, M. *et al.* Biologically informed deep learning to query gene programs in single-cell atlases. *Nature Cell Biology* (2023).
- [372] AlQuraishi, M. & Sorger, P. K. Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. *Nature Methods* **18**, 1169–1180 (2021).
- [373] Lobentanzer, S. *et al.* Democratizing knowledge representation with BioCypher. *Nature Biotechnology* (2023).
- [374] Ruiz, C., Zitnik, M. & Leskovec, J. Identification of disease treatment mechanisms through the multiscale interactome. *Nature Communications* **12** (2021).
- [375] Li, M. M., Huang, K. & Zitnik, M. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering* **6**, 1353–1369 (2022).

- [376] Liu, Q., Hu, Z., Jiang, R. & Zhou, M. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics* **36**, i911–i918 (2020).
- [377] Nguyen, T., Nguyen, G. T. T., Nguyen, T. & Le, D.-H. Graph convolutional networks for drug response prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **19**, 146–154 (2022).
- [378] Zagidullin, B. *et al.* DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Research* **47**, W43–W51 (2019).
- [379] Menden, M. P. *et al.* Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature Communications* **10** (2019).
- [380] Ye, C. *et al.* DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nature Communications* **9** (2018).
- [381] Dixit, A. *et al.* Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866.e17 (2016).
- [382] Lotfollahi, M. *et al.* Biologically informed deep learning to query gene programs in single-cell atlases. *Nature Cell Biology* (2023).
- [383] Roohani, Y., Huang, K. & Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology* (2023).
- [384] Chen, J. *et al.* Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nature Communications* **13** (2022).
- [385] Fröhlich, F. *et al.* Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. *Cell Systems* **7**, 567–579.e6 (2018).
- [386] Yuan, B. *et al.* CellBox: Interpretable machine learning for perturbation biology with application to the design of cancer combination therapy. *Cell Systems* **12**, 128–140.e4 (2021).
- [387] Raimúndez, E. *et al.* Model-based analysis of response and resistance factors of cetuximab treatment in gastric cancer cell lines. *PLOS Computational Biology* **16**, e1007147 (2020).
- [388] Fröhlich, F., Gerosa, L., Muhlich, J. & Sorger, P. K. Mechanistic model of *MAPK* signaling reveals how allostery and rewiring contribute to drug resistance. *Molecular Systems Biology* **19** (2023).
- [389] Arulraj, T., Wang, H., Emens, L. A., Santa-Maria, C. A. & Popel, A. S. A transcriptome-informed qsp model of metastatic triple-negative breast cancer identifies predictive biomarkers for pd-1 inhibition. *Science Advances* **9** (2023).
- [390] Hahn, P. R., Murray, J. S. & Carvalho, C. M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis* **15** (2020).
- [391] Scott, J. G. & Berger, J. O. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* **38** (2010).
- [392] Yoon, J., Jordon, J. & van der Schaar, M. GANITE: Estimation of individualized treatment effects using generative adversarial nets. *International Conference on Learning Representations* (2018).
- [393] Curth, A., Lee, C. & van der Schaar, M. SurvITE: Learning heterogeneous treatment effects from time-to-event data. *Advances in Neural Information Processing Systems* (2021).
- [394] Schwab, P., Linhardt, L., Bauer, S., Buhmann, J. M. & Karlen, W. Learning counterfactual representations for estimating individual dose-response curves. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 5612–5619 (2020).

-
- [395] Liu, R., Wei, L. & Zhang, P. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. *Nature Machine Intelligence* **3**, 68–75 (2021).
- [396] Bica, I., Alaa, A. M., Jordon, J. & van der Schaar, M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *International Conference on Learning Representations* (2020).
- [397] Qian, Z., Zhang, Y., Bica, I., Wood, A. & van der Schaar, M. Synctwin: Treatment effect estimation with longitudinal outcomes. *Advances in Neural Information Processing Systems* **34**, 3178–3190 (2021).
- [398] Melnychuk, V., Frauen, D. & Feuerriegel, S. Causal transformer for estimating counterfactual outcomes. *Proceedings of the 39th International Conference on Machine Learning* **162**, 15293–15329 (2022).
- [399] Crabbé, J., Curth, A., Bica, I. & van der Schaar, M. Benchmarking heterogeneous treatment effect models through the lens of interpretability. *Advances in Neural Information Processing Systems* **35**, 12295–12309 (2022).
- [400] Athey, S. & Wager, S. Policy learning with observational data. *arXiv* (2017).
- [401] Pace, A., Chan, A. & van der Schaar, M. POETREE: Interpretable policy learning with adaptive decision trees. *International Conference on Learning Representations* (2022).
- [402] Reisle, C. *et al.* A platform for oncogenomic reporting and interpretation. *Nature Communications* **13** (2022).
- [403] Ivanenkov, Y. A. *et al.* Chemistry42: An AI-driven platform for molecular design and optimization. *Journal of Chemical Information and Modeling* **63**, 695–701 (2023).
- [404] Wagner, A. H. *et al.* A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. *Nature Genetics* **52**, 448–457 (2020).

REFERENCES

Appendix A

Abbreviations

The used abbreviations for subjects of this work are listed and described here. These include the introduced acronyms and naming conventions for genes and proteins in the main text.

A.1 List of acronyms

Acronym	Description
(e)QTL	(Expression) quantitative trait loci
(m)RNA	(Messenger) ribonucleic acid
5mC	5-methylcytosine DNA methylation
ACF	Auto-correlation function
ALL	Acute lymphoblastic leukaemia
AML	Acute myeloid leukaemia
AMP	Copy number amplification
ANOVA	Analysis of variance
ATE	Average treatment effect
ATP	Adenosine triphosphate
AUC	Area under the curve
BART	Bayesian additive regression tree
BH	Benjamini-Hochberg
BRCA	Breast cancer
CATE	Conditional average treatment effect
CCLC	Cancer Cell Line Encyclopedia
ChEMBL	Chemical database by the European Molecular Biology Laboratory (EMBL)
chr20q	Chromosome arm 20q
CI	Confidence interval
CIMP	CpG island methylator phenotype
CMAP	Connectivity map
CMS	Consensus molecular subtypes
COREAD	Colorectal adenocarcinoma
COSMIC	Catalogue of Somatic Mutations in Cancer
CpG	Region of DNA with cytosine followed by guanine along the 5' to 3' direction
CRF	Causal random forest
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTC	Circulating tumour cells
CTRP	Cancer Therapeutic Response Portal
CV	Cross-validation
d.o.f.	Degree of freedom
dDMR	Drug differentially methylated regions
DEL	Copy number deletion
DFS	Disease-free survival
DGE	Differential gene expression
DHS	DNAase I hypersensitive site
DML	Double machine learning
DMP	Differentially methylated probe
DMR	Differentially methylated region
DREAM	Dialogue for Reverse Engineering Assessment and Methods
EHR	Electronic health records
ELMER	Enhancer linking by methylation/expression relationships
EMA	European Medicines Agency
EMT	Epithelial-mesenchymal transition
ENCODE	Encyclopedia of DNA Elements
ESMO	European Society for Medical Oncology

EWAS	Epigenome-wide association study
FDA	U.S. Food and Drug Administration
FDR	False discovery rate
FOLFIRI	5-fluorouracil, folinic acid and irinotecan treatment backbone
FOLFOX	5-fluorouracil, folinic acid and oxaliplatin treatment backbone
FOLFOXIRI	5-fluorouracil, oxaliplatin, folinic acid and irinotecan treatment backbone
FWER	Family-wise error rate
GDC	Genomic Data Commons
GDSC	Genomics of Drug Sensitivity in Cancer
GO	Gene Ontology
GSVA	Gene set variation analysis
GTEX	Genotype-Tissue Expression project
GWAS	Genome-wide association study
HNSC	Head- and neck cancer
HR	Hazard ratio
HTS	(Drug) high-throughput screen
IC ₅₀	Half maximal inhibitory concentration
ICGC	International Cancer Genome Consortium
KEGG	Kyoto Encyclopedia of Genes and Genomes
LINCS	Library of Integrated Network-Based Cellular Signatures
LSTM	Long-short term memory
LUAD	Lung adenocarcinoma
MAK	Mak <i>et al.</i> EMT score
mCRC	Metastatic colorectal cancer
MET	Mesenchymal-epithelial transition
MMRd	Mismatch repair deficiency
MOA	Mechanism of action
MOB	Model-based partitioning
MSI	Microssatellite instability
NCBI	National Center for Biotechnology Information
NCG	Network of Cancer Genes
NCI	National Cancer Institute
NGS	Next-generation sequencing
NIH	National Institutes of Health
NSCLC	Non-small cell lung carcinoma
OncoBird	Oncology Biomarker Discovery
OR	Odds ratio
ORR	Objective response rate
OS	Overall survival
OWE	Outcome-weighted estimation
PAM50	Prediction Analysis of Microarray with 50 genes
PCAWG	Pan-Cancer Analysis of Whole Genomes
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PFS	Progression-free survival
POL	Policy learning
PubChem	Chemical database by the National Center for Biotechnology Information (NCBI)
RCT	Randomised controlled clinical trial
RECIST	Response Evaluation Criteria in Solid Tumours
RNAi	RNA interference

RNN	Recurrent neural network
RRBS	Reduced representation bisulfite sequencing
SCLC	Small cell lung carcinoma
SKCM	Skin cutaneous melanoma
SMILES	Simplified Molecular Input Line Entry System
SNP	Single-nucleotide polymorphism
SNV/SV	Single-nucleotide variant/variation
STAD	Stomach adenocarcinoma
STITCH	Search Tool for Interactions of CHemicals
TAN	Tan <i>et al.</i> EMT score
TARGET	Therapeutically Applicable Research to Generate Effective Treatments program
TCGA	The Cancer Genome Atlas
TF	Transcription factor
tgDMR	Tumour-generalisable drug differentially methylated regions
TKI	Tyrosine kinase inhibitor
TMB	Tumour mutational burden
TS	Targeted sequencing
TSS200	200 bases upstream of the transcription start site
TW	Tagliazucchi and Wiecek <i>et al.</i> EMT score
UMAP	Uniform manifold approximation and projection
VAE	Variational autoencoder
VT	Virtual twins
WES	Whole-exome sequencing
WGBS	Whole-genome bisulfite sequencing
WGS	Whole-genome sequencing

A.2 List of proteins

Symbol	Description
ABL1	ABL1 tyrosine kinase, name from <i>ABL</i> proto-oncogene 1 extracted from the Abelson murine leukaemia virus
AKT	Protein kinase B (PKB)
ALK	Anaplastic lymphoma kinase
BCR	Breakpoint cluster region protein
BCR-ABL1	Fusion protein from gene fusion of <i>BCR</i> and <i>ABL1</i> due to Philadelphia translocation
BIRC5	Baculoviral IAP repeat containing 5
BRAF	Serine/threonine-protein kinase B-Raf
CASP3	Caspase 3
CDC25A	Cell division cycle 25A
CDK2/4	Cyclin dependent kinase 2/4
CHK1/2	Checkpoint kinase 1/2
CRYAB	Crystallin alpha B; alias HSPB5
DNMT	DNA-methyltransferase protein family
DNMT3A	DNA (cytosine-5)-methyltransferase 3A
EGFR	Epidermal growth factor receptor
ERBB2	Receptor tyrosine-protein kinase erb-B2, name from erythroblastic leukaemia viral oncogene, alias human epidermal growth factor receptor 2 (HER2)

ERBB3	Receptor tyrosine-protein kinase erb-B3
ERK	Extracellular signal-regulated kinase
ESR1	Estrogen receptor 1
GSK3 β	Glycogen synthase kinase 3 beta
HDAC	Histone deacetylase protein family
HOX	Homeobox protein family
HSP	Heat shock protein family
HSP90	Heat shock protein 90
IAP	Inhibitor of apoptosis protein family
IDH1	Isocitrate dehydrogenase 1
IRF4	Interferon regulatory Factor 4
KRAS	Ras GTPase isoform K-Ras
MAPK	Mitogen-activated protein family of kinases
MEK	Mitogen-activated protein kinase kinase
MGMT	O-6-methylguanine-DNA methyltransferase
MITF	Melanocyte inducing transcription factor
MYC	MYC proto-oncogene protein
NAE	NEDD8 activating enzyme
NEDD8	Neural precursor cell expressed developmentally down-regulated 8
NEK9	NIMA (never in mitosis gene A)- related kinase 9
P16	Cyclin-dependent kinase inhibitor 2A
P53	Tumour protein p53
PARP	Poly (ADP-ribose) polymerase
PD-1	Programmed cell death protein 1
PD-L1	Programmed cell death 1 ligand 1
PI3K	Phosphoinositide 3-kinase
RAS	Small GTPase Ras
ROCK1	Rho associated coiled-coil containing protein kinase 1
SMAD	Contraction of C. elegans Sma and Drosophila Mad gene family
SOX2/9	SRY (sex determining region Y)-box transcription factor 2/9
TET2	Tet methylcytosine dioxygenase 2
TGF- β	Transforming growth factor beta
TOP1	DNA topoisomerase 1
VEGF	Vascular endothelial growth factor

A.3 List of genes

Symbol	Description
<i>APC</i>	Adenomatous polyposis coli protein tumour suppressor
<i>ARFRP1</i>	ADP (adenosine diphosphate) ribosylation factor related protein 1
<i>ARID1A</i>	AT-rich interaction domain 1A
<i>ATM</i>	ATM (ataxia telangiectasia mutated) serine/threonine kinase
<i>AURKA</i>	Aurora kinase A
<i>BCL2L1</i>	BCL2 (B-cell lymphoma 2) like 1
<i>BRAF</i>	v-Raf murine sarcoma viral oncogene homolog B
<i>BRCA1/2</i>	Breast cancer 1/2, DNA repair associated
<i>CCNB1</i>	Cyclin B1

