Pharmacoproteomic characterisation of human colon and rectal cancer



Martin Frejno
Keble College
University of Oxford

Thesis submitted for the degree of
Doctor of Philosophy
Trinity 2016

All models are wrong but some are useful (Box, 1979)

Acknowledgements

First and foremost, I would like to thank my supervisors Stephan Feller and Stefan Knapp, as well as my collaborator and clandestine third supervisor Bernhard Küster. Without their help, guidance and support throughout my DPhil, I would have not been able to complete this project. Special thanks to Stephan Feller for his preliminary work with the cell line panel this thesis revolves around, as well as for countless hours spent in the cell culture in preparation of this study. I would like to specifically thank Stefan Knapp for taking me on as a DPhil student after Stephan Feller left Oxford, for helpful discussions and financial support. My sincere thanks also goes to Bernhard Küster for allowing me to spend time in his laboratory, for introducing me to mass spectrometry and for allowing me to use up more than 200 days worth of mass spectrometer measurement time.

I would also like to thank all my colleagues in the Feller, Knapp and Küster laboratories, who enriched the past four years both personally and professionally. Special thanks goes to Jenny, Katharina, Marc and Philip for their advice, to Julia Meier for help with western blots, to Riccardo for help with the hSAX fractionations and desalting, to Mathias and Chen for introducing me to and helping me with R, to Susan, Jana, Heiner, Runsheng and Julia Mergner for taking care of mass spectrometric measurements, to Anna and Steffi for help with Kinobeads pulldowns, to Ben, Guillaume and Li-Hua for their preliminary work enabling this thesis, to Karl for follow-up experiments and to all other members of these three laboratories, as well as to my transfer examiners Benedikt Kessler and Val Macaulay for fruitful discussion and welcome distractions. Thanks also to David Kerr and Elaine Johnstone, to Marcus Green and Helen Scott, to Julia Slotta-Huspenina and Moritz Jesinghaus, as well as to Enric Domingo for provision, as well as help with the immunohistochemistry, the analysis and survival modelling of the QUASAR2 TMAs, respectively. I would also like to thank Michaela Krötz-Fahning, Andrea Hubauer and Andreas Klaus for their laboratory assistance.

I am indebted to the Medical Research Council, the Department of Oncology and the late Mrs Jane Ledig-Rowohlt (nee Scatcherd) for financing my DPhil.

Many thanks also to all my close friends for listening to boring explanations about what I do and why I do it. I am sorry for not having spent more time with you during the past four years and hope I can make up for it in the years to come.

Last but not least, I would like to thank my family. I sincerely thank my parents Anton and Johanna for always believing in me and supporting me no matter what the cost or effort. Many thanks also to my brother for many humorous discussions and his boundless honesty. Finally, a special thanks goes to my wife Andrea for her often moral, sometimes statistical and always unconditional support, her endurance, as well as her praise and critique. I dedicate this thesis to our history together, the collective presence we enjoy and the shared future we will cherish.

I love you.

Content

Acknowledgements v
Abbreviations xi
Figures xiv
Tables xiv

Abstract 1

Chapter 1 Introduction 3
1.1 Colorectal cancer 3
1.1.1 Epidemiology 3
1.1.2 Prognosis 3
1.1.3 Staging 4
1.1.4 Screening 4
1.1.5 Management 5
1.1.6 Differential drug sensitivity and predictive biomarkers 7
1.1.7 Advances in cancer genomics 9
1.1.7.1 Microarrays 9
1.1.7.2 mRNA-Seq 10
1.1.7.3 Tumour heterogeneity and gene signatures 11
1.1.7.4 Molecular subtyping 12
1.2 Associating drug response and omics data of cell lines 13
1.2.1 Pharmacogenomics 13
1.2.2 Pharmacoproteomics 17
1.2.3 Elastic net regression 19
1.3 Mass-spectrometry-based bottom-up proteomics 21
1.3.1 Instrumentation 23
1.3.1.1 History 23
1.3.1.2 Soft ionisation techniques 24
1.3.1.3 Ion optics 26
1.3.1.4 Mass analysers 27
1.3.1.5 Gas-phase fragmentation and tandem mass spectrometry 29
1.3.2 MaxQuant 32

1.3.2.1 Peptide and protein identification by database search 32
1.3.2.2 Protein quantification 34
1.4 Aim of this work 36

Chapter 2 Materials and Methods 43
2.1 Chemicals 43
2.2 Cell lines 43
2.3 Cell culture and lysis 45
2.4 Patient samples 47
2.5 Sample processing for mass spectrometry 48
2.5.1 Full proteomes 48
2.5.1.1 Acetone precipitation 48
2.5.1.2 Resolubilisation 48
2.5.1.3 In-solution digestion 49
2.5.1.4 Desalting 49
2.5.1.5 Hydrophilic strong anion exchange chromatography 50
2.5.1.6 Post-hSAX desalting 51
2.5.2 Kinobeads 52
2.5.2.1 Kinobeads in microplates 52
2.5.2.2 Kinobeads in microcolumns 53
2.5.2.3 Kinobeads in 15 mL centrifuge tubes 55
2.5.2.4 Detergent removal from Kinobeads eluates 55
2.5.2.5 In-gel digestion 56
2.6 LC-MS/MS data acquisition 57
2.6.1 Full proteomes 57
2.6.2 Kinobeads 58
2.7 Analysis of LC-MS/MS raw data & giBAQ quantification 59
2.7.1 Full proteomes and reanalysis of the CPTAC patient data 59
2.7.2 giBAQ quantification 60
2.7.3 Kinobeads 63
2.7.4 LFQ quantification 63
2.8 Postprocessing of quantitative proteomics data 64
2.8.1 Full Proteomes 64

2.8.2 Kinobeads 64
2.9 Reanalysis of public mRNA datasets 65
2.9.1 GSE36133 66
2.9.2 E-MTAB-783 67
2.9.3 E-MTAB-2706 67
2.9.4 GSE28567 68
2.9.5 GSE59857 68
2.9.6 Supplementary Table S6 from Mouradov et al 69
2.9.7 Platform code "IlluminaGA_RNASeq" from TCGA 70
2.9.8 Platform code "AgilentG4502A_07_3" from TCGA 71
2.9.9 GSE8332 72
2.9.10 GSE24795 72
2.9.11 Combination of transcriptomics datasets 72
2.10 Multi-omics data integration strategy 76
2.10.1 Computation of protein/mRNA ratios 76
2.10.2 Integration of transcriptomics and proteomics data 76
2.10.3 mRNA-guided missing value imputation 78
2.10.4 Minimum-guided missing value imputation 79
2.10.5 Combination of the CRC65 and CPTAC datasets 80
2.11 Identification and characterisation of Full Proteome & Kinobeads Subtypes 81
2.11.1 Identification of subtypes 81
2.11.2 Discovery of differentially expressed gene groups 83
2.11.3 Functional annotation enrichment and clustering 83
2.12 Reanalysis of public Drug Sensitivity datasets 85
2.12.1 Pre-processing 85
2.12.2 GDSC 86
2.12.3 CCLE 86
2.12.4 CTRP 87
2.12.5 Cetuximab 88
2.12.6 Dose-response modelling and parameter extraction 88
2.13 Modelling of Drug Sensitivity 89
2.14 Cell viability assays 92

2.15 Western blots 95
2.16 Immunohistochemistry 96
2.17 Cox proportional hazards regression 98
2.18 Annotation information 99

Chapter 3 Transcriptome and full proteome of the CRC65 cell line panel and the CPTAC patient dataset 101
3.1 Introduction 101
3.2 Transcriptome 102
3.2.1 Combination of multiple mRNA datasets 102
3.2.2 Prediction of Consensus Molecular Subtypes (CMS) 110
3.3 Full proteome 113
3.3.1 Optimization of hSAX fractionation 113
3.3.2 Protein identifications in the CRC65 and CPTAC datasets 115
3.4 Discussion 122

Chapter 4 Kinome of the CRC65 cell line panel 127
4.1 Introduction 127
4.2 Kinobeads in Micro-spin columns 129
4.3 Buffer comparison: CP versus RIPA100 134
4.4 Optimisation of the pulldown volume 137
4.5 Kinase identifications in the CRC65 dataset 139
4.6 Discussion 146

Chapter 5 Multi-omics data integration and validation 149
5.1 Introduction 149
5.2 Integration of full proteome and transcriptome data 151
5.2.1 Protein/mRNA ratios 151
5.2.2 Protein/mRNA correlation and mRNA-guided missing value imputation 153
5.2.3 Minimum-guided missing value imputation 158
5.3 Full Proteome Subtypes (FPSs) of CRC 164
5.4 Kinobeads Subtypes (KSs) of CRC 171
5.5 Modelling drug sensitivity as a function of full proteome profile 173

5.6 Modelling drug sensitivity as a function of Kinase profile....177
5.7 MERTK is a predictive biomarker in CRC....182
5.8 MERTK is a prognostic biomarker in CRC....185
5.9 Discussion....187

Chapter 6 General discussion 195

References 203

Appendix 255
6.1 Supplemental tables....255

Abbreviations

2D-PAGE	Two-dimensional polyacrylamide gel electrophoresis
5-FU	Fluorouracil
ACN	Acetonitrile
AcOH	Acetic acid
AGC	Automatic gain control
AIC	Akaike's Information Criterion
ATCC	American Type Culture Collection
ATP	Adenosine triphosphate
AUC	Area under the dose-response curve
Bis-Tris	Bis(2-hydroxyethyl)amino-tris(hydroxymethyl) methane
BMI	Body mass index
C18	octadecyl carbon chain
ca	ComBat-adjusted
CCLE	Cancer Cell Line Encyclopedia
cDNA	Complementary DNA
CID	Collision-induced dissociation
CIN	Chromosomal instability
CMS	Consensus molecular subtype
COAD/READ	Colorectal adenocarcinoma/Rectal adenocarcinoma
COXPH	Cox proportional hazards regression
CP buffer	Compound pulldown buffer
CPTAC	Clinical Proteomics Tumor Analysis Consortium
CRC	Colorectal cancer
CTRP	Cancer Therapeutics Response Portal
CV	Coefficient of variation
DAB	3,3'-diaminobenzidine tetrahydrochloride hydrate
DDA	Data-dependent acquisition
DFS	Disease-free survival
DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DPYD	Dihydropyrimidine dehydrogenase
DSMZ	Deutsche Sammlung von Mikroorganismen und Zellkulturen
DTT	Dithiothreitol
Dukes	Dukes' staging system specific to bowel cancer (A-D)
ECACC	European Collection of Authenticated Cell Cultures
ED50	Effective dose resulting in a half-maximal response
EDTA	Ethylenediaminetetraacetic acid
EGTM	European Group on Tumour Markers
EtOH	Ethanol
FA	Formic acid
FAP	Familial adenomatous polyposis
FAB	Fast atom bombardment
FBS	Fetal bovine serum
FDA	US Food and Drug Administration
FDR	False discovery rate
FFPE	Formalin-fixed, paraffin-embedded
FOV	Field of view
FPS	Full Proteome Subtype
FTMS	Fourier transformation mass spectrometry
FTP	File transfer protocol
GA	Genome Analyzer
GC	Gas chromatography
GC-MS	Gas chromatography coupled to mass spectrometry
GDSC	Genomics of Drug Sensitivity in Cancer
GEO	Gene expression omnibus
giBAQ	Gene-centric intensity-based absolute quantification
GO	Gene ontology
HCD	Higher energy C-trap dissociation
HGNC	HUGO Gene Nomenclature Committee

HILO	Kinobeads compound order 19,15, 1, 7, 5
HNPCC	Hereditary nonpolyposis colorectal cancer or Lynch Syndrome
HPLC	High performance liquid chromatography
HRP	Horseradish peroxidase
hSAX	Hydrophilic strong anion exchange
HSRRB	Health Science Research Resource Bank
i.d.	Inner diameter
iBAQ	Intensity-based absolute quantification
IC50	Inhibitory dose resulting in a half-maximal inhibition
IHC	Immunohistochemistry
ITMS	Ion trap mass spectrometry
KBγ	Kinobeads gamma
KS	Kinobeads Subtype
LC	Liquid chromatography
LC-MS/MS	Liquid chromatography tandem mass spectrometry
LDS	Lithium dodecyl sulfate
LFQ	Label-free quantification
lm	Linear model
log2/log10	Logarithm to the base of 2/10
LOHI	Kinobeads compound order 5, 7, 1, 15, 19
m/z	Mass-to-charge ratio
MA-plot	Log-ratio versus mean average plot
MAD	Median absolute deviation
MALDI	Matrix-assisted laser desorption/ionization
mAU	Milli arbitrary units
mc	Median-centred
mca	MComBat-adjusted
MeOH	Methanol
MG	Mapping group
MGH	Massachusetts General Hospital
MinDet	Deterministic Minimum Imputation
MinProb	Probabilistic Minimum Imputation
MMR	DNA Mismatch repair
MNAR	Missing not at random
MOPS	3-(N-morpholino) propanesulfonic acid
mRNA	Messenger ribonucleic acid
mRNA-Seq	mRNA Sequencing
MS	Mass spectrometry or mass spectrometer
MS/MS	Tandem mass spectrometry
MS1 spectrum	Precursor mass spectrum
MS2 spectrum	Fragment mass spectrum
MSI	Microsatellite instability
MSI-H	Microsatellite instability high
MSI-L	Microsatellite instability low
MSS	Microsatellite stability
NCBI	National Center for Biotechnology Information
normexp	Normal-exponential
OLS	Ordinary least squares
OS	Overall survival
PBS	Phosphate buffer saline
PD	Pulldown
PD/PD	Pulldown of pulldown
Pen-Strep	Penicillin/Streptomycin mixture
perseus	Perseus-type Imputation
PFA	Paraformaldehyde
PMF	Peptide-mass fingerprinting
ppm	Parts per million
PSM	Peptide spectrum match
PVDF	Polyvinylidene fluoride
qn	Quantile-normalised
R^2	Coefficient of determination
Rep	Replicate
RF	Radio frequency
RFS	Recurrence-free survival
RIPA100 buffer	Radioimmunoprecipitation assay buffer containing 100 mM NaCl
RMA	Robust multi-array average
RP	Reversed-phase
RPKM	Reads per kilobase of exon model per million mapped reads
RT	Room temperature
SAM	Significance analysis of microarrays
SDS	Sodium dodecyl sulfate

SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
Stage	Number staging system of tumours (I-IV)
TBS	Tris-buffered saline
TBST	Tris-buffered saline with Tween 20
TCGA	The Cancer Genome Atlas
TCL	Total cell lysate
TEAB	Triethylammonium bicarbonate
TFA	Trifluoroacetic acid
TMA	Tissue microarray
TNM staging	Staging based on tumour size and spread to nearby tissue (T; 1-4), number of involved lymph nodes (N; 0-3) and metastasis (M; 0-1)
TOP2A	DNA topoisomerase 2-alpha
TOP2B	DNA topoisomerase 2-beta
Tris	Tris(hydroxymethyl) aminomethane
ts	Total-sum normalised
UniProtKB	UniProt Knowledgebase
WB	Western blot
XIC	Extracted ion chromatogram (also extracted ion current)

Figures

Figure 1.4.1 Multi-omics data integration strategy 41
Figure 3.2.1 Exploration of normalisation options for GSE59857 103
Figure 3.2.2 MA-plots of tumour 8D after background correction using four different methods 105
Figure 3.2.3 Maxcomet distributions of the different mRNA datasets 106
Figure 3.2.4 Bimodal distribution of transcript abundance after dataset aggregation 107
Figure 3.2.5 Sample mix-ups in the TCGA mRNA datasets of tumours 109
Figure 3.2.6 Bimodal distribution of the sum of mRNA abundance for each gene across all samples 110
Figure 3.2.7 Overview of the datasets used for multi-omics data integration 112
Figure 3.3.1 Optimisation of the hSAX fractionation scheme 114
Figure 3.3.2 Retention time recalibration – merged or separate processing 117
Figure 3.3.3 Correlation of MS1-based quantification and spectral counting 118
Figure 3.3.4 Full proteome identifications in the CRC65 dataset 119
Figure 3.3.5 Full proteome identifications in the CPTAC dataset 120
Figure 3.3.6 Venn diagrams of full proteome identifications 121
Figure 4.2.1 Experimental setup for Kinobeads pulldowns in microcolumns 131
Figure 4.2.2 Scatterplots of LFQ Intensities for kinases enriched in microcolumns or batch-mode 132
Figure 4.2.3 Venn diagrams of kinases identified in enrichments using microcolumns or batch-mode 133
Figure 4.3.1 Scatterplots of LFQ Intensities for kinases from cells lysed with CP or RIPA100 buffer 135
Figure 4.3.2 Venn diagrams of kinases identified in enrichments from cells lysed with CP or RIPA100 buffer 136
Figure 4.4.1 Scatterplots of LFQ Intensities for kinases enriched from varying volumes of RIPA100 buffer 138
Figure 4.4.2 Venn diagrams of kinases identified in enrichments from varying volumes of RIPA100 buffer 139
Figure 4.5.1 Batch effect in the kinome profiles of the CRC65 cell line panel 140

Figure 4.5.2 Batch effect in the kinome profiles of the CRC65 cell line panel removed 141
Figure 4.5.3 Example of the systematic fragment mass deviation observed in the Kinobeads dataset 142
Figure 4.5.4 Kinome identifications in the CRC65 and CPTAC datasets 143
Figure 4.5.5 Venn diagram of kinome identifications 144
Figure 4.5.6 Western blots of EPHA4 and ABL1 across the CRC65 cell line panel 145
Figure 4.5.7 Correlation of LFQ and densitometry and reproducibility of biological triplicates 146
Figure 5.2.1 Stable Protein/mRNA ratios across the CRC65 and CPTAC datasets 152
Figure 5.2.2 Correlation of median mRNA Expression and Protein Expression 154
Figure 5.2.3 Correlation heat maps of median mRNA Expression versus Protein Expression 157
Figure 5.2.4 Relationship between giBAQ intensity and the number of missing values 159
Figure 5.2.5 Evaluation of imputation methods informed by the detection limit 162
Figure 5.3.1 Integrated Full Proteome Subtypes of CRC 165
Figure 5.3.2 Process networks and pathway maps enriched in proteins differentially expressed between Full Proteome Subtypes 168
Figure 5.3.3 GO-terms enriched in proteins differentially expressed between Full Proteome Subtypes 169
Figure 5.4.1 Kinobeads Subtypes of CRC 172
Figure 5.5.1 Prediction of drug sensitivity for cell lines and patients 175
Figure 5.6.1 Comparison of drug sensitivity predictions based on kinome/full proteome profiles 178
Figure 5.6.2 Confirmation that high EPHA2 is expression is associated with resistance towards cetuximab 179
Figure 5.6.3 MAP2K1 is a predictive marker for resistance towards EGFR inhibitors in cell lines 180

Figure 5.6.4 MERTK is a predictive marker for resistance towards MEK1/2 inhibitors in cell lines 181
Figure 5.6.5 MERTK is recurrently associated with drug sensitivity 182
Figure 5.7.1 Co-treatment with UNC569 (MERTK inhibitor) and MEK inhibitors 184
Figure 5.8.1 Western blot and immunohistochemistry of MERTK in cell lines and tumours 185
Figure 5.8.2 Survival analysis of the QUASAR2 patient cohort, stratified by MERTK expression 187

Tables

Table 2.2.1 Cell lines used in this study including references/commercial sources 44
Table 6.1.1 Associations of Full Proteome Subtypes with other clinical and molecular data 255
Table 6.1.2 Associations of Kinobeads Subtypes with other subtypes and molecular data 258
Table 6.1.3 Kinases differentially expressed between Kinobeads subtypes 259
Table 6.1.4 Target space enrichment 261
Table 6.1.5 Univariate Cox proportional hazards regression 271
Table 6.1.6 Multivariate Cox proportional hazards regression 273
Table 6.1.7 Associations of high membranous MERTK expression with other clinical and molecular data 274
Table 6.1.8 Annotation of drug sensitivity datasets 275
Table 6.1.9 Annotation of the CRC65 and CPTAC datasets 292

Abstract

Colorectal cancer (CRC) is one of the top three most common cancers and among the top four causes of cancer-related deaths worldwide (Torre et al., 2015). CRC patients are well characterised on the transcriptome and proteome level, but proteomics data on representative cell lines as model systems for pre-clinical drug sensitivity studies lag behind. Here, label-free quantitative mass spectrometry was used to characterise the kinomes and full proteomes of 65 CRC cell lines, collectively termed the CRC65 cell line panel. This data was integrated with proteomics data on patient samples, as well as public transcriptome and drug sensitivity datasets, which were reanalysed from raw data in order to unify and streamline the data processing. Protein/mRNA ratios were constant across these datasets, enabling linear prediction of protein abundance from mRNA abundance after appropriate adjustment, which was used for mRNA-guided missing value imputation. An exploration of secondary imputation methods prompted the development of a complementary method for minimum-guided missing value imputation. Combining the proteomics datasets on cell lines and patients led to the discovery of integrated proteomic subtypes of CRC and enabled the identification of representative cell lines for each of them. Modelling publicly available dose-response data generated by four large-scale drug sensitivity studies as a function of kinome/full proteome profiles fuelled the prediction of drug sensitivity for cell lines and patients, allowed the identification of drugs differentially effective between the different integrated proteomic subtypes and revealed MERTK as a predictive biomarker for resistance towards MEK1/2 inhibitors. This predictive role of MERTK was subsequently confirmed using *in vitro* experiments, while immunohistochemistry of TMAs from 1,074 tumours generated as part of the QUASAR2 clinical trial unveiled that MERTK is also a prognostic biomarker in CRC. This dataset will be made available to the scientific community to facilitate the design of prospective clinical studies.

Chapter 1
Introduction

1.1 Colorectal cancer

1.1.1 Epidemiology

Globally, cancer is among the top ten causes of premature death in high-income countries (Mortality and Causes of Death, 2015) and encompasses a very heterogeneous group of diseases defined as neoplasms by the ICD-10-CM codes C00-D49 (Bielby, 2011). Neoplasms are vastly different from each other not only with respect to their location (e.g. lung, breast, gut), but also considering their cellular and molecular characteristics (Futreal et al., 2004; Kandoth et al., 2013; Vogelstein et al., 2013). Worldwide, colorectal cancer (CRC) is the second and third most commonly diagnosed cancer and the third and fourth most common cause of cancer-related death in females and males, respectively (Torre et al., 2015). In high-income countries, CRC was even found to be the second most common cause of cancer-related premature death after lung cancer (Mortality and Causes of Death, 2015).

1.1.2 Prognosis

In general, the prognosis of CRC patients has improved over the course of the past decades (Holleczek et al., 2015; Sankaranarayanan et al., 2010; Siegel et al., 2014), with the cumulative lifetime risk to develop CRC before the age of 75 estimated to be between 0.4% and 5.2%, depending on sex and country (Ferlay et al., 2015). However, genetic predispositions like relatively rare mutations in APC – giving rise to familial adenomatous polyposis or FAP (Ichii et al., 1993; Kikuchi-Yanoshita et al., 1992; Laken et al., 1997; Nagase et al., 1992) – increase

the risk for CRC (Galiatsatos and Foulkes, 2006). The same is true for more frequently occurring mutations in genes involved in the DNA mismatch repair pathway (MMR) such as MLH1, MSH2, MSH6 and PMS2, which give rise to hereditary nonpolyposis colorectal cancer/HNPCC or Lynch Syndrome, characterised by microsatellite instability or MSI (Aaltonen et al., 1993; Leach et al., 1993; Peltomaki et al., 1993). These genetic predispositions – together with other forms of familial aggregation – account for about 20% of all CRC cases (Ahsan et al., 1998; Bonelli et al., 1988; Hemminki and Chen, 2004; Hemminki and Eng, 2004).

1.1.3 Staging

Upon presentation, CRC is commonly staged according to the TNM system (Compton and Greene, 2004; Edge, 2010; Gunderson et al., 2010a; Gunderson et al., 2010b; Jessup et al., 2011), which takes the presence, size and extend of the primary tumour (TX, Tis, T0-3, T4a-b), the involvement and number of regional lymph nodes (NX, N0, N1a-c, N2a-c) and the presence and number of distant metastases into consideration (M0, M1a-b). These classifications are subsequently summarised in stage groups (0, I, IIA-C, IIIA-C, IVA-B). Older, CRC-specific staging systems like Dukes' classification exist as well, however they are not used as regularly anymore (Astler and Coller, 1954; Dukes and Bussey, 1958; Haslam et al., 2006; Jass and Morson, 1987; Kirklin et al., 1949).

1.1.4 Screening

Survival in CRC – a cancer developing over the course of several years (Brenner et al., 2007; Kuntz et al., 2011; Stryker et al., 1987) – is highly dependent on the

stage of the tumour at diagnosis (Siegel et al., 2014). Since the cost for chemotherapy of advanced disease is high (Lansdorp-Vogelaar et al., 2009) and because adenomatous polyps (polypectomy) are removed easily during colonoscopy (Winawer et al., 1993; Zauber et al., 2012), population-based screening programmes are cost-effective (Habbema et al., 2014; Huang et al., 2014; Lansdorp-Vogelaar et al., 2010; Lansdorp-Vogelaar et al., 2009; Lee and Park, 2016; Patel and Kilgore, 2015; Shaukat et al., 2013; van Hees et al., 2015; Wong et al., 2015). The utility of such screening programmes employing the guaiac-based (gFOBT) or immunochemical detection of faecal occult blood (FIT), as well as flexible sigmoidoscopy and colonoscopy was proven by randomised controlled trials (Brenner et al., 2014; Hewitson et al., 2008; Shaukat et al., 2013), leading to their incorporation into international guidelines (Canadian Task Force on Preventive Health, 2001; Group et al., 2013; Levin et al., 2008; Sung et al., 2015). In 2016, the US Food and Drug Administration or FDA approved Epi proColon®, a blood-based test for methylated SEPT9 as a biomarker for CRC (Church et al., 2014; deVos et al., 2009; Grutzmann et al., 2008; Lofton-Day et al., 2008; Model et al., 2007), adding a novel screening procedure to the clinical repertoire.

1.1.5 Management

However, even though various screening techniques (Kahi et al., 2013), the preventive effect of anti-inflammatory agents (Wang and DuBois, 2013) and increased awareness of avoidable risk factors (Chan and Giovannucci, 2010) like consumption of processed or red meat (Bouvard et al., 2015) are able to reduce incidence and mortality of CRC, still more than half of all patients presented to

their physician with advanced (i.e. stage III/IV) disease during a recent study in the UK (Lyratzopoulos et al., 2013). As mentioned above, treatment of CRC is highly dependent on the stage of the cancer: polypectomy (Jass, 1995; Morson et al., 1984; Wolff and Shinya, 1975) and colectomy (surgical resection of the colon, Belderbos et al., 2014; Cooper et al., 1995; Cranley et al., 1986; Haggitt et al., 1985) or laparoscopic surgery (Ota et al., 1994) with removal of regional lymph nodes (summarised under the term "surgical resection") are the methods of choice for stage I and low-risk stage II CRC, while high-risk stage II patients, as well as patients with involvement of regional lymph nodes (stage III) should undergo surgical resection in combination with adjuvant chemotherapy according to the current NCCN guidelines (Benson et al., 2004; Compton et al., 2000; Des Guetz et al., 2010; National Comprehensive Cancer Network, 2016). Patients suffering from stage IV disease with distant metastases on the other hand should be treated with chemotherapy and/or surgical resection (Bartlett et al., 2006; Goldberg, 2006; Goldberg et al., 2007; Kelly and Goldberg, 2005; Van Cutsem, 2006). Chemotherapeutics recommended by the NCCN for non-metastatic (high-risk stage II and stage III) CRC in the adjuvant setting are 5-FU and its pro-drug capecitabine with or without leucovorin modulation (Machover et al., 1982; Poon et al., 1989), oxaliplatin and combinations thereof (Andre et al., 2009; Andre et al., 1999; Haller et al., 2005; Haller et al., 2011; Kuebler et al., 2007; Labianca et al., 1995; Schmoll et al., 2007; Twelves et al., 2005), while the same drugs plus more targeted (example targets in parentheses) agents like bevacizumab (VEGF), cetuximab (EGFR), panitumumab (EGFR), irinotecan (TOP1), ziv-aflibercept (VEGF), ramucirumab (VEGF), regorafenib (BRAF) and trifluridine-tipiracil (TYMP) are recommended for the treatment of metastatic

(stage IV) disease, the response-rates of which are highly variable (Cunningham et al., 2004; Douillard et al., 2000; Goldberg et al., 2004; Grothey et al., 2013; Mayer et al., 2015; Peeters et al., 2010; Tabernero et al., 2015; Tol et al., 2008; Van Cutsem et al., 2009; Van Cutsem et al., 2012).

1.1.6 Differential drug sensitivity and predictive biomarkers

Variable response rates to chemotherapy of CRC prompted studies investigating the cause of this differential drug sensitivity. For 5-FU, proteins involved in its metabolism as well as thymidylate synthase or TYMS – a target of fluorodeoxyuridine monophosphate, which is an active metabolite of 5-FU – were considered as potentially predictive biomarkers (Jensen et al., 2012). Of those, TYMS was investigated in greatest depth and high expression of TYMS was found to be predictive for resistance to 5-FU, while gene polymorphisms of TYMS were shown to influence 5-FU toxicity (Aschele et al., 2002; Cascinu et al., 1999; Davies et al., 1999; Ichikawa et al., 2003; Lecomte et al., 2004; Leichman et al., 1995). However, studies investigating this relationship often lacked an appropriate control group not treated with 5-FU and were underpowered (Duffy et al., 2007; Jensen et al., 2012), which is one of the reasons why the European Group on Tumour Markers (EGTM) does not recommend its use as a predictive biomarker (Duffy et al., 2014; Duffy et al., 2007). Similarly, the predictive role of p53 in the context of 5-FU chemotherapy is still debated (Brett et al., 1996; Cascinu et al., 2000; Elsaleh et al., 2001; Liang et al., 2002a; Paradiso et al., 2002; Tominaga et al., 2010), hence it is also not included on the list of recommended biomarkers, alongside SMAD4 (Boulay et al., 2002) and COX-2 (Uchida et al., 2005). The MSI status of a tumour was also proposed as a predictive biomarker for response to

5-FU (Arnold et al., 2003; Carethers et al., 1999; Carethers et al., 2004; Iwaizumi et al., 2011; Jover et al., 2009; Jover et al., 2006; Muller et al., 2008; Ribic et al., 2003; Sargent et al., 2010), but because several studies showed conflicting evidence (Bendardaf et al., 2007; Elsaleh et al., 2001; Liang et al., 2002b; Watanabe et al., 2001), it is also not recommended as a predictive biomarker (Duffy et al., 2014; Duffy et al., 2007). It is however recommended as a pre-screening method for patients with Lynch Syndrome, as MSI is one of its major characteristics (see above, Duffy et al., 2014; Duffy et al., 2007; National Comprehensive Cancer Network, 2016). Other notable candidates of predictive biomarkers are TOP1 (primary target) and UGT1A1 (irinotecan-inactivating enzyme) for response towards irinotecan (Braun et al., 2008; Cecchin et al., 2009; Toffoli et al., 2006), as well as ERCC1 (part of the nucleotide excision repair or NER machinery) for response towards oxaliplatin (Arnould et al., 2003; Kim et al., 2009; Shirota et al., 2001; Yin et al., 2011). However, they are also not yet recommended in clinical practice due to insufficient data (Duffy et al., 2014; Duffy et al., 2007). Currently, the only predictive biomarker routinely used in clinical practice is the mutation status of KRAS, which predicts response towards EGFR-targeted chemotherapy, with KRAS wild type patients benefitting more from treatment with drugs like cetuximab or panitumumab than patients with various KRAS mutations (Amado et al., 2008; De Roock et al., 2010; Duffy et al., 2014; Karapetis et al., 2008; Khambata-Ford et al., 2007; Lievre et al., 2006; Loupakis et al., 2009; Malapelle et al., 2012; National Comprehensive Cancer Network, 2016; Peeters et al., 2013; Soeda et al., 2013). Other proteins in the EGFR pathway like AREG and EREG ligands (Jacobs et al., 2009; Khambata-Ford et al., 2007), EGFR itself (Cappuzzo et al., 2008; Li et al., 2011; Licitra et al., 2013; Lievre et al., 2006;

Lin et al., 2011; Personeni et al., 2008), NRAS (De Roock et al., 2010), BRAF (De Roock et al., 2010; Lin et al., 2011; Loupakis et al., 2009; Soeda et al., 2013), PIK3CA (De Roock et al., 2010; Soeda et al., 2013) and PTEN (Frattini et al., 2007; Lin et al., 2011) were also considered to be potential biomarkers, however more studies are needed to increase confidence in them.

1.1.7 Advances in cancer genomics

The search for predictive biomarkers like the ones described above was accelerated in the past two decades by advances in genomics in general and transcriptomics in particular. With methods such as spotted (DeRisi et al., 1996; Schena et al., 1995) or *in situ* synthesised microarrays (Fodor et al., 1991; Pease et al., 1994) as well as mRNA sequencing (mRNA-Seq, Mortazavi et al., 2008), the genome-wide quantification of mRNA expression became routine and saw widespread application, also in cancer research.

1.1.7.1 Microarrays

In the case of microarrays, a typical gene expression experiment starts with the extraction of RNA from biological samples, followed by the enrichment of mRNA by exploiting the interaction of its polyadenylated tails with resin-bound oligo(dT) (Aviv and Leder, 1972; Chirgwin et al., 1979). Afterwards, mRNA is reverse transcribed to form complementary DNA or cDNA, incorporating special deoxynucleotides suitable for subsequent fluorescent labelling (Xiang et al., 2002). Labelled mRNA is then hybridised to a microarray, which consists of a collection of oligonucleotides or probes complementary to cDNA sequences targeted for quantification. These probes are arranged in a pre-defined grid or

are otherwise spatially encoded (Gunderson et al., 2004). Because each cDNA species will preferentially bind to its corresponding complementary probe, the amount of fluorescence from specific regions on the microarray can be recorded with a laser scanner and is afterwards related to specific cDNA species and their amounts in the original sample.

1.1.7.2 mRNA-Seq

In the case of mRNA-Seq, a typical gene expression experiment also starts with the isolation of mRNA and its reverse transcription to cDNA, leaving out the labelling step. Afterwards, next generation sequencing (mostly on Illumina instruments; see Goodwin et al., 2016 for review) is used to identify and quantify cDNAs. Briefly, cDNA is fragmented, followed by adapter ligation, binding to primers attached to a solid support (flow cell) and bridge amplification via PCR, resulting in cDNA clusters of the same sequence. These clusters of amplified cDNA are then "sequenced by synthesis" using a mixture of all four deoxynucleotides with blocked 3'-OH groups, resulting in the addition of only one deoxynucleotide to each cDNA fragment per cycle (Guo et al., 2008; Ju et al., 2006). Each of the four different deoxynucleotides is labelled with a specific fluorophore, the fluorescence of which is recorded at the end of each cycle, before both the blocking group and the fluorophore are removed and a new cycle begins. Here, cluster formation prior to sequencing ensures signal amplification. This process results in a collection of short sequences or "reads", which are then mapped to a reference genome of the organism under investigation, essentially assigning these reads to exon models they could have originated from. Transcripts are then quantified using RPKM (Mortazavi et al., 2008) values,

which are the number of reads per exon model per million mapped reads, thereby accounting for length differences of exon models. While other quantification methods like for example "RNA-Seq by Expectation Maximization" or RSEM exist (Li and Dewey, 2011), this thesis only uses data expressed as RPKM.

1.1.7.3 Tumour heterogeneity and gene signatures

Profiting from these breakthroughs in genomics, the field of oncology nowadays embraces a systems-biology view on "oncogene addiction" as the "Achilles heel" of cancer – originally proposed by Bernard Weinstein at the beginning of the 21st century (Weinstein, 2000, 2002) – and appreciates that the emergence of secondary resistance to drugs targeting oncogenes, as well as the pathway context most oncogenes are embedded in play crucial roles when trying to predict treatment success (Pagliarini et al., 2015; Vivanco, 2014; Weinstein and Joe, 2008). Therefore, the traditional model of CRC tumourigenesis (Fearon and Vogelstein, 1990; Kinzler and Vogelstein, 1996) made way for the concept of tumour heterogeneity, representing the first steps towards a system-level understanding of tumour biology (Blanco-Calvo et al., 2015; Linnekamp et al., 2015). In view of tumour heterogeneity, many prognostic and/or predictive gene expression signatures were developed with the idea in mind, that the quantification of transcription with respect to a selection of relevant genes will be more prognostic and/or predictive than measurements of single genes, enabling the stratification of patients into clinically relevant subgroups. Examples of commercial gene signatures are the Oncotype DX® breast cancer (Albain et al., 2010; Paik et al., 2004), MammaPrint® (Bueno-de-Mesquita et al., 2007; van 't Veer et al., 2002; van de Vijver et al., 2002) and Prosigna® (Nielsen et al., 2010;

Parker et al., 2009) assays for breast cancer, as well as the Oncotype DX® colon cancer (Gray et al., 2011; O'Connell et al., 2010; Srivastava et al., 2014), ColoPrint® (Maak et al., 2013; Salazar et al., 2011) and ColDX® (Kennedy et al., 2011) assays for CRC, of which the MammaPrint® and Prosigna® assays were also approved by the FDA. All of these gene signatures are mainly used to estimate patient prognosis, even though for example the Oncotype DX® colon cancer assay also includes the option to calculate a treatment score to predict patient benefit to 5-FU in combination with leucovorin. However, this predictive value of the gene signature could not be confirmed in an independent clinical study (Gray et al., 2011). Other gene signatures developed on a smaller number of patients or using previously published data were predictive of patient response to combinations of 5-FU and leucovorin with either irinotecan (Del Rio et al., 2007) or oxaliplatin (Watanabe et al., 2011b), as well as of success of cetuximab monotherapy in KRAS wild type patients (Balko and Black, 2009), but these signatures lack independent validation.

1.1.7.4 Molecular subtyping

One way to develop such a gene signature is to first classify patients into intrinsic molecular subtypes based on their mRNA expression profiles, since molecular subtypes might already be associated with prognosis. Particularly in the case of CRC, recent years have seen an increase in efforts to classify patients into different subtypes based on transcriptional profiling in order to find classifiers, which are of both prognostic and predictive value (Budinska et al., 2013; De Sousa et al., 2013; Marisa et al., 2013; Perez-Villamil et al., 2012; Roepman et al., 2014; Sadanandam et al., 2013; Schlicker et al., 2012; TCGA Network, 2012b).

These classifiers were used by Medico and colleagues (2015) to assign a panel of CRC cell lines to the different molecular subtypes, in order to find model systems amenable for drug sensitivity screens. However, such endeavours were complicated by inconsistencies among the different subtype models and the fact that the proteomics-based CRC subtypes published by the NCI's Clinical Proteomic Tumor Analysis Consortium (CPTAC, Zhang et al., 2014a) were similar to mRNA-based CRC subtypes, but also highlighted features only detectable in the proteome. In 2015, Guinney et al. compared the different mRNA-based subtypes in order to establish Consensus Molecular Subtypes of CRC (CMS, 2015), which showed partial concordance with the proteomic subtypes published by Zhang et al. and were also prognostic of patient survival. This resolved the aforementioned inconsistencies, however the apparent discrepancies between transcriptomic and proteomic subtypes – especially with respect to CRC cell lines, for which there is currently no comprehensive proteomic dataset – remain to date.

1.2 Associating drug response and omics data of cell lines

1.2.1 Pharmacogenomics

The foundation for many if not all of the above-mentioned gene signature studies was laid at the end of the 20th century by the laboratories of Eric S. Lander and John N. Weinstein (Golub et al., 1999; Scherf et al., 2000). The pioneering work of Todd R. Golub and colleagues demonstrated the utility of gene expression profiling by microarrays for the classification of acute myeloid leukaemia and acute lymphoblastic leukaemia patients. In addition to patients, cell lines were also used extensively in studies trying to identify biomarkers or gene signatures predicting response to chemotherapy. Cell lines have the big advantage that they

can be used to screen a plethora of drugs for interesting pre-clinical compounds. The Developmental Therapeutics Program for example screened thousands of drugs against the famous NCI-60 cell line panel (Monks et al., 1991; Weinstein et al., 1997) and subsequently showed that correlating steady-state mRNA expression profiles to drug sensitivity profiles can be used to identify genomic markers of drug response (Scherf et al., 2000). While Scherf and colleagues were able to identify low DPYD expression as a marker for 5-FU sensitivity, a later study compiled mRNA expression and drug sensitivity data into an integrated chemosensitivity database and identified high expression of TOP2A and TOP2B as a marker for etoposide sensitivity (Dan et al., 2002). Not long after their breakthrough publication in 1999, Golubs laboratory demonstrated that steady-state mRNA expression profiles of the NCI-60 cell line panel could also be used to predict drug sensitivity of these cell lines towards a variety of compounds (Staunton et al., 2001). The concept that correlating steady-state mRNA expression to drug sensitivity facilitates the identification of predictive biomarkers and also enables the prediction of drug response forms the basis of modern large-scale pharmacogenomics and -proteomics. Over the course of the following years after the seminal publication by Staunton and colleagues, many more studies investigated the predictive potential of genomics profiles from cell lines. Some studies used cell lines from multiple entities to identify prognostic and predictive genomic markers (Dry et al., 2010; Garraway et al., 2005; Greshock et al., 2010; McDermott et al., 2007; Solit et al., 2006), while others focused on a single entity (Lin et al., 2008; Neve et al., 2006; Sos et al., 2009). CRC cell lines were also used to identify genomic markers of drug sensitivity and predictive gene signatures, which were for example associated with response towards 5-FU

(Boyer et al., 2006; Bracht et al., 2010; Mariadason et al., 2003; Zheng et al., 2015), oxaliplatin (Arango et al., 2004; Boyer et al., 2006; Jensen et al., 2015; Zheng et al., 2015), irinotecan (Allen et al., 2008; Allen et al., 2012; Jensen et al., 2015; Zheng et al., 2015), camptothecin (Mariadason et al., 2003), cetuximab (Ashraf et al., 2012; Jhawer et al., 2008; Medico et al., 2015), trastuzumab (Ashraf et al., 2012), pertuzumab (Ashraf et al., 2012) and lapatinib (Ashraf et al., 2012). Particularly the recent study by Zheng and colleagues is notable, since they developed three gene signatures based on differentially expressed genes between HCT 116 cells made resistant to 5-FU, oxaliplatin or irinotecan and their corresponding parental cell line (Zheng et al., 2015). They analysed public microarray data (Allen et al., 2012; Boyer et al., 2006) and were able to stratify patients into groups with distinct survival outcomes. The discovery of these gene signatures was based on only a single CRC cell line, which the authors described as advantageous, since the genetic background of different cancer cell lines might influence drug resistance. However, in a clinical setting, patients will also harbour different genetic backgrounds and mutational landscapes (Kandoth et al., 2013; Sjoblom et al., 2006; Vogelstein et al., 2013; Wood et al., 2007), which is why the development of predictive gene signatures should ideally be based on a sizeable number of cell lines of a single entity, in order to be able to consider the different molecular subtypes and their influence on drug sensitivity (Wilding and Bodmer, 2014). More recently, high-throughput drug sensitivity screens increased in size, both in terms of the number of drugs analysed, as well as considering the number of cell lines screened. In 2012, two landmark studies in the field of pharmacogenomics from the Broad Institute (Cancer Cell Line Encyclopedia or CCLE, Barretina et al., 2012) and from the Sanger Institute in collaboration with

the Massachusetts General Hospital or MGH (Genomics of Drug Sensitivity in Cancer or GDSC, Garnett et al., 2012) were published back to back in Nature. Both studies combined high-throughput drug sensitivity screening with targeted sequencing, array-based estimation of copy number variations and microarray-based gene expression profiling of 947 and 639 cell lines, respectively, in order to find genomic correlates of drug sensitivity. While both studies successfully and reproducibly identified known genomic predictors of drug response like ERBB2 amplification/overexpression for sensitivity to lapatinib (ERBB1 and ERBB2 inhibitor, Konecny et al., 2006; Xia et al., 2002) and BRAF mutations for sensitivity to PLX4720 (RAF inhibitor, Tsai et al., 2008) and PD-0325901 (MEK1/2 inhibitor, Barrett et al., 2008; Solit et al., 2006), their suggestions for novel predictors and general methodology was called into question by a subsequent meta-analysis, which claimed that – while the genomics measurements of the two studies agreed reasonably well – their quantification of drug sensitivity was highly discordant (Haibe-Kains et al., 2013). Haibe-Kains et al. reported poor correlation between the respective drug sensitivity metrics used by each of the two studies: the IC50 or the drug concentration necessary to achieve half-maximal inhibition and the area under the activity curve measuring dose-response. They also consulted dose-response data from a third large-scale study by GlaxoSmithKline, the IC50 values of which were more consistent with the ones from the study by Barretina et al., presumably because a comparable viability readout was used (Greshock et al., 2010). The authors of the CCLE and GDSC studies subsequently published a similar comparison of their own, in which they highlighted the agreement between their 2012 publications with respect to which genetic biomarkers they identified for which drug and also considering the

correlation of their drug sensitivity measurements, provided that the highly discontinuous nature of IC50 distributions in both datasets was accounted for during the comparison (Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium, 2015). These findings were subsequently reinforced by data from a fourth large-scale drug-sensitivity screen by Haverty et al. (2016). The Genentech Cell Line Screening Initiative or gCSI independently profiled 16 drugs previously evaluated by the CCLE and GDSC studies across 410 cell lines and also found for example ERBB2 amplification/overexpression to be associated with drug sensitivity of lapatinib. Haverty et al. stressed that both the seeding density of the cell lines at the beginning of the assay, as well as the medium composition might lead to underestimation of drug sensitivity and called for the use of appropriate and consistent data generation and analysis protocols. Meanwhile, both the Broad Institute and the Sanger Institute in collaboration with the MGH updated their initial pharmacogenomics datasets from 2012, increasing the amount of publicly available drug sensitivity data on cancer cell lines even further (Basu et al., 2013; Iorio et al., 2016; Rees et al., 2016; Seashore-Ludlow et al., 2015).

1.2.2 Pharmacoproteomics

While numerous pharmacogenomics studies were published over the course of the past decade, equivalent studies using proteomics data were lagging behind. One of the first systematic pharmacoproteomic studies made use of two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) to create a protein expression database of the NCI-60 cell line panel (Myers et al., 1997; Weinstein et al., 1997). Proteins were separated by isoelectric focusing in the first

dimension and conventional SDS–polyacrylamide gel electrophoresis in the second dimension (Anderson and Anderson, 1978), before protein spots were identified either by matrix-assisted laser desorption ionisation mass spectrometry (MALDI-MS, Li et al., 1997) and peptide-mass fingerprinting (PMF) or by western blots (see also section 1.3.1.5). Identified protein spots could then be quantified with respect to position, shape and density using specialised software (Myers et al., 1997). However, this method was quickly replaced by protein microarrays at the beginning of the 21st century (Lueking et al., 1999; MacBeath et al., 1999; MacBeath and Schreiber, 2000), when Sreekumar et al. succeeded in identifying various proteins, which were up-regulated upon irradiation of a CRC cell line (LoVo, Sreekumar et al., 2001). They used forward-phase protein or antibody arrays, which are glass slides spotted with different antibodies. These arrays were then probed with fluorescently labelled cell lysates to detect proteins binding to these antibodies. Nishizuka et al. subsequently used a similar method termed reverse-phase protein arrays (Paweletz et al., 2001) to profile the NCI-60 proteome again and identified proteomic markers distinguishing colon and ovarian tumours (Nishizuka et al., 2003b), while being able to compare protein and mRNA expression patterns across the cell line panel (Nishizuka et al., 2003a; Shankavaram et al., 2007). Instead of spotting antibodies onto glass slides, they spotted lysates of all NCI-60 cell lines in a range of dilutions onto a single slide and probed each copy of this slide with a specific primary and secondary antibody, allowing simultaneous detection of one protein across all cell lines. This dataset was subsequently updated (Park et al., 2010) and served as the basis for the identification of several proteomic markers of drug response to paclitaxel, some of which could be validated using data from a whole-genome

siRNA synthetic lethality screen (Whitehurst et al., 2007). Innovations in the field of mass spectrometry recently enabled the characterisation of the NCI-60 cell line panel using label-free quantitative mass spectrometry, resulting in the identification of 10,350 proteins in total (Gholami et al., 2013), which was impossible to do at the time using protein microarrays of any kind. Moghaddas Gholami et al. succeeded in identifying proteins recurrently associated with drug response, which inspired a recent study investigating the proteome of triple-negative breast cancer (Lawrence et al., 2015); however, a comparable study for CRC does not exist as of yet. Moghaddas Gholami and colleagues associated their proteomics data with the drug sensitivity data generated by the Developmental Therapeutics Program using a regularised regression method termed elastic net regression (Zou and Hastie, 2005), which was applied before in pharmacogenomics studies (Barretina et al., 2012; Basu et al., 2013; Garnett et al., 2012; Iorio et al., 2016) and is particularly well suited for this type of problem.

1.2.3 Elastic net regression

One of the more modern regression methods used for the modelling of drug sensitivity as a function of various omics data is elastic net regression. Together with a bootstrapping approach, this regularised multivariate linear regression method was previously successfully applied to identify genomic, transcriptomic and proteomic markers of drug sensitivity and resistance (Barretina et al., 2012; Basu et al., 2013; Garnett et al., 2012; Gholami et al., 2013; Iorio et al., 2016). The advantages of elastic net regression compared to other regression methods are discussed in detail in the Supplementary Methods associated with the publication by Barretina et al. (2012), as well as in the original publication by Zou and Hastie

(2005). The algorithm is especially useful for "large p, small N" ($p>>N$) regression problems involving for example data generated using current omics technologies of various kind, which measure many, possibly highly correlated analytes (p) across a small number of samples (N). In settings like these with varying degrees of multicollinearity, the ordinary least squares (OLS) estimates of the regression coefficients and the resulting models frequently perform poorly both in terms of prediction accuracy on new data, as well as regarding the interpretability of the model itself (Zou and Hastie, 2005). The elastic net aims at improving both prediction performance and model interpretability by consolidating L2-regularised Ridge regression (Hoerl and Kennard, 1970) and L1-regularised Lasso regression (Tibshirani, 1996), in order to combine the advantages of the different penalty terms they use to constrain the coefficient vector. On its own, Ridge regression often has a higher predictive accuracy than OLS in the above-mentioned situations by striking a balance between bias and variance of predicted values via its bound on the L2-norm of the coefficient vector, thereby aiming at improving the overall prediction accuracy (Tibshirani, 1996). However, it always keeps all predictors in the model, which is not useful if the goal is to increase its interpretability compared to OLS. Lasso regression on the other hand does automatic variable selection via its bound on the L1-norm of the coefficient vector, thereby shrinking some coefficients while forcing others to be exactly zero, which results in a more parsimonious model consisting of a reduced subset of good predictors, which is easier to interpret (Tibshirani, 1996). However, this comes at the cost of only selecting at most N variables in settings where $p>>N$, as well as only selecting one variable from a group of correlated variables to be included in the model while excluding the others in settings with

multicollinearity (Zou and Hastie, 2005). The authors also observed that Ridge regression performs far better than Lasso regression with respect to predictions on new data in settings with multicollinearity where $N>p$. By combining these two penalty terms, the elastic net now promotes parsimony of models through the L1-penalty, while at the same time encouraging a grouping effect through the L2-penalty (Barretina et al., 2012), meaning that highly collinear predictors are usually either in the model or dropped from it together. Double shrinkage of coefficients is prevented by the introduction of a scaling factor (Zou and Hastie, 2005) and the higher level parameter α with $0 \leq \alpha \leq 1$ is used to control the balance between the L2-penalty (α=0) and the L1-penalty (α=1), while λ controls the degree of regularisation (Friedman et al., 2010). Since the elastic net is indifferent with respect to which predictors are consulted to fit the model, it can be used for both pharmacogenomics and pharmacoproteomics studies.

1.3 Mass-spectrometry-based bottom-up proteomics

The completion of the Human Genome Project at the beginning of the 21st century (Lander et al., 2001; Venter et al., 2001) heralded the onset of the 'genomic era' in modern medicine (Guttmacher and Collins, 2003). More than a decade later, we arrived in the 'post-genomic era' (awaited as early as 2001; see Brower, et al.) and the scientific community realised that having a blueprint for which proteins can potentially be expressed is often not enough to initiate drug development for a given disease. Therefore, fields like proteomics (James, 1997) are gaining momentum, the foundation of which was laid by genomics. Proteomics in analogy to genomics is the study of the proteome – defined as the protein complement of the genome (Wasinger et al., 1995; Wilkins et al., 1996a; Wilkins et al., 1996b) –

and its dynamic changes in response to biological cues. Mass spectrometry is coming of age and is now the dominant technology in the field of proteomics, culminating in the recent publication of the first drafts of the human proteome (Kim et al., 2014; Wilhelm et al., 2014).

Mass-spectrometry-based bottom-up or shotgun proteomics investigates the proteome by first digesting proteins with sequence-specific proteases like trypsin, which is an endoprotease cleaving proteins carboxy-terminal of arginine and lysine residues (Giansanti et al., 2016). The resulting peptides are then commonly separated according to their hydrophobicity by reversed-phase high-performance liquid-chromatography or RP-HPLC, a technique developed in the 1950s (Howard and Martin, 1950) and adapted to the analysis of peptides by introduction of ion-pairing agents like trifluoroacetic acid (TFA) or formic acid (FA) in the 1970s (Hancock et al., 1976; Hancock et al., 1978a; Hancock et al., 1978b; Mahoney and Hermodson, 1980). Coupled to a mass spectrometer via electrospray ionisation (see section 1.3.1.2), this analytical setup is the current workhorse of proteomics. This approach is contrasted by top-down proteomics, which employs a similar analytical setup for the analysis of intact proteins, but is used much less frequently due to various reasons discussed elsewhere (Steen and Mann, 2004). Briefly, sequence information necessary for protein identification is generated more easily using small peptides rather than large proteins, but identification and quantification of proteins in undigested mixtures are recently becoming more feasible (Durbin et al., 2016).

1.3.1 Instrumentation

In layman's terms, a mass spectrometer can be described as a scale, sensitive enough to "weigh" particles not visible to the naked eye (reviewed in Savaryn et al., 2016). However – since the mass of these particles is vanishingly small – the force of gravity exerted on them is not strong enough to be measured using scales in the conventional sense. Therefore, a mass spectrometer exploits the fact that electromagnetic and electrostatic fields influence charged analytes or ions in the gas-phase. They change their trajectories in response to changes in these fields, enabling the measurement of their mass-to-charge ratios (m/z) and – upon inspection of their isotope distributions – the inference of their neutral masses. This implies that neutral molecules first need to be charged and transferred to the gas phase before they can be separated according to their m/z and finally quantified or "counted" using a mass spectrometer. These tasks are accomplished by the core parts of each mass spectrometer: the ion source – in charge of ionisation and vaporisation of analytes – the mass analyser – responsible for resolving different ions by m/z – and the detector – taking care of ion quantification (Aebersold and Mann, 2003).

1.3.1.1 History

Mass spectrometry is based on findings of the late Sir Joseph John Thomson using cathode rays more than 100 years ago (Thomson, 1913). As an analytical technique, MS exists since the early 1900s (Aston, 1919a; Dempster, 1918) and was initially used to study isotope compositions of gases and metals (Aston, 1919b; Dempster, 1921). Mass-spectrometry-based analysis of peptides was developed about half a century later, resulting in the publication of the primary

structure of bacteriorhodopsin (Gerber et al., 1979; Khorana et al., 1979) and alanyl-tRNA synthetase (Putney et al., 1981) by the laboratory of Klaus Biemann, who is widely regarded as the father of mass-spectrometry-based proteomics. In the early 1960s, the Biemann laboratory popularised the coupling of gas chromatography or GC to MS for the analysis of peptides (Watson and Biemann, 1964) and developed early methods for the computer-aided acquisition (Hites and Biemann, 1967, 1968) and computerised interpretation (Hertz et al., 1971) of mass spectra. One development particularly important for this thesis is the "mass chromatogram" (Hites and Biemann, 1970) or the plot of the intensity with respect to a particular ion versus spectrum index across a GC-MS experiment.

1.3.1.2 Soft ionisation techniques

In many ways, MS analysis of peptides and proteins changed with the advent of soft ionisation techniques, which – in contrast to hard ionisation techniques like for example electron ionisation (Dempster, 1918; Smyth, 1922) – leave relatively large biomolecules like peptides and proteins intact during the process of ionisation. Examples of soft ionisation techniques are fast atom bombardment or FAB (Barber et al., 1981; Morris et al., 1981), MALDI (Karas et al., 1987; Karas et al., 1985; Tanaka et al., 1988) and electrospray ionisation or ESI (Bruins et al., 1987; Whitehouse et al., 1985; Yamashita and Fenn, 1984). Particularly the development of ESI resulted in a revolution with respect to MS-based analysis of peptides and proteins, since it enabled the direct coupling of high performance liquid chromatography or HPLC to MS in conjunction with ionisation at atmospheric pressure, giving rise to multiply charged ions (Covey et al., 1988;

Fenn et al., 1989). This thesis makes extensive use of mass chromatograms generated by coupling an HPLC to an MS via an ESI source.

An ESI source works by channelling analytes dissolved in chromatography solvent downstream from the HPLC column through a stainless steel needle (emitter), which is maintained at a few thousand volts relative to the orifice of the MS. At flow rates between 1 and 20 µL/min for regular ESI (Fenn et al., 1989) and about 20 nL/min for the more sensitive nanoESI configuration using smaller emitters (Wilm and Mann, 1996), the surface of the chromatography solvent becomes electrostatically charged as it leaves the needle. This causes the formation of a so-called Taylor cone if the applied electric potential exceeds the critical potential of the solvent, which is determined by its surface tension (Taylor, 1964; Wilson and Taylor, 2008; Zeleny, 1914). A Taylor cone is characterised by the projection of an axial liquid jet from the tip of the cone, which is further dispersed by Coulomb forces into a fine spray consisting of charged droplets containing the analytes. Propelled by the difference in electric potential, these droplets migrate towards the MS orifice and shrink due to evaporation of the chromatography solvent, which increases their surface charge. Once their Rayleigh limit (Rayleigh, 1882) is reached at which Coulomb repulsion is of the same order as droplet surface tension, droplet instability culminates in what is called a Coulomb explosion, which results in droplet fission and increased rates of evaporation. According to the ion evaporation model by Iribarne and Thomson (1976), this process repeats until the droplets are small enough for ions to be desorbed from the solvent into the gas phase by a strong electric field resulting from the high density of droplet surface charges. This model deviates from the charged residue model proposed earlier by Dole and colleagues (Dole et al., 1968;

Mack et al., 1970), who postulated that progressive droplet fission would eventually lead to droplets consisting of only a single solute molecule, which will ultimately evaporate and transfer some or all of its charges to the analytes, thereby ionising them. However, evidence from experiments with vapour of polar solvents in bath gas suggested that ionisation follows the ion evaporation model (Nguyen and Fenn, 2007).

1.3.1.3 Ion optics

Once ionised, charged analytes enter the MS through its orifice and are subsequently manipulated using electromagnetic fields, generated by a variety of different ion optics (see Savaryn et al., 2016). In general, ions are propelled axially through the MS by DC offsets between consecutive disc-shaped electrodes or lenses, which can focus ions analogous to optical lenses. During the analysis of cations, this means that as charged analytes migrate further into the MS, these electrodes are increasingly negatively charged, attracting positively charged analytes into the MS and towards the mass analyser. Meanwhile, they are confined radially with the help of electromagnetic fields generated by rod-shaped electrodes. These are arranged axially in juxtaposed pairs to form various "poles" like quadrupoles (two pairs) or octapoles (four pairs). Application of AC – alternating at 1.2 MHz[1] – with equal amplitude and opposite sign to pairs of juxtaposed electrodes results in a quadrupolar electromagnetic field, which induces secular (corkscrew-like) motion of ions and thereby confines their trajectories radially.

[1] It is worth mentioning that AC frequencies between 3 kHz and 300 GHz are also called radio frequencies (RF).

1.3.1.4 Mass analysers

Since this thesis made use of linear or 2D ion traps, C-traps and orbitraps, this section focuses on these three types of mass analysers. Like most modern mass analysers at the time of writing, the former two work by manipulating the trajectories of ions electromagnetically, while the orbitrap does so electrostatically (Makarov, 2000).

A quadrupole segmented axially into three parts and filled with a dampening gas like helium (Makarov et al., 2006; Olsen et al., 2009b) is also capable of confining ions axially by creating a potential well using DC offsets of the outer segments. Resonance excitation can then be used to isolate and fragment ions of particular m/z (see section 1.3.1.5) or propel particular ion species out of the exit rods of the quadrupole, which contain slits for ions to pass through (Savaryn et al., 2016).

C-traps on the other hand are special RF-only ion traps, which were specifically developed to deliver compact ion packets to orbitraps (Makarov et al., 2006). They are c-shaped quadrupoles, filled with a dampening gas like N_2 and capped by two flat lenses. Upon entering a C-trap through the gate lens, ions are "cooled" through collisions with N_2, bouncing back and forth between the trap lens and the DC offset in front of the C-trap, while being confined radially by the quadrupolar field. The energy of these collisions is fine-tuned so as to not result in fragmentation of the charged analytes. Applying 200 V DC to both the gate and trap lenses compresses ions axially before they are ejected laterally through a slit in one of the rod-shaped electrodes towards an orbitrap. This is done by rapidly ramping down the RF voltage, followed by application of DC pulses to the electrodes of the C-trap, which create appropriate DC offsets.

Orbitraps are based on the Kingdon trap (Kingdon, 1923) and consist of two barrel-like outer electrodes facing towards each other with their inner volumes, which are held together by a central spindle electrode, placed axially (Makarov, 2000). All electrodes are isolated from one another, so that application of a voltage between the outer and inner electrodes results in a radial electrostatic field, causing attraction of ions towards the spindle electrode. Coming from a C-trap, ions enter an orbitrap tangentially through a small aperture, which is offset relative to the equator of the orbitrap by 7.5 mm (Makarov et al., 2006). This tangential velocity generates a centrifugal force, which counteracts the attractive electrostatic field. As they enter the orbitrap, ions are captured by a rapid increase in this electrostatic field and are thereby "squeezed" towards the spindle electrode. The lower the m/z of a particular ion, the closer it is squeezed towards the centre of the orbitrap, since less force is needed to influence its trajectory. This results in the formation of thin, rotating rings of different diameters, each containing a separate ion species. These rings subsequently oscillate axially along the spindle electrode in response to centrifugal and electrostatic forces, without colliding with each other (Savaryn et al., 2016). The frequency of this oscillation is proportional to $(m/z)^{-1/2}$ and is captured for all ion species at once as an image current at the barrel-like outer electrodes of the orbitrap (Makarov, 2000). Fourier transformation can then be used to decompose this signal into the underlying oscillation frequencies and their corresponding amplitudes, which are subsequently used to calculate the m/z values and intensities of the ions in the orbitrap, respectively (Makarov et al., 2006).

Modern mass spectrometers are usually equipped with combinations of mass analysers. This thesis made use of two different mass spectrometers with a linear ion trap-orbitrap configuration (Michalski et al., 2012a; Olsen et al., 2009a; Olsen et al., 2009b; Pachl et al., 2013b).

1.3.1.5 Gas-phase fragmentation and tandem mass spectrometry

Until the end of the 1980s, peptide sequences were determined using automated Edman degradation (Edman, 1956; Edman and Begg, 1967). Soft ionisation together with innovations in gas-phase peptide biochemistry subsequently made the determination of peptide primary structures using Edman sequencing obsolete. The mass spectrometric instrumentation described up to this point is capable of measuring masses of intact ions. In the case of peptides ions, this information was initially used to identify proteins by PMF (Henzel et al., 1993; James et al., 1993; Mann et al., 1993; Pappin et al., 1993; Yates et al., 1993). The idea behind PMF is that the masses of a limited number (3-4) of tryptic peptides harbour enough discriminating power – without any sequence information – to distinguish even closely related proteins when compared to a database of *in silico* digested protein sequences. However, this discriminating power is conditional upon analysing only simple mixtures of proteins, such as the ones present in spots of a 2D-PAGE (Henzel et al., 2003). However, tryptic digests from complex protein mixtures routinely extracted from biological samples contain too many peptides with indistinguishable masses, making unambiguous protein identification based only on intact peptide masses unfeasible. This problem only becomes worse when considering post-translational modifications. In order to alleviate this problem of ambiguity related to the search space of theoretical peptide masses

as dictated by the protein database being queried, some amount of peptide sequence information was required in addition to intact peptide masses. One of the earliest approaches making use of peptide sequence information in this context were peptide sequence tags, which are short sequences of amino acids (2-3) derived from tandem mass spectra (Mann and Wilm, 1994). Nowadays, tandem MS or MS/MS is routinely used to generate peptide sequence information.

During an MS/MS experiment on a mass spectrometer of the linear ion trap-orbitrap type, the mass spectrometer cycles between two modes of operation, while the HPLC continuously delivers peptides to the ESI source. First, so called precursor mass spectra or MS1 spectra are acquired, which record the intact masses of peptides eluting from the HPLC column at any given time, as well as their intensities. Next, based on some pre-defined heuristic, the mass spectrometer isolates a peptide ion species of defined m/z by modulating the AC amplitude in the linear ion trap (Savaryn et al., 2016). Afterwards, this ion species is subjected to gas-phase fragmentation using one of the methods listed below, before the resulting peptide fragment ions are recorded in what is called a fragment mass spectrum or MS2 spectrum. Currently, popular fragmentation methods are collision-induced dissociation or CID (Hunt et al., 1981; Kondrat and Cooks, 1978; Yost and Enke, 1978), higher energy C-trap dissociation or HCD (Olsen et al., 2007), electron transfer dissociation or ETD (Chi et al., 2007; Frese et al., 2013a; Mikesh et al., 2006; Syka et al., 2004) and a combination of ETD and HCD termed EThcD (Frese et al., 2012; Frese et al., 2013b). In this thesis, both CID and HCD data were used, which is why these fragmentation methods are showcased below.

CID and HCD are conceptually similar to each other in that isolated peptide ion species are excited by some mechanism and acquire kinetic energy. This kinetic energy is subsequently converted to vibrational energy by multiple collisions with dampening gas molecules like He (linear ion trap) or N_2 (C-trap), ultimately causing peptide ions to break, forming characteristic fragment ions (Roepstorff and Fohlman, 1984). In addition to classical b-type and y-type ions, neutral losses of water or ammonia can occur, producing additional informative peaks in the MS2 spectrum (Michalski et al., 2012b). Internal fragment ions or "fragments of fragments" are much more common in HCD than they are in CID, which is the result of different excitation mechanisms used by these methods. While CID uses resonance excitation, HCD uses what is known as beam-type excitation. During beam-type excitation, ions are accelerated by DC offsets into a quadrupole filled with dampening gas (Xia et al., 2006), causing peptide ion fragments to retain kinetic energy and therefore the potential to fragment again. Resonance excitation on the other hand is dependent on the m/z value of the isolated peptide ion species. As soon as peptide ions break apart in CID, they are not subject to excitation anymore, making secondary fragmentation less likely. However, this also results in a downside of CID fragmentation in a linear ion trap: fragment ions below a certain m/z value cannot be confined radially anymore by the AC field of the quadrupole and are frequently lost, resulting in a low m/z cutoff of MS2 spectra (Schwartz et al., 2002).

In this thesis, data-depend acquisition or DDA was the heuristic used to decide which peptide precursor to isolate and fragment. This means that the mass spectrometer is instructed to isolate and fragment the top x most abundant peptides from a preceding MS1 spectrum one after the other. In order to prevent

sequencing the same peptide multiple times and to prioritise peptides of interest, dynamic exclusion and static inclusion lists are used, which contain m/z values and corresponding HPLC retention times employed by the mass spectrometer to fine-tune its acquisition. One drawback of DDA is its stochastic nature, meaning that slight variations in MS1 intensities will result in a different population of peptide precursors to be isolated and fragmented, effectively reducing the reproducibility of the measurement and introducing missing values in the final protein expression matrix.

1.3.2 MaxQuant

In this thesis, data generated using MaxQuant (Cox and Mann, 2008) together with its search engine Andromeda (Cox et al., 2011) were used to process LC-MS/MS raw data, which is why this workflow is described below.

1.3.2.1 Peptide and protein identification by database search

With MS1 and MS2 spectra for specific peptide ions at hand, the above-mentioned problem with respect to ambiguity of peptide identifications is reduced tremendously. However, a typical MS experiment results in the acquisition of 10,000s of MS2 spectra, manual interpretation of which is unfeasible. Therefore, a variety of algorithms for the automatic interpretation of MS2 spectra were developed. Some of these algorithms perform *de novo* sequencing (Ma et al., 2005), while others compare acquired MS2 spectra to previously identified ones in an approach similar to the one developed by the Biemann laboratory (Hertz et al., 1971; Yates et al., 1998). MaxQuant – among other algorithms – compares MS2 spectra to theoretical fragment ion spectra (Cox et al., 2011; Eng et al., 1994;

Pappin et al., 1993; Perkins et al., 1999; Tabb et al., 2007). First, an *in silico* proteolytic digest of a protein database (including common contaminants) is performed using appropriate parameters corresponding to the biological sample and question at hand. Afterwards, theoretical fragment ion spectra are generated from the resulting peptides, the collection of which form the target database (Nesvizhskii et al., 2007). The protease used for digestion, the number of allowed missed cleavages, expected post-translational modifications and other parameters specified during this process influence the size of this target database, since every possible peptide corresponding to these parameters is generated. In addition, the protein sequences in the original database are reversed and the above-mentioned process is repeated, in order to generate a decoy database of theoretical fragment ion spectra, necessary for estimating the false discovery rate or FDR of peptide spectrum matches or PSMs (Elias and Gygi, 2007). These two databases are concatenated and form the search space. After peak detection and following isotope and charge state deconvolution of MS1 data, an initial Andromeda search with a relatively high precursor ion tolerance (e.g. 20 ppm) and a user-specified fragment ion tolerance is carried out (mass accuracy restrictions). Andromeda only considers the most abundant peaks per 100 Th interval in each MS2 spectrum up to a user-specified number, in order to prioritise fragment ions of high abundance. For each MS2 spectrum and theoretical fragment ion spectrum with the correct precursor mass, it then counts the number of matches k between observed peaks in the MS2 spectrum and the n theoretical fragment masses, given the mass accuracy restrictions specified above. The Andromeda score for this comparison is then given by -10 times the logarithm of the binomial probability of matching at least k out of the n

theoretical fragment masses by chance (Cox et al., 2011). Only the highest scoring theoretical fragment ion spectrum among the ones with the correct precursor mass is retained as the PSM for each MS2 spectrum. Based on these PSMs, MaxQuant performs non-linear mass- and retention-time-recalibration, after which the main Andromeda search is performed (usually with lower precursor ion tolerance). Optionally, a "Second peptide search" is carried out after subtraction of explained peaks in all MS2 spectra, in order to identify co-fragmented peptide ions. After the application of PSM and peptide FDR filters, "Match between runs" (Cox et al., 2014) is performed if multiple raw files are analysed at the same time. Here, identifications from peptides are transferred to unidentified precursor ions with closely matching masses and recalibrated retention times in an approach similar to the accurate mass and time tag (Stanley et al., 2011). The degree of accepted deviation is user-specified. Subsequently, proteins present in the sample are inferred based on identified peptides (Nesvizhskii and Aebersold, 2005), applying the principle of parsimony (Nesvizhskii et al., 2007). Therefore, only the smallest number of proteins or protein groups is reported, which can account for all observed peptides, followed by the final application of a protein FDR and subsequent protein quantification.