<i>CDC25A</i>	Cell division cycle 25A
<i>CDH1/2</i>	Cadherin 1/2
<i>CDKN2A</i>	Cyclin-dependent kinase inhibitor 2A
<i>CRYAB</i>	Crystallin alpha B
<i>CUL3</i>	Cullin 3
<i>CYP2D6</i>	Cytochrome P450 family 2 subfamily D member 6
<i>DKK1</i>	Dickkopf WNT signalling pathway inhibitor 1
<i>EGFR</i>	Epidermal growth factor receptor
<i>EML4-ALK</i>	Fusion gene of echinoderm microtubule-associated protein-like 4 and anaplastic lymphoma kinase
<i>ERBB2</i>	v-erb-b2 avian erythroblastic leukaemia viral oncogene homolog 2
<i>ESR1</i>	Estrogen receptor 1
<i>FAM123B</i>	APC membrane recruitment protein 1
<i>FN1</i>	Fibronectin 1
<i>GNAS</i>	GNAS (G-protein subunit alpha S) complex locus
<i>HOXB2</i>	Homeobox B2
<i>IDH1</i>	Isocitrate dehydrogenase 1
<i>IRF4</i>	Interferon Regulatory Factor 4
<i>IRS2</i>	Insulin receptor substrate 2
<i>KRAS</i>	Kirsten rat sarcoma viral oncogene homolog
<i>LRP1B</i>	LDL (low density lipoprotein) receptor-related protein 1B
<i>MAP2K1</i>	Mitogen-activated protein kinase kinase 1 (ERK)
<i>MAPK1</i>	Mitogen-activated protein kinase 1 (MEK)
<i>MET</i>	MET proto-oncogene
<i>MGMT</i>	O-6-methylguanine-DNA methyltransferase
<i>MITF</i>	Melanocyte inducing transcription factor
<i>MYC</i>	MYC (myelocytomatosis) proto-oncogene
<i>NEK9</i>	NIMA (Never in mitosis gene A)- related kinase 9
<i>NF1</i>	Neurofibromin 1
<i>NKX2-1</i>	NK2 homeobox 1
<i>NQO1</i>	NAD(P)H quinone dehydrogenase 1
<i>NRAS</i>	Neuroblastoma RAS viral (v-Ras) oncogene
<i>OPLAH</i>	5-oxoprolinase, ATP-hydrolysing
<i>PIK3CA</i>	Phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha
<i>PITX2</i>	Paired like homeodomain 2
<i>PLK1</i>	Polo like kinase 1
<i>PTPRN2</i>	Protein tyrosine phosphatase receptor type N2
<i>PXN</i>	Paxillin
<i>RAS</i>	Rat sarcoma viral oncogene
<i>RB1</i>	RB (retinoblastoma) transcriptional corepressor 1
<i>RNF43</i>	Ring finger protein 43
<i>SHC1</i>	Src homology 2 domain-containing adaptor protein 1
<i>SKI</i>	Sloan-Kettering Institute proto-oncogene
<i>SLFN11</i>	Schlafen family member 11
<i>SMAD3/4</i>	SMAD family member 3/4
<i>SRC</i>	Src (Rous sarcoma virus) proto-oncogene
<i>TGFB1</i>	Transforming growth factor beta 1
<i>TOP1</i>	DNA topoisomerase 1
<i>TP53</i>	Tumour protein P53
<i>VIM</i>	Vimentin

ZNF217 Zinc finger protein 217

Appendix B

Supplementary material

This chapter contains supplementary information to the three research articles presented in Chapter 2. The supplementary information for the articles in Sections 2.1 and 2.3 were peer-reviewed and published open-access jointly with their respective article. The supplementary information to the presented preprint in Section 2.2 is publicly available with its preprint prior to peer review.

B.1 The pharmacoepigenomic landscape of cancer cell lines reveals the epigenetic component of drug sensitivity in cancer, *Communications Biology* (2023), supplementary information

This material is supplementary to the peer-reviewed and published open-access article in *Communications Biology* presented in Section 2.1 [1] and is reproduced with permission from Springer Nature. It is publicly available at <https://doi.org/10.1038/s42003-023-05198-y>.

Supplementary information for

The pharmacoepigenomic landscape of cancer cell lines reveals the epigenetic component of drug sensitivity

Ohnmacht AJ, Rajamani A, Avar G, Kutkaite G, Gonçalves E, Saur D, Menden MP

Corresponding Menden MP.

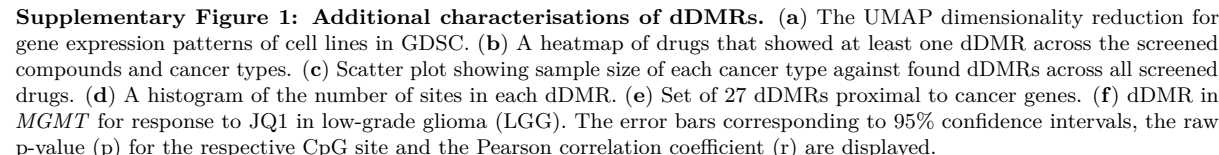
E-mail: michael.menden@helmholtz-munich.de

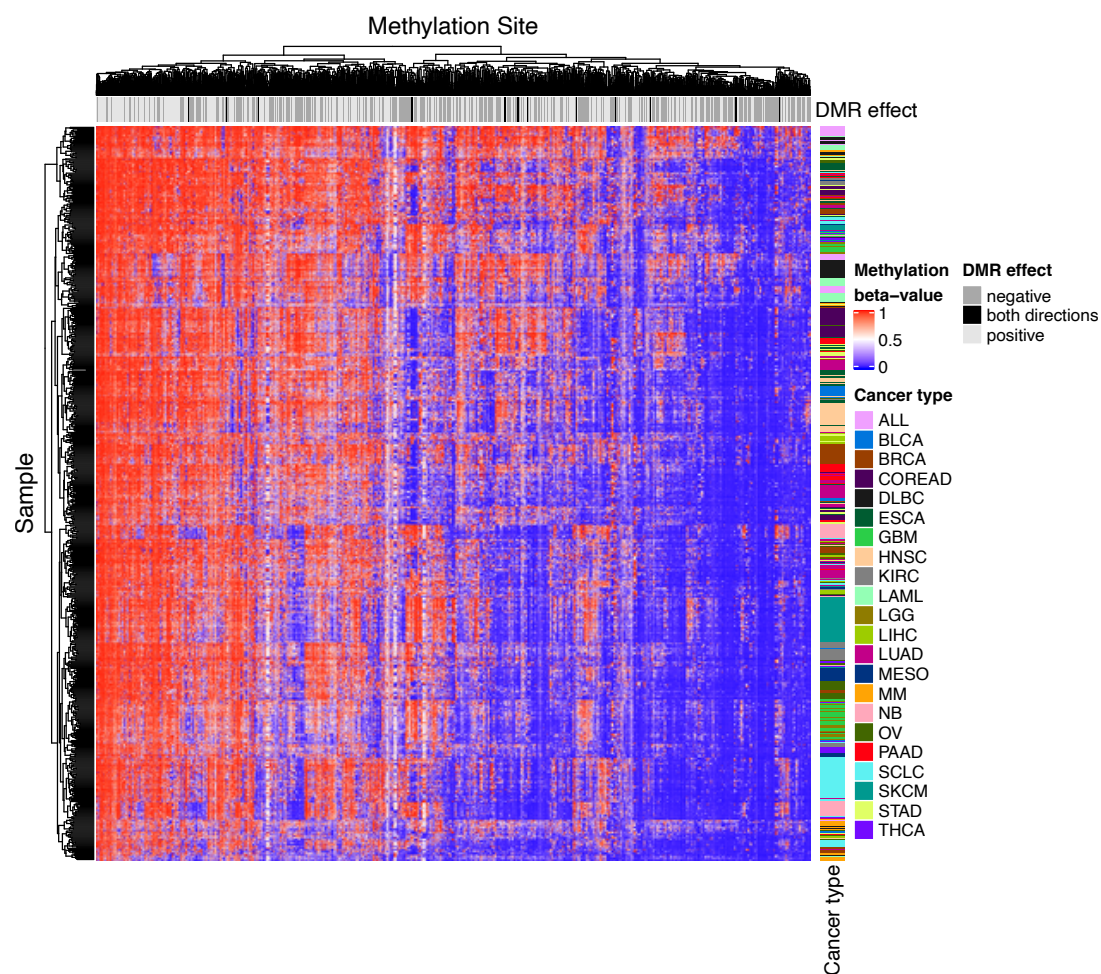
This PDF file includes:

Supplementary Figure 1 to 7

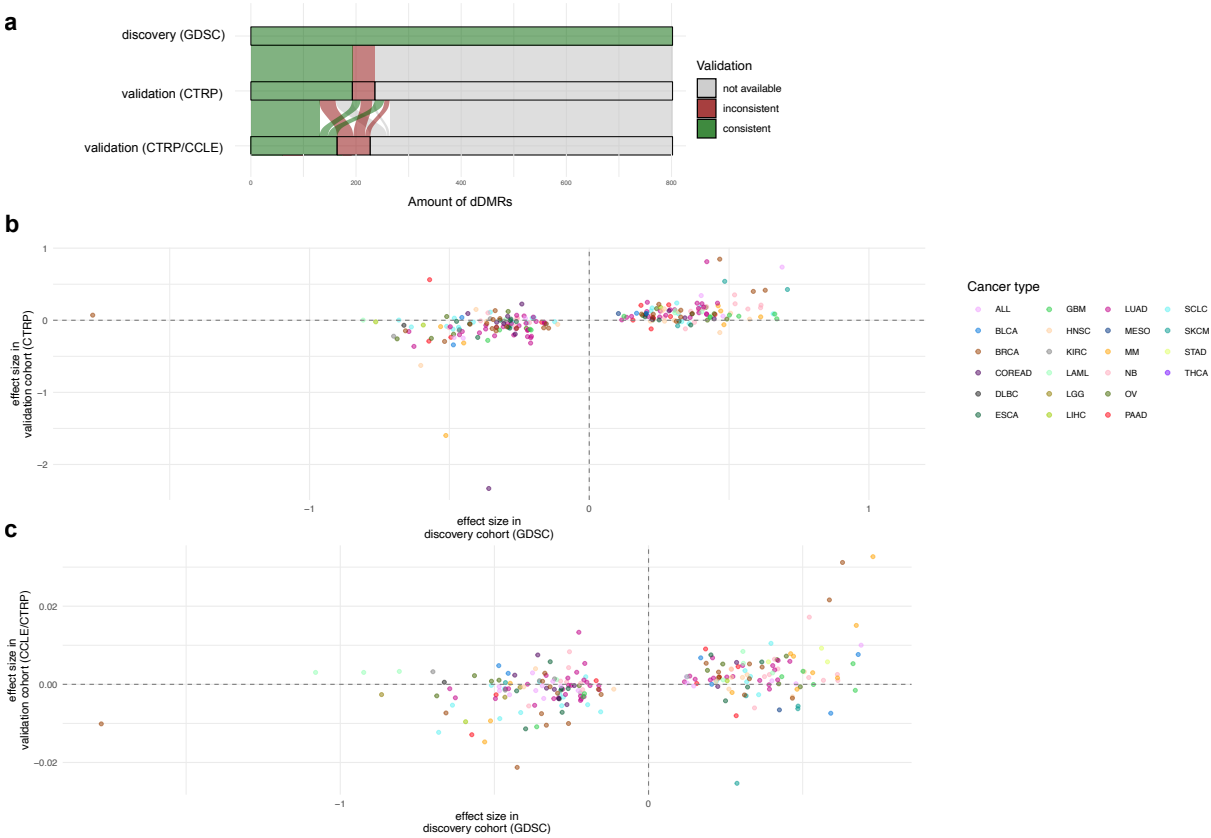
Other supplementary information for this manuscript include:

Supplementary Data 1 to 3

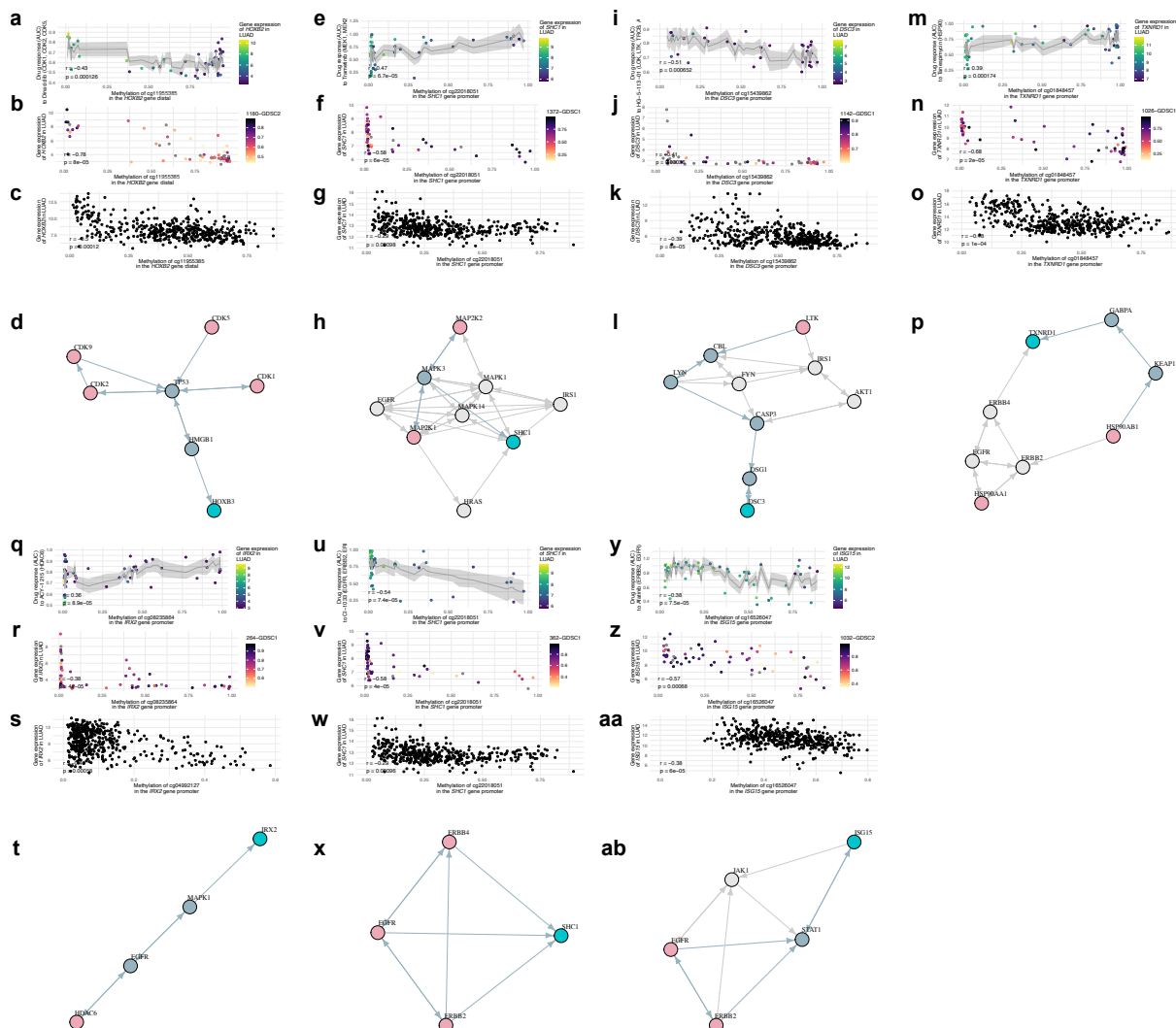




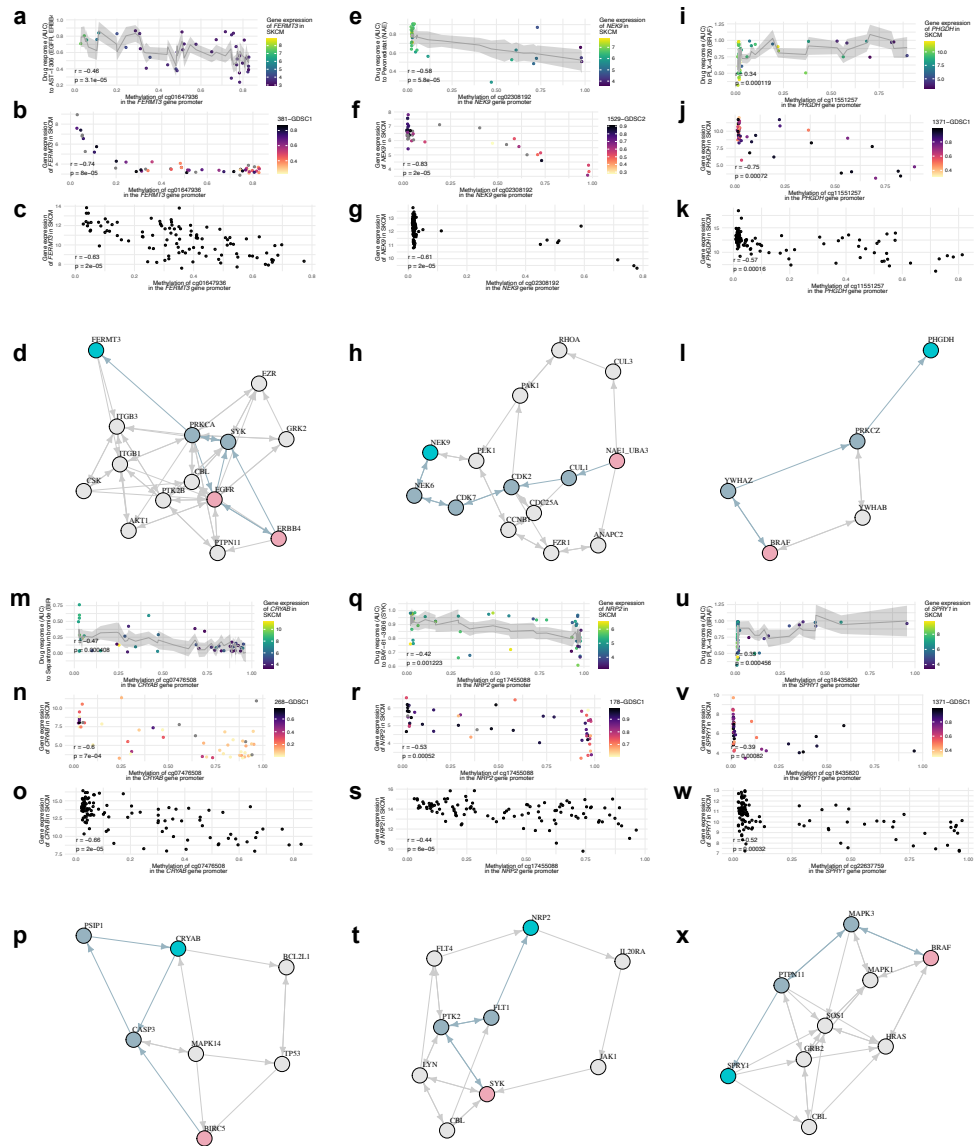
Supplementary Figure 2: Methylation pattern of dDMRs. Heatmap of dDMR methylation across cancer types.



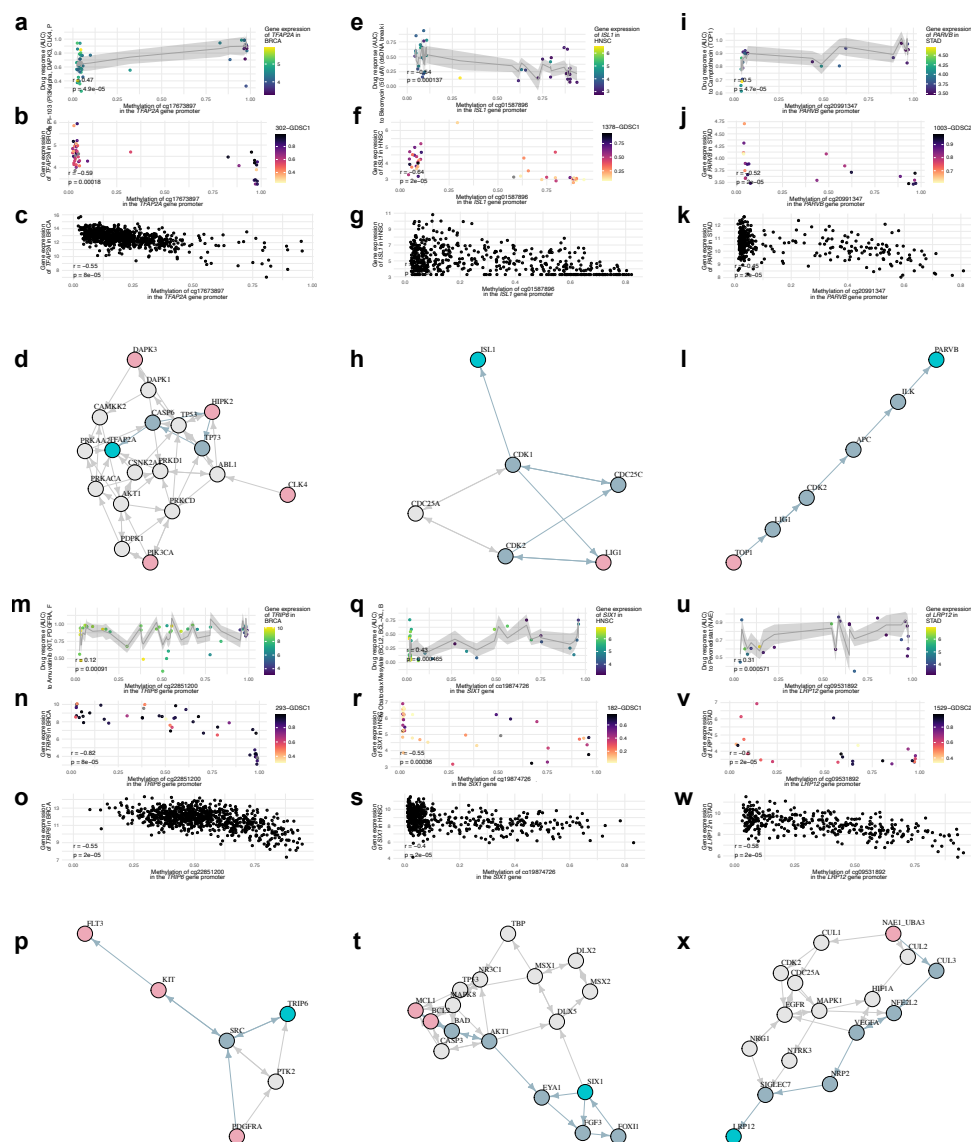
Supplementary Figure 3: Scatter plot of effect sizes from validation cohorts. (a) Consistency of effect sizes of 802 dDMRs from the GDSC discovery cohort validated in either the CTRP and CCLE datasets. (b) Effect sizes of dDMRs validated with the CTRP drug response HTS. (c) Effect sizes of overlapping dDMRs with the CCLE RRBS and CTRP data.

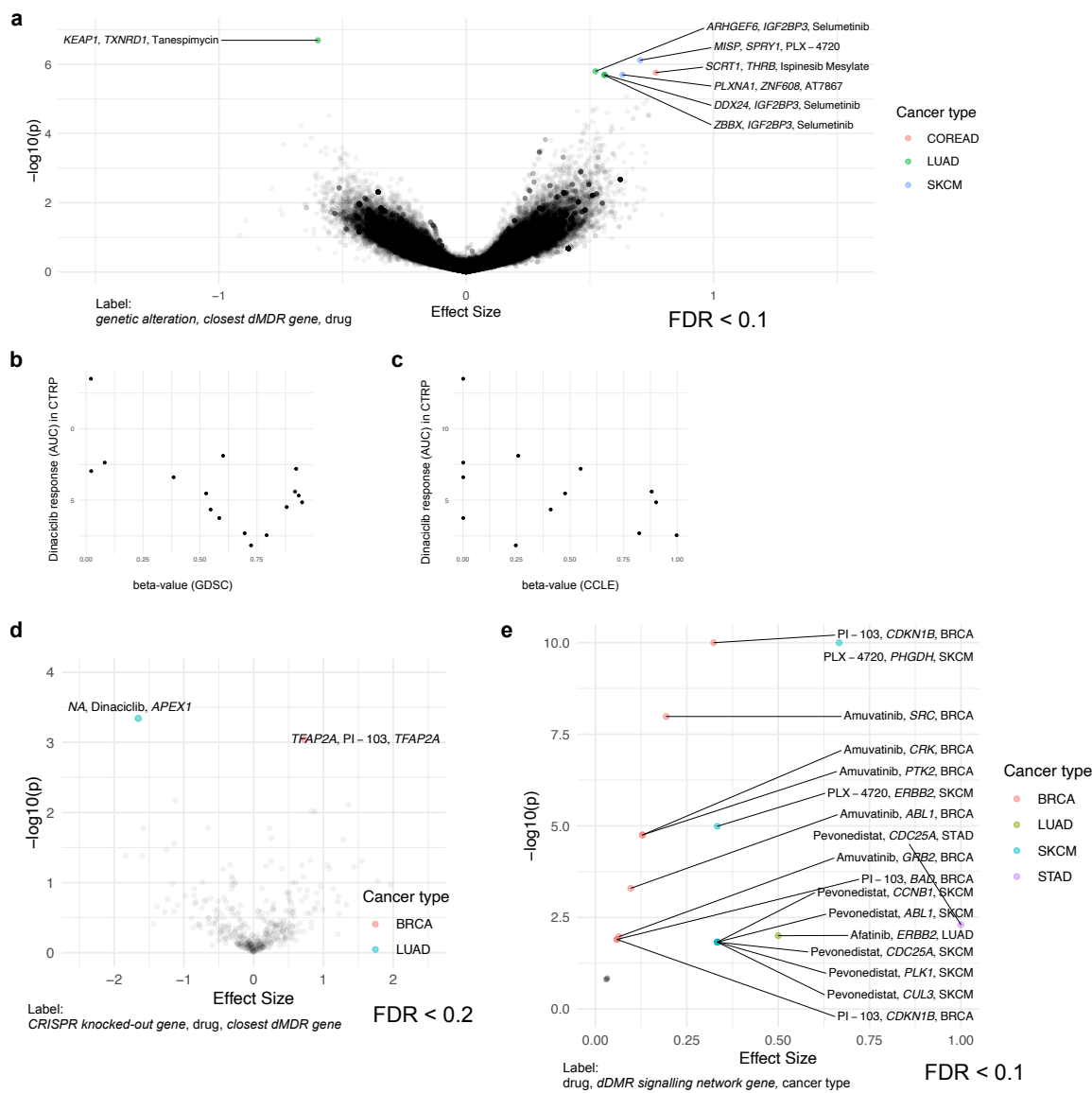


Supplementary Figure 4: tgdDMRs in LUAD. (a)-(ab) Correlation between drug response quantified by area-under-the-curve, DNA methylation and gene expression plus the corresponding protein-protein interaction network between putative drug target (pink) and tgdDMR-associated gene encoding protein (light blue). In the graph, nodes that are traversed with a shortest path are highlighted by the blue-grey colour among the alternative paths. For analysing DNA methylation and drug response, the error bars corresponding to 95% confidence intervals, the raw p-value (p) for each CpG site and the Pearson correlation coefficient (r) are reported. For analysing DNA methylation and gene expression, the empirical adjusted p-value (p) and the Pearson correlation coefficient (r) are reported.