1.3.2.2 Protein quantification

Protein quantification in MaxQuant can either be based on MS1 or MS2 spectra, with special experimental setups allowing quantification using MS3 (not discussed here, McAlister et al., 2014). Since this thesis made extensive use of MS1-based label-free quantification (LFQ, Cox et al., 2014) and intensity-based absolute quantification (iBAQ, Schwanhausser et al., 2011), these methods are

outlined below. With the exception of spectral counting (number of MS2 spectra recorded for the same peptide, Liu et al., 2004), methods for MS2-based quantification usually involve isobaric labels like tandem mass tags (TMT, Thompson et al., 2003) or isobaric tags for relative and absolute quantitation (iTRAQ, Wiese et al., 2007), which made them expensive for the study at hand. The same is true for MS1-based quantification using stable isotope labelling by amino acids in cell culture (SILAC, Ong et al., 2002), which also suffered from limitations in terms of multiplexing capability.

MS1-based label-free quantification using MaxLFQ (Cox et al., 2014) in MaxQuant relies on extracted ion chromatograms or XICs of peptides. In order to generate XICs, the MS1 signal intensity of a given peptide is integrated over its chromatographic retention time, analogous to a "mass chromatogram" for a single peptide (Hites and Biemann, 1970). This method is highly dependent on the resolution of the MS1 spectrum, since signals from peptide precursors with similar masses need to be resolved in order to achieve accurate quantification (Andersen et al., 2003). In order to compare multiple mass spectrometry runs – even from experimental designs requiring pre-fractionation of proteins and/or peptides – MaxQuant makes use of "delayed normalisation" (Cox et al., 2014). Under the assumption that the majority of peptides should not change in abundance between different biological samples, MaxQuant uses a global optimisation procedure to calculate a normalisation factor for each MS run. For each protein and sample pair, peptides corresponding to this protein that are present in both samples are then used to calculate peptide ratios, the median of which corresponds to the pairwise protein ratio for the respective sample pair. The resulting matrix of pairwise protein ratios is finally treated as an

overdetermined system of equations, for which an optimal solution is found by least-squares analysis, preserving the total summed intensity for a given protein across all samples by rescaling its profile to the sum of raw peptide intensities corresponding to said protein across samples.

In comparison to LFQ, iBAQ is much simpler. In order to calculate iBAQ values, MaxQuant only uses non-normalised raw peptide intensities based on XICs and sums them up for each protein and biological sample. These intensities will however depend on the length of a given protein, since longer proteins are more likely to yield more identified peptides. In order to compensate for this, MaxQuant divides these protein intensities by the number of theoretically observable peptides, considering only fully tryptic peptides between 6 and 30 amino acids without missed cleavages (Schwanhausser et al., 2011).

1.4 Aim of this work

As indicated by the title of this thesis, the central aim of this project was the pharmacoproteomic characterisation of human colon and rectal cancer. In other words, this study attempted to explain the variation in the observed drug sensitivity between different CRC tumours and cell lines (see section 1.1.6) by variations in their proteome profiles (see section 1.2.1). The entire project revolved around the notion that appropriately trained pharmacoproteomic models relating drug sensitivity to steady-state protein expression can be used to accurately predict the response of tumours and cell lines towards drugs of various kinds and allow the identification of proteomic markers of drug sensitivity and resistance, as exemplified by previous studies (Gholami et al., 2013; Lawrence et al., 2015). This concept was adapted from more established

pharmacogenomic approaches, which aimed at accomplishing the same task using steady-state mRNA expression profiles or other genomic features like DNA mutations or methylation (see section 1.2.1). However, since drugs almost always target proteins and not genes or mRNA, this project employed a pharmacoproteomic approach, which was complemented by transcriptomics data in order to supplement evidence on the protein level by performing multi-omics data integration. The aims of this project were: 1) establish and characterise a training and testing dataset suitable for pharmacoproteomic modelling of CRC; 2) model drug sensitivity as a function of steady-state protein expression; 3) use these models to a) predict the drug sensitivity of cell lines and patients and b) identify proteomic markers of drug sensitivity and resistance; 4) use *in vitro* experiments to a) test drug sensitivity predictions on independent cell lines and b) validate proteomic markers of drug sensitivity; 5) use immunohistochemistry on patient samples to gauge the clinical relevance of proteomic markers of drug sensitivity. Combined with a molecular subtyping approach on the protein level, the training dataset used to generate these pharmacoproteomic models could then be employed to identify drugs differentially effective in different proteomic subtypes of CRC, thereby informing potential prospective clinical studies and promoting personalised medicine. To achieve these goals, a combination of proteomics, chemoproteomics, bioinformatics, cell biology and immunohistochemistry was used.

A central requirement for a study of this kind was a reasonably sized and diverse set of CRC samples characterised with respect to their sensitivity towards relevant drugs, as well as considering their steady-state protein expression profiles. These samples should ideally cover the different molecular subtypes of

CRC so that they could serve as a representative training dataset during the model-building process (Chen et al., 2015; Wilding and Bodmer, 2014). The gold standard in this case would be to use tumour samples if the goal is to eventually translate findings into the clinic. However, since drug sensitivities of tumours cannot be determined on a large scale in clinical studies, appropriate model systems were required. This project took advantage of the fact that CRC cell lines were recently extensively characterised with respect to drug sensitivity as part of major screening efforts (Barretina et al., 2012; Basu et al., 2013; Garnett et al., 2012; Iorio et al., 2016; Medico et al., 2015; Rees et al., 2016; Seashore-Ludlow et al., 2015). With a collection of CRC cell lines at hand (see section 2.2) – many of which were included in the drug sensitivity datasets mentioned above – the characterisation of these 65 cell lines (collectively called "CRC65 cell line panel") with respect to their subtype membership was the first step. Most of the molecular subtyping in CRC in the past decade was done on the mRNA level, with the consensus molecular subtypes (CMS) published by Guinney et al. representing the current subtyping consensus (Guinney et al., 2015). The authors provided the community with a single-sample predictor suitable for assigning novel samples to the established CMS based on transcriptomics data, but their subtype model did not contain any information on cell lines. However, CRC cell lines were shown to capture the main molecular subtypes in CRC with respect to the majority of subtype models combined to form the CMS (Medico et al., 2015; Mouradov et al., 2014). Hence, the CRC65 cell line panel first needed to be characterised on the transcriptome level in order to be able to assign the cell lines to the CMS using the above-mentioned predictor, which is described in section 3.2. However, the CMS subtype model showed only partial concordance with the

proteomic subtype model published by the "Clinical Proteomics Tumor Analysis Consortium" (CPTAC, Zhang et al., 2014a). Following a pharmacoproteomic approach as described above, the second step was therefore to characterise the CRC65 cell line panel on the proteome level to a comparable depth as the CPTAC tumour dataset. This was done in order to lay the foundations for the discovery of integrated proteomic subtypes of CRC consisting of cell lines and patients, which is detailed in section 3.3. This enabled the identification of representative cell lines as model systems for the pre-clinical evaluation of drugs in CRC. In addition to the full proteome profiling described in section 3.3, the amount of cell lysate generated as part of this project from three biological replicates of the CRC65 cell line panel also allowed the characterisation of these cell lines with respect to their steady-state kinase expression profiles or kinomes, described in section 4.5. Following a multi-omics data integration approach, a comparison of transcriptome and full proteome data of the CRC65 and CPTAC datasets is illustrated in sections 5.2.2 and 5.2.1. Because of the similarities between transcriptomics and proteomics data once systematic differences between them were accounted for, mRNA abundance could be used to impute some of the missing values in the full proteome data of both datasets (see section 5.2.2), while a newly developed method for minimum-guided missing value imputation (see section 2.10.4) was used to impute the remainder of the missing values after examination of several alternatives, as described in section 5.2.3. Afterwards, the CRC65 and CPTAC datasets were combined and used together in section 5.3 to identify and characterise integrated Full Proteome Subtypes (FPSs) of CRC, which partially recapitulated the CMS subtype model. The kinomes of the CRC65 cell line panel were then used to identify Kinobeads subtypes (KSs) of CRC in section

5.4, which were also compared to the integrated FPSs and the CMS. Since the KSs fully recapitulated the FPSs, the FPSs were used as the reference subtype model in later stages of the project. It should be stressed that the FPSs consisted of patients and cell lines, some of which were included in the drug sensitivity datasets mentioned above, while others were not, forming an ideal training and testing dataset for modelling drug sensitivity as a function of steady-state protein expression.

With a training and testing dataset at hand, the next step was to build the above-mentioned pharmacoproteomic models of drug sensitivity and use them to predict the drug response of cell lines and patients towards a wide variety of drugs and combination treatments, detailed in section 5.5. Models based on the full proteome profiles of the CRC65 cell line panel were used together with the FPSs for target space enrichment, while models based on the kinome profiles were used for the discovery of kinases associated with drug sensitivity or resistance, illustrated in section 5.6. This represented a more focused approach centred on a class of proteins frequently targeted by drugs and often deregulated in cancer (Brognard and Hunter, 2011; Giancotti, 2014; Kolch et al., 2015; Manning, 2009; Zhang et al., 2009). Some of the drug sensitivity predictions were then tested on independent cell lines and MERTK was validated as a predictive marker for resistance towards MEK1/2 inhibitors using co-treatment experiments (section 5.7). Finally, MERTK was also confirmed to have a prognostic role in CRC using immunohistochemistry of a large collection of tumours in section 5.8. Apart from the validation experiments, this multi-omics data integration strategy is summarised in Figure 1.4.1, which visualises the streamlined data analysis workflow.

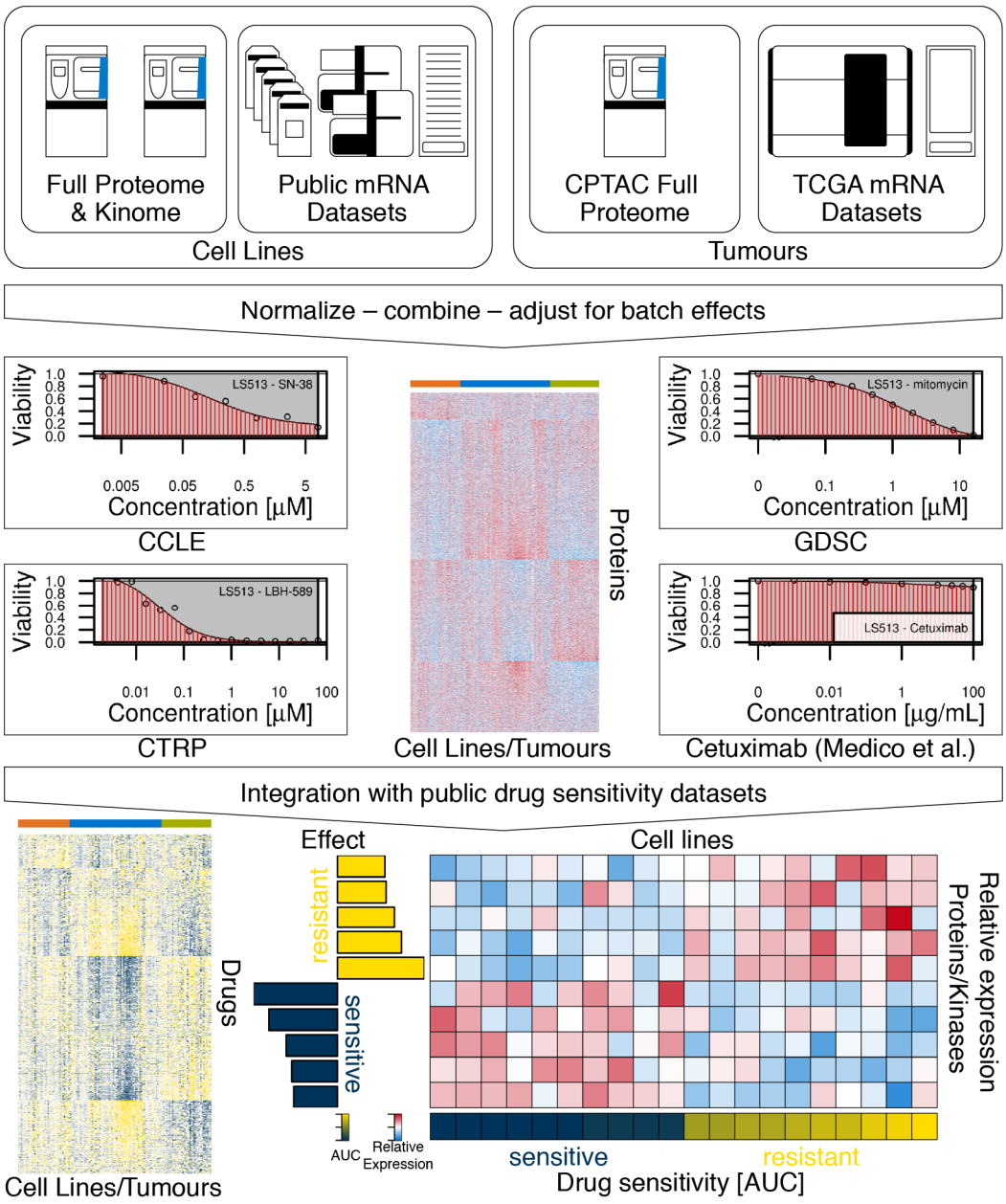


Figure 1.4.1

Multi-omics data integration strategy

Overview of the multi-omics data integration workflow. CCLE, CTRP, GDSC and Cetuximab (Medico et al.) indicate the different drug sensitivity datasets, which were overlaid onto the newly-generated proteomics data (Barretina et al., 2012; Basu et al., 2013; Garnett et al., 2012; Iorio et al., 2016; Medico et al., 2015; Rees et al., 2016; Seashore-Ludlow et al., 2015). The area under the normalized dose-response curve or AUC between the lowest and highest tested concentration of four different compounds is shown as an example (red shading lines). An example of a so-called effect-size heat map is shown at the bottom-right of the figure (see section 2.13 for details on the modelling procedure and an explanation of this type of figure).

Chapter 2
Materials and Methods

2.1 Chemicals

All chemicals were from Sigma-Aldrich if not stated otherwise.

2.2 Cell lines

All cell lines in the CRC65 panel apart from SW480 (kind gift from Ulrike Stein from the MDC, Berlin) were collected by the "Cancer and Immunogenetics Laboratory" headed by Prof. Sir Walter Bodmer FRS at the University of Oxford and their sources are described in Supplementary Table 1 of the publication by Emaduddin et al. (2008). MV-4-11 – a macrophage-derived cell line used for the optimisation of Kinobeads pulldowns – was sourced from the American Tissue Culture Collection.

Table 2.2.1

Cell lines used in this study including references/commercial sources

Cell line	Reference/Commercial Source	Cell line	Reference/Commercial Source
CaCo-2	ATCC, USA (www.atcc.org)	HDC-8	Brüderlein et al. (1990)
Colo 320DM		HDC-9	
DLD-1		HDC-54	
HCT 116		HDC-57	
HT-29		HDC-73	
LoVo		HDC-82	
LS123		HDC-111	
LS 174T		HDC-114	
LS 180		HDC-135	
LS411		HDC-142	
LS513		HDC-143	
LS1034		Colo-678	DSMZ, Germany (www.dsmz.de)
MV-4-11		Colo 741	ECACC, UK (www.ecacc.org.uk)
NCI-H548		CC20	Greenhalgh and Kinsella (1985)
NCI-H716		CaR-1	HSRRB, Japan (www.jhsf.or.jp)
NCI-H747		CCK-81	
RKO		CoCM-1	
SK-CO-1		RCM-1	
SNU-C2B		HCA-7	Kirkland (1985)
SW48		HCA-46	Kirkland and Bailey (1986)
SW403		HRA-19	Kirkland (1986)
SW620		C125-PM	Liu and Bodmer (2006)
SW837		VACO 4A	McBain et al. (1984)
SW948		GP2d	Solic et al. (1995)
SW1116		SW480	Ulrike Stein, MDC, Berlin, Germany
SW1417		OXCO-1	Vincenzo Cerundolo and Khoon Lin Ling, WIMM, Oxford, UK
T84		OXCO-3	
PC/JW	Berry and Paraskeva (1988)	HT55	Watkins and Sanger (1977)
C10	Browning et al. (1993)	C99	Wheeler et al. (1999)
C32		C106	
C70		LIM1863	Whitehead et al. (1987)
C75			
C80			
C84			
CC07			

2.3 Cell culture and lysis

Slightly altering the culture conditions described by Emaduddin et al. (2008), cells in the CRC65 panel were grown in high glucose Dulbecco's Modified Eagle Medium (DMEM, including GlutaMAX and pyruvate; PAA) containing 1% Pen-Strep (Penicillin at 100 units/mL and Streptomycin at 100 µg/mL final concentration, PAA) and 10% fetal bovine serum (FBS, SLI EU-000F Batch 503005) in a humidified incubator at 37°C and 10% CO_2 using T-175 flasks (Corning). Hereafter, this medium composition is referred to as "culture medium" and the culture conditions apart from the culture vessel are referred to as "standard culture conditions". Adherent cells were harvested at approximately 80-90% confluency by first washing three times with 50 mL ice-cold phosphate buffer saline (PBS, 0.003 M Na_2HPO_4, 0.001 M KH_2PO_4, 150 mM NaCl) and subsequently scraping them into 1.5 mL pre-cooled lysis buffer per 175 cm^2 flask. Suspension cells (e.g. HDC-135) growing in 250 mL of culture medium were harvested by centrifugation in 250 mL centrifuge tubes (Corning) at 300 xg and 4°C for 5 min and washed with the same volume of chilled PBS after discarding the medium. Cells were then pelleted again at 2,600 xg and 4°C for 15 min, the supernatant was discarded and the cells were lysed in about two pellet volumes of lysis buffer after chilling the pellet for 15 min on ice. Lysis was carried out for 30 min at 4°C using radioimmunoprecipitation assay buffer with 100 mM NaCl (RIPA100 buffer; 20 mM Tris/HCl pH 7.5, 1 mM EDTA, 100 mM NaCl, 1% Triton X-100, 0.5% sodium deoxycholate and 0.1% SDS) containing protease (Complete™ mini with EDTA; Roche) and phosphatase inhibitors (Phosphatase Inhibitor cocktail 1 and 2; Sigma-Aldrich) at 2x and 5x the final concentration recommended by the manufacturer, respectively. After clearing the total cell

lysates (TCLs) at 22,000 xg for 30 min at 4°C, the protein concentration was determined using the Bradford assay (Bradford, 1976) and the TCLs were stored at -80°C until further use. These TCLs are designated "TCLs in RIPA100 buffer" throughout this thesis.

MV-4-11 cells were grown in Roswell Park Memorial Institute 1640 medium (RPMI1640, Thermo Fisher) containing 1% Pen-Strep and 10% FBS in a humidified incubator at 37°C and 5% CO_2 using 1 L roller bottles (Corning) under constant agitation. At a density of approximately $1.2x10^6$ cells/mL (600 mL flask volume), 400-450 mL of cell suspension were harvested and the rest of the suspension was used for continuing the culture. These conditions enabled two harvests per week of more than 150 mg of protein per flask. First, the cell suspension was distributed to 250 mL centrifuge tubes (Corning), the cells were pelleted at 300 xg for 5 min at room temperature (RT) and the majority of the supernatant was discarded. Afterwards, the remaining cell suspension (1-2 mL per 250 mL centrifuge tube) was pooled in 50 mL centrifuge tubes (Thermo Fisher), the cells were pelleted again at 300 xg for 5 min and the supernatant was discarded. Next, the cells were washed twice with 20 mL ice-cold PBS, pelleting the cells at 300 xg for 5 min after each washing step and discarding the supernatant. After aspirating the remainder of the supernatant, the wet cell pellet was weighed, followed by the addition of four times the amount of ice-cold lysis buffer (e.g. 12 mL for 3 g of cells), which resulted in protein concentrations of about 12 mg/mL. Depending on the Kinobead optimisation experiment, cells were lysed for 30 min at 4°C with either RIPA100 buffer as described above or with compound pulldown buffer (CP buffer, 50 mM Tris/HCl pH 7.5, 5% glycerol, 1.5 mM $MgCl_2$, 150 mM NaCl, 1 mM Na_3VO_4) containing 0.8% Igepal CA-630

(Thermo Fisher). Both lysis buffers contained protease (Complete™ mini with EDTA; Roche) and phosphatase inhibitors (Phosphatase Inhibitor cocktail 1 and 2; Sigma-Aldrich) at 2x and 5x the final concentration recommended by the manufacturer, as well as 20 nM calyculin A. Afterwards, the TCLs were transferred to 2 mL microcentrifuge tubes and were cleared by centrifugation at 22,000 xg and 4°C for 30 min. Finally, the TCLs were distributed to clean microcentrifuge tubes, the protein concentration was determined using the Bradford assay and the TCLs were stored at -80°C until further use. These TCLs are designated "TCLs in CP buffer" throughout this thesis.

2.4 Patient samples

Analysis of MERTK expression in 1,074 patients from the QUASAR2 clinical trial cohort was performed in accordance with the Metropolitan Multi-centre Research Ethics Committee (Ref. 04/MRE11/18). QUASAR2 is a phase III international randomised controlled trial, which collected data on toxicity, overall survival (OS), disease-free survival (DFS) and recurrence-free survival (RFS) for 1,941 stage II/III CRC patients, with the aim to determine the efficacy of adjuvant capecitabine $\pm$ bevacizumab after resection of the primary tumour. A biobank comprising 1,350 formalin-fixed, paraffin embedded (FFPE) blocks was established and tissue microarrays (TMAs) were generated from 1.2 mm cores by Haitao Wang (research assistant in Prof. Rachel Kerrs group).

2.5 Sample processing for mass spectrometry

2.5.1 Full proteomes

2.5.1.1 Acetone precipitation

TCLs were first acetone-precipitated to remove detergents. For this, twice the amount of protein lysate required for in-solution digestion was distributed into 2 mL microcentrifuge tubes to account for losses during the precipitation. Subsequently, four volumes of pre-cooled (-40°C) acetone were added to each tube, followed by vigorous mixing and subsequent incubation over night at -40°C. Precipitates were then washed twice with an excess of acetone after the samples were centrifuged for 10 min at 13,000 xg and 4°C and the pellets were collected in a single 2 mL microcentrifuge tube per TCL sample. For each washing step, the supernatant was taken off and replaced with 1 mL fresh, cold acetone, the pellets were dislodged using a pipette and after an incubation at -40°C for 1h, the precipitates were pelleted for 10 min at 13,000 xg and 4°C. After the final washing step, the supernatant was taken off and the samples were left to dry in a fume hood at RT.

2.5.1.2 Resolubilisation

Following acetone precipitation, the samples were re-suspended in urea-buffer (40 mM Tris/HCl pH = 7.6, 8 M Urea) containing protease (Complete™ mini without EDTA; Roche) and phosphatase inhibitors (Phosphatase Inhibitor cocktail 1 and 2; Sigma-Aldrich) at 1x and 5x the final concentration recommended by the manufacturer, respectively, as well as 20 nM calyculin A. In order to ensure proper re-solubilisation of proteins, the precipitates were first mixed thoroughly by pipetting up and down, followed by sonication of each

sample for 5 min on ice using an HF generator GM mini20, equipped with an ultrasonic converter UW mini20 and a microtip MS 2.5 sonotrode (Bandelin), which was set to 3 s pulses at 30% intensity with 3 s pause in between. After re-solubilisation, the protein concentration of the lysate was determined again using the Bradford assay (Bradford, 1976).

2.5.1.3 In-solution digestion

For in-solution digestion of proteins, 3.5 mg of TCL per sample were transferred to new microcentrifuge tubes and disulfide bonds were reduced with 10 mM dithiothreitol (DTT) for 1 h at 56°C and 700 rpm in a thermomixer (Thermo Fisher). Subsequently, samples were cooled to RT and free sulfhydryl groups were alkylated using chloroacetamide at 55 mM for 30 min at RT protected from light. Following reduction and alkylation, samples were diluted with 40 mM Tris/HCl pH 7.6 to reduce the urea concentration to 1.5 M, 1.5 µL $CaCl_2$ were added to each sample and proteins were digested over night at 37°C and 700 rpm in a thermomixer using trypsin (Roche) at a protease to protein ratio of 1:50.

2.5.1.4 Desalting

Desalting of peptide mixtures was carried out at RT using Sep-Pak cartridges (50 mg sorbent per cartridge, Waters) and a vacuum manifold. Briefly, samples were acidified to pH 2 using TFA to stop the digestion, followed by 5 min of centrifugation at 5,000 xg and RT. Cartridges were wetted twice with 1 mL solvent B (0.07% TFA in 50% acetonitrile) and subsequently equilibrated using three times 1 mL solvent A (0.07% TFA). Samples were then loaded onto cartridges under reduced flow, the flow-through was collected and re-loaded a

second time, followed by five washing steps with 1 mL solvent A each. Finally, peptides were eluted slowly into new microcentrifuge tubes using 0.3 mL solvent B, the eluate was diluted with 0.2 mL solvent A to reduce the concentration of acetonitrile (ACN) to 30% and the samples were stored at -80°C until further use.

2.5.1.5 Hydrophilic strong anion exchange chromatography

For hydrophilic strong anion exchange (hSAX) chromatography, peptide solutions were first dried down in a Speed-Vac, re-solubilised in hSAX solvent A (5 mM Tris/HCl, pH 8.5) to a concentration of 2.73 µg/µL peptide and then centrifuged for 30 s at 5,000 xg to spin down insoluble debris. Next, the samples were placed in the autosampler of a Dionex Ultimate 3000 HPLC system (Thermo Fisher), which was equipped with an IonPac AG24 guard column (Thermo Fisher), as well as an IonPac AS24 strong anion exchange column (Thermo Fisher). Chromatography was performed at 30°C and a flow rate of 250 µL/min. Following the injection of 100 µL sample and 3 min of equilibration with 100% hSAX solvent A, peptides were eluted using a two-step linear gradient from 0% to 25% hSAX solvent B (5 mM Tris/HCl, pH 8.5, 1 M NaCl) in 24 min and from 25% to 100% solvent B in 13 min. Solvent B was kept at 100% for another 4 min to flush the column before returning to 0% in 1 min and additional 5 min of equilibration with 100% hSAX solvent A. Collection of 48 fractions of 250 µL each was started after 2 min of gradient time, subsequently combining them to form 24 fractions after consulting the 214 nm chromatography-trace, as well as separate LC-MS/MS measurements of each hSAX fraction from an in-solution digest of lysate from human placenta (obtained from Freising hospital following informed consent), performed by Lihua Li (former lab member). For that,

fractions 5-7 were pooled to form fraction 5, fractions 24-25 were pooled to form fraction 22, fractions 26-28 were pooled to form fraction 23 and fractions 29-48 were pooled to form fraction 24, respectively. Hydrophilic strong anion exchange chromatography of the majority of cell lines was performed by Riccardo Zenezini Chiozzi (former lab member).

2.5.1.6 Post-hSAX desalting

All 24 hSAX fractions were desalted using StageTips as described earlier (Rappsilber et al., 2007), with minor modifications. Briefly, five C18-discs (3M Empore) of about 1.5 mm diameter in 200 µL pipette tips were used, which were placed into 1.5 mL microcentrifuge tubes through a pre-cut hole in the lid. Unless stated otherwise, 250 µL of solvent was used and forced through the StageTips by centrifugation at 850 xg for 10 min, emptying the tube after every other centrifugation step. Fractions were acidified with TFA to pH 2 while wetting the StageTips with MeOH and subsequently with 5% TFA in 80% ACN. Next, StageTips were equilibrated twice with 0.1% TFA, samples were loaded and slowly passed through the StageTips at 370 xg for 10 min, repeating the loading procedure for samples with volumes exceeding 250 µL. Subsequently, StageTips were washed twice with 0.1% TFA and the remainder of the solvent was forced through the StageTips with a syringe without drying them completely. Finally, the microcentrifuge tubes were replaced with fresh ones and the desalted peptides were eluted using 200 µL of 0.1% TFA in 60% ACN, centrifuging the StageTips at 370 xg for 10 min. The remainder of the solvent was forced through the StageTips using a syringe, thereby drying it completely. Eluates were transferred to 96-well plates, evaporated to dryness using a Speed-Vac and subsequently stored at -20°C

until further use. Post-hSAX desalting of the majority of cell lines was performed by Riccardo Zenezini Chiozzi (former lab member).

2.5.2 Kinobeads

2.5.2.1 Kinobeads in microplates

Kinobeads (KB) pulldowns (PD) in biological triplicates were performed as described elsewhere (Medard et al., 2015), with minor modifications. TCLs were quickly thawed at 37°C in a water bath and subsequently kept on ice. For each final pulldown, 3 mL of 2.2 mg/mL TCL were transferred to pre-cooled ultracentrifuge tubes (Beckman & Coulter) and cleared by ultracentrifugation at 192,000 xg and 4°C for 20 min. Volumes and concentrations were optimised as described in section 4.4. Meanwhile, 35 µL of settled Kinobeads-γ (KBγ) per TCL were distributed to separate wells of deep 96-well filter plates (Porvair Sciences), washed twice with 1 mL CP buffer each and subsequently equilibrated using 1 mL CP buffer each, supplemented with 0.4% Igepal CA-630. Solvents were drained from the beads via gravitational flow through the filter unless stated otherwise. Afterwards, the plates were placed on regular 96-well plates for waste collection, centrifuged for 2 min at 240 xg and 4°C and sealed with the corresponding bottom sealing mat. Next, approximately 1.8 mL (equivalent of 4 mg of protein) of each TCL was transferred to its corresponding well, the plates were sealed with their top sealing mats and incubated for 60 min at 4°C on a head-over-end shaker. After lysate incubation, the plates were centrifuged for 3 min at 240 xg and 4°C, followed by drainage of the TCLs from the beads and three washing steps with CP buffer containing 0.4% Igepal CA-630 and two washing steps with CP buffer containing 0.2% Igepal CA-630. Subsequently the plates

were centrifuged for 2 min at 240 xg on a 96-well plate for waste collection and sealed again with their bottom sealing mat. For elution of proteins bound to KBγ, 60 µL of 1x NuPAGE LDS sample buffer (Thermo Fisher) per well containing 50 mM DTT were distributed to the filter plates, the plates were sealed with their top sealing mats and incubated for 30 min at 50°C and 700 rpm in a thermomixer. Afterwards, the top and bottom sealing mats were removed, the filter plates were placed onto 96-well plates for eluate collection and centrifuged for 2 min at 670 xg and 4°C. Free sulfhydryl groups were then alkylated for 30 min at RT protected from light by addition of chloroacetamide to a final concentration of 55 mM.

2.5.2.2 Kinobeads in microcolumns

Amending the standard protocol for KBγ pulldowns, 2 mL of 15 mg/mL TCL in CP buffer per pulldown were diluted with 2 mL CP buffer containing the same concentration of protease and phosphatase inhibitors to bring the concentration of Igepal CA-630 to 0.4%. The diluted lysates were cleared by ultracentrifugation at 192,000 xg and 4°C for 20 min and subsequently diluted a second time with CP buffer containing 0.4% Igepal CA-630 and the same concentration of protease and phosphatase inhibitors, in order to bring the protein concentration to 5 mg/mL. For each microcolumn, a micro-spin column (Thermo Fisher) was then sequentially packed with 20 µL settled bead volume of each of the five different bead types contained in the KBγ mixture, resulting in a total bed volume of 100 µL settled beads. After each addition of 40 µL 1:1 bead slurry in ethanol/EtOH, the columns were centrifuged in microcentrifuge tubes for 5 min at 240 xg and 4°C in order to settle the beads before the next type of bead was added on top,

discarding the flow-through. Beads were stacked from the bottom to the top of the column in either increasing (LOHI: compound 5, 7, 1, 15, 19) or decreasing (HILO) order with respect to the number of kinases they enrich in single-compound pulldowns (Medard et al., 2015). Packed microcolumns were then connected via Luer Lock connections to an open circuit of Tygon Standard tubing (4.8 mm i.d.), attached to a BVP Standard peristaltic pump through MS/CA Click'n'go cassettes, which were seated in an MS/CA 8-6 pump head (Ismatec). The entire experimental setup was kept at 4°C. The tubing was cut to a final inner volume of 5 mL and was pre-filled with CP buffer. The beads were then washed and equilibrated using the same sequence of buffers as described in section 2.5.2.1 by pumping 3 mL of the respective solution at 1 mL/min through the microcolumns. After equilibration, the lysates were added to the syringes and pumped through the circuit at 1 mL/min for 60 min, followed by the collection of the lysate in 15 mL centrifuge tubes for pulldown of pulldown (PD/PD) experiments. The microcolumns were then washed using the same sequence of buffers as described in section 2.5.2.1 by pumping 3 mL of the respective solution at 1 mL/min through the microcolumns. Afterwards, the microcolumns were detached from the circuit and centrifuged for 1 min at 240 xg and 4°C before proceeding with the elution of proteins from the beads. For that, 60 µl of 2x NuPAGE LDS sample buffer (Thermo Fisher) containing 50 mM DTT were added to each microcolumn, which were then placed into microcentrifuge tubes and incubated for 30 min at 50°C and 700 rpm in a thermomixer (Thermo Fisher). Lysates were finally collected by centrifugation for 2 min at 480 xg and free sulfhydryl groups were then alkylated for 30 min at RT protected from light by addition of chloroacetamide to a final concentration of 55 mM. For PD/PD

experiments, lysates were processed as described above using a fresh set of pre-equilibrated microcolumns.

2.5.2.3 Kinobeads in 15 mL centrifuge tubes

In order to compare Kinobeads enrichments using microcolumns with batch enrichments, the volumes of the batch enrichment protocol described in section 2.5.2.1 had to be adjusted to the volumes used with the microcolumn setup. Therefore, all steps up to and including the lysate incubation previously performed in 96-well filter plates were instead performed in 15 mL centrifuge tubes, keeping all other parameters the same as in the microcolumn setup described in section 2.5.2.2. After incubating the lysate with the beads in batch-mode, the beads were transferred to mobicol columns (M1002-90, MoBiTec) for subsequent washing and elution steps as described in section 2.5.2.2.

2.5.2.4 Detergent removal from Kinobeads eluates

After alkylating free sulfhydryl groups using chloroacetamide, detergents and salts were removed from Kinobeads eluates by gel electrophoresis, followed by tryptic in-gel digestion. Proteins in LDS sample buffer were run into 4−12% NuPAGE gels (Thermo Fisher) at 200 V constant for 5 min (approximately 0.5 cm). Subsequently, gels were fixed for 1 h in fixation solution (2% acetic acid/AcOH in 40% methanol/MeOH), stained for 10 min in staining solution (7.5% Rotiphorese Blue R from Carl Roth in 20% MeOH), partially de-stained by washing once with washing solution (5% AcOH in 25% MeOH) and de-stained further by incubating them three times for 1 h with 25% EtOH under constant agitation at RT. Gels were stored in 1% AcOH at 4°C until further use.

2.5.2.5 In-gel digestion

In-gel digestion of proteins after Kinobeads pulldowns was carried out by Andreas Klaus or Michaela Krötz-Fahning (technical assistants in the laboratory of Prof. Küster) as described earlier (Hellman et al., 1995; Rosenfeld et al., 1992), with minor modifications. Briefly, protein bands were cut out after gel electrophoresis and further processed in modified 96-well plates containing small holes for liquid drainage by centrifugation in a Speed-Vac (UNIVAPO 150 ECH, equipped with the Swing-out-rotor for 4 microtiterplates) at 1,250 rpm and 1 atm for 1 min after each processing step, all of which took place at RT unless specified otherwise. Gel pieces were de-stained twice – for 2 h and 1 h, respectively – at 55°C with de-staining solution (5 mM Triethylammonium bicarbonate/TEAB in 50% EtOH), de-hydrated in absolute EtOH for 10 min, washed with 5 mM TEAB for 20 min, de-hydrated two more times with absolute EtOH for 10 min and incubated with 10 ng/mL sequencing grade modified trypsin (Promega) in 5 mM TEAB for 15 min at 4°C, followed by the removal of residual trypsin by centrifugation as described above. After the addition of 5 mM TEAB, the modified 96-well plate was placed on a collection plate and digestion was carried out at 37°C for 4 h. Peptides were then acidified with 5% formic acid (FA) to stop the digestion, extracted twice with 1% FA and once with 0.1% FA in 60% ACN for 30 min each, followed by two drying steps with 100% ACN for 15 min each, after which peptides were evaporated to dryness under vacuum. Dried peptides were stored at -20°C until further use.

2.6 LC-MS/MS data acquisition

2.6.1 Full proteomes

Full proteome fractions were measured using nanoflow LC-MS/MS by directly coupling an Eksigent nanoLC-Ultra 1D+ (Eksigent) to an Orbitrap Velos mass spectrometer (Thermo Fisher). Peptides were dissolved in 20 µL buffer A (0.1% FA), injecting 5 µL per measurement. Using a flow rate of 5 µL/min, peptides were then loaded onto a 2 cm trap column (100 µm i.d., ReproSil-Pur C18-AQ 5 µm, Dr. Maisch) and washed for 10 min with 100% buffer A. Subsequently, peptides were separated on a 40 cm analytical column (75 µm i.d., ReproSil-Pur C18-AQ 5 µm, Dr. Maisch) using a flow rate of 300 nL/min and a gradient from 2% buffer B (0.1% FA and 5% DMSO in ACN) to 4% in 2 min (buffer A now also contained 5% DMSO) and from 4% to 32% in 96 min. Buffer B was then ramped from 32% to 80% in 1 min and the column was flushed with 80% buffer B for 4 min, before returning to 2% buffer B in 2 min and a final equilibration step with 2% buffer B for 5 min. This resulted in a turnaround time of 120 min per full proteome fraction, totalling 130 days of measurement time for the entire CRC65 cell line panel.

Using electrospray ionization (Fenn et al., 1989), the eluate from the analytical column was sprayed via stainless steel emitters (Thermo) at a source voltage of 2.6 kV towards the orifice of the mass spectrometer, a transfer capillary heated to 275°C. The Orbitrap Velos was set to data-dependent acquisition in positive ion mode, automatically selecting the top 10 most intense precursor ions from the preceding full MS (MS1) spectrum with an isolation width of 2.0 Th (Cooks and Rockwood, 1991) for fragmentation using higher energy C-trap dissociation

(HCD) at 30% normalised collision energy and subsequent identification by MS/MS. MS1 (360-1,300 m/z) and MS2 (precursor-dependent m/z-range, starting at m/z 100) spectra were acquired in the Orbitrap using a resolution of 30,000 and 7,500 at m/z 400, with an automatic gain control (AGC) target value of $1x10^6$ and $3x10^4$ charges and a maximum injection time of 100 ms and 200 ms, respectively. Dynamic exclusion was set to 60 s and internal calibration was carried out using the signal of a DMSO-containing ion at m/z 401.922720 (Hahne et al., 2013).

2.6.2 Kinobeads

Kinobeads eluates were measured using nanoflow LC-MS/MS by directly coupling a nanoLC-Ultra 1D+ (Eksigent) to an Orbitrap Elite mass spectrometer (Thermo Fisher). Chromatography as well as data acquisition were similar to the settings used for full proteome fractions, hence the description below is restricted to differences between the two setups. Instead of injecting 5 µL per measurement, each injection was done using 10 µL of sample, separating peptides using a slightly steeper gradient from 2% buffer B to 4% in 2 min and from 4% to 32% in 88 min instead of 96 min. The rest of the gradient was kept the same, resulting in a turnaround time of 110 min per Kinobeads pulldown, totalling approximately 15 days of measurement time for the entire CRC65 cell line panel in biological triplicates.

The source voltage was at 2.2 kV instead of 2.6 kV and the Orbitrap Elite selected the top 15 most intense precursor ions for MS/MS instead of the top 10 most intense ones because of its higher scanning speed. While the MS1 m/z range, resolution, AGC target value and maximum injection time was the same as for the

Orbitrap Velos, the Orbitrap Elite acquired MS2 spectra at a resolution of 15,000 at m/z 400 with an AGC target value of $2x10^4$ charges and a maximum injection time of 100 ms. Dynamic exclusion was set to 20 s instead of 60 s and the Orbitrap Elite also made use of a global inclusion list, which contained precursor ions and retention times from frequently observed kinase peptides.

2.7 Analysis of LC-MS/MS raw data & giBAQ quantification

2.7.1 Full proteomes and reanalysis of the CPTAC patient data

MaxQuant v.1.5.3.30 (Cox and Mann, 2008) was used together with Andromeda (Cox et al., 2011) to search newly generated LC-MS/MS raw data, as well as the raw data from the original CPTAC publication on human colon and rectal cancer (Zhang et al., 2014a) against UniProtKB (v25.11.2015, 92,011 sequences; UniProt Consortium, 2015), concatenated with a list of common contaminants supplied by MaxQuant (245 sequences) in two separate runs with identical settings. The digestion mode was set to fully tryptic, allowing for cleavage before proline (Trypsin/P) and a maximum of two missed cleavages. Carbamidomethylation of cyteines was set as a fixed modification and oxidation of methionines, as well as acetylation of protein N-termini were set as variable modifications, allowing for a maximum number of 5 modifications per peptide. Candidate peptides were required to have a length of at least 7 amino acids, with a maximum peptide mass of 4,600 Da. The fragment ion tolerance was set to 20 ppm for FTMS and 0.4 Da for ITMS spectra, respectively. A first search with a precursor ion tolerance of 20 ppm was used to recalibrate raw data based on all peptide-spectrum-matches (PSMs) without filtering using hard score cut-offs. After recalibration, the data were searched with a precursor ion tolerance of 4.5 ppm, while chimeric spectra

were searched a second time using MaxQuant's "Second peptides" option to identify co-fragmented peptide precursors. "Match between runs" was used with an alignment time window of 30 min and a match time window of 1.1 min to transfer identifications between raw files of the same and neighbouring fractions ($\pm$ 1 fraction). Using the classical target-decoy approach with a concatenated database of reversed peptide sequences, data were filtered using a PSM and protein FDR of 1%. Protein groups were required to have at least one unique or razor peptide, with each razor peptide being used only once during the calculation of the protein FDR. Razor peptides are shared by multiple protein groups, which means they are not unique for a single protein group. In MaxQuant, their intensity contributes only to the protein group with the highest number of non-razor peptides. No score cut-offs were applied in addition to the target-decoy FDR.

2.7.2 giBAQ quantification

Unique and razor peptides were used for the quantification of proteins, discarding the unmodified counterparts of peptides harbouring oxidised methionines and acetylated protein N-termini. MaxQuant was set to calculate iBAQ intensities (Schwanhausser et al., 2011), however a modified version of the classical iBAQ approach termed giBAQ was instead used to estimate absolute quantities of gene groups as opposed to protein groups, excluding reverse and contaminant hits from the proteinGroups.txt output of MaxQuant. Unless otherwise stated, the CRC65 and CPTAC datasets were processed separately from one another. In order to ensure comparability with transcriptomics data, every protein group was assigned to a single gene group using the following heuristic

with four mapping groups (MGs): first, a mapping of Uniprot identifiers in the "Majority protein IDs" column to HGNC symbols was retrieved from Ensembl (Aken et al., 2016) using biomaRt v2.25.2 (Durinck et al., 2005; Durinck et al., 2009), collapsing isoforms to their parent Uniprot identifier. This resulted in two mapping groups, one for the Swiss-Prot entries (MG1) and one for the TrEMBL entries (MG2) contained in Uniprot, which were subsequently filtered to remove Swiss-Prot and TrEMBL identifiers mapping to more than one HGNC symbol. Each protein group was then represented by the first Swiss-Prot entry in the "Majority protein IDs" column, subsequently assigning the corresponding HGNC symbol from MG1 to the protein group. Afterwards, protein groups left without an HGNC symbol were represented by the first TrEMBL entry in the "Majority protein IDs" column and subsequently assigned the corresponding HGNC symbol from MG2. For protein groups still left without an HGNC symbol, Uniprot.ws v2.10.2 (Carlson, 2016) was used to retrieve a mapping of Uniprot identifiers in the "Majority protein IDs" column to HGNC symbols from Uniprot (MG3), collapsing isoforms to their parent Uniprot identifier. Protein groups left without an HGNC symbol were then represented by the first Uniprot entry in the "Majority protein IDs" column and subsequently assigned the first corresponding HGNC symbol from MG3. Afterwards, all protein groups still left without an HGNC symbol were assigned the first HGNC symbol from the "Gene names" column of the proteinGroups.txt output file (MG4). This heuristic guaranteed that the majority of HGNC-symbol-assignments were done using biomaRt (99% for both the CRC65 and CPTAC datasets), a system which – as opposed to Uniprot.ws or MaxQuant – could also be used to map HGNC symbols to the various identifiers encountered in the transcriptomics datasets described below, ensuring

maximum comparability of transcriptomics and proteomics data. MG1 was favoured over MG2 during assignment of HGNC symbols, since it was based on Swiss-Prot entries, which are manually curated. Selecting the first Swiss-Prot (MG1), TrEMBL (MG2) or Uniprot (MG3) entry in the "Majority protein IDs" column to represent a given protein group during HGNC-symbol-assignment in their respective mapping group further made sure that these assignments were always based on the identifier corresponding to the protein with the highest number peptides in the respective protein group. It is worth noting that 97% and 96% of these heuristic HGNC-symbol assignments for the CRC65 and CPTAC datasets were identical to the corresponding values in the "Gene names" column of the proteinGroups.txt output file, even though it only served as the source for 0.1% and 1% of the HGNC-symbols, respectively.

Next, the number of theoretical peptides per protein group used by MaxQuant for the calculation of iBAQ intensities was extracted as the ratio of raw intensities to iBAQ intensities, rounded to the nearest integer. As expected, these values were constant for each protein group across all samples in the two datasets and were equal to the values determined by *in silico* digestion (Protein Digestion Simulator, Monroe, 2015) of the fasta file of UniprotKB used in the database search by MaxQuant. Protein groups with only missing intensity values across one of the two datasets or no assigned HGNC-symbol were excluded from the respective dataset. For each gene group as defined by a common HGNC-symbol, the raw intensities of protein groups in this gene group were then summed up in a sample-wise manner and divided by the sum of the number of theoretical peptides for these proteins. For a gene group only containing proteins which do not share any theoretical peptides, this quantity termed giBAQ value will be equal

to the iBAQ value obtained for a hypothetical protein constructed from them, thereby normalising to the length of the portion of the gene accessible by mass spectrometry, rather than the individual protein lengths.

2.7.3 Kinobeads

Raw files from triplicate Kinobeads pulldowns were processed using a pipeline similar to the one employed for the analysis of full proteomes, adapting some of the parameters described hereafter. The fragment ion tolerance was set to 120 ppm for FTMS spectra instead of 20 ppm, since systematic fragment mass deviations were observed, which were linearly dependent on the m/z of the fragment ions in ppm space. Because this was likely due to problems during data acquisition, it had to be compensated for during the processing of raw data. Since Kinobeads pulldowns were not fractionated, "Match between runs" was used with the same parameters as described above to transfer identifications between all raw files.

2.7.4 LFQ quantification

Unique and razor peptides were used for quantification, discarding the unmodified counterparts of peptides harbouring oxidised methionines and acetylated protein N-termini as described above. MaxQuant was set to calculate LFQ intensities with an LFQ minimum ratio count of two. Fast LFQ was used to determine normalisation factors, with the minimum number of neighbours set to three, the average number of neighbours set to six and the "Stabilize large LFQ ratios" option enabled (Cox et al., 2014). The acquisition of MS2 spectra was not required for pairwise peptide intensity comparisons during the calculation of

LFQ intensities. Gene names associated with protein groups were remapped as described above for full proteome data. For each gene group and sample, the LFQ intensities of protein groups in this gene group were summed up to generate a gene-level quantification, which was used in all subsequent analyses.

2.8 Postprocessing of quantitative proteomics data

2.8.1 Full Proteomes

The CRC65 and CPTAC datasets were processed separately from one another. For the CPTAC dataset, the analysis was restricted to CPTAC sample IDs not removed in the original publication (Zhang et al., 2014a) due to duplication. Following common practice, giBAQ values were log2-transformed and median-centred. Since the CRC65 cell line panel contained two pairs of concordant cell lines (HDC-54/HDC-57 and LS 180/LS 174T, Bracht et al., 2010; Klijn et al., 2015; Medico et al., 2015), missing values in HDC-54 and LS 180 were replaced with measured values from HDC-57 and LS 174T, respectively. HDC-57 and LS 174T were subsequently excluded from the analysis. It is worth noting that the analysis did not involve just calculating the arithmetic mean of these cell line pairs, since this would prevent minimum-guided missing value imputation on the peptide level with raw-file-specific backgrounds, as described in its respective section.

2.8.2 Kinobeads

Since batch effects were observed between the three different biological replicates (see Figure 4.5.1), they were log2-transformed and median-centred separately from one another, followed by batch-effect removal using ComBat (Johnson et al., 2007). Briefly, the dataset was first restricted to gene groups

consistently measured between the different replicates and across the dataset, selecting gene groups with at least two out of three reported LFQ values per cell line in at least 23 cell lines. Next, all gene groups with one or fewer reported LFQ values per replicate were removed in order to eliminate noise and enable subsequent parametric batch adjustment using ComBat, implemented in the sva R package v3.18.0 (Leek and Storey, 2007), with MSI status as a covariate as described by Guinney and colleagues (2015). The median across all three replicates was used to summarise the LFQ values for each cell line, taking the median across all six replicates for the two pairs of concordant cell lines described above; HDC-54 was selected to represent HDC-54/HDC-57 and LS 180 was selected to represent LS 180/LS 174T.

2.9 Reanalysis of public mRNA datasets

Ten mRNA datasets were downloaded and processed as follows, automatically obtaining the required chip definition files (.cdf) from Bioconductor (Huber et al., 2015) for microarray datasets, if not stated otherwise (GSE36133 - Barretina et al., 2012; E-MTAB-783 - Garnett et al., 2012; E-MTAB-2706 - Klijn et al., 2015; GSE28567 - Loboda et al., 2011; GSE59857 - Medico et al., 2015; Supplementary Table S6 - Mouradov et al., 2014; platform codes "IlluminaGA_RNASeq" + "AgilentG4502A_07_3" - TCGA Network, 2012a; GSE8332 - Wagner et al., 2007; GSE24795 - Wilding et al., 2010, respectively). The same cell lines have inconsistent names in the different transcriptomics datasets, which is why the first pre-processing step always involved the mapping of all cell line names to a set of consistent sample names (see also Table 6.1.9).

2.9.1 GSE36133

GEOquery v2.35.6 (Davis and Meltzer, 2007) was used to download annotation information associated with GSE36133 and raw .CEL files, as well as the "Expression arrays samples info file" were obtained from the CCLE portal of the Broad Institute (http://portals.broadinstitute.org/ccle/data/browseData?conversationPropagation=begin, 31 December 2015). The dataset was restricted to cell lines from the large intestine, as well as C32 and Colo 741, which were derived from skin according to the "Expression arrays samples info file". The preproPara() function from the affyPara package v1.29.0 (Schmidberger et al., 2013) was then used with background correction through RMA (Irizarry et al., 2003), quantile normalisation (Bolstad et al., 2003) and probe summarisation through medianpolish of perfect match probes to calculate the classical RMA expression measure. Probes with only missing values across the dataset were excluded. Next, a mapping of AffyIDs to HGNC symbols was retrieved from Ensembl (Aken et al., 2016) using biomaRt (Durinck et al., 2005; Durinck et al., 2009) and filtered to remove AffyIDs mapping to more than one HGNC symbol. This mapping was then used to assign each AffyID to its corresponding HGNC symbol. AffyIDs, which could not be mapped this way, were excluded from the dataset. This was one of two datasets for which primary IDs (in this case AffyIDs) could not be mapped to GeneIDs first using the default annotation data downloaded from GEO or related sources. For consistency reasons, GeneIDs served as the basis for HGNC symbol assignment during the analysis of other transcriptomics datasets whenever possible. For GSE36133, this resulted in an expression matrix of 36,877 AffyIDs across 63 cell lines.

2.9.2 E-MTAB-783

Raw .CEL files, as well as the corresponding eSet.r and adf.txt files were obtained from ArrayExpress (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-783/files/, 31 December 2015), while the "gdsc_manova_input_w5.csv" file was downloaded from the GDSC portal in order to be able to map cell lines to tissues of origin (ftp://ftp.sanger.ac.uk/pub4/cancerrxgene/releases/release-5.0/, 12 March 2016). The dataset was restricted to cell lines from the large intestine, as well as C32 and SW626, which were derived from skin and ovary, respectively, according to the tissue labels in the "gdsc_manova_input_w5.csv" file. The rest of the pre-processing was similar to GSE36133, with the only difference being that AffyIDs were first mapped to Ensembl GeneIDs using the adf.txt file, followed by the assignment of each GeneID to an HGNC symbol using biomaRt as described above. This resulted in an expression matrix of 20,106 AffyIDs across 37 cell lines.

2.9.3 E-MTAB-2706

Supplementary Table 1 was obtained from the original publication by Klijn et al., while RPKM data for coding genes were downloaded from the website accompanying the publication (http://research-pub.gene.com/KlijnEtAl2014/, 06 October 2015). Supplementary Table 1 was used to assign cell lines to their tissues of origin and the dataset was subsequently restricted to CRC cell lines as well as C32, which was derived from skin. RPKM values equal to zero were treated as missing values and GeneIDs only containing missing values were removed from the dataset. Afterwards, the data was quantile normalised and log2-transformed, followed by assigning each GeneID to an HGNC symbol using

biomaRt as described above. This resulted in an expression matrix of 20,965 GeneIDs across 56 cell lines.

2.9.4 GSE28567

Annotation information for GSE28567 was obtained through GEOquery (Davis and Meltzer, 2007) and the appropriate custom chip definition file (.cdf) and the corresponding raw .CEL files were downloaded from NCBI GEO (Edgar et al., 2002, 06 November 2015). The R package makecdfenv v1.46.0 (Irizarry et al., 2006) was used to construct an annotation package from the .cdf file, which was subsequently used to relate probes to locations on the microarray. The rest of the pre-processing was similar to GSE36133, with the only difference being that the custom probe IDs were first mapped to Ensembl GeneIDs using the annotation information from GEO, followed by assigning each GeneID to an HGNC symbol using biomaRt as described above. This resulted in an expression matrix of 33,749 probe IDs across 67 cell lines.

2.9.5 GSE59857

GEOquery was used to obtain a normalised expression set object and annotation information associated with GSE59857, non-normalised expression data was downloaded from NCBI GEO (04 November 2015) and further annotation information for Illumina expression arrays was retrieved from Bioconductor (Barbosa-Morais et al., 2010). The R packages limma v3.26.2 (Ritchie et al., 2007) and beadarray v2.19.1 (Dunning et al., 2007) were used for background correction, normalisation and transformation of raw data. First, an ExpressionSetIllumina was generated from the expression set object downloaded

from NCBI GEO, the normalised expression data of which was replaced with the non-normalised data. Subsequently, background correction was performed using the backgroundCorrect function from limma, estimating the parameters of the normal-exponential (normexp) convolution model adapted for microarrays other than Affymetrix arrays by Ritchie et al. (2007) using the maximum-likelihood estimator developed by Silver et al. (2009) with an offset of 10 in order to ensure that expression values were positive after background correction. Afterwards, the corrected expression data were quantile normalised using the normaliseIllumina function from the beadarray package, followed by log2-transformation of the dataset. This method was used in favour of the neqc method described by Shi and colleagues (Shi et al., 2010) after comparison of the distributions of expression values produced by the two algorithms (see section 3.2.1). The dataset was restricted to probes from coding regions, excluding negative control and bad quality probes, but keeping probes associated with GeneIDs contained in the single-sample predictor for the CMS subtypes published by Guinney et al. (Guinney et al., 2015) regardless of probe quality annotation. Ensembl GeneIDs were mapped to HGNC symbols using biomaRt as described above. This resulted in an expression matrix of 27,293 probe IDs across 155 cell lines.

2.9.6 Supplementary Table S6 from Mouradov et al.

Supplementary Table S6 was downloaded from the original publication by Mouradov et al. (2014, 05 August 2014), cell line RW2982 was excluded from the dataset as the only cell line not present in the CRC65 panel, RPKM values equal to zero were treated as missing values and the dataset was restricted to RefSeq IDs

containing at least one measured value. Subsequently, RPKM values were quantile normalised and log2-transformed, followed by re-mapping RefSeq IDs to HGNC symbols using biomaRt as described above. This resulted in an expression matrix of 29,973 RefSeq IDs across 12 cell lines.

2.9.7 Platform code "IlluminaGA_RNASeq" from TCGA

The TCGAbiolinks R package v1.0.4 (Colaprico et al., 2016) was used to obtain and prepare level 3 data associated with COAD/READ tumors and platform code "IlluminaGA_RNASeq", only keeping samples for which proteomics data from the CPTAC study on CRC (Zhang et al., 2014a) were also present and which were not excluded from this dataset due to duplication as described above (16 December 2015). This dataset was favoured over level 3 data associated with platform code "IlluminaGA_RNASeqV2" for reasons of consistency, since these were normalised with the RSEM algorithm (Li and Dewey, 2011) instead of the classical RPKM method (Mortazavi et al., 2008). Duplicate rows were removed from the RPKM expression matrix, TCGA barcodes were re-mapped to the sample names described in Table 6.1.9, RPKM values equal to zero were treated as missing values and the dataset was restricted to GeneIDs containing at least one measured value, as described above. Afterwards, the dataset was log2-transformed and median-centred, followed by re-mapping of GeneIDs to HGNC symbols using biomaRt as described above. This resulted in an expression matrix of 19,670 GeneIDs across 87 tumours.

2.9.8 Platform code "AgilentG4502A_07_3" from TCGA

The TCGAbiolinks R package v1.0.4 (Colaprico et al., 2016) was again used to obtain and prepare level 1 data and MAGE-TAB files associated with COAD/READ tumors and platform code "*AgilentG4502A_07_3*", for which proteomics data from the CPTAC study on CRC was present, as well (Zhang et al., 2014a, 25 January 2016). The corresponding platform design file (.adf files) was downloaded from the TCGA portal at https://tcga-data.nci.nih.gov/docs/publications/tcga/platformdesign.html (26 January 2016). The R package limma v3.26.2 was used for the processing of txt files generated from microarray scans using Agilent's Feature Extraction Software (TCGA Network, 2011). Briefly, an RGList was generated, followed by annotation of the probes using the platform design file. Subsequently, background correction was performed using the backgroundCorrect function from limma with the method described by Edwards et al. (2003) and an offset of 50, since it produced the most stable results as judged by inspection of MA-plots following within-array loess normalisation with standard parameters (Smyth and Speed, 2003) for several background correction methods (see section 3.2.1). Afterwards, between-array normalisation using the "Gquantile" method implemented in the normalizeBetweenArrays function of limma was carried out under the assumption that the signal distributions of the green channels across all arrays should be the same, since the green channel always contained the Universal Human Reference RNA (TCGA Network, 2011, 2012a). The median of replicate probes was used to summarise their expression, probes without GeneID annotation were filtered out and GeneIDs were finally mapped to HGNC symbols

using biomaRt as described above. This resulted in an expression matrix of 66,728 Agilent IDs across 73 tumours.

2.9.9 GSE8332

GEOquery was used to download annotation information associated with GSE8332 and raw .CEL files were obtained from NCBI GEO (20 November 2015). The dataset was pre-processed similarly to GSE36133 and subsequently restricted to CRC cell lines based on Figure 1 of the original publication (Wagner et al., 2007), as well as C32 and Colo 741, which were derived from skin as described above. It is worth noting that the authors categorised Colo 741 as a CRC cell line, which contradicts the annotations found in the other mRNA datasets analysed thus far. Using the annotation information downloaded from GEO, AffyIDs were then annotated with GeneIDs, which were subsequently mapped to HGNC symbols using biomaRt as described above. This resulted in an expression matrix of 39,546 AffyIDs across 36 cell lines.

2.9.10 GSE24795

Raw data for GSE24795 were obtained (05 November 2015) and pre-processed as described for GSE8332, with the only difference that no samples needed to be omitted, since all cell lines included in the study were CRC cell lines. This resulted in an expression matrix of 39,546 AffyIDs across 30 cell lines.

2.9.11 Combination of transcriptomics datasets

The different transcriptomics datasets described above were aggregated into one expression matrix using a scheme similar to the one published by Guinney and colleagues (2015). As mentioned in the original publication, strong batch effects

were expected between the different transcriptomics datasets and a single probe or primary ID needed to be selected to represent the expression of a given gene in each dataset in order to avoid inconsistent mRNA quantification. But since the intention was to compare the abundance of as many mRNA species to as many protein species across as many samples as possible, respectively, outlier samples were not removed from each dataset separately and the list of quantifiable mRNA species was not restricted to the ones with the largest median absolute deviation (MAD). Moreover, a reference dataset was not selected *a priori*. Instead, similar consistency measures as the ones described by Guinney and colleagues were computed for each dataset and used for the selection of the reference dataset afterwards, thereby maximising the consistency between the different datasets as described below. All datasets were first reduced to a common set of 10,044 reference genes based on the HGNC-symbol-assignments described in the previous sections. Next, each of the reference genes was temporarily represented by the primary identifier (AffyID, Probe ID, etc.) with the largest MAD in each of the datasets (reference identifiers). For each dataset, the correlation between all primary identifiers and these reference identifiers was then calculated, generating ten correlation matrices with a number of rows equal to the number of primary identifiers in the respective dataset and 10,044 columns each; one for each reference identifier. For each dataset, this correlation matrix is equivalent to the collection of correlation vectors C of length $|G_{REF}|$ mentioned in the publication by Guinney and colleagues. For each of the reference genes and each dataset, the correlation of all correlation vectors associated with primary identifiers quantifying the respective reference gene to all other correlation vectors from the different datasets associated with primary identifiers

quantifying the same reference gene was then calculated. This resulted in a collection of 10,044 "gene-lists", one for each reference gene. Each of these "gene-lists" itself consisted of ten "reference-dataset-lists", one for each dataset. Each of these "reference-dataset-lists" itself contained a "dataset-list" of correlation matrices, one for each dataset. These correlation matrices finally contained the correlations of all correlation vectors associated with primary IDs quantifying the respective reference gene in the respective dataset against all corresponding correlation vectors associated with primary IDs quantifying the same reference gene in all other datasets. This enabled avoiding having to select a reference dataset *a priori* as described by Guinney et al., thereby allowing different selection criteria to be consulted before deciding on a reference dataset. In order to select a reference dataset, the following heuristic was used: for each potential reference dataset and each primary ID in it, a consistency metric *comet* was calculated by first determining which primary IDs in the other datasets quantifying the same reference gene show the highest "correlation of correlations" value with the respective primary ID in the potential reference dataset and subsequently summing up these "correlations of correlations" for each primary ID in the potential reference dataset. For each dataset and reference gene, this resulted in a vector of *comet* values, one for each primary ID quantifying the respective reference gene. The primary ID with the maximum *comet* value (*maxcomet* value) was selected to represent the respective reference gene in the respective dataset. These primary IDs were the ones resulting in the highest consistency between the different datasets if the respective dataset was selected as the reference. The distribution of these *maxcomet* values was then compared across all datasets in order to find the

reference dataset resulting in the highest consistency between the different datasets. Four of the ten datasets showed similarly promising distributions of *maxcomet* values (E-MTAB-2706 by Klijn et al., GSE28567 by Loboda et al., GSE8332 by Wagner et al. and GSE36133 by Barretina et al., see section 3.2.1) and GSE36133 by Barretina et al. was finally used as the reference dataset, since it was the biggest dataset in terms of both primary IDs and cell lines. In this reference dataset, primary IDs associated with *maxcomet* values were selected to represent their corresponding reference genes. For each of the other datasets, the primary ID with the maximum "correlation of correlations" value with the primary ID chosen to represent a given reference gene in the reference dataset was selected to represent the respective reference gene in the respective dataset. This resulted in ten datasets with 10,044 genes each, which were represented by a combination of primary IDs ensuring maximum consistency between the different datasets.

Next, the eight datasets quantifying transcription in cell lines were merged into one expression matrix (for the CRC65 dataset) and a second expression matrix was generated containing all ten datasets (for the CPTAC dataset). These two expression matrices were treated separately from one another during adjustment of batch effects in order to maximise the number of genes in each of them. First, these matrices were reduced to genes quantified in at least two samples per dataset, followed by the adjustment of batch effects between the different datasets using ComBat (Johnson et al., 2007) with MSI status as a covariate. Next, the expression matrix containing all ten datasets was restricted to samples also present in the CPTAC patient dataset. This resulted in two mRNA expression matrices, which – together with the giBAQ quantitation for cell lines

and patients described above – served as the basis for the multi-omics data integration strategy described below.

2.10 Multi-omics data integration strategy

2.10.1 Computation of protein/mRNA ratios

The median expression for each transcript and sample in the two mRNA expression matrices described above was calculated and divided by the giBAQ values of the corresponding proteins in the CRC65 and CPTAC dataset in a sample-wise manner, respectively. This resulted in two ratio matrices, one for the CRC65 dataset and one for the CPTAC dataset, which were basis for the analysis described in section 5.2.1. The median mRNA expression values were the basis for the results described in section 5.2.2.

2.10.2 Integration of transcriptomics and proteomics data

Stable protein/mRNA ratios across both datasets provoked the hypothesis that they could be treated as batch effects and therefore be adjusted using e.g. ComBat as described above. ComBat however shifts genes in each dataset to the overall mean and pooled variance across datasets, thereby altering the location and scale of each dataset. Here, the goal was to avoid altering the giBAQ expression values as they represent actual measurements of gene group abundance on the protein level. Therefore, MComBat (Stein et al., 2015) was used instead of ComBat, which allows the specification of a "gold-standard" batch towards which all other datasets are adjusted and which is not changed. Stein and colleagues published the R code for MComBat alongside their manuscript at https://github.com/SteinCK/M-ComBat/blob/master/M-ComBat%20R%20Script, however this

script is setup to stop if the combined expression matrix used as input contains any number of missing values. This script was modified as part of this project to instead work similarly to the original ComBat function implemented in the sva R package v3.18.0 with respect to missing values. With this modified MComBat function at hand, transcriptomics and proteomics data for the CRC65 cell line and CPTAC patient datasets were integrated, keeping the two datasets separate from one another. After merging proteomics and transcriptomics data for each dataset separately, these matrices were reduced to gene groups quantified in at least two samples on both the protein and transcript level, followed by the adjustment of batch effects between proteomics and transcriptomics data using the aforementioned modified MComBat function without MSI status as a covariate. The resulting adjusted expression matrices were the basis for Figure 5.2.2 and Figure 5.2.3, as well as for mRNA-guided missing value imputation.

In order to produce Figure 5.2.3, all pairwise correlations between protein and mRNA samples were first calculated for both datasets. This resulted in two correlation matrices, one for the CRC65 dataset with 59 rows (protein samples) and 294 columns (mRNA samples) and one for CPTAC dataset with 89 rows (protein samples) and 160 columns (mRNA samples), respectively. In order to restrict these matrices to mRNA samples showing an above-average correlation with their cognate protein measurements, these matrices were standardised in a column-wise manner so that all correlations corresponding to a given mRNA sample had a mean of zero and a standard deviation of one. For each protein sample, mRNA samples were restricted to those corresponding to the same cell line or patient, and whose standardised correlation with the respective protein sample was at least two standard deviations bigger than the mean across all

correlations corresponding to the same protein sample. All other mRNA samples were excluded from the non-standardised correlation matrices. Protein samples lacking cognate mRNA samples following the aforementioned filtering were excluded as well. This resulted in the exclusion of 22 and 35 mRNA samples (7% & 22%, respectively), as well as 8 and 19 protein samples (13% & 21%, respectively) from the non-standardised correlation matrices of the CRC65 and CPTAC datasets, respectively. These reduced correlation matrices were the basis for Figure 5.2.3. Rows in these heat maps were ordered using hierarchical clustering of the mRNA samples with "1-Pearson correlation" as the distance metric and "ward.D2" as the agglomeration method (Murtagh and Legendre, 2014). Columns on the other hand were ordered according to the occurrence of their cognate mRNA samples in the rows of the correlation matrices following reordering of rows.

2.10.3 mRNA-guided missing value imputation

The increase in the overall correlation between transcriptomics and proteomics data after adjusting for batch effects between them enabled linear prediction of protein abundance from adjusted mRNA abundance. In order to do so, the CRC65 and CPTAC datasets had to be processed separately from one another. For each protein sample and all of its cognate mRNA samples, all pairs of protein and mRNA measurements corresponding to the same gene group were assembled, subsequently modelling protein abundance as a function of adjusted mRNA abundance using a single linear model. For each cognate mRNA sample, the corresponding linear model was then used to predict the protein abundance of gene groups with missing values in the respective protein sample. For each gene

group with missing values in the respective protein sample, the mean across all predictions based on cognate mRNA samples was then used to impute the giBAQ value of the corresponding gene group. Imputed giBAQ values smaller than zero were treated as missing values.

2.10.4 Minimum-guided missing value imputation

Since not all missing values in the proteomics datasets could be imputed using mRNA-guided missing value imputation, a new imputation method termed "minimum-guided missing value imputation" on the protein level was implemented. This method ensured that no imputed value for a given protein was bigger than a measured value for the same protein. In order to do so, missing values were sampled with replacement from the distribution of measured values, truncated to the right at the minimum for the respective protein. This was done across the entire dataset, which is why each dataset needed to be normalised and handled separately. For each protein with one or more missing values across the dataset, the minimum expression value was first determined. In a protein-wise fashion, missing values were subsequently replaced with values, which were sampled from the truncated distribution of measured values as described above. This resulted in a much more favourable behaviour with reduced bimodality of the intensity distributions of each dataset if applied after mRNA-guided missing value imputation. Imputation on the protein level however should not be favoured over imputation on the peptide level, since aggregating peptide-level information to form protein expression values already represents a form of implicit missing value imputation, as discussed by Lazar et al. (2016). Therefore, the concept of minimum-guided missing value imputation was transferred to the

peptide level, only imputing missing values for peptides if the corresponding protein abundance was missing as well. With a stable protein-group-to-HGNC-symbol mapping at hand (see above), MaxQuant's relational database output was used to map each peptide and its intensity in the evidence.txt output file to the protein group and by extension also to the gene group to the quantification of which the respective peptide intensity contributed. Since razor peptides per definition map to more than one protein group, their intensity was only mapped to those protein and gene groups for the quantification of which MaxQuant actually used them. For each protein with one or more missing values across the dataset, the sample with the lowest expression of this protein was then found, subsequently identifying the peptides contributing to this value. For each missing value in each of these reference peptides, a window of 1.1 min – centred at the recalibrated retention time of the respective reference peptide in the matching fractions – was then searched for peptides with a lower intensity than the reference peptide. From the intensities of these low-intensity peptides, the missing values for the corresponding reference peptide were then sampled with replacement. All imputed peptide intensities corresponding to a given protein with missing values were subsequently summed up in a sample-wise fashion to form protein intensities and calculate giBAQ values as described above. Missing values in the Kinobeads dataset were imputed using minimum-guided missing value imputation on the protein level.

2.10.5 Combination of the CRC65 and CPTAC datasets

After missing value imputation, the CRC65 cell line and CPTAC patient datasets had to be combined into one expression matrix in order to be able to identify

integrated proteomic subtypes of CRC consisting of cell lines and patients. However, the applied imputation might influence downstream analyses, since some of the proteomics samples had more missing values than others and because transcriptomics data were not available for all samples. Missing values in samples for which data on the mRNA level was not available were imputed using minimum-guided missing value imputation only, possibly generating outlier samples, which will distort any unsupervised clustering. Therefore, outlier samples were first removed from both datasets separately using arrayQualityMetrics (Kauffmann et al., 2009), only keeping samples not marked as an outlier by any of the outlier detection methods performed ("Distance between arrays", "Boxplots" and "MA plots"). Afterwards, the two expression matrices were merged into one joined matrix. Finally, batch effects between the two datasets were adjusted using ComBat (Johnson et al., 2007) with MSI status as a covariate as described above.

2.11 Identification and characterisation of Full Proteome & Kinobeads Subtypes

2.11.1 Identification of subtypes

After combining the full proteome datasets into one expression matrix and following the removal of batch effects between them, consensus clustering (Monti et al., 2003) as implemented in the ConsensusClusterPlus R package v1.24.0 (Wilkerson and Hayes, 2010) was used to determine Full Proteome Subtypes (FPSs) and Kinobeads Subtypes (KSs) of CRC. The method described below was in many aspects similar to the method used by Zhang et al. (2014a), but made use of a different agglomeration method, while being more stringent

during the identification of core samples. The main difference however was that the entire expression matrix was used for the discovery of FPSs, only restricting the Kinobeads data to kinases during the identification of KSs. Hierarchical clustering was used to assign samples to one of k=4 clusters, with "1-Pearson correlation" as the distance metric, "ward.D2" as the agglomeration method and 1,000 resampling iterations using 80% of all gene groups and samples, while varying k from two to eight clusters during exploratory data analysis. Four clusters were chosen based on the four known Consensus Molecular Subtypes published by Guinney et al. (2015). Clusters with less than six members were removed, followed by the calculation of item-consensus values and silhouette widths (Rousseeuw, 1987) for each sample using the calcICL function from the ConsenusClusterPlus R package v1.24.0 and the silhouette function of the R package cluster v2.0.3, respectively. Being more stringent than Zhang et al., core samples were defined as samples with a positive silhouette width, provided that they also show the highest item-consensus with the cluster they were finally assigned to. Clusters only consisting of core samples for which no mRNA data were available were excluded in order to avoid artefacts due to missing value imputation. The columns in the final heat maps were restricted to core samples, which were ordered according to the final consensus tree. This method was also used to cluster the reduced mRNA expression matrix described in section 3.2.1 after the exclusion of cell lines not in the CRC65 dataset and transcripts not in the CMS classifier by Guinney et al. (2015), in order to produce Figure 3.2.7.

2.11.2 Discovery of differentially expressed gene groups

Significance analysis of microarrays (SAM, Tusher et al., 2001) as implemented in the samr R package v2.0 was used in order to discover gene groups differentially expressed between the different subtypes, restricting the analysis to core samples while using 100 permutations, the Wilcoxon test statistic and a target FDR of 0.01 (1%). The rows in the final heat maps were ordered according to a similar metric, which was intended to visualise which gene groups were higher or lower expressed in one subtype compared to the remaining subtypes. First, the expression matrix was standardised in a row-wise fashion so that each gene group had a mean of zero and a standard deviation of one. For each gene group, all pairwise combinations of two-sided Wilcoxon test statistics between each subtype and the remaining subtypes were then calculated. Each gene group was then assigned to the subtype with the lowest p-value. For each subtype, gene groups assigned to it were then ordered in increasing order of their summed expression across all samples in this subtype. Gene groups with higher expression in the respective subtype were followed by gene groups with lower expression in the respective subtype. These blocks of gene groups were then sorted according to the order of the subtypes they were assigned to in the final consensus tree, creating the checkerboard-like pattern seen in Figure 5.3.1 and Figure 5.4.1.

2.11.3 Functional annotation enrichment and clustering

MetaCore v6.26.68498 from Thomson Reuters was used to perform enrichment analyses of functional annotations for the gene groups differentially expressed between the different FPSs. In order to be more stringent, the background

MetaCore uses to calculate its p-values for the enrichment of functional annotations using the hypergeometric test was set to the list of network objects defined by the HGNC symbols detected in at least two samples in both the CRC65 cell line and CPTAC patient datasets. The standardised mean difference between subtype-specific gene group expression and overall mean expression as computed using SAM served as the raw input for the enrichment analysis, which was further filtered to remove values between the 2.5% and 97.5% quantiles of these values across all permutations, in order to restrict the analysis to gene groups with large class contrasts. This filtered matrix was uploaded to https://portal.genego.com/, followed by the calculation of enrichment analyses with respect to "Pathway Maps", "Process Networks", "GO Processes", "GO Molecular Functions" and "GO Localizations" for up- and down-regulated gene groups, respectively. The entire table of annotations for each of these functional annotation enrichments was exported and further processed in R. For "Pathway Maps" and "Process Networks", only annotations with a minimum FDR≤0.05 (5%) as provided by MetaCore were imported, the tables of annotations enriched in up- and down-regulated gene groups from any subtype were merged into a single table, rodent-specific annotations were excluded and finally barplots of $-\log_{10}$(FDR) values for up- and down-regulated gene groups were assembled into a single figure. For each of the different Gene Ontology (GO) categories, only annotations with a minimum FDR≤0.05 (5%) as provided by MetaCore were imported, as well. Subsequently, tables for the different GO categories were combined while keeping GO-terms enriched in up- and down-regulated gene groups specific to a certain subtype separate from one another. This resulted in six different tables; two for each subtype containing GO-terms associated with

up- and down-regulated gene groups, respectively. Afterwards, GO-term names were re-mapped to GO Accessions using a mapping downloaded from AmiGO (Carbon et al., 2009) using GOOSE (http://amigo.geneontology.org/goose, 19 May 2016) and GO Accessions with an FDR of more than 0.05 (5%) were removed from the corresponding table. Finally, these tables were sequentially uploaded to REViGO (http://revigo.irb.hr/, Supek et al., 2011) for summarisation and visualisation of significant GO-terms using treemaps, with an allowed similarity of 0.7, interpreting associated numbers as p-values, restricting the database to *Homo sapiens* and employing the SimRel semantic similarity measure (Schlicker et al., 2006, 01 July 2016).

2.12 Reanalysis of public Drug Sensitivity datasets

2.12.1 Pre-processing

The same cell lines and drugs have inconsistent names in the different drug sensitivity datasets, which is why the first pre-processing step always involved the mapping of all cell line and drug names to a set of consistent sample and drug names, respectively (see also Table 6.1.8 and Table 6.1.9). For reasons of transparency, each drug was annotated with the corresponding dataset from which it was derived. Concordant cell lines were treated as replicates during model fitting, which was kept as consistent as possible between the different datasets. In essence, a relative response measure R was always modelled as a function of the final drug concentration on the assay plate as described below.

2.12.2 GDSC

Raw dose-response data (gdsc_drug_sensitivity_raw_data_w5.csv), annotation information (gdsc_tissue_output_w5.csv) and drug concentrations (gdsc_compounds_conc_w5.csv) were obtained from ftp://ftp.sanger.ac.uk/pub4/cancerrxgene/releases/release-5.0/ (04 September 2015) prior to the publication by Iorio et al. (2016), which is why all analyses were based on release v5.0. First, the raw data was annotated with the corresponding drug names and "fold-dilution" information, as well as with data on the maximum screening concentration used for each drug. This allowed the calculation of the actual concentrations at which each compound was used in the drug screen. The mean background signal for each compound and cell line was then calculated and subtracted from each corresponding raw signal, setting negative values to zero afterwards. Next, the mean control signal for each compound and cell line was determined, subsequently dividing each corresponding background-corrected raw signal by its respective control signal, which resulted in normalised and background-corrected viability data. Finally, for each compound, compound concentration and cell line, the relative response measures R were then calculated as the mean of these values, restricting the dataset to cell lines from the large intestine, as well as C32 and SW626.

2.12.3 CCLE

Supplementary Table 11 from the original publication by Barretina et al. (2012) was downloaded and restricted to cell lines from the large intestine, as well as C32 and Colo 741. Drugs for which the authors were not able to fit more than one sigmoid dose-response model among the cell lines also present in the CRC65 cell

line panel as indicated in the FitType column were filtered out, removing dose-response data for both “Nutlin-3” and “PHA-665752”. Subsequently, raw activity values A (“Activity Data (median)”) were converted back to relative response measures R=T/U=1+A/100 (T represents the response for the compound-treated well, U represents the median response of the untreated wells across the plate) by reverting the corresponding equation in the addendum to the original publication, in order to make the CCLE dataset more comparable to the other drug sensitivity datasets.

2.12.4 CTRP

The expanded CTRP v2.0 dataset was obtained from the FTP server of the CTD[2] data portal at ftp://caftpd.nci.nih.gov/pub/dcc_ctd2/Broad/CTRPv2.0_2015_ctd2_ExpandedDataset/ (23 December 2015). The results described here are therefore partially based upon data generated by the Cancer Target Discovery and Development (CTD2) Network (https://ctd2.nci.nih.gov/dataPortal/) established by the National Cancer Institute’s Office of Cancer Genomics. First, the raw data was annotated with the corresponding compound and cell line names, as well as information on the tissue of origin for the cell lines used in the screen. Based on these annotations, the dataset was restricted to cell lines from the large intestine, as well as C32 and Colo 741. Viability data annotated with the same cell line name, compound concentration and compound name were treated as replicates during model fitting. The weighted percent-viability with error propagation (“cpd_avg_pv”) was used as a relative response measure R.

2.12.5 Cetuximab

Supplementary Data 1 from the original publication by Medico et al. (2015, 03 January 2016) were downloaded, converting the percentage of growth inhibition P to a relative response measure R=1-P/100 in order to make the data comparable to the other drug sensitivity datasets.

2.12.6 Dose-response modelling and parameter extraction

Using the drc package v2.5-12 in R (Ritz et al., 2015), the classical symmetric four-parameter log-logistic model was fitted to each drug in each dataset. The four parameters c="Lower Limit" (i.e. lower limit of the relative response as the drug concentration approaches infinity), d="Upper Limit" (i.e. upper limit of the relative response as the drug concentration approaches zero), b="Slope" (Hill slope, i.e. slope around the inflection point of the dose-response curve), and e="ED50" (i.e. the dose required to reduce the relative response to c+(d-c)/2) were estimated from the data and subsequently extracted from the fitted model. The computeAUC function from the drexplorer package v1.1.2 (Tong et al., 2015) was modified to accept the output of the drm function from the drc R package as an input, using it to extract the standardised area under the dose-response curve (AUC) for each of the fitted models across the tested dose-range. Here, the AUC across the tested dose-range was defined as the area under the dose-response curve between zero and one, divided by the area under y=1 from the lowest to the highest concentration tested. It is worth noting that the relative response measures described above varied between approximately zero and one, enabling unified model fitting and parameter extraction.

2.13 Modelling of Drug Sensitivity

In order to find proteins and kinases associated with drug sensitivity or resistance, elastic net regression (Zou and Hastie, 2005) was used to model drug sensitivity as a function of protein and kinase profile, respectively.

If $y_i \in \mathbb{R}$ are drug responses expressed as AUC for $i=1,\ldots,N$ cell lines and $x_i \in \mathbb{R}^p$ are standardised measurements for $i=1,\ldots,N$ cell lines across $j=1,\ldots,p$ proteins or kinases with zero mean and unit variance forming N observation pairs (x_i, y_i), then the cost function of the elastic net – minimised using glmnet v2.0-5 (Friedman et al., 2010; Simon et al., 2011) – is given by

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - {x_i}^T\beta)^2 + \lambda\left[(1-\alpha)\frac{1}{2}\|\beta\|_{l_2}^2 + \alpha\|\beta\|_{l_1}\right]$$

Leave-one-out cross-validation over a grid of 100 equally-spaced values of α with $0 \leq \alpha \leq 1$ by 100 values of λ automatically selected by the algorithm for each α was used to find the combination of α and λ resulting in the lowest mean-squared error between the fitted mean and the response. A bootstrapping approach similar to the ones used before was then employed in order to be able to sort the different proteins or kinases according to their importance in predicting the response to a given drug (Barretina et al., 2012; Basu et al., 2013; Garnett et al., 2012; Gholami et al., 2013; Iorio et al., 2016). Briefly, 1000 bootstrap datasets were generated by sampling the above-mentioned observation pairs with replacement, subsequently solving the elastic net optimisation problem for each of these datasets using the optimal values of α and λ. This resulted in a $1000 \times (p+1)$ matrix of regression coefficients capturing the solutions for the bootstrap datasets as rows and the regression coefficients

as columns. The intercept column $\beta_{int}^{BS} \in \mathbb{R}^{1000,1}$ was stored separately from the remainder of the matrix, which reduced it to a $1000 \times p$ matrix of coefficients $\beta_{coef}^{BS} \in \mathbb{R}^{1000,p}$ for j=1,...,p proteins or kinases across 1000 bootstrap datasets. Each column of β_{coef}^{BS} was subsequently multiplied by the standard deviation of the corresponding protein or kinase across all x_i, as well as by its corresponding weight ω described in Equation 11 in the publication by Moghaddas Gholami et al. (2013) if $\omega \in \{-1,1\}$ or by zero if $\omega \notin \{-1,1\}$. This resulted in a $1000 \times p$ matrix of effect sizes β_{eff}^{BS} analogous to Equation 12 in the publication by Moghaddas Gholami et al. It is worth stressing that columns of effect sizes corresponding to proteins or kinases whose coefficients did not have a consistent sign across all 1000 bootstrap solutions were set to zero through the multiplication with the weight ω. For each column in β_{int}^{BS}, β_{coef}^{BS} and β_{eff}^{BS}, the mean across all bootstrap solutions was then calculated to form β_{int}^{mean}, β_{coef}^{mean} and β_{eff}^{mean}, respectively, while also counting the number of times the absolute effect size of a given protein or kinase was bigger than zero to form $F_{eff} \in \mathbb{N}^p$. For each protein or kinase j=1,...,p, it was then determined whether (a) $F_{eff,j}$ was bigger than or equal to the 9th decile of F_{eff}, while also checking whether $\beta_{eff,j}^{mean}$ was either (b_1) bigger than or equal to the 19th ventile of β_{eff}^{mean} or (b_2) smaller than or equal to the 1st ventile of β_{eff}^{mean}. Only proteins or kinases for which both condition (a) and one of condition (b_1) or (b_2) were true were considered to be robust predictive markers of drug response, while all proteins or kinases j=1,...,p with $\left|\beta_{eff,j}^{mean}\right| > 0$ were designated as being associated with drug response. For each drug, all N available observation pairs (x_i, y_i) were used to train the model with three different input matrices each:

1) The combined full proteome expression matrix after batch adjustment
2) All proteins in the Kinobeads expression matrix after missing value imputation
3) All kinases in the Kinobeads expression matrix after missing value imputation

For each drug, the corresponding β_{int}^{mean} and β_{coef}^{mean} from scenario 1) were then used to predict the drug response $\hat{y}_i$ of each cell line and patient in the combined dataset based on the corresponding full proteome expression profile x_i using the following formula:

$$\hat{y}_i = \beta_{int}^{mean} + \sum_{j=1}^{p} \beta_{coef,j}^{mean} \times x_{i,j}$$

These predictions were subsequently standardised to have zero mean and unit variance. Moreover, $\hat{y}_i$ was also predicted using β_{int}^{mean} and β_{coef}^{mean} from scenario 3), subsequently standardising the predictions to have zero mean and unit variance in order to generate kinase-centric drug sensitivity hypotheses. Cell lines with a standardised AUC bigger than zero were designated as "resistant", while cell lines with a standardised AUC smaller than zero were designated as "sensitive". The type of figure used to visualise the result of an elastic net regression is termed "effect-size heat map". In effect-size heat maps, up to five of the most robust predictive markers of drug resistance and sensitivity with respect to β_{eff}^{mean} were displayed, respectively, while restricting the columns to at most the top 10 and bottom 10 cell lines with respect to their drug response as measured by AUC. In such a figure, these top and bottom ten cell lines are shown as columns ordered from left to right in increasing order of drug resistance, while

the predictors (proteins or kinases) with respect to absolute effect-size are shown as rows. In the bar plot to the left, predictors associated with drug resistance are visualised by yellow bars, while predictors associated with drug sensitivity are shown as dark blue bars in increasing or decreasing order of absolute effect-size from top to bottom, respectively. The heat map itself shows standardized expression values (z-scores) of the predictors with blue=”low expression” and red=”high expression”. Below the heat map, a heat strip visualises the AUC of the drug in question from dark-blue=”low AUC” to yellow=”high AUC” (see Figure 1.4.1 for an example of an effect-size heat map).

2.14 Cell viability assays

With categorical drug sensitivity predictions at hand, *in vitro* cell viability assays were performed in order to test these *in silico* predictions. Assays were performed as described previously (Garnett et al., 2012), with minor modifications. Briefly, the optimal seeding density for six cell lines was determined in dilution experiments mimicking the actual assay setup in technical quadruplicates. On day zero, cells were seeded at 500, 1,000, 2,000, 4,000 and 8,000 cells per well of a 96-well plate in 100 µL culture medium, filling the wells forming the outer rim of the plate with medium only and the interstitial space between the wells with PBS to avoid edge effects. Plates were incubated over night under standard culture conditions in order for the cells to attach. On day one, 50 µL of fresh medium containing 1% DMSO were added to all wells and the cells were incubated for additional 72 h under standard culture conditions. On day four, all wells were inspected using standard cell culture microscopes and the seeding density resulting in approximately 70% confluency was recorded or

extrapolated based on all measurements for each cell line (C10=4,000, CaCo-2=2,000, CC07=4,000, HDC-143=12,000, SK-CO-1=8,000, T84=11,000). Cell viability assays were carried out in technical triplicates in order to generate 10-point dose-response curves for drugs with the resistance of which MERTK was associated. For each drug, three cell lines predicted to be resistant and three cell lines predicted to be sensitive towards the respective drug (see above) were seeded in 100 µL culture medium at their optimal seeding density on day zero. Following an over night incubation under standard culture conditions, 50 µL of fresh medium containing either 1% DMSO or the respective drug in a 9-point two-fold dilution series in 1% DMSO were added to the corresponding wells. This resulted in a final DMSO concentration of 0.33%, while the highest final drug concentration was 10 µM for UNC569 (Merck), as well as 0.5 µM for PD-0325901 and 5 µM for RDEA119 (all from Cambridge Bioscience), respectively. It should be noted that the highest final drug concentration for PD-0325901 was twice as high as in the publication by Garnett et al. (2012) in order to better capture the expected dynamic range of drug sensitivities. For drug co-treatments, the respective compounds were combined at constant ratios over the entire concentration range used for the single agent treatments, keeping the final DMSO concentration in the assay at 0.33%. After 72 h of incubation time under standard culture conditions, the medium was removed from the cells and replaced with 150 µL of fresh culture medium containing 1 µM of Hoechst 33342 (Thermo Fisher #H3570). Cells were incubated under standard culture conditions for 1 h before they were imaged using an In Cell Analyzer 6000 automated confocal microscope (GE Healthcare), which was adjusted to the dimensions of the 96-well plates (Corning CellBind #3300) in order to ensure proper focus (SAC collar

setting of 1.16 mm) and accurate placement of fields of view (FOVs). Four FOVs per well arranged in the corners of a 1 mm^2 square centred on the well were acquired using the CFI Plan Fluor ELWD 20xC objective (Nikon) and the 405 nm laser to excite Hoechst 33342 fluorescence, which was captured by the CMOS camera without binning after passing through a 455/50 nm bandpass filter. Both the laser (10% power) and software autofocus (40 µM) was used, setting the exposure to 100 ms at 80% laser power with open aperture and imaging wells in a horizontal serpentine pattern to save acquisition time. The Hoechst 33342 channel of all images was subsequently analysed with Columbus v2.6.0 (PerkinElmer) using two of the built-in algorithms "B" and "C" to automatically segment nuclei, which were adapted to dispersed (C10, CC07, SK-CO-1 and CaCo-2) or clustered (remainder of the cell lines) cell growth characteristics, respectively. The area was set to bigger than 60 $µM^2$ and the split factor was seven for both methods, while the common threshold was set to 0.4 and 0.01, the individual threshold was 0.35 and 0.3 and the contrast was 0.1 and -0.3 for algorithm "B" and "C", respectively. Only nuclei completely contained in the FOV were considered while also restricting the analysis to nuclei with a mean Hoechst 33342 intensity of less than or equal to 8,000 and 10,000, as well as an area of more than or equal to 150 $µM^2$ and 100 $µM^2$ for algorithm "B" and "C", respectively. For each well, the number of nuclei satisfying these conditions in all four FOVs were counted and normalised to the mean of the corresponding DMSO controls. Afterwards, dose-response modelling and parameter extraction was performed as described above.

2.15 Western blots

Total cell lysates were diluted to 1 mg/mL protein concentration and 1x final sample buffer concentration with 4x sample buffer (70 mM Tris pH 6.8, 5% v/v 2-mercaptoethanol, 40% v/v glycerol, 3% w/v SDS, 0.05% w/v bromophenol blue) and stored at -80°C until further use. Samples were thawed quickly, heated to 95°C for 10 min in a heat block and subsequently separated using either custom 7-10% polyacrylamide gels in Protean II Xi Cells (Bio-Rad) and a Tris/HCl discontinuous buffer system at 100 V constant for 14 h (100 µg per sample; EPHA4, ABL1 and ERK1/2) or using 4-12% NuPAGE Bis-Tris midi gels in XCell4 SureLock Midi-Cells (Thermo Fisher) and MOPS running buffer at 180 V constant for 45 min (14 µg per sample; MERTK and ACTB). Proteins were transferred from custom or NuPAGE gels to Hybond-P PVDF (Amersham) or nitrocellulose (iBlot Transfer Stack, Thermo Fisher) membranes using the Trans-Blot SD Semi-Dry Transfer Cell (Bio-Rad) at 20 V constant for 90 min or the iBlot Dry Blotting System (Thermo Fisher) at 20 V (program 3) for 7 min, respectively. PVDF or nitrocellulose membranes were subsequently blocked for 1 h with 5% nonfat dry milk or 5% BLOT-QuickBlocker (Merck) in TBST (20 mM Tris/HCl pH 7.5, 100 mM NaCl, 0.1% Tween 20), respectively. Afterwards, PVDF membranes were incubated with primary antibodies against EPHA4 (ab157588, 1:500, Abcam), ABL1 (OP20, 1:1000, Merck) and ERK1/2 (06-182, 1:1000, Merck), while nitrocellulose membranes were incubated with primary antibodies against MERTK (ab52968, 1:2000, Abcam) and ACTB (sc-47778, 1:1000, Santa Cruz) in their corresponding blocking buffers over night at 4°C, respectively. Following three 10 min washes with TBST, PVDF membranes were incubated with appropriate HRP-coupled secondary antibodies (donkey anti-mouse

#715-036-151 and donkey anti-rabbit #711-036-152, Jackson Immuno Research), while nitrocellulose membranes were incubated in the dark with appropriate fluorophore-coupled secondary antibodies (IRDye 800CW goat anti-rabbit IgG #926-32211 and IRDye 680RD goat anti-mouse IgG #925-68070) in their corresponding blocking buffers for 1 h at RT, respectively. After three 10 min washes with TBST and following the addition of SuperSignal West Dura Extended Duration Substrate (Thermo Fisher) to PVDF membranes, they were measured using the Fusion Capt Advance FX7 chemiluminescence imager (VILBER), while nitrocellulose membranes were measured using the Odyseey CLx (LI-COR). Densitometry of western blots was carried out using Fiji v2.0.0-rc-43/1.50i (Schindelin et al., 2012), expressing EPHA4 and ABL1 expression relative to the respective ERK1/2 signal, followed by dividing all expression values by the expression value of C10 and log2-transformation.

2.16 Immunohistochemistry

For each of the four control cell lines (RKO, CC07, HT55 and CaCo-2), one T-175 flask was grown under standard culture conditions in culture medium to approximately 70% confluency. After removing the medium, the cells were washed once with 10 mL ice-cold PBS (without Ca^{2+}/Mg^{2+}, Thermo Fisher), scraped into 2 mL 4% PFA in PBS, pelleted at 400 xg for 10 min and subsequently fixed over night at 4°C. The next day, the cells were again pelleted at 400 xg for 10 min, the supernatant was discarded and replaced with 2 mL 70% EtOH in PBS, followed by an incubation over night at 4°C before the cell pellets were embedded in paraffin by Marcus Green using standard procedures (research assistant in the Preclinical Validation Core, University of Oxford). Immunohistochemistry (IHC)

of QUASAR2 TMAs and FFPE control cell pellets was carried out by Helen Scott from the Department of Oncology at the University of Oxford as already described (Schlegel et al., 2013), with minor modifications. Stains quantifying MERTK expression were performed using the Bond-MAX automated IHC system (Leica Biosystems) at RT if not specified otherwise. After de-paraffinising slides using Bond Dewax Solution (AR9222) and following re-hydration in Bond Wash Solution (AR9590), heat induced epitope retrieval (HIER) was performed for 10 min at 100°C with epitope retrieval solution 1 (AR9961, pH 6). Using the Bond Polymer Refine Detection Kit (DS9800), endogenous peroxidase activity was blocked for 5 min with 3-4% (v/v) hydrogen peroxide, followed by incubating slides for 15 min with primary antibody against MERTK (ab52968, 1:1000, Abcam) diluted in Bond Primary Antibody Diluent (AR9352). Primary mouse antibodies were localised by incubation with Post Primary solution for 8 min (<10 μg/mL rabbit anti mouse IgG containing 10% v/v animal serum and 85.5 ppm 2-methyl-4-isothiazolin-3-one in TBS), followed by incubating slides with HRP-polymer (<25μg/mL anti-rabbit poly-HRP-IgG containing 10% v/v animal serum and 85.5 ppm 2-methyl-4-isothiazolin-3-one in TBS) for 8 min. The staining reaction was initiated by incubating slides with a mixture of DAB Part 1 (66 mM 3,3'-diaminobenzidine tetrahydrochloride hydrate in a stabilizer solution) and DAB Part B (≤0.1% v/v hydrogen peroxide in a stabilizer solution) for 10 min, followed by counterstaining nuclei with Hematoxylin solution (<0.1% Hematoxylin) for 5 min. Slides were washed with de-ionised water for 10 min, de-hydrated with an increasing alcohol series (30 s in 50%, 70%, absolute and absolute EtOH), cleared with two 30 s washes in 100% xylene and subsequently mounted using DPX mountant (Sigma-Aldrich). Finally, slides were scanned at 40

x magnification using a ScanScope (Aperio) and examined by trained pathologists from the Technical University of Munich (Julia Slotta-Huspenina and Moritz Jesinghaus).

2.17 Cox proportional hazards regression

In order to assess whether or not the membranous expression level of MERTK is a prognostic biomarker in CRC, the different outcome variables of the QUASAR2 trial were modelled as a function of MERTK expression using univariate and multivariate Cox proportional hazards regression (COXPH), restricted to the first 5 years of follow-up. First, univariate COXPH was carried out using the coxph function of the R package survival v2.38-3 (Therneau, 2015; Therneau and Grambsch, 2000) with default parameters and “Gender”, “Age”, “BMI”, "Treatment" (capecitabine $\pm$ bevacizumab), "Location" (right colon or left colon), "T" (from TNM staging system; T2/T3, T3 or T4), "N" (from TNM staging system; N0, N1 or N2) , "Stage" (II or III), "Dukes" (Dukes' staging system; B or C), "MSI" (microsatellite instability), "CIN" (chromosomal instability) and "MERTK" (membranous MERTK expression; $\leq$5% or $>$5%) as covariates. Significant ($p<0.05$) covariates from univariate COXPH apart from “Dukes” (just a different label for the “Stage” variable here) were then used as explanatory variables in multivariate COXPH. The stepAIC function from the R package MASS v7.3-45 (Venables and Ripley, 2002) was used to perform stepwise backward model selection starting with the full model and removing explanatory variables in order to minimise Akaike’s Information Criterion (AIC, Akaike, 1973). For each outcome variable, final models were then built using significant ($p<0.05$) covariates from the stepwise selected models. These final models all fulfilled the

proportional hazards assumption (data not shown). Results from the COXPH analyses can be found in Table 6.1.5 and Table 6.1.6.

2.18 Annotation information

Annotation information for the CRC65 and CPTAC datasets, as well as target annotations for drugs in the four drug sensitivity datasets (GDSC, CCLE, CTRP and cetuximab) was compiled from several sources (Supplementary Table 11 & GSE36133, Barretina et al., 2012; CTRP v2.0 expanded dataset, Basu et al., 2013; Supplementary Table 1, Bracht et al., 2010; Supplementary Table 3, De Sousa et al., 2013; COSMIC, Forbes et al., 2015; GDSC release 5.0 & E-MTAB-783, Garnett et al., 2012; cms_labels_public_all.txt, Guinney et al., 2015; GDSC release 5.0, Iorio et al., 2016; Figure 1 & E-MTAB-2706, Klijn et al., 2015; Table 1, Ku et al., 1999; Table S1, Liu and Zhang, 2016; GSE28567, Loboda et al., 2011; Supplementary Data 1 & 2 & GSE59857, Medico et al., 2015; Supplementary Table 1 & 6, Mouradov et al., 2014; CTRP v2.0 expanded dataset, Rees et al., 2016; Supplementary Table 2, Sadanandam et al., 2013; CTRP v2.0 expanded dataset, Seashore-Ludlow et al., 2015; Supplementary Table 1 & platform codes "IlluminaGA_RNASeq" & "AgilentG4502A_07_3", TCGA Network, 2012a; GSE8332, Wagner et al., 2007; Table 1, Wheeler et al., 1999; GSE24795, Wilding et al., 2010; Supplementary Table 1, Zhang et al., 2014a) and summarised in Table 6.1.8 and Table 6.1.9. Entries for cell lines in the columns “Sadanandam subtype”, “Marisa subtype”, “De Sousa E Melo subtype” and “Roepman subtype” were based on the publication by Medico and colleagues and were not restricted by an FDR filter. The “Goblet-like”, “Goblet” and “Enterocyte” subtype of the classifier by Sadanandam and colleagues were merged into the “Entererocyte/Goblet-like”

subtype, while “MSI-L” was recoded to “MSI stable”=TRUE and “MSI-H” was recoded to “MSI stable”=FALSE according to the publication by Liu and Zhang. Some of the other annotations were also recoded to fit the unified nomenclature. For drug targets, the following heuristic was applied: after mapping compounds to a consensus set of names, the annotation from the CTRP dataset was used as the first source of target information. Drugs for which this dataset lacked annotation were assigned targets based on the annotation from the GDSC dataset. Drugs the target information of which was not included in this dataset either were then assigned targets based on the annotation from the CCLE dataset. Drugs still left without target information were finally assigned labels based on the information contained in the “target_or_activity_of_compound” column of the “v20.meta.per_compound.txt” file from the expanded CTRP v2.0 dataset. Amendments to these mappings in order to reduce redundancy are documented in Table 6.1.8.

Chapter 3
Transcriptome and full proteome of the CRC65 cell line panel and the CPTAC patient dataset

3.1 Introduction

As mentioned in section 1.4, the first aim of this project was to establish and characterise a training and testing dataset suitable for pharmacoproteomic modelling in CRC. Ideally, this dataset should consist of appropriate cell-line-based model systems of CRC, as well as CRC tumours, so as to increase the clinical relevance of drug sensitivity predictions based on these models, but retaining the option to test some of the predictions *in vitro*. As described in Figure 1.4.1, this project integrated data on the CRC65 cell line panel with data on tumours, which were characterised as part of the TCGA/CPTAC CRC tumour profiling efforts (TCGA Network, 2012a; Zhang et al., 2014a), in order to transfer drug sensitivity information from cell lines to tumours. This chapter focuses on the characterisation of the CRC65 cell line panel and the CPTAC tumour dataset with respect to their transcriptomes and full proteomes. With the mRNA-based CMS of tumours from the CPTAC dataset known (Guinney et al., 2015), the first step was to combine multiple public mRNA datasets and determine the CMS of the CRC65 panel, explained in section 3.2. Mass-spectrometry-based bottom-up proteomics (see section 1.3) combined with giBAQ quantification of proteins (described in section 2.7.2) was then applied to characterise the full proteomes of the CRC65 cell line panel, because of its ability to identify and quantify large numbers of proteins from a single sample, which is detailed in section 3.3. This section also contains a re-analysis of the CPTAC tumour dataset in order to unify the data analysis workflow. As mentioned in section 1.4, the CMS subtype model

and the proteomic subtype model published by the CPTAC (Zhang et al., 2014a) only showed partial concordance, which is why this project strived to establish the basis for the discovery of integrated full proteome subtypes of CRC by combining data on cell lines and patients.

3.2 Transcriptome

3.2.1 Combination of multiple mRNA datasets

CRC cell lines were part of several large-scale studies, which generated transcriptomics data of non-treated cell lines growing under steady-state conditions to answer a variety of questions, which ranged from pharmacogenomics of cell line panels covering multiple entities (Barretina et al., 2012; Garnett et al., 2012) over transcriptomic resources (Klijn et al., 2015) to systems biology approaches examining epithelial-mesenchymal transition in colon cancer (Loboda et al., 2011). Other studies focused on molecular subtyping of CRC (Medico et al., 2015; Mouradov et al., 2014) systematically analysed apoptosis across several different tumour types (Wagner et al., 2007) or characterised differences in gene expression between CRC cell lines with or without microsatellite instability (Wilding et al., 2010). These eight publicly available mRNA datasets on cell lines, as well as two public transcriptomics datasets on CRC tumours published by TCGA (TCGA Network, 2012a) were downloaded, analysed from the data-level closest to raw data available (see section 2.9) and aggregated using a scheme similar to the one used by Guinney et al. (see section 2.9.11), which supported mRNA data generated using a variety of different technologies. The analysis of raw data was straightforward in most cases, with some datasets requiring special attention. The recommended

normalisation method for GSE59857 for example was the neqc method described by Shi et al. (2010), however simply quantile-normalising and log2-transforming the data resulted in a more favourable distribution of expression values, visualised in Figure 3.2.1.

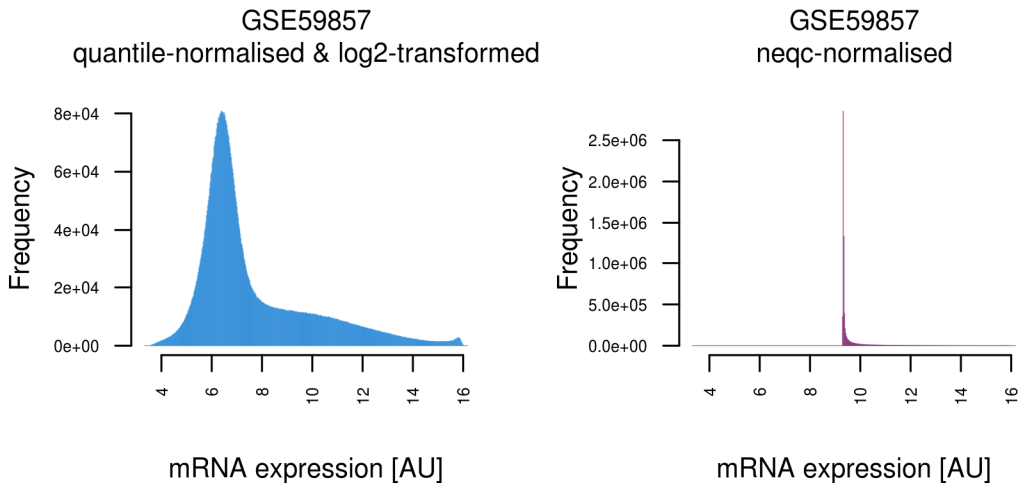


Figure 3.2.1

Exploration of normalisation options for GSE59857

Histograms visualising the effect of different normalisation methods on the distribution of mRNA expression values from the GSE59857 dataset. The same bandwidth was used for both histograms.

The Agilent microarray data generated as part of TCGAs molecular characterisation of CRC (TCGA Network, 2012a) could also be analysed in several different ways. In the original publication, the authors lowess-normalised the data and subsequently calculated log2-ratios of the Cy5 channel (containing the sample) and the Cy3 channel (containing the Universal Human Reference RNA), performing quality-control using a variety of different bioinformatics techniques (TCGA Network, 2011). However, they did not do any background correction or between-array-normalisation, even though their experimental design lent itself specifically to between-array-normalisation using the Gquantile method of the

normalizeBetweenArrays function of the R package limma v3.26.2. This method equalises the empirical distributions of the green (G) channels without changing the log2-ratios of the red (R) and green channels in each array. Since the green channel always contained the Universal Human Reference RNA, equalising its empirical distribution across all arrays and adjusting the red channels accordingly was the method of choice, as described in section 2.9.8. Background correction on the other hand required the exploration of several different methods before MA-plots could be used to evaluate their performance, as depicted in Figure 3.2.2.

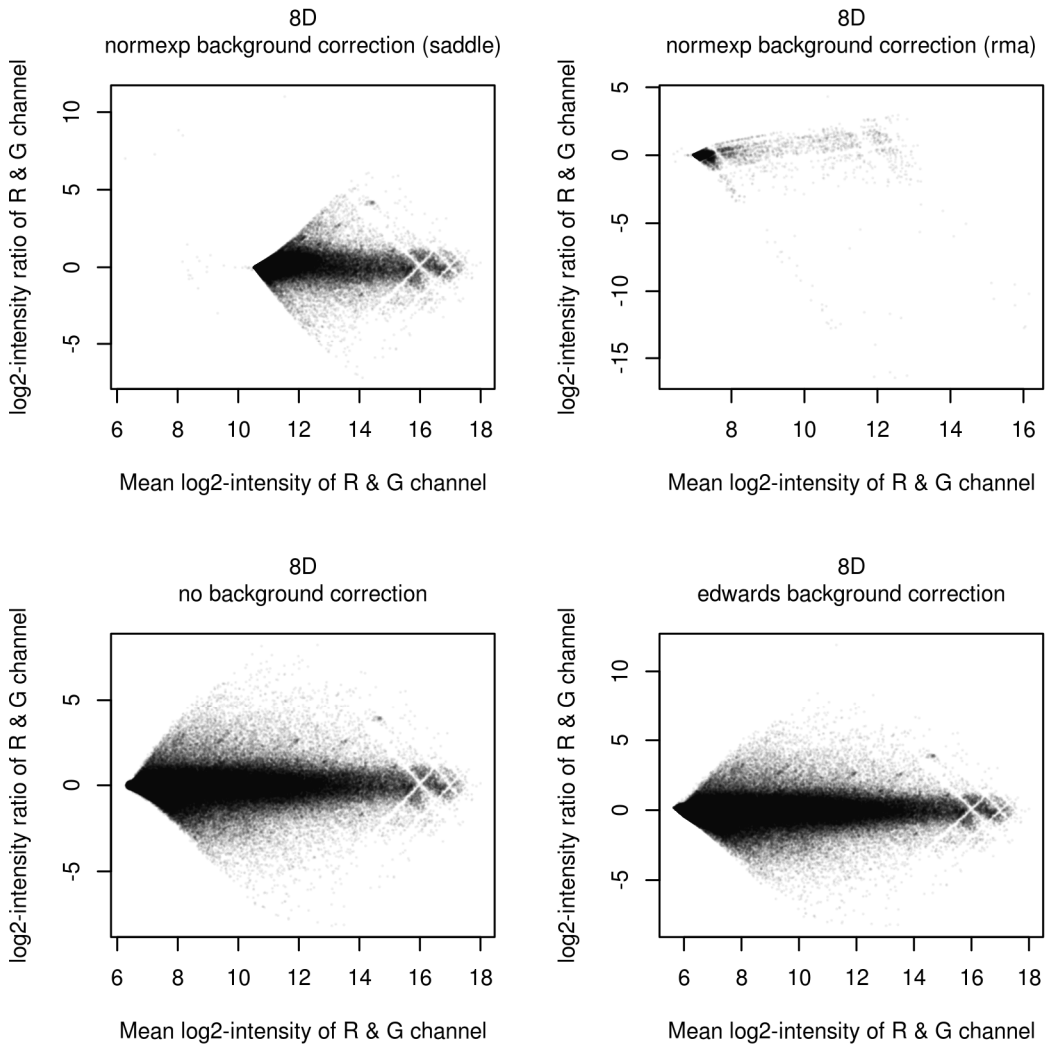


Figure 3.2.2

MA-plots of tumour 8D after background correction using four different methods

Scatterplots visualising log2-ratios of the intensities in the red (R) and green (G) channels (M-values) as a function of the mean log2-intensity of those channels (A-values) with respect to tumour 8D from the "AgilentG4502A_07_3" platform code in the original genomic characterisation of CRC by TCGA (TCGA Network, 2012b). The original publication did not contain any background correction (bottom left panel). Several different alternatives were explored, with the model-based normexp methods using saddle-point approximation (Ritchie et al., 2007; Silver et al., 2009, top left panel) or the classical RMA method (Irizarry et al., 2003, top right panel) being outperformed by the method according to Edwards (2003, bottom right panel). M-values were loess-normalised using the function normalizeWithinArrays from the R package limma v3.26.2 before plotting. Each dot corresponds to one probe.

A log-linear interpolation as proposed by Edwards (2003) was finally used to adjust lower intensities since it showed the best performance across all samples, which resulted in strictly positive mRNA measurements.

After the streamlined pre-processing of the different transcriptomics datasets, they had to be combined in order to generate consensus measurements for all transcripts in each sample, which was a requirement for the prediction of CMS subtype membership using the classifier by Guinney and colleagues (Guinney et al., 2015). One central amendment made to the original aggregation method proposed by Guinney et al. was the delayed selection of a reference dataset based on the distributions of *maxcomet* values for each dataset, visualised in Figure 3.2.3 and described in section 2.9.11.

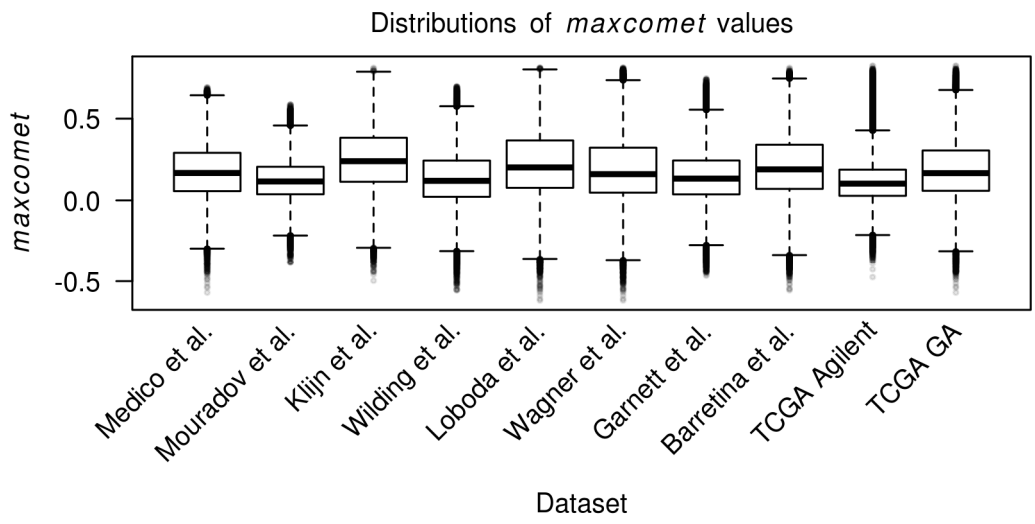


Figure 3.2.3

Maxcomet distributions of the different mRNA datasets

Boxplots visualising the distributions of maxcomet values across all mRNA datasets integrated in this study. The dataset by Barretina et al. was finally selected as the reference, since it showed a favourable distribution of maxcomet values, while also being the biggest dataset both in terms of samples and detected probes (see section 2.9.11).

Based on these distributions, the dataset by Barretina et al. was chosen to be the reference, since its distribution was shifted towards higher *maxcomet* values and because it was the biggest dataset considering samples and detected probes. The ten mRNA datasets were then aggregated as described in section 2.9.11, which resulted in a combined expression matrix of 9,737 transcripts across 448 and 160 redundant or 145 and 89 non-redundant cell lines and tumours, respectively (including also CRC cell lines not in the CRC65 panel). As observed previously for transcriptomics datasets, mRNA expression followed a bimodal distribution of transcript abundance, which is shown in Figure 3.2.4 (Hart et al., 2013; Hebenstreit et al., 2011; Nagaraj et al., 2011; Sharma et al., 2015).

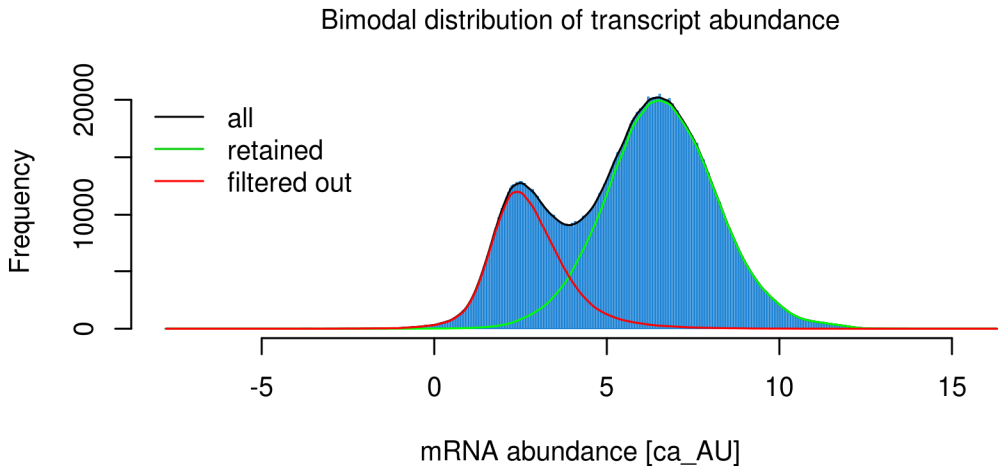


Figure 3.2.4

Bimodal distribution of transcript abundance after dataset aggregation

Histogram visualising the bimodal distribution of transcript abundance after aggregation of all ten mRNA datasets as described in section 2.9.11. Genes were either retained or filtered out using a heuristic based on the sum of mRNA abundance across all samples for each gene, visualised in Figure 3.2.6.

With a combined consensus mRNA expression matrix at hand, it was possible to examine the correlation structure of the different datasets with the aim of identifying possible sample mix-ups. Focusing on the correlation structure of the two TCGA datasets following the integration of the above-mentioned public mRNA datasets, two pairwise sample mix-ups became apparent. All pairwise correlations between samples present in both the Agilent microarray dataset (platform code "AgilentG4502A_07_3") and the Genome Analyzer mRNA-Seq dataset were calculated and the distribution of correlation coefficients from sample pairs with differing sample annotations was defined as the background of false correlations. Sample pairs the correlations of which were two standard deviations bigger than the mean of the background of false correlations were treated as cognate sample pairs. To no surprise, most of these sample pairs lay on the diagonal of the symmetrical image plot in Figure 3.2.5, which meant that

cognate sample pairs also had conforming sample annotations, as indirectly defined above. However, four of these cognate sample pairs did not coincide with the diagonal, meaning that they had differing sample annotations even though their correlations were significantly higher than the background of false correlations. Since these were two reciprocal pairs of samples, it was highly likely that they were mislabelled in one of the two mRNA quantification workflows of TCGA. Under the assumption that the older dataset ("Agilent") contains correct annotations, these samples were re-labelled in the TCGA mRNA-Seq dataset ("GA") to adjust for this mislabelling.



Figure 3.2.5

Sample mix-ups in the TCGA mRNA datasets of tumours

Image plot visualising two pairwise sample mix-ups in the TCGA mRNA datasets with the platform codes "AgilentG4502A_07_3" (rows "Agilent") and "IlluminaGA_RNASeq" (columns "GA"). Only samples present in both datasets are shown. Blue squares mark sample pairs the correlations of which were significantly higher than the background of false correlations (see main text). Bold red arrows visualise the direction of the mix-ups under the assumption that the older dataset ("Agilent") contains correct annotations, which is also the direction chosen for the correction. Faint red arrows visualise the alternative direction of the mix-ups under the assumption that the younger dataset ("GA") contains correct annotations.

However, without a proper reference, there is now way of knowing which of the two datasets contains correct annotations. Afterwards, the combined mRNA expression matrix was restricted to high abundance transcripts based on the distribution of the sum of mRNA abundance across all samples for each gene (see Figure 3.2.6). This reduced the dataset further to a matrix of 7,113 transcripts across 145 cell lines and 89 tumours, respectively ("retained" in Figure 3.2.4), which could then be used to predict the CMS of cell lines and tumours, as described in section 3.2.2.

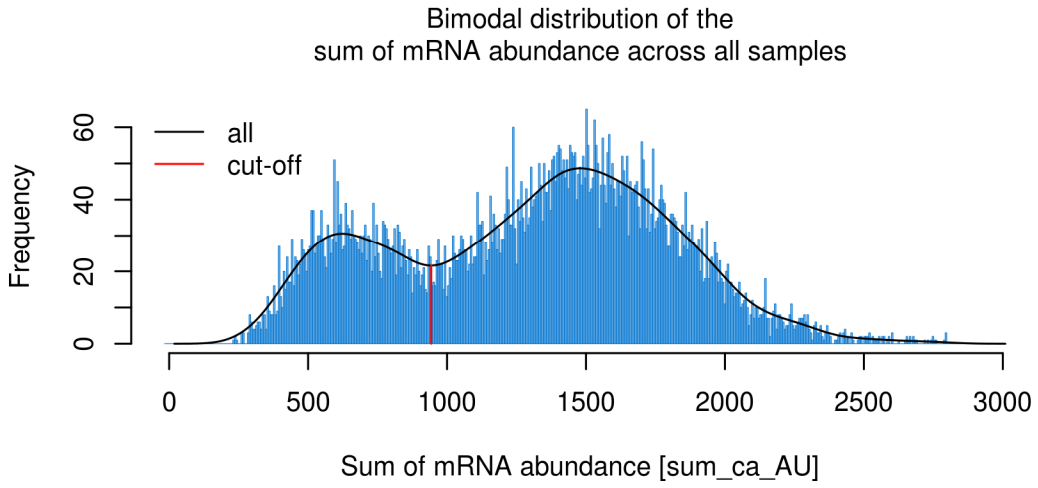


Figure 3.2.6

Bimodal distribution of the sum of mRNA abundance for each gene across all samples

Histogram and smoothed density plot (black line "all") visualising the bimodal distribution with respect to the sum of mRNA abundance for each gene across all samples, which was used as a heuristic to filter out genes with consistently low abundance. The vertical red line shows the cut-off used to distinguish low-abundance (to the left of the line) and high abundance (to the right of the line) transcripts.

3.2.2 Prediction of Consensus Molecular Subtypes (CMS)

In order to be able to use the single-sample predictor supplied by Guinney et al. for the assignment of cell lines and tumours to the CMS based on the aggregated mRNA data, the classifier had to be adapted to accept HGNC symbols as opposed to Ensembl GeneIDs as identifiers. This predictor used gene expression centroids of the four CMS with respect to 693 GeneIDs (classifier genes) – selected based on GSE39582 (Marisa et al., 2013) to maximise their predictive performance as evaluated using area-under-the-ROC-curve and fold-change criteria – and measured the similarity of new samples to these centroids using Pearson correlation coefficients. It is worth noting that the centroids used by the predictor were not restricted to the ones from GSE39582 alone, but also included centroids of the four CMS from four additional datasets, some of which belong to the TCGA datasets described in section 3.2.1. This resulted in a collection of centroid vectors $c_{i,j} \in \mathbb{R}^{g_c}$ for i=1,…,4 CMS and j=1,…,5 datasets across g=1,…,g_c classifier genes, with g_c=693. The original publication by Guinney et al. (2015) contains a detailed description of the prediction algorithm. Briefly, for each new sample $x \in \mathbb{R}^G$ with p=1,…,G genes, the Pearson correlation between $x \in \mathbb{R}^{g_c}$ and all $c_{i,j} \in \mathbb{R}^{g_c}$ was calculated while restricting x to only the classifier genes, which resulted in 20 similarity measures $S_{i,j} \in \mathbb{R}$. Afterwards, for each CMS i=1,…,4, minimal

similarities $S_i = \min_j S_{i,j}$ and their maxima $S_{i^*} = \max_i S_i$, as well as similarity differences $D_{i^*} = S_{i^*} - max_{i'} S_{i'}$ with $i' = i \backslash i^*$ were determined, followed by the assignment of CMS i^* to x given that $S_{i^*} \geq 0.15$ and $D_{i^*} \geq 0.06$, or "undetermined" otherwise. In order for this assignment to be possible, the same classifier genes needed to be used for both $c_{i,j}$ and x. Therefore, biomaRt (Durinck et al., 2005; Durinck et al., 2009) was employed to map Ensembl GeneIDs associated with classifier genes also present in the combined mRNA expression matrix described in section 3.2.1 to HGNC symbols, as described in section 2.9 for similar mappings during the pre-processing of public mRNA datasets. After assigning these HGNC symbols to the classifier genes, the aggregated mRNA dataset was employed for the assignment of cell lines and tumours to the CMS using the classifyCMS.SSP function from the CMSclassifier R package v1.0.0, published alongside the original paper (Guinney et al., 2015). This resulted in an overall prediction accuracy for the tumours of 0.80 (65/81). The calculation of the overall prediction accuracy for the cell lines was not possible due to the lack of reference annotations in the original study by Guinney and colleagues. The subtype labels of cell lines (excluding CRC cell lines not in the CRC65 panel) and tumours are shown in Figure 3.2.7, alongside a summary of which proteomic datatypes (kinome and full proteome), mRNA technologies (Agilent, Genome Analyzer/GA, etc.) and drug sensitivity data were included in this study for each sample. The subtype labels can also be found in Table 6.1.9, which summarises the annotation information for the CRC65 and CPTAC datasets (see section 2.18).

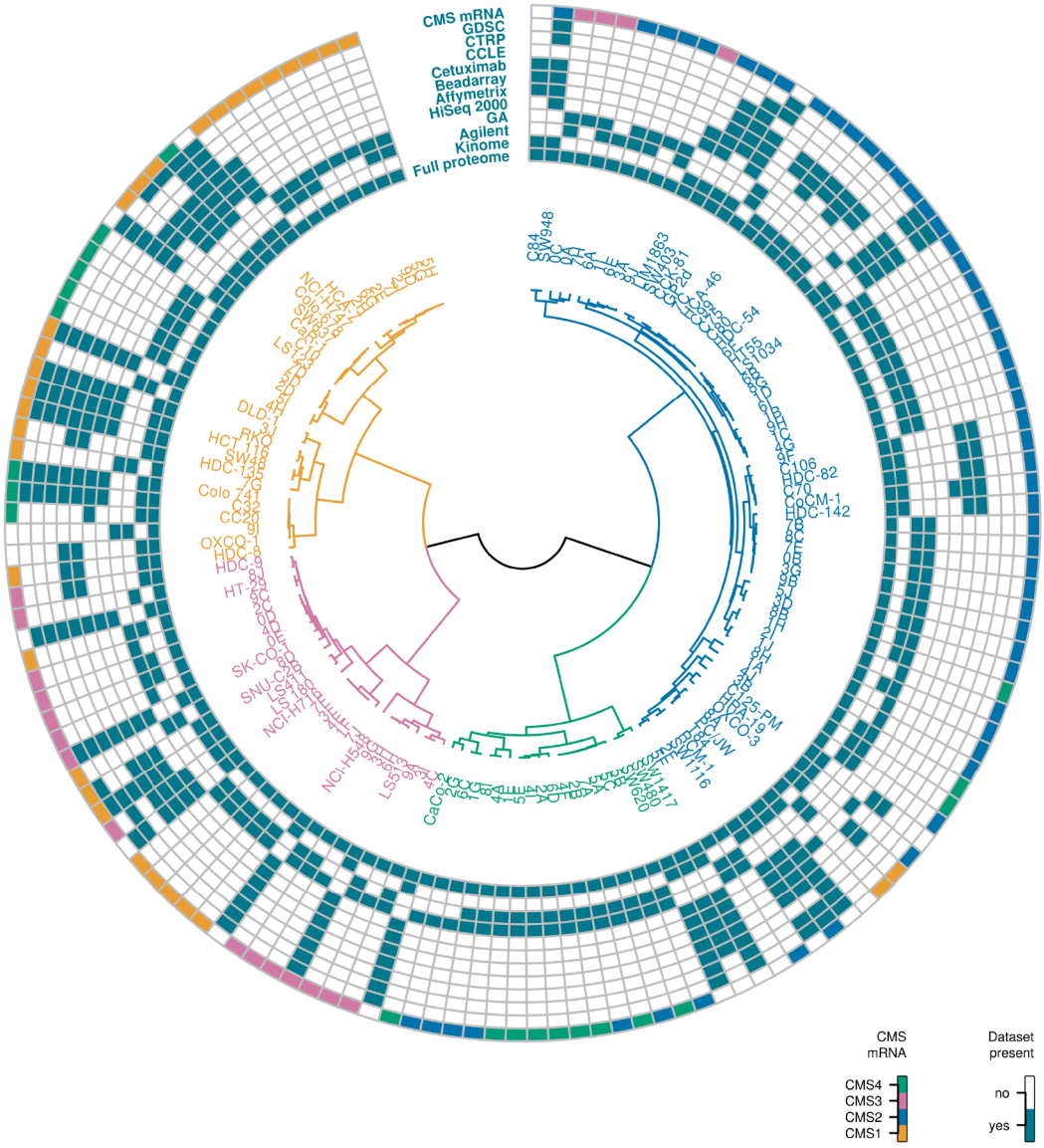


Figure 3.2.7

Overview of the datasets used for multi-omics data integration

Circos plot visualising the different datasets, which were integrated throughout this thesis. Patients (two-character labels) and cell lines (remainder of the samples) are ordered hierarchically based on their mRNA expression profiles (distance metric: 1-Pearson correlation; see section 2.11). The dendrogram is coloured according to the CMS representing the majority of samples in the respective cluster and a set of rings around it indicates which proteomics data (Full proteome, Kinome), mRNA technologies (Agilent microarray, Genome-Analyser-based mRNA-Seq, HiSeq-2000-based mRNA-Seq, Affymetrix microarrays or Illumina Beadarrays) and drug sensitivity datasets (Cetuximab, CCLE, CTRP or GDSC) were included in this study. The outermost ring highlights the different CMS subtype predictions based on mRNA abundance (see section 3.2.2). "Undetermined" CMS class labels are left white.

3.3 Full proteome

3.3.1 Optimization of hSAX fractionation

Following the transcriptomic characterisation of the cell lines in the CRC65 panel, the next step was their characterisation with respect to full proteome profiles, as the goal was to establish integrated full proteome subtypes of CRC and compare them to the CMS subtype model. Since even modern mass spectrometers cannot quantify all proteins contained in a conventional TCL in a single measurement because of the sheer complexity and high dynamic range of such a sample, methods to reduce sample complexity had to be applied. Focusing on the full proteome profiling, the number of proteins quantified per full proteome was maximised by offline fractionation of peptides from tryptic digests of TCLs using hSAX chromatography because of its high orthogonality to classical RP chromatography (Ritorto et al., 2013), which was used for online peptide separation. The established hSAX fractionation protocol used in the laboratory of Prof. Küster involved the collection of 48 fractions per sample, each of which was measured using a 2 h online RP gradient (see section 2.5.1.5). Following this standard procedure would have resulted in a prohibitive amount of measurement time for the entire CRC65 panel (260 days). In order to reduce the measurement time spent on the acquisition of full proteome data, separate LC-MS/MS measurements of each hSAX fraction from an in-solution digest of lysate from human placenta (provided by Lihua Li, former lab member) were analysed and a pooling scheme was developed based on the number of uniquely identified peptides per hSAX fraction. The aim was to combine hSAX fractions containing a below-average number of unique peptides, in order to roughly equalise the complexity of the different fractions, as visualised in Figure 3.3.1.

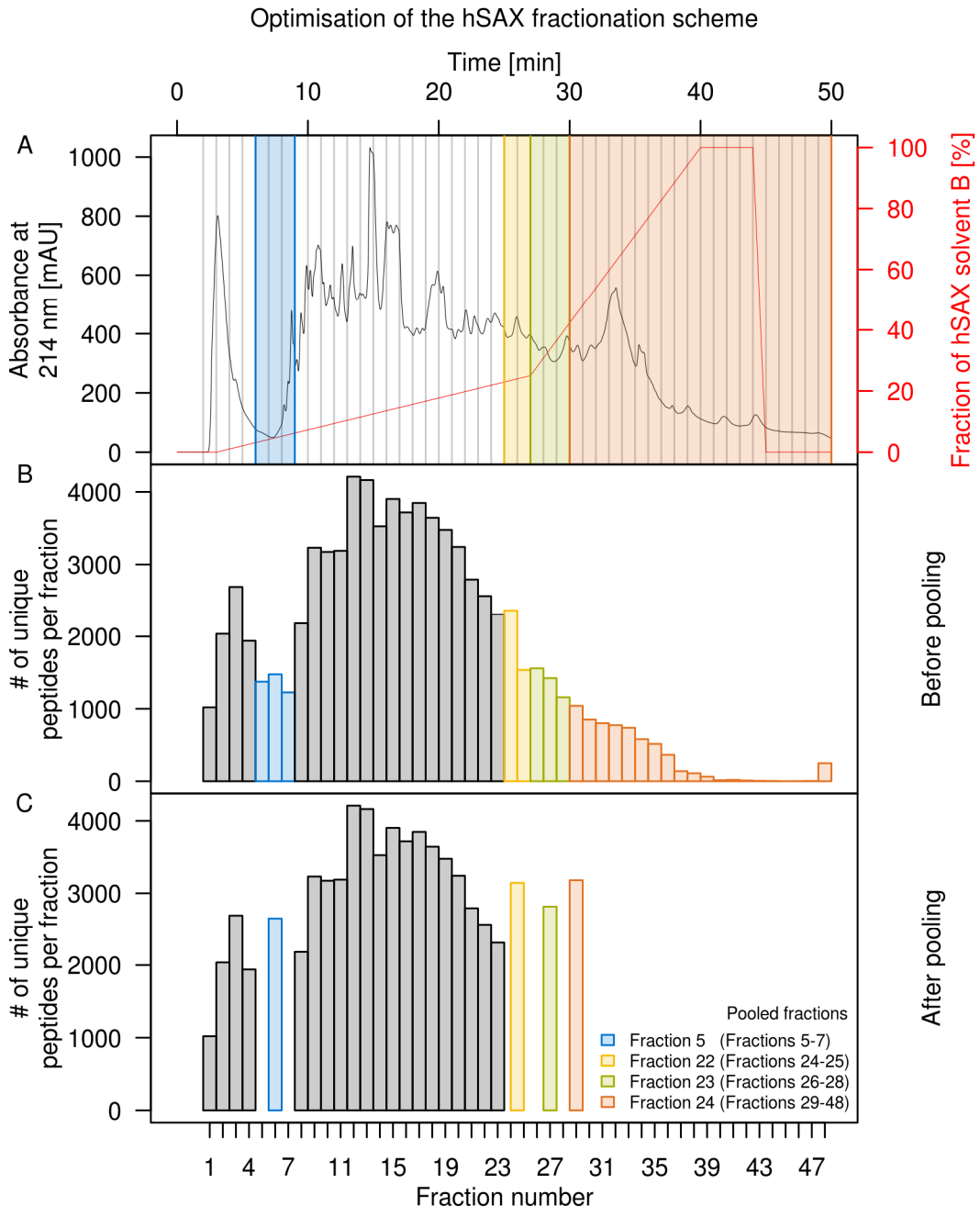


Figure 3.3.1

Optimisation of the hSAX fractionation scheme

(A) Chromatogram of an offline hSAX fractionation from an in-solution digest of lysate from human placenta (provided by Lihua Li, former lab member). The absorbance at 214 nm (black trace) is shown on the primary y-axis in mAU, while the fraction of hSAX solvent B (red trace) is shown on the secondary y-axis in per cent as a function of the chromatography time in minutes. Vertical grey bars show the fraction borders, while coloured stripes visualise which fractions were pooled. (B-C) Bar chart of the number of unique peptides per fraction (B) before or (C) after pooling of specific fractions. Fractions that were selected to be pooled

are coloured according to the legend in (C), while non-pooled fractions are of grey colour.

Based on these results, the CRC65 panel was fractionated as described in section 2.5.1.5, pooling fractions 5-7 to form fraction 5, fractions 24-25 to form fraction 22, fractions 26-28 to form fraction 23 and fractions 29-48 to form fraction 24, respectively. This reduced the measurement time per cell line from four days down to two, effectively cutting the measurement time needed for the full proteome characterisation of the CRC65 panel in half.

3.3.2 Protein identifications in the CRC65 and CPTAC datasets

Having established an efficient pre-fractionation scheme, the CRC65 cell line panel was processed as described in section 2.5.1 with a lot of help from Riccardo Zenezini Chiozzi (former lab member) during the hSAX fractionation and the post-hSAX desalting step. Raw data were processed as detailed in section 2.7.1 in order to unify the processing of the CRC65 and CPTAC datasets. The data were therefore searched using the same parameters, but not in a single MaxQuant run, since this resulted in the breakdown of MaxQuant's retention time alignment. Figure 3.3.2 shows retention time data for the CRC65 (left column) and CPTAC datasets (right column) before (top row) or after retention time recalibration in two separate (middle row) or one merged MaxQuant run (bottom row). Depicted are the (un-) calibrated retention times with respect to peptides identified by MS/MS in all fractions of one reference sample (CRC65=C10, CPTAC=0A) on the x-axes of the various plots. They are plotted against the (un-) calibrated retention times (y-axes of the various plots) with respect to peptides identified by MS/MS in all fractions of all other samples, against which these reference fractions would

be matched during “Match between runs” (the same fraction ± one fraction, i.e. the same and neighbouring fractions). Even though in principle, all fractions of all samples were matched to the same and neighbouring fractions in all other samples, only reference fractions for one sample are shown in order to reduce the complexity of the figure. Since this is still a significant number of data points to be shown in one plot, they were summarised as a two-dimensional histogram to avoid over-plotting. All peptides in the grey “alignment time window” were used for retention time recalibration. After recalibration, identifications were transferred from peptides identified by MS/MS to unidentified peptide precursor ions with less than 10 ppm mass deviation and less than 1.1 min difference in recalibrated retention time (red “match time window”). Looking at the middle and bottom row of plots in Figure 3.3.2, it can be seen that only processing the datasets separately from one another resulted in satisfactory retention time alignment of both datasets, even though completely different fraction numbers were used for the two datasets in both analyses. Particularly the CPTAC dataset suffers from a merged analysis, since the majority of peptides are outside the “match time window” in the bottom right plot.

Processing the two datasets separately from one another according to section 2.7.1 resulted in the quantification of gene groups using giBAQ values (see section 2.7.2), which were based on the MS1 raw intensities extracted from raw files with the help of MaxQuant. These were subsequently compared to the spectral-count-based quantification as reported by CPTAC (Zhang et al., 2014a).

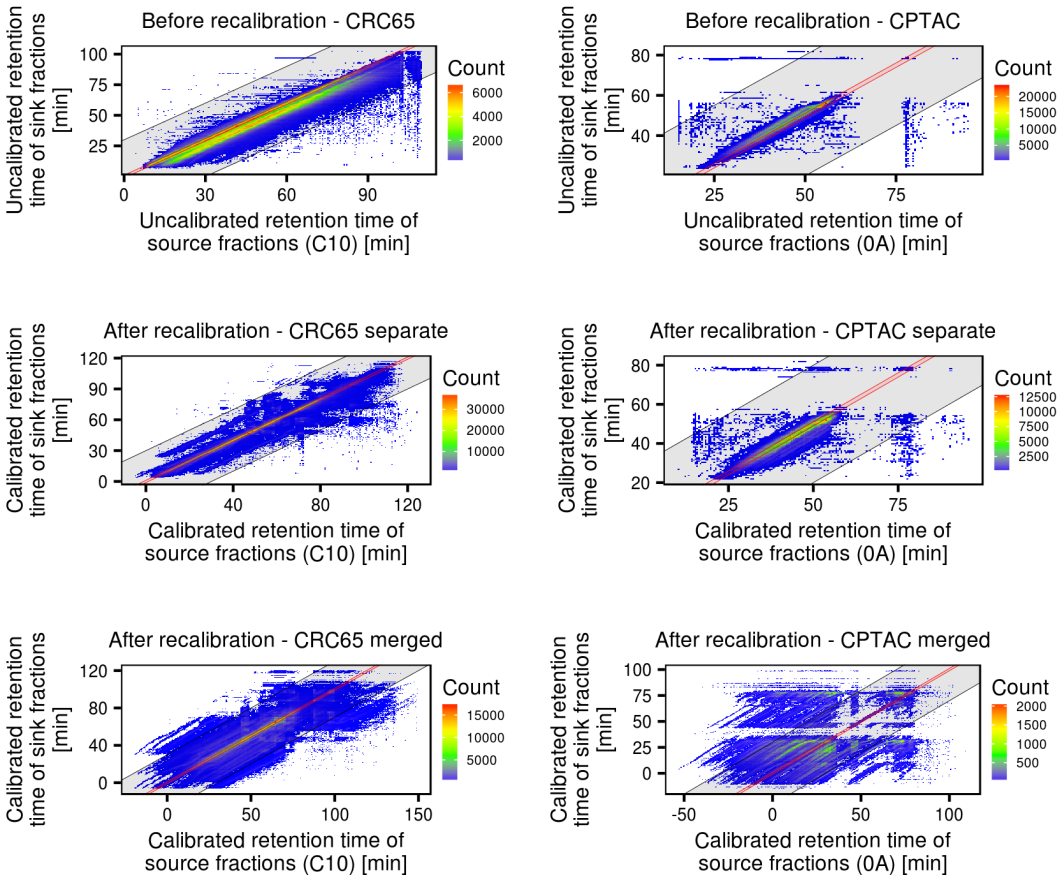


Figure 3.3.2

Retention time recalibration – merged or separate processing

Two-dimensional histograms (600 bins in each dimension) visualising uncalibrated (top row) or calibrated (bottom two rows) retention times of peptides identified by MS/MS in fractions receiving identifications via "Match between runs" (y-axes, "sink fractions") as a function of the (un-) calibrated retention times of the same peptides identified by MS/MS in fractions serving as the source for these identifications (x-axes, "source fractions"). Only fractions eventually "matched" to each other (the same fraction ± one fraction) are plotted against each other. For simplicity, only one reference cell line/tumour and its source fractions are shown. The left column shows the CRC65 data, while the right column shows the CPTAC data. To recalibrate retention times, the two datasets were either processed separately from one another (middle row) or in one merged MaxQuant run (bottom row). In order to prevent over-plotting, the number of data points in a particular area of the plot (Count) was mapped to a colour-scale as opposed to plotting each data point separately. Grey area: "alignment time window". Red area: "match time window".