Supplementary Figure 5: tgDDMRs in SKCM. (a)-(x) Correlation between drug response quantified by area-under-the-curve, DNA methylation and gene expression plus the corresponding protein-protein interaction network between putative drug target (pink) and tgDDMR-associated gene encoding protein (light blue). In the graph, nodes that are traversed with a shortest path are highlighted by the blue-grey colour among the alternative paths. For analysing DNA methylation and drug response, the error bars corresponding to 95% confidence intervals, the raw p-value (p) for each CpG site and the Pearson correlation coefficient (r) are reported. For analysing DNA methylation and gene expression, the empirical adjusted p-value (p) and the Pearson correlation coefficient (r) are reported.





Supplementary Figure 7: tgdDMRs in the context of genetic alterations, CRISPR screens and drug signatures. (a) A volcano plot summarising associations between tgdDMRs and somatic mutations in cancer cell lines. (b,c) Scatter plot for drug response to dinaciclib (AUC) in the CTRP validation set with the HumanMethylation450 BeadChip array in GDSC and reduced representation bisulfite sequencing in the CCLE independent dataset. (d) A volcano plot summarising associations between tgdDMRs and CRISPR knockout screens of the genes associated with tgdDMRs and their signalling network neighbourhood. (e) A volcano plot summarising enrichments of genes associated with tgdDMRs and their signalling network neighbourhood in the LINCS drug signatures for the matching compound and cancer type.

B.2 The pharmacogenomic assessment of molecular epithelial-mesenchymal transition signatures reveals drug susceptibilities in cancer cell lines, *bioRxiv* (2024), supplementary information

This material is supplementary to the open-access preprint article in *bioRxiv* presented in Section 2.2 [2]. It is publicly available at <https://doi.org/10.1101/2024.01.16.575190>.

Supplementary information for

The pharmacogenomic assessment of molecular epithelial-mesenchymal transition signatures reveals drug susceptibilities in cancer cell lines

Alexander J. Ohnmacht, Göksu Avar, Marisa K. Schübel, Thomas J. O'Neill, Daniel Krappmann, Michael P. Menden

Corresponding:

Daniel Krappmann. (daniel.krappmann@helmholtz-munich.de),

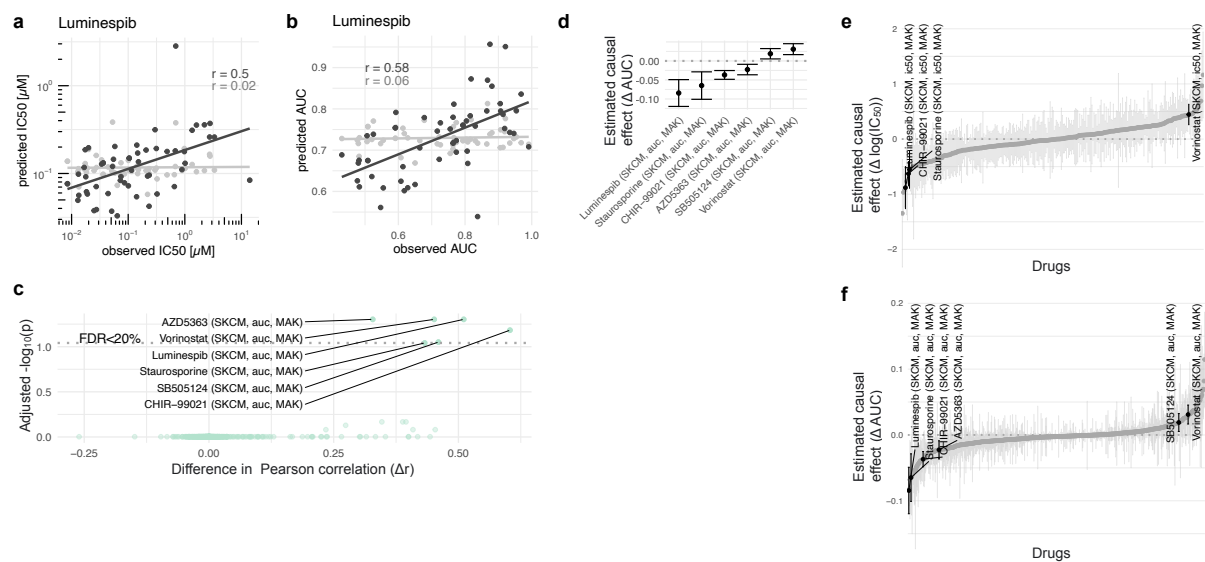
Michael P. Menden. (michael.menden@unimelb.edu.au)

This PDF file includes:

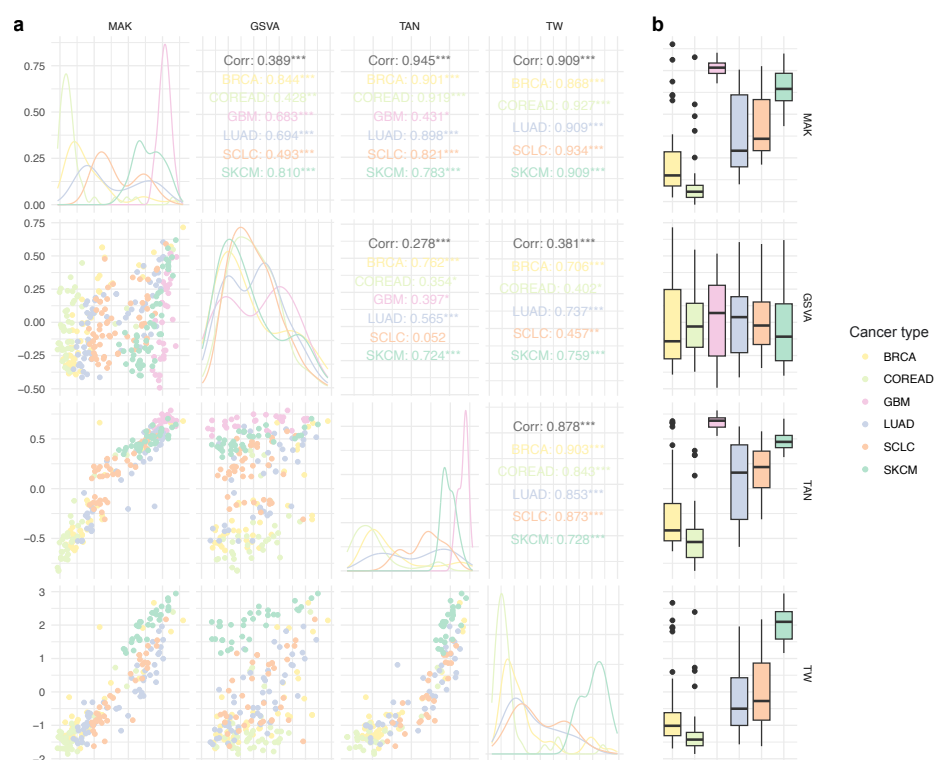
Supplementary Figure 1 to 6

Other supplementary information for this manuscript include:

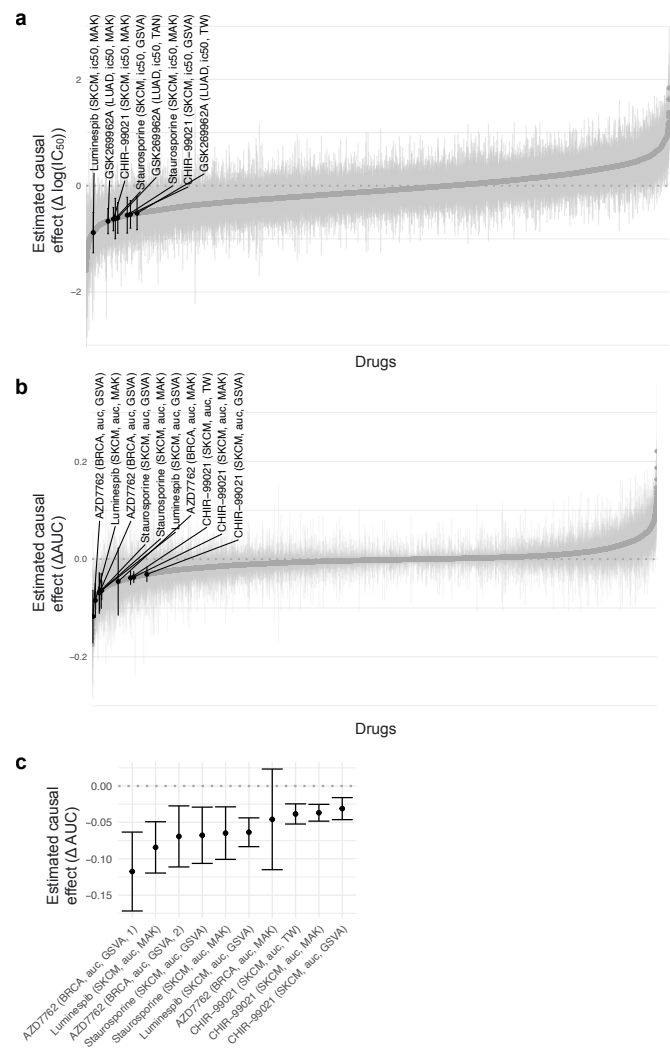
Supplementary Data 1 to 7



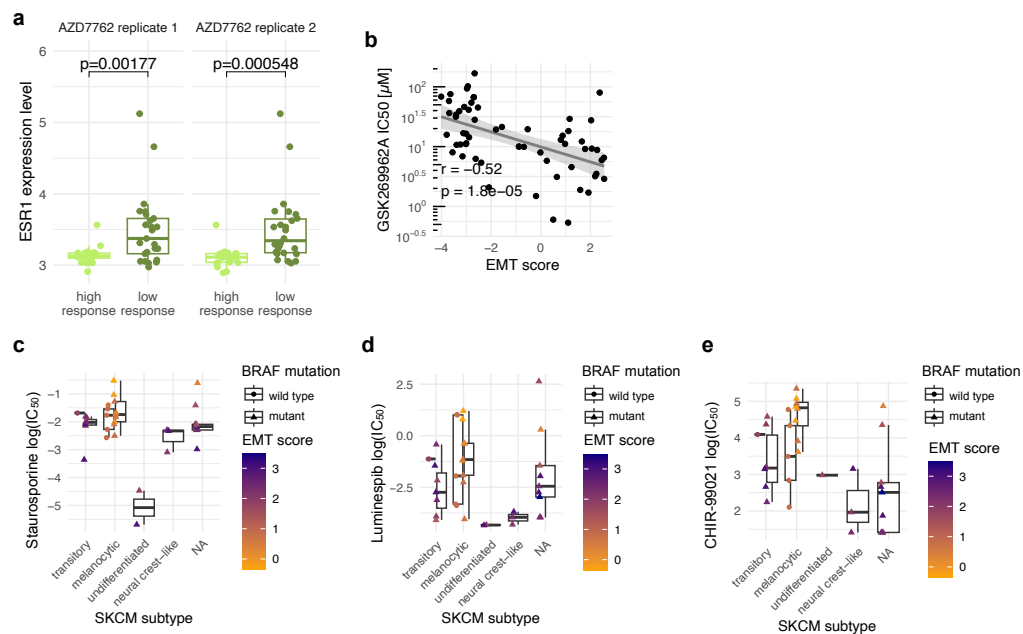
Supplementary Figure 1: Predictive and causal modelling in SKCM. The Scatter plot shows observed versus predicted (a) IC_{50} and (b) AUC values of luminespib for the full model (black: leveraging EMT scores and mutational background) and with ablation of the MAK EMT score (grey). (c) The volcano plot illustrates the ablation study in SKCM with AUC quantifying significance and effect size with adjusted p -values of a t -test for performance metrics and the difference in Pearson's correlation Δr , respectively. (d) The inferred EMT effects and 95% confidence intervals (CI) on drug responses in terms of AUC for the identified set of compounds in SKCM are shown. Furthermore, the estimated EMT effects with 95% CIs for all compounds in SKCM using (e) IC_{50} and (f) AUC values coloured in grey with significant compounds in black are highlighted.