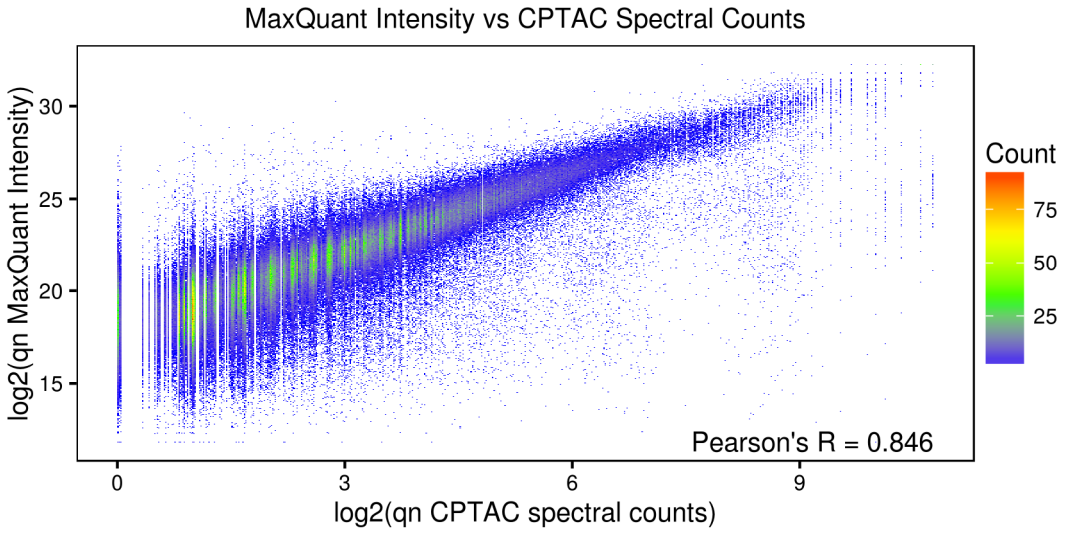


Figure 3.3.3

Correlation of MS1-based quantification and spectral counting

Two-dimensional histogram (600 bins in each dimension) visualising the quantile-normalised (qn) and log2-transformed (log2) MS1 raw intensity (MaxQuant Intensity) for all gene groups across all tumour samples in the CPTAC dataset determined using MaxQuant as a function of the quantile-normalised and log2-transformed spectral-count-based quantification extracted from Supplementary Table 13 of the original paper (Zhang et al., 2014a). In order to prevent over-plotting, the number of data points in a particular area of the plot (Count) was mapped to a colour-scale as opposed to plotting each data point separately.

Looking at the correlation of these two quantification approaches in Figure 3.3.3, the higher dynamic range of MS1-based quantification at the lower end of the detection limit is evident, emphasising the necessity for a unified processing strategy in order to make the CRC65 and CPTAC datasets comparable to each other.

Up to this point, offline peptide fractionation followed by classical LC-MS/MS-based shotgun proteomics (Experimental Procedures) was used to measure full proteomes of 65 CRC lines. A total of 187,045 peptides were identified using this approach (median/cell line=88,527), which were assembled into 11,796 protein

groups (median/cell line=9,447) and subsequently summarised to form 10,951 gene groups (median/cell line=9,068), as depicted in Figure 3.3.4.

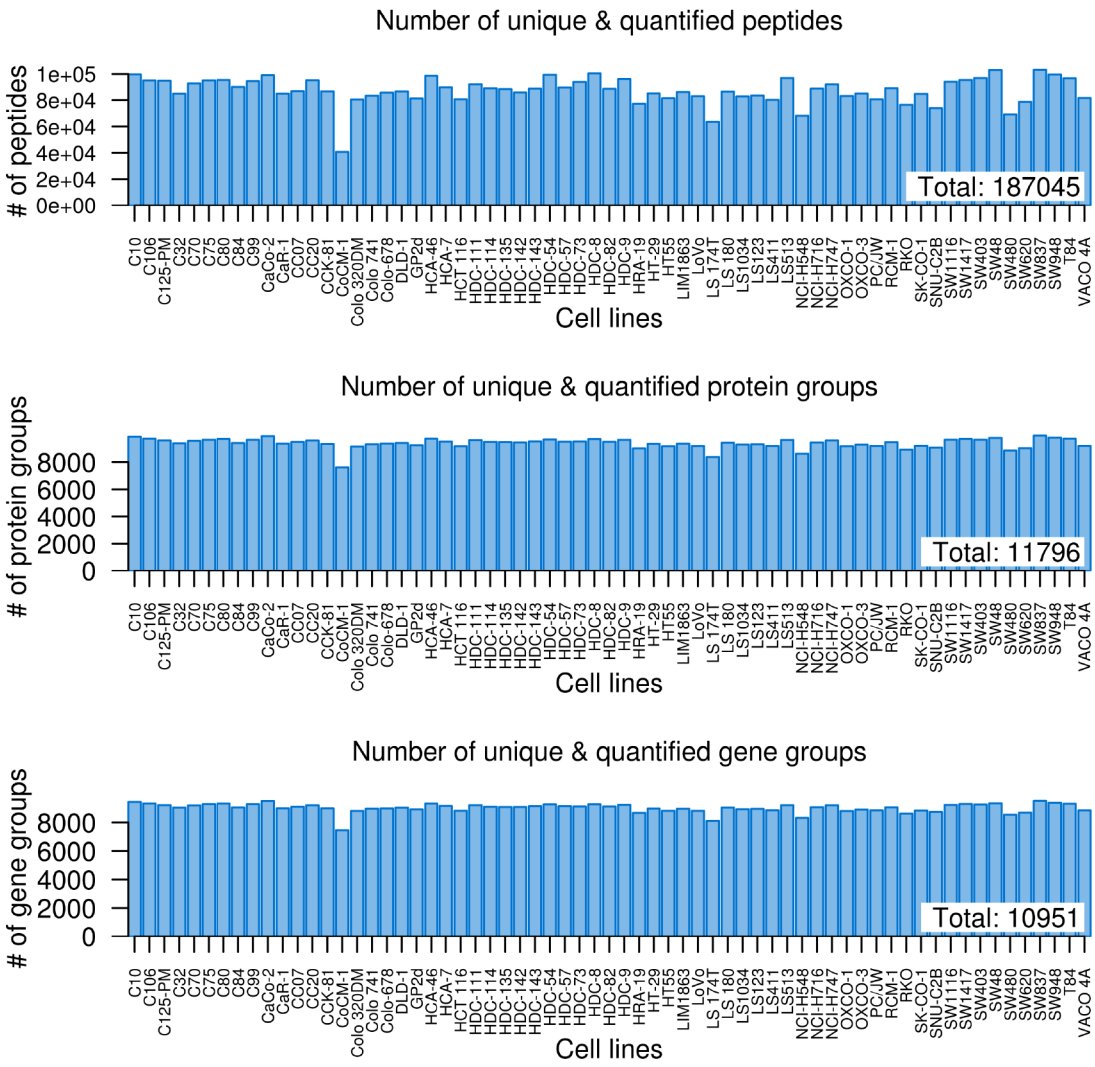


Figure 3.3.4

Full proteome identifications in the CRC65 dataset

Bar charts visualising the number of unique identified and quantified peptides, protein groups and gene groups in the full proteomes of the CRC65 cell line panel. Insets indicate the corresponding number of identifications across the entire dataset.

In order to transfer knowledge on drug sensitivities from cell lines to patients, this dataset was compared to a reanalysis of the CPTAC proteomics dataset on CRC (Zhang et al., 2014a).

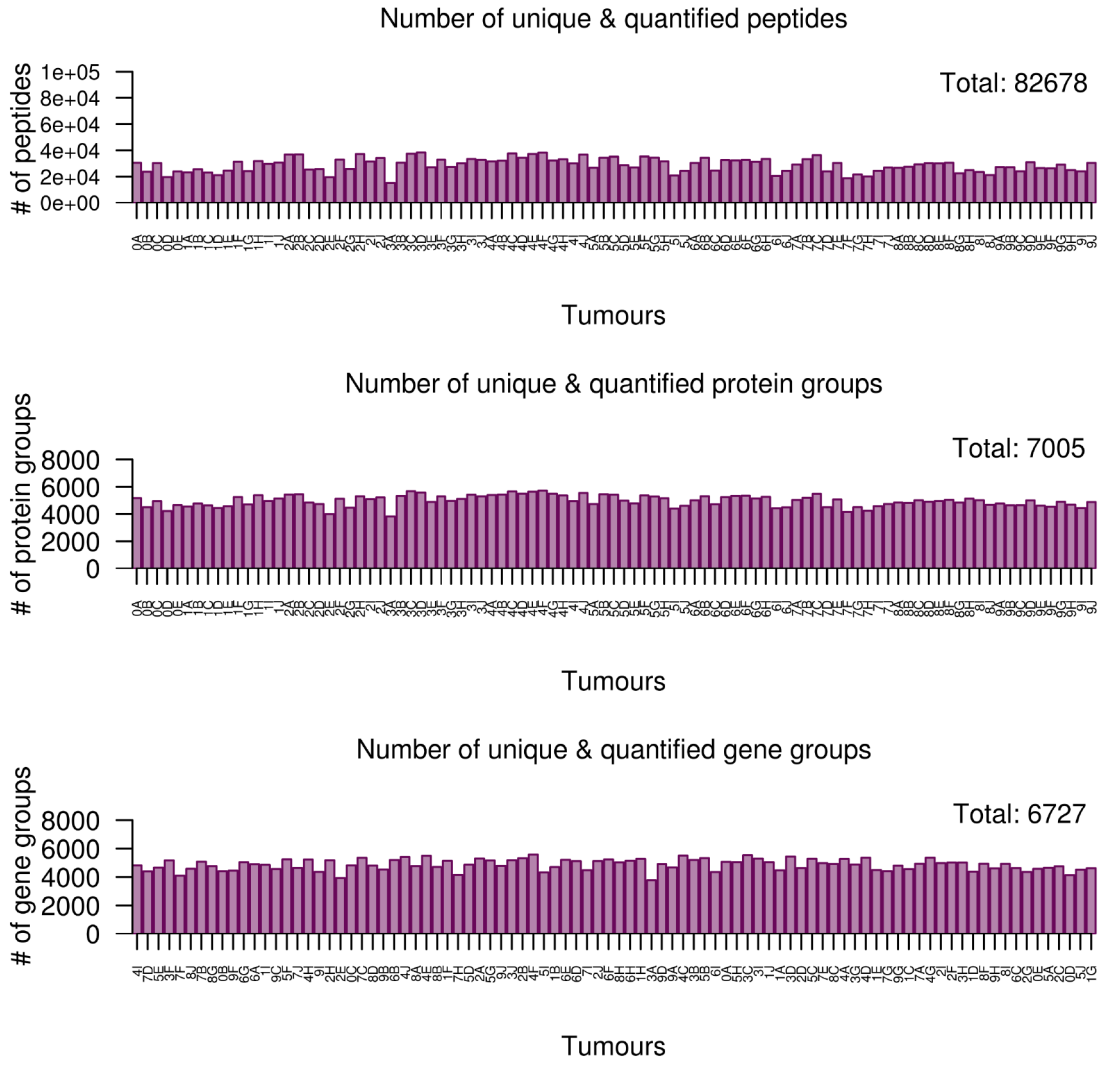


Figure 3.3.5

Full proteome identifications in the CPTAC dataset

Bar charts visualising the number of unique identified and quantified peptides, protein groups and gene groups in the full proteomes of the CPTAC patient dataset. Insets indicate the corresponding number of identifications across the entire dataset. The scale of the y-axes is the same as for Figure 3.3.4.

Using the same analysis pipeline as for the CRC65 panel (see section 2.7), 82,678 peptides were identified (median/patient=30,126), which were assembled into 7,005 protein groups (median/patient=4,980) and subsequently summarised to form 6,727 gene groups (median/patient=4,901), visualised in Figure 3.3.5.

In order to confirm that the CRC65 cell line panel was indeed characterised to a comparable proteomic depth as the CPTAC patient dataset – a prerequisite for the discovery of proteomic subtypes of CRC consisting of cell lines and patients – the y-axes in Figure 3.3.4 and Figure 3.3.5 were kept the same, while Venn diagrams were used to illustrate the overlap, as well as the complements with respect to the identifications in both datasets.

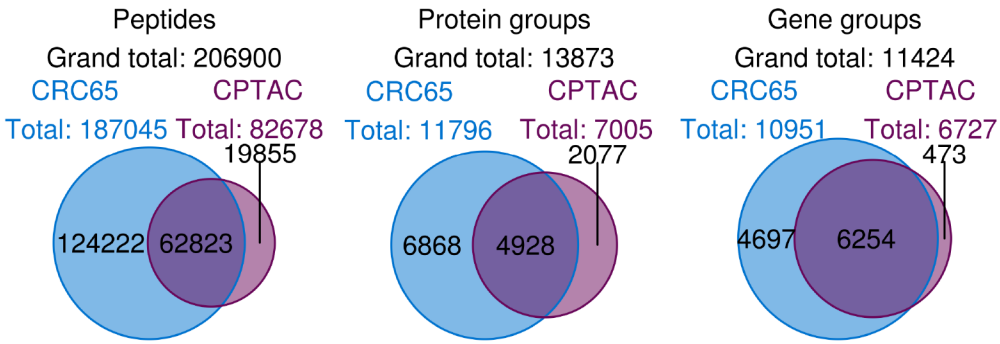


Figure 3.3.6

Venn diagrams of full proteome identifications

Venn diagrams showing the intersection and complements with respect to peptide, protein group and gene group identifications across both the CRC65 cell line and CPTAC patient dataset.

As seen in Figure 3.3.6, the CRC65 dataset contained most of the CPTAC dataset and showed a much higher number of unique identifications in all categories examined. This in turn enabled the discovery of integrated proteomics subtypes of CRC, as outlined in section 5.3. It is worth mentioning that gene groups unique to the CPTAC dataset were enriched in extracellular matrix proteins and IgGs, while gene groups unique to the CRC65 dataset were enriched in transcription factors and DNA binding proteins. In all subsequent proteomics data analyses, only data on the gene group level with gene symbols as identifiers was used and referred to as protein expression/abundance throughout this thesis in order to be able to compare proteomics and transcriptomics data.

3.4 Discussion

It is known that cell lines included in the CRC65 panel capture the main molecular subtypes of CRC (Medico et al., 2015; Mouradov et al., 2014), however the CMS of these cell lines was not known as of yet, precluding integrated analyses based on the current molecular subtyping consensus in CRC. By combining multiple public mRNA datasets, it was possible to assign 42 of them to the CMS subtypes with the help of the single-sample predictor by Guinney et al. (2015), confirming that these cell lines indeed represent the main subtypes of CRC. This represents the most comprehensive meta-analysis on transcriptomic data from CRC cell lines so far. The pre-processing of the different datasets was streamlined and tightly controlled using common quality control measures, resulting in consistent mRNA abundance measurements, which were needed for the accurate assignment of these cells to the CMS subtypes. It should be noted that the transcriptional profiles of the same cell lines and tumours were surprisingly consistent across the different datasets if systematic differences between them were accounted for, underscoring the importance of suitable cross-platform data aggregation methods for the success of meta-analyses in transcriptomics research. It was shown before by numerous studies that mRNA abundance measurements from the same sample generated on different platforms in the same or multiple distinct laboratories agree well with each other (Agarwal et al., 2010; Fu et al., 2009; Kogenaru et al., 2012; Malone and Oliver, 2011; MAQC Consortium et al., 2006; SEQC MAQC-III Consortium, 2014; Xu et al., 2013), particularly when highly expressed genes are considered (Barnes et al., 2005; Marioni et al., 2008; Su et al., 2011; Wang et al., 2014; Zhao et al., 2014) and even more so if batch effects between datasets are accounted for using one of several algorithms (Singular

Value Decomposition or SVD - Alter et al., 2000; Distance Weighted Discrimination or DWD - Benito et al., 2004; Quantile Normalisation or QN - Bolstad et al., 2003; Remove Unwanted Variation, 2-step or RUV-2 - Gagnon-Bartsch and Speed, 2012; Distribution Transformation or DisTran - Jiang et al., 2004; Empirical Bayes or ComBat - Johnson et al., 2007; Scaling to reference dataset - Kim et al., 2007; Surrogate Variable Analysis or SVA - Leek and Storey, 2007; Gene-wise Standardisation or GWS - Li and Wong, 2001; Ratio-based batch reference - Luo et al., 2010; fRMA - McCall et al., 2010; fRMA barcode - McCall et al., 2011; Cross-Platform Normalization or XPN - Shabalin et al., 2008; Batch Mean Centering or BMC - Sims et al., 2008; MComBat - Stein et al., 2015; Quantile Discretization or QD - Warnat et al., 2005; Gene Quantiles or GQ - Xia et al., 2009). With such a variety of methods at hand, the task of choosing an appropriate one for the removal of systematic differences between two given datasets becomes difficult, but amongst the recommended methods from the list of techniques above as judged by comparative analyses (Chen et al., 2011; Luo et al., 2010; Rudy and Valafar, 2011), ComBat stood out as not relying on large batch sizes (Johnson et al., 2007). With the dataset by Mouradov et al. (2014) only containing 12 cell lines, this method laid the foundation for the integrative transcriptomics analysis performed as part of this thesis. The approach taken here was adapted from the scheme by Guinney et al. (2015), which in turn resembles the system described by Heider and Alt (2013). Their R package virtualArray is designed for meta-analyses of transcriptomics datasets from different microarray platforms and consists of the steps 1) loading and storage of raw data, 2) transformations of raw data, 3) annotation of raw data, 4) collapsing of redundant probesets, 5) compilation of the virtual array, 6) construction of new ExpressionSet and 7)

removal of batch effects. Steps 1)-3) are similar to the pre-processing described in section 2.9 of this thesis, with the default common identifier used here also being HGNC symbols. In contrast to this, step 4) was replaced with a modified reference dataset method analogous to the one used by Guinney and colleagues, which does not collapse redundant probesets to a single value and instead uses a consistency criterion to decide which probeset to use in order to represent a given HGNC symbol. The remainder of the aggregation method outlined in section 2.9.11 again resembled steps 5)-7) of the virtualArray R package, with the default batch adjustment technique being ComBat in both cases. This aggregation scheme resulted in a consensus mRNA dataset, which could finally be used for the assignment of cell lines and patients to the CMS subtype model, as well as for the comparison of mRNA and protein abundance, detailed in sections 5.2.2 and 5.2.1.

In order to perform these comparisons as part of this project and to fuel the pharmacoproteomic approach described in section 1.4, the hitherto biggest LC-MS/MS-based shotgun proteomics dataset on cancer cell lines in general and CRC cell lines in particular was generated, as described in section 3.3. Other datasets on CRC cell lines were much smaller (Fanayan et al., 2013b; Halvey et al., 2014; Mathivanan et al., 2012; McAlister et al., 2014; Piersma et al., 2015) or used reverse-phase protein arrays instead of mass spectrometry (Briffa et al., 2015). Some datasets on collections of cell lines from different tumour entities were comparable in size (Coscia et al., 2016; Grundner-Culemann et al., 2016; Lawrence et al., 2015; Rikova et al., 2007) and only very few studies on cell lines from multiple different tumour entities reached similar numbers in terms of cell lines investigated or proteins quantified (Geiger et al., 2012; Gholami et al., 2013). However, in order to be able to reliably model drug sensitivity as a function of full

proteome profile and predict *in vitro* preclinical drug sensitivity, a sizeable number of CRC cell lines covering the main molecular subtypes of CRC needed to be analysed (Wilding and Bodmer, 2014). The first small-scale proteomic studies on CRC cell lines used 2D-PAGE to study a hand full of proteins in a limited number of cell lines (Ji et al., 1994; Simpson et al., 2000; Ward et al., 1990), but as LC-MS/MS became more powerful over the following years and after the introduction of multi-dimensional separation techniques (Washburn et al., 2001), research groups started to move away from gel-based comparisons and began to quantify CRC proteomes, phosphoproteomes and secretomes using mass spectrometry (Ali and Molloy, 2011; Bauer et al., 2012; Fanayan et al., 2013a; Ghosh et al., 2011; Ikonomou et al., 2012; Kim et al., 2005b; Leroy et al., 2009; Luque-Garcia et al., 2010; Marin-Vicente et al., 2013; Organ et al., 2011; Slebos et al., 2008; Tan et al., 2008; Wang et al., 2009; Wu et al., 2010). However, it was not until very recently that proteomics datasets reached an analytical depth of 5,000-10,000 quantified proteins per sample (Beck et al., 2011; Coscia et al., 2016; Geiger et al., 2012; Gholami et al., 2013; Halvey et al., 2014; Lawrence et al., 2015; McAlister et al., 2014; Nagaraj et al., 2011), enabling pharmacoproteomic studies like the one described herein. The data processing workflow of the full proteomes was tightly controlled and optimised in order to make the CRC65 dataset as comparable as possible to the CPTAC dataset, enabling the transfer of knowledge from cell lines to patients. This comprehensive proteomic dataset on CRC served as the basis for the first in-depth pharmacoproteomic characterisation of human colon and rectal cancer.

Chapter 4
Kinome of the CRC65 cell line panel

4.1 Introduction

Kinases are pharmacologically well tractable proteins (Knapp et al., 2013), which play essential roles in various types of cancer (Brognard and Hunter, 2011; Giancotti, 2014; Kolch et al., 2015) and are therefore frequently inhibited by targeted agents developed to treat it (Gross et al., 2015; Zhang et al., 2009). However, it is not uncommon that tumours are either intrinsically resistant to a given targeted drug or quickly acquire resistance to such inhibitors through a variety of mechanisms (Holohan et al., 2013). In recent years, personalised medicine and combinatorial therapy emerged as viable tactics to overcome intrinsic or acquired resistance respectively (Al-Lazikani et al., 2012; Burris, 2013; Garraway and Janne, 2012; Gonzalez de Castro et al., 2013; McCubrey et al., 2012). The mainstay of such approaches is the availability of viable targets to the inhibition of which the tumour in question is likely to respond. As such, the kinome represents a promising sub-proteome to consider while searching for novel therapeutic targets, which is why this study included a dedicated approach for studying this protein family during the pharmacoproteomic characterisation of CRC.

While some kinases are frequently mutated (Bardelli et al., 2003; Kandoth et al., 2013), others become deregulated at the level of protein expression or through post-translational modifications and thus cannot be effectively analysed by DNA or RNA sequencing. A variety of different techniques can be used to study kinases at the protein level, some of which are *in vitro* kinase assays, (phosphorylation-

specific) western blots (WB) and chemoproteomics. These methods either focus only on a single kinase and its activity (kinase assays), can measure the abundance of multiple kinases and their activity on different substrates (WBs), or measure a composite of abundance and activity (Ruprecht et al., 2015) of a number of kinases at once by first enriching them from crude cell extracts and then quantifying them using mass spectrometry (Hess, 2013). Since the latter enables the quantification of many kinases in parallel from several different sources in a comparably small timeframe, it was the method of choice for the characterisation with respect to the kinome of the CRC65 cell line panel.

As kinases are structurally similar to each other – especially considering their functionally important ATP binding site (Fedorov et al., 2010) – it is possible to enrich them using ATP-competitive inhibitors or ATP analogues, coupled to solid supports (Daub et al., 2008; Medard et al., 2015; Oppermann et al., 2009; Patricelli et al., 2011). The present study made use of the latest generation of Kinobeads, which started out as a mixed affinity matrix consisting of seven ATP-competitive kinase inhibitors coupled to Sepharose beads (Kinobeads-α). This technology was developed by Bantscheff et al. and led to the revelation of the mechanisms of action with respect to clinical ABL kinase inhibitors (Bantscheff et al., 2007). This bead mixture was subsequently improved by the addition of beads coated with an AKT probe, giving rise to the next generation of Kinobeads (Kinobeads-β, Pachl et al., 2013a). The current iteration of beads – omitting staurosporine-, sunitinib- and bisindolylmaleimide III-coated beads from the mixture and adding beads coated with Cpd 19 – was used in this study and is termed Kinobeads-γ or KBγ (Medard et al., 2015). If these beads are incubated with a protein mixture containing natively folded kinases – such as total cell

lysate (TCL) – the inhibitors on the surface of the Sepharose beads will interact with the ATP binding pocket of kinases present in the mixture. Depending on the K_d of these interactions, kinases are retained on the beads during subsequent washing steps, which results in the enrichment of kinases from the original protein mixture. The different kinases retained on the affinity matrix can then be identified and quantified using mass spectrometry. This chapter focuses on the optimisation of the Kinobeads workflow for the kinome profiling of the CRC65 panel in sections 4.2-4.4. Section 4.2 describes an alternative to the classical batch enrichment protocol used in the laboratory of Prof. Kuster, section 4.3 focuses on the comparison of two different lysis buffers and section 4.4 evaluates the effect of different pulldown volumes on the number of kinases identified and quantified by mass spectrometry. The results of the final profiling are subsequently laid out in section 4.5.

4.2 Kinobeads in Micro-spin columns

The standard protocol for KBγ pulldowns in the Küster lab is well established for the profiling of as many kinases as possible from total cell lysate, without actually depleting them entirely (Medard et al., 2015). However, it could be advantageous to deplete some of the more abundant kinases from a given total cell lysate first, in order to be able to capture kinases which otherwise would not be able to bind to KBγ due to lower affinities, if the goal is to maximise the number of kinases quantified in a single experiment. Multicolumn affinity chromatography was amongst the first methods successfully applied for the enrichment of kinases from cell lysates (Daub et al., 2008; Oppermann et al., 2009; Wissing et al., 2007). With one advantage being the possibility to process more starting material in

order to increase the sensitivity of the assay (Oppermann et al., 2009), a multicolumn setup could have been a viable alternative to the standard batch enrichment procedure with KBγ. However, using an HPLC system for serial enrichments as in the original publication would have prohibitively increased the time to process a single sample, which is why microcolumns (Thermo Fisher) were instead investigated as a possible alternative to Tricorn 5/20 columns (GE Healthcare). These could be loaded and washed with the help of a six-channel peristaltic pump and instead of using separate microcolumns for each of the different bead types included in the KBγ mixture, the beads were stacked in a single microcolumn as described in section 2.5.2.2. Briefly, they were stacked from the bottom to the top of the column in either increasing (LOHI: compound 5, 7, 1, 15, 19) or decreasing (HILO) order with respect to the number of kinases they enrich in single-compound pulldowns (Medard et al., 2015). The experimental setup for a single pulldown experiment using microcolumns (PD LOHI/HILO) is visualised in Figure 4.2.1. This setup was compared to a classical batch enrichment procedure (PD Batch), outlined in section 2.5.2.3.

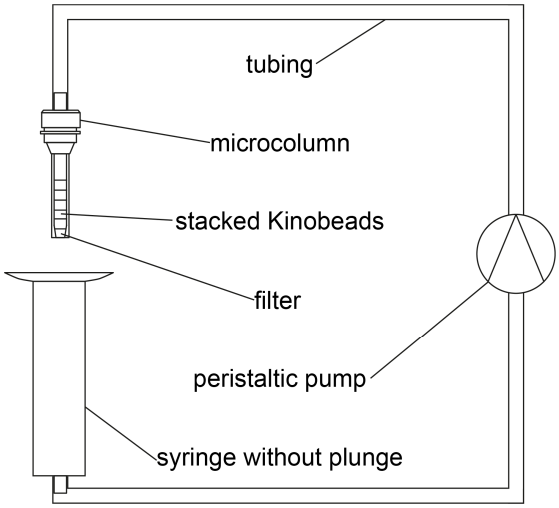


Figure 4.2.1

Experimental setup for Kinobeads pulldowns in microcolumns

Schematic of the experimental setup for Kinobeads pulldowns in microcolumns; only one of six channels is displayed for visual clarity.

In order to find out which of the two setups is superior, these two enrichment strategies were compared in technical duplicates (Rep 1 and Rep 2) with respect to the intensities of kinases quantified using both methods, as well as considering the total number of quantified kinases using either method from a common sample. For that, MV-4-11 cells were used as a model system and both lysis and pulldowns were carried out using CP buffer. Pulldown of pulldown experiments (PD/PD) were performed as well, in order to assess the degree of kinase depletion from total cell lysates.

Looking at scatterplots of LFQ kinase intensities for all samples against PD Batch Rep 1 as a common reference (Figure 4.2.2), it became apparent that all samples including the different PD/PDs were very similar, suggesting that enrichments using KBγ in microcolumns do not perform better than the regular batch enrichment method and do not result in kinase depletion, either.

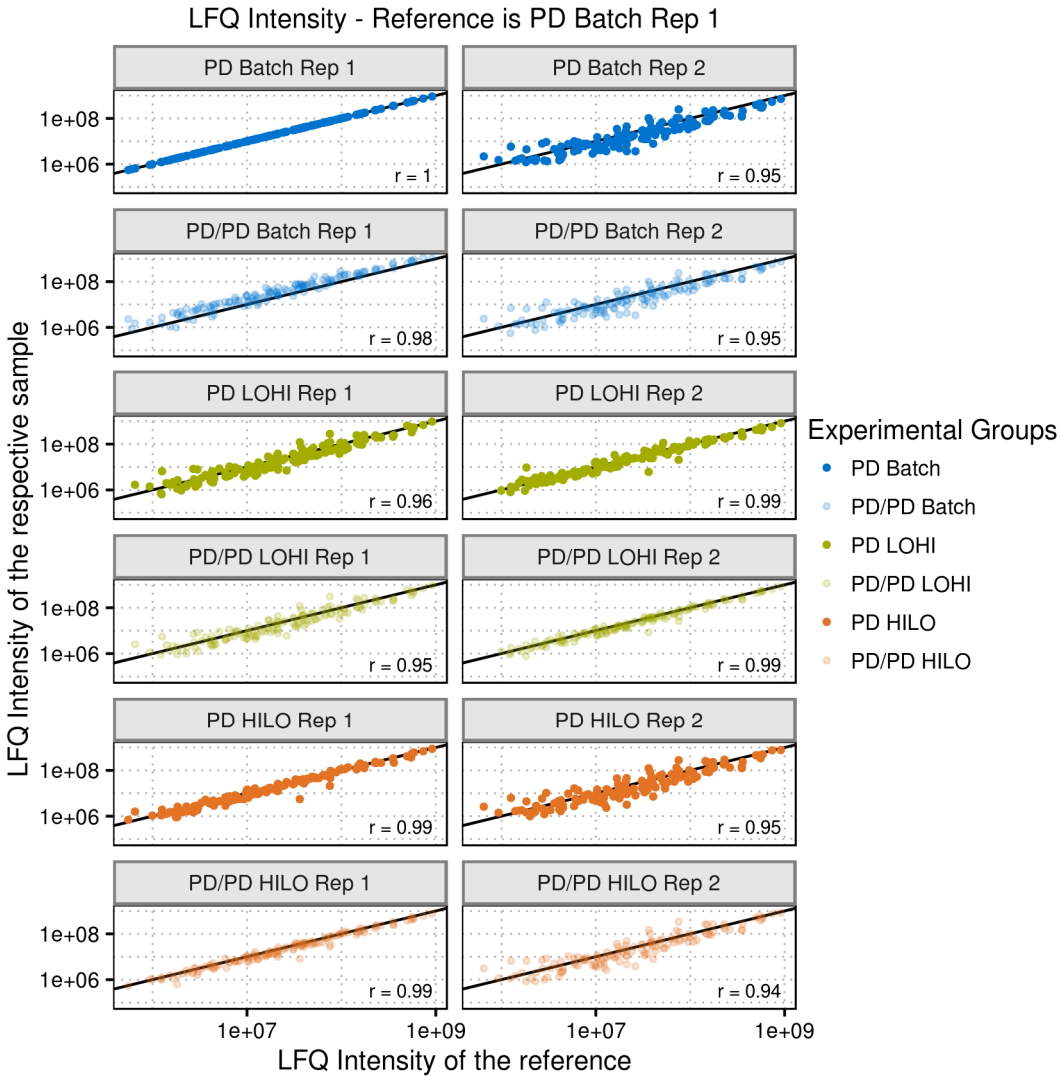


Figure 4.2.2

Scatterplots of LFQ Intensities for kinases enriched in microcolumns or batch-mode

Scatterplots visualising LFQ Intensities from kinases identified in Kinobeads pulldowns (PD) and the corresponding pulldowns of pulldowns (PD/PD) as a function of a common reference (PD Batch Rep 1). MV-4-11 cells were used for the pulldown experiments. Only data points present in both the reference and the corresponding experiment are shown. Data from the microcolumn configurations described in section 2.5.2.2 or the classical batch enrichment method outlined in section 2.5.2.3 are colour-coded. Columns represent technical replicates. The value of the Pearson correlation coefficient (r) for each of the comparisons is depicted in the lower right corner of each panel.

Venn diagrams were then used to visualise the reproducibility of the different pulldown methods in Figure 4.2.3, highlighting that they were also very similar

in how many kinases they were able to enrich reproducibly, with the batch enrichment method performing slightly better than the enrichments using microcolumns. Taken together, these results suggest that enrichments using stacked KBγ in microcolumns do not have an advantage over standard batch enrichment protocols. Considering that the pulldowns using microcolumns were considerably more time consuming, complicated and therefore error-prone, the CRC65 cell line panel was profiled using the standard batch enrichment procedure in microplates, described in section 2.5.2.1.

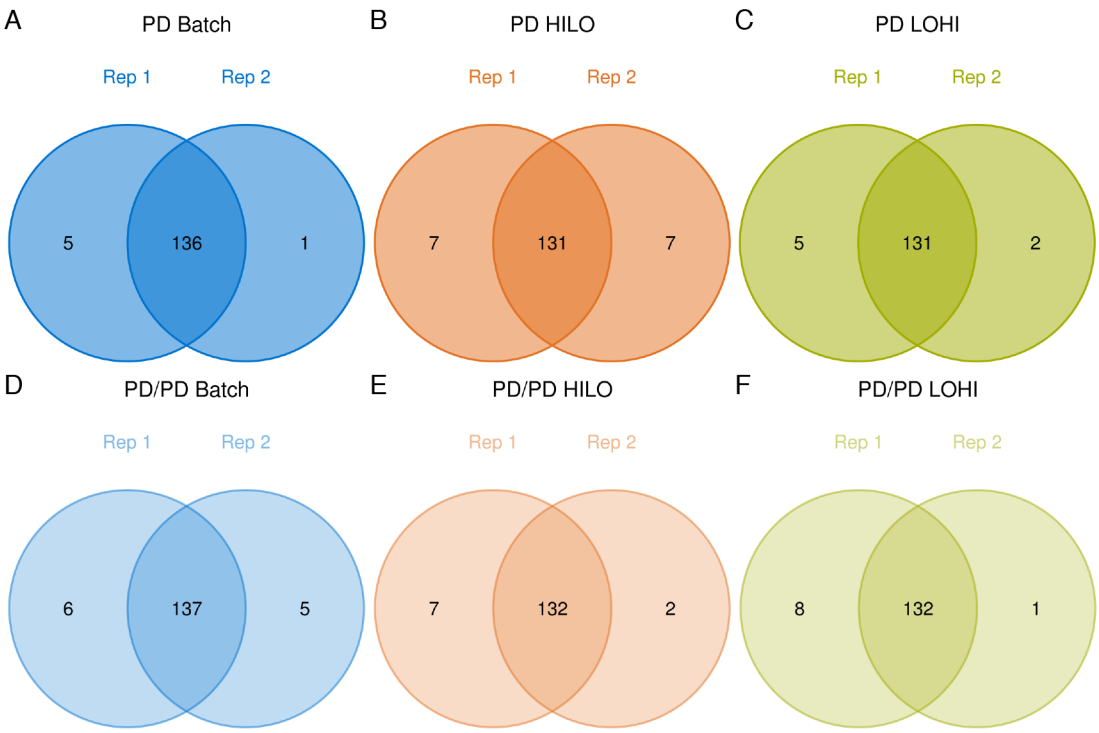


Figure 4.2.3

Venn diagrams of kinases identified in enrichments using microcolumns or batch-mode

Venn diagrams depicting the number of identified and quantified kinase gene groups in enrichments using microcolumns or classical batch-mode enrichment (PD; see sections 2.5.2.2 and 2.5.2.3), as well as in the corresponding pulldowns of pulldowns (PD/PD). MV-4-11 cells were used for the pulldown experiments.

4.3 Buffer comparison: CP versus RIPA100

The standard KBγ protocol in the Küster lab involved performing both the lysis of cells and the actual pulldown using CP buffer (Medard et al., 2015), with a 1:1 dilution step of the lysate before the pulldown to reduce the concentration of detergent, as described elsewhere (Pachl et al., 2013a). In the past, 5 mg of protein was used for each pulldown. The lysis of cells in the CRC65 cell line panel was performed using RIPA100 and the lowest protein concentration achieved was 2.2 mg/mL. This limited the amount of protein to 4.39 mg per pulldown if constant protein amounts should be used (deep-well filter plates had a maximum capacity of 2 mL) and prohibited diluting the lysates 1:1 prior to kinase enrichments using KBγ. Therefore, a comparison of the two buffers and protein amounts was needed. Again, MV-4-11 cells were used as a model system and lysed using either CP or RIPA100 buffer (see section 2.5.2.1). Pulldowns were performed in technical triplicates using 5 mg or 4.39 mg of protein in 2 mL lysate, following the standard protocol for CP buffer (including the initial 1:1 dilution step; see Pachl et al.) and an amended protocol for RIPA100 buffer, which omitted the initial 1:1 dilution.

Figure 4.3.1 depicts scatterplots of LFQ kinase intensities for all samples against the first replicate of pulldowns using 5 mg of protein in CP buffer (CP 5 mg Rep 1) as a common reference. Focusing on the pulldowns in CP buffer using 5 mg (blue) and 4.39 mg (green) of protein in 2 mL lysis buffer (top two rows), no striking differences in the LFQ intensities between the different samples could be observed. In contrast to this, pulldowns in RIPA100 buffer using 5 mg (orange) or 4.39 mg protein (red) seemed to result in lower LFQ intensities for some

kinases compared to pulldowns using CP buffer, as evident by the shift of a portion of the data points away from the diagonal in the direction of the x-axis. However, this shift seemed to be consistent between the different pulldowns in RIPA100 buffer, independent of the amount of protein used for each pulldown.

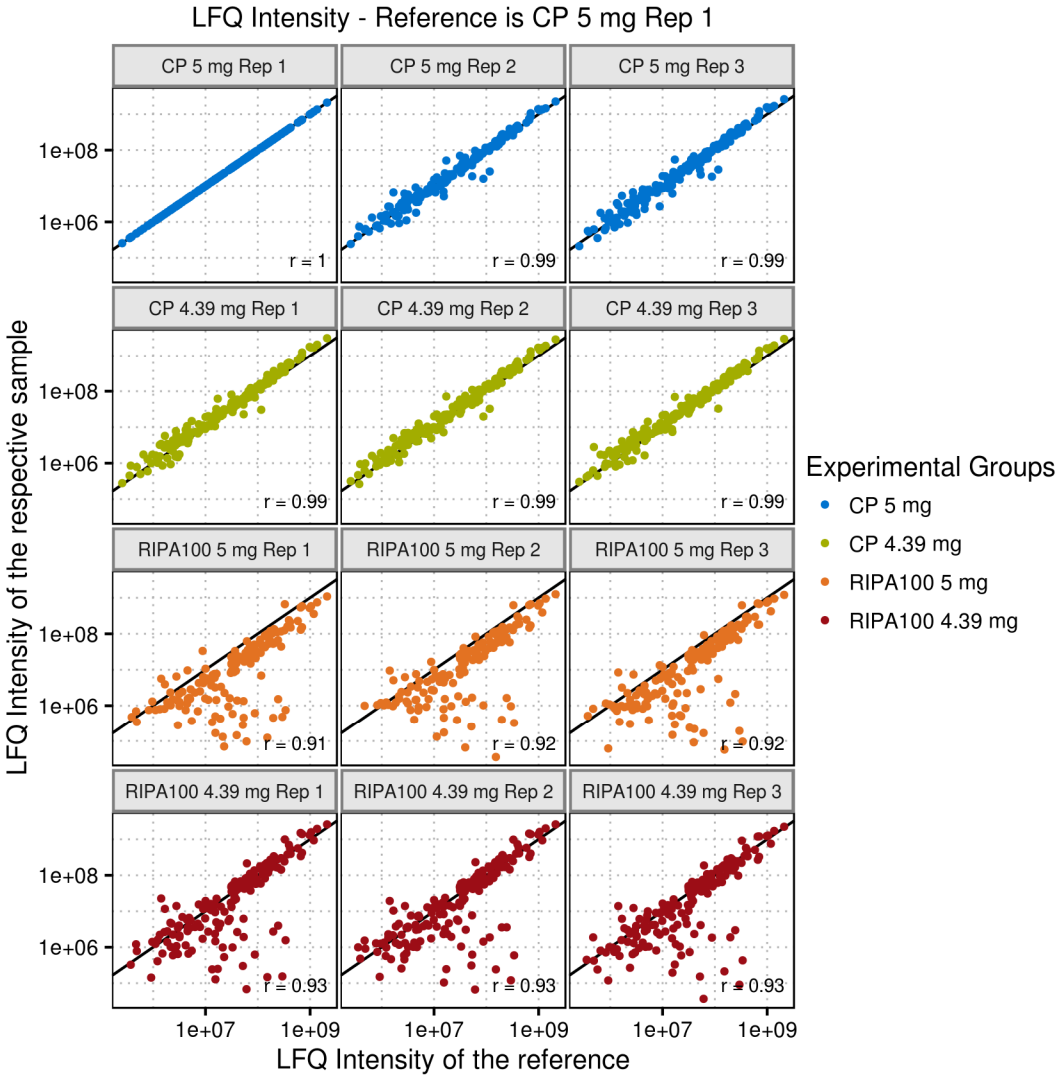


Figure 4.3.1

Scatterplots of LFQ Intensities for kinases from cells lysed with CP or RIPA100 buffer

Scatterplots visualising LFQ Intensities for kinases identified in Kinobeads pulldowns (microplates; see section 2.5.2.1) from cells lysed with either CP or RIPA100 buffer as a function of a common reference (CP 5 mg Rep 1). Only data points present in both the reference and the corresponding experiment are shown. MV-4-11 cells were used for the pulldown experiments. The total protein

amount in the lysate during the incubation with KBγ was either 5 mg or 4.39 mg. Columns represent technical replicates. The value of the Pearson correlation coefficient (r) for each of the comparisons is depicted in the lower right corner of each panel.

Visualising the reproducibility of kinase quantifications for the different experimental setups using Venn diagrams (Figure 4.3.2) showed that using CP buffer for both cell lysis and pulldowns not only resulted in higher LFQ intensities for some kinases independent of the protein amount in the pulldown, but also led to the quantification of significantly more kinases when using 4.39 mg (p=0.026; paired t-test) and 5 mg of protein per pulldown (p=0.014; paired t-test). The amount of protein per pulldown on the other hand did not seem to significantly affect the number of kinases quantified using CP (p=0.63; paired t-test) or RIPA100 buffer (p=0.069; paired t-test). This suggested that it would have been better to lyse the CRC65 cell line panel using CP instead of RIPA100 buffer, while at the same time implying that the amount of protein in each pulldown does not affect the number of kinases quantified, as long as it is between 4.39 mg and 5 mg.

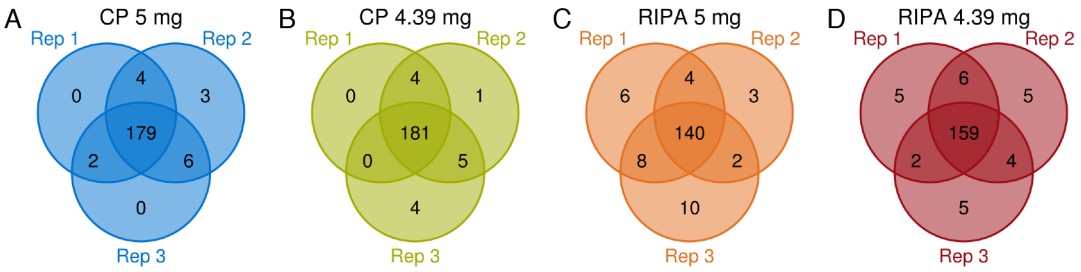


Figure 4.3.2

Venn diagrams of kinases identified in enrichments from cells lysed with CP or RIPA100 buffer

Venn diagrams depicting the number of identified and quantified kinase gene groups in enrichments from cells lysed with CP or RIPA100 buffer. MV-4-11 cells

were used for the pulldown experiments. The total protein amount in the lysate during the incubation with KBγ was either 5 mg or 4.39 mg.

4.4 Optimisation of the pulldown volume

While the total amount of protein in each pulldown did not seem to have an effect on the number of kinases quantified, the final protein concentration could have had an influence. Therefore, pulldowns were performed using 4 mg of protein in 1.8 mL, 1.4 mL, 1 mL and 0.33 mL of RIPA100 buffer. The maximum volume was set to 1.8 mL in order to avoid overflowing and cross-contamination of separate wells on the 96-well filter plates. MV-4-11 cells were again used as a model system, diluting one TCL in RIPA100 to the protein concentrations described above prior to ultracentrifugation in technical triplicates.

Figure 4.4.1 visualises the scatter with respect to kinase LFQ intensities of the different samples as a function of a common reference (1.8 mL RIPA100 Rep 1). As seen by the shift of the entire data cloud away from the diagonal for pulldown volumes smaller than 1.8 mL, it seemed to be beneficial to keep the pulldown volume as low as possible. This was confirmed by Venn diagrams focusing on the reproducibility of kinase quantifications in the four experimental groups, shown in Figure 4.4.2. Going from 1.8 mL to 0.33 mL final pulldown volume, a successive increase in the number of quantified kinases could be observed, which was significant for each volume step if all pairwise comparisons were performed (maximum FDR-adjusted $p=0.024$).

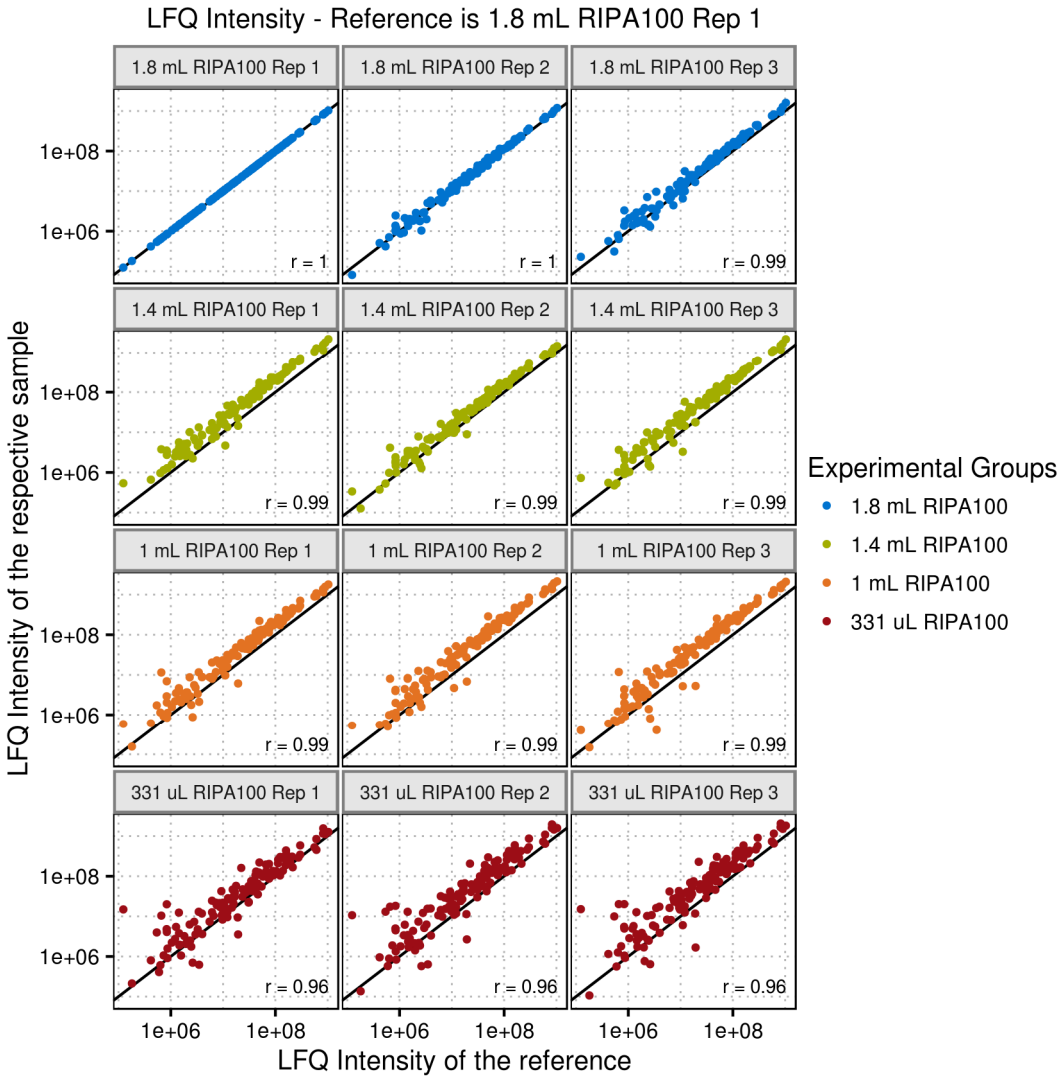


Figure 4.4.1

Scatterplots of LFQ Intensities for kinases enriched from varying volumes of RIPA100 buffer

Scatterplots visualising LFQ Intensities for kinases identified in Kinobeads pulldowns (microplates; see section 2.5.2.1) from variable volumes of RIPA100 buffer as a function of a common reference (1.8 mL RIPA100). Only data points present in both the reference and the corresponding experiment are shown. MV-4-11 cells were used for the pulldown experiments. The total protein amount in the lysate during the incubation with KBγ was 4 mg. Columns represent technical replicates. The value of the Pearson correlation coefficient (r) for each of the comparisons is depicted in the lower right corner of each panel.

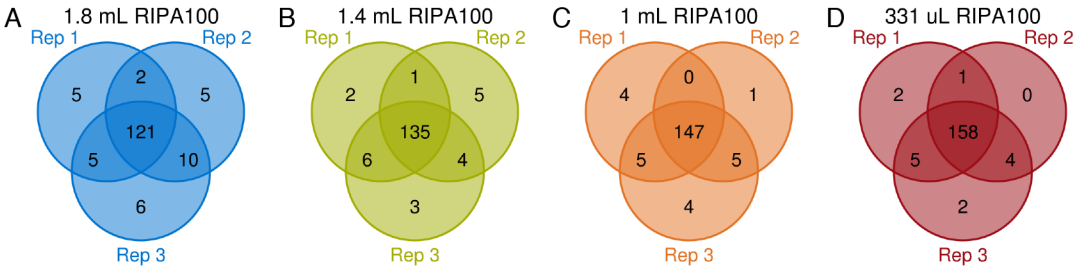


Figure 4.4.2

Venn diagrams of kinases identified in enrichments from varying volumes of RIPA100 buffer

Venn diagrams depicting the number of identified and quantified kinase gene groups in enrichments from varying volumes of RIPA100 buffer. MV-4-11 cells were used for the pulldown experiments. The total protein amount in the lysate during the incubation with KBγ was 4 mg.

Therefore, the CRC65 cell line panel was processed using KBγ and the standard microplate enrichment protocol, keeping the total amount of protein per pulldown, as well as the pulldown volume and by extension also the protein concentration in each pulldown the same, as described in section 2.5.2.1. Hence, these experiments were limited by the sample with the lowest protein concentration (2.2 mg/mL), which is why the final pulldowns were performed using 4 mg of protein in 1.8 mL of RIPA100 buffer.

4.5 Kinase identifications in the CRC65 dataset

The Kinobeads pulldowns for the final kinome profiling of the CRC65 cell line panel were carried out in three batches; one for each biological replicate. Because these experiments were performed over the course of several weeks, the possibility of batch effects between the different biological replicates was explored following the quantification of the pulldown eluates by mass spectrometry as described in sections 2.7.3 and 2.7.4. The distributions of gene group intensities from the different Kinobeads pulldowns were median-centred

separately for each batch, followed by the visualisation of intensity distributions with respect to kinase gene groups using boxplots, as depicted in Figure 4.5.1.

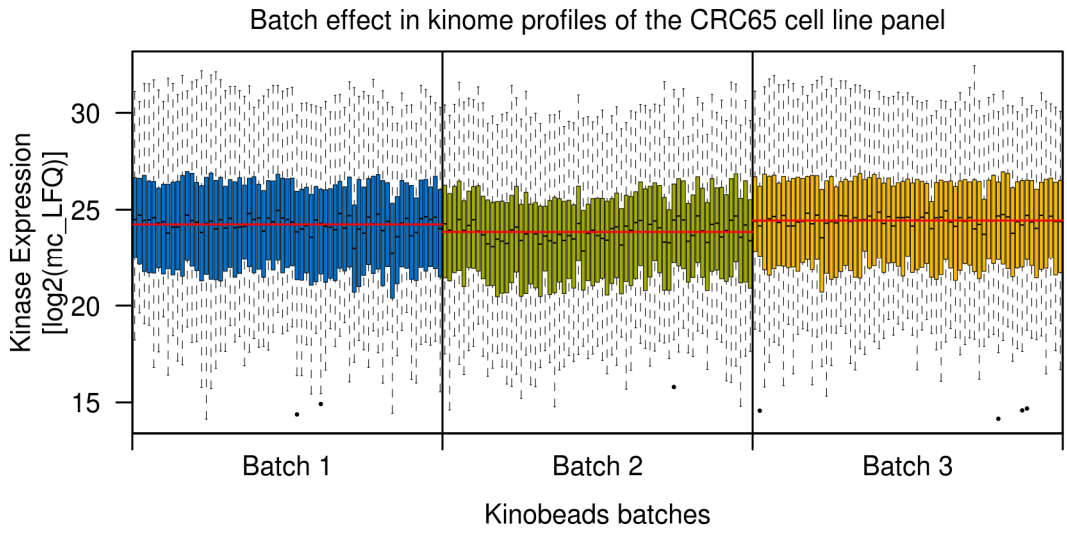


Figure 4.5.1

Batch effect in the kinome profiles of the CRC65 cell line panel

Boxplots visualising the distribution of kinase expression (log-2 transformed and median-centred LFQ values) for each cell line in the CRC65 panel. The different biological replicates are coloured according to the batches they were processed in. For each batch, a red horizontal line visualises the median kinase expression in this batch, which is different for each batch, even after LFQ normalisation by MaxQuant.

As seen by a shift in the batch-wise medians (red lines) relative to each other, the second batch seemed to have consistently lower kinase LFQ intensities than the other two batches. Once batch effects between the biological replicates were removed using ComBat (Johnson et al., 2007), this difference in the batch-wise medians was negligible (see Figure 4.5.2).

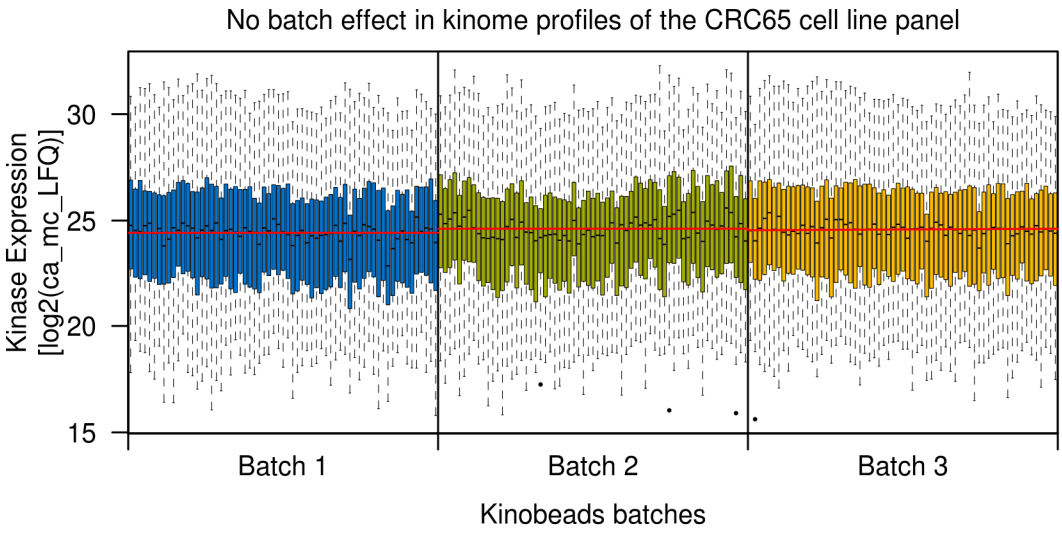


Figure 4.5.2

Batch effect in the kinome profiles of the CRC65 cell line panel removed

Boxplots visualising the distribution of kinase expression (log-2 transformed and median-centred LFQ values after ComBat adjustment) for each cell line in the CRC65 panel. The different biological replicates are coloured according to the batches they were processed in. For each batch, a red horizontal line visualises the median kinase expression in this batch, which is similar for each batch after ComBat adjustment.

This systematic batch effect could have been the result of differences between the biological replicates, differences in the sample preparation between the batches or variation in the performance of the mass spectrometer over time. While the former was relatively unlikely because the different biological replicates of lysates for each cell line were generated successively before moving to the next cell line, differences in the sample preparation or the machine performance were more likely, since these steps were actually performed in batches.

In order to investigate whether machine performance varied between the different Kinobeads batches, an extensive QC pipeline was implemented for MaxQuant output data, part of which was the visualisation of peptide fragment ion mass deviations from their theoretical peptide fragment masses as a function

of the m/z-ratio of the respective fragments. This led to the discovery of systematic fragment mass deviations for the majority of samples in the first and second processing batch (see Figure 4.5.3 for an example), which is why the MS raw data was searched with 120 ppm fragment mass tolerance instead of 20 ppm. Since these deviations occurred in both the first and the second processing batch but not the third, the batch effect outlined above was likely caused by differences in the sample preparation between the processing batches, however an involvement of differences in machine performance can not formally be excluded.

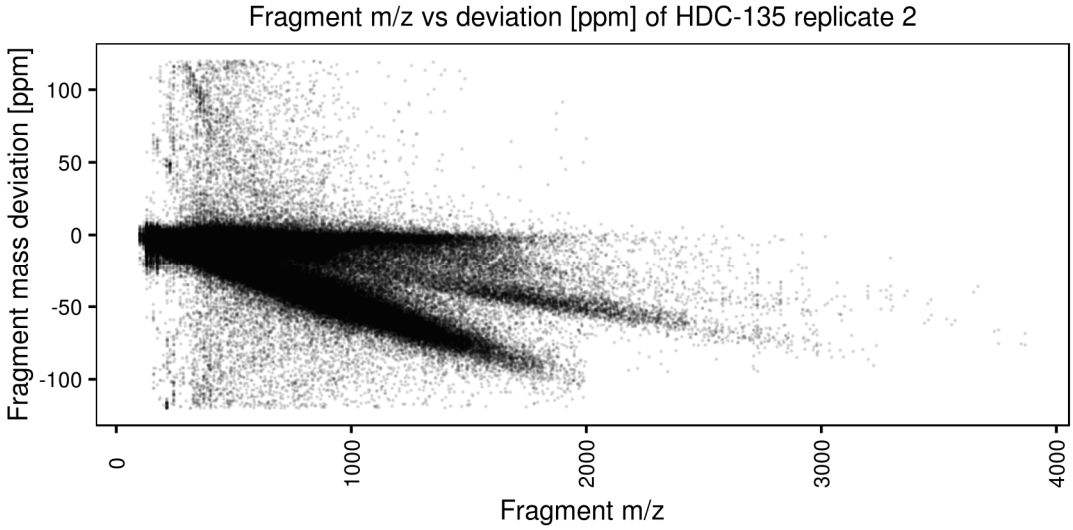


Figure 4.5.3

Example of the systematic fragment mass deviation observed in the Kinobeads dataset

Scatterplot visualising the deviation of the measured fragment mass from the theoretical fragment mass in ppm as a function of the fragment m/z-ratio for all measured peptide fragment ions in the Kinobeads pulldown of the second biological replicate of HDC-135. Each dot represents a single measured peptide fragment ion used by MaxQuant for the comparison of measured and theoretical MS2 spectra during the database search. A linear dependency of the fragment mass deviation on the fragment m/z-ratio is evident, which was observed for several measurements and required the opening of the fragment ion tolerance window to 120 ppm.

Processing the Kinobeads eluates as described in the methods section finally resulted in the identification of 235 of more than 500 known (Manning et al., 2002) human kinases (median/cell line=155, see Figure 4.5.4). The CPTAC dataset did not contain Kinobeads experiments, however 181 of the 6,727 gene groups were kinase gene groups (median/patient=122, see Figure 4.5.4). As already observed for the full proteome data, the CRC65 dataset contained most of the CPTAC dataset and showed a complement of twice as many kinases as the complement of the CPTAC dataset, visualised in Figure 4.5.5. This highlights the utility of Kinobeads experiments for the quantification of kinases in a reasonable timeframe.

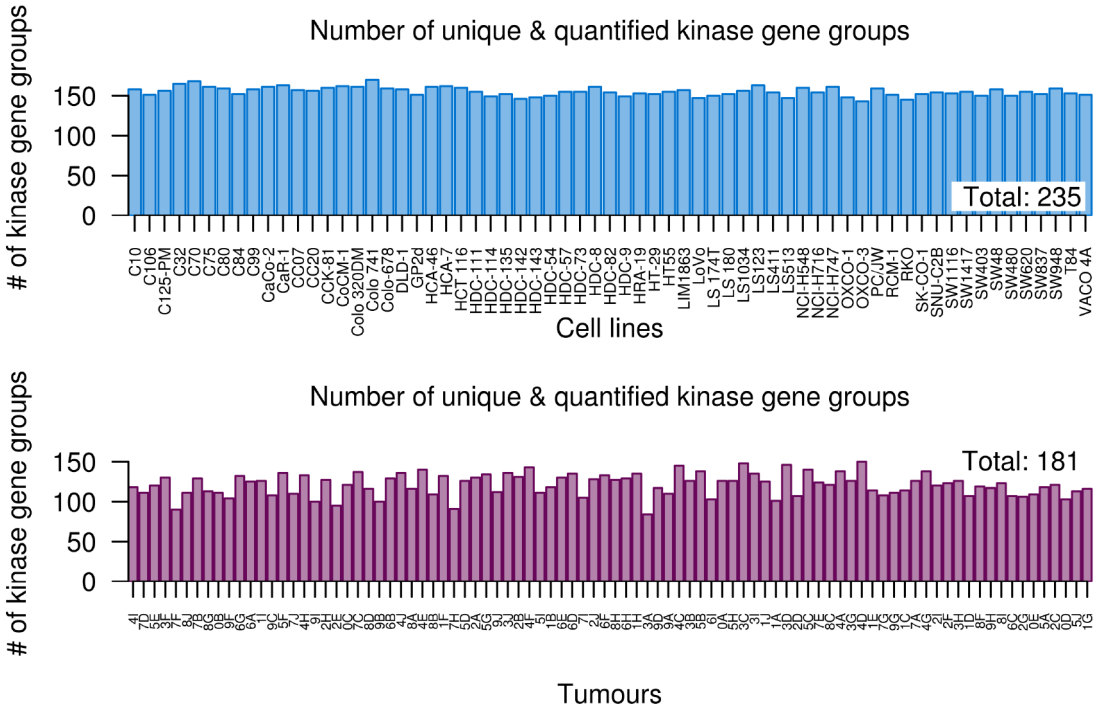


Figure 4.5.4

Kinome identifications in the CRC65 and CPTAC datasets

Bar charts visualising the number of unique identified and quantified kinase gene groups in the Kinobeads enrichments of the CRC65 cell line panel (blue) and extracted from the full proteome quantification of the CPTAC dataset (purple).

Insets indicate the corresponding number of identifications across the entire dataset.

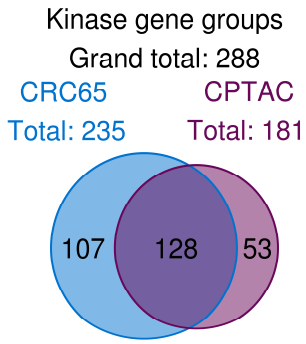


Figure 4.5.5

Venn diagram of kinome identifications

Venn diagram showing the intersection and complements with respect to kinase gene group identifications across both the CRC65 and CPTAC dataset.

Findings by Ruprecht et al. (2015) suggested that binding of kinases to kinase inhibitors on solid supports is mostly independent of kinase activity, which is why western blots were performed in order to compare kinase abundance measured by densitometry and Kinobeads/mass spectrometry for selected kinases (see Figure 4.5.6).

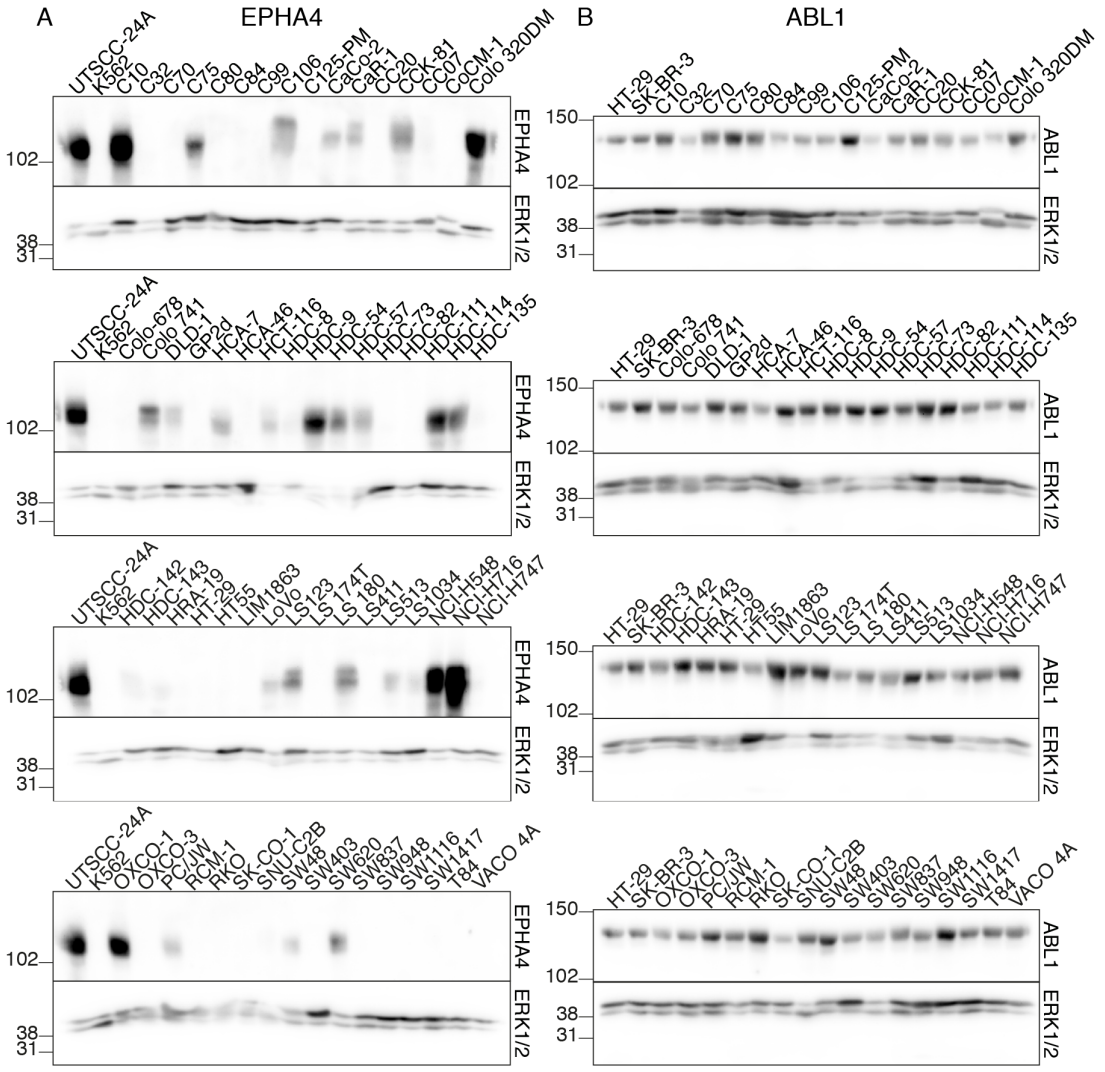


Figure 4.5.6

Western blots of EPHA4 and ABL1 across the CRC65 cell line panel

Western blots visualising the expression of (A) EPHA4 and (B) ABL1 across the CRC65 cell line panel; ERK1/2 was used as a loading control. Control cell lysates with previously documented presence or absence of target proteins are UTSCC-24A (head and neck squamous cell carcinoma, Wu et al., 2011) and SK-BR-3 (breast adenocarcinoma, Wang et al., 2016), as well as K562 cells (chronic myeloid leukaemia, Medard et al., 2015), respectively.

As seen in Figure 4.5.7 A, the quantification of ABL1 and EPHA4 using Kinobeads/mass spectrometry correlated well with the quantification using western blots and densitometry, suggesting that kinase abundance may indeed be the main factor influencing the binding of these kinases to Kinobeads. Since the Kinobeads-based quantification was highly reproducible across biological

replicates (Figure 4.5.7 B), it represented an excellent addition to the pharmacoproteomic characterisation of human colon and rectal cancer, described in this thesis.

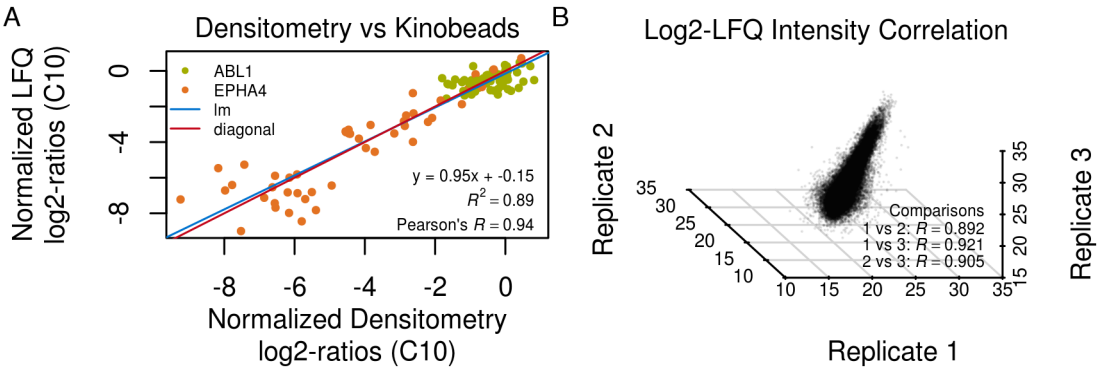


Figure 4.5.7

Correlation of LFQ and densitometry and reproducibility of biological triplicates

(A) Scatterplot visualising the correlation between normalised densitometry log2-ratios relative to C10 and normalised LFQ log2-ratios relative to C10 for EPHA4 and ABL1 (diagonal: x=y; lm: linear model; R^2: coefficient of determination). Densitometry was performed using Fiji (Schindelin et al., 2012), expressing EPHA4 and ABL1 expression relative to the respective ERK1/2 signal, followed by dividing all expression values by the expression value of C10 and log2-transformation. LFQ log2-ratios were based on log2-transformed & median-centred LFQ intensities after ComBat adjustment. Log2-transformation was reverted and all LFQ values were divided by the LFQ value of C10, followed by log2-transformation of the resulting ratios. (B) 3-D scatterplot showing the correlation of LFQ values for all proteins across all cell lines in the three biological replicates of the Kinobeads pulldowns after ComBat adjustment (R: Pearson's R).