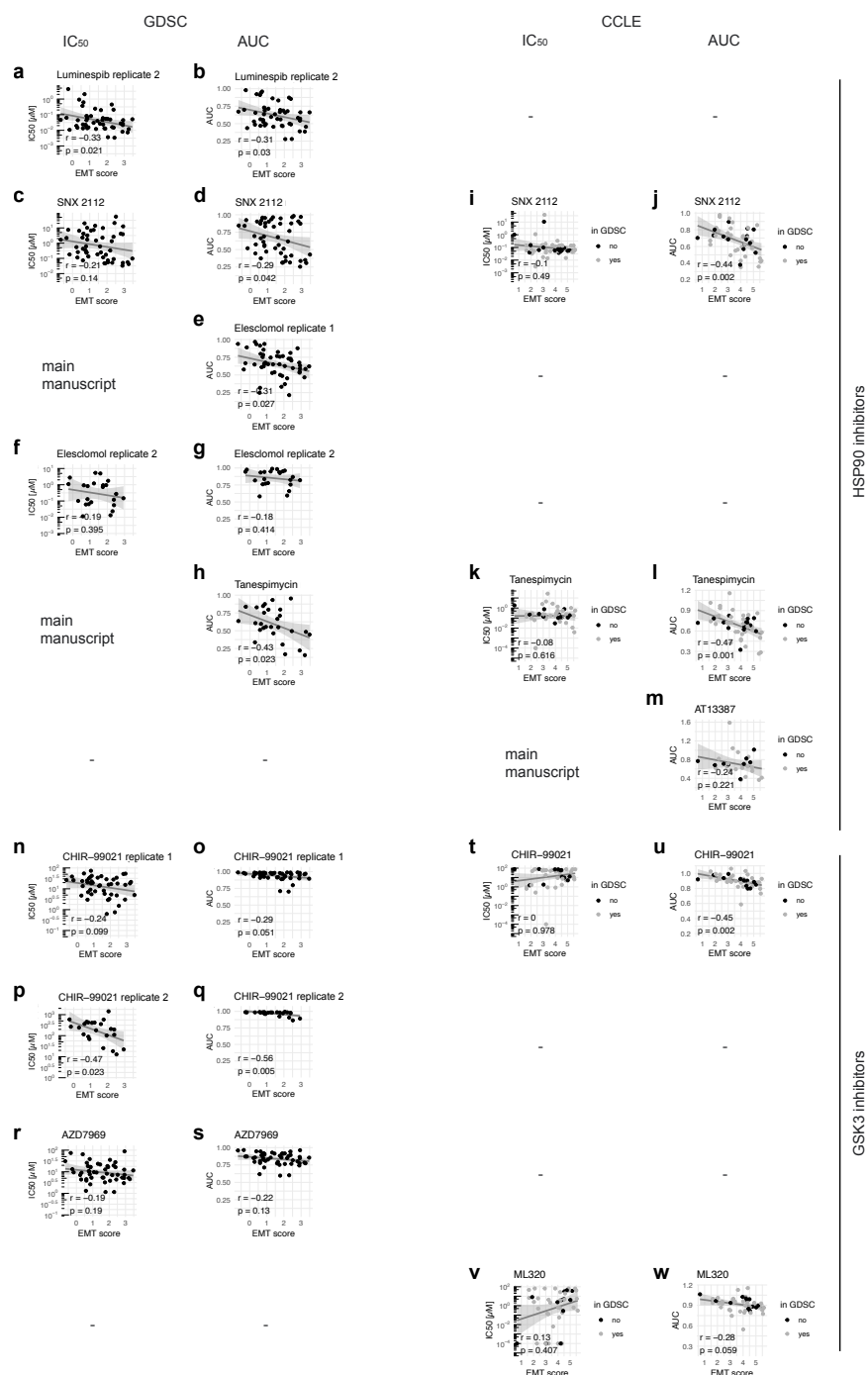
Supplementary Figure 2: Comparison of EMT scores. (a) Pairwise plots of the four derived EMT scores display their scatter plots (lower off-diagonal), their Pearson's correlation coefficient including significance of the correlation test, i.e. '***' for $p < 0.001$, '**' for $p < 0.01$, '*' for $p < 0.05$, '.' for $p < 0.1$ and no asterisk otherwise (upper off-diagonal), and their distribution (diagonal). (b) Boxplots demonstrating EMT scores of six cancer types that show at least one significant compound (FDR < 0.2) in the predictive modelling approach. The centre represents the median, while the box illustrates the interquartile range (IQR). The whiskers show a range that is 1.5 times the size of the IQR.



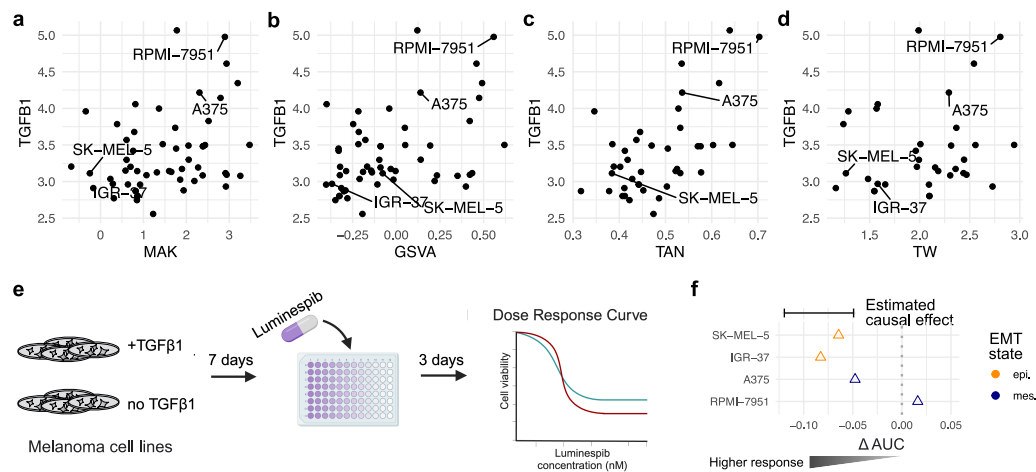
Supplementary Figure 3: Causal modelling across cancer types. The estimated EMT effects with 95% confidence intervals (CI) for all compounds across all cancer types using (a) IC_{50} and (b) AUC values coloured in grey with significant compounds in black are highlighted. (c) Furthermore, models for AUC values across all cancer types and EMT scores, the estimated EMT effect plus 95% CIs for the significant compounds are shown.



Supplementary Figure 4: Characterisation of EMT in the context of drug response. (a) The boxplots show *ESR1* expression of BRCA cell lines stratified by higher versus lower AZD7762 responding cell lines, i.e. discretisation by median IC₅₀ quantified by a two-sided *t*-test and its derived *p*-value. (b) The scatter plot highlights IC₅₀ values of LUAD cell lines treated with the ROCK inhibitor GSK269962A dependent on their MAK EMT score. The Pearson's correlation coefficient (*r*) and the associated *p*-value of the correlation test (*p*) are displayed. The boxplots show the response to (c) staurosporine, (d) luminespib and (e) CHIR-99021 depending on BRCA PAM50 subtypes, the MAK EMT score and mutations in BRCA1/2. The centre on the boxplot represents the median, while the box illustrates the interquartile range (IQR). The whiskers show a range that is 1.5 times the size of the IQR.



Supplementary Figure 5: Consistency of HSP90 and GSK3 inhibitors in GDSC and CCLE. (a-h) The scatter plots show drug responses to HSP90 inhibitors in the GDSC quantified by either IC₅₀ or AUC and their MAK EMT scores for the second luminespib replicate, SNX 2112, both elesclomol replicates and tanespimycin. The Pearson's correlation coefficient (r) and the associated p -value of the correlation test (p) are displayed. (i-m) The analogous scatter plots show HSP90 inhibitors SNX 2112, tanespimycin and AT13387 (onalespib) in the CCLE/CTRP. Cell lines in grey were also screened in the GDSC datasets, whereas cell lines in black were exclusively screened in the CCLE/CTRP dataset. (n-s) The scatter plots show drug responses to GSK3 inhibitors in the GDSC quantified by either IC₅₀ or AUC and their MAK EMT scores for the two CHIR-99021 replicates and AZD7969. (t-w) The analogous scatter plots show GSK3 inhibitors CHIR-99021 and ML320 in the CCLE/CTRP. Cell lines in grey were also screened in the GDSC datasets, while cell lines in black were exclusively screened in the CCLE/CTRP dataset. Dashes in the shown figure indicate missing drugs in the GDSC and CCLE/CTRP datasets, respectively.



Supplementary Figure 6: Validation experiments for luminespib with TGF- β pretreatment. *TGF β 1* expression is demonstrated in SKCM cell lines depending on their EMT scores: (a) MAK, (b) GSVA, (c) TAN, and (d) TW. (e) The cartoon for luminespib screens illustrates TGF- β 1 pretreatment. (f) TGF- β treated epithelial cell lines demonstrate higher responses (decreased AUC) to luminespib within the inferred causal effect 95% confidence interval.

B.3 The Oncology Biomarker Discovery framework reveals cetuximab and bevacizumab response patterns in metastatic colorectal cancer, *Nature Communications* (2023), supplementary information

This material is supplementary to the peer-reviewed and published open-access article in *Nature Communications* presented in Section 2.3 [3] and is reproduced with permission from Springer Nature. It is publicly available at <https://doi.org/10.1038/s41467-023-41011-4>.

Supplementary information for

The Oncology Biomarker Discovery framework reveals cetuximab and bevacizumab response patterns in metastatic colorectal cancer

Ohnmacht AJ, Stahler A, Stintzing S, Modest DP, Holch JW, Westphalen CB, Hölzel L, Schübel MK, Galhoz A, Farnoud A, Ud-Dean M, Vehling-Kaiser U, Decker T, Moehler M, Heinig M, Heinemann V, Menden MP

Corresponding:

Menden MP. (michael.menden@helmholtz-munich.de),
Heinemann V. (volker.heinemann@med.uni-muenchen.de)

This PDF file includes:

Supplementary Figure 1 to 16
Supplementary Table 1

Other supplementary information for this manuscript include:

Supplementary Data 1 to 4

B.3 The Oncology Biomarker Discovery framework reveals cetuximab and bevacizumab response patterns in metastatic colorectal cancer, *Nature Communications* (2023), supplementary information

Method abbreviation	Name	Implementation	Modelling type	Control of false positive rate	Output type	Found subgroups (OS)	Hazard ratio on OS for subgroup not in standard treatment and discovered by method [p-value; n = amount of patients]	Execution time [mins]
OncoBird	Oncology Biomarker Discovery	R package 'OncoBird'	stratified linear regression with interactions	Benjamini-Hochberg correction; permutations and bootstrapping for treatment effect correction	subgroups with interaction effects in predefined subgroups	<i>KRAS</i> mutations in CMS4; <i>KRAS/NRAS/SRC/BRAF</i> alterations in CMS4; <i>KRAS/NRAS/SRC</i> alterations in CMS4; <i>TOP1</i> amplification in CMS2; <i>ARFRP1</i> amplifications in CMS2	HR = 0.57 (p = 0.16, n = 29)	2.2
VT	Virtual twins	R package 'randomforestSRC'	random forests and regression tree	cross-validation	decision tree on biomarkers from global outcomes	<i>KRAS/NRAS/BRAF</i> alterations not in CMS2; CMS3	HR = 0.65 (p = 0.17, n = 48)	1.2
MOB	Model-based recursive partitioning	R package 'model4you'	recursive chi-squared independence tests	Bonferroni-correction	decision tree on recursive treatment effects	<i>KRAS/NRAS/BRAF/IRS2/NF1</i> alterations	HR = 0.82 (p = 0.62, n = 28)	0.1
OWE	Outcome weighting estimation	R package 'personalized'	outcome weighting or A-learning	cross-validation; bootstrapping	subgroups with interaction effects	<i>NRAS</i> mutations in left-sided; <i>SOX9</i> mutations in left-sided	HR = 0.87 (p = 0.44, n = 150)	14.7
CRF	Causal random forests	R package 'grf'	generalised random forests	honest trees	variable importance scores for treatment effects	<i>KRAS/NRAS/BRAF/IRS2/NF1</i> alterations; <i>KRAS/NRAS/BRAF</i> alterations; <i>ATM/TOP1</i> alterations	HR = 0.63 (p = 0.12, n = 50)	0.1
POL	Policy learning	R package 'policytree'	semi-parametrically efficient estimation	honest trees	decision tree on found subgroups	<i>KRAS/NRAS/BRAF</i> alterations; <i>PIK3CA/PTEN/GNAS/ERBB2</i> alterations in left-sided; <i>ARID1A/SMAD2/CUL4A</i> alterations	HR = 0.76 (p = 0.18, n = 117)	0.1
GUIDE	Generalized, unbiased, Interaction detection and estimation	Binary executable https://pages.cs.wisc.edu/~loh/guide.html	interaction tests	cross-validation and bootstrapping for treatment effect correction	subgroups with interaction effects	<i>KRAS/NRAS/BRAF</i> alterations; CMS3	HR = 0.76 (p = 0.30, n = 60)	0.4
PRISM	Patient response identifiers for Stratified Medicine	R package 'StratifiedMedicine'	virtual twins or model-based partitioning with post parameter estimation	bootstrapping	tree from bayesian posterior distribution in found subgroups	<i>KRAS/NRAS/BRAF</i> alterations in left-sided	no new subgroup	0.2
SIDES	Subgroup identification based on differential effect search	R package 'SIDES'	differential effects search	Šidák multiplicity adjustment and independent validation	subgroups with interaction effects	none	-	1.9

Supplementary Table 1: Benchmark of OncoBird with other methods. Qualitative and quantitative comparison between OncoBird and previously published data-driven subgroup analysis methods, i.e. virtual twins (VT), model-based partitioning (MOB), an outcome-weighted method (OWE), causal random forests (CRF), policy learning (POL), GUIDE, PRISM and SIDES.

b Uploading Files

Upload File

Subtype Enrichment

Genomic Enrichments

Mutual Exclusivity

Ontoprint

Treatment Specific Biomarkers

Treatment Specific Biomarker Subtypes

Predictive Biomarkers

Predictive Biomarker Subtypes

Predictive Comparison

Plot Example

Summary

Uploading Files

Choose "data_mutations"

Browse... data_mutations.csv

Upload complete

Choose "data_clinical"