4.6 Discussion

As mentioned in section 4.1, kinases are a family of proteins with various crucial roles in heath and disease. Targeting for example EGFR in advanced, chemorefractory CRC has already proven effective (Cunningham et al., 2004; Van Cutsem et al., 2007), but suffers from limited clinical impact in patients with activating mutations of proteins downstream of the receptor (mainly KRAS; Amado et al., 2008; De Roock et al., 2010; Di Nicolantonio et al., 2008; Karapetis

et al., 2008) or low copy numbers of EGFR (Cappuzzo et al., 2008; Moroni et al., 2005). Almost all patients treated with EGFR-targeted monoclonal antibodies like for example cetuximab or panitumumab eventually develop resistance to the treatment through mutation of the receptor (Bertotti et al., 2015), mutation of other downstream signalling proteins like KRAS (Misale et al., 2012), or via expansion of pre-existing clones with mutations in other signalling proteins like KRAS (Diaz Jr et al., 2012; Siravegna et al., 2015). Dual targeting of EGFR using two non-overlapping monoclonal antibodies (Sanchez-Martin et al., 2016) or a single monoclonal antibody combined with small molecule kinase inhibitors targeting EGFR has led to promising results in both cell lines and patients with wild type KRAS (Matar et al., 2004; Weickhardt et al., 2012), while co-treatment of cell lines with cetuximab targeting EGFR and dasatinib targeting Src-family kinases was shown to sensitise even KRAS mutant cell lines to cetuximab therapy (Dunn et al., 2011). Using predictive biomarkers like KRAS as a selection criterion for EGFR-targeted therapy with monoclonal antibodies in the spirit of personalised medicine ensures that only patients responsive to the treatment actually receive it (Siena et al., 2009), but targeted combination therapy is emerging as the method of choice to overcome acquired resistance (Bozic et al., 2013). Identifying new kinases as drug targets in CRC is therefore likely to open up additional windows for therapeutic intervention and will possibly also facilitate the development of novel combination therapies. Kinobeads have proven to be an effective method to achieve this goal (Wu et al., 2011), which is why their inclusion in this study enriched the pharmacoproteomic characterisation of human colon and rectal cancer with a dedicated approach focused on this important protein family. The pulldown protocol was optimised

for the enrichment of kinases from TCLs in RIPA100 buffer, tightly controlling both the volume of and the protein amount in each pulldown. This resulted in high quality kinome profiles of the CRC65 cell line panel, which were reproducible across biological replicates and correlated well with the densitometry-based quantification of selected kinases. This Kinobeads dataset therefore served as an integral part of the multi-omics data integration strategy, described in the next chapter.

Chapter 5
Multi-omics data integration and validation

5.1 Introduction

With a comprehensive characterisation of the transcriptomes and full proteomes with respect to the CRC65 cell line and CPTAC patient cohort at hand, the next step was to integrate these datasets and combine them with the kinome profiles of the cell lines described in the preceding chapter, in order to be able to finalise the characterisation of this pharmacoproteomic training and testing dataset for the modelling of drug sensitivity in CRC (see also section 1.4). Recent multi-omics data integration efforts in CRC focused on the identification of differentially expressed protein-protein-interaction sub-networks using proteomics and genomics data (Nibbe et al., 2010), studied prognostic biomarkers for relapse after surgery in genomic data from tissues and metabolomic data from biofluids (Madhavan et al., 2013) or investigated global gene expression, copy number variation and protein expression changes of several CRC patients (Zhang et al., 2014b). Other studies integrated multiple genomic readouts to find functionally important regions of differential DNA methylation (Fan et al., 2015), developed online resources for the visualisation and integrated analysis of a variety of omics data (Shi et al., 2013; Zhu et al., 2015) or applied fuzzy logic modelling of gene expression, copy number variation and mutation data to gain insight into gene activity (Pavel et al., 2016). Some recent publications even combined several transcriptomic datasets in search of a classifier predicting response to 5-FU based on relative expression orderings (Tong et al., 2016), used a resampling-based regression model on transcriptomics and proteomics data to detect differential post-transcriptional regulation between MSI and MSS tumours (Liu and Zhang,

2016) or made use of consensus principal component analysis to define a set of joint latent variables between different omics data types in order to establish an overarching subtype model of CRC (Meng et al., 2016). While most of these studies looked at patients *or* cell lines, none of them attempted to apply multi-omics data integration between cell lines *and* patients in order to find *integrated* subtypes of CRC and establish cell-line-based model systems suitable for studying differential drug sensitivity between these subtypes. Here, for the first time, access to in-depth transcriptomics and proteomics data on the CRC65 cell line and CPTAC patient cohort enabled just that. The availability of these large-scale datasets fuelled the most extensive comparison of steady-state mRNA and protein abundance for human colon and rectal cancer to date, which is described in section 5.2 for the CRC65 and CPTAC datasets. It starts out with an exploration of protein/mRNA ratios as proxies for translation rates in section 5.2.1 based on results published by Wilhelm et al. (2014), followed by an evaluation of the global protein/mRNA correlation for both datasets in section 5.2.2, which in turn led to the development of a method for mRNA-guided missing value imputation based on the results outlined in section 5.2.1. Section 5.2.3 then focuses more on the detection limit of mass spectrometers and how it influenced the likelihood of encountering missing values in the proteomics experiments described in this thesis, followed by the proposal and application of minimum-guided missing value imputation. After imputing missing values in the CRC65 and CPTAC datasets and following their combination as described in section 2.10, the combined full proteome dataset on CRC was used for the discovery and characterisation of integrated FPSs consisting of cell lines and patients, laid out in section 5.3. In section 5.4, these subtypes were then compared to Kinobeads

Subtypes defined based on the kinome profiles of the CRC65 cell line panel (see Chapter 4), followed by their characterisation with respect to kinases differentially expressed between them, which concluded the characterisation of the pharmacoproteomic training and testing dataset. The following sections then focus on the integration of the proteomic data with public drug sensitivity datasets and the pharmacoproteomic modelling of drug sensitivity as a function of full proteome (section 5.5) and kinome profile (section 5.6), resulting in the identification of differential drug sensitivities between FPSs and the discovery of MERTK as a predictive proteomic marker of drug response. This predictive role of MERTK was subsequently validated using *in vitro* co-treatment experiments (section 5.7), while immunohistochemistry suggested that MERTK is also a prognostic biomarker in CRC (section 5.8).

5.2 Integration of full proteome and transcriptome data

5.2.1 Protein/mRNA ratios

As mentioned above, the first step with respect to the integration of transcriptomics and proteomics data was an exploration of protein/mRNA ratios in both the CRC65 and CPTAC dataset. Wilhelm et al. showed that protein/mRNA ratios were constant across a number of different tissues and that median ratios could be used as proxies for translation rates in order to predict protein abundance from mRNA abundance (Wilhelm et al., 2014). Since these ratios were calculated based on RPKM values as determined by mRNA-Seq, the question arose as to whether this also holds true when transcript abundance is estimated using integrated data from both mRNA-Seq and microarray experiments. Therefore, the giBAQ value of each protein in each sample was de-logarithmised

and divided by the corresponding de-logarithmised median expression of each transcript in each sample across the different mRNA datasets following their combination using the scheme by Guinney et al. (2015), as described in section 2.10.1.

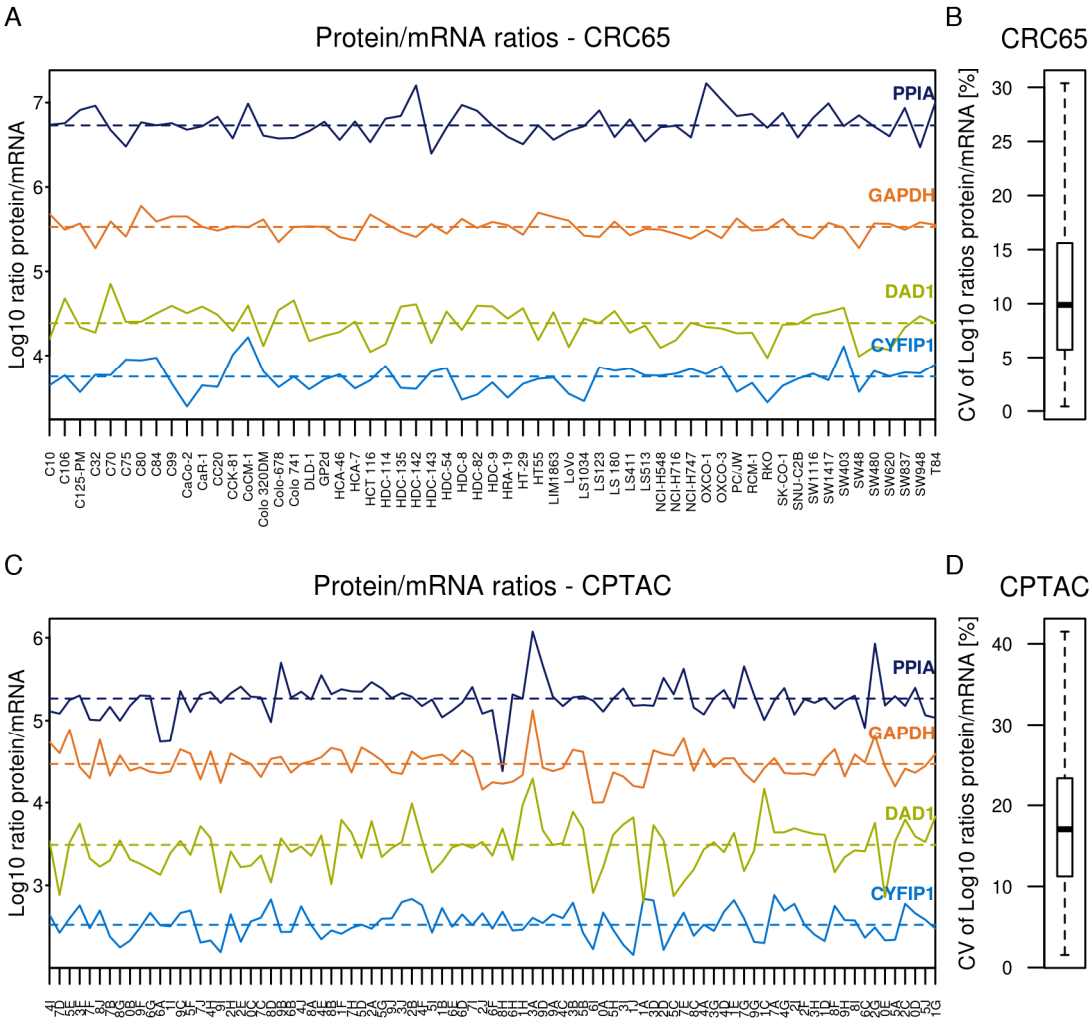


Figure 5.2.1

Stable Protein/mRNA ratios across the CRC65 and CPTAC datasets

(A & C) Protein/mRNA log-ratios of four example proteins across the (A) CRC65 cell line and (C) CPTAC patient dataset. (B & D) Distribution of the coefficients of variation (CV) of Protein/mRNA log-ratios across the (B) CRC65 and (D) CPTAC datasets in percent. Outliers were removed for visual clarity.

Figure 5.2.1 depicts log-ratios for four example proteins (PPIA, GAPDH, DAD1 and CYFIP1) across the CRC65 (A) and CPTAC datasets (C) and also visualises the distribution of the coefficients of variation for all protein/mRNA log-ratios across these two datasets as boxplots (B and D). As indicated by these ratio examples and the accompanying boxplots, protein/mRNA ratios in the CRC65 and CPTAC datasets were surprisingly stable, suggesting that they could be treated as systematic differences between protein and mRNA measurements and therefore adjusted using batch effect removal methods such as ComBat and MComBat (Johnson et al., 2007; Stein et al., 2015).

5.2.2 Protein/mRNA correlation and mRNA-guided missing value imputation

After establishing that protein/mRNA log-ratios were stable across the two datasets, the next step was the evaluation of the protein/mRNA correlation for both datasets. Therefore, both the sample-specific, as well as the overall correlation of median mRNA and protein expression was calculated for each dataset. The overall correlation was defined as the correlation of all median mRNA measurements for all samples against all corresponding protein measurements for all samples in their respective datasets.

Figure 5.2.2 A and B visualises the overall correlations for the CRC65 and CPTAC dataset as two-dimensional histograms in order to reduce over-plotting and increase visual clarity.

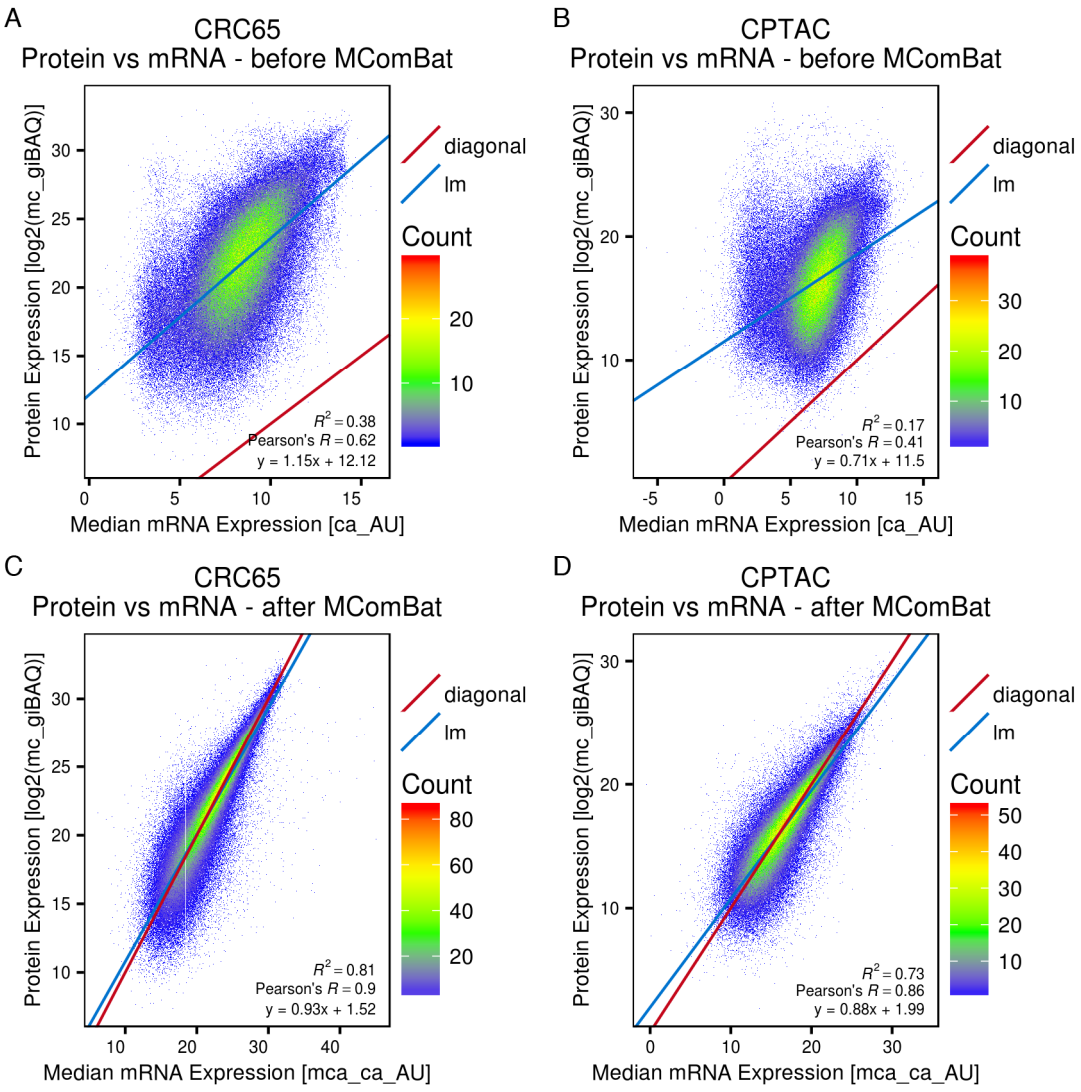


Figure 5.2.2

Correlation of median mRNA Expression and Protein Expression

Two-dimensional histogram (600 bins in each dimension) visualising the median mRNA Expression (A-B) before and (C-D) after MComBat adjustment for all transcripts across all cell lines versus Protein Expression in the (A & C) CRC65 and (B & D) CPTAC datasets (log2-transformed & median-centred giBAQ; diagonal: x=y; lm: linear model; R^2: coefficient of determination). In order to prevent over-plotting, the number of data points in a particular area of the plot (Count) was mapped to a colour-scale as opposed to plotting each data point separately.

As noted previously by Zhang et al. (2014a), all patient samples showed a significant positive protein/mRNA correlation (maximum FDR-adjusted p=1.33x10-29, Pearson's product-moment correlation), with an overall

correlation of 0.41; this was also true for the cell lines (maximum FDR-adjusted p=3.14x10-241, Pearson's product-moment correlation), which showed an overall correlation of 0.62. When constant, batch-effect-like differences such as stable protein/mRNA ratios between the proteomics and transcriptomics datasets were adjusted using a modified version of MComBat (Stein et al., 2015) and the proteomics data as the "gold-standard" batch as described in section 2.10.2, the protein/mRNA correlation increased substantially, as visualised by the two-dimensional histograms in Figure 5.2.2 C and D. With an overall Pearson correlation coefficient of 0.9 for the CRC65 dataset and 0.86 for the CPTAC dataset, this enabled the linear prediction of protein abundance from mRNA abundance, which in turn fuelled mRNA-guided missing value imputation (see section 2.10.3). One concern was that the adjustment of systematic differences between the proteomics and transcriptomics datasets using MComBat blurred differences between individual samples, making it difficult to extract meaningful biological information from them. While the proteomics and transcriptomics datasets indeed became more similar to each other, which was expected, these concerns were unfounded for the proteomics data, since they were not changed at all by MComBat because of their designation as the "gold-standard" batch. This dispelled concerns for an artificial increase in homogeneity for the proteomics datasets, but the concerns for the mRNA datasets remained. Therefore, the internal correlation structure of the mRNA datasets was examined for the CRC65 and CPTAC datasets and subsequently compared to the correlation structure between mRNA and proteomics data, visualised in Figure 5.2.3.

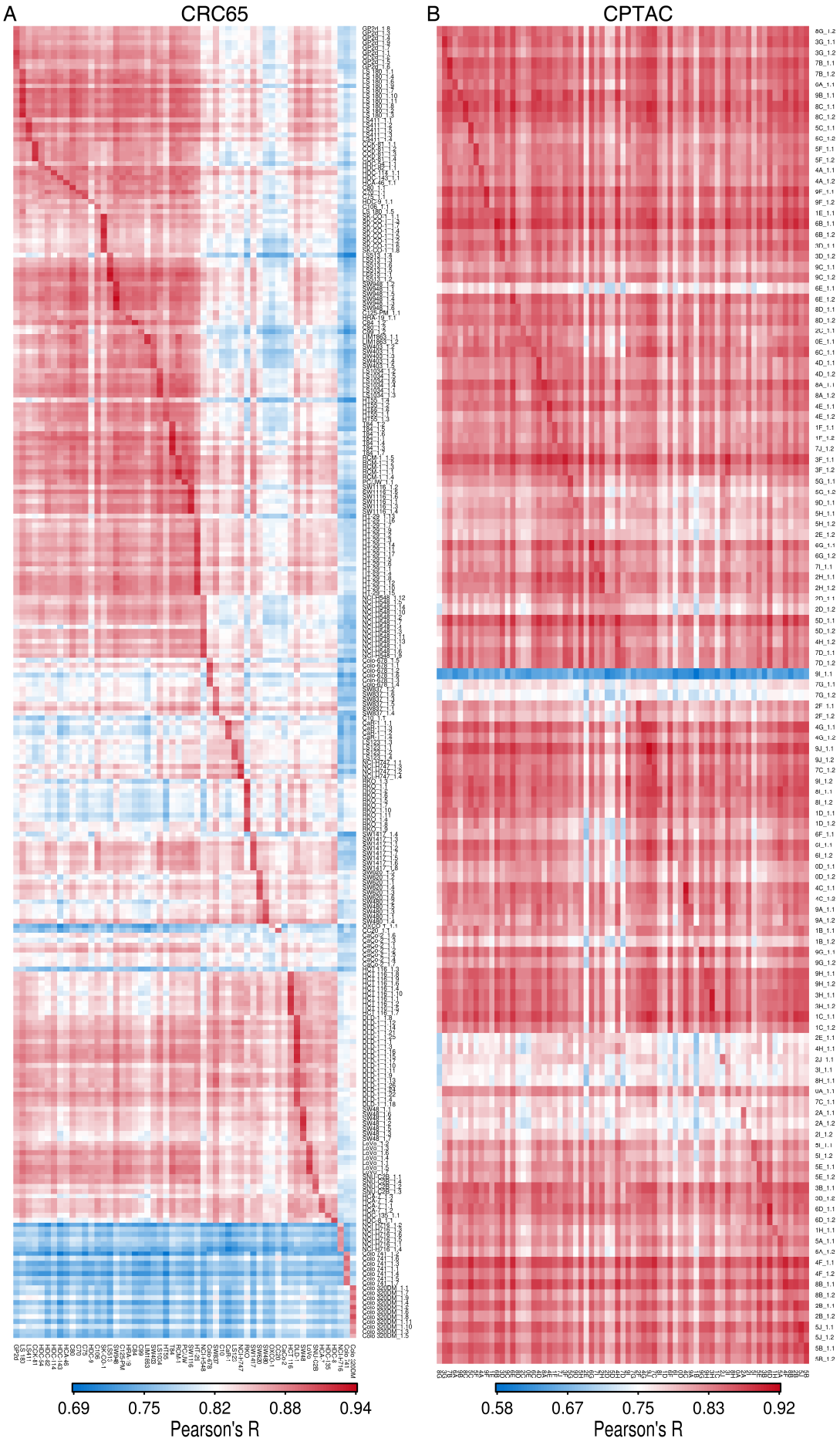

A
CRC65
0.69
0.75
0.82
0.88
0.94
Pearson's R
B
CPTAC
0.58
0.67
0.75
0.83
0.92
Pearson's R

Figure 5.2.3

Correlation heat maps of median mRNA Expression versus Protein Expression

Heat maps of Pearson correlation coefficients between mRNA Expression (y-axes) and Protein expression (x-axes) for the (A) CRC65 (55 protein samples and 272 mRNA samples) and (B) CPTAC datasets (71 protein samples and 125 mRNA samples). Samples in mRNA-space are clustered hierarchically based on their mRNA expression profile (distance metric: 1-Pearson correlation), while samples in Protein-space are ordered according to their occurrence in mRNA-clusters (see also section 2.10.2).

A detailed explanation of how these correlation heat maps were generated can be found in section 2.10.2. Briefly, all pairwise correlations (Pearson's R) between protein and mRNA samples were calculated for both datasets, followed by the exclusion of mRNA samples with a below-average correlation with their cognate protein measurement and the subsequent removal of protein samples without cognate mRNA samples. This way, 22 and 35 mRNA samples (7% & 22%, respectively), as well as 8 and 19 protein samples (13% & 21%, respectively) were excluded from the correlation matrices of the CRC65 and CPTAC dataset, respectively. These correlation matrices were subsequently used to generate Figure 5.2.3 by clustering the mRNA samples (rows) hierarchically using "1-Pearson correlation" as the distance metric (calculated based on the mRNA measurements and not on the correlation matrix) and "ward.D2" as the agglomeration method, while ordering the columns according to the occurrence of the first cognate mRNA sample corresponding to a given protein sample in the rows of the correlation matrices after reordering of the rows. Therefore, the row ordering represents the similarity of different mRNA samples (potentially from different mRNA datasets) and the column ordering matches the ordering of the first cognate mRNA sample for each protein sample in the rows of the correlation heat map after hierarchical clustering of mRNA samples. The heat map itself

depicts Pearson correlation coefficients between protein and mRNA samples. The conclusions drawn from this figure were two-fold. First, one can appreciate that the same mRNA samples from different datasets clustered together, meaning they were more similar to each other than they were to mRNA samples from different cell lines/tumours, dispelling concerns of artificial sample homogeneity in the mRNA datasets. Second, it became evident that the majority of mRNA samples had the highest correlation with their cognate protein samples after MComBat adjustment, which essentially meant that this method enabled the unsupervised matching of protein and mRNA samples. In other words, following MComBat adjustment, it would have been possible to assign the majority of mRNA measurements to the correct cell line based only on the annotation of their cognate protein samples, further emphasising the agreement of protein and mRNA measurements after MComBat-based adjustment of systematic differences between proteomics and transcriptomics data. Encouraged by these results, mRNA-guided missing value imputation as described in section 2.10.3 was carried out to impute as many missing values as possible based on evidence on the transcript level.

5.2.3 Minimum-guided missing value imputation

Following mRNA-guided missing value imputation, some missing values in the proteomics datasets still remained. Since pharmacoproteomic modelling of drug sensitivities as a function of full proteome and kinome profiles required complete data matrices, a selection of secondary imputation methods was explored, which could complement mRNA-guided missing value imputation. All of these methods shared the common assumption that most missing values in proteomics datasets

are “missing not at random” (MNAR), i.e. that the likelihood of “encountering” a missing value increases as the abundance of the respective analyte approaches the detection limit of the instrument (Lazar et al., 2016). First, this assumption was confirmed for both the CRC65 and CPTAC dataset by calculating the frequency and fraction of missing values per intensity bin for both datasets after total-sum-normalisation.

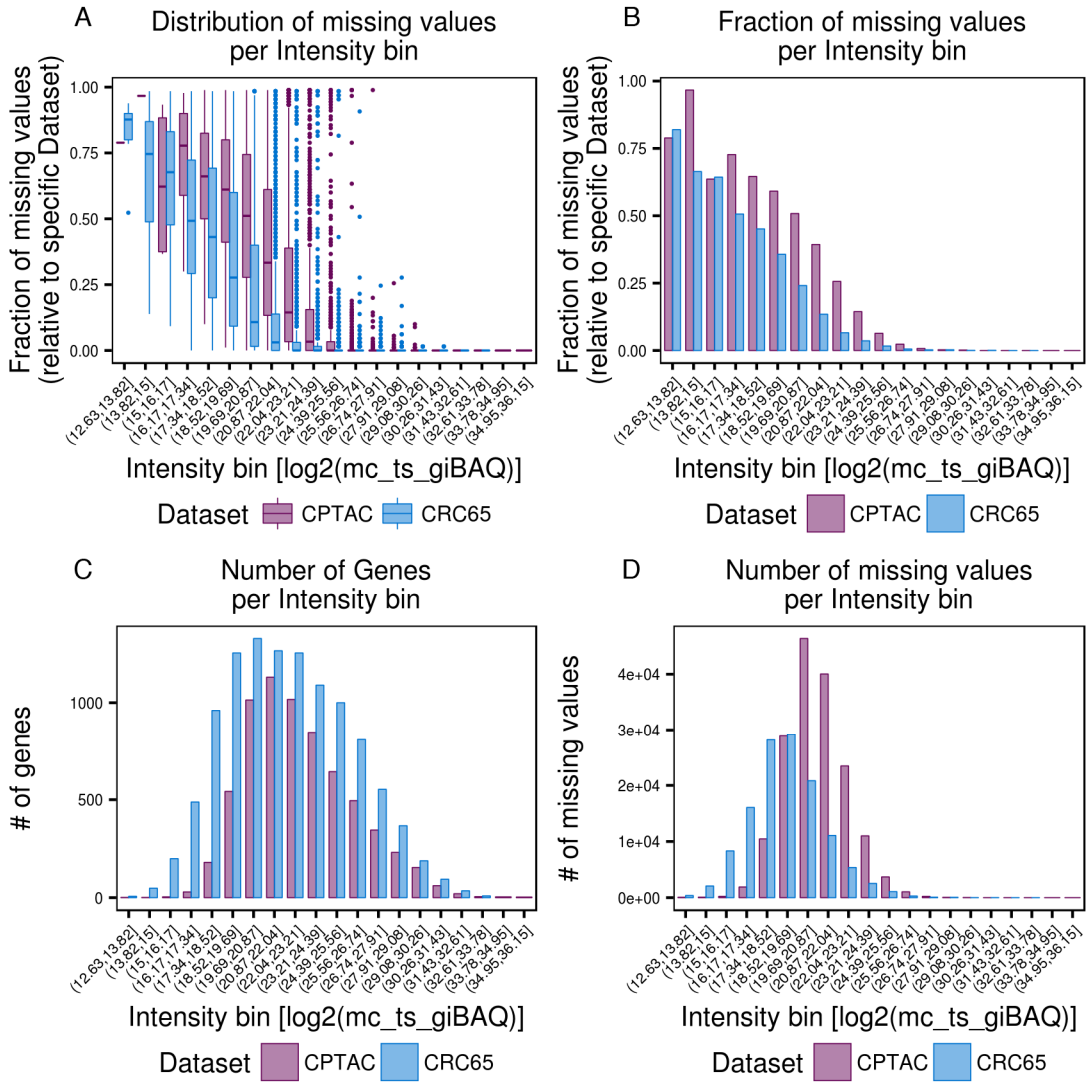


Figure 5.2.4

Relationship between giBAQ intensity and the number of missing values

Plots visualising the relationship between total-sum-normalised giBAQ intensity and the number of missing values. The CRC65 and CPTAC datasets were total-sum-normalised and multiplied by a constant factor (10^{12}) prior to log2-

transformation and median-centring to bring their intensities onto the same scale. (A) Boxplots showing the distribution of the fraction of missing values per intensity bin and gene (total-sum-normalised, log2-transformed & median-centred giBAQ) in both the CRC65 (N=65) and CPTAC (N=90) dataset, relative to size of the respective dataset. (B) Bar chart of the fraction of missing values per intensity bin, relative to the respective dataset. (C) Bar chart of the number of genes per intensity bin. (D) Bar chart of the number of missing values per intensity bin.

For that, each giBAQ intensity of samples in the two datasets was divided by the sum of all giBAQ intensities for the respective sample, before multiplying them with a constant factor (10^{12}), log2-transformation and median-centring, in order to bring the two datasets onto the same scale. Subsequently, the intensity scale was divided into equally sized bins and each protein was assigned to one of these bins based on its median expression across the respective dataset.

Figure 5.2.4 visualises the two datasets side-by-side and shows distributions of the fraction of missing values per intensity bin using boxplots (A; itemising each protein separately), as well as the total fraction of missing values (B; not itemising each protein separately), the number of proteins (C) and the number of missing values per intensity bin (D) using bar charts, respectively. As evident by Figure 5.2.4 A and B, the fraction of missing values per intensity bin – a proxy for the likelihood of "encountering" a missing value – indeed increases as the median protein expression decreases, enabling imputation approaches which take this relationship into consideration.

The first imputation method which was tested ("perseus") fell into the MinProb category described by Lazar et al. (2016) and was the method described in Supplementary Figure 3 of the recent publication by Tyanova et al. (2016), using a down-shift of 1.8 and a width of 0.3. When this method was applied without preceding mRNA-guided missing value imputation, this resulted in strong

bimodality of the intensity distribution for both datasets (see Figure 5.2.5 A and B), but even applying it after mRNA-guided missing value imputation (see Figure 5.2.5 C and D) did not circumvent this undesirable behaviour completely. Since imputing missing values with the sample-wise (MinDet method described by Lazar et al.) or protein-wise minimum introduced too much bias by reducing the variability of the measurements, two new imputation methods termed "minimum-guided missing value imputation on the protein level" and "minimum-guided missing value imputation on the peptide level" were implemented, as described in section 2.10.4. Briefly, the concept these two methods revolved around the idea that no imputed value for a given protein should be bigger than the smallest measured value for the same protein. On the protein level, this was achieved by sampling missing values with replacement from the distribution of measured values, truncated to the right at the minimum for the respective protein.

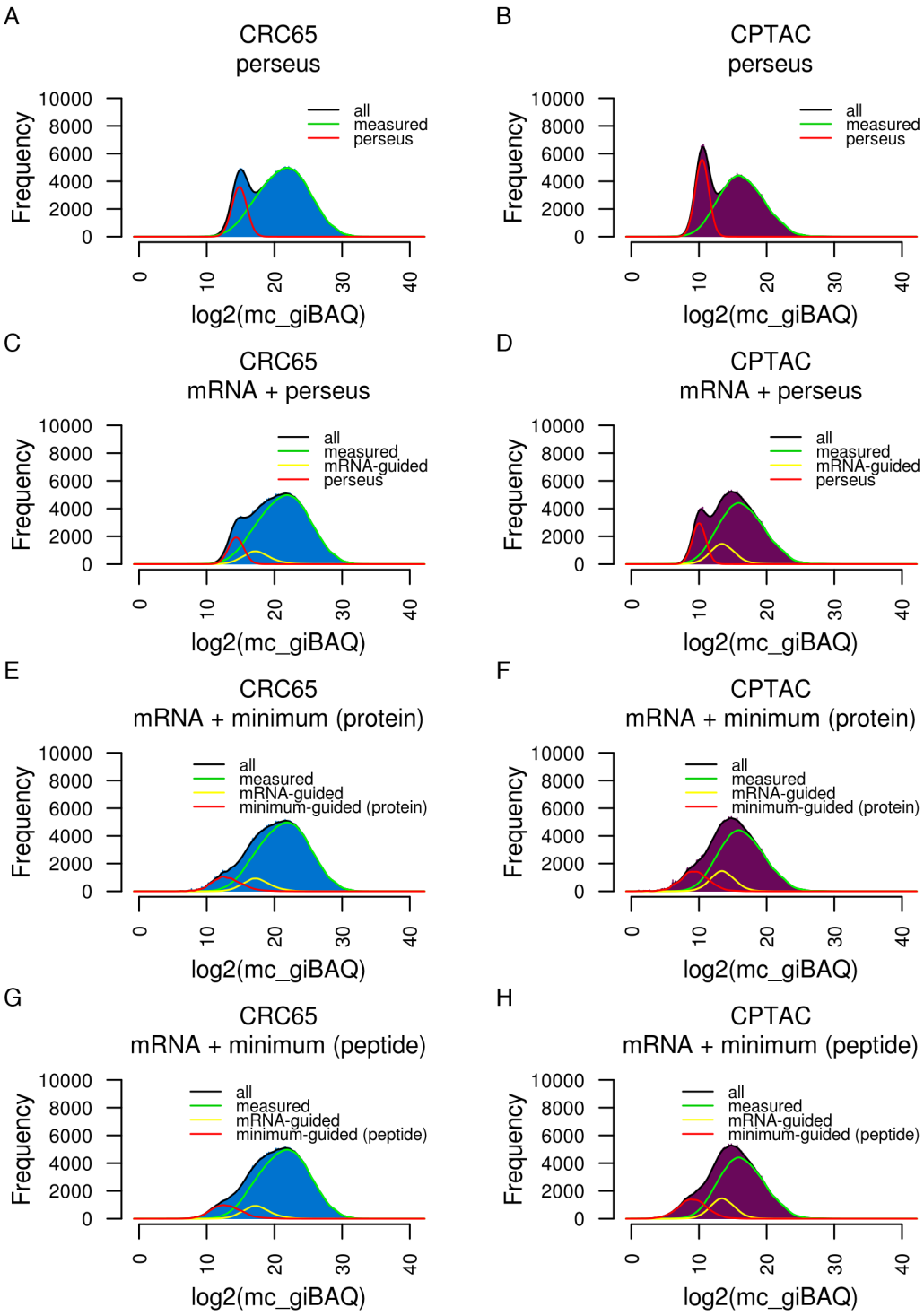


Figure 5.2.5

Evaluation of imputation methods informed by the detection limit

Histograms visualising the distribution of log2-transformed & median-centred giBAQ values of the CRC65 (left column) and CPTAC (right column) datasets after application of (A-B) perseus-type, (C-D) sequential mRNA-guided and perseus-type, as well as sequential mRNA-guided and minimum-guided missing value imputation on the (E-F) protein or (G-H) peptide level, respectively. Additional lines visualise the contribution of measured values (green), values imputed by

mRNA-guided missing value imputation (yellow) and values imputed by either perseus-type or minimum-guided missing-value imputation (red, see also sections 2.10.3 and 2.10.4).

On the peptide level, the algorithm was more complicated, since missing values were sampled from peptide intensities (see section 2.10.4). Briefly, peptides and their intensities making up the minimum protein intensity for any given protein with missing values were found in their respective fraction and at their respective retention times. Each missing value on the protein level was assumed to only consist of these protein-specific reference peptides. For each protein and its corresponding reference peptides, missing values were sampled from the distribution of peptide intensities smaller than or equal to the intensity of the respective reference peptide in a user-defined retention time window, centred on the recalibrated retention time of this reference peptide. Afterwards, missing values on the protein level were calculated by summing up the sampled peptide intensities. By adjusting the retention time window queried for low-intensity peptides, one could control how similar the imputed values were to one another, thereby reducing the variability across the dataset in the low-intensity range. This was useful especially when considering for example clustering algorithms using correlation as a measure of distance between samples, since low-intensity features did not contribute as much to the distance between samples as they would have if they were imputed with a different method, giving more weight to reliably measured features. The user-defined retention time window was the same window MaxQuant used for transferring identifications by "Match between runs", since the concept is somewhat similar. As indicated by Figure 5.2.5 E-H, both minimum-guided imputation methods performed well in that they did not introduce strong bimodality into the distribution of giBAQ values, with

imputation on the peptide level resulting in a smoother intensity histogram. Since imputation on the protein level should not be favoured over imputation on the peptide level because the aggregation of peptide-level information to form protein expression values already represents a form of implicit missing value imputation (discussed by Lazar et al., 2016), missing values in the full proteome datasets were first imputed using mRNA-guided missing value imputation, followed by minimum-guided missing value imputation on the peptide level. This further ensured that subsequent clustering and other machine learning algorithms were less influenced by low-confidence quantifications at the lower end of the detection limit of the mass spectrometer, as described above. Missing values in the Kinobeads dataset were imputed using minimum-guided missing value imputation on the protein level, since LFQ quantitation was used, which involved an additional normalisation step that would interfere with imputation on the peptide level.

5.3 Full Proteome Subtypes (FPSs) of CRC

Since substantially more peptides per sample were identified in the CRC65 dataset than in the CPTAC dataset (Figure 3.3.4 and Figure 3.3.5) and because MaxQuant calculates protein intensities as the sum of peptide intensities, strong batch effects were expected between the two datasets even after missing value imputation, which would have had an unfavourable impact on any clustering algorithm applied to the joined expression matrix. Therefore, batch effects between the two datasets were adjusted using ComBat (Johnson et al., 2007) with MSI status as a covariate, described in section 2.10.5. Afterwards, the combined dataset was used to discover integrated Full Proteome Subtypes of CRC (see

section 2.11.1) and to compare them to the CMS, since the CMS subtype model and the proteomic subtype model published by the CPTAC (Zhang et al., 2014a) only showed partial concordance.

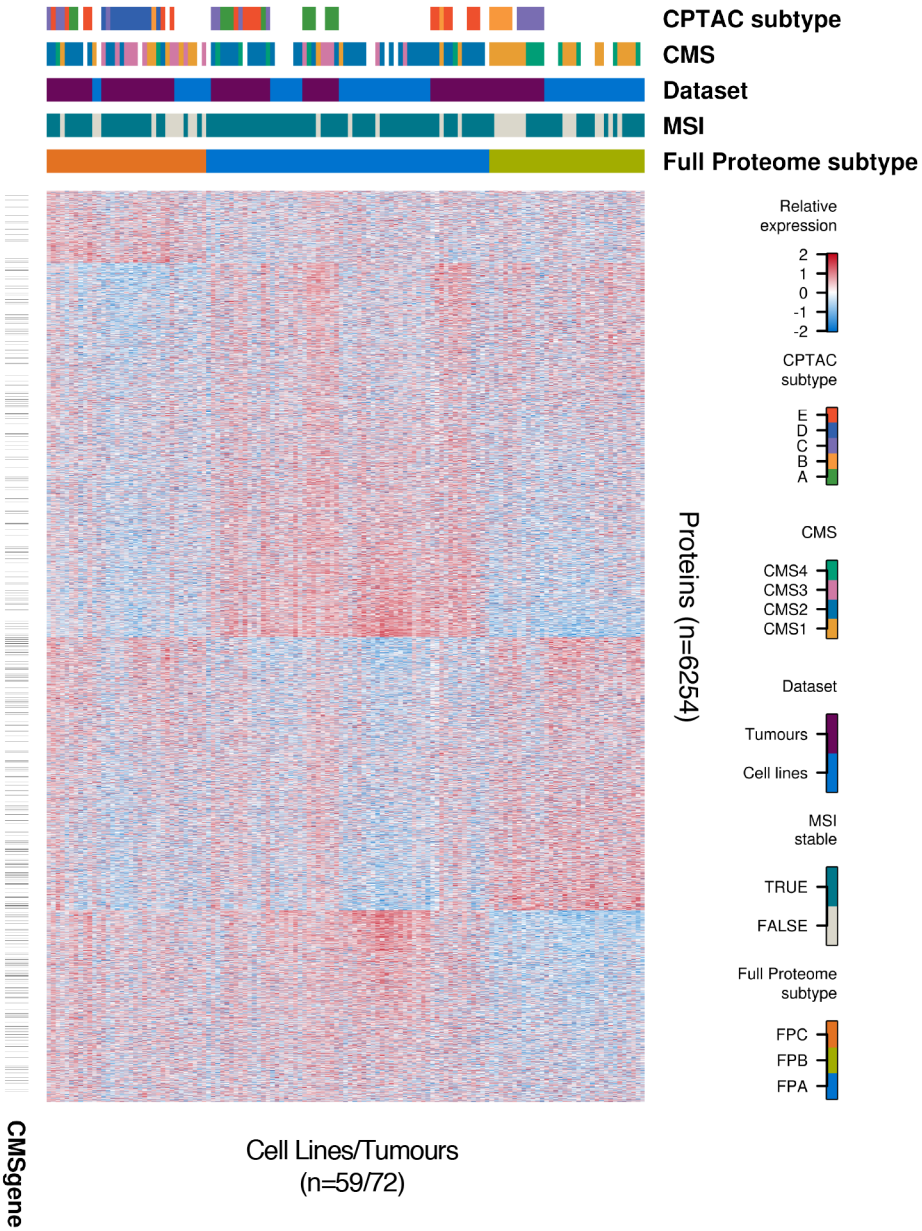


Figure 5.3.1

Integrated Full Proteome Subtypes of CRC

Heat map of standardised, log2-transformed and median-centred giBAQ protein quantification (z-scores) across the combined CRC65/CPTAC dataset. Cell lines/patients are displayed as columns and proteins are shown as rows (all proteins were used to find subtypes). Black bars to the left of the heat map

indicate the presence of the respective protein in the CMS classifier. Annotation bars on top of the heat map visualise the membership of the different cell lines/patients in five annotation categories. Columns are ordered hierarchically according to the consensus clustering results (see section 2.11.1), while rows are ordered according to their association with a particular FPS (see section 2.11.2).

Consensus clustering (Monti et al., 2003) was used in an approach similar to the one employed by Zhang et al. (see section 2.11.1) and identified three Full Proteome Subtypes (FPSs) FPA-FPC (Figure 5.3.1). Each subtype consisted of cell lines as well as patients in a ratio of 28/34 for FPA, 22/12 for FPB and 9/26 for FPC, indicating that the CRC65 cell line panel contained cell lines, which were more similar to tumours than they were to other cell lines.

The association of FPSs with previously published subtypes, CMS predictions, as well as genomic and epigenomic features was measured using Fisher's exact test (Table 6.1.1). Most notably, the proteomics-based classification mapped partially (multiple-test adjusted $p<0.05$) to the CMS published for the CPTAC patient samples, the CMS mRNA predictions and the proteomics-based subtypes by Zhang and colleagues, while mapping completely to other mRNA-based subtypes such as the ones by Roepman et al. (2014) and TCGA (TCGA Network, 2012b). As such, FPB was associated with microsatellite instability and the MSI/CIMP subtype, whereas FPA was associated with 18q loss and the CIN subtype of the TCGA classification, highlighting the distinct biology of these two subtypes. FPA was associated with CMS2, FPB was associated with CMS1 and CMS4, while FPC was associated with CMS3, respectively. FPC also showed association with mutations in BCL9L and RNF43, as well as with mutations in EP300, pointing towards deregulated WNT signalling in conjunction with aberrant histone acetylation. Interestingly, FPA was associated with TP53 mutations while FPB

was associated with ATM mutations, suggesting that p53 signalling in response to DNA damage is perturbed through distinct mechanisms in these two subtypes. Since the FPSs agreed well with both mRNA-based and protein-based subtypes, the CMS were also predicted for the CRC65 cell line and CPTAC patient datasets using the proteomics data, which resulted in an overall prediction accuracy of 0.66 (49/74) for the patients, which was somewhat lower than the accuracy achieved using mRNA data (0.8). The CMS predictions, as well as the FPS subtype labels can also be found in Table 6.1.9.

With integrated FPS of CRC at hand, the next step was their characterisation with respect to proteins differentially expressed between them, as well as pathways and other functional annotations enriched in these proteins. Significance analysis of microarrays or SAM (Tusher et al., 2001) was used to discover proteins with significant differential expression between the different FPSs at a median FDR of 0.01, as described in section 2.11.2. Afterwards, enrichment analysis of various functional classifications was carried out on differentially expressed proteins using MetaCore as outlined in section 2.11.3.

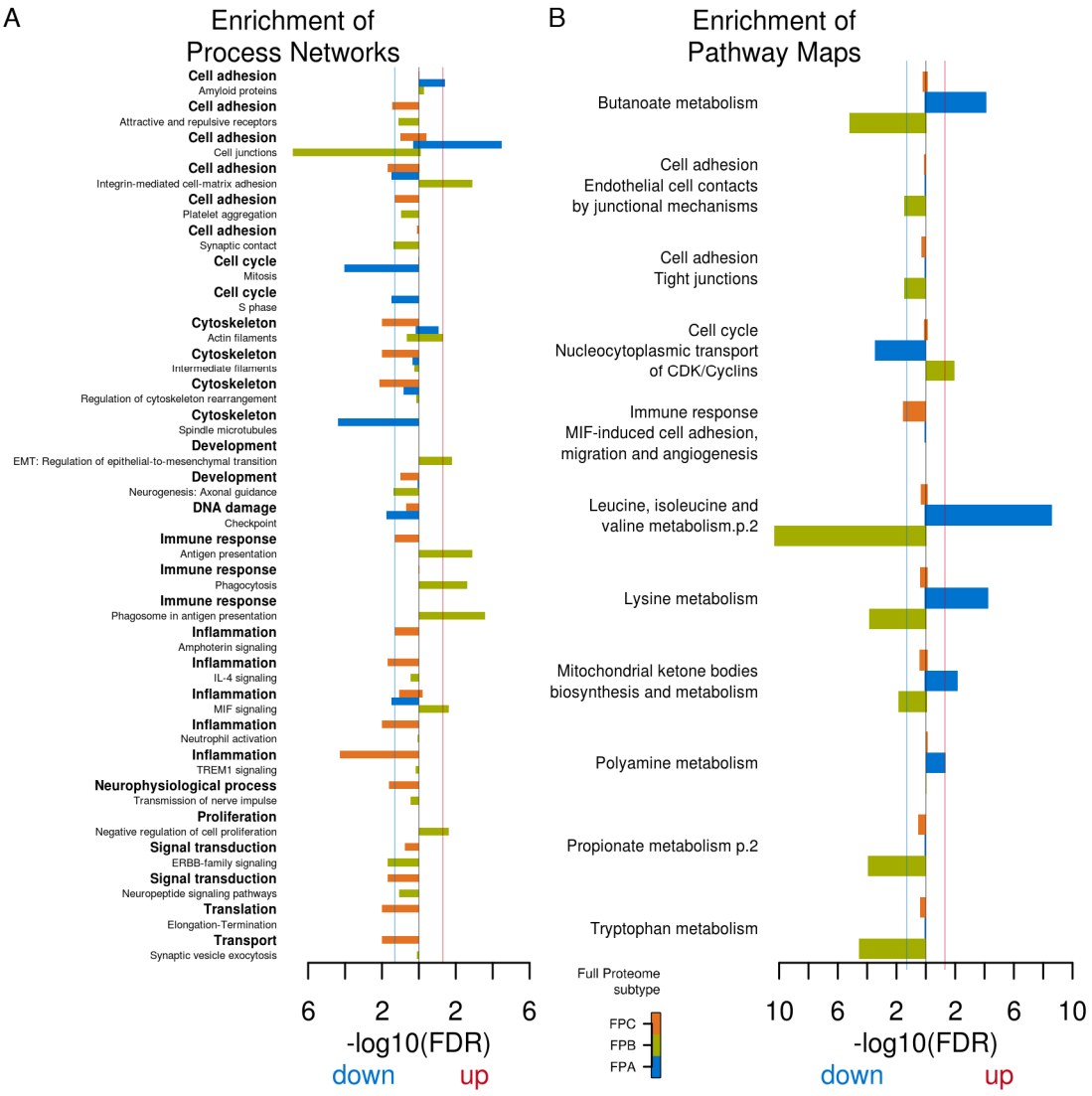


Figure 5.3.2

Process networks and pathway maps enriched in proteins differentially expressed between Full Proteome Subtypes

(A) Bar chart visualising the negative logarithm to the base of ten with respect to the enrichment FDR of Process Networks significantly (minimum FDR≤0.05, MetaCore) enriched in proteins down-regulated (to the left of the origin) or up-regulated (to the right of the origin) in the different FPSs. (B) Same as (A) for Pathway Maps. See section 2.11.3 for details.

This analysis tested whether the list of up- or down-regulated proteins contained more proteins annotated with a given functional classification than would be expected by chance using a hypergeometric test, which is essentially a one-sided Fisher's exact test (Rivals et al., 2007).

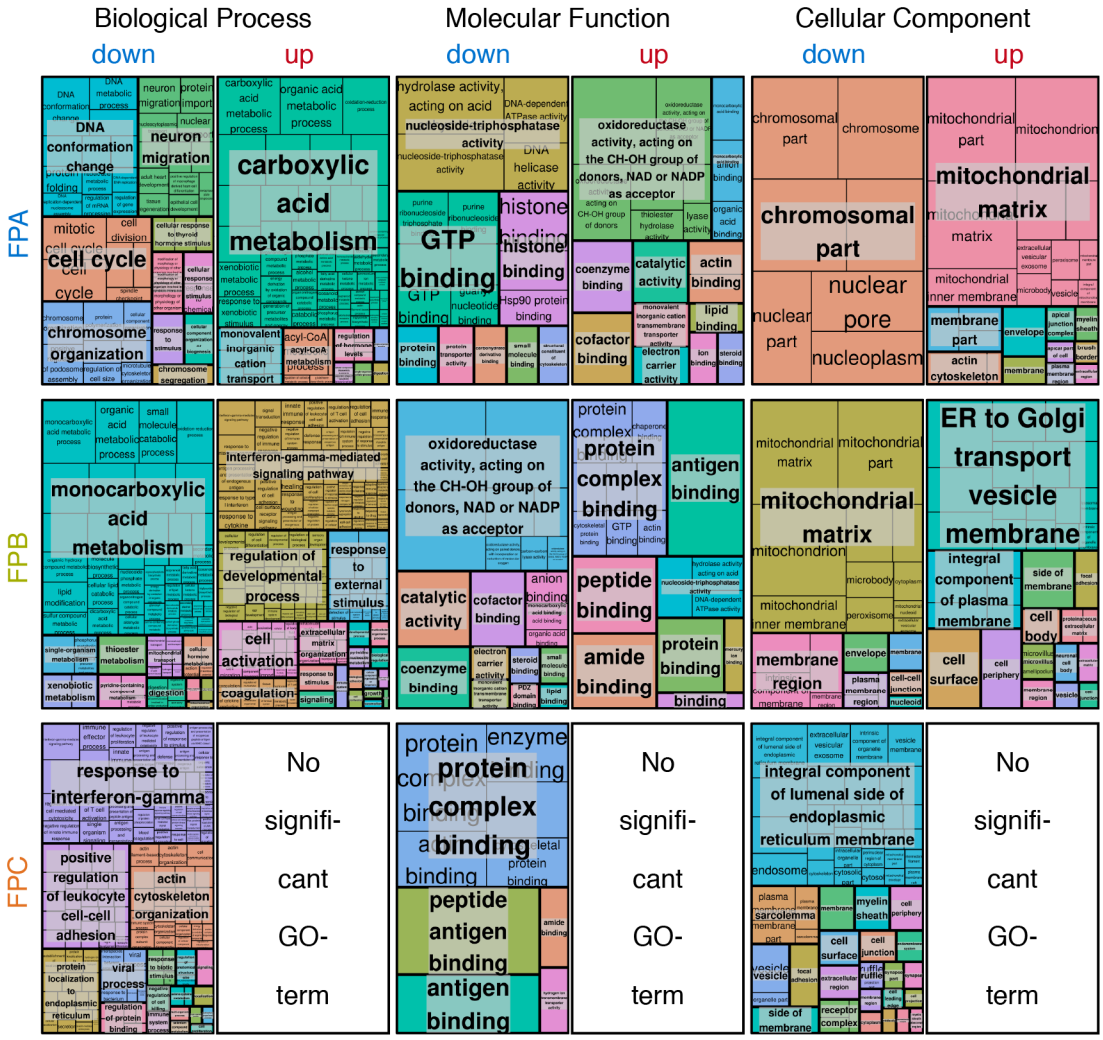


Figure 5.3.3

GO-terms enriched in proteins differentially expressed between Full Proteome Subtypes

Tree maps generated from R scripts as downloaded from REViGO (Supek et al., 2011) after summarisation of GO-terms significantly (minimum FDR≤0.05, MetaCore) enriched in proteins down- or up-regulated in the different FPSs. The size of each area corresponds to the enrichment FDR of the respective GO-term as reported by MetaCore. See section 2.11.3 for details.

While the results of this enrichment analysis with respect to process networks and pathway maps were directly visualised using bar charts of the enrichment FDRs (see Figure 5.3.3), enriched GO-terms were subsequently uploaded to REViGO (http://revigo.irb.hr/, Supek et al., 2011) for summarisation and visualisation using tree maps (see Figure 5.3.3). As indicated by Figure 5.3.2 and

Figure 5.3.3, this analysis revealed that proteins significantly down-regulated in FPA were significantly (multiple-test adjusted $p<0.05$) enriched in process networks associated with "Cell adhesion" (Integrin-mediated cell-matrix adhesion), "Cell cycle" (S phase & Mitosis), "Cytoskeleton" (Spindle microtubules), "DNA damage" and "Inflammation" (MIF signalling, Figure 5.3.2 A), as well as "Cell cycle" pathway maps (Nucleocytoplasmic transport of CDK/Cyclins) and GO-terms associated with "DNA conformation change", "cell cycle", "nucleoside-triphosphatase activity" (DNA helicase activity) and "chromosomal part" (Figure 5.3.3). Proteins up-regulated in FPA on the other hand were enriched in metabolism-associated pathway maps (Figure 5.3.2 B, several of them involving amino acids) and "Cell adhesion" process networks (Cell junctions & Amyloid proteins), in conjunction with GO-terms associated with "carboxylic-acid metabolism", "oxidoreductase activity" and "mitochondrial matrix". FPA was strongly contrasted by FPB, as down-regulated proteins in FPB showed similar enriched functional classifications as up-regulated proteins in FPA and vice versa, with the exception of a lack of enrichment with respect to process networks involved in "Cell adhesion" (Amyloid proteins), "Cell cycle", "Cytoskeleton" (Spindle microtubules) and "DNA damage". Process networks specifically enriched in proteins up-regulated in FPB were related to "Cytoskeleton" (Actin filaments), "Development" (EMT), "Immune response" (among them "Antigen presentation") and "Proliferation", while GO-terms enriched in proteins up-regulated in FPB were associated with "interferon-gamma signalling", "protein complex binding" and "antigen binding", as well as Cellular Component GO-terms related to secretion and the plasma membrane. In contrast to FPB, all these GO-terms as well as the "Cell adhesion" process network

were enriched in proteins down-regulated in FPC, in addition to a variety of other process networks involved in e.g. “Cell adhesion”, “Cytoskeleton” and prominently also “Inflammation”, as well as pathway maps involved in immune cell signalling. Interestingly, proteins up-regulated in FPC did not show any significant enrichment of functional classifications.

5.4 Kinobeads Subtypes (KSs) of CRC

The discovery of integrated FPSs of CRC based on full proteome data from cell lines and patients prompted the question whether the sub-proteome enriched by Kinobeads could recapitulate these subtypes. This represented the last remaining step with respect to the characterisation of the pharmacoproteomic training and testing dataset established as part of this thesis, outlined in section 1.4. Consensus clustering was again used to identify three Kinobeads Subtypes (KSs) KA-KC on the basis of kinases enriched by Kinobeads (Figure 5.4.1). Afterwards, their association with FPSs, as well as with CMS predictions, subtypes published by De Sousa E Melo et al. and a reduced set of genomic features was measured using Fisher’s exact test (Table 6.1.2). The KSs fully recapitulated the FPSs and the subtypes by De Sousa E Melo et al. in that KA was significantly (multiple-test adjusted $p<0.05$) associated with FPC and CCS2, KB was associated with FPA and CCS1 and KC was associated with FPB and CCS3, respectively. KA was also significantly associated with microsatellite instability, which was previously observed for both CCS2 (De Sousa et al., 2013) and CMS1 (Guinney et al., 2015), with which KA was associated, too.

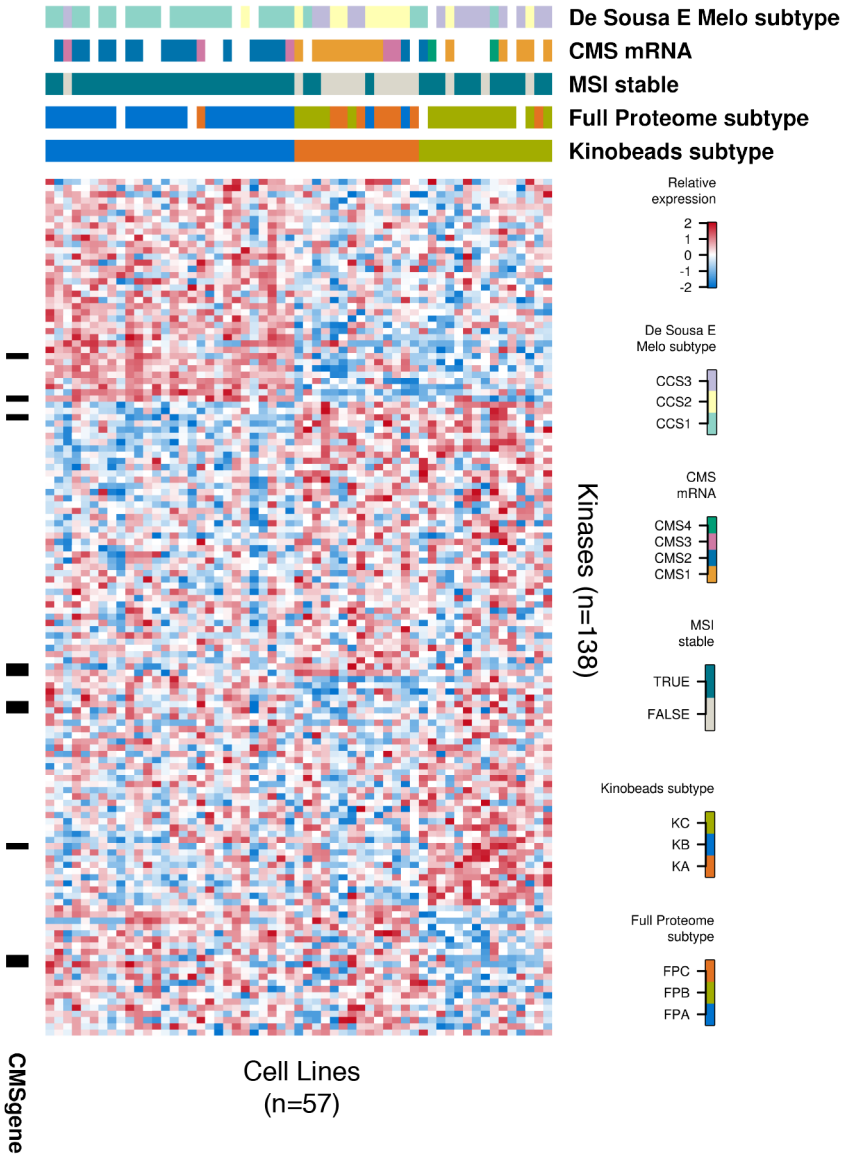


Figure 5.4.1

Kinobeads Subtypes of CRC

Heat map of standardised, log2-transformed and median-centred LFQ kinase quantification (z-scores) across the CRC65 dataset. Cell lines are displayed as columns and kinases are shown as rows (kinases measured in at least two biological replicates across at least one third of the cell lines were used to find subtypes). Black bars to the left of the heat map indicate the presence of the respective kinase in the CMS classifier. Annotation bars on top of the heat map visualise the membership of the different cell lines/patients in five annotation categories. Columns are ordered hierarchically according to the consensus clustering results (see section 2.11.1), while rows are ordered according to their association with a particular KS (see section 2.11.2).

Afterwards, SAM was used for the discovery of kinases with significant differential expression between the different Kinobeads Subtypes at a median FDR of 0.01. This approach uncovered kinases frequently mutated in MSI tumours with concomitant decrease in expression like ACVR2A and TGFBR2 (Jung et al., 2004; Markowitz et al., 1995), as well as EPHA2, overexpression of which is known to predict resistance towards cetuximab in CRC (Strimpakos et al., 2013). SAM also identified MERTK, a receptor tyrosine kinase correlating with disease progression in melanoma (Schlegel et al., 2013), to be differentially expressed between the different Kinobeads Subtypes (q-value<0.01). MERTK is down-regulated in KRAS mutant CRC (Watanabe et al., 2011a) and up-regulated in pancreatic cancer cell lines resistant towards selumetinib (Beech and Kelly, 2014). Kinases with differential expression between the different KSs can be found in Table 6.1.3.

5.5 Modelling drug sensitivity as a function of full proteome profile

The discovery and characterisation of integrated Full Proteome and Kinobeads subtypes concluded the characterisation of the comprehensive pharmacoproteomic training and testing dataset established as part of this thesis. This dataset was subsequently used to model drug sensitivity as a function of full proteome profiles, taking advantage of the overlap between the CRC65 cell line dataset and the cell lines used in recent large-scale drug screening studies (Barretina et al., 2012; Basu et al., 2013; Garnett et al., 2012; Iorio et al., 2016; Medico et al., 2015; Rees et al., 2016; Seashore-Ludlow et al., 2015). One goal of this project was to tightly control and streamline each part of the data analysis workflow in order to avoid drawing conclusions based on differences in the data

processing. Therefore, instead of simply downloading and using the processed drug sensitivity data from the "Cancer Cell Line Encyclopedia" (CCLE), the "Genomics of Drug Sensitivity in Cancer" (GDSC), the "Cancer Therapeutics Response Portal" (CTRP) and the study focusing on cetuximab by Medico et al. (2015) – all of which used different dose-response models and ways to estimate their measures of drug sensitivity – these datasets were re-analysed from raw data where possible. Unifying the dose-response modelling, relative response measures R were calculated for each dataset as described in section 2.12, which expressed drug response as residual cell viability relative to a control ("relative response"), the response of which was set to one. Therefore, all dose-response measures varied between zero and one, bringing the different datasets onto the same scale. Streamlining the data processing even further, these relative response measures were subsequently used for dose-response modelling using a single, unified framework to fit the classical symmetric four-parameter log-logistic model to each drug in each dataset (see section 2.12.6). Instead of the IC50, which does not take the effect strength into account, this study made use of the area under the dose response curve (AUC), which accounts for both the IC50, as well as the strength of a drug's effect (Basu et al., 2013). Together with the combined proteomic dataset described above, the aggregated AUC values for all drugs across all four datasets were used for pharmacoproteomic modelling of drug response as a function of full proteome profiles using elastic net regression, as described in section 2.13. The elastic net regression models were subsequently used to predict the drug sensitivity of both cell lines and tumours, which is visualised in Figure 5.5.1.

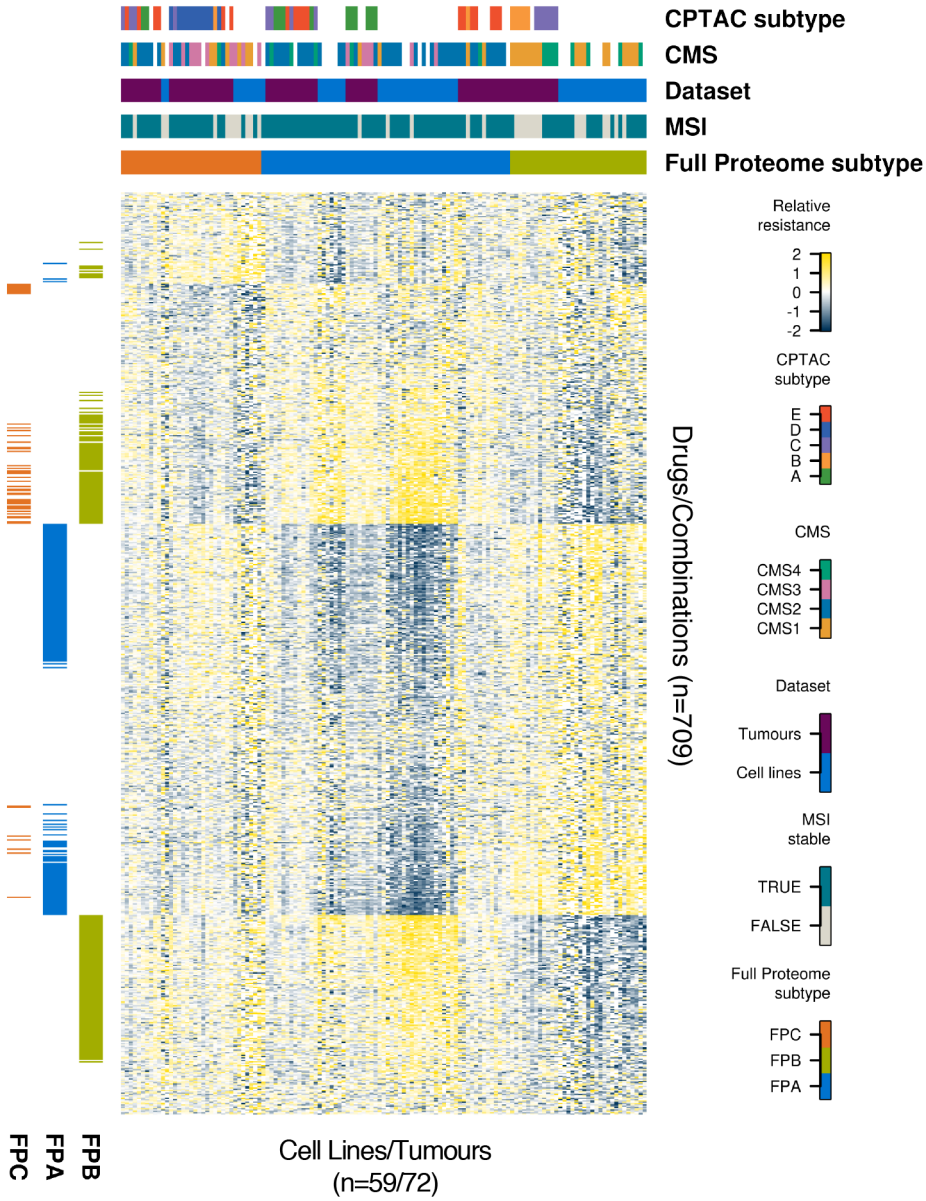


Figure 5.5.1

Prediction of drug sensitivity for cell lines and patients

Heat map of standardised predicted AUC values (z-scores of the area under the dose-response curve). Elastic net regression was used to model drug sensitivity as a function of full protein profile with the CCLE, CTRP, GDSC and Cetuximab datasets as input. The resulting models were used to predict the drug sensitivity of cell lines/patients (see also section 2.13). Cell lines/patients are displayed as columns with the same annotations and ordering as in Figure 5.3.1, while drugs are shown as rows. Coloured bars to the left of the heat map indicate drugs which are more effective in the respective subtype, i.e. the standardised mean difference between the drug's AUC in that subtype versus its overall mean AUC is smaller than the 2.5 % quantile over all drugs, as determined using SAM.

Combining these drug sensitivity predictions with the integrated FPSs of CRC enabled target space enrichment, which is essentially a method to discover proteins or pathways frequently targeted by drugs differentially effective between the different FPSs. In order to do so, SAM (median FDR of 0.001) was used to identify drugs, which were differentially effective between the FPSs. Most of these drugs were annotated with cellular targets and other mode-of-action-relevant information, which is compiled in Table 6.1.8 (see also section 2.18).

Using this information, target space enrichment was performed by measuring whether specific targets or other annotations appear more frequently than expected by chance in drugs differentially effective in the different FPSs using Fisher's exact test (Table 6.1.4). Cetuximab for example was one of the drugs more effective in FPA than in the other FPSs, with the EGFR target enrichment almost reaching significance (multiple-test adjusted $p=0.0505$). Interestingly, FPA was also significantly (multiple-test adjusted $p<0.05$) associated with drugs targeting MAP2K1 and MAP2K2 (mainly selumetinib/AZD6244 and its combinations), as well as with drugs and combinations targeting BRDT – a bromodomain-containing protein (Jones et al., 1997) – (mainly JQ-1), histone deacetylases or HDACs (among them vorinostat) and NAMPT – a protein suggested to be involved in CRC progression (Lv et al., 2015) – (among them daporinad). Even though more drugs were differentially effective in FPB than in FPA (205 versus 170), most of these drugs did not show significant enrichment in their target space. However, FPB did show positive association with drugs targeting inhibitors of apoptosis or IAPs (all are SMAC mimetics), as well as negative association with drugs targeting EGFR, while FPC showed association with drugs targeting DHFR – a protein possibly associated with response towards

5-FU in CRC (Shimamoto et al., 2016) – (among them methotrexate). In summary, pharmacoproteomic modelling of drug sensitivity as a function of full proteome profiles resulted in the prediction of drug sensitivity with respect to both patients and cell lines towards a variety of drugs and drug combinations and enabled the identification of proteins, which were disproportionally often targeted by drugs more effective in one FPS than in the remaining FPSs. These predictions could be used to help in the design of prospective studies focusing on personalised medicine.

5.6 Modelling drug sensitivity as a function of Kinase profile

While the preceding section focused on the modelling of drug sensitivity as a function of full proteome profiles, this section describes the results considering the modelling of drug sensitivity as a function of kinome profiles of the CRC65 cell line panel.

An initial comparison of the predictions based on full proteome and kinome profiles was carried out by calculating z-scores of the AUC for each drug in each prediction dataset and subsequently examining the correlation of these z-scores. As visualised in Figure 5.6.1 A, the overall correlation between these two datasets was 0.63, with a median correlation across the different drugs of 0.65 (see Figure 5.6.1 B), indicating that the predictions based on the two different omics datasets were in agreement with each other for the majority of all drugs. Closer inspection of drugs the Pearson correlation coefficient of which was disproportionally small (equal to or smaller than the mean minus two standard deviations) revealed that the models used for these drug sensitivity predictions were based on significantly (Welch t-test $p=0.0003$) fewer cell lines, suggesting that the training dataset for

these drugs might have been too small for accurate and stable drug sensitivity predictions.