Browse... data_clinical.csv

Upload complete

Select Clinical Endpoint column

OS

Select treatment column

treatment

Select patientID column

sample

Prepare data

Choose multiple columns for as Tumour subtypes

CMS primary.site

Choose 2 Treatments

FOLFIRI Bevacizumab FOLFIRI Cetuximab

Choose up to 2 Covariates

resected.1stline metastatic.site

Submit

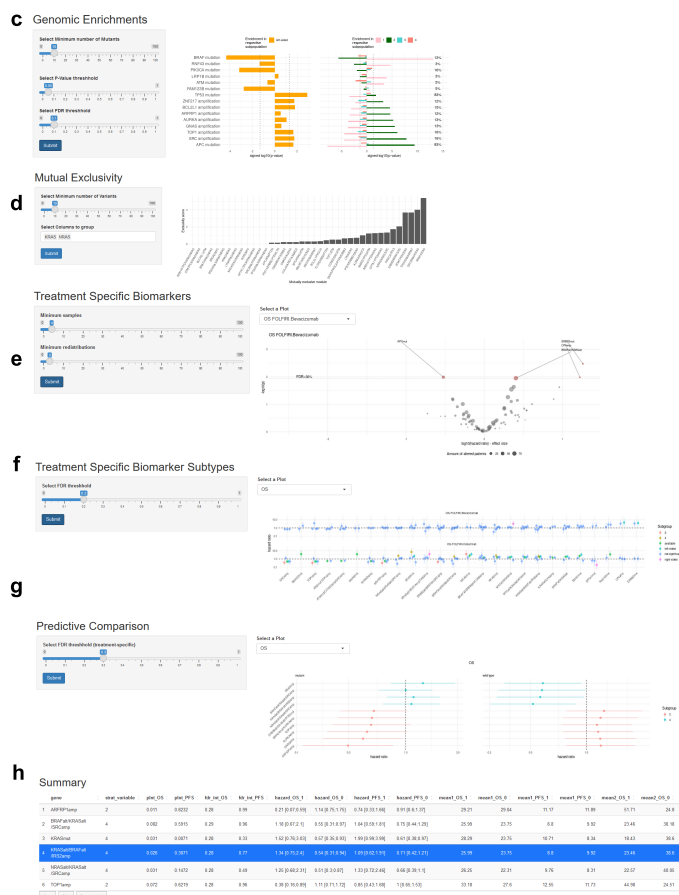
☐ Recalculate Mutex

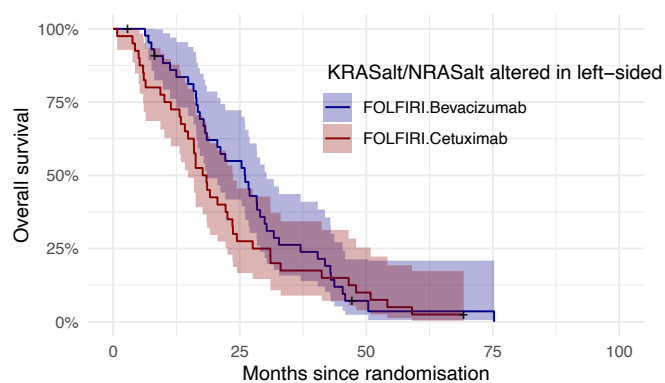
c Genomic Enrichments

d Mutual Exclusivity

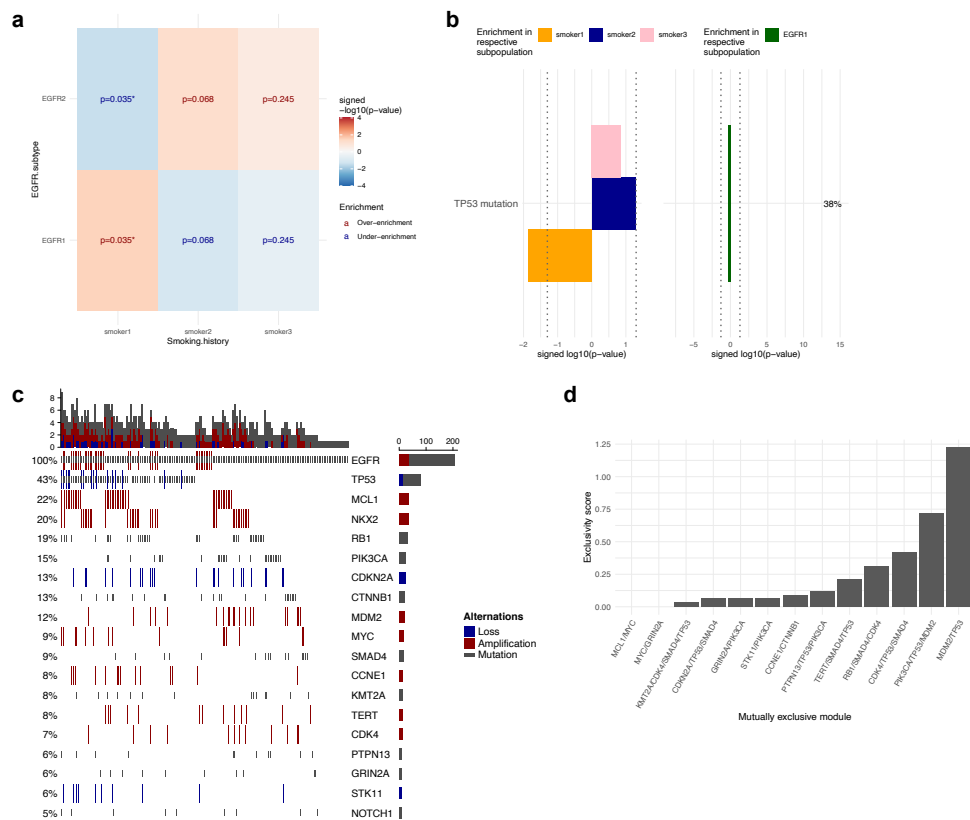
e Treatment Specific Biomarkers

f Treatment Specific Biomarker Subtypes

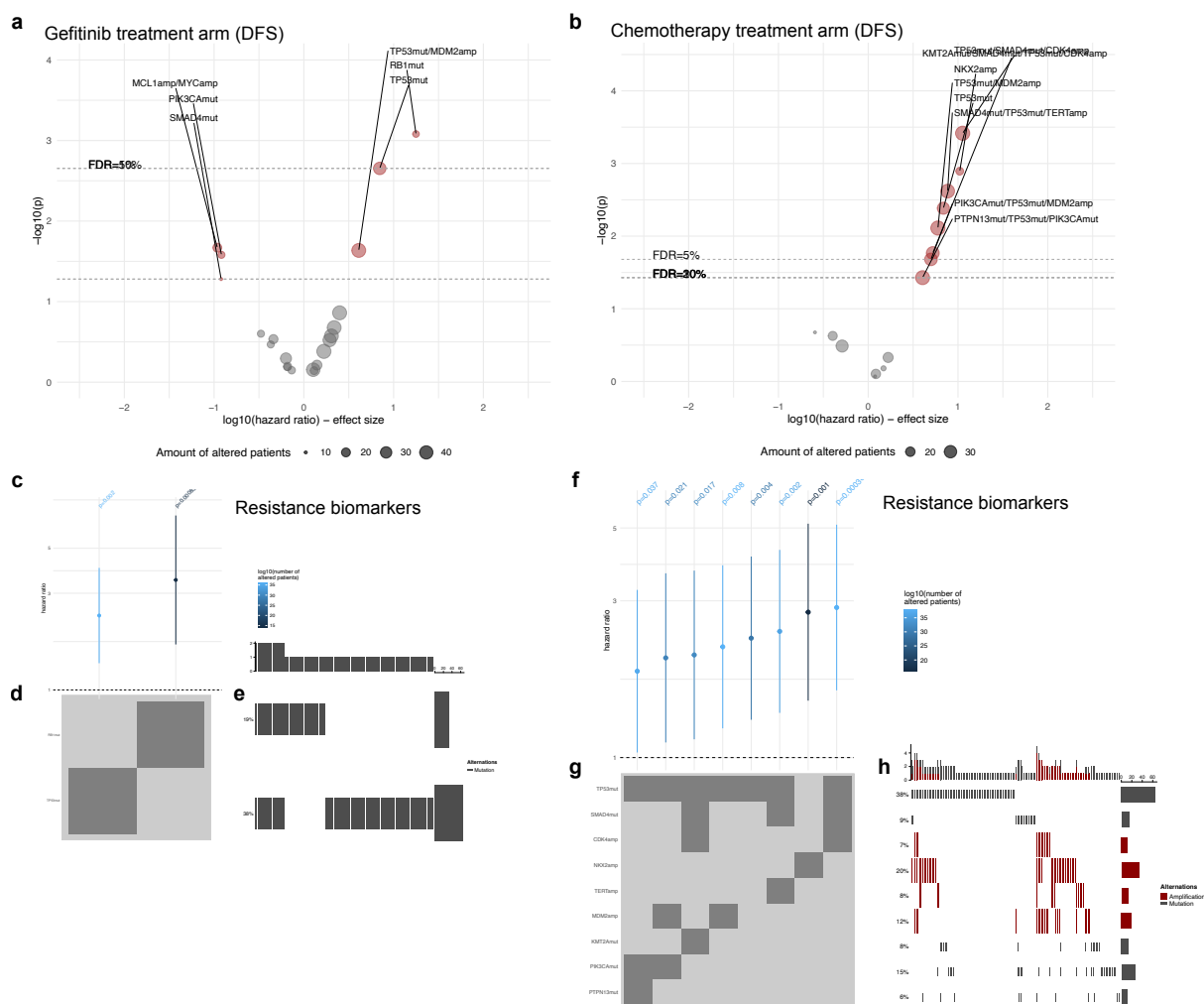




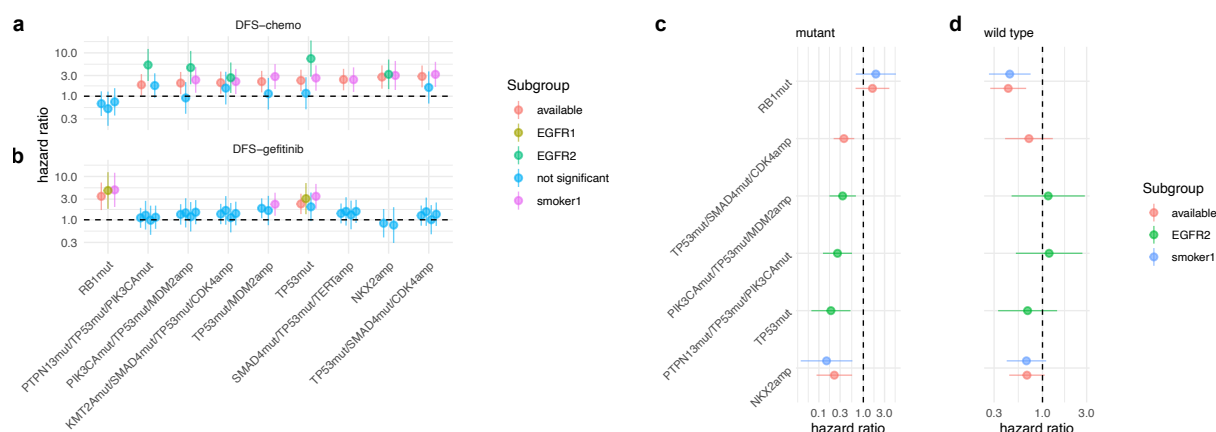
Supplementary Figure 2: Predictive biomarkers in metastatic colorectal cancer. Kaplan-Meier plot including 95% confidence intervals (CI) showing left-sided tumours that are mutated in either *KRAS* or *NRAS* treated with either cetuximab or bevacizumab.



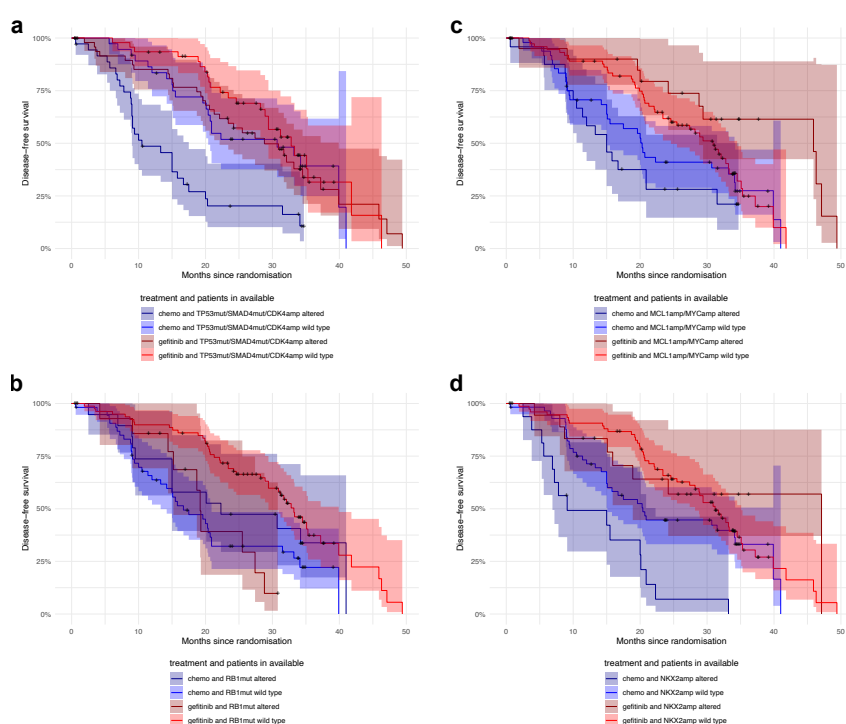
Supplementary Figure 3: The molecular landscape of non-small cell lung cancer. (a) Interactions between EGFR subtype and smoking status as tumour subtypes showing raw p-values derived from one-sided hypergeometric tests. (b) Altered cancer genes tested for enrichment in these tumour subtypes using one-sided hypergeometric tests showing raw p-values. (c) Oncoprint of 171 tumours, including mutations and copy number alterations detected in more than ten tumours. (d) The mutually exclusive alteration patterns were derived with the Mutex algorithm. Source data for the figure panels are provided as **Source Data** file.



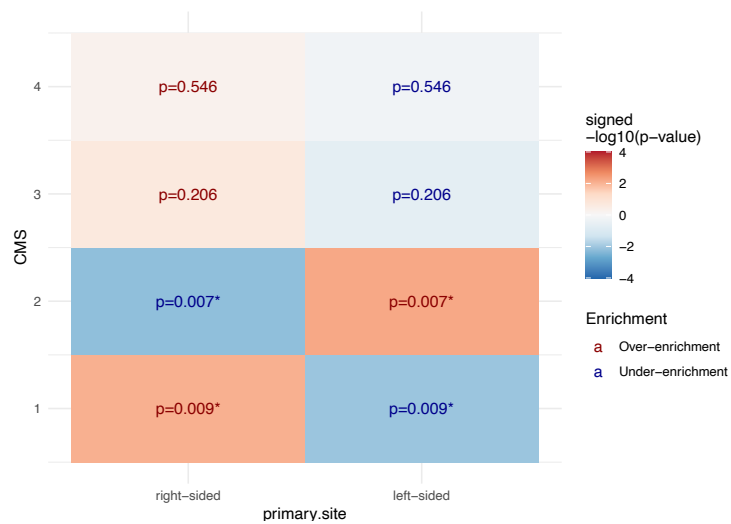
Supplementary Figure 4: Genetic biomarkers of DFS for gefitinib or chemotherapy. Prognostic value of mutually exclusive gene modules or single gene mutations for (a) gefitinib or (b) chemotherapy. Each point shows the effect of a particular group of alterations summarised by its hazard ratio derived by the Cox regression models and its raw p-value derived by a Wald test. We focus on (c) resistance biomarkers of gefitinib with $FDR_{gef/che} < 0.1$, showing their hazard ratios and 95% confidence intervals (CI), (d) the composition of mutually exclusive gene modules indicated by dark grey colour, and (e) an oncoprint highlighting mutational frequencies and patterns. In like manner, (f) chemotherapy resistance biomarkers with their hazard ratios and 95% CIs and their (g) composition are shown with (h) their oncoprint. A **Source Data** file is provided, which contains the source data for the figure panels and the sample sizes of the conducted statistical tests.



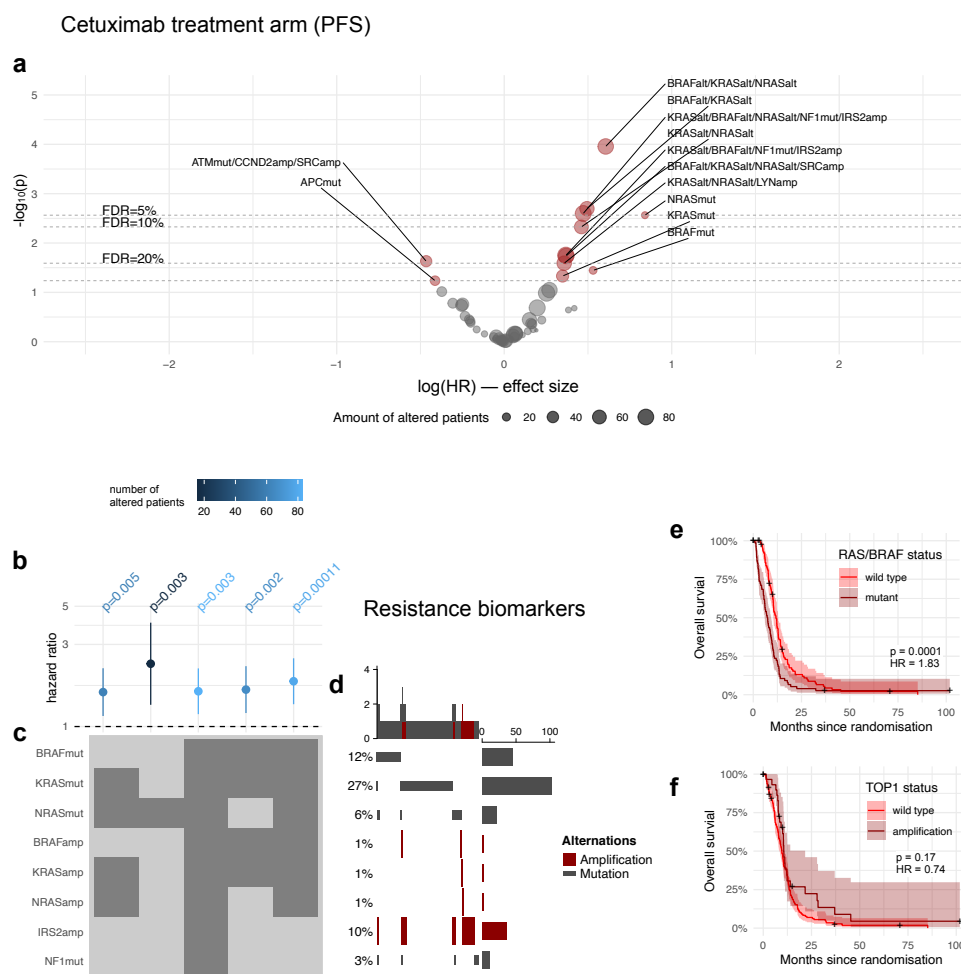
Supplementary Figure 5: Identification of subtype-specific predictive biomarkers for gefitinib or chemotherapy. Subtype-specific genetic biomarkers for DFS of (a) chemotherapy and (b) gefitinib using hazard ratios including 95% confidence intervals (CI) derived from single Cox regression models. Subtypes are defined by either the EGFR subtype, smoking status or unstratified (reference model). A **Source Data** file is provided, which contains the source data for the figure panels and the sample sizes of the conducted statistical tests. Overview of interaction biomarkers ($FDR_{int} < 0.2$) focusing on (c) mutant and (d) wild type tumours, using hazard ratios including 95% CIs derived from single Cox regression models fitted on DFS when comparing gefitinib or chemotherapy treatment. For the conducted statistical tests, the sample sizes are given in **Supplementary Data 2**. Source data for the figure panels are provided as **Source Data** file.



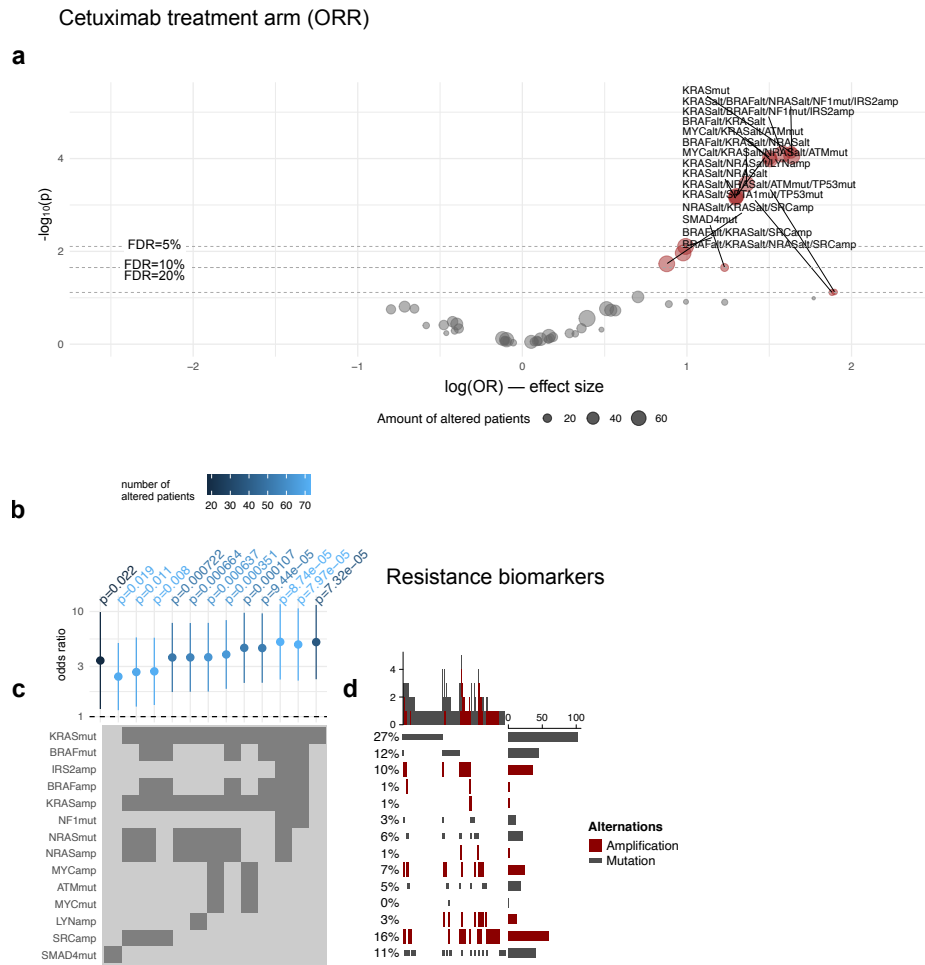
Supplementary Figure 6: Predictive biomarkers for gefitinib. Kaplan-Meier plots including 95% confidence intervals (CI) of genetic biomarkers for (a) mutually exclusive module consisting of mutations in *TP53*, *SMAD4* or *CDK4* amplifications, (b) *RB1* mutations, (c) amplifications in *MCL1* or *MYC*, and (d) *NKX2-1* amplifications across all patients in either the gefitinib or chemotherapy treatment arm.



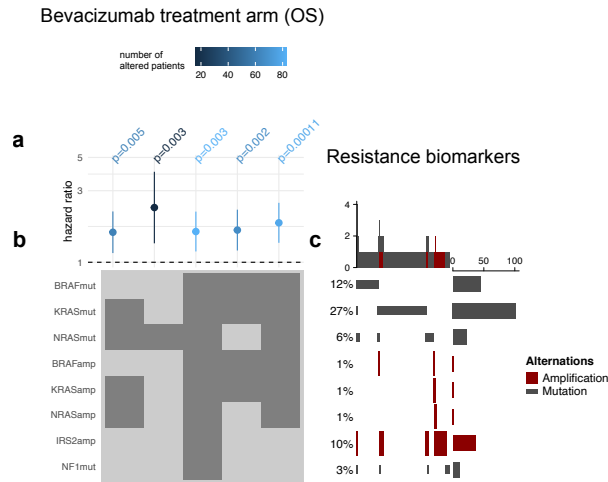
Supplementary Figure 7: Tumour subtypes in metastatic colorectal cancer. Interactions between tumour sidedness and CMS subtypes showing raw p-values derived from one-sided hypergeometric tests.



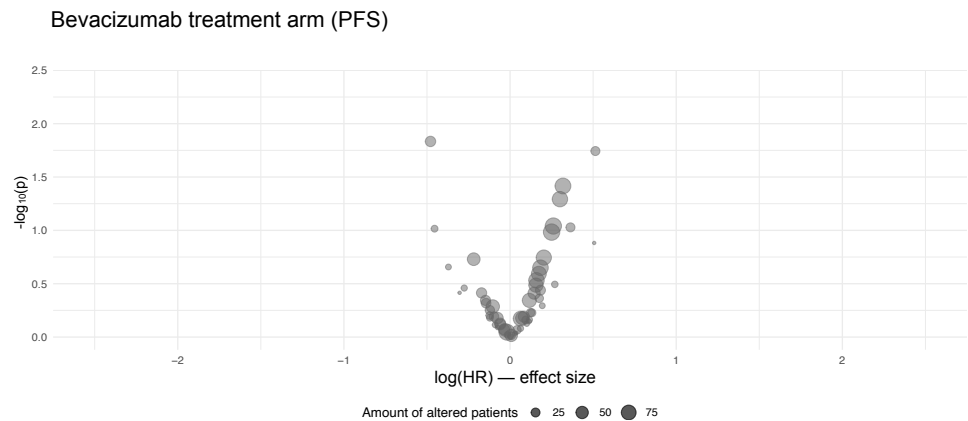
Supplementary Figure 8: Genetic biomarkers of PFS for cetuximab. (a) Prognostic value of mutually exclusive gene modules or single gene mutations for cetuximab. Each point shows the effect of a particular group of alterations summarised by its hazard ratio derived by the Cox regression models and its raw p-value derived by a Wald test. We focus on (b) resistance biomarkers of FOLFIRI plus cetuximab with $\text{FDR}_{\text{cet}} < 0.1$, showing their hazard ratios and 95% confidence intervals (CI), (c) the composition of mutually exclusive gene modules indicated by dark grey colour, and (d) an oncoprint highlighting mutational frequencies and patterns. Kaplan-Meier plot including 95% CIs and summary statistics of the Cox regression models for (e) *RAS* or *BRAF* mutations and (f) *TOP1* amplifications. No sensitivity biomarkers were found with $\text{FDR}_{\text{cet}} < 0.1$. A **Source Data** file is provided, which contains the source data for the figure panels and the sample sizes of the conducted statistical tests.