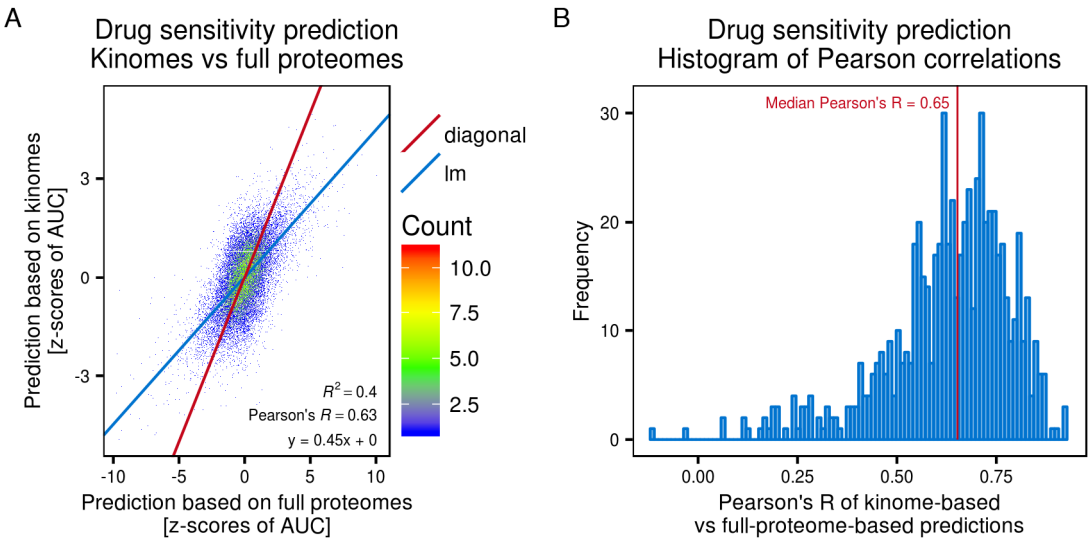


Figure 5.6.1

Comparison of drug sensitivity predictions based on kinome/full proteome profiles

(A) Two-dimensional histogram (600 bins in each dimension) visualising the AUC of drug sensitivity predictions (z-scores) based on the kinomes as a function of the drug sensitivity predictions (z-scores) based on the full proteomes of the CRC65 cell line panel (z-scores were calculated separately for each drug and each dataset of predictions; diagonal: x=y; lm: linear model; R^2: coefficient of determination). In order to prevent over-plotting, the number of data points in a particular area of the plot (Count) was mapped to a colour-scale as opposed to plotting each data point separately. (B) One-dimensional histogram (100 bins) visualising the distribution of Pearson correlation coefficients (R) between predictions for each drug based on kinome and full proteome profiles. Only correlations for drugs with more than one valid prediction per dataset are shown (n=659).

As explained in section 2.13, elastic net regression could also be used to identify proteins associated with drug sensitivity or resistance. Since many targeted drugs act on kinases, this section focuses on the identification of kinases associated with drug sensitivity or resistance. As visualised in Figure 5.6.2, elastic

net regression could for example confirm findings by Strimpakos et al. (2013) as EPHA2 was strongly associated with resistance towards cetuximab.

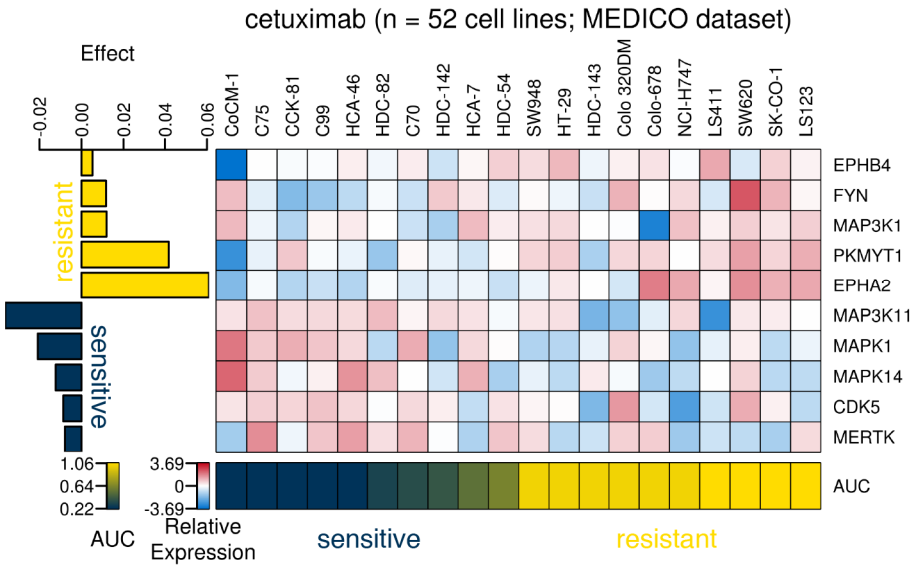


Figure 5.6.2

Confirmation that high EPHA2 is expression is associated with resistance towards cetuximab

Effect-size heat map of cetuximab visualising high EPHA2 expression as a robust predictive marker of resistance. See section 2.13 for an explanation of effect-size heat maps.

Strikingly, MAP2K1 was found to be amongst the top five kinases associated with resistance in two-thirds (12/18) of all inhibitors targeting EGFR for which the elastic net reported at least one non-zero coefficient, six of which are shown in Figure 5.6.3.

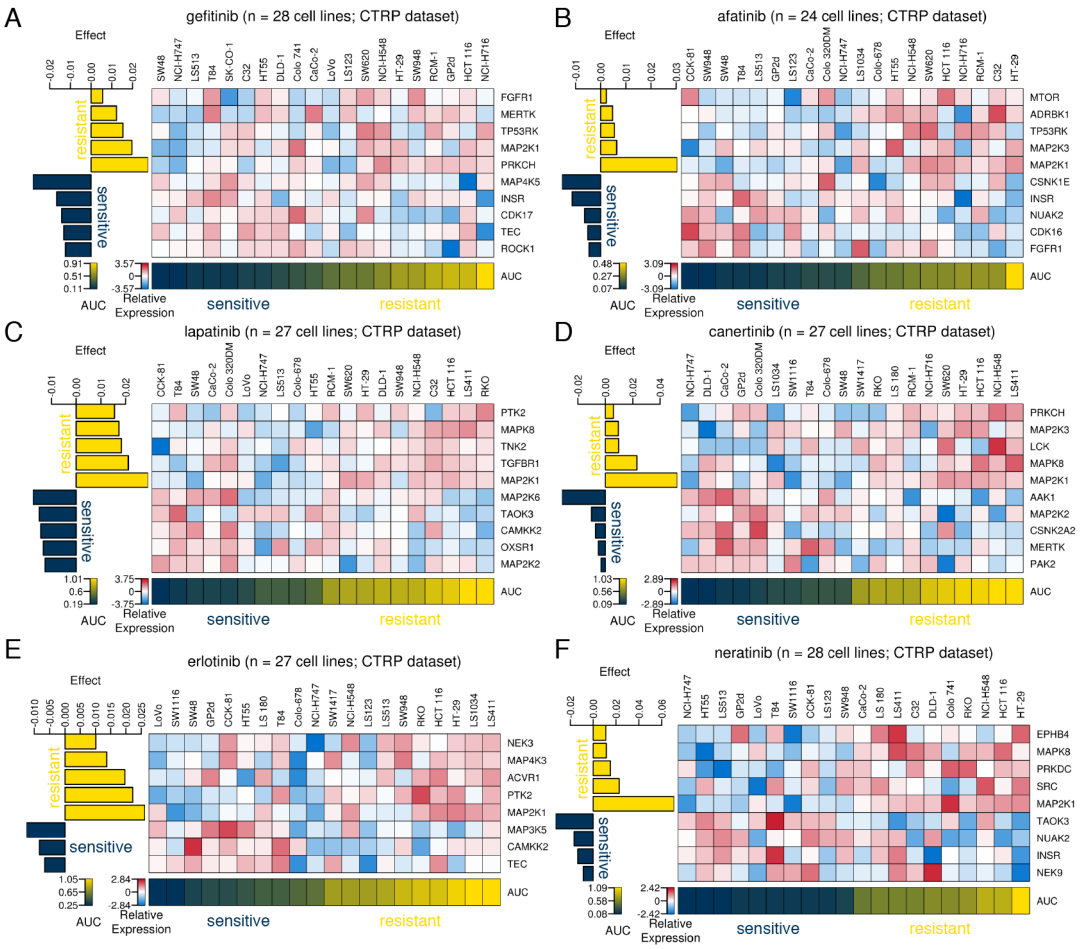


Figure 5.6.3

MAP2K1 is a predictive marker for resistance towards EGFR inhibitors in cell lines

Effect-size heat maps of six drugs targeting EGFR show consistent association of high MAP2K1 expression with drug resistance. See section 2.13 for an explanation of effect-size heat maps.

This prompted further investigations into potentially novel kinase-drug associations. Interestingly, elastic net regression recurrently found MERTK associated with resistance towards MEK1/2 inhibitors from multiple drug sensitivity datasets (Figure 5.6.4), irrespective of whether only kinases or all proteins were used as predictors.

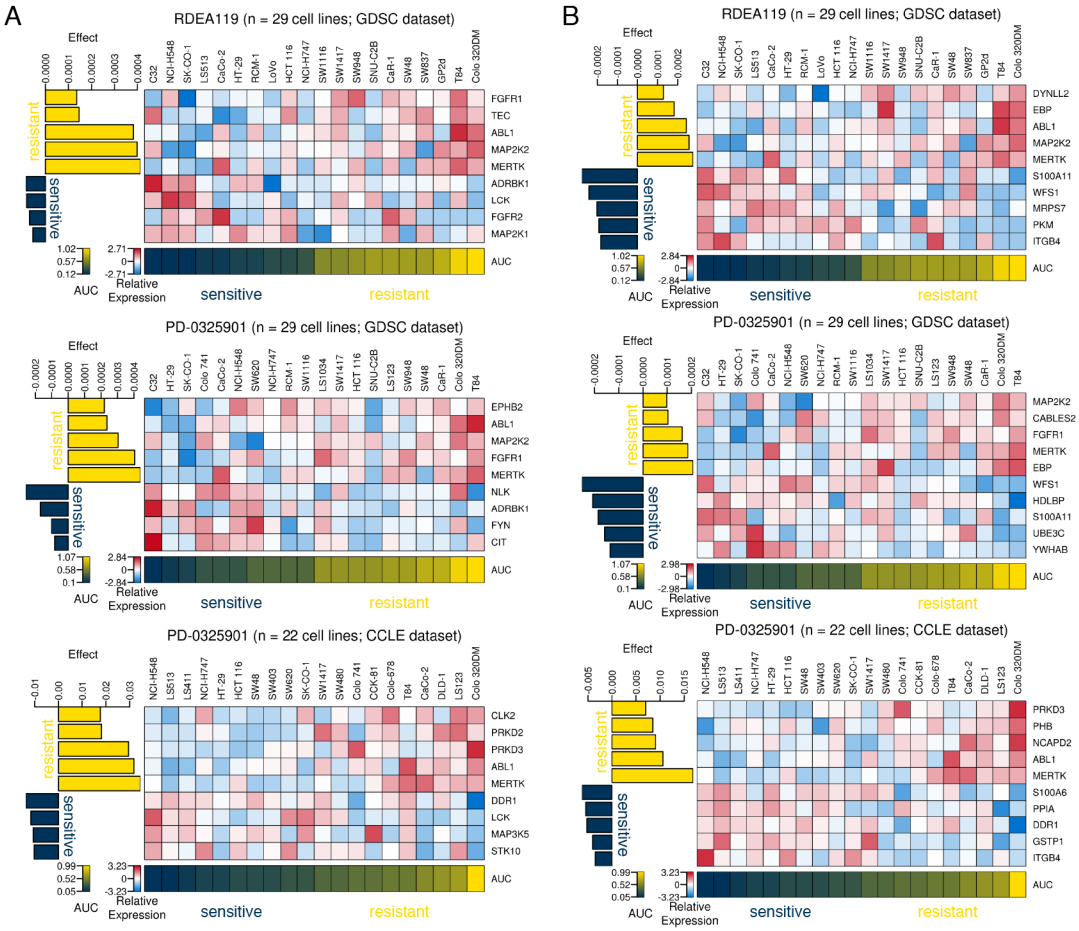


Figure 5.6.4

MERTK is a predictive marker for resistance towards MEK1/2 inhibitors in cell lines

Effect-size heat maps of two drugs (one from two datasets) targeting MEK1/2 (different rows of heat maps) show consistent association of high MERTK expression with drug resistance when drug sensitivity is modelled based on (A) kinases or (B) all proteins in the Kinobeads expression matrix (see section 2.13). See section 2.13 for an explanation of effect-size heat maps.

Since this was not yet reported for CRC and because MERTK was frequently associated with resistance towards drugs in the CCLE and GDSC datasets (Figure 5.6.5 A), three cell lines predicted to be sensitive to two MEK1/2 inhibitors (RDEA119 and PD-0325901), as well as three cell lines predicted to be resistant with differential expression of MERTK (Figure 5.6.5 B) were selected in order to

confirm these *in silico* predictions *in vitro*, which represented one of the last aims of this study (see section 1.4) and is detailed in section 5.7.

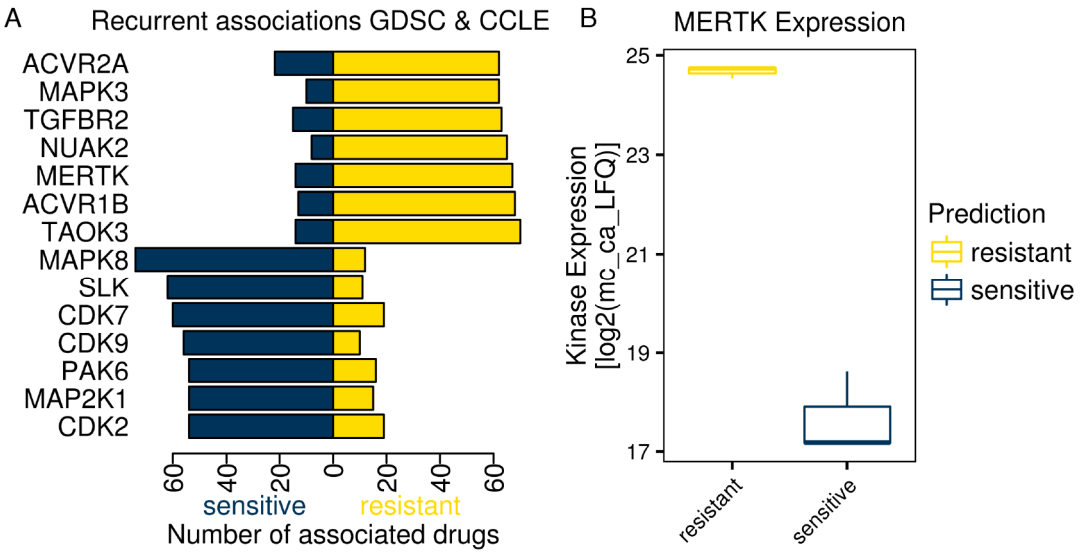


Figure 5.6.5

MERTK is recurrently associated with drug sensitivity

(A) Bar chart visualising the top seven kinases recurrently associated (absolute effect-size>0) with drug resistance (top seven bars) and sensitivity (bottom seven bars), respectively (only drugs from the GDSC and CCLE datasets). X-axis: Number of drugs for which effect-size>0 (to the right of the origin, "resistance") or effect-size<0 (to the left of the origin, "sensitivity"). (B) Boxplots visualising the expression of MERTK (log2-transformed & median-centred LFQ values after ComBat adjustment) as measured using Kinobeads for cell lines predicted to be resistant (yellow) or sensitive (dark blue) towards RDEA119 and PD-0325901 (resistant: C10, CaCo-2, T84; sensitive: CC07, HDC-143 and SK-CO-1). Resistance towards these MEK1/2 inhibitors was associated with high expression of MERTK.

5.7 MERTK is a predictive biomarker in CRC

As described in the preceding section, elastic net regression identified MERTK as recurrently associated with drug resistance, particularly considering MEK1/2 inhibitors. In order to confirm the above-mentioned pharmacoproteomic predictions of drug sensitivity, as well as the association of MERTK with resistance towards MEK1/2 inhibitors, *in vitro* single-agent and co-treatment experiments with two drugs targeting MEK1/2 (RDEA119 and PD-0325901) and

one drug targeting MERTK (UNC569) were performed as described in section 2.14. First, cell lines predicted to be resistant or sensitive to MEK1/2 inhibition were treated with these compounds in a 9-point two-fold dilution series or with medium containing the same DMSO concentration as a control, in order to confirm the aforementioned *in silico* predictions in independent cell lines (C10, CC07, HDC-143), as well as cell lines included in the training dataset (CaCo-2, SK-CO-1, T84). Next, constant-ratio combination-treatments across the entire concentration range were used to test whether inhibition of MERTK can re-sensitise resistant cell lines to treatment with MEK1/2 inhibitors. The number of nuclei detected by automated confocal microscopy was chosen as the viability readout, followed by dose-response modelling as described in section 2.12.6.

As shown in Figure 5.7.1 B and C and summarised in Figure 5.7.1 A, it was possible to confirm the *in silico* pharmacoproteomic drug sensitivity predictions for RDEA119 and PD-0325901, as cell lines predicted to be resistant (red) showed a consistently higher AUC than cell lines predicted to be sensitive (black). Moreover, co-treatment of the same cell lines with MEK1/2 inhibitors and UNC569 (MERTK inhibitor) reduced the AUC of MEK1/2-inhibitor-resistant cell lines to levels comparable to sensitive cell lines, as visualised by Figure 5.7.1 D and summarised in Figure 5.7.1 A. However, this reduction in the AUC seemed to be mainly due to UNC569 on its own, since both MEK1/2-inhibitor sensitive and resistant cell lines showed AUCs comparable to the co-treatment when treated with UNC569 alone. Taken together, pharmacoproteomic predictions of drug sensitivity could be confirmed on independent cell lines, while co-treatment experiments suggested that MERTK is a predictive biomarker for response

towards MEK1/2 inhibitors. More experiments are required in order to quantify to which extend co-inhibition of MERTK and MEK1/2 can be synergistic.

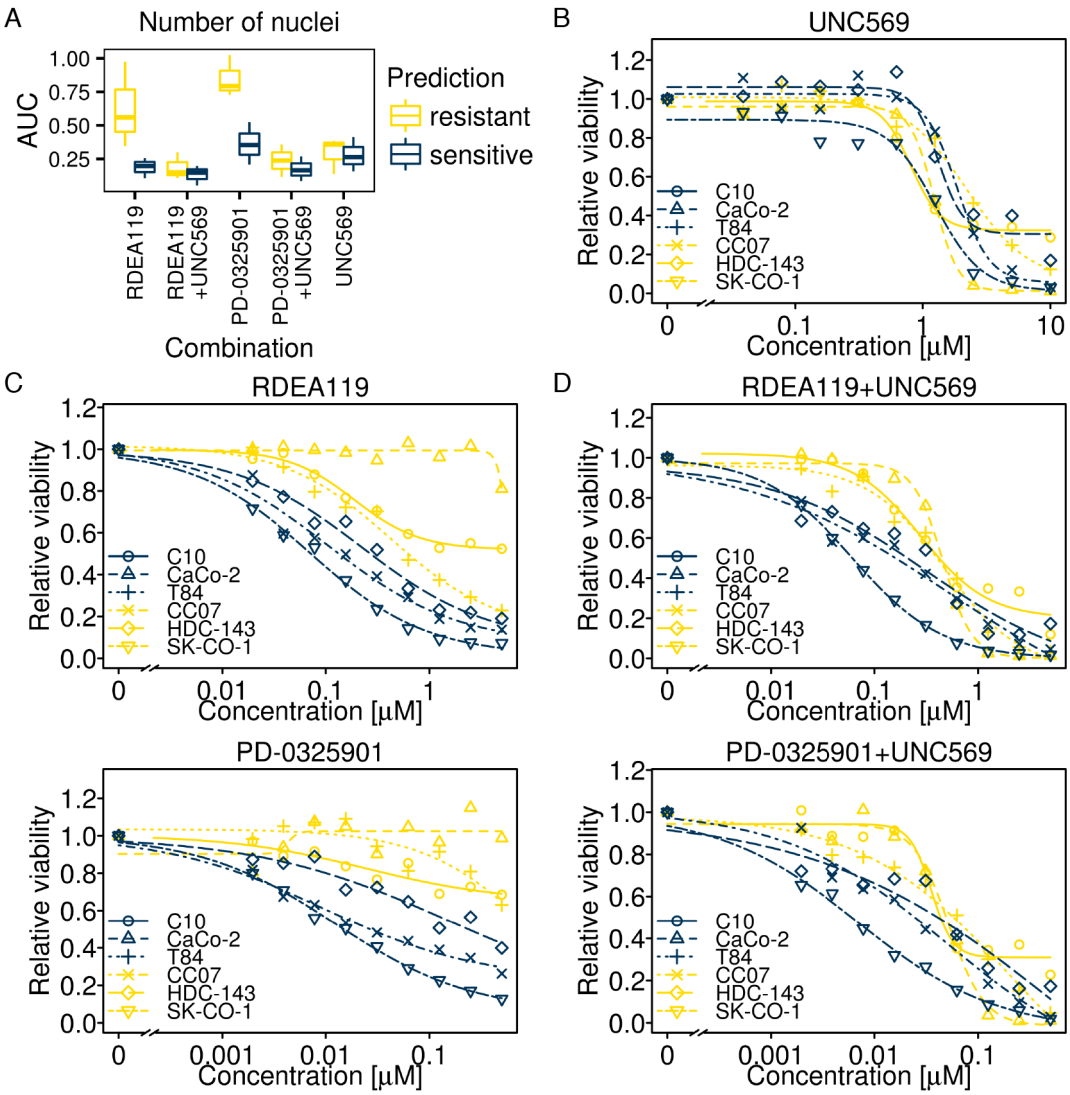


Figure 5.7.1

Co-treatment with UNC569 (MERTK inhibitor) and MEK inhibitors

(A) Boxplots and scatter of the AUC from (B-F) 10-point dose-response curves with respect to (B) the MERTK inhibitor UNC569, (C) two MEK1/2 inhibitors RDEA119 and PD-0325901 and (D) their constant-ratio combination (see section 2.14) for the same cell lines as depicted in Figure 5.6.5 B. Relative response is expressed as the mean of three technical replicates.

5.8 MERTK is a prognostic biomarker in CRC

In order to evaluate the prognostic potential of MERTK in CRC, its expression was quantified in 1,074 patients enrolled in the QUASAR2 clinical trial by immunohistochemistry (IHC, see section 2.16).

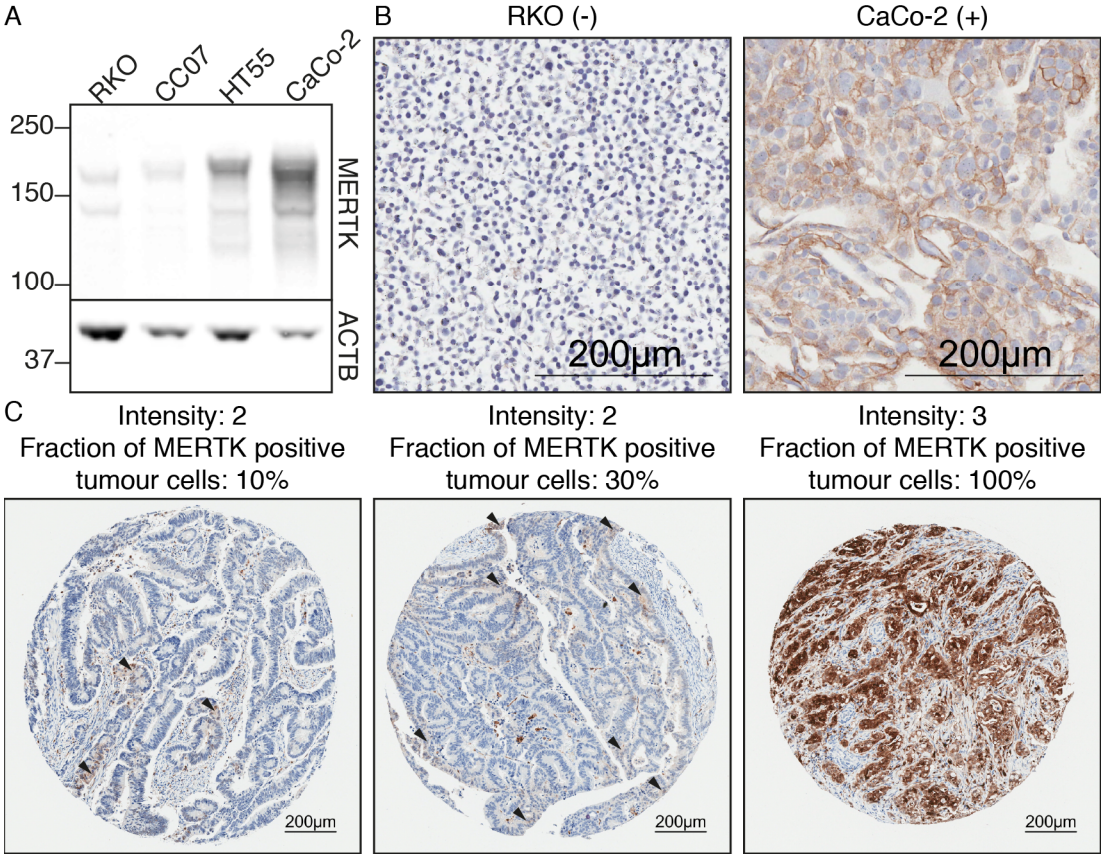


Figure 5.8.1

Western blot and immunohistochemistry of MERTK in cell lines and tumours

(A) Western Blot visualising MERTK expression in two positive (CaCo-2 & HT55) and two negative control cell lines (RKO, CC07). (B) IHC staining of MERTK expression in RKO (negative control) and CaCo-2 (positive control) cell pellets. (C) IHC staining of three representative TMAs (performed by Helen Scott from the Department of Oncology, University of Oxford) and the corresponding quantification of the signal by pathologists from the Technical University of Munich (Julia Slotta-Huspenina and Moritz Jesinghaus).

After establishing robust, cell-pellet-based positive and negative controls with high and low expression of MERTK (Figure 5.8.1 A and B), TMAs were stained

with a monoclonal antibody previously used for IHC of TMAs (performed by Helen Scott from the Department of Oncology at the University of Oxford; Schlegel et al., 2013) and the fraction of membrane-positive cells, as well as the staining intensity were quantified by trained pathologists (Julia Slotta-Huspenina and Moritz Jesinghaus from the Technical University of Munich). Figure 5.8.1 C shows three representative TMA samples alongside their quantitation of MERTK, which highlights the variability of MERTK expression in primary tumours.

Modelling the outcome variables of the QUASAR2 trial as a function of MERTK expression showed that high membranous MERTK expression was prognostic of worse 5-year overall survival (OS), disease-free survival (DFS) and recurrence-free survival (RFS) in both univariate (Table 6.1.5) and multivariate Cox proportional hazards regression (Figure 5.8.2 and Table 6.1.6), with a multivariate hazard ratio (HR) of 1.6 (over 5% membrane-positive tumour cells versus less than or equal to 5% membrane-positive tumour cells; $CI_{0.95}$=1.09-2.35; p=0.0163), 1.7 ($CI_{0.95}$=1.24-2.33; p=0.00095) and 1.77 ($CI_{0.95}$=1.26-2.48; p=0.00087), respectively. High membranous expression of MERTK was also significantly associated with $BRAF^{V600E}$ mutations and T4 tumours, while low membranous expression was associated with wild-type BRAF and T2/T3 tumours (two-sided Fisher's exact test $p<0.05$, Table 6.1.7), suggesting that either aberrantly activated MAPK-signalling or tumour progression in general might lead to up-regulation of MERTK. However, more studies need to be done in order to determine the direction of causality for both of these observations. Taken together, this suggests that MERTK is a prognostic biomarker in CRC patients.

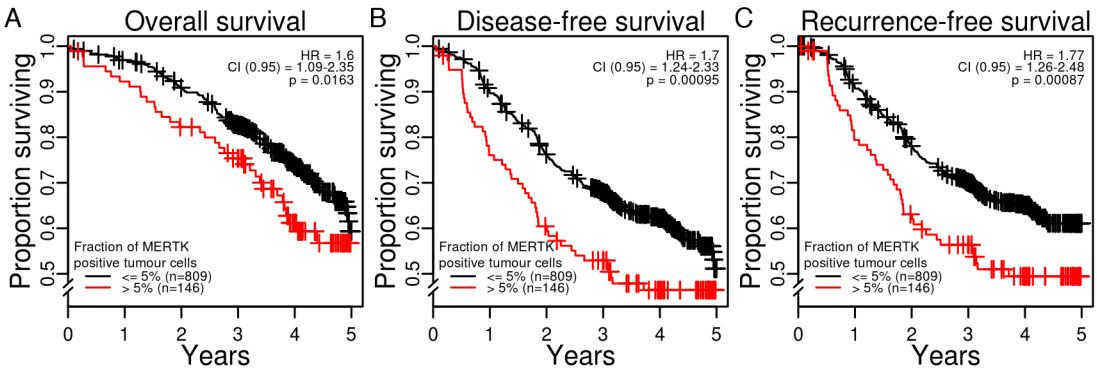


Figure 5.8.2

Survival analysis of the QUASAR2 patient cohort, stratified by MERTK expression

Kaplan-Meier plots showing worse (A) overall, (B) disease-free and (C) recurrence-free survival of patients with high (>5% of tumour cells) membranous expression of MERTK.

5.9 Discussion

This chapter laid out the results of the multi-omics data integration of transcriptomics, proteomics and drug sensitivity information across the CRC65 and CPTAC datasets, which resulted in the most comprehensive pharmacoproteomic characterisation of human colon and rectal cancer to date. Studying the relationship between mRNA and protein abundance confirmed findings by Wilhelm et al. (2014), as protein/mRNA ratios were constant across both the CRC65 and CPTAC datasets. Constant protein/mRNA ratios were also observed before, for example across various microorganisms by studying phosphorous:nitrogen ratios (Karpinets et al., 2006), in yeast by using microarrays and antibody-based detection of TAP-tagged proteins (Ghaemmaghami et al., 2003), across different *Pseudomonas aeruginosa* strains by using microarrays and mass spectrometry (Kwon et al., 2014) and across other human cell lines by using mRNA-Seq and mass spectrometry (Stevens and Brown, 2013). Especially the data of the last study were in very good agreement

with the data presented here, as their median CV was also between 10% and 20%; possibly because the two approaches were similar in how protein/mRNA ratios were determined. This body of evidence suggests that protein/mRNA ratios can vary for different protein species, but are constant between different biological samples under certain circumstances, especially considering steady-state growth conditions as reviewed by Liu et al. (2016). Therefore, MComBat (Stein et al., 2015) could be used to adjust for these ratios in the CRC65 and CPTAC datasets in a protein/gene-specific fashion, effectively removing translation efficiency as a confounding factor of protein/mRNA correlation without changing the location or scale of the proteomics data.

Preceding this adjustment, the protein/mRNA correlation of the two datasets agreed well with previous publications for prokaryotes (Kwon et al., 2014), maize (Ponnala et al., 2014), yeast (Ghaemmaghami et al., 2003; Gygi et al., 1999; Ideker et al., 2001; Lee et al., 2011; Marguerat et al., 2012), drosophila tissue (Juschke et al., 2013), mouse tissue (Chick et al., 2016; Huttlin et al., 2009; Kislinger et al., 2006; Ning et al., 2012; Sharma et al., 2015), human tissue (Wilhelm et al., 2014; Zhang et al., 2014b) and cell lines (Baek et al., 2008; Gholami et al., 2013; Lundberg et al., 2010; Ly et al., 2014; Nagaraj et al., 2011; Schwanhausser et al., 2011; Stevens and Brown, 2013; Vogel et al., 2010; Wu et al., 2013). However, these correlations were modest and would not allow the prediction of protein abundance from mRNA abundance. Only recently, some studies found that steady-state protein abundance is mainly determined by mRNA abundance and therefore transcription (Jovanovic et al., 2015; Li et al., 2014), which is in line with the data presented herein following MComBat adjustment and enabled mRNA-guided missing value imputation; essentially, a

multi-omics data imputation strategy. Strategies of this kind are becoming powerful as different omics technologies mature and generate more and more correlated datasets. Lin et al. for example already developed a method for multi-omics data imputation (Lin et al., 2016), which takes advantage of correlated multi-omics datasets. The fact that transcriptomics and proteomics data correlate sufficiently after MComBat adjustment for imputation across omics technologies was impressively demonstrated by correlation heat maps, visualising how the same samples from different transcriptomics datasets cluster together and how these mRNA measurements show the highest correlation with their cognate protein measurements in most cases, enabling semi-supervised assignment of mRNA measurements to the correct cell lines. After MComBat adjustment, biology as opposed to technology clusters together, which was demonstrated before for ComBat (Johnson et al., 2007; Kitchen et al., 2010; Turnbull et al., 2012), on which MComBat is based.

However, not all missing values in the proteomics datasets could be imputed using mRNA-guided missing value imputation, which motivated the development of novel algorithms for minimum-guided missing value imputation. Other imputation methods used with proteomics data were previously developed for microarrays (Bo et al., 2004; Kim et al., 2005a; Oba et al., 2003; Troyanskaya et al., 2001) or other data types (Schneider, 2001; Tipping and Bishop, 1999), while some imputation methods were specifically developed for proteomics data (Almeida et al., 2005; Chen et al., 2014; Chich et al., 2007; Deeb et al., 2012; Karpievitch et al., 2012; Meleth et al., 2005; Tyanova et al., 2016; Wang et al., 2006). Many of these methods perform well on proteomics data (Lazar et al., 2016; Webb-Robertson et al., 2015), however they cannot always guarantee that

imputed values are always smaller than a measured value for the same protein in the dataset under investigation, which is achieved with the novel minimum-guided missing value imputation presented here. Lazar et al. rightfully argues that the nature of the missing values should be investigated prior to deciding on the use of a particular method (Lazar et al., 2016), however this is often only possible for bigger datasets capable of cross-comparisons between different samples. For the datasets described in this thesis, the MNAR assumption held true for the majority of missing values, which was already observed previously for proteomics data (Webb-Robertson et al., 2015). Another desired property of any imputation approach is that it does not alter the overall intensity distribution of the data. However, previously published MinDet imputation approaches that take the MNAR nature of missing values into account resulted in strong bimodality of the intensity distribution, which was not the case for the methods developed as part of this thesis. Especially minimum-guided missing value imputation on the peptide level had desirable properties, since it reduced the variability of low-abundance proteins by shrinking imputed values towards each other, which is useful for clustering and machine learning algorithms. The regularised log-transformation published as part of the DESeq2 package achieves a similar effect in order to improve unsupervised class discovery (Love et al., 2014).

In this thesis, hierarchical consensus clustering – previously used for CRC subtyping (Marisa et al., 2013; Zhang et al., 2014b) – was employed in favour of other tools previously used for unsupervised class discovery in CRC (Budinska et al., 2013; Medico et al., 2015; Perez-Villamil et al., 2012; Sadanandam et al., 2013; Salazar et al., 2011; Schlicker et al., 2012; TCGA Network, 2012b), resulting in three integrated FPSs consisting of cell lines and patients, even though the

selection of the number of clusters using consensus clustering is debated (Senbabaoglu et al., 2014). However, instead of using the consensus matrix to choose the optimal number of clusters, this project made use of the current CRC subtyping consensus in order to guide the search for the optimal number of clusters under specific constraints (Guinney et al., 2015). Using the item consensus together with silhouette values subsequently increased the stringency of assigning samples as core samples (Monti et al., 2003; Rousseeuw, 1987). It is worth mentioning that cell lines and patients still cluster together in their respective subtypes, suggesting that they are still very similar to each other. Similarity expressed as the number of clusters present in a given dataset is an "elbow" problem difficult to solve objectively (Ketchen and Shook, 1996). However, the association of the FPS with previously published subtypes suggests that the scientific community gravitates towards an integrated consensus subtype model of CRC. MSI for example was enriched in FPB, the subtype also associated with the MSI/CIMP subtype of the TCGA subtype model (TCGA Network, 2012b), reciprocally reinforcing conclusions made based on transcriptomics and proteomics data. The presence of additional subtypes in other subtype models of CRC might simply reflect the subjective nature of this "elbow" problem, which could be circumvented by less commonly used algorithms based on local point densities (Fu and Medico, 2007; Rodriguez and Laio, 2014). However, these were not investigated here due to the availability of more established approaches. Nevertheless, they represent interesting starting points for future studies.

Comparisons of FPSs with the CMS published by Guinney et al. and CMS assignments based on the combined proteomics and transcriptomics datasets

revealed a relatively low prediction accuracy for proteomics data, which might be due to the fact that the combined proteomics dataset was somewhat smaller than the combined mRNA dataset (6,254 proteins versus 7,113 genes). The proteomics dataset covered 58 fewer classifier genes than the mRNA dataset (321 versus 379), which might explain the reduced prediction accuracy of 0.66 compared to 0.8. One would think that 58 classifier genes is not a lot, but considering that a classification based on only 138 kinases (KSs) was able to fully recapitulate the classification based on 6,254 proteins (FPSs) helps putting the value of only a few proteins during unsupervised class discovery in perspective. It was reassuring to see that proteins known to be down-regulated in MSI tumours are also found to be down-regulated in KA/FPC, which were associated with MSI. This highlights the utility of Kinobeads pulldowns, as both ACVR2A and TGFBR2 were not detected in the full proteome profiles of the CPTAC dataset. However, there is also slight disagreement between the KSs and the FPSs regarding associations with MSI status, as KC/FPB is associated with MSI in the combined dataset, while KA/FPC is associated with MSI status in the kinome dataset, which only consisted of cell lines. This is likely due to the fact that – unlike FPB tumours – FPC tumours were not as enriched with tumours characterised by MSI, which was expected considering that FPC was significantly associated with CMS3 in mRNA-based CMS predictions; a subtype characterised by the original study as "Mixed MSI status" (Guinney et al., 2015). The fact that KA/FPC is associated with MSI in the kinome dataset might therefore reflect the fact that MSI is more common in cell lines than in tumours, possibly due to MSI tumours being propagated more easily *in vitro* (Medico et al., 2015).

While the integrated omics data processing workflow described above allowed the combination of transcriptomics and proteomics data from cell lines and patients, a streamlined reanalysis of public drug sensitivity datasets featuring a unified dose-response modelling framework enabled the cohesive pharmacoproteomic modelling of drug sensitivity as a function of full proteome and kinome profiles of the CRC65 cell line panel. In general, predictions based on the full proteome and kinome profiles of the CRC65 cell line panel were in good agreement with each other, however there were a few drugs with only low agreement or even anti-correlation between the two prediction methods. Close examination of the models used to generate these predictions revealed that they were based on significantly fewer cell lines than models with well-correlated predictions, underscoring the importance of a sufficiently sized training dataset (Wilding and Bodmer, 2014).

As mentioned in section 5.4, MERTK was found to be differentially expressed between the different KSs and was likewise not identified in the CPTAC dataset. It was significantly down-regulated in KC/FPB; the same subtype, in which PD-0325901 – a MEK1/2 inhibitor – was more effective. In fact, elastic net modelling of drug sensitivity as a function of kinome profiles revealed that MERTK was recurrently associated with resistance towards MEK1/2 inhibitors, confirming data on pancreatic cell lines by Beech and Kelly (2014). MERTK was previously shown to be down-regulated in KRAS mutant CRC (Watanabe et al., 2011a), however the KC/FPB subtype – in which MERTK was down-regulated – did not show an association with KRAS mutation status. Nevertheless, the association of MERKT with resistance towards MEK1/2 inhibitors, as well as *in silico* pharmacoproteomic drug sensitivity predictions for RDEA119 and PD-0325901

were subsequently confirmed using *in vitro* single-agent and co-treatment experiments, while IHC of 1,074 tumours from the QUASAR2 patient cohort could confirm a significant prognostic role for MERTK in CRC, achieving the final goal of this project as outlined in section 1.4.

In a nutshell, the in-depth multi-omics characterisation of a pharmacoproteomic training and testing dataset for human colon and rectal cancer resulted in integrated Full Proteome Subtypes of CRC consisting of cell lines and patients, which were in line with previously published subtypes, correlated with other genomic features and were fully recapitulated by Kinobeads Subtypes based on only 138 kinases. This dataset enabled the pharmacoproteomic modelling of drug sensitivity as a function of full proteome and kinome profiles and allowed the prediction of drug sensitivity for cell lines and tumours, which resulted in the identification of drugs differentially active between the different FPSs. Drug sensitivity models based on kinome profiles fuelled the identification of MERTK as a predictive biomarker for the response towards MEK1/2 inhibitors, which was subsequently confirmed using *in vitro* experiments. Gauging the relevance of high MERTK expression for clinical decisions using IHC finally revealed a significant prognostic role of MERTK in CRC.

Chapter 6
General discussion

Recent landmark studies established resources like the "Cancer Cell Line Encyclopedia" (CCLE, Barretina et al., 2012), the "Genomics of Drug Sensitivity in Cancer" (GDSC, Garnett et al., 2012; Iorio et al., 2016) and the "Cancer Therapeutics Response Portal" (CTRP, Basu et al., 2013; Rees et al., 2016; Seashore-Ludlow et al., 2015), all of which focus on the identification of genomic and transcriptomic associations with drug sensitivity. However, since drugs often target proteins, one possible determinant of drug sensitivity is currently often overlooked. Here, these large-scale drug sensitivity datasets and data on cetuximab (Medico et al., 2015) were overlaid onto the most comprehensive proteomics dataset on CRC cell lines to date. A previous study with comparable proteome coverage made use of isobaric labelling and quantified 8,378 proteins in a panel of eight CRC cell lines (McAlister et al., 2014). However, in order to be able to reliably model drug sensitivity as a function of proteome profile and predict in vitro preclinical drug sensitivity, more CRC cell lines covering the main molecular subtypes of CRC needed to be analysed (Wilding and Bodmer, 2014).

It was shown before that cell lines included in the CRC65 panel capture the main molecular subtypes of CRC (Medico et al., 2015; Mouradov et al., 2014), however the CMS of these cell lines was not known as of yet. By combining multiple public mRNA datasets acquired using a variety of technologies across a number of laboratories involving several different batches of cell lines cultivated under a diverse set of conditions, it was possible to assign 42 of them to the CMS subtypes using the single-sample predictor by Guinney et al. (2015), confirming that these cell lines represent the main subtypes of CRC. Using the combined mRNA data, a

streamlined data processing pipeline enabled the confirmation of previous findings for tissue samples (Wilhelm et al., 2014) in that protein/mRNA ratios were stable across both the CRC65 cell line dataset, as well as the CPTAC patient dataset. However, without the use of appropriate standards, there is no way of knowing whether the observed protein/mRNA ratios are due to varying mRNA translation efficiencies or protein degradation propensities, respectively, or whether they are a mere consequence of technique-inherent systematic bias. One such bias could stem from the fact that trypsin was used as the protease of choice and a modified version of iBAQ for label-free absolute protein quantification. It was observed before that the protease used in workflows like the one explained herein can bias protein quantification across the entire dynamic range; sometimes by more than a factor of 1,000 (Peng et al., 2012). A far more likely explanation is that the protein-specific predicted number of observable peptides used in the iBAQ approach to scale the sum of peak intensities of all peptides matching to a specific protein deviates from the true number of detectable peptides and thereby introduces a bias, which itself will be dependent on the protease used. Since stable protein/mRNA ratios and therefore systematic differences between protein and mRNA measurements were observed, these differences could however be accounted for using standard batch-adjustment techniques such as ComBat or MComBat (Johnson et al., 2007; Stein et al., 2015). When these systematic differences were adjusted, the overall Pearson correlation between protein and mRNA abundance increased to 0.9 for the CRC65 dataset and 0.86 for the CPTAC dataset, which is in line with recent literature showing that differences between steady-state protein levels are largely explained by varying transcript levels (reviewed in Liu et al., 2016). This allowed the linear

prediction of protein abundance from mRNA abundance and the development of a method for mRNA-guided missing value imputation.

Combining the CRC65 and CPTAC datasets then enabled the identification of integrated proteomic subtypes of CRC (consisting of cell lines and patients), which were associated with previously published mRNA-based and proteomics-based subtypes. Since the CMS represent the current CRC subtyping consensus, it is worth noting that even though significant association was observed between the FPSs – discovered as part of this thesis – and the CMS, as well as with CMS predictions based on transcriptomics and proteomics data, they were only partially recovered from proteomics data. The fact that CMS1 and CMS4 were grouped together into FPB might be explained by the fact that mesenchymal genes up-regulated in CMS4 are mainly of stromal rather than epithelial origin (Calon et al., 2015; Isella et al., 2015) and will therefore evade detection in cell culture experiments. Perhaps unsurprisingly, FPB was associated with MSI and the MSI/CIMP TCGA subtype, while showing up-regulation of proteins involved in immune response, EMT and integrin-mediated cell-matrix adhesion; characteristics associated with both CMS1 and CMS4 (Guinney et al., 2015). FPB also showed down-regulation of proteins involved in functional classifications associated with carboxylic acid and amino acid metabolism, which was also true for proteins up-regulated in FPA. This exemplifies the additional value of subtyping at the protein level: the proteomics data generated as part of this thesis showed that FPC is associated with CMS3, for which Guinney et al. reported enrichment of metabolism signatures on the mRNA level but not on the protein level, while functional annotation enrichment on the proteomics data was only able to detect significant enrichment of metabolism-associated functional

classifications in proteins down-regulated in FPB and up-regulated in FPA, both of which were not associated with CMS3. Taken together, this data suggests that a “high metabolism, low cell cycle, MSS” signature characterizes FPA, FPB shows a “high immune response, low metabolism, MSI” signature and FPC displays a “low immune response, low inflammation, low adhesion (invasive)” signature. In summary, three integrated (cell lines and patients) FPSs of CRC were identified and characterised, which were associated with previously published mRNA subtypes, but at the same time highlighted the different biology which separates these subtypes on the protein level.

Due to the substantial overlap of the CRC65 full proteome data on cell lines with the CPTAC patient data, it was possible to predict the drug sensitivity of both cell lines and patients towards 709 drugs and/or drug combinations as a function of protein profile. Recently, Iorio et al. (2016) reported that models using genomic features like mutations as predictors of drug sensitivity had a higher predictive accuracy than models based on gene expression data when focusing on specific cancer types. However, models using gene expression data to predict drug sensitivity were more informative in their pan-cancer analysis. Unfortunately, the question whether proteomics-based models result in more accurate predictions than genomics-based ones cannot be answered using the data described herein, since only a limited number of cell lines of a single entity were studied. Nevertheless, the predictions described in section 5.5 can be used to help in the development of targeted therapies for the different CRC subtypes and allowed the suggestion of a number of known and novel promising drugs and targets specific to the different FPSs. One example of a known subtype-specific drug is cetuximab, which was more effective in FPA than in the other FPSs. Since FPA was associated

with the transit-amplifying (TA) subtype (Table 6.1.8), this confirms previous findings on cell lines and patients where cetuximab sensitivity was recurrently observed in the TA subtype (Medico et al., 2015; Sadanandam et al., 2013). Interestingly, modelling drug sensitivity as a function of kinase expression also confirmed previous findings by Strimpakos et al. (2013), as high expression of EPHA2 was associated with resistance towards cetuximab. Since the Kinobeads-based expression level of EPHA2 was significantly lower in KB/FPA than in the other subtypes, one might speculate that co-treatment with cetuximab and EPHA2-targeted drugs might re-sensitise cetuximab-resistant patients and cell lines. Together with the target space enrichment analysis described in section 5.5, this suggests that patients of the FPA subtype might profit most from drugs targeting the classical MAPK signalling cascade or acetylation-based epigenetic modifications, patients of the FPB subtype might benefit most from drugs promoting apoptosis and patients of the FPC subtype might best take advantage of DHFR-targeted drugs.

The kinase-centric drug sensitivity analysis described in section 5.6 found high expression of MAP2K1 frequently associated with resistance towards EGFR inhibitors, and high expression of MERTK recurrently associated with resistance towards MEK1/2 inhibitors. While activating mutations of K57 in MAP2K1 are known to play a role in the development of resistance towards EGFR-targeted treatments (Bertotti et al., 2015; Marks et al., 2008; Siravegna et al., 2015), MERTK was so far only shown to be up-regulated on the mRNA level in pancreatic cancer cell lines resistant towards the MEK1/2 inhibitor selumetinib (Beech and Kelly, 2014). It was possible to confirm this predictive role of MERTK for CRC cell lines *in vitro*, while an exploration of MERTK expression levels in tumours using

immunohistochemistry on TMAs showed that high membranous MERTK expression is a significant prognostic biomarker for worse overall, disease-free and recurrence-free survival in CRC patients. Ectopic expression of MERTK was observed previously in a variety of cancers (reviewed in Graham et al., 2014) and miR-335 – a microRNA known to be down-regulated in tumours of patients with aggressive CRC and poor prognosis (Sun et al., 2014) – was shown to inhibit the expression of MERTK, taking the role of a metastasis suppressor microRNA in breast cancer (Tavazoie et al., 2008). In light of high membranous expression of MERTK being more prevalent in T4 tumours, this suggests that ectopic expression of MERTK in CRC may be due to down-regulation of miR-335 and together with the *in vitro* data from section 5.7 implies that MERTK could be a viable target for the treatment of advanced CRC, perhaps in combination with MEK1/2 inhibitors.

Taken together, this thesis established and characterised the biggest pharmacoproteomic training and testing dataset to date, which allowed the modelling of drug sensitivity as a function of full proteome and kinome profiles, facilitating the discovery of novel, CRC-specific proteomic markers of drug sensitivity and resistance. Combined with molecular subtyping on the protein level, this study enabled the identification of drugs differentially effective between integrated proteomic subtypes, thereby informing potential prospective clinical studies and promoting personalised medicine. In order to validate some of the *in silico* predictions based on the above-mentioned pharmacoproteomic models of drug sensitivity, as well as to verify MERTK as a predictive and prognostic biomarker in CRC, *in vitro* experiments and immunohistochemical staining of patient samples were carried out, confirming the role of MERTK as a

biomarker. However, more studies are needed to evaluate whether co-inhibition with MEK1/2- and MERTK-inhibitors has clinical potential in CRC.

Currently, phosphoproteomic profiling of the CRC65 cell line panel is carried out in order to shed light onto differentially active pathways in CRC. This might help explain some of the protein/drug associations reported as part of this project and will be an invaluable resource for research into cancer cell signalling.

References

Aaltonen, L.A., Peltomaki, P., Leach, F.S., Sistonen, P., Pylkkanen, L., Mecklin, J.P., Jarvinen, H., Powell, S.M., Jen, J., Hamilton, S.R.*, et al.* (1993). Clues to the pathogenesis of familial colorectal cancer. Science *260*, 812-816.

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198-207.

Agarwal, A., Koppstein, D., Rozowsky, J., Sboner, A., Habegger, L., Hillier, L.W., Sasidharan, R., Reinke, V., Waterston, R.H., and Gerstein, M. (2010). Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics *11*, 383.

Ahsan, H., Neugut, A.I., Garbowski, G.C., Jacobson, J.S., Forde, K.A., Treat, M.R., and Waye, J.D. (1998). Family history of colorectal adenomatous polyps and increased risk for colorectal cancer. Ann Intern Med *128*, 900-905.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory, 267-281.

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., Garcia Giron, C., Hourlier, T.*, et al.* (2016). The Ensembl gene annotation system. Database : the journal of biological databases and curation *2016*, baw093.

Al-Lazikani, B., Banerji, U., and Workman, P. (2012). Combinatorial drug therapy for cancer in the post-genomic era. Nat Biotechnol *30*, 679-692.

Albain, K.S., Barlow, W.E., Shak, S., Hortobagyi, G.N., Livingston, R.B., Yeh, I.T., Ravdin, P., Bugarini, R., Baehner, F.L., Davidson, N.E.*, et al.* (2010). Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. The Lancet Oncology *11*, 55-65.

Ali, N.A., and Molloy, M.P. (2011). Quantitative phosphoproteomics of transforming growth factor-beta signaling in colon cancer cells. Proteomics *11*, 3390-3401.

Allen, W.L., Coyle, V.M., Jithesh, P.V., Proutski, I., Stevenson, L., Fenning, C., Longley, D.B., Wilson, R.H., Gordon, M., Lenz, H.J.*, et al.* (2008). Clinical determinants of response to irinotecan-based therapy derived from cell line models. Clin Cancer Res *14*, 6647-6655.

Allen, W.L., Turkington, R.C., Stevenson, L., Carson, G., Coyle, V.M., Hector, S., Dunne, P., Van Schaeybroeck, S., Longley, D.B., and Johnston, P.G. (2012). Pharmacogenomic profiling and pathway analyses identify MAPK-dependent

migration as an acute response to SN38 in p53 null and p53-mutant colorectal cancer cells. Mol Cancer Ther *11*, 1724-1734.

Almeida, J.S., Stanislaus, R., Krug, E., and Arthur, J.M. (2005). Normalization and analysis of residual variation in two-dimensional gel electrophoresis for quantitative differential proteomics. Proteomics *5*, 1242-1249.

Alter, O., Brown, P.O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A *97*, 10101-10106.

Amado, R.G., Wolf, M., Peeters, M., Van Cutsem, E., Siena, S., Freeman, D.J., Juan, T., Sikorski, R., Suggs, S., Radinsky, R.*, et al.* (2008). Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol *26*, 1626-1634.

Andersen, J.S., Wilkinson, C.J., Mayor, T., Mortensen, P., Nigg, E.A., and Mann, M. (2003). Proteomic characterization of the human centrosome by protein correlation profiling. Nature *426*, 570-574.

Anderson, N.G., and Anderson, N.L. (1978). Analytical techniques for cell fractions. XXI. Two-dimensional analysis of serum and tissue proteins: multiple isoelectric focusing. Anal Biochem *85*, 331-340.

Andre, T., Boni, C., Navarro, M., Tabernero, J., Hickish, T., Topham, C., Bonetti, A., Clingan, P., Bridgewater, J., Rivera, F.*, et al.* (2009). Improved overall survival with oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment in stage II or III colon cancer in the MOSAIC trial. J Clin Oncol *27*, 3109-3116.

Andre, T., Louvet, C., Maindrault-Goebel, F., Couteau, C., Mabro, M., Lotz, J.P., Gilles-Amar, V., Krulik, M., Carola, E., Izrael, V.*, et al.* (1999). CPT-11 (irinotecan) addition to bimonthly, high-dose leucovorin and bolus and continuous-infusion 5-fluorouracil (FOLFIRI) for pretreated metastatic colorectal cancer. GERCOR. Eur J Cancer *35*, 1343-1347.

Arango, D., Wilson, A.J., Shi, Q., Corner, G.A., Aranes, M.J., Nicholas, C., Lesser, M., Mariadason, J.M., and Augenlicht, L.H. (2004). Molecular mechanisms of action and prediction of response to oxaliplatin in colorectal cancer cells. Br J Cancer *91*, 1931-1946.

Arnold, C.N., Goel, A., and Boland, C.R. (2003). Role of hMLH1 promoter hypermethylation in drug resistance to 5-fluorouracil in colorectal cancer cell lines. Int J Cancer *106*, 66-73.

Arnould, S., Hennebelle, I., Canal, P., Bugat, R., and Guichard, S. (2003). Cellular determinants of oxaliplatin sensitivity in colon cancer cell lines. Eur J Cancer *39*, 112-119.

Aschele, C., Debernardis, D., Bandelloni, R., Cascinu, S., Catalano, V., Giordani, P., Barni, S., Turci, D., Drudi, G., Lonardi, S.*, et al.* (2002). Thymidylate synthase protein expression in colorectal cancer metastases predicts for clinical outcome

to leucovorin-modulated bolus or infusional 5-fluorouracil but not methotrexate-modulated bolus 5-fluorouracil. Ann Oncol *13*, 1882-1892.

Ashraf, S.Q., Nicholls, A.M., Wilding, J.L., Ntouroupi, T.G., Mortensen, N.J., and Bodmer, W.F. (2012). Direct and immune mediated antibody targeting of ERBB receptors in a colorectal cancer cell-line panel. Proc Natl Acad Sci U S A *109*, 21046-21051.

Astler, V.B., and Coller, F.A. (1954). The prognostic significance of direct extension of carcinoma of the colon and rectum. Ann Surg *139*, 846-852.

Aston, F.W. (1919a). LXXIV. A positive ray spectrograph. The London.

Aston, F.W. (1919b). Neon. Nature.

Aviv, H., and Leder, P. (1972). Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid-cellulose. Proc Natl Acad Sci U S A *69*, 1408-1412.

Baek, D., Villen, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. Nature *455*, 64-71.

Balko, J.M., and Black, E.P. (2009). A gene expression predictor of response to EGFR-targeted therapy stratifies progression-free survival to cetuximab in KRAS wild-type metastatic colorectal cancer. BMC Cancer *9*, 145.

Bantscheff, M., Eberhard, D., Abraham, Y., Bastuck, S., Boesche, M., Hobson, S., Mathieson, T., Perrin, J., Raida, M., Rau, C.*, et al.* (2007). Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. Nature biotechnology *25*, 1035-1044.

Barber, M., Bordoli, R.S., Sedgwick, R.D., and Tyler, A.N. (1981). Fast Atom Bombardment of Solids as an Ion-Source in Mass-Spectrometry. Nature *293*, 270-275.

Barbosa-Morais, N.L., Dunning, M.J., Samarajiwa, S.A., Darot, J.F., Ritchie, M.E., Lynch, A.G., and Tavare, S. (2010). A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. Nucleic Acids Res *38*, e17.

Bardelli, A., Parsons, D.W., Silliman, N., Ptak, J., Szabo, S., Saha, S., Markowitz, S., Willson, J.K., Parmigiani, G., Kinzler, K.W.*, et al.* (2003). Mutational analysis of the tyrosine kinome in colorectal cancers. Science *300*, 949.

Barnes, M., Freudenberg, J., Thompson, S., Aronow, B., and Pavlidis, P. (2005). Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic Acids Res *33*, 5914-5923.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D.*, et al.* (2012). The Cancer Cell Line

Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603-607.

Barrett, S.D., Bridges, A.J., Dudley, D.T., Saltiel, A.R., Fergus, J.H., Flamme, C.M., Delaney, A.M., Kaufman, M., LePage, S., Leopold, W.R.*, et al.* (2008). The discovery of the benzhydroxamate MEK inhibitors CI-1040 and PD 0325901. Bioorganic & medicinal chemistry letters *18*, 6501-6504.

Bartlett, D.L., Berlin, J., Lauwers, G.Y., Messersmith, W.A., Petrelli, N.J., and Venook, A.P. (2006). Chemotherapy and regional therapy of hepatic colorectal metastases: expert consensus statement. Ann Surg Oncol *13*, 1284-1292.

Basu, A., Bodycombe, N.E., Cheah, J.H., Price, E.V., Liu, K., Schaefer, G.I., Ebright, R.Y., Stewart, M.L., Ito, D., Wang, S.*, et al.* (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell *154*, 1151-1161.

Bauer, K.M., Lambert, P.A., and Hummon, A.B. (2012). Comparative label-free LC-MS/MS analysis of colorectal adenocarcinoma and metastatic cells treated with 5-fluorouracil. Proteomics *12*, 1928-1937.

Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011). The quantitative proteome of a human cell line. Mol Syst Biol *7*, 549.

Beech, J., and Kelly, K. (2014). Abstract 5594: JNK1 and MERTK are markers of MEK inhibitor resistance and new targets for therapy. Cancer Research *73*, 5594-5594.

Belderbos, T.D., Leenders, M., Moons, L.M., and Siersema, P.D. (2014). Local recurrence after endoscopic mucosal resection of nonpedunculated colorectal lesions: systematic review and meta-analysis. Endoscopy *46*, 388-402.

Bendardaf, R., Lamlum, H., Ristamaki, R., Korkeila, E., Syrjanen, K., and Pyrhonen, S. (2007). Mismatch repair status is a predictive factor of tumour response to 5-fluorouracil and irinotecan chemotherapy in patients with advanced colorectal cancer. Tumour Biol *28*, 212-220.

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C.M., and Marron, J.S. (2004). Adjustment of systematic microarray data biases. Bioinformatics *20*, 105-114.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met *57*, 289-300.

Benson, A.B., 3rd, Schrag, D., Somerfield, M.R., Cohen, A.M., Figueredo, A.T., Flynn, P.J., Krzyzanowska, M.K., Maroun, J., McAllister, P., Van Cutsem, E.*, et al.* (2004). American Society of Clinical Oncology recommendations on adjuvant chemotherapy for stage II colon cancer. J Clin Oncol *22*, 3408-3419.

Berry, R.D., and Paraskeva, C. (1988). Expression of carcinoembryonic antigen by adenoma and carcinoma derived epithelial cell lines: Possible marker of tumour progression and modulation of expression by sodium butyrate. Carcinogenesis *9*, 447-450.

Bertotti, A., Papp, E., Jones, S., Adleff, V., Anagnostou, V., Lupo, B., Sausen, M., Phallen, J., Hruban, C.A., Tokheim, C.*, et al.* (2015). The genomic landscape of response to EGFR blockade in colorectal cancer. Nature *526*, 263-267.

Bielby, J.A. (2011). Coding neoplasms in ICD-10-CM. J AHIMA *82*, 72-74; quiz 76.

Blanco-Calvo, M., Concha, A., Figueroa, A., Garrido, F., and Valladares-Ayerbes, M. (2015). Colorectal Cancer Classification and Cell Heterogeneity: A Systems Oncology Approach. Int J Mol Sci *16*, 13610-13632.

Bo, T.H., Dysvik, B., and Jonassen, I. (2004). LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Res *32*, e34.

Boland, G.M., Chang, G.J., Haynes, A.B., Chiang, Y.-J., Chagpar, R., Xing, Y., Hu, C.-Y., Feig, B.W., You, Y.N., and Cormier, J.N. (2012). Association between adherence to National Comprehensive Cancer Network treatment guidelines and improved survival in patients with colon cancer. Cancer *119*, 1593-1601.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185-193.

Bonelli, L., Martines, H., Conio, M., Bruzzi, P., and Aste, H. (1988). Family history of colorectal cancer as a risk factor for benign and malignant tumours of the large bowel. A case-control study. Int J Cancer *41*, 513-517.

Boulay, J.L., Mild, G., Lowy, A., Reuter, J., Lagrange, M., Terracciano, L., Laffer, U., Herrmann, R., and Rochlitz, C. (2002). SMAD4 is a predictive marker for 5-fluorouracil-based chemotherapy in patients with colorectal cancer. Br J Cancer *87*, 630-634.

Bouvard, V., Loomis, D., Guyton, K.Z., Grosse, Y., Ghissassi, F.E., Benbrahim-Tallaa, L., Guha, N., Mattock, H., Straif, K., and International Agency for Research on Cancer Monograph Working, G. (2015). Carcinogenicity of consumption of red and processed meat. The Lancet Oncology *16*, 1599-1600.

Box, G.E.P. (1979). Robustness in the Strategy of Scientific Model Building. In Robustness in Statistics, R.L. Launer, and G.N. Wilkinson, eds. (Academic Press CY -), pp. 201-236.

Boyer, J., Allen, W.L., McLean, E.G., Wilson, P.M., McCulla, A., Moore, S., Longley, D.B., Caldas, C., and Johnston, P.G. (2006). Pharmacogenomic identification of novel determinants of response to chemotherapy in colon cancer. Cancer Res *66*, 2765-2777.

Bozic, I., Reiter, J.G., Allen, B., Antal, T., Chatterjee, K., Shah, P., Moon, Y.S., Yaqubie, A., Kelly, N., Le, D.T.*, et al.* (2013). Evolutionary dynamics of cancer in response to targeted combination therapy. Elife *2*, e00747.

Bracht, K., Nicholls, A.M., Liu, Y., and Bodmer, W.F. (2010). 5-Fluorouracil response in a large panel of colorectal cancer cell lines is associated with mismatch repair deficiency. Br J Cancer *103*, 340-346.

Bradford, M.M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem *72*, 248-254.

Braun, M.S., Richman, S.D., Quirke, P., Daly, C., Adlard, J.W., Elliott, F., Barrett, J.H., Selby, P., Meade, A.M., Stephens, R.J.*, et al.* (2008). Predictive biomarkers of chemotherapy efficacy in colorectal cancer: results from the UK MRC FOCUS trial. J Clin Oncol *26*, 2690-2698.

Brenner, H., Hoffmeister, M., Stegmaier, C., Brenner, G., Altenhofen, L., and Haug, U. (2007). Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840,149 screening colonoscopies. Gut *56*, 1585-1589.

Brenner, H., Stock, C., and Hoffmeister, M. (2014). Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomised controlled trials and observational studies. Bmj *348*, g2467.

Brett, M.C., Pickard, M., Green, B., Howel-Evans, A., Smith, D., Kinsella, A., and Poston, G. (1996). p53 protein overexpression and response to biomodulated 5-fluorouracil chemotherapy in patients with advanced colorectal cancer. European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology *22*, 182-185.

Briffa, R., Um, I., Faratian, D., Zhou, Y., Turnbull, A.K., Langdon, S.P., and Harrison, D.J. (2015). Multi-Scale Genomic, Transcriptomic and Proteomic Analysis of Colorectal Cancer Cell Lines to Identify Novel Biomarkers. PLoS One *10*, e0144708.

Brognard, J., and Hunter, T. (2011). Protein kinase signaling networks in cancer. Current opinion in genetics & development *21*, 4-11.

Brower, V. (2001). Proteomics: biology in the post-genomic era. Companies all over the world rush to lead the way in the new post-genomics race. EMBO Rep *2*, 558-560.

Browning, M.J., Krausa, P., Rowan, A., Bicknell, D.C., Bodmer, J.G., and Bodmer, W.F. (1993). Tissue typing the HLA-A locus from genomic DNA by sequence-specific PCR: comparison of HLA genotype and surface expression on colorectal tumor cell lines. PNAS *90*, 2842-2845.

Brüderlein, S., van der Bosch, K., Schlag, P., and Schwab, M. (1990). Cytogenetics and DNA amplification in colorectal cancers. Genes, chromosomes & cancer *2*, 63-70.

Bruins, A.P., Covey, T.R., and Henion, J.D. (1987). Ion Spray Interface for Combined Liquid Chromatography/Atmospheric Pressure Ionization Mass-Spectrometry. Analytical Chemistry *59*, 2642-2646.

Budinska, E., Popovici, V., Tejpar, S., D'Ario, G., Lapique, N., Sikora, K.O., Di Narzo, A.F., Yan, P., Hodgson, J.G., Weinrich, S.*, et al.* (2013). Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. J Pathol *231*, 63-76.

Bueno-de-Mesquita, J.M., van Harten, W.H., Retel, V.P., van't Veer, L.J., van Dam, F.S., Karsenberg, K., Douma, K.F., van Tinteren, H., Peterse, J.L., Wesseling, J.*, et al.* (2007). Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER). The Lancet Oncology *8*, 1079-1087.

Burris, H.A., 3rd (2013). Overcoming acquired resistance to anticancer therapy: focus on the PI3K/AKT/mTOR pathway. Cancer Chemother Pharmacol *71*, 829-842.

Calon, A., Lonardo, E., Berenguer-Llergo, A., Espinet, E., Hernando-Momblona, X., Iglesias, M., Sevillano, M., Palomo-Ponce, S., Tauriello, D.V., Byrom, D.*, et al.* (2015). Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. Nat Genet *47*, 320-329.

Canadian Task Force on Preventive Health, C. (2001). Colorectal cancer screening. Recommendation statement from the Canadian Task Force on Preventive Health Care. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne *165*, 206-208.

Cancer Cell Line Encyclopedia Consortium, and Genomics of Drug Sensitivity in Cancer Consortium (2015). Pharmacogenomic agreement between two cancer cell line data sets. Nature *528*, 84-87.

Cappuzzo, F., Finocchiaro, G., Rossi, E., Janne, P.A., Carnaghi, C., Calandri, C., Bencardino, K., Ligorio, C., Ciardiello, F., Pressiani, T.*, et al.* (2008). EGFR FISH assay predicts for response to cetuximab in chemotherapy refractory colorectal cancer patients. Ann Oncol *19*, 717-723.

Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S., Ami, G.O.H., and Web Presence Working, G. (2009). AmiGO: online access to ontology and annotation data. Bioinformatics *25*, 288-289.

Carethers, J.M., Chauhan, D.P., Fink, D., Nebel, S., Bresalier, R.S., Howell, S.B., and Boland, C.R. (1999). Mismatch repair proficiency and in vitro response to 5-fluorouracil. Gastroenterology *117*, 123-131.

Carethers, J.M., Smith, E.J., Behling, C.A., Nguyen, L., Tajima, A., Doctolero, R.T., Cabrera, B.L., Goel, M., Arnold, C.A., Miyai, K., *et al.* (2004). Use of 5-fluorouracil and survival in patients with microsatellite-unstable colorectal cancer. Gastroenterology *126*, 394-401.

Carlson, M. (2016). UniProt.ws: R Interface to UniProt Web Services.

Cascinu, S., Aschele, C., Barni, S., Debernardis, D., Baldo, C., Tunesi, G., Catalano, V., Staccioli, M.P., Brenna, A., Muretto, P., *et al.* (1999). Thymidylate synthase protein expression in advanced colon cancer: correlation with the site of metastasis and the clinical response to leucovorin-modulated bolus 5-fluorouracil. Clin Cancer Res *5*, 1996-1999.

Cascinu, S., Catalano, V., Aschele, C., Barni, S., Debernardis, D., Gallo, L., Bandelloni, R., Staccioli, M.P., Baldelli, A.M., Brenna, A., *et al.* (2000). Immunohistochemical determination of p53 protein does not predict clinical response in advanced colorectal cancer with low thymidylate synthase expression receiving a bolus 5-fluorouracil-leucovorin combination. Ann Oncol *11*, 1053-1056.

Cecchin, E., Innocenti, F., D'Andrea, M., Corona, G., De Mattia, E., Biason, P., Buonadonna, A., and Toffoli, G. (2009). Predictive role of the UGT1A1, UGT1A7, and UGT1A9 genetic variants and their haplotypes on the outcome of metastatic colorectal cancer patients treated with fluorouracil, leucovorin, and irinotecan. J Clin Oncol *27*, 2457-2465.

Chan, A.T., and Giovannucci, E.L. (2010). Primary prevention of colorectal cancer. Gastroenterology *138*, 2029-2043 e2010.

Chen, B.J., Litvin, O., Ungar, L., and Pe'er, D. (2015). Context Sensitive Modeling of Cancer Drug Sensitivity. PLoS One *10*, e0133850.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., and Liu, C. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One *6*, e17238.

Chen, L.S., Prentice, R.L., and Wang, P. (2014). A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. Biometrics *70*, 312-322.

Chi, A., Huttenhower, C., Geer, L.Y., Coon, J.J., Syka, J.E., Bai, D.L., Shabanowitz, J., Burke, D.J., Troyanskaya, O.G., and Hunt, D.F. (2007). Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. Proc Natl Acad Sci U S A *104*, 2193-2198.

Chich, J.F., David, O., Villers, F., Schaeffer, B., Lutomski, D., and Huet, S. (2007). Statistics for proteomics: experimental design and 2-DE differential analysis. J Chromatogr B Analyt Technol Biomed Life Sci *849*, 261-272.

Chick, J.M., Munger, S.C., Simecek, P., Huttlin, E.L., Choi, K., Gatti, D.M., Raghupathy, N., Svenson, K.L., Churchill, G.A., and Gygi, S.P. (2016). Defining the consequences of genetic variation on a proteome-wide scale. Nature *534*, 500-505.

Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J., and Rutter, W.J. (1979). Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry *18*, 5294-5299.

Church, T.R., Wandell, M., Lofton-Day, C., Mongin, S.J., Burger, M., Payne, S.R., Castanos-Velez, E., Blumenstein, B.A., Rosch, T., Osborn, N.*, et al.* (2014). Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. Gut *63*, 317-325.

Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I.*, et al.* (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res *44*, e71.

Compton, C.C., Fielding, L.P., Burgart, L.J., Conley, B., Cooper, H.S., Hamilton, S.R., Hammond, M.E., Henson, D.E., Hutter, R.V., Nagle, R.B.*, et al.* (2000). Prognostic factors in colorectal cancer. College of American Pathologists Consensus Statement 1999. Archives of pathology & laboratory medicine *124*, 979-994.

Compton, C.C., and Greene, F.L. (2004). The staging of colorectal cancer: 2004 and beyond. CA Cancer J Clin *54*, 295-308.

Cooks, R.G., and Rockwood, A.L. (1991). The Thomson - a Suggested Unit for Mass Spectroscopists. Rapid Communications in Mass Spectrometry *5*, 93-93.

Cooper, H.S., Deppisch, L.M., Gourley, W.K., Kahn, E.I., Lev, R., Manley, P.N., Pascal, R.R., Qizilbash, A.H., Rickert, R.R., Silverman, J.F.*, et al.* (1995). Endoscopically removed malignant colorectal polyps: clinicopathologic correlations. Gastroenterology *108*, 1657-1665.

Coscia, F., Watters, K.M., Curtis, M., Eckert, M.A., Chiang, C.Y., Tyanova, S., Montag, A., Lastra, R.R., Lengyel, E., and Mann, M. (2016). Integrative proteomic profiling of ovarian cancer cell lines reveals precursor cell associated proteins and functional status. Nat Commun *7*, 12645.

Covey, T.R., Bonner, R.F., Shushan, B.I., and Henion, J. (1988). The determination of protein, oligonucleotide and peptide molecular weights by ion-spray mass spectrometry. Rapid Commun Mass Spectrom *2*, 249-256.

Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics *13*, 2513-2526.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol *26*, 1367-1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res *10*, 1794-1805.

Cranley, J.P., Petras, R.E., Carey, W.D., Paradis, K., and Sivak, M.V. (1986). When is endoscopic polypectomy adequate therapy for colonic polyps containing invasive carcinoma? Gastroenterology *91*, 419-427.

Cunningham, D., Humblet, Y., Siena, S., Khayat, D., Bleiberg, H., Santoro, A., Bets, D., Mueser, M., Harstrick, A., Verslype, C.*, et al.* (2004). Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. N Engl J Med *351*, 337-345.

Dan, S., Tsunoda, T., Kitahara, O., Yanagawa, R., Zembutsu, H., Katagiri, T., Yamazaki, K., Nakamura, Y., and Yamori, T. (2002). An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines. Cancer Res *62*, 1139-1147.

Daub, H., Olsen, J.V., Bairlein, M., Gnad, F., Oppermann, F.S., Korner, R., Greff, Z., Keri, G., Stemmann, O., and Mann, M. (2008). Kinase-selective enrichment enables quantitative phosphoproteomics of the kinome across the cell cycle. Mol Cell *31*, 438-448.

Davies, M.M., Johnston, P.G., Kaur, S., and Allen-Mersh, T.G. (1999). Colorectal liver metastasis thymidylate synthase staining correlates with response to hepatic arterial floxuridine. Clin Cancer Res *5*, 325-328.

Davis, S., and Meltzer, P.S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics *23*, 1846-1847.

De Roock, W., Claes, B., Bernasconi, D., De Schutter, J., Biesmans, B., Fountzilas, G., Kalogeras, K.T., Kotoula, V., Papamichael, D., Laurent-Puig, P.*, et al.* (2010). Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. The Lancet Oncology *11*, 753-762.

De Sousa, E.M.F., Wang, X., Jansen, M., Fessler, E., Trinh, A., de Rooij, L.P., de Jong, J.H., de Boer, O.J., van Leersum, R., Bijlsma, M.F.*, et al.* (2013). Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nat Med *19*, 614-618.

Deeb, S.J., D'Souza, R.C., Cox, J., Schmidt-Supprian, M., and Mann, M. (2012). Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. Mol Cell Proteomics *11*, 77-89.

Del Rio, M., Molina, F., Bascoul-Mollevi, C., Copois, V., Bibeau, F., Chalbos, P., Bareil, C., Kramar, A., Salvetat, N., Fraslon, C.*, et al.* (2007). Gene expression signature in advanced colorectal cancer patients select drugs and response for the use of leucovorin, fluorouracil, and irinotecan. J Clin Oncol *25*, 773-780.

Dempster, A.J. (1918). A new Method of Positive Ray Analysis. Physical Review *11*, 316-325.

Dempster, A.J. (1921). Positive Ray Analysis of Lithium and Magnesium. Physical Review *18*, 415-422.

DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A., and Trent, J.M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet *14*, 457-460.

Des Guetz, G., Uzzan, B., Morere, J.F., Perret, G., and Nicolas, P. (2010). Duration of adjuvant chemotherapy for patients with non-metastatic colorectal cancer. Cochrane Database Syst Rev, CD007046.

deVos, T., Tetzner, R., Model, F., Weiss, G., Schuster, M., Distler, J., Steiger, K.V., Grutzmann, R., Pilarsky, C., Habermann, J.K.*, et al.* (2009). Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. Clin Chem *55*, 1337-1346.

Di Nicolantonio, F., Martini, M., Molinari, F., Sartore-Bianchi, A., Arena, S., Saletti, P., De Dosso, S., Mazzucchelli, L., Frattini, M., Siena, S.*, et al.* (2008). Wild-type BRAF is required for response to panitumumab or cetuximab in metastatic colorectal cancer. J Clin Oncol *26*, 5705-5712.

Diaz Jr, L.A., Williams, R.T., Wu, J., Kinde, I., Hecht, J.R., Berlin, J., Allen, B., Bozic, I., Reiter, J.G., Nowak, M.A.*, et al.* (2012). The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. Nature *486*, 537-540.

Dole, M., Mack, L.L., and Hines, R.L. (1968). Molecular Beams of Macroions. Journal of Chemical Physics *49*, 2240-&.

Douillard, J.Y., Cunningham, D., Roth, A.D., Navarro, M., James, R.D., Karasek, P., Jandik, P., Iveson, T., Carmichael, J., Alakl, M.*, et al.* (2000). Irinotecan combined with fluorouracil compared with fluorouracil alone as first-line treatment for metastatic colorectal cancer: a multicentre randomised trial. Lancet *355*, 1041-1047.

Dry, J.R., Pavey, S., Pratilas, C.A., Harbron, C., Runswick, S., Hodgson, D., Chresta, C., McCormack, R., Byrne, N., Cockerill, M.*, et al.* (2010). Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244). Cancer Res *70*, 2264-2273.

Duffy, M.J., Lamerz, R., Haglund, C., Nicolini, A., Kalousova, M., Holubec, L., and Sturgeon, C. (2014). Tumor markers in colorectal cancer, gastric cancer and gastrointestinal stromal cancers: European group on tumor markers 2014 guidelines update. Int J Cancer *134*, 2513-2522.

Duffy, M.J., van Dalen, A., Haglund, C., Hansson, L., Holinski-Feder, E., Klapdor, R., Lamerz, R., Peltomäki, P., Sturgeon, C., and Topolcan, O. (2007). Tumour markers in colorectal cancer: European Group on Tumour Markers (EGTM) guidelines for clinical use. In European Journal of Cancer, pp. 1348-1360.

Dukes, C.E., and Bussey, H.J. (1958). The spread of rectal cancer and its effect on prognosis. Br J Cancer *12*, 309-320.

Dunn, E.F., Iida, M., Myers, R.A., Campbell, D.A., Hintz, K.A., Armstrong, E.A., Li, C., and Wheeler, D.L. (2011). Dasatinib sensitizes KRAS mutant colorectal tumors to cetuximab. Oncogene *30*, 561-574.

Dunning, M.J., Smith, M.L., Ritchie, M.E., and Tavare, S. (2007). beadarray: R classes and methods for Illumina bead-based data. Bioinformatics *23*, 2183-2184.

Durbin, K.R., Fornelli, L., Fellers, R.T., Doubleday, P.F., Narita, M., and Kelleher, N.L. (2016). Quantitation and Identification of Thousands of Human Proteoforms below 30 kDa. J Proteome Res *15*, 976-982.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439-3440.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc *4*, 1184-1191.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res *30*, 207-210.

Edge, S.B. (2010). AJCC cancer staging manual, 7th edn (New York, London: Springer).

Edman, P. (1956). On the Mechanism of the Phenyl Isothiocyanate Degradation of Peptides. Acta Chemica Scandinavica *10*, 761-768.

Edman, P., and Begg, G. (1967). A protein sequenator. Eur J Biochem *1*, 80-91.

Edwards, D. (2003). Non-linear normalization and background correction in one-channel cDNA microarray studies. Bioinformatics *19*, 825-833.

Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods *4*, 207-214.

Elsaleh, H., Powell, B., McCaul, K., Grieu, F., Grant, R., Joseph, D., and Iacopetta, B. (2001). P53 alteration and microsatellite instability have predictive value for survival benefit from chemotherapy in stage III colorectal carcinoma. Clin Cancer Res *7*, 1343-1349.

Emaduddin, M., Bicknell, D.C., Bodmer, W.F., and Feller, S.M. (2008). Cell growth, global phosphotyrosine elevation, and c-Met phosphorylation through Src family kinases in colorectal cancer cells. Proc Natl Acad Sci U S A *105*, 2358-2362.

Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom *5*, 976-989.

Fan, H., Zhao, H., Pang, L., Liu, L., Zhang, G., Yu, F., Liu, T., Xu, C., Xiao, Y., and Li, X. (2015). Systematically Prioritizing Functional Differentially Methylated Regions (fDMRs) by Integrating Multi-omics Data in Colorectal Cancer. Sci Rep *5*, 12789.

Fanayan, S., Smith, J.T., Lee, L.Y., Yan, F., Snyder, M., Hancock, W.S., and Nice, E. (2013a). Proteogenomic analysis of human colon carcinoma cell lines LIM1215, LIM1899, and LIM2405. J Proteome Res *12*, 1732-1742.

Fanayan, S., Smith, J.T., Sethi, M.K., Cantor, D., Goode, R., Simpson, R.J., Baker, M.S., Hancock, W.S., and Nice, E. (2013b). Chromosome 7-centric analysis of proteomics data from a panel of human colon carcinoma cell lines. J Proteome Res *12*, 89-96.

Fearon, E.R., and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. Cell *61*, 759-767.

Fedorov, O., Müller, S., and Knapp, S. (2010). The (un)targeted cancer kinome. Nature Chemical Biology *6*, 166-169.

Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., and Whitehouse, C.M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. Science *246*, 64-71.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer *136*, E359-386.

Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J.W., Comber, H., Forman, D., and Bray, F. (2013). Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. European Journal of Cancer *49*, 1374-1403.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. Science *251*, 767-773.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S.*, et al.* (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res *43*, D805-811.

Frattini, M., Saletti, P., Romagnani, E., Martin, V., Molinari, F., Ghisletta, M., Camponovo, A., Etienne, L.L., Cavalli, F., and Mazzucchelli, L. (2007). PTEN loss of expression predicts cetuximab efficacy in metastatic colorectal cancer patients. Br J Cancer *97*, 1139-1145.

Frese, C.K., Altelaar, A.F., van den Toorn, H., Nolting, D., Griep-Raming, J., Heck, A.J., and Mohammed, S. (2012). Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. Anal Chem *84*, 9668-9673.

Frese, C.K., Nolting, D., Altelaar, A.F., Griep-Raming, J., Mohammed, S., and Heck, A.J. (2013a). Characterization of electron transfer dissociation in the Orbitrap Velos HCD cell. J Am Soc Mass Spectrom *24*, 1663-1670.

Frese, C.K., Zhou, H., Taus, T., Altelaar, A.F., Mechtler, K., Heck, A.J., and Mohammed, S. (2013b). Unambiguous phosphosite localization using electron-transfer/higher-energy collision dissociation (EThcD). J Proteome Res *12*, 1520-1525.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw *33*, 1-22.

Fu, L., and Medico, E. (2007). FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinformatics *8*, 3.

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R.*, et al.* (2009). Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics *10*, 161.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nature reviews Cancer *4*, 177-183.

Gagnon-Bartsch, J.A., and Speed, T.P. (2012). Using control genes to correct for unwanted variation in microarray data. Biostatistics *13*, 539-552.

Galiatsatos, P., and Foulkes, W.D. (2006). Familial adenomatous polyposis. Am J Gastroenterol *101*, 385-398.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J.*, et al.* (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570-575.

Garraway, L.A., and Janne, P.A. (2012). Circumventing cancer drug resistance in the era of personalized medicine. Cancer Discov *2*, 214-226.

Garraway, L.A., Widlund, H.R., Rubin, M.A., Getz, G., Berger, A.J., Ramaswamy, S., Beroukhim, R., Milner, D.A., Granter, S.R., Du, J.*, et al.* (2005). Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature *436*, 117-122.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics *11*, M111 014050.

Gerber, G.E., Anderegg, R.J., and Herlihy, W.C. (1979). Partial primary structure of bacteriorhodopsin: sequencing methods for membrane proteins. Proc Natl Acad Sci U S A *76*, 227-231.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737-741.

Gholami, A.M., Hahne, H., Wu, Z., Auer, F.J., Meng, C., Wilhelm, M., and Kuster, B. (2013). Global proteome analysis of the NCI-60 cell line panel. Cell Rep *4*, 609-620.

Ghosh, D., Yu, H., Tan, X.F., Lim, T.K., Zubaidah, R.M., Tan, H.T., Chung, M.C., and Lin, Q. (2011). Identification of key players for colorectal cancer metastasis by iTRAQ quantitative proteomics profiling of isogenic SW480 and SW620 cell lines. J Proteome Res *10*, 4373-4387.

Giancotti, F.G. (2014). Deregulation of cell signaling in cancer. FEBS Lett *588*, 2558-2570.

Giansanti, P., Tsiatsiani, L., Low, T.Y., and Heck, A.J. (2016). Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. Nat Protoc *11*, 993-1006.

Goldberg, R.M. (2006). Therapy for metastatic colorectal cancer. Oncologist *11*, 981-987.

Goldberg, R.M., Rothenberg, M.L., Van Cutsem, E., Benson, A.B., 3rd, Blanke, C.D., Diasio, R.B., Grothey, A., Lenz, H.J., Meropol, N.J., Ramanathan, R.K.*, et al.* (2007). The continuum of care: a paradigm for the management of metastatic colorectal cancer. Oncologist *12*, 38-50.

Goldberg, R.M., Sargent, D.J., Morton, R.F., Fuchs, C.S., Ramanathan, R.K., Williamson, S.K., Findlay, B.P., Pitot, H.C., and Alberts, S.R. (2004). A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. J Clin Oncol *22*, 23-30.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A.*, et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531-537.

Gonzalez de Castro, D., Clarke, P.A., Al-Lazikani, B., and Workman, P. (2013). Personalized cancer medicine: molecular diagnostics, predictive biomarkers, and drug resistance. Clin Pharmacol Ther *93*, 252-259.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet *17*, 333-351.

Graham, D.K., DeRyckere, D., Davies, K.D., and Earp, H.S. (2014). The TAM family: phosphatidylserine sensing receptor tyrosine kinases gone awry in cancer. Nat Rev Cancer *14*, 769-785.

Gray, R.G., Quirke, P., Handley, K., Lopatin, M., Magill, L., Baehner, F.L., Beaumont, C., Clark-Langone, K.M., Yoshizawa, C.N., Lee, M.*, et al.* (2011). Validation study of a quantitative multigene reverse transcriptase-polymerase chain reaction assay for assessment of recurrence risk in patients with stage II colon cancer. J Clin Oncol *29*, 4611-4619.

Greenhalgh, D.A., and Kinsella, A.R. (1985). c-Ha-ras not c-Ki-ras activation in three colon tumour cell lines. Carcinogenesis *6*, 1533-1535.

Greshock, J., Bachman, K.E., Degenhardt, Y.Y., Jing, J., Wen, Y.H., Eastman, S., McNeil, E., Moy, C., Wegrzyn, R., Auger, K.*, et al.* (2010). Molecular target class is predictive of in vitro response profile. Cancer Res *70*, 3677-3686.

Gross, S., Rahal, R., Stransky, N., Lengauer, C., and Hoeflich, K.P. (2015). Targeting cancer with kinase inhibitors. J Clin Invest *125*, 1780-1789.

Grothey, A., Van Cutsem, E., Sobrero, A., Siena, S., Falcone, A., Ychou, M., Humblet, Y., Bouche, O., Mineur, L., Barone, C.*, et al.* (2013). Regorafenib monotherapy for previously treated metastatic colorectal cancer (CORRECT): an international, multicentre, randomised, placebo-controlled, phase 3 trial. Lancet *381*, 303-312.

Group, E.C.C.S.G.W., von Karsa, L., Patnick, J., Segnan, N., Atkin, W., Halloran, S., Lansdorp-Vogelaar, I., Malila, N., Minozzi, S., Moss, S.*, et al.* (2013). European guidelines for quality assurance in colorectal cancer screening and diagnosis: overview and introduction to the full supplement publication. In Endoscopy, pp. 51-59.

Grundner-Culemann, K., Dybowski, J.N., Klammer, M., Tebbe, A., Schaab, C., and Daub, H. (2016). Comparative proteome analysis across non-small cell lung cancer cell lines. J Proteomics *130*, 1-10.

Grutzmann, R., Molnar, B., Pilarsky, C., Habermann, J.K., Schlag, P.M., Saeger, H.D., Miehlke, S., Stolz, T., Model, F., Roblick, U.J.*, et al.* (2008). Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. PLoS One *3*, e3759.

Guinney, J., Dienstmann, R., Wang, X., de Reynies, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P.*, et al.* (2015). The consensus molecular subtypes of colorectal cancer. Nat Med *21*, 1350-1356.

Gunderson, K.L., Kruglyak, S., Graige, M.S., Garcia, F., Kermani, B.G., Zhao, C., Che, D., Dickinson, T., Wickham, E., Bierle, J.*, et al.* (2004). Decoding randomly ordered DNA arrays. Genome research *14*, 870-877.

Gunderson, L.L., Jessup, J.M., Sargent, D.J., Greene, F.L., and Stewart, A. (2010a). Revised tumor and node categorization for rectal cancer based on surveillance, epidemiology, and end results and rectal pooled analysis outcomes. J Clin Oncol *28*, 256-263.

Gunderson, L.L., Jessup, J.M., Sargent, D.J., Greene, F.L., and Stewart, A.K. (2010b). Revised TN categorization for colon cancer based on national survival outcomes data. J Clin Oncol *28*, 264-271.

Guo, J., Xu, N., Li, Z., Zhang, S., Wu, J., Kim, D.H., Sano Marma, M., Meng, Q., Cao, H., Li, X.*, et al.* (2008). Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proc Natl Acad Sci U S A *105*, 9145-9150.

Guttmacher, A.E., and Collins, F.S. (2003). Welcome to the genomic era. N Engl J Med *349*, 996-998.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Mol Cell Biol *19*, 1720-1730.

Habbema, J.D., Wilt, T.J., Etzioni, R., Nelson, H.D., Schechter, C.B., Lawrence, W.F., Melnikow, J., Kuntz, K.M., Owens, D.K., and Feuer, E.J. (2014). Models in the development of clinical practice guidelines. Ann Intern Med *161*, 812-818.

Haggitt, R.C., Glotzbach, R.E., Soffer, E.E., and Wruble, L.D. (1985). Prognostic factors in colorectal carcinomas arising in adenomas: implications for lesions removed by endoscopic polypectomy. Gastroenterology *89*, 328-336.

Hahne, H., Pachl, F., Ruprecht, B., Maier, S.K., Klaeger, S., Helm, D., Medard, G., Wilm, M., Lemeer, S., and Kuster, B. (2013). DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. Nat Methods *10*, 989-991.

Haibe-Kains, B., El-Hachem, N., Birkbak, N.J., Jin, A.C., Beck, A.H., Aerts, H.J., and Quackenbush, J. (2013). Inconsistency in large pharmacogenomic studies. Nature *504*, 389-393.

Haller, D.G., Catalano, P.J., Macdonald, J.S., O'Rourke, M.A., Frontiera, M.S., Jackson, D.V., and Mayer, R.J. (2005). Phase III study of fluorouracil, leucovorin, and levamisole in high-risk stage II and III colon cancer: final report of Intergroup 0089. J Clin Oncol *23*, 8671-8678.

Haller, D.G., Tabernero, J., Maroun, J., de Braud, F., Price, T., Van Cutsem, E., Hill, M., Gilberg, F., Rittweger, K., and Schmoll, H.J. (2011). Capecitabine plus oxaliplatin compared with fluorouracil and folinic acid as adjuvant therapy for stage III colon cancer. J Clin Oncol *29*, 1465-1471.

Halvey, P.J., Wang, X., Wang, J., Bhat, A.A., Dhawan, P., Li, M., Zhang, B., Liebler, D.C., and Slebos, R.J. (2014). Proteogenomic analysis reveals unanticipated adaptations of colorectal tumor cells to deficiencies in DNA mismatch repair. Cancer Res *74*, 387-397.

Hancock, W.S., Bishop, C.A., and Hearn, M.T. (1976). High pressure liquid chromatography in the analysis of underivatised peptides using a sensitive and rapid procedure. FEBS Lett *72*, 139-142.

Hancock, W.S., Bishop, C.A., Prestidge, R.L., Harding, D.R., and Hearn, M.T. (1978a). Reversed-phase, high-pressure liquid chromatography of peptides and proteins with ion-pairing reagents. Science *200*, 1168-1170.

Hancock, W.S., Bishop, C.A., Prestidge, R.L., Harding, D.R.K., and Hearn, M.T.W. (1978b). High-Pressure Liquid-Chromatography of Peptides and Proteins .2. Use of Phosphoric-Acid in Analysis of Un-Derivatized Peptides by Reversed-Phase High-Pressure Liquid-Chromatography. Journal of Chromatography *153*, 391-398.

Hart, T., Komori, H.K., LaMere, S., Podshivalova, K., and Salomon, D.R. (2013). Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics *14*, 778.

Haslam, A.C., Pearson, J.M., Bisset, D.L., Michie, H.R., and Hobbiss, J.H. (2006). A Dukes/Jass combination--is it more discriminating? Colorectal Dis *8*, 418-422.

Haverty, P.M., Lin, E., Tan, J., Yu, Y., Lam, B., Lianoglou, S., Neve, R.M., Martin, S., Settleman, J., Yauch, R.L.*, et al.* (2016). Reproducible pharmacogenomic profiling of cancer cell line panels. Nature *533*, 333-337.

Hebenstreit, D., Fang, M., Gu, M., Charoensawan, V., van Oudenaarden, A., and Teichmann, S.A. (2011). RNA sequencing reveals two major classes of gene expression levels in metazoan cells. Mol Syst Biol *7*, 497.

Heider, A., and Alt, R. (2013). virtualArray: a R/bioconductor package to merge raw data from different microarray platforms. BMC Bioinformatics *14*, 75.

Hellman, U., Wernstedt, C., Gonez, J., and Heldin, C.H. (1995). Improvement of an "In-Gel" digestion procedure for the micropreparation of internal protein fragments for amino acid sequencing. Anal Biochem *224*, 451-455.

Hemminki, K., and Chen, B.W. (2004). Familial risk for colorectal cancers are mainly due heritable causes. Cancer Epidem Biomar *13*, 1253-1256.

Hemminki, K., and Eng, C. (2004). Clinical genetic counselling for familial cancers requires reliable data on familial cancer risks and general action plans. J Med Genet *41*, 801-807.

Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C., and Watanabe, C. (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. Proc Natl Acad Sci U S A *90*, 5011-5015.

Henzel, W.J., Watanabe, C., and Stults, J.T. (2003). Protein identification: the origins of peptide mass fingerprinting. J Am Soc Mass Spectrom *14*, 931-942.

Hertz, H.S., Hites, R.A., and Biemann, K. (1971). Identification of Mass Spectra by Computer-Searching a File of Known Spectra. Analytical Chemistry *43*, 681-&.

Hess, S. (2013). The emerging field of chemo- and pharmacoproteomics. Proteomics Clin Appl *7*, 171-180.

Hewitson, P., Glasziou, P., Watson, E., Towler, B., and Irwig, L. (2008). Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): an update. Am J Gastroenterol *103*, 1541-1549.

Hites, R.A., and Biemann, K. (1967). A Computer-Compatible Digital Data Acquisition System for Fast-Scanning Single-Focusing Mass Spectrometers. Analytical Chemistry *39*, 965-&.

Hites, R.A., and Biemann, K. (1968). Mass Spectrometer-Computer System Particularly Suited for Gas Chromatography of Complex Mixtures. Analytical Chemistry *40*, 1217-&.

Hites, R.A., and Biemann, K. (1970). Computer Evaluation of Continuously Scanned Mass Spectra of Gas Chromatographic Effluents. Analytical Chemistry *42*, 855-&.

Hoerl, A.E., and Kennard, R.W. (1970). Ridge Regression - Biased Estimation for Nonorthogonal Problems. Technometrics *12*, 55-&.

Holleczek, B., Rossi, S., Domenic, A., Innos, K., Minicozzi, P., Francisci, S., Hackl, M., Eisemann, N., Brenner, H., and Group, E.-W. (2015). On-going improvement and persistent differences in the survival for patients with colon and rectum cancer across Europe 1999-2007 - Results from the EUROCARE-5 study. Eur J Cancer.

Holohan, C., Van Schaeybroeck, S., Longley, D.B., and Johnston, P.G. (2013). Cancer drug resistance: an evolving paradigm. Nat Rev Cancer *13*, 714-726.

Howard, G.A., and Martin, A.J. (1950). The separation of the C12-C18 fatty acids by reversed-phase partition chromatography. Biochem J *46*, 532-538.

Huang, W., Liu, G., Zhang, X., Fu, W., Zheng, S., Wu, Q., Liu, C., Liu, Y., Cai, S., and Huang, Y. (2014). Cost-effectiveness of colorectal cancer screening protocols in urban Chinese populations. PLoS One *9*, e109150.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T.*, et al.* (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods *12*, 115-121.

Hunt, D.F., Buko, A.M., Ballard, J.M., Shabanowitz, J., and Giordani, A.B. (1981). Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer. Biomed Mass Spectrom *8*, 397-408.

Huttlin, E.L., Chen, X., Barrett-Wilt, G.A., Hegeman, A.D., Halberg, R.B., Harms, A.C., Newton, M.A., Dove, W.F., and Sussman, M.R. (2009). Discovery and validation of colonic tumor-associated proteins via metabolic labeling and stable isotopic dilution. Proc Natl Acad Sci U S A *106*, 17235-17240.

Ichii, S., Takeda, S., Horii, A., Nakatsuru, S., Miyoshi, Y., Emi, M., Fujiwara, Y., Koyama, K., Furuyama, J., Utsunomiya, J.*, et al.* (1993). Detailed analysis of genetic alterations in colorectal tumors from patients with and without familial adenomatous polyposis (FAP). Oncogene *8*, 2399-2405.

Ichikawa, W., Uetake, H., Shirota, Y., Yamada, H., Nishi, N., Nihei, Z., Sugihara, K., and Hirayama, R. (2003). Combination of dihydropyrimidine dehydrogenase and thymidylate synthase gene expressions in primary tumors as predictive parameters for the efficacy of fluoropyrimidine-based chemotherapy for metastatic colorectal cancer. Clin Cancer Res *9*, 786-791.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science *292*, 929-934.

Ikonomou, G., Kostourou, V., Shirasawa, S., Sasazuki, T., Samiotaki, M., and Panayotou, G. (2012). Interplay between oncogenic K-Ras and wild-type H-Ras in Caco2 cell transformation. J Proteomics *75*, 5356-5369.

Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Goncalves, E., Barthorpe, S., Lightfoot, H.*, et al.* (2016). A Landscape of Pharmacogenomic Interactions in Cancer. Cell *166*, 740-754.

Iribarne, J.V., and Thomson, B.A. (1976). On the evaporation of small ions from charged droplets. The Journal of Chemical Physics *64*, 2287.

Irizarry, R.A., Gautier, L., Huber, W., and Bolstad, B. (2006). makecdfenv: CDF Environment Maker.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249-264.

Isella, C., Terrasi, A., Bellomo, S.E., Petti, C., Galatola, G., Muratore, A., Mellano, A., Senetta, R., Cassenti, A., Sonetto, C.*, et al.* (2015). Stromal contribution to the colorectal cancer transcriptome. Nat Genet *47*, 312-319.

Iwaizumi, M., Tseng-Rogenski, S., and Carethers, J.M. (2011). DNA mismatch repair proficiency executing 5-fluorouracil cytotoxicity in colorectal cancer cells. Cancer biology & therapy *12*, 756-764.

Jacobs, B., De Roock, W., Piessevaux, H., Van Oirbeek, R., Biesmans, B., De Schutter, J., Fieuws, S., Vandesompele, J., Peeters, M., Van Laethem, J.L.*, et al.* (2009). Amphiregulin and epiregulin mRNA expression in primary tumors predicts outcome in metastatic colorectal cancer treated with cetuximab. J Clin Oncol *27*, 5068-5074.

James, P. (1997). Protein identification in the post-genome era: the rapid rise of proteomics. Q Rev Biophys *30*, 279-331.

James, P., Quadroni, M., Carafoli, E., and Gonnet, G. (1993). Protein identification by mass profile fingerprinting. Biochem Biophys Res Commun *195*, 58-64.

Jass, J.R. (1995). Malignant colorectal polyps. Gastroenterology *109*, 2034-2035.

Jass, J.R., and Morson, B.C. (1987). Reporting colorectal cancer. J Clin Pathol *40*, 1016-1023.

Jensen, N.F., Smith, D.H., Nygard, S.B., Romer, M.U., Nielsen, K.V., and Brunner, N. (2012). Predictive biomarkers with potential of converting conventional chemotherapy to targeted therapy in patients with metastatic colorectal cancer. Scand J Gastroenterol *47*, 340-355.

Jensen, N.F., Stenvang, J., Beck, M.K., Hanakova, B., Belling, K.C., Do, K.N., Viuff, B., Nygard, S.B., Gupta, R., Rasmussen, M.H.*, et al.* (2015). Establishment and characterization of models of chemotherapy resistance in colorectal cancer: Towards a predictive signature of chemoresistance. Mol Oncol *9*, 1169-1185.

Jessup, J.M., Gunderson, L.L., Greene, F.L., Washington, M.K., Compton, C.C., Sobin, L.H., Minsky, B., Goldberg, R.M., and Hamilton, S.R. (2011). 2010 Staging System for Colon and Rectal Carcinoma. Annals of Surgical Oncology *18*, 1513-1517.

Jhawer, M., Goel, S., Wilson, A.J., Montagna, C., Ling, Y.H., Byun, D.S., Nasser, S., Arango, D., Shin, J., Klampfer, L.*, et al.* (2008). PIK3CA mutation/PTEN expression status predicts response of colon cancer cells to the epidermal growth factor receptor inhibitor cetuximab. Cancer Res *68*, 1953-1961.

Ji, H., Whitehead, R.H., Reid, G.E., Moritz, R.L., Ward, L.D., and Simpson, R.J. (1994). Two-dimensional electrophoretic analysis of proteins expressed by normal and cancerous human crypts: application of mass spectrometry to peptide-mass fingerprinting. Electrophoresis *15*, 391-405.

Jiang, H., Deng, Y., Chen, H.S., Tao, L., Sha, Q., Chen, J., Tsai, C.J., and Zhang, S. (2004). Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics *5*, 81.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118-127.

Jones, M.H., Numata, M., and Shimane, M. (1997). Identification and characterization of BRDT: A testis-specific gene related to the bromodomain genes RING3 and Drosophila fsh. Genomics *45*, 529-534.

Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., Raychowdhury, R.*, et al.* (2015). Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. Science *347*, 1259038.

Jover, R., Zapater, P., Castells, A., Llor, X., Andreu, M., Cubiella, J., Balaguer, F., Sempere, L., Xicola, R.M., Bujanda, L.*, et al.* (2009). The efficacy of adjuvant

chemotherapy with 5-fluorouracil in colorectal cancer depends on the mismatch repair status. Eur J Cancer *45*, 365-373.

Jover, R., Zapater, P., Castells, A., Llor, X., Andreu, M., Cubiella, J., Pinol, V., Xicola, R.M., Bujanda, L., Rene, J.M.*, et al.* (2006). Mismatch repair status in the prediction of benefit from adjuvant fluorouracil chemotherapy in colorectal cancer. Gut *55*, 848-855.

Ju, J., Kim, D.H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M.S., Shi, S., Wu, J.*, et al.* (2006). Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci U S A *103*, 19635-19640.

Jung, B., Doctolero, R.T., Tajima, A., Nguyen, A.K., Keku, T., Sandler, R.S., and Carethers, J.M. (2004). Loss of activin receptor type 2 protein expression in microsatellite unstable colon cancers. Gastroenterology *126*, 654-659.

Juschke, C., Dohnal, I., Pichler, P., Harzer, H., Swart, R., Ammerer, G., Mechtler, K., and Knoblich, J.A. (2013). Transcriptome and proteome quantification of a tumor model provides novel insights into post-transcriptional gene regulation. Genome Biol *14*, r133.

Kahi, C.J., Anderson, J.C., and Rex, D.K. (2013). Screening and surveillance for colorectal cancer: state of the art. Gastrointest Endosc *77*, 335-350.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A.*, et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339.

Karapetis, C.S., Khambata-Ford, S., Jonker, D.J., O'Callaghan, C.J., Tu, D., Tebbutt, N.C., Simes, R.J., Chalchal, H., Shapiro, J.D., Robitaille, S.*, et al.* (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med *359*, 1757-1765.

Karas, M., Bachmann, D., Bahr, U., and Hillenkamp, F. (1987). Matrix-assisted ultraviolet laser desorption of non-volatile compounds. International Journal of Mass Spectrometry and Ion Processes *78*, 53-68.

Karas, M., Bachmann, D., and Hillenkamp, F. (1985). Influence of the Wavelength in High-Irradiance Ultraviolet-Laser Desorption Mass-Spectrometry of Organic-Molecules. Analytical Chemistry *57*, 2935-2939.

Karpievitch, Y.V., Dabney, A.R., and Smith, R.D. (2012). Normalization and missing value imputation for label-free LC-MS analysis. BMC Bioinformatics *13 Suppl 16*, S5.

Karpinets, T.V., Greenwood, D.J., Sams, C.E., and Ammons, J.T. (2006). RNA:protein ratio of the unicellular organism as a characteristic of phosphorous and nitrogen stoichiometry and of the cellular requirement of ribosomes for protein synthesis. BMC Biol *4*, 30.

Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. Bioinformatics *25*, 415-416.

Kelly, H., and Goldberg, R.M. (2005). Systemic therapy for metastatic colorectal cancer: current options, current evidence. J Clin Oncol *23*, 4553-4560.

Kennedy, R.D., Bylesjo, M., Kerr, P., Davison, T., Black, J.M., Kay, E.W., Holt, R.J., Proutski, V., Ahdesmaki, M., Farztdinov, V.*, et al.* (2011). Development and independent validation of a prognostic assay for stage II colon cancer using formalin-fixed paraffin-embedded tissue. J Clin Oncol *29*, 4620-4626.

Ketchen, D.J., and Shook, C.L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. Strategic Management Journal *17*, 441-458.

Khambata-Ford, S., Garrett, C.R., Meropol, N.J., Basik, M., Harbison, C.T., Wu, S., Wong, T.W., Huang, X., Takimoto, C.H., Godwin, A.K.*, et al.* (2007). Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. J Clin Oncol *25*, 3230-3237.

Khorana, H.G., Gerber, G.E., Herlihy, W.C., Gray, C.P., Anderegg, R.J., Nihei, K., and Biemann, K. (1979). Amino acid sequence of bacteriorhodopsin. Proc Natl Acad Sci U S A *76*, 5046-5050.

Kikuchi-Yanoshita, R., Konishi, M., Ito, S., Seki, M., Tanaka, K., Maeda, Y., Iino, H., Fukayama, M., Koike, M., Mori, T.*, et al.* (1992). Genetic changes of both p53 alleles associated with the conversion from colorectal adenoma to early carcinoma in familial adenomatous polyposis and non-familial adenomatous polyposis patients. Cancer Res *52*, 3965-3971.

Kim, H., Golub, G.H., and Park, H. (2005a). Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics *21*, 187-198.

Kim, J.E., Tannenbaum, S.R., and White, F.M. (2005b). Global phosphoproteome of HT-29 human colon adenocarcinoma cells. J Proteome Res *4*, 1339-1346.

Kim, K.Y., Kim, S.H., Ki, D.H., Jeong, J., Jeong, H.J., Jeung, H.C., Chung, H.C., and Rha, S.Y. (2007). An attempt for combining microarray data sets by adjusting gene expressions. Cancer Res Treat *39*, 74-81.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S.*, et al.* (2014). A draft map of the human proteome. Nature *509*, 575-581.

Kim, S.H., Kwon, H.C., Oh, S.Y., Lee, D.M., Lee, S., Lee, J.H., Roh, M.S., Kim, D.C., Park, K.J., Choi, H.J.*, et al.* (2009). Prognostic value of ERCC1, thymidylate synthase, and glutathione S-transferase pi for 5-FU/oxaliplatin chemotherapy in advanced colorectal cancer. Am J Clin Oncol *32*, 38-43.

Kingdon, K.H. (1923). A Method for the Neutralization of Electron Space Charge by Positive Ionization at Very Low Gas Pressures. Physical Review *21*, 408-418.

Kinzler, K.W., and Vogelstein, B. (1996). Lessons from hereditary colorectal cancer. Cell *87*, 159-170.

Kirkland, S.C. (1985). Dome formation by a human colonic adenocarcinoma cell line (HCA-7). Cancer Research *45*, 3790-3795.

Kirkland, S.C. (1986). Endocrine differentiation by a human rectal adenocarcinoma cell line (HRA-19). Differentiation *33*, 148-155.

Kirkland, S.C., and Bailey, I.G. (1986). Establishment and characterisation of six human colorectal adenocarcinoma cell lines. British journal of cancer *53*, 779-785.

Kirklin, J.W., Dockerty, M.B., and Waugh, J.M. (1949). The role of the peritoneal reflection in the prognosis of carcinoma of the rectum and sigmoid colon. Surg Gynecol Obstet *88*, 326-331.

Kislinger, T., Cox, B., Kannan, A., Chung, C., Hu, P., Ignatchenko, A., Scott, M.S., Gramolini, A.O., Morris, Q., Hallett, M.T.*, et al.* (2006). Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. Cell *125*, 173-186.

Kitchen, R.R., Sabine, V.S., Sims, A.H., Macaskill, E.J., Renshaw, L., Thomas, J.S., van Hemert, J.I., Dixon, J.M., and Bartlett, J.M. (2010). Correcting for intra-experiment variation in Illumina BeadChip data is necessary to generate robust gene-expression profiles. BMC Genomics *11*, 134.

Klijn, C., Durinck, S., Stawiski, E.W., Haverty, P.M., Jiang, Z., Liu, H., Degenhardt, J., Mayba, O., Gnad, F., Liu, J.*, et al.* (2015). A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol *33*, 306-312.

Knapp, S., Arruda, P., Blagg, J., Burley, S., Drewry, D.H., Edwards, A., Fabbro, D., Gillespie, P., Gray, N.S., Kuster, B.*, et al.* (2013). A public-private partnership to unlock the untargeted kinome. Nature Chemical Biology *9*, 3-6.

Kogenaru, S., Qing, Y., Guo, Y., and Wang, N. (2012). RNA-seq and microarray complement each other in transcriptome profiling. BMC Genomics *13*, 629.

Kolch, W., Halasz, M., Granovskaya, M., and Kholodenko, B.N. (2015). The dynamic control of signal transduction networks in cancer cells. Nat Rev Cancer *15*, 515-527.

Kondrat, R.W., and Cooks, R.G. (1978). Direct analysis of mixtures by mass spectrometry. Analytical Chemistry.

Konecny, G.E., Pegram, M.D., Venkatesan, N., Finn, R., Yang, G., Rahmeh, M., Untch, M., Rusnak, D.W., Spehar, G., Mullin, R.J.*, et al.* (2006). Activity of the dual kinase

inhibitor lapatinib (GW572016) against HER-2-overexpressing and trastuzumab-treated breast cancer cells. Cancer Res *66*, 1630-1639.

Ku, J.L., Yoon, K.A., Kim, D.Y., and Park, J.G. (1999). Mutations in hMSH6 alone are not sufficient to cause the microsatellite instability in colorectal cancer cell lines. European Journal of Cancer *35*, 1724-1729.

Kuebler, J.P., Wieand, H.S., O'Connell, M.J., Smith, R.E., Colangelo, L.H., Yothers, G., Petrelli, N.J., Findlay, M.P., Seay, T.E., Atkins, J.N.*, et al.* (2007). Oxaliplatin combined with weekly bolus fluorouracil and leucovorin as surgical adjuvant chemotherapy for stage II and III colon cancer: results from NSABP C-07. J Clin Oncol *25*, 2198-2204.

Kuntz, K.M., Lansdorp-Vogelaar, I., Rutter, C.M., Knudsen, A.B., van Ballegooijen, M., Savarino, J.E., Feuer, E.J., and Zauber, A.G. (2011). A systematic comparison of microsimulation models of colorectal cancer: the role of assumptions about adenoma progression. Med Decis Making *31*, 530-539.

Kwon, T., Huse, H.K., Vogel, C., Whiteley, M., and Marcotte, E.M. (2014). Protein-to-mRNA ratios are conserved between Pseudomonas aeruginosa strains. J Proteome Res *13*, 2370-2380.

Labianca, R., Marsoni, S., Pancera, G., Torri, V., Zaniboni, A., Erlichman, C., Pater, J., Shepherd, L., Zee, B., Seitz, J.F.*, et al.* (1995). Efficacy of adjuvant fluorouracil and folinic acid in colon cancer. International Multicentre Pooled Analysis of Colon Cancer Trials (IMPACT) investigators. Lancet *345*, 939-944.

Laken, S.J., Petersen, G.M., Gruber, S.B., Oddoux, C., Ostrer, H., Giardiello, F.M., Hamilton, S.R., Hampel, H., Markowitz, A., Klimstra, D.*, et al.* (1997). Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. Nat Genet *17*, 79-83.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lansdorp-Vogelaar, I., Knudsen, A.B., and Brenner, H. (2010). Cost-effectiveness of colorectal cancer screening - an overview. Best practice & research Clinical gastroenterology *24*, 439-449.

Lansdorp-Vogelaar, I., van Ballegooijen, M., Zauber, A.G., Habbema, J.D., and Kuipers, E.J. (2009). Effect of rising chemotherapy costs on the cost savings of colorectal cancer screening. J Natl Cancer Inst *101*, 1412-1422.

Lawrence, R.T., Perez, E.M., Hernandez, D., Miller, C.P., Haas, K.M., Irie, H.Y., Lee, S.I., Blau, C.A., and Villen, J. (2015). The proteomic landscape of triple-negative breast cancer. Cell Rep *11*, 630-644.

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

Leach, F.S., Nicolaides, N.C., Papadopoulos, N., Liu, B., Jen, J., Parsons, R., Peltomaki, P., Sistonen, P., Aaltonen, L.A., Nystrom-Lahti, M., *et al.* (1993). Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. Cell *75*, 1215-1225.

Lecomte, T., Ferraz, J.M., Zinzindohoue, F., Loriot, M.A., Tregouet, D.A., Landi, B., Berger, A., Cugnenc, P.H., Jian, R., Beaune, P., *et al.* (2004). Thymidylate synthase gene polymorphism predicts toxicity in colorectal cancer patients receiving 5-fluorouracil-based chemotherapy. Clin Cancer Res *10*, 5880-5888.

Lee, K.S., and Park, E.C. (2016). Cost Effectiveness of Colorectal Cancer Screening Interventions with Their Effects on Health Disparity Being Considered. Cancer Res Treat *48*, 1010-1019.

Lee, M.V., Topper, S.E., Hubler, S.L., Hose, J., Wenger, C.D., Coon, J.J., and Gasch, A.P. (2011). A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. Mol Syst Biol *7*, 514.

Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet *3*, 1724-1735.

Leichman, L., Lenz, H.J., Leichman, C.G., Groshen, S., Danenberg, K., Baranda, J., Spears, C.P., Boswell, W., Silberman, H., Ortega, A., *et al.* (1995). Quantitation of intratumoral thymidylate synthase expression predicts for resistance to protracted infusion of 5-fluorouracil and weekly leucovorin in disseminated colorectal cancers: preliminary report from an ongoing trial. Eur J Cancer *31A*, 1306-1310.

Leroy, C., Fialin, C., Sirvent, A., Simon, V., Urbach, S., Poncet, J., Robert, B., Jouin, P., and Roche, S. (2009). Quantitative phosphoproteomics reveals a cluster of tyrosine kinases that mediates SRC invasive activity in advanced colon carcinoma cells. Cancer Res *69*, 2279-2286.

Levin, B., Lieberman, D.A., McFarland, B., Smith, R.A., Brooks, D., Andrews, K.S., Dash, C., Giardiello, F.M., Glick, S., Levin, T.R., *et al.* (2008). Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. In CA: a cancer journal for clinicians, pp. 130-160.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, C., and Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A *98*, 31-36.

Li, G., Waltham, M., Anderson, N.L., Unsworth, E., Treston, A., and Weinstein, J.N. (1997). Rapid mass spectrometric identification of proteins from two-dimensional polyacrylamide gels after in gel proteolytic digestion. Electrophoresis *18*, 391-402.

Li, J.J., Bickel, P.J., and Biggin, M.D. (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ *2*, e270.

Li, Y.H., Wang, F., Shen, L., Deng, Y.M., Shao, Q., Feng, F., An, X., Wang, F.H., Wang, Z.Q., Xu, R.H.*, et al.* (2011). EGFR fluorescence in situ hybridization pattern of chromosome 7 disomy predicts resistance to cetuximab in KRAS wild-type metastatic colorectal cancer patients. Clin Cancer Res *17*, 382-390.

Liang, J.T., Huang, K.C., Cheng, Y.M., Hsu, H.C., Cheng, A.L., Hsu, C.H., Yeh, K.H., Wang, S.M., and Chang, K.J. (2002a). P53 overexpression predicts poor chemosensitivity to high-dose 5-fluorouracil plus leucovorin chemotherapy for stage IV colorectal cancers after palliative bowel resection. Int J Cancer *97*, 451-457.

Liang, J.T., Huang, K.C., Lai, H.S., Lee, P.H., Cheng, Y.M., Hsu, H.C., Cheng, A.L., Hsu, C.H., Yeh, K.H., Wang, S.M.*, et al.* (2002b). High-frequency microsatellite instability predicts better chemosensitivity to high-dose 5-fluorouracil plus leucovorin chemotherapy for stage IV sporadic colorectal cancer after palliative bowel resection. Int J Cancer *101*, 519-525.

Licitra, L., Storkel, S., Kerr, K.M., Van Cutsem, E., Pirker, R., Hirsch, F.R., Vermorken, J.B., von Heydebreck, A., Esser, R., Celik, I.*, et al.* (2013). Predictive value of epidermal growth factor receptor expression for first-line chemotherapy plus cetuximab in patients with head and neck and colorectal cancer: analysis of data from the EXTREME and CRYSTAL studies. Eur J Cancer *49*, 1161-1168.

Lievre, A., Bachet, J.B., Le Corre, D., Boige, V., Landi, B., Emile, J.F., Cote, J.F., Tomasic, G., Penna, C., Ducreux, M.*, et al.* (2006). KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. Cancer Res *66*, 3992-3995.

Lin, D., Zhang, J., Li, J., Xu, C., Deng, H.W., and Wang, Y.P. (2016). An integrative imputation method based on multi-omics datasets. BMC Bioinformatics *17*, 247.

Lin, J.K., Lin, A.J., Lin, C.C., Lan, Y.T., Yang, S.H., Li, A.F., and Chang, S.C. (2011). The status of EGFR-associated genes could predict the outcome and tumor response of chemo-refractory metastatic colorectal patients using cetuximab and chemotherapy. J Surg Oncol *104*, 661-666.

Lin, W.M., Baker, A.C., Beroukhim, R., Winckler, W., Feng, W., Marmion, J.M., Laine, E., Greulich, H., Tseng, H., Gates, C.*, et al.* (2008). Modeling genomic diversity and tumor dependency in malignant melanoma. Cancer Res *68*, 664-673.

Linnekamp, J.F., Wang, X., Medema, J.P., and Vermeulen, L. (2015). Colorectal cancer heterogeneity and targeted therapy: a case for molecular disease subtypes. Cancer Res *75*, 245-249.

Liu, H., Sadygov, R.G., and Yates, J.R., 3rd (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem *76*, 4193-4201.

Liu, Q., and Zhang, B. (2016). Integrative Omics Analysis Reveals Post-Transcriptionally Enhanced Protective Host Response in Colorectal Cancers with Microsatellite Instability. J Proteome Res *15*, 766-776.

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell *165*, 535-550.

Liu, Y., and Bodmer, W.F. (2006). Analysis of P53 mutations and their expression in 56 colorectal cancer cell lines. PNAS *103*, 976-981.

Loboda, A., Nebozhyn, M.V., Watters, J.W., Buser, C.A., Shaw, P.M., Huang, P.S., Van't Veer, L., Tollenaar, R.A., Jackson, D.B., Agrawal, D.*, et al.* (2011). EMT is the dominant program in human colon cancer. BMC Med Genomics *4*, 9.

Lofton-Day, C., Model, F., Devos, T., Tetzner, R., Distler, J., Schuster, M., Song, X., Lesche, R., Liebenberg, V., Ebert, M.*, et al.* (2008). DNA methylation biomarkers for blood-based colorectal cancer screening. Clin Chem *54*, 414-423.

Loupakis, F., Ruzzo, A., Cremolini, C., Vincenzi, B., Salvatore, L., Santini, D., Masi, G., Stasi, I., Canestrari, E., Rulli, E.*, et al.* (2009). KRAS codon 61, 146 and BRAF mutations predict resistance to cetuximab plus irinotecan in KRAS codon 12 and 13 wild-type metastatic colorectal cancer. Br J Cancer *101*, 715-721.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

Lueking, A., Horn, M., Eickhoff, H., Bussow, K., Lehrach, H., and Walter, G. (1999). Protein microarrays for gene expression and antibody screening. Anal Biochem *270*, 103-111.

Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Algenas, C., Lundeberg, J., Mann, M., and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. Mol Syst Biol *6*, 450.

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., Shi, T., Tong, W., Shi, L., Hong, H.*, et al.* (2010). A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics J *10*, 278-291.

Luque-Garcia, J.L., Martinez-Torrecuadrada, J.L., Epifano, C., Canamero, M., Babel, I., and Casal, J.I. (2010). Differential protein expression on the cell surface of colorectal cancer cells associated to tumor metastasis. Proteomics *10*, 940-952.

Lv, X., Zhang, L., Zhu, Y., Said, H.M., Shi, J., and Xu, G. (2015). Regulative Effect of Nampt on Tumor Progression and Cell Viability in Human Colorectal Cancer. J Cancer *6*, 849-858.

Ly, T., Ahmad, Y., Shlien, A., Soroka, D., Mills, A., Emanuele, M.J., Stratton, M.R., and Lamond, A.I. (2014). A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. Elife *3*, e01630.

Lyratzopoulos, G., Abel, G.A., Brown, C.H., Rous, B.A., Vernon, S.A., Roland, M., and Greenberg, D.C. (2013). Socio-demographic inequalities in stage of cancer diagnosis: evidence from patients with female breast, lung, colon, rectal, prostate, renal, bladder, melanoma, ovarian and endometrial cancer. Ann Oncol *24*, 843-850.

Ma, B., Zhang, K.Z., and Liang, C.Z. (2005). An effective algorithm for peptide de novo sequencing from MS/MS spectra. Journal of Computer and System Sciences *70*, 418-430.

Maak, M., Simon, I., Nitsche, U., Roepman, P., Snel, M., Glas, A.M., Schuster, T., Keller, G., Zeestraten, E., Goossens, I.*, et al.* (2013). Independent validation of a prognostic genomic signature (ColoPrint) for patients with stage II colon cancer. Ann Surg *257*, 1053-1058.

MacBeath, G., Koehler, A.N., and Schreiber, S.L. (1999). Printing small molecules as microarrays and detecting protein-ligand interactions en masse. Journal of the American Chemical Society *121*, 7967-7968.

MacBeath, G., and Schreiber, S.L. (2000). Printing proteins as microarrays for high-throughput function determination. Science *289*, 1760-1763.

Machover, D., Schwarzenberg, L., Goldschmidt, E., Tourani, J.M., Michalski, B., Hayat, M., Dorval, T., Misset, J.L., Jasmin, C., Maral, R.*, et al.* (1982). Treatment of advanced colorectal and gastric adenocarcinomas with 5-FU combined with high-dose folinic acid: a pilot study. Cancer Treat Rep *66*, 1803-1807.

Mack, L.L., Kralik, P., Rheude, A., and Dole, M. (1970). Molecular Beams of Macroions .2. Journal of Chemical Physics *52*, 4977-&.

Madhavan, S., Gusev, Y., Natarajan, T.G., Song, L., Bhuvaneshwar, K., Gauba, R., Pandey, A., Haddad, B.R., Goerlitz, D., Cheema, A.K.*, et al.* (2013). Genome-wide multi-omics profiling of colorectal cancer identifies immune determinants strongly associated with relapse. Front Genet *4*, 236.

Mahoney, W.C., and Hermodson, M.A. (1980). Separation of Large Denatured Peptides by Reverse Phase High-Performance Liquid-Chromatography - Trifluoroacetic-Acid as a Peptide Solvent. Journal of Biological Chemistry *255*, 1199-1203.

Makarov, A. (2000). Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal Chem *72*, 1156-1162.

Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K., and Horning, S. (2006). Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. Anal Chem *78*, 2113-2120.

Malapelle, U., Carlomagno, C., Salatiello, M., De Stefano, A., De Luca, C., Bianco, R., Marciano, R., Cimminiello, C., Bellevicine, C., De Placido, S.*, et al.* (2012). KRAS mutation detection by high-resolution melting analysis significantly predicts

clinical benefit of cetuximab in metastatic colorectal cancer. Br J Cancer *107*, 626-631.

Malone, J.H., and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol *9*, 34.

Mann, M., Hojrup, P., and Roepstorff, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol Mass Spectrom *22*, 338-345.

Mann, M., and Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem *66*, 4390-4399.

Manning, B.D. (2009). Challenges and opportunities in defining the essential cancer kinome. Sci Signal *2*, pe15.

Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science *298*, 1912-1934.

MAQC Consortium, Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., de Longueville, F., Kawasaki, E.S.*, et al.* (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol *24*, 1151-1161.

Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., and Bahler, J. (2012). Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. Cell *151*, 671-683.

Mariadason, J.M., Arango, D., Shi, Q., Wilson, A.J., Corner, G.A., Nicholas, C., Aranes, M.J., Lesser, M., Schwartz, E.L., and Augenlicht, L.H. (2003). Gene expression profiling-based prediction of response of colon carcinoma cells to 5-fluorouracil and camptothecin. Cancer Res *63*, 8791-8812.

Marin-Vicente, C., Lyutvinskiy, Y., Romans Fuertes, P., Zubarev, R.A., and Visa, N. (2013). The effects of 5-fluorouracil on the proteome of colon cancer cells. J Proteome Res *12*, 1969-1979.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome research *18*, 1509-1517.

Marisa, L., de Reynies, A., Duval, A., Selves, J., Gaub, M.P., Vescovo, L., Etienne-Grimaldi, M.C., Schiappa, R., Guenot, D., Ayadi, M.*, et al.* (2013). Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med *10*, e1001453.

Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L., Lutterbaugh, J., Fan, R.S., Zborowska, E., Kinzler, K.W., Vogelstein, B.*, et al.* (1995). Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. Science *268*, 1336-1338.

Marks, J.L., Gong, Y., Chitale, D., Golas, B., McLellan, M.D., Kasai, Y., Ding, L., Mardis, E.R., Wilson, R.K., Solit, D.*, et al.* (2008). Novel MEK1 mutation identified by mutational analysis of epidermal growth factor receptor signaling pathway genes in lung adenocarcinoma. Cancer Res *68*, 5524-5528.

Matar, P., Rojo, F., Cassia, R., Moreno-Bueno, G., Di Cosimo, S., Tabernero, J., Guzman, M., Rodriguez, S., Arribas, J., Palacios, J.*, et al.* (2004). Combined epidermal growth factor receptor targeting with the tyrosine kinase inhibitor gefitinib (ZD1839) and the monoclonal antibody cetuximab (IMC-C225): superiority over single-agent receptor targeting. Clin Cancer Res *10*, 6487-6501.

Mathivanan, S., Ji, H., Tauro, B.J., Chen, Y.S., and Simpson, R.J. (2012). Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry. Journal of Proteomics *76*, 141-149.

Mayer, R.J., Van Cutsem, E., Falcone, A., Yoshino, T., Garcia-Carbonero, R., Mizunuma, N., Yamazaki, K., Shimada, Y., Tabernero, J., Komatsu, Y.*, et al.* (2015). Randomized trial of TAS-102 for refractory metastatic colorectal cancer. N Engl J Med *372*, 1909-1919.

McAlister, G.C., Nusinow, D.P., Jedrychowski, M.P., Wuhr, M., Huttlin, E.L., Erickson, B.K., Rad, R., Haas, W., and Gygi, S.P. (2014). MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. Anal Chem *86*, 7150-7158.

McBain, J.A., Weese, J.L., Meisner, L.F., Wolberg, W.H., and Willson, J.K. (1984). Establishment and characterization of human colorectal cancer cell lines. Cancer Research *44*, 5813-5821.

McCall, M.N., Bolstad, B.M., and Irizarry, R.A. (2010). Frozen robust multiarray analysis (fRMA). Biostatistics *11*, 242-253.

McCall, M.N., Uppal, K., Jaffee, H.A., Zilliox, M.J., and Irizarry, R.A. (2011). The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. Nucleic Acids Res *39*, D1011-1015.

McCubrey, J.A., Steelman, L.S., Chappell, W.H., Abrams, S.L., Franklin, R.A., Montalto, G., Cervello, M., Libra, M., Candido, S., Malaponte, G.*, et al.* (2012). Ras/Raf/MEK/ERK and PI3K/PTEN/Akt/mTOR cascade inhibitors: how mutations can result in therapy resistance and how to overcome resistance. Oncotarget *3*, 1068-1111.

McDermott, U., Sharma, S.V., Dowell, L., Greninger, P., Montagut, C., Lamb, J., Archibald, H., Raudales, R., Tam, A., Lee, D.*, et al.* (2007). Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. Proc Natl Acad Sci U S A *104*, 19936-19941.

Medard, G., Pachl, F., Ruprecht, B., Klaeger, S., Heinzlmeir, S., Helm, D., Qiao, H., Ku, X., Wilhelm, M., Kuehne, T.*, et al.* (2015). Optimized chemical proteomics assay for kinase inhibitor profiling. J Proteome Res *14*, 1574-1586.

Medico, E., Russo, M., Picco, G., Cancelliere, C., Valtorta, E., Corti, G., Buscarino, M., Isella, C., Lamba, S., Martinoglio, B.*, et al.* (2015). The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. Nat Commun *6*, 7002.

Meleth, S., Deshane, J., and Kim, H. (2005). The case for well-conducted experiments to validate statistical protocols for 2D gels: different pre-processing = different lists of significant proteins. BMC Biotechnol *5*, 7.

Meng, C., Helm, D., Frejno, M., and Kuster, B. (2016). moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. J Proteome Res *15*, 755-765.

Michalski, A., Damoc, E., Lange, O., Denisov, E., Nolting, D., Muller, M., Viner, R., Schwartz, J., Remes, P., Belford, M.*, et al.* (2012a). Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. Mol Cell Proteomics *11*, O111 013698.

Michalski, A., Neuhauser, N., Cox, J., and Mann, M. (2012b). A systematic investigation into the nature of tryptic HCD spectra. J Proteome Res *11*, 5479-5491.

Mikesh, L.M., Ueberheide, B., Chi, A., Coon, J.J., Syka, J.E., Shabanowitz, J., and Hunt, D.F. (2006). The utility of ETD mass spectrometry in proteomic analysis. Biochim Biophys Acta *1764*, 1811-1822.

Misale, S., Yaeger, R., Hobor, S., Scala, E., Janakiraman, M., Liska, D., Valtorta, E., Schiavo, R., Buscarino, M., Siravegna, G.*, et al.* (2012). Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. Nature *486*, 532-536.

Model, F., Osborn, N., Ahlquist, D., Gruetzmann, R., Molnar, B., Sipos, F., Galamb, O., Pilarsky, C., Saeger, H.D., Tulassay, Z.*, et al.* (2007). Identification and validation of colorectal neoplasia-specific methylation markers for accurate classification of disease. Mol Cancer Res *5*, 153-163.

Monks, A., Scudiero, D., Skehan, P., Shoemaker, R., Paull, K., Vistica, D., Hose, C., Langley, J., Cronise, P., Vaigrowolff, A.*, et al.* (1991). Feasibility of a High-Flux Anticancer Drug Screen Using a Diverse Panel of Cultured Human Tumor-Cell Lines. Journal of the National Cancer Institute *83*, 757-766.

Monroe, M. (2015). Protein Digestion Simulator.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn *52*, 91-118.

Moroni, M., Veronese, S., Benvenuti, S., Marrapese, G., Sartore-Bianchi, A., Di Nicolantonio, F., Gambacorta, M., Siena, S., and Bardelli, A. (2005). Gene copy number for epidermal growth factor receptor (EGFR) and clinical response to

antiEGFR treatment in colorectal cancer: a cohort study. The Lancet Oncology *6*, 279-286.

Morris, H.R., Panico, M., Barber, M., Bordoli, R.S., Sedgwick, R.D., and Tyler, A. (1981). Fast atom bombardment: a new mass spectrometric method for peptide sequence analysis. Biochem Biophys Res Commun *101*, 623-631.

Morson, B.C., Whiteway, J.E., Jones, E.A., Macrae, F.A., and Williams, C.B. (1984). Histopathology and prognosis of malignant colorectal polyps treated by endoscopic polypectomy. Gut *25*, 437-444.

Mortality, G.B.D., and Causes of Death, C. (2015). Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet *385*, 117-171.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*, 621-628.

Mouradov, D., Sloggett, C., Jorissen, R.N., Love, C.G., Li, S., Burgess, A.W., Arango, D., Strausberg, R.L., Buchanan, D., Wormald, S.*, et al.* (2014). Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. Cancer Res *74*, 3238-3247.

Muller, C.I., Schulmann, K., Reinacher-Schick, A., Andre, N., Arnold, D., Tannapfel, A., Arkenau, H., Hahn, S.A., Schmoll, S.H., Porschen, R.*, et al.* (2008). Predictive and prognostic value of microsatellite instability in patients with advanced colorectal cancer treated with a fluoropyrimidine and oxaliplatin containing first-line chemotherapy. A report of the AIO Colorectal Study Group. Int J Colorectal Dis *23*, 1033-1039.

Murtagh, F., and Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification *31*, 274-295.

Myers, T.G., Anderson, N.L., Waltham, M., Li, G., Buolamwini, J.K., Scudiero, D.A., Paull, K.D., Sausville, E.A., and Weinstein, J.N. (1997). A protein expression database for the molecular pharmacology of cancer. Electrophoresis *18*, 647-653.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol *7*, 548.

Nagase, H., Miyoshi, Y., Horii, A., Aoki, T., Ogawa, M., Utsunomiya, J., Baba, S., Sasazuki, T., and Nakamura, Y. (1992). Correlation between the location of germ-line mutations in the APC gene and the number of colorectal polyps in familial adenomatous polyposis patients. Cancer Res *52*, 4055-4057.

National Comprehensive Cancer Network (2016). NCCN Clinical Practice Guidelines in Oncology: Colon Cancer (Version 2.2016).

Nesvizhskii, A.I., and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics *4*, 1419-1440.

Nesvizhskii, A.I., Vitek, O., and Aebersold, R. (2007). Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods *4*, 787-797.

Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.P., Tong, F.*, et al.* (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell *10*, 515-527.

Nguyen, S., and Fenn, J.B. (2007). Gas-phase ions of solute species from charged droplets of solutions. Proc Natl Acad Sci U S A *104*, 1111-1117.

Nibbe, R.K., Koyuturk, M., and Chance, M.R. (2010). An integrative -omics approach to identify functional sub-networks in human colorectal cancer. PLoS Comput Biol *6*, e1000639.

Nielsen, T.O., Parker, J.S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S.R., Snider, J., Stijleman, I.J., Reed, J.*, et al.* (2010). A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. Clin Cancer Res *16*, 5222-5232.

Ning, K., Fermin, D., and Nesvizhskii, A.I. (2012). Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. J Proteome Res *11*, 2261-2271.

Nishizuka, S., Charboneau, L., Young, L., Major, S., Reinhold, W.C., Waltham, M., Kouros-Mehr, H., Bussey, K.J., Lee, J.K., Espina, V.*, et al.* (2003a). Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. Proc Natl Acad Sci U S A *100*, 14229-14234.

Nishizuka, S., Chen, S.T., Gwadry, F.G., Alexander, J., Major, S.M., Scherf, U., Reinhold, W.C., Waltham, M., Charboneau, L., Young, L.*, et al.* (2003b). Diagnostic markers that distinguish colon and ovarian adenocarcinomas: identification by genomic, proteomic, and tissue array profiling. Cancer Res *63*, 5243-5250.

O'Connell, M.J., Lavery, I., Yothers, G., Paik, S., Clark-Langone, K.M., Lopatin, M., Watson, D., Baehner, F.L., Shak, S., Baker, J.*, et al.* (2010). Relationship between tumor gene expression and recurrence in four independent studies of patients with stage II/III colon cancer treated with surgery alone or surgery plus adjuvant fluorouracil plus leucovorin. J Clin Oncol *28*, 3937-3944.

Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K., and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. Bioinformatics *19*, 2088-2096.

Olsen, J.V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007). Higher-energy C-trap dissociation for peptide modification analysis. Nat Methods *4*, 709-712.

Olsen, J.V., Nielsen, M.L., Damoc, N.E., Griep-Raming, J., Moehring, T., Makarov, A., Schwartz, J., Horning, S., and Mann, M. (2009a). Characterization of the Velos, an Enhanced LTQ Orbitrap, for Proteomics. Molecular & Cellular Proteomics *8*, S40-S40.

Olsen, J.V., Schwartz, J.C., Griep-Raming, J., Nielsen, M.L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M*., et al.* (2009b). A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. Mol Cell Proteomics *8*, 2759-2769.

Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics *1*, 376-386.

Oppermann, F.S., Gnad, F., Olsen, J.V., Hornberger, R., Greff, Z., Keri, G., Mann, M., and Daub, H. (2009). Large-scale proteomics analysis of the human kinome. Mol Cell Proteomics *8*, 1751-1764.

Organ, S.L., Tong, J., Taylor, P., St-Germain, J.R., Navab, R., Moran, M.F., and Tsao, M.S. (2011). Quantitative phospho-proteomic profiling of hepatocyte growth factor (HGF)-MET signaling in colorectal cancer. J Proteome Res *10*, 3200-3211.

Ota, D.M., Nelson, H., and Weeks, J.C. (1994). Controversies regarding laparoscopic colectomy for malignant diseases. Curr Opin Gen Surg, 208-213.

Pachl, F., Plattner, P., Ruprecht, B., Medard, G., Sewald, N., and Kuster, B. (2013a). Characterization of a chemical affinity probe targeting Akt kinases. J Proteome Res *12*, 3792-3800.

Pachl, F., Ruprecht, B., Lemeer, S., and Kuster, B. (2013b). Characterization of a high field Orbitrap mass spectrometer for proteome analysis. Proteomics *13*, 2552-2562.

Pagliarini, R., Shao, W., and Sellers, W.R. (2015). Oncogene addiction: pathways of therapeutic response, resistance, and road maps toward a cure. EMBO Rep *16*, 280-296.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T*., et al.* (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med *351*, 2817-2826.

Pappin, D.J., Hojrup, P., and Bleasby, A.J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. Curr Biol *3*, 327-332.

Paradiso, A., Ranieri, G., Simone, G., Silvestris, N., Costa, A., De Lena, M., Leone, A., Vallejo, C., and Lacava, J. (2002). mdm2-p53 Interaction: lack of correlation with the response to 5-fluorouracil in advanced colorectal cancer. Oncology *62*, 278-285.

Park, E.S., Rabinovsky, R., Carey, M., Hennessy, B.T., Agarwal, R., Liu, W., Ju, Z., Deng, W., Lu, Y., Woo, H.G.*, et al.* (2010). Integrative analysis of proteomic signatures, mutations, and drug responsiveness in the NCI 60 cancer cell line set. Mol Cancer Ther *9*, 257-267.

Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z.*, et al.* (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol *27*, 1160-1167.

Patel, S.S., and Kilgore, M.L. (2015). Cost Effectiveness of Colorectal Cancer Screening Strategies. Cancer Control *22*, 248-258.

Patricelli, M.P., Nomanbhoy, T.K., Wu, J., Brown, H., Zhou, D., Zhang, J., Jagannathan, S., Aban, A., Okerberg, E., Herring, C.*, et al.* (2011). In situ kinase profiling reveals functionally relevant properties of native kinases. Chemistry & biology *18*, 699-710.

Pavel, A.B., Sonkin, D., and Reddy, A. (2016). Integrative modeling of multi-omics data to identify cancer drivers and infer patient-specific gene activity. BMC Syst Biol *10*, 16.

Paweletz, C.P., Charboneau, L., Bichsel, V.E., Simone, N.L., Chen, T., Gillespie, J.W., Emmert-Buck, M.R., Roth, M.J., Petricoin, I.E., and Liotta, L.A. (2001). Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. Oncogene *20*, 1981-1989.

Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P., and Fodor, S.P. (1994). Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci U S A *91*, 5022-5026.

Peeters, M., Douillard, J.Y., Van Cutsem, E., Siena, S., Zhang, K., Williams, R., and Wiezorek, J. (2013). Mutant KRAS codon 12 and 13 alleles in patients with metastatic colorectal cancer: assessment as prognostic and predictive biomarkers of response to panitumumab. J Clin Oncol *31*, 759-765.

Peeters, M., Price, T.J., Cervantes, A., Sobrero, A.F., Ducreux, M., Hotko, Y., Andre, T., Chan, E., Lordick, F., Punt, C.J.*, et al.* (2010). Randomized phase III study of panitumumab with fluorouracil, leucovorin, and irinotecan (FOLFIRI) compared with FOLFIRI alone as second-line treatment in patients with metastatic colorectal cancer. J Clin Oncol *28*, 4706-4713.