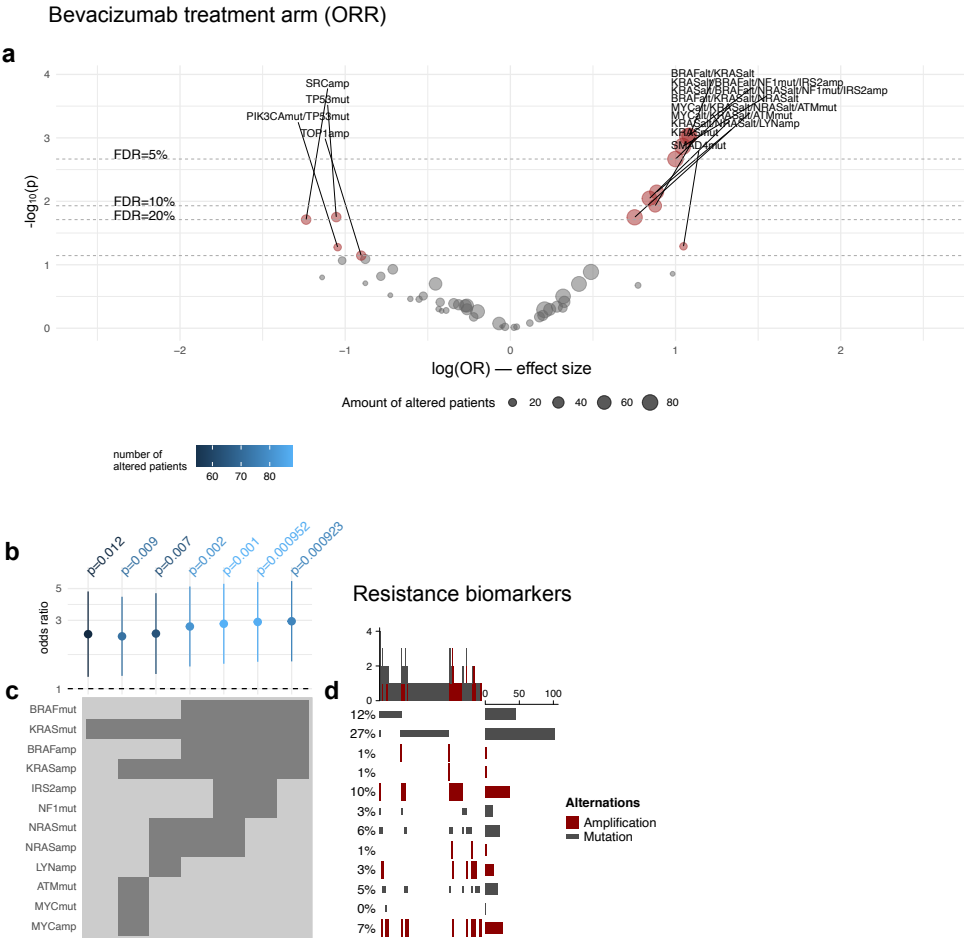
Supplementary Figure 9: Genetic biomarkers of ORR for cetuximab. (a) Prognostic value of mutually exclusive gene modules or single gene mutations for cetuximab. Each point shows the effect of a particular group of alterations summarised by its odds ratio derived by the logistic regression models and its raw p-value derived by a Wald test. We focus on (b) resistance biomarkers of FOLFIRI plus cetuximab with $FDR_{cet} < 0.1$ showing their odds ratios and 95% confidence intervals (CI), (c) the composition of mutually exclusive gene modules indicated by dark grey colour, and (d) an oncoprint highlighting mutational frequencies and patterns. No sensitivity biomarkers were found with $FDR_{cet} < 0.1$. A **Source Data** file is provided, which contains the source data for the figure panels and the sample sizes of the conducted statistical tests.



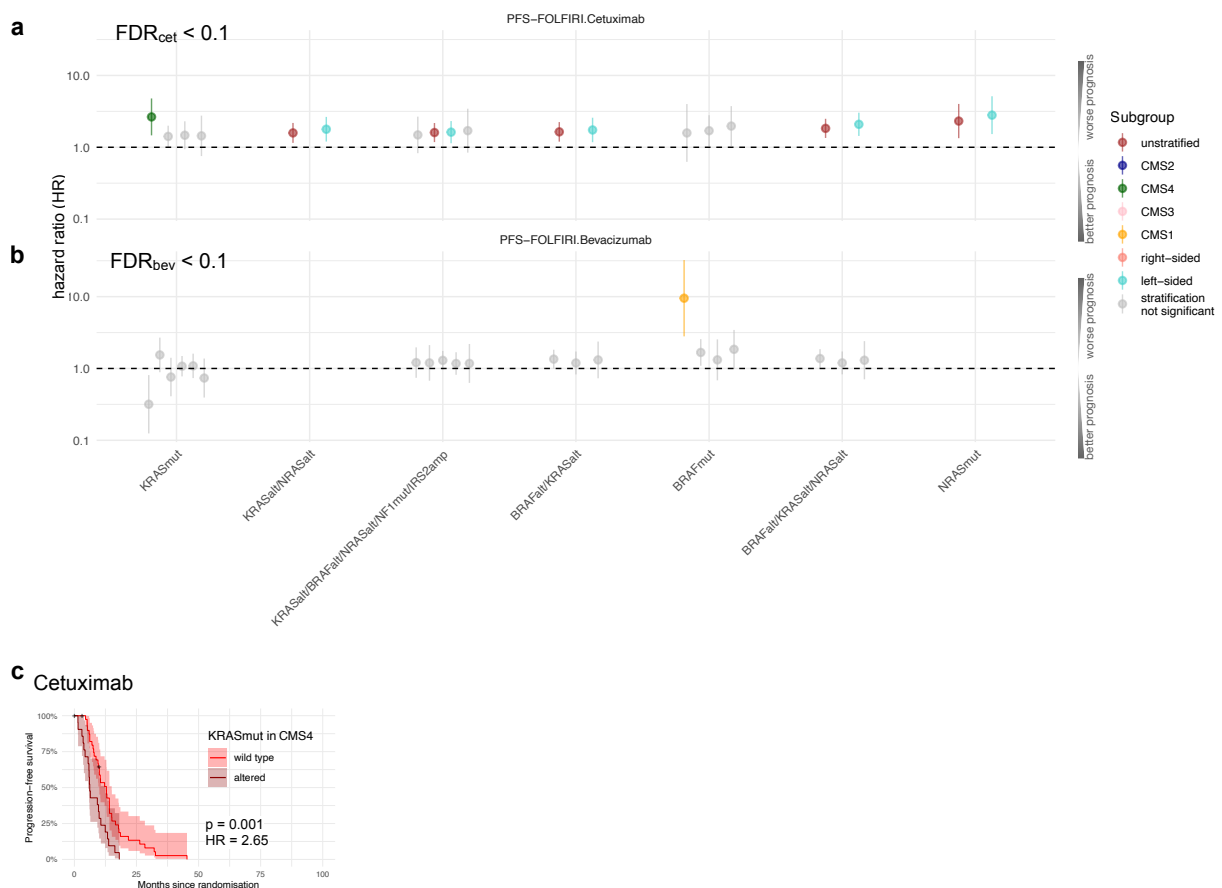
Supplementary Figure 10: Genetic biomarkers of OS for bevacizumab. We focus on (a) resistance biomarkers of FOLFIRI plus bevacizumab with $FDR_{bev} < 0.3$, showing their hazard ratios and 95% confidence intervals (CI), (b) the composition of mutually exclusive gene modules indicated by dark grey colour, and (c) an oncoprint highlighting mutational frequencies and patterns. Apart from *APC* mutations, no further sensitivity biomarkers were found with $FDR_{bev} < 0.3$. A **Source Data** file is provided, which contains the source data for the figure panels and the sample sizes of the conducted statistical tests.



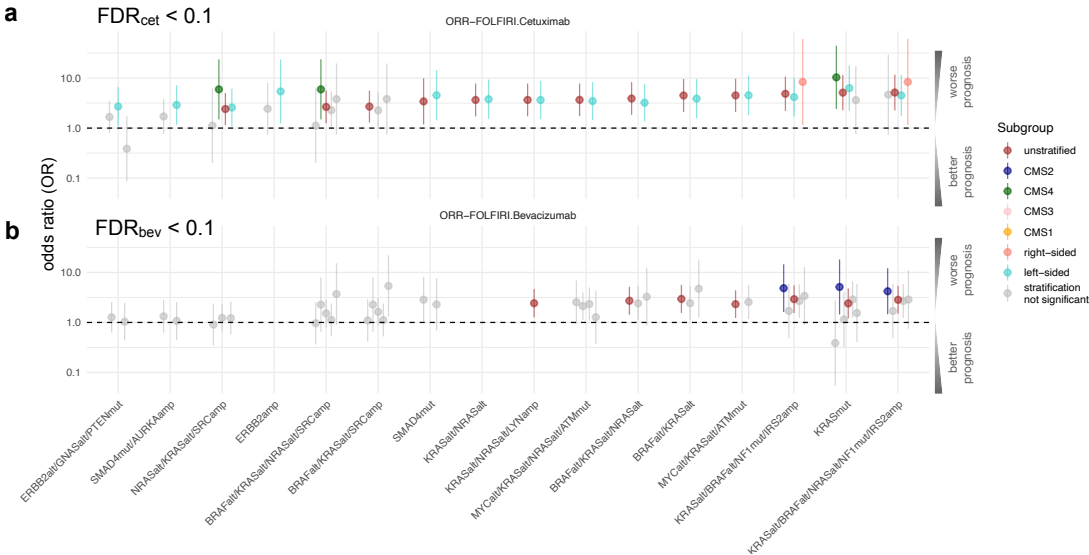
Supplementary Figure 11: Genetic biomarkers of PFS for bevacizumab. No prognostic value of mutually exclusive gene modules or single gene mutations for bevacizumab. Each point shows the effect of a particular group of alterations summarised by its hazard ratio derived by the Cox regression models and its raw p-value derived by a Wald test. A **Source Data** file is provided, which contains the source data for the figure panels and the sample sizes of the conducted statistical tests.



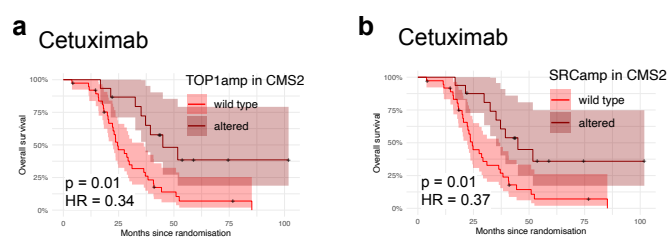
Supplementary Figure 12: Genetic biomarkers of ORR for bevacizumab. (a) Prognostic value of mutually exclusive gene modules or single gene mutations for bevacizumab. Each point shows the effect of a particular group of alterations summarised by its odds ratio derived by the logistic regression models and its raw p-value derived by a Wald test. We focus on (b) resistance biomarkers of FOLFIRI plus bevacizumab with $FDR_{bev} < 0.1$, showing their hazard ratios and 95% confidence intervals (CI), (c) the composition of mutually exclusive gene modules indicated by dark grey colour, and (d) an oncoprint highlighting mutational frequencies and patterns. No sensitivity biomarkers were found with $FDR_{bev} < 0.1$. A **Source Data** file is provided, which contains the source data for the figure panels and the sample sizes of the conducted statistical tests.



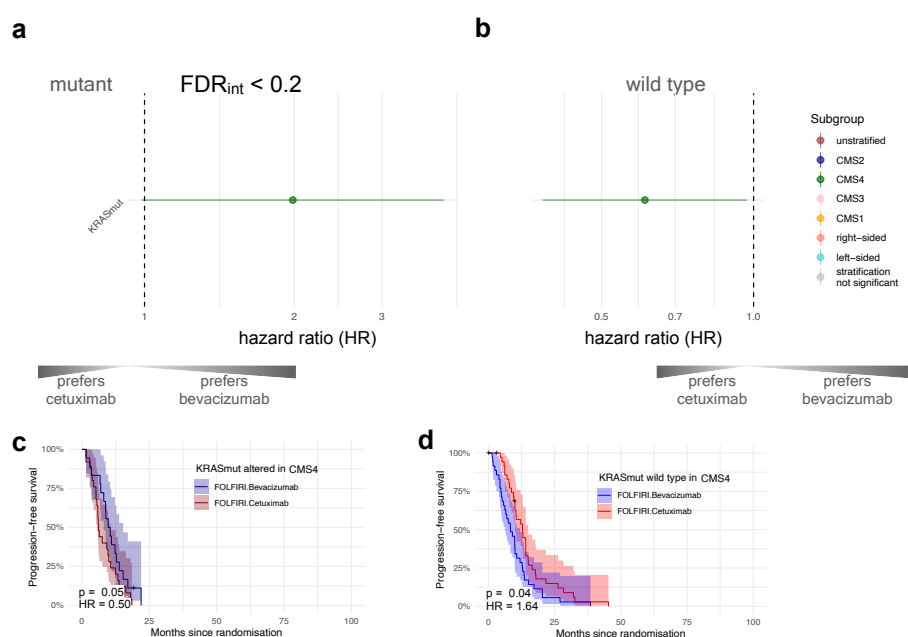
Supplementary Figure 13: Identification of subtype-specific biomarkers for FOLFIRI plus cetuximab or bevacizumab in terms of PFS. Subtype-specific prognostic value of genetic markers for PFS of (a) cetuximab and (b) bevacizumab using hazard ratios including 95% confidence intervals (CI) derived from single Cox regression models. Subtypes are defined by either the primary tumour side, CMS or unstratified (reference model). (c) Kaplan-Meier plot including 95% CIs, hazard ratios and raw p-values derived by Wald tests from the Cox regression models showing PFS of CMS4 tumours stratified by *KRAS* mutations treated with cetuximab. A **Source Data** file is provided, which contains the source data for the figure panels and the sample sizes of the conducted statistical tests.



Supplementary Figure 14: Identification of subtype-specific biomarkers for FOLFIRI plus cetuximab or bevacizumab in terms of ORR. Subtype-specific prognostic value of biomarkers for ORR of (a) cetuximab and (b) bevacizumab using odds ratios including 95% confidence intervals (CI) derived from single logistic regression models. Subtypes are defined by either the primary tumour side, CMS or unstratified (reference model). A **Source Data** file is provided, which contains the source data for the figure panels and the sample sizes of the conducted statistical tests.



Supplementary Figure 15: Amplifications in chr20q in tumours treated with cetuximab. Kaplan-Meier plots including 95% CIs, hazard ratios and raw p-values derived by Wald tests from the Cox regression models of subtype-specific prognostic biomarkers for OS and cetuximab showing (a) *TOP1* amplifications or *SRC* amplifications in CMS2, respectively.



Supplementary Figure 16: Predictive interaction biomarkers in the context of tumour subtypes in terms of PFS. Overview of interaction biomarkers focusing on (a) mutant and (b) wild type populations when comparing cetuximab and bevacizumab with an interaction $FDR_{int} < 0.2$ using hazard ratios including 95% confidence intervals (CI) derived from single Cox regression models fitted on PFS. For the conducted statistical tests, the sample sizes are given in **Supplementary Data 3**. Here exemplified, Kaplan-Meier plots including 95% CIs, hazard ratios and raw p-values derived by Wald tests from the Cox regression models comparing mutant and wild type cohorts of (c,d) *KRAS* mutations in CMS4. Source data for the figure panels are provided as **Source Data** file.