Peltomaki, P., Lothe, R.A., Aaltonen, L.A., Pylkkanen, L., Nystrom-Lahti, M., Seruca, R., David, L., Holm, R., Ryberg, D., Haugen, A.*, et al.* (1993). Microsatellite instability is associated with tumors that characterize the hereditary non-polyposis colorectal carcinoma syndrome. Cancer Res *53*, 5853-5855.

Peng, M., Taouatas, N., Cappadona, S., van Breukelen, B., Mohammed, S., Scholten, A., and Heck, A.J. (2012). Protease bias in absolute protein quantitation. Nat Methods *9*, 524-525.

Perez-Villamil, B., Romera-Lopez, A., Hernandez-Prieto, S., Lopez-Campos, G., Calles, A., Lopez-Asenjo, J.A., Sanz-Ortega, J., Fernandez-Perez, C., Sastre, J., Alfonso, R.*, et al.* (2012). Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior. BMC Cancer *12*, 260.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis *20*, 3551-3567.

Personeni, N., Fieuws, S., Piessevaux, H., De Hertogh, G., De Schutter, J., Biesmans, B., De Roock, W., Capoen, A., Debiec-Rychter, M., Van Laethem, J.L.*, et al.* (2008). Clinical usefulness of EGFR gene copy number as a predictive marker in colorectal cancer patients treated with cetuximab: a fluorescent in situ hybridization study. Clin Cancer Res *14*, 5869-5876.

Piersma, S.R., Knol, J.C., de Reus, I., Labots, M., Sampadi, B.K., Pham, T.V., Ishihama, Y., Verheul, H.M., and Jimenez, C.R. (2015). Feasibility of label-free phosphoproteomics and application to base-line signaling of colorectal cancer cell lines. J Proteomics *127*, 247-258.

Ponnala, L., Wang, Y., Sun, Q., and van Wijk, K.J. (2014). Correlation of mRNA and protein abundance in the developing maize leaf. Plant J *78*, 424-440.

Poon, M.A., O'Connell, M.J., Moertel, C.G., Wieand, H.S., Cullinan, S.A., Everson, L.K., Krook, J.E., Mailliard, J.A., Laurie, J.A., Tschetter, L.K.*, et al.* (1989). Biochemical modulation of fluorouracil: evidence of significant improvement of survival and quality of life in patients with advanced colorectal carcinoma. J Clin Oncol *7*, 1407-1418.

Putney, S.D., Royal, N.J., Neuman de Vegvar, H., Herlihy, W.C., Biemann, K., and Schimmel, P. (1981). Primary structure of a large aminoacyl-tRNA synthetase. Science *213*, 1497-1501.

Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc *2*, 1896-1906.

Rayleigh, L. (1882). XX. On the equilibrium of liquid conducting masses charged with electricity. Philosophical Magazine Series 5 *14*, 184-186.

Rees, M.G., Seashore-Ludlow, B., Cheah, J.H., Adams, D.J., Price, E.V., Gill, S., Javaid, S., Coletti, M.E., Jones, V.L., Bodycombe, N.E.*, et al.* (2016). Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol *12*, 109-116.

Ribic, C.M., Sargent, D.J., Moore, M.J., Thibodeau, S.N., French, A.J., Goldberg, R.M., Hamilton, S.R., Laurent-Puig, P., Gryfe, R., Shepherd, L.E.*, et al.* (2003). Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. N Engl J Med *349*, 247-257.

Rikova, K., Guo, A., Zeng, Q., Possemato, A., Yu, J., Haack, H., Nardone, J., Lee, K., Reeves, C., Li, Y.*, et al.* (2007). Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell *131*, 1190-1203.

Ritchie, M.E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., and Smyth, G.K. (2007). A comparison of background correction methods for two-colour microarrays. Bioinformatics *23*, 2700-2707.

Ritorto, M.S., Cook, K., Tyagi, K., Pedrioli, P.G., and Trost, M. (2013). Hydrophilic strong anion exchange (hSAX) chromatography for highly orthogonal peptide separation of complex proteomes. J Proteome Res *12*, 2449-2457.

Ritz, C., Baty, F., Streibig, J.C., and Gerhard, D. (2015). Dose-Response Analysis Using R. PLoS One *10*, e0146021.

Rivals, I., Personnaz, L., Taing, L., and Potier, M.C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics *23*, 401-407.

Rodriguez, A., and Laio, A. (2014). Machine learning. Clustering by fast search and find of density peaks. Science *344*, 1492-1496.

Roepman, P., Schlicker, A., Tabernero, J., Majewski, I., Tian, S., Moreno, V., Snel, M.H., Chresta, C.M., Rosenberg, R., Nitsche, U.*, et al.* (2014). Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. Int J Cancer *134*, 552-562.

Roepstorff, P., and Fohlman, J. (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed Mass Spectrom *11*, 601.

Rosenfeld, J., Capdevielle, J., Guillemot, J.C., and Ferrara, P. (1992). In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. Anal Biochem *203*, 173-179.

Rousseeuw, P.J. (1987). Silhouettes - a Graphical Aid to the Interpretation and Validation of Cluster-Analysis. Journal of Computational and Applied Mathematics *20*, 53-65.

Rudy, J., and Valafar, F. (2011). Empirical comparison of cross-platform normalization methods for gene expression data. BMC Bioinformatics *12*, 467.

Ruprecht, B., Zecha, J., Heinzlmeir, S., Medard, G., Lemeer, S., and Kuster, B. (2015). Evaluation of Kinase Activity Profiling Using Chemical Proteomics. ACS Chem Biol *10*, 2743-2752.

Sadanandam, A., Lyssiotis, C.A., Homicsko, K., Collisson, E.A., Gibb, W.J., Wullschleger, S., Ostos, L.C., Lannon, W.A., Grotzinger, C., Del Rio, M.*, et al.* (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat Med *19*, 619-625.

Salazar, R., Roepman, P., Capella, G., Moreno, V., Simon, I., Dreezen, C., Lopez-Doriga, A., Santos, C., Marijnen, C., Westerga, J.*, et al.* (2011). Gene expression

signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol *29*, 17-24.

Sanchez-Martin, F.J., Bellosillo, B., Gelabert-Baldrich, M., Dalmases, A., Canadas, I., Vidal, J., Martinez, A., Argiles, G., Siravegna, G., Arena, S.*, et al.* (2016). The First-in-class Anti-EGFR Antibody Mixture Sym004 Overcomes Cetuximab Resistance Mediated by EGFR Extracellular Domain Mutations in Colorectal Cancer. Clin Cancer Res *22*, 3260-3267.

Sankaranarayanan, R., Swaminathan, R., Brenner, H., Chen, K., Chia, K.S., Chen, J.G., Law, S.C., Ahn, Y.O., Xiang, Y.B., Yeole, B.B.*, et al.* (2010). Cancer survival in Africa, Asia, and Central America: a population-based study. The Lancet Oncology *11*, 165-173.

Sargent, D.J., Marsoni, S., Monges, G., Thibodeau, S.N., Labianca, R., Hamilton, S.R., French, A.J., Kabat, B., Foster, N.R., Torri, V.*, et al.* (2010). Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. J Clin Oncol *28*, 3219-3226.

Savaryn, J.P., Toby, T.K., and Kelleher, N.L. (2016). A researcher's guide to mass spectrometry-based proteomics. Proteomics *16*, 2435-2443.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-470.

Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T.*, et al.* (2000). A gene expression database for the molecular pharmacology of cancer. Nat Genet *24*, 236-244.

Schlegel, J., Sambade, M.J., Sather, S., Moschos, S.J., Tan, A.C., Winges, A., DeRyckere, D., Carson, C.C., Trembath, D.G., Tentler, J.J.*, et al.* (2013). MERTK receptor tyrosine kinase is a therapeutic target in melanoma. J Clin Invest *123*, 2257-2267.

Schlicker, A., Beran, G., Chresta, C.M., McWalter, G., Pritchard, A., Weston, S., Runswick, S., Davenport, S., Heathcote, K., Castro, D.A.*, et al.* (2012). Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. BMC Med Genomics *5*, 66.

Schlicker, A., Domingues, F.S., Rahnenfuhrer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics *7*, 302.

Schmidberger, M., Vicedo, E., and Mansmann, U. (2013). affyPara: Parallelized preprocessing methods for Affymetrix Oligonucleotide Arrays.

Schmoll, H.J., Cartwright, T., Tabernero, J., Nowacki, M.P., Figer, A., Maroun, J., Price, T., Lim, R., Van Cutsem, E., Park, Y.S.*, et al.* (2007). Phase III trial of capecitabine plus oxaliplatin as adjuvant therapy for stage III colon cancer: a planned safety analysis in 1,864 patients. J Clin Oncol *25*, 102-109.

Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate *14*, 853-871.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337-342.

Schwartz, J.C., Senko, M.W., and Syka, J.E. (2002). A two-dimensional quadrupole ion trap mass spectrometer. J Am Soc Mass Spectrom *13*, 659-669.

Seashore-Ludlow, B., Rees, M.G., Cheah, J.H., Cokol, M., Price, E.V., Coletti, M.E., Jones, V., Bodycombe, N.E., Soule, C.K., Gould, J.*, et al.* (2015). Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov *5*, 1210-1223.

Senbabaoglu, Y., Michailidis, G., and Li, J.Z. (2014). Critical limitations of consensus clustering in class discovery. Sci Rep *4*, 6207.

SEQC MAQC-III Consortium (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol *32*, 903-914.

Shabalin, A.A., Tjelmeland, H., Fan, C., Perou, C.M., and Nobel, A.B. (2008). Merging two gene-expression studies via cross-platform normalization. Bioinformatics *24*, 1154-1160.

Shankavaram, U.T., Reinhold, W.C., Nishizuka, S., Major, S., Morita, D., Chary, K.K., Reimers, M.A., Scherf, U., Kahn, A., Dolginow, D.*, et al.* (2007). Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. Mol Cancer Ther *6*, 820-832.

Sharma, K., Schmitt, S., Bergner, C.G., Tyanova, S., Kannaiyan, N., Manrique-Hoyos, N., Kongi, K., Cantuti, L., Hanisch, U.K., Philips, M.A.*, et al.* (2015). Cell type- and brain region-resolved mouse brain proteome. Nat Neurosci *18*, 1819-1831.

Shaukat, A., Mongin, S.J., Geisser, M.S., Lederle, F.A., Bond, J.H., Mandel, J.S., and Church, T.R. (2013). Long-term mortality after screening for colorectal cancer. N Engl J Med *369*, 1106-1114.

Shi, W., Oshlack, A., and Smyth, G.K. (2010). Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. Nucleic Acids Res *38*, e204.

Shi, Z., Wang, J., and Zhang, B. (2013). NetGestalt: integrating multidimensional omics data over biological networks. Nat Methods *10*, 597-598.

Shimamoto, Y., Nukatsuka, M., Takechi, T., and Fukushima, M. (2016). Association between mRNA expression of chemotherapy-related genes and clinicopathological features in colorectal cancer: A large-scale population analysis. Int J Mol Med *37*, 319-328.

Shirota, Y., Stoehlmacher, J., Brabender, J., Xiong, Y.P., Uetake, H., Danenberg, K.D., Groshen, S., Tsao-Wei, D.D., Danenberg, P.V., and Lenz, H.J. (2001). ERCC1 and thymidylate synthase mRNA levels predict survival for colorectal cancer patients receiving combination oxaliplatin and fluorouracil chemotherapy. J Clin Oncol *19*, 4298-4304.

Siegel, R., Desantis, C., and Jemal, A. (2014). Colorectal cancer statistics, 2014. CA Cancer J Clin *64*, 104-117.

Siena, S., Sartore-Bianchi, A., Di Nicolantonio, F., Balfour, J., and Bardelli, A. (2009). Biomarkers predicting clinical outcome of epidermal growth factor receptor-targeted therapy in metastatic colorectal cancer. J Natl Cancer Inst *101*, 1308-1324.

Silver, J.D., Ritchie, M.E., and Smyth, G.K. (2009). Microarray background correction: maximum likelihood estimation for the normal-exponential convolution. Biostatistics *10*, 352-363.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J Stat Softw *39*, 1-13.

Simpson, R.J., Connolly, L.M., Eddes, J.S., Pereira, J.J., Moritz, R.L., and Reid, G.E. (2000). Proteomic analysis of the human colon carcinoma cell line (LIM 1215): development of a membrane protein database. Electrophoresis *21*, 1707-1732.

Sims, A.H., Smethurst, G.J., Hey, Y., Okoniewski, M.J., Pepper, S.D., Howell, A., Miller, C.J., and Clarke, R.B. (2008). The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis. BMC Med Genomics *1*, 42.

Siravegna, G., Mussolin, B., Buscarino, M., Corti, G., Cassingena, A., Crisafulli, G., Ponzetti, A., Cremolini, C., Amatu, A., Lauricella, C.*, et al.* (2015). Clonal evolution and resistance to EGFR blockade in the blood of colorectal cancer patients. Nat Med *21*, 795-801.

Sjoblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N.*, et al.* (2006). The consensus coding sequences of human breast and colorectal cancers. Science *314*, 268-274.

Slebos, R.J., Brock, J.W., Winters, N.F., Stuart, S.R., Martinez, M.A., Li, M., Chambers, M.C., Zimmerman, L.J., Ham, A.J., Tabb, D.L.*, et al.* (2008). Evaluation of strong cation exchange versus isoelectric focusing of peptides for multidimensional liquid chromatography-tandem mass spectrometry. J Proteome Res *7*, 5286-5294.

Smyth, G.K., and Speed, T. (2003). Normalization of cDNA microarray data. Methods *31*, 265-273.

Smyth, H.D. (1922). A new method for studying ionising potentials. In Proceedings of the Royal Society of London Series A.

Soeda, H., Shimodaira, H., Watanabe, M., Suzuki, T., Gamoh, M., Mori, T., Komine, K., Iwama, N., Kato, S., and Ishioka, C. (2013). Clinical usefulness of KRAS, BRAF, and PIK3CA mutations as predictive markers of cetuximab efficacy in irinotecan- and oxaliplatin-refractory Japanese patients with metastatic colorectal cancer. Int J Clin Oncol *18*, 670-677.

Solic, N., Collins, J.E., Richter, A., Holt, S.J., Campbell, I., Alexander, P., and Davies, D.E. (1995). Two newly established cell lines derived from the same colonic adenocarcinoma exhibit differences in EGF-receptor ligand and adhesion molecule expression. International journal of cancer Journal international du cancer *62*, 48-57.

Solit, D.B., Garraway, L.A., Pratilas, C.A., Sawai, A., Getz, G., Basso, A., Ye, Q., Lobo, J.M., She, Y., Osman, I.*, et al.* (2006). BRAF mutation predicts sensitivity to MEK inhibition. Nature *439*, 358-362.

Sos, M.L., Michel, K., Zander, T., Weiss, J., Frommolt, P., Peifer, M., Li, D., Ullrich, R., Koker, M., Fischer, F.*, et al.* (2009). Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions. J Clin Invest *119*, 1727-1740.

Sreekumar, A., Nyati, M.K., Varambally, S., Barrette, T.R., Ghosh, D., Lawrence, T.S., and Chinnaiyan, A.M. (2001). Profiling of cancer cells using protein microarrays: discovery of novel radiation-regulated proteins. Cancer Res *61*, 7585-7593.

Srivastava, G., Renfro, L.A., Behrens, R.J., Lopatin, M., Chao, C., Soori, G.S., Dakhil, S.R., Mowat, R.B., Kuebler, J.P., Kim, G.*, et al.* (2014). Prospective multicenter study of the impact of oncotype DX colon cancer assay results on treatment recommendations in stage II colon cancer patients. Oncologist *19*, 492-497.

Stanley, J.R., Adkins, J.N., Slysz, G.W., Monroe, M.E., Purvine, S.O., Karpievitch, Y.V., Anderson, G.A., Smith, R.D., and Dabney, A.R. (2011). A statistical method for assessing peptide identification confidence in accurate mass and time tag proteomics. Anal Chem *83*, 6135-6140.

Staunton, J.E., Slonim, D.K., Coller, H.A., Tamayo, P., Angelo, M.J., Park, J., Scherf, U., Lee, J.K., Reinhold, W.O., Weinstein, J.N.*, et al.* (2001). Chemosensitivity prediction by transcriptional profiling. Proc Natl Acad Sci U S A *98*, 10787-10792.

Steen, H., and Mann, M. (2004). The ABC's (and XYZ's) of peptide sequencing. Nature reviews Molecular cell biology *5*, 699-711.

Stein, C.K., Qu, P., Epstein, J., Buros, A., Rosenthal, A., Crowley, J., Morgan, G., and Barlogie, B. (2015). Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. BMC Bioinformatics *16*, 63.

Stevens, S.G., and Brown, C.M. (2013). In silico estimation of translation efficiency in human cell lines: potential evidence for widespread translational control. PLoS One *8*, e57625.

Strimpakos, A., Pentheroudakis, G., Kotoula, V., De Roock, W., Kouvatseas, G., Papakostas, P., Makatsoris, T., Papamichael, D., Andreadou, A., Sgouros, J.*, et al.*

(2013). The prognostic role of ephrin A2 and endothelial growth factor receptor pathway mediators in patients with advanced colorectal cancer treated with cetuximab. Clinical colorectal cancer *12*, 267-274 e262.

Stryker, S.J., Wolff, B.G., Culp, C.E., Libbe, S.D., Ilstrup, D.M., and MacCarty, R.L. (1987). Natural history of untreated colonic polyps. Gastroenterology *93*, 1009-1013.

Su, Z., Li, Z., Chen, T., Li, Q.Z., Fang, H., Ding, D., Ge, W., Ning, B., Hong, H., Perkins, R.G.*, et al.* (2011). Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. Chem Res Toxicol *24*, 1486-1493.

Sun, Z., Zhang, Z., Liu, Z., Qiu, B., Liu, K., and Dong, G. (2014). MicroRNA-335 inhibits invasion and metastasis of colorectal cancer by targeting ZEB2. Med Oncol *31*, 982.

Sung, J.J.Y., Ng, S.C., Chan, F.K.L., Chiu, H.M., Kim, H.S., Matsuda, T., Ng, S.S.M., Lau, J.Y.W., Zheng, S., Adler, S.*, et al.* (2015). An updated Asia Pacific Consensus Recommendations on colorectal cancer screening. In Gut, pp. 121-132.

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One *6*, e21800.

Syka, J.E., Coon, J.J., Schroeder, M.J., Shabanowitz, J., and Hunt, D.F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc Natl Acad Sci U S A *101*, 9528-9533.

Tabb, D.L., Fernando, C.G., and Chambers, M.C. (2007). MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res *6*, 654-661.

Tabernero, J., Yoshino, T., Cohn, A.L., Obermannova, R., Bodoky, G., Garcia-Carbonero, R., Ciuleanu, T.E., Portnoy, D.C., Van Cutsem, E., Grothey, A.*, et al.* (2015). Ramucirumab versus placebo in combination with second-line FOLFIRI in patients with metastatic colorectal carcinoma that progressed during or after first-line therapy with bevacizumab, oxaliplatin, and a fluoropyrimidine (RAISE): a randomised, double-blind, multicentre, phase 3 study. The Lancet Oncology *16*, 499-508.

Tan, H.T., Tan, S., Lin, Q., Lim, T.K., Hew, C.L., and Chung, M.C. (2008). Quantitative and temporal proteome analysis of butyrate-treated colorectal cancer cells. Mol Cell Proteomics *7*, 1174-1185.

Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T., and Matsuo, T. (1988). Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid Communications in Mass Spectrometry *2*, 151-153.

Tavazoie, S.F., Alarcon, C., Oskarsson, T., Padua, D., Wang, Q., Bos, P.D., Gerald, W.L., and Massague, J. (2008). Endogenous human microRNAs that suppress breast cancer metastasis. Nature *451*, 147-152.

Taylor, G. (1964). Disintegration of water drops in an electric field. In Proceedings of the Royal Society of ….

TCGA Network (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609-615.

TCGA Network (2012a). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330-337.

TCGA Network (2012b). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330-337.

Therneau, T.M. (2015). A Package for Survival Analysis in S.

Therneau, T.M., and Grambsch, P.M. (2000). Modeling Survival Data: Extending the Cox Model (New York: Springer).

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

Thomson, J.J. (1913). Bakerian lecture: Rays of positive electricity. In Proceedings of the Royal Society of London Series A.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met *58*, 267-288.

Tipping, M.E., and Bishop, C.M. (1999). Probabilistic principal component analysis. J Roy Stat Soc B *61*, 611-622.

Toffoli, G., Cecchin, E., Corona, G., Russo, A., Buonadonna, A., D'Andrea, M., Pasetto, L.M., Pessa, S., Errante, D., De Pangher, V.*, et al.* (2006). The role of UGT1A1*28 polymorphism in the pharmacodynamics and pharmacokinetics of irinotecan in patients with metastatic colorectal cancer. J Clin Oncol *24*, 3061-3068.

Tol, J., Koopman, M., Rodenburg, C.J., Cats, A., Creemers, G.J., Schrama, J.G., Erdkamp, F.L., Vos, A.H., Mol, L., Antonini, N.F.*, et al.* (2008). A randomised phase III study on capecitabine, oxaliplatin and bevacizumab with or without cetuximab in first-line advanced colorectal cancer, the CAIRO2 study of the Dutch Colorectal Cancer Group (DCCG). An interim analysis of toxicity. Ann Oncol *19*, 734-738.

Tominaga, T., Iwahashi, M., Takifuji, K., Hotta, T., Yokoyama, S., Matsuda, K., Higashiguchi, T., Oku, Y., Nasu, T., and Yamaue, H. (2010). Combination of p53 codon 72 polymorphism and inactive p53 mutation predicts chemosensitivity to 5-fluorouracil in colorectal cancer. Int J Cancer *126*, 1691-1701.

Tong, M., Zheng, W., Li, H., Li, X., Ao, L., Shen, Y., Liang, Q., Li, J., Hong, G., Yan, H.*, et al.* (2016). Multi-omics landscapes of colorectal cancer subtypes discriminated by an individualized prognostic signature for 5-fluorouracil-based chemotherapy. Oncogenesis *5*, e242.

Tong, P., Coombes, K.R., Johnson, F.M., Byers, L.A., Diao, L., Liu, D.D., Lee, J.J., Heymach, J.V., and Wang, J. (2015). drexplorer: A tool to explore dose-response relationships and drug-drug interactions. Bioinformatics *31*, 1692-1694.

Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. CA Cancer J Clin *65*, 87-108.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics *17*, 520-525.

Tsai, J., Lee, J.T., Wang, W., Zhang, J., Cho, H., Mamo, S., Bremer, R., Gillette, S., Kong, J., Haass, N.K.*, et al.* (2008). Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity. Proc Natl Acad Sci U S A *105*, 3041-3046.

Turnbull, A.K., Kitchen, R.R., Larionov, A.A., Renshaw, L., Dixon, J.M., and Sims, A.H. (2012). Direct integration of intensity-level data from Affymetrix and Illumina microarrays improves statistical power for robust reanalysis. BMC Med Genomics *5*, 35.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A *98*, 5116-5121.

Twelves, C., Wong, A., Nowacki, M.P., Abt, M., Burris, H., 3rd, Carrato, A., Cassidy, J., Cervantes, A., Fagerberg, J., Georgoulias, V.*, et al.* (2005). Capecitabine as adjuvant treatment for stage III colon cancer. N Engl J Med *352*, 2696-2704.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods *13*, 731-740.

Uchida, K., Schneider, S., Yochim, J.M., Kuramochi, H., Hayashi, K., Takasaki, K., Yang, D., Danenberg, K.D., and Danenberg, P.V. (2005). Intratumoral COX-2 gene expression is a predictive factor for colorectal cancer response to fluoropyrimidine-based chemotherapy. Clin Cancer Res *11*, 3363-3368.

UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res *43*, D204-212.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T.*, et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530-536.

Van Cutsem, E. (2006). Challenges in the use of epidermal growth factor receptor inhibitors in colorectal cancer. Oncologist *11*, 1010-1017.

Van Cutsem, E., Kohne, C.H., Hitre, E., Zaluski, J., Chang Chien, C.R., Makhson, A., D'Haens, G., Pinter, T., Lim, R., Bodoky, G.*, et al.* (2009). Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. N Engl J Med *360*, 1408-1417.

Van Cutsem, E., Peeters, M., Siena, S., Humblet, Y., Hendlisz, A., Neyns, B., Canon, J.L., Van Laethem, J.L., Maurel, J., Richardson, G.*, et al.* (2007). Open-Label Phase III Trial of Panitumumab Plus Best Supportive Care Compared With Best Supportive Care Alone in Patients With Chemotherapy-Refractory Metastatic Colorectal Cancer. Journal of Clinical Oncology *25*, 1658-1664.

Van Cutsem, E., Tabernero, J., Lakomy, R., Prenen, H., Prausova, J., Macarulla, T., Ruff, P., van Hazel, G.A., Moiseyenko, V., Ferry, D.*, et al.* (2012). Addition of aflibercept to fluorouracil, leucovorin, and irinotecan improves survival in a phase III randomized trial in patients with metastatic colorectal cancer previously treated with an oxaliplatin-based regimen. J Clin Oncol *30*, 3499-3506.

van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J.*, et al.* (2002). A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med *347*, 1999-2009.

van Hees, F., Zauber, A.G., van Veldhuizen, H., Heijnen, M.L., Penning, C., de Koning, H.J., van Ballegooijen, M., and Lansdorp-Vogelaar, I. (2015). The value of models in informing resource allocation in colorectal cancer screening: the case of The Netherlands. Gut *64*, 1985-1997.

Venables, W.N., and Ripley, B.D. (2002). Modern Applied Statistics with S, Fourth edn (New York: Springer).

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A.*, et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Vivanco, I. (2014). Targeting molecular addictions in cancer. Br J Cancer *111*, 2033-2038.

Vogel, C., Abreu Rde, S., Ko, D., Le, S.Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M., and Penalva, L.O. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol *6*, 400.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546-1558.

Wagner, K.W., Punnoose, E.A., Januario, T., Lawrence, D.A., Pitti, R.M., Lancaster, K., Lee, D., von Goetz, M., Yee, S.F., Totpal, K.*, et al.* (2007). Death-receptor O-

glycosylation controls tumor-cell sensitivity to the proapoptotic ligand Apo2L/TRAIL. Nat Med *13*, 1070-1077.

Wang, C., Gong, B., Bushel, P.R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z.*, et al.* (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat Biotechnol *32*, 926-932.

Wang, D., and DuBois, R.N. (2013). The role of anti-inflammatory drugs in colorectal cancer. Annu Rev Med *64*, 131-144.

Wang, J., Rouse, C., Jasper, J.S., and Pendergast, A.M. (2016). ABL kinases promote breast cancer osteolytic metastasis by modulating tumor-bone interactions through TAZ and STAT5 signaling. Sci Signal *9*, ra12.

Wang, P., Lo, A., Young, J.B., Song, J.H., Lai, R., Kneteman, N.M., Hao, C., and Li, L. (2009). Targeted quantitative mass spectrometric identification of differentially expressed proteins between Bax-expressing and deficient colorectal carcinoma cells. J Proteome Res *8*, 3403-3414.

Wang, P., Tang, H., Zhang, H., Whiteaker, J., Paulovich, A.G., and Mcintosh, M. (2006). Normalization regarding non-random missing values in high-throughput mass spectrometry data. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 315-326.

Ward, L.D., Hong, J., Whitehead, R.H., and Simpson, R.J. (1990). Development of a database of amino acid sequences for human colon carcinoma proteins separated by two-dimensional polyacrylamide gel electrophoresis. Electrophoresis *11*, 883-891.

Warnat, P., Eils, R., and Brors, B. (2005). Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. BMC Bioinformatics *6*, 265.

Washburn, M.P., Wolters, D., and Yates, J.R., 3rd (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol *19*, 242-247.

Wasinger, V.C., Cordwell, S.J., Cerpa-Poljak, A., Yan, J.X., Gooley, A.A., Wilkins, M.R., Duncan, M.W., Harris, R., Williams, K.L., and Humphery-Smith, I. (1995). Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. Electrophoresis *16*, 1090-1094.

Watanabe, T., Kobunai, T., Yamamoto, Y., Matsuda, K., Ishihara, S., Nozawa, K., Iinuma, H., Ikeuchi, H., and Eshima, K. (2011a). Differential gene expression signatures between colorectal cancers with and without KRAS mutations: crosstalk between the KRAS pathway and other signalling pathways. Eur J Cancer *47*, 1946-1954.

Watanabe, T., Kobunai, T., Yamamoto, Y., Matsuda, K., Ishihara, S., Nozawa, K., Iinuma, H., Konishi, T., Horie, H., Ikeuchi, H.*, et al.* (2011b). Gene expression

signature and response to the use of leucovorin, fluorouracil and oxaliplatin in colorectal cancer patients. Clinical & translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico *13*, 419-425.

Watanabe, T., Wu, T.T., Catalano, P.J., Ueki, T., Satriano, R., Haller, D.G., Benson, A.B., 3rd, and Hamilton, S.R. (2001). Molecular predictors of survival after adjuvant chemotherapy for colon cancer. N Engl J Med *344*, 1196-1206.

Watkins, J.F., and Sanger, C. (1977). Properties of a cell line from human adenocarcinoma of the rectum. British journal of cancer *35*, 785-794.

Watson, J.T., and Biemann, K. (1964). High-Resolution Mass Spectra of Compounds Emerging from a Gas Chromatograph. Analytical Chemistry *36*, 1135-1137.

Webb-Robertson, B.J., Wiberg, H.K., Matzke, M.M., Brown, J.N., Wang, J., McDermott, J.E., Smith, R.D., Rodland, K.D., Metz, T.O., Pounds, J.G.*, et al.* (2015). Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. J Proteome Res *14*, 1993-2001.

Weickhardt, A.J., Price, T.J., Chong, G., Gebski, V., Pavlakis, N., Johns, T.G., Azad, A., Skrinos, E., Fluck, K., Dobrovic, A.*, et al.* (2012). Dual targeting of the epidermal growth factor receptor using the combination of cetuximab and erlotinib: preclinical evaluation and results of the phase II DUX study in chemotherapy-refractory, advanced colorectal cancer. J Clin Oncol *30*, 1505-1512.

Weinstein, I.B. (2000). Disorders in cell circuitry during multistage carcinogenesis: the role of homeostasis. Carcinogenesis *21*, 857-864.

Weinstein, I.B. (2002). Cancer. Addiction to oncogenes--the Achilles heal of cancer. Science *297*, 63-64.

Weinstein, I.B., and Joe, A. (2008). Oncogene addiction. Cancer Res *68*, 3077-3080; discussion 3080.

Weinstein, J.N., Myers, T.G., O'Connor, P.M., Friend, S.H., Fornace, A.J., Jr., Kohn, K.W., Fojo, T., Bates, S.E., Rubinstein, L.V., Anderson, N.L.*, et al.* (1997). An information-intensive approach to the molecular pharmacology of cancer. Science *275*, 343-349.

Wheeler, J.M., Beck, N.E., Kim, H.C., Tomlinson, I.P., Mortensen, N.J., and Bodmer, W.F. (1999). Mechanisms of inactivation of mismatch repair genes in human colorectal cancer cell lines: the predominant role of hMLH1. Proc Natl Acad Sci U S A *96*, 10296-10301.

Whitehead, R.H., Jones, J.K., Gabriel, A., and Lukies, R.E. (1987). A new colon carcinoma cell line (LIM1863) that grows as organoids with spontaneous differentiation into crypt-like structures in vitro. Cancer Research *47*, 2683-2689.

Whitehouse, C.M., Dreyer, R.N., Yamashita, M., and Fenn, J.B. (1985). Electrospray interface for liquid chromatographs and mass spectrometers. Anal Chem *57*, 675-679.

Whitehurst, A.W., Bodemann, B.O., Cardenas, J., Ferguson, D., Girard, L., Peyton, M., Minna, J.D., Michnoff, C., Hao, W., Roth, M.G.*, et al.* (2007). Synthetic lethal screen identification of chemosensitizer loci in cancer cells. Nature *446*, 815-819.

Wiese, S., Reidegeld, K.A., Meyer, H.E., and Warscheid, B. (2007). Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. Proteomics *7*, 340-350.

Wilding, J.L., and Bodmer, W.F. (2014). Cancer cell lines for drug discovery and development. Cancer Res *74*, 2377-2384.

Wilding, J.L., McGowan, S., Liu, Y., and Bodmer, W.F. (2010). Replication error deficient and proficient colorectal cancer gene expression differences caused by 3'UTR polyT sequence deletions. Proc Natl Acad Sci U S A *107*, 21058-21063.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H.*, et al.* (2014). Mass-spectrometry-based draft of the human proteome. Nature *509*, 582-587.

Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572-1573.

Wilkins, M.R., Pasquali, C., Appel, R.D., Ou, K., Golaz, O., Sanchez, J.C., Yan, J.X., Gooley, A.A., Hughes, G., Humphery-Smith, I.*, et al.* (1996a). From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Bio/technology *14*, 61-65.

Wilkins, M.R., Sanchez, J.C., Gooley, A.A., Appel, R.D., Humphery-Smith, I., Hochstrasser, D.F., and Williams, K.L. (1996b). Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnology & genetic engineering reviews *13*, 19-50.

Wilm, M., and Mann, M. (1996). Analytical properties of the nanoelectrospray ion source. Anal Chem *68*, 1-8.

Wilson, C.T.R., and Taylor, G.I. (2008). The bursting of soap-bubbles in a uniform electric field. Mathematical Proceedings of the Cambridge Philosophical Society *22*, 728.

Winawer, S.J., Zauber, A.G., Ho, M.N., O'Brien, M.J., Gottlieb, L.S., Sternberg, S.S., Waye, J.D., Schapiro, M., Bond, J.H., Panish, J.F.*, et al.* (1993). Prevention of colorectal cancer by colonoscopic polypectomy. The National Polyp Study Workgroup. N Engl J Med *329*, 1977-1981.

Wissing, J., Jansch, L., Nimtz, M., Dieterich, G., Hornberger, R., Keri, G., Wehland, J., and Daub, H. (2007). Proteomics analysis of protein kinases by target class-

selective prefractionation and tandem mass spectrometry. Mol Cell Proteomics *6*, 537-547.

Wolff, W.I., and Shinya, H. (1975). Definitive treatment of "malignant" polyps of the colon. Ann Surg *182*, 516-525.

Wong, M.C., Ching, J.Y., Chan, V.C., and Sung, J.J. (2015). The comparative cost-effectiveness of colorectal cancer screening using faecal immunochemical test vs. colonoscopy. Sci Rep *5*, 13568.

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjoblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J.*, et al.* (2007). The genomic landscapes of human breast and colorectal cancers. Science *318*, 1108-1113.

Wu, C.C., Hsu, C.W., Chen, C.D., Yu, C.J., Chang, K.P., Tai, D.I., Liu, H.P., Su, W.H., Chang, Y.S., and Yu, J.S. (2010). Candidate serological biomarkers for cancer identified from the secretomes of 23 cancer cell lines and the human protein atlas. Mol Cell Proteomics *9*, 1100-1117.

Wu, L., Candille, S.I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M. (2013). Variation and genetic control of protein abundance in humans. Nature *499*, 79-82.

Wu, Z., Doondeea, J.B., Gholami, A.M., Janning, M.C., Lemeer, S., Kramer, K., Eccles, S.A., Gollin, S.M., Grenman, R., Walch, A.*, et al.* (2011). Quantitative chemical proteomics reveals new potential drug targets in head and neck cancer. Mol Cell Proteomics *10*, M111 011635.

Xia, W., Mullin, R.J., Keith, B.R., Liu, L.H., Ma, H., Rusnak, D.W., Owens, G., Alligood, K.J., and Spector, N.L. (2002). Anti-tumor activity of GW572016: a dual tyrosine kinase inhibitor blocks EGF activation of EGFR/erbB2 and downstream Erk1/2 and AKT pathways. Oncogene *21*, 6255-6263.

Xia, X.Q., McClelland, M., Porwollik, S., Song, W., Cong, X., and Wang, Y. (2009). WebArrayDB: cross-platform microarray data analysis and public data repository. Bioinformatics *25*, 2425-2429.

Xia, Y., Liang, X., and McLuckey, S.A. (2006). Ion trap versus low-energy beam-type collision-induced dissociation of protonated ubiquitin ions. Anal Chem *78*, 1218-1227.

Xiang, C.C., Kozhich, O.A., Chen, M., Inman, J.M., Phan, Q.N., Chen, Y., and Brownstein, M.J. (2002). Amine-modified random primers to label probes for DNA microarrays. Nat Biotechnol *20*, 738-742.

Xu, X., Zhang, Y., Williams, J., Antoniou, E., McCombie, W.R., Wu, S., Zhu, W., Davidson, N.O., Denoya, P., and Li, E. (2013). Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets. BMC Bioinformatics *14 Suppl 9*, S1.

Yamashita, M., and Fenn, J.B. (1984). Electrospray Ion-Source - Another Variation on the Free-Jet Theme. Journal of Physical Chemistry *88*, 4451-4459.

Yates, J.R., 3rd, Morgan, S.F., Gatlin, C.L., Griffin, P.R., and Eng, J.K. (1998). Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. Anal Chem *70*, 3557-3565.

Yates, J.R., 3rd, Speicher, S., Griffin, P.R., and Hunkapiller, T. (1993). Peptide mass maps: a highly informative approach to protein identification. Anal Biochem *214*, 397-408.

Yin, M., Yan, J., Martinez-Balibrea, E., Graziano, F., Lenz, H.J., Kim, H.J., Robert, J., Im, S.A., Wang, W.S., Etienne-Grimaldi, M.C.*, et al.* (2011). ERCC1 and ERCC2 polymorphisms predict clinical outcomes of oxaliplatin-based chemotherapies in gastric and colorectal cancer: a systemic review and meta-analysis. Clin Cancer Res *17*, 1632-1640.

Yost, R.A., and Enke, C.G. (1978). Selected Ion Fragmentation with a Tandem Quadrupole Mass-Spectrometer. Journal of the American Chemical Society *100*, 2274-2275.

Zauber, A.G., Winawer, S.J., O'Brien, M.J., Lansdorp-Vogelaar, I., van Ballegooijen, M., Hankey, B.F., Shi, W., Bond, J.H., Schapiro, M., Panish, J.F.*, et al.* (2012). Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. N Engl J Med *366*, 687-696.

Zeleny, J. (1914). The Electrical Discharge from Liquid Points, and a Hydrostatic Method of Measuring the Electric Intensity at Their Surfaces. Physical Review *3*, 69-91.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S.*, et al.* (2014a). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382-387.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S.*, et al.* (2014b). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382-387.

Zhang, J., Yang, P.L., and Gray, N.S. (2009). Targeting cancer with small molecule kinase inhibitors. Nat Rev Cancer *9*, 28-39.

Zhao, S., Fung-Leung, W.P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One *9*, e78644.

Zheng, Y., Zhou, J., and Tong, Y. (2015). Gene signatures of drug resistance predict patient survival in colorectal cancer. Pharmacogenomics J *15*, 135-143.

Zhu, J., Shi, Z., Wang, J., and Zhang, B. (2015). Empowering biologists with multi-omics data: colorectal cancer as a paradigm. Bioinformatics *31*, 1436-1443.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *67*, 301-320.

Appendix

6.1 Supplemental tables

Table 6.1.1

Associations of Full Proteome Subtypes with other clinical and molecular data

Results of two-sided Fisher's exact tests for the association of FPSs with other subtypes and genomic/epigenomic features. Depicted are the p-values for the various tests, which were FDR-adjusted according to the method by Benjamini and Hochberg (1995). Red/blue colour highlights cases with significantly (FDR≤0.05) more/fewer occurrences of the corresponding value of a given category in one subtype compared to the remaining subtypes.

		BH-adjusted association p-value		
Category	**Values**	**FPA**	**FPB**	**FPC**
CPTAC subtype	A	0.00879	0.106	0.106
	B	0.188	0.00374	0.232
	C	0.353	0.0346	0.535
	D	0.00214	0.188	0.0000147
	E	0.0381	0.0381	0.775
Transcriptomic subtype	CIN	0.00834	0.0302	0.256
	Invasive	0.399	0.0688	0.0279
	MSI/CIMP	0.0808	0.0000401	0.263
Methylation subtype	CIMP-high	0.556	0.172	0.76
	CIMP-low	0.698	0.698	0.698
	Cluster 3	0.676	0.676	1
	Cluster 4	0.772	0.772	0.772
CMS	CMS1	0.0315	0.000826	0.547
	CMS2	0.0213	0.000821	1
	CMS3	0.521	0.287	0.129
	CMS4	1	0.282	0.457
CMS mRNA	CMS1	0.000166	0.00000343	0.629
	CMS2	0.00000381	0.000000142	0.826
	CMS3	0.567	0.0278	0.0138
	CMS4	0.405	0.031	0.33
CMS Protein	CMS1	0.0214	0.000537	0.34
	CMS2	0.00127	0.0000519	1
	CMS3	0.307	0.00695	0.182
	CMS4	0.0423	0.00531	0.575
Marisa subtype	C1	0.00872	0.0123	0.667
	C2	0.127	1	0.0464
	C3	0.647	0.314	0.314
	C4	0.000122	0.0000103	0.417
	C5	0.183	0.455	0.455

		BH-adjusted association p-value		
Category	Values	FPA	FPB	FPC
	C6	1	1	1
Budinska subtype	Subtype A	1	0.109	0.109
	Subtype B	0.00039	0.00039	0.709
	Subtype C	1	1	1
	Subtype D	0.0003	0.0000467	0.666
	Subtype E	1	1	1
Roepman subtype	A-Type	0.0458	1	0.0149
	B-Type	1.12E-08	0.00167	0.00197
	C-Type	0.0000648	0.000343	0.703
Sadanandam subtype	Enterocyte_or_Goblet-like	1	0.0357	0.0403
	Inflammatory	0.0822	0.0169	0.79
	Stem-like	0.111	0.000538	0.109
	TA	0.00307	0.00015	1
De Sousa E Melo subtype	CCS1	0.000000115	7.25E-10	1
	CCS2	0.0526	0.2	0.334
	CCS3	0.0000643	0.00000026	0.352
POLE wt	FALSE	0.635	0.401	0.635
18q loss	FALSE	0.0111	0.0111	0.617
MSI stable	FALSE	0.000396	0.0118	0.167
CIMP Pos	FALSE	0.0319	0.182	0.182
MLH1 methylation Pos	FALSE	0.0728	0.646	0.123
MLH1 wt	FALSE	0.358	1	0.289
MLH3 wt	FALSE	1	1	1
MSH3 wt	FALSE	0.0806	1	0.0806
MSH6 wt	FALSE	1	1	0.6
APC wt	FALSE	0.369	0.246	1
SOX9 wt	FALSE	1	1	1
BCL9L wt	FALSE	0.0218	0.718	0.0218
AXIN2 wt	FALSE	1	0.81	1
FBXW7 wt	FALSE	0.475	0.486	0.547
FAT1 wt	FALSE	0.714	1	0.714
CREBBP wt	FALSE	0.641	1	0.385
CTNNB1 wt	FALSE	0.319	0.319	1
TCF7L2 wt	FALSE	0.27	0.773	0.27
RNF43 wt	FALSE	0.0218	0.718	0.0218
DKK2 wt	FALSE	0.364	0.138	1
DKK3 wt	FALSE	1	1	1
DKK4 wt	FALSE	0.506	0.0494	0.506
FZD10 wt	FALSE	0.543	1	0.437
FAM123B wt	FALSE	0.654	0.654	1
KRAS wt	FALSE	0.502	0.82	0.46
NRAS wt	FALSE	0.403	1	0.403
BRAF wt	FALSE	0.248	0.248	1

		BH-adjusted association p-value		
Category	Values	FPA	FPB	FPC
ERBB2 wt	FALSE	0.742	0.742	1
ERBB3 wt	FALSE	0.00752	0.00431	1
IGF2 wt	FALSE	1	0.293	0.293
IRS2 wt	FALSE	1	0.727	1
PIK3CA wt	FALSE	0.142	1	0.12
PIK3C2B wt	FALSE	1	0.48	0.48
PIK3R1 wt	FALSE	0.546	0.546	0.561
PTEN wt	FALSE	0.742	0.742	1
TIAM1 wt	FALSE	1	1	1
TP53 wt	FALSE	0.0325	0.172	0.211
ATM wt	FALSE	0.154	0.0491	1
SMAD2 wt	FALSE	1	0.727	1
SMAD3 wt	FALSE	1	1	1
SMAD4 wt	FALSE	1	1	1
ACVR1B wt	FALSE	1	1	1
ACVR2A wt	FALSE	0.0541	0.0541	1
TGFBR1 wt	FALSE	1	0.138	0.388
TGFBR2 wt	FALSE	0.357	0.755	0.39
ARID1A wt	FALSE	0.766	0.766	0.766
CHD6 wt	FALSE	1	1	1
SRCAP wt	FALSE	0.295	1	0.246
ASH1L wt	FALSE	1	1	1
EP300 wt	FALSE	0.000957	0.469	0.00385
EP400 wt	FALSE	0.0806	1	0.0806
MLL2 wt	FALSE	0.188	0.736	0.188
MLL3 wt	FALSE	0.66	0.66	0.66
PRDM2 wt	FALSE	0.134	0.718	0.134
TRRAP wt	FALSE	0.0797	0.694	0.0892
Tumor stage	I	0.656	0.656	0.656
	IIA	1	1	1
	IIB	1	0.478	1
	IIIA	0.675	1	0.797
	IIIB	1	1	1
	IIIC	0.162	1	0.162
	IV	0.286	0.286	0.734

Table 6.1.2

Associations of Kinobeads Subtypes with other subtypes and molecular data

Results of two-sided Fisher's exact tests for the association of KSs with other subtypes and genomic/epigenomic features. Depicted are the p-values for the various tests, which were FDR-adjusted according to the method by Benjamini and Hochberg (1995). Red/blue colour highlights cases with significantly (FDR≤0.05) more/fewer occurrences of the corresponding value of a given category in one subtype compared to the remaining subtypes.

		BH-adjusted association p-value		
Category	**Values**	**KA**	**KB**	**KC**
MSI stable	FALSE	0.00000626	0.000235	0.735
KRAS wt	FALSE	1	1	1
BRAF wt	FALSE	0.611	0.611	0.653
TP53 wt	FALSE	0.0992	0.445	0.445
CMS mRNA	CMS1	0.00188	0.00001	0.109
	CMS2	0.00281	0.0000101	0.0489
	CMS3	0.661	0.661	0.661
	CMS4	1	0.731	0.113
CMS Protein	CMS1	0.185	0.0969	0.319
	CMS2	0.238	0.108	0.238
	CMS3	0.218	0.174	0.54
	CMS4	0.0846	0.00575	0.0846
De Sousa E Melo subtype	CCS1	0.00542	0.000000214	0.0083
	CCS2	0.00183	0.00961	0.704
	CCS3	1	0.000821	0.000821
Full Proteome subtype	FPA	0.00176	3.04E-11	0.0000185
	FPB	0.748	0.000000537	0.000000537
	FPC	0.0018	0.0365	0.424

Table 6.1.3

Kinases differentially expressed between Kinobeads subtypes

Results of significance analysis of microarrays (SAM, Tusher et al., 2001), aimed at the discovery of differentially expressed kinases between the different Kinobeads subtypes. Depicted are the standardised mean differences between subtype-specific kinase expression and overall mean expression of a given kinase and the corresponding local FDR (q-value) as computed using SAM. Only kinases with significant differences in expression (q-value<0.01) between the different Kinobeads subtypes are shown. Kinases are ordered first by q-value and then by the sum of their absolute standardised mean differences.

		Standardised mean difference		
Category	**q-value**	**KB**	**KA**	**KC**
SYK	0	2.255	-1.707	-2.616
CDK7	0	-2.181	1.094	3.05
ACVR2A	0	1.985	-2.762	-1.129
MARK2	0	1.867	-2.12	-1.507
TGFBR2	0	1.829	-2.782	-0.818
MAPK8	0	-1.714	2.248	1.1
LIMK2	0	1.726	-0.623	-2.641
SRC	0	1.711	-1.076	-2.19
STK24	0	1.643	-3.31	0.022
NUAK2	0	1.692	-0.983	-2.24
EPHB2	0	1.299	0.542	-2.93
NEK3	0	1.633	-1.277	-1.856
STK4	0	-0.17	-2.213	2.382
CDK1	0	-1.599	1.63	1.463
PKN1	0	-1.529	-0.039	2.89
TAOK3	0	1.519	-0.917	-1.979
FGFR2	0	0.218	1.947	-2.224
CSNK1E	0	-1.425	1.591	1.175
FRK	0	1.457	-0.104	-2.623
PRKG1	0	1.433	-1.001	-1.74
CDK17	0	-0.807	-0.881	2.328
MAPK3	0	1.358	-1.687	-0.96
PRKCD	0	1.312	-1.724	-0.84
EPHA1	0	0.56	1.171	-2.139
TNK2	0	-0.115	-1.789	1.884
CHEK1	0	-1.311	0.146	2.311
STK10	0	-0.078	-1.809	1.834
EPHB3	0	0.512	1.157	-2.036
PTK6	0	1.244	-1.887	-0.56
ACVR1B	0	1.22	-1.92	-0.485
CDK2	0	-1.214	0.851	1.471
MAP3K1	0	-0.772	-0.643	2.042

		Standardised mean difference		
Category	q-value	KB	KA	KC
FER	0	-1.145	1.444	0.791
PAK6	0	-0.553	1.968	-0.805
INSR	0	1.124	-0.202	-1.909
IRAK4	0	1.08	-1.553	-0.566
PKMYT1	0	-1.095	0.893	1.21
MAP2K6	0	1.109	-0.221	-1.864
ACVR1	0	-1.07	1.715	0.397
EPHA2	0	-1.089	0.325	1.73
STK16	0	1.086	-0.298	-1.749
CIT	0	-0.566	-0.765	1.77
OXSR1	0	1.028	-0.361	-1.582
PTK2B	0	0.976	-1.844	-0.1
ABL2	0	-0.786	-0.309	1.756
PRKAA1	0	0.919	0.064	-1.774
PRKCI	0	-0.853	-0.119	1.704
MAP3K6	0.00532	0.12	1.365	-1.497
RPS6KA1	0.00532	0.943	-0.357	-1.427
WEE1	0.00932	-0.088	1.483	-1.22
CLK1	0.00932	-0.955	0.78	1.055
CDK10	0.00932	-0.1	-1.291	1.391
SIK2	0.00932	0.495	-1.642	0.609
MAP3K5	0.00932	0.043	1.356	-1.345
SLK	0.00932	-0.922	1.139	0.658
LYN	0.00932	0.632	0.45	-1.601
PDPK1	0.00932	0.917	-0.765	-0.998
DYRK1A	0.00932	-0.904	0.488	1.233
MERTK	0.00932	0.824	-0.097	-1.448
EPHA4	0.00932	-0.72	1.514	-0.069

Table 6.1.4

Target space enrichment

Results of two-sided Fisher's exact tests for the association of FPSs with targets or similar annotations of the 709 drugs included in the four different datasets (GDSC, CCLE, CTRP, Cetuximab) integrated with the proteomics data in this study. Depicted are the p-values for the various tests, which were FDR-adjusted according to the method by Benjamini and Hochberg (1995). Red/blue colour highlights cases with significantly (FDR≤0.05) more/fewer occurrences of the corresponding drug target in drugs more effective in a given subtype compared to the remaining subtypes.

		BH-adjusted association p-value		
Category	**Values**	**FPA**	**FPB**	**FPC**
ABCG2	TRUE	0.719	1	1
ABL1	TRUE	0.741	0.514	0.514
ACAT1	TRUE	1	1	1
ACER1	TRUE	0.719	1	1
ACER2	TRUE	0.719	1	1
ACER3	TRUE	0.719	1	1
ADORA2A	TRUE	1	1	1
AHR	TRUE	1	1	1
AKT1	TRUE	0.45	0.45	0.29
AKT2	TRUE	0.561	0.561	0.561
AKT3	TRUE	1	1	1
ALK	TRUE	0.675	0.584	0.584
analog of natural product cortistatin	TRUE	0.719	1	1
analog of tetracycline	TRUE	1	1	1
analog of the natural product piperlongumine	TRUE	1	1	1
ANPEP	TRUE	1	1	1
anti-fungal agent	TRUE	1	1	1
APH1A	TRUE	1	1	1
aphrocallistin derivative	TRUE	1	1	1
AR	TRUE	0.719	1	1
ARAF	TRUE	1	0.867	1
ARF1	TRUE	0.719	1	1
ARFGAP1	TRUE	1	1	1
ARG1	TRUE	0.719	1	1
ARG2	TRUE	0.719	1	1
ASAH1	TRUE	0.719	1	1
ASAH2	TRUE	0.719	1	1
ASAH2B	TRUE	0.719	1	1
ATM	TRUE	0.866	1	0.735
ATP1A1	TRUE	1	1	1
ATP1A2	TRUE	1	1	1

		BH-adjusted association p-value		
Category	Values	FPA	FPB	FPC
ATP1A3	TRUE	1	1	1
ATP1A4	TRUE	1	1	1
ATP1B1	TRUE	1	1	1
ATP1B2	TRUE	1	1	1
ATP1B3	TRUE	1	1	1
ATP1B4	TRUE	1	1	1
ATP5L2	TRUE	1	0.867	1
ATP6V0A1	TRUE	0.719	1	1
ATPase	TRUE	1	0.867	1
ATR	TRUE	1	0.867	1
AURKA	TRUE	1	0.493	0.493
AURKB	TRUE	1	1	1
AURKC	TRUE	0.842	1	0.57
AXL	TRUE	1	1	1
BAX	TRUE	1	1	1
BCL2	TRUE	0.172	0.16	0.714
BCL2L1	TRUE	0.248	0.248	0.704
BCL2L10	TRUE	1	1	1
BCL2L2	TRUE	0.279	0.134	0.389
BCR	TRUE	0.518	0.63	0.518
BCR-ABL1	TRUE	1	1	1
BIRC5	TRUE	1	0.867	1
BLK	TRUE	1	0.25	1
BMX	TRUE	1	1	1
BRAF	TRUE	0.654	0.0758	1
BRD2	TRUE	0.172	1	1
BRD3	TRUE	0.172	1	1
BRD4	TRUE	0.172	1	1
BRDT	TRUE	0.00298	0.303	1
BTK	TRUE	1	1	1
Ca2plus transporting	TRUE	1	0.867	1
CA9	TRUE	0.719	1	1
cardiac muscle	TRUE	1	0.867	1
CASP3	TRUE	0.561	0.561	0.561
CASP6	TRUE	1	1	1
CASP7	TRUE	1	1	1
CCNB1	TRUE	1	1	1
CDC25A	TRUE	1	0.867	1
CDC25B	TRUE	1	0.867	1
CDC25C	TRUE	1	0.867	1
CDK1	TRUE	1	1	1
CDK16	TRUE	1	1	1
CDK17	TRUE	1	1	1

		BH-adjusted association p-value		
Category	Values	FPA	FPB	FPC
CDK2	TRUE	1	1	1
CDK4	FALSE	1	1	1
CDK5	TRUE	1	1	1
CDK5R1	TRUE	1	1	1
CDK6	FALSE	1	1	1
CDK7	TRUE	1	1	1
CDK9	TRUE	1	1	1
CDKL5	TRUE	1	1	1
CDKs	TRUE	1	1	1
CERK	TRUE	0.719	1	1
CERS1	TRUE	1	0.867	1
CERS2	TRUE	1	0.867	1
CERS3	TRUE	1	0.867	1
CERS4	TRUE	1	0.867	1
CERS5	TRUE	1	0.867	1
CERS6	TRUE	1	0.867	1
CHEK1	TRUE	1	1	0.393
CHEK2	TRUE	1	1	0.393
CHK1	TRUE	1	1	1
CLK1	TRUE	1	0.867	1
CLK4	TRUE	1	0.867	1
CLTA	TRUE	1	1	1
CLTB	TRUE	1	1	1
CLTC	TRUE	1	1	1
CLTCL1	TRUE	1	1	1
CMPK1	TRUE	0.584	0.584	0.584
CNR2	TRUE	0.719	1	1
CPT1A	TRUE	1	0.867	1
CRBN	TRUE	1	0.867	1
CSNK1D	TRUE	1	1	1
CSNK1E	TRUE	1	1	1
CSNK2A1	TRUE	1	0.867	1
CSNK2A2	TRUE	1	0.867	1
CTSL1	TRUE	1	0.867	1
CUL1	TRUE	1	1	1
CXCR2	TRUE	1	1	1
CYP17A1	TRUE	0.719	1	1
DCLK1	TRUE	0.719	1	1
DHEA analog	TRUE	1	0.867	1
DHFR	TRUE	1	1	0.0387
DHX9	TRUE	0.719	1	1
DIABLO	TRUE	1	0.25	1
dimer of piperlongumine	TRUE	0.719	1	1

		BH-adjusted association p-value		
Category	Values	FPA	FPB	FPC
DNA alkylator	TRUE	1	1	1
DNA crosslinker	TRUE	1	1	1
DNA damage	TRUE	0.678	0.678	0.678
DNAPK	TRUE	1	1	1
DNA synthesis	TRUE	0.633	1	0.393
DNM1	TRUE	1	1	1
DNMT1	TRUE	1	1	1
DOT1L	TRUE	1	1	1
DPP4	TRUE	1	0.867	1
DRD2	TRUE	1	1	1
EGFR	TRUE	0.0505	0.0373	0.389
EHMT1	TRUE	1	1	1
EHMT2	TRUE	1	1	1
EIF4A2	TRUE	0.172	1	1
EIF4E	TRUE	0.172	1	1
EIF4G1	TRUE	0.172	1	1
electrophilic inducer of the NFE2L2-KEAP1 pathway	TRUE	0.719	1	1
EP300	TRUE	1	0.867	1
EPHA2	TRUE	0.742	0.742	1
ERBB2	TRUE	0.218	0.118	1
ERG	TRUE	0.719	1	1
ESR1	TRUE	1	1	1
ESR2	TRUE	1	1	1
ETV1	TRUE	0.719	1	1
EZH2	TRUE	0.842	0.606	1
F2R	TRUE	1	1	1
FAAH	TRUE	1	1	1
FAK	TRUE	1	1	1
FASN	TRUE	1	1	1
FGFR1	TRUE	0.866	0.866	1
FGFR2	TRUE	1	1	1
FGFR3	TRUE	1	1	1
FGR	TRUE	1	0.25	1
flavonoid antioxidant	TRUE	1	0.867	1
FLT1	TRUE	0.225	0.225	0.615
FLT3	TRUE	0.475	0.475	0.639
FLT4	TRUE	1	0.25	1
FNTA	TRUE	1	0.945	0.118
FNTB	TRUE	1	0.867	1
FYN	TRUE	1	0.25	1
GLS	TRUE	1	0.867	1
GPER1	TRUE	1	1	1

		BH-adjusted association p-value		
Category	Values	FPA	FPB	FPC
GPX4	TRUE	1	1	1
g-secretase	TRUE	1	1	1
GSK3A	TRUE	1	1	1
GSK3B	TRUE	1	1	1
HCK	TRUE	1	0.25	1
HDAC1	TRUE	0.0131	0.0501	0.664
HDAC2	TRUE	0.00285	0.0762	0.647
HDAC3	TRUE	0.0105	0.173	0.347
HDAC6	TRUE	0.0509	0.647	0.647
HDAC8	TRUE	0.0105	0.347	0.347
HMGCR	TRUE	1	1	1
HMGCS1	TRUE	1	1	1
HNF4A	TRUE	1	1	1
HRH2	TRUE	1	0.867	1
HSP70	TRUE	1	0.434	0.203
HSP90	TRUE	1	0.434	0.203
HSP90AA1	TRUE	0.759	1	0.759
HSP90B1	TRUE	1	1	1
HSPA1A	TRUE	1	1	1
HSPA1B	TRUE	1	1	1
HSPA1L	TRUE	1	1	1
IAP	TRUE	0.345	0.00585	0.345
IDH1	TRUE	1	1	1
IGF1R	TRUE	0.121	0.689	1
IKBKB	TRUE	0.842	0.842	1
IKBKE	TRUE	1	0.867	1
IKK	TRUE	1	0.434	0.203
IKKE	TRUE	1	1	1
inducer of ROS	TRUE	0.733	0.875	1
induces cell death through a non-apoptotic and mitochondrial-dependent mechanism	TRUE	1	0.867	1
induction of rat HCN cell differentiation as judged by induction of neuronal reporter neuroD activity	TRUE	0.719	1	1
inhibitor of actin polymerization	TRUE	0.719	1	1
inhibitor of antigen receptor-mediated NFkappaB activity	TRUE	1	1	1
inhibitor of breast cancer stem cell proliferation	TRUE	1	1	0.203
inhibitor of cellular respiration	TRUE	0.719	1	1
inhibitor of downstream signaling of RIP3 associated with MLKL	TRUE	0.719	1	1
inhibitor of hedgehog signaling pathway	TRUE	1	1	1
inhibitor of JAK/STAT signaling	TRUE	0.719	1	1

		BH-adjusted association p-value		
Category	Values	FPA	FPB	FPC
inhibitor of leukemic stem cells	TRUE	1	0.434	0.203
inhibitor of LSF1-mediated transcription	TRUE	1	1	1
inhibitor of multiple kinases	TRUE	1	0.867	1
inhibitor of necroptosis	TRUE	1	0.867	1
inhibitor of p53-dependent signaling	TRUE	1	0.867	1
inhibitor of protein translation by preventing protein elongation	TRUE	1	1	1
inhibitor of RNA polymerase II	TRUE	1	1	1
inhibitor of specificity protein 1 transcription factor in cells	TRUE	1	0.867	1
inhibitor of WNT signaling	TRUE	1	0.25	1
INSR	TRUE	0.134	0.875	1
ITK	TRUE	1	1	1
JAK1	TRUE	1	1	1
JAK2	TRUE	0.678	0.678	0.678
JNK	TRUE	1	0.841	0.57
KDM1A	TRUE	1	1	1
KDM3A	TRUE	1	1	1
KDM4A	TRUE	1	1	1
KDM4B	TRUE	1	1	1
KDM4C	TRUE	1	1	1
KDM4D	TRUE	1	1	1
KDM6A	TRUE	0.719	1	1
KDM6B	TRUE	0.719	1	1
KDR	TRUE	1	0.402	1
KEAP1	TRUE	1	0.867	1
KIF11	TRUE	1	1	1
KIT	TRUE	0.654	0.654	1
KLF5	TRUE	1	1	1
KPNB1	TRUE	1	0.867	1
LAP3	TRUE	1	1	1
LCK	TRUE	1	0.225	1
LDHA	TRUE	1	1	1
LDHB	TRUE	1	1	1
LDHC	TRUE	1	1	1
LRRK2	TRUE	0.719	1	1
LYN	TRUE	1	0.25	1
LYPLA1	TRUE	1	1	1
MAP2K1	TRUE	0.00154	0.116	0.389
MAP2K2	TRUE	0.00154	0.116	0.389
MAP3K8	TRUE	0.719	1	1
MAPK1	TRUE	0.434	0.841	1
MAPK10	TRUE	1	1	1

		BH-adjusted association p-value		
Category	Values	FPA	FPB	FPC
MAPK11	TRUE	1	1	1
MAPK12	TRUE	1	1	1
MAPK13	TRUE	1	1	1
MAPK14	TRUE	0.866	1	0.735
MAPK7	TRUE	1	0.867	1
MAPK8	TRUE	1	1	1
MAPK9	TRUE	1	1	1
MCL1	TRUE	1	1	1
MDM2	TRUE	1	1	1
MDM4	TRUE	1	1	1
MEN1	TRUE	1	1	1
MET	TRUE	0.858	1	0.858
MGMT	TRUE	1	1	1
Microtubules	TRUE	0.766	0.766	0.654
modulator of NF-kappa-B signaling	TRUE	0.719	1	1
modulator of NFKB1 and STAT3 signaling	TRUE	0.719	1	1
modulator of ROS	TRUE	1	1	1
MTOR	TRUE	1	1	1
MYH1	TRUE	0.719	1	1
MYH2	TRUE	0.719	1	1
NAE1	TRUE	1	1	1
NAMPT	TRUE	0.00965	0.875	1
natural product	TRUE	0.435	1	1
NCEH1	TRUE	1	0.867	1
NCSTN	TRUE	1	1	1
NFE2L2	TRUE	1	0.867	1
NFKB1	TRUE	0.742	0.742	1
NOTCH1	TRUE	1	1	1
NPEPPS	TRUE	1	1	1
NPM1	TRUE	1	0.867	1
NR0B2	TRUE	0.719	1	1
NR1D1	TRUE	1	1	1
NR1H4	TRUE	0.719	1	1
NR3C1	TRUE	1	1	1
NTRK1	TRUE	1	0.742	0.393
organoplatinum reagent	TRUE	0.742	0.742	1
P2RX7	TRUE	1	0.867	1
PAK	TRUE	1	0.867	1
PAK4	TRUE	1	1	1
PARP1	TRUE	0.345	0.222	0.118
PARP2	TRUE	0.345	0.222	0.118
PDE4A	TRUE	1	1	1

		BH-adjusted association p-value		
Category	Values	FPA	FPB	FPC
PDE4B	TRUE	1	1	1
PDE4D	TRUE	1	1	1
PDE5A	TRUE	1	1	1
PDGFRA	TRUE	1	0.39	1
PDGFRB	TRUE	1	0.263	1
PDK1	TRUE	1	0.196	0.196
PDK2	TRUE	1	0.867	1
PDPK1	TRUE	1	0.867	1
PICK1	TRUE	1	0.867	1
PIK3CA	TRUE	0.788	1	1
PIK3CB	TRUE	1	1	1
PIK3CD	TRUE	0.752	0.752	0.752
PIK3CG	TRUE	0.917	1	1
PIM3	TRUE	1	1	1
PKM	TRUE	1	1	1
PLA2G4A	TRUE	1	1	1
PLA2G4B	TRUE	1	1	1
PLA2G4C	TRUE	1	1	1
PLA2G4D	TRUE	1	1	1
PLAUR	TRUE	1	0.867	1
PLD1	TRUE	1	0.867	1
PLD2	TRUE	1	0.25	1
PLK1	TRUE	0.182	0.68	0.584
PLK3	TRUE	1	1	1
POLA1	TRUE	0.719	1	1
PPID	TRUE	1	1	1
PPM1D	TRUE	1	1	1
PPP1R15A	TRUE	1	0.434	0.203
PPP2CA	TRUE	1	1	1
PPP3CA	TRUE	1	1	1
PPP3CB	TRUE	1	1	1
PPP3CC	TRUE	1	1	1
PPP3R1	TRUE	1	1	1
PPP3R2	TRUE	1	1	1
PRKAA1	TRUE	1	0.867	1
PRKAA2	TRUE	1	0.867	1
PRKAB1	TRUE	1	0.867	1
PRKC	TRUE	1	1	1
PRKCB	TRUE	1	1	1
PRKDC	TRUE	1	1	1
product of diversity oriented synthesis	TRUE	0.114	0.368	0.496
Prolyl-4-Hydroxylase	TRUE	1	1	1

		BH-adjusted association p-value		
Category	Values	FPA	FPB	FPC
PSEN1	TRUE	1	1	1
PSENEN	TRUE	1	1	1
PSMB1	TRUE	1	0.646	0.646
PSMB2	TRUE	1	0.646	0.646
PSMB5	TRUE	1	0.705	0.705
PSMD1	TRUE	1	0.646	0.646
PSMD2	TRUE	1	0.646	0.646
PTGS2	TRUE	1	1	1
PTK2	TRUE	1	1	1
PTP4A3	TRUE	1	0.867	1
PTPN11	TRUE	1	1	1
PTPN6	TRUE	1	1	1
RAC1	TRUE	1	0.867	1
Rac GTPases	TRUE	1	1	1
RAD51	TRUE	1	0.867	1
RAF1	TRUE	0.518	0.0792	1
RARA	TRUE	1	0.713	1
RARB	TRUE	1	0.871	1
RARG	TRUE	1	1	1
RASAL1	TRUE	0.172	1	1
RET	TRUE	1	0.494	0.494
RHOA	TRUE	1	1	1
RIPK1	TRUE	1	1	1
ROCK	TRUE	1	0.867	1
RORA	TRUE	1	0.867	1
RORC	TRUE	1	0.867	1
RPS6KB1	TRUE	1	1	1
RPS6KB2	TRUE	0.719	1	1
RRM1	TRUE	1	0.493	0.493
RSK	TRUE	1	1	1
RXRA	TRUE	1	1	0.393
RXRB	TRUE	1	1	0.393
RXRG	TRUE	1	1	0.393
S1PR1	TRUE	0.719	1	1
S1PR2	TRUE	1	0.867	1
S1PR3	TRUE	1	1	1
SCARB1	TRUE	1	0.867	1
screening hit	TRUE	0.537	0.0919	0.704
selectively kills engineered cells expressing mutant HRAS	TRUE	0.172	1	1
SGK3	TRUE	1	1	1
SIRT1	TRUE	0.866	0.866	1
SIRT2	TRUE	1	0.606	1

		BH-adjusted association p-value		
Category	Values	FPA	FPB	FPC
SKP1	TRUE	1	1	1
SLC5A1	TRUE	1	0.867	1
SLC7A11	TRUE	1	1	1
slow twitch 2	TRUE	1	0.867	1
SMO	TRUE	0.842	0.606	1
SPHK1	TRUE	1	1	1
SPTLC1	TRUE	1	0.867	1
SPTLC2	TRUE	1	0.867	1
SPTLC3	TRUE	1	0.867	1
SRC	TRUE	0.71	0.488	0.71
SRD5A2	TRUE	1	0.867	1
stapled helical peptide	TRUE	0.309	0.309	1
STAT3	TRUE	1	0.25	1
STK33	TRUE	1	1	1
SYK	TRUE	1	1	0.393
TACC3	TRUE	0.719	1	1
TASP1	TRUE	1	0.867	1
TBK1	TRUE	1	0.434	0.203
TEC	TRUE	1	1	1
TERT	TRUE	1	1	1
TGFBR1	TRUE	1	1	1
TIAM1	TRUE	1	0.867	1
TNF	TRUE	1	0.434	0.203
TNKS	TRUE	1	1	1
TOP1	TRUE	1	0.945	0.889
TOP2A	TRUE	1	1	1
TOP2B	TRUE	1	1	1
TP53	TRUE	1	1	1
TRIO	TRUE	1	0.867	1
TRPC6	TRUE	1	0.867	1
TRPV4	TRUE	1	1	1
TSPO	TRUE	0.719	1	1
TTK	TRUE	1	1	1
TXK	TRUE	1	1	1
TXN	TRUE	0.719	1	1
TYMS	TRUE	0.345	0.345	0.345
UBA1	TRUE	1	0.867	1
UCHL5	TRUE	0.719	1	1
UGCG	TRUE	1	1	1
unknown	TRUE	1	1	1
USP10	TRUE	1	0.867	1
USP13	TRUE	1	1	1
USP14	TRUE	1	1	1

		BH-adjusted association p-value		
Category	Values	FPA	FPB	FPC
USP5	TRUE	1	1	1
USP7	TRUE	1	1	1
USP9X	TRUE	0.719	1	1
VCP	TRUE	0.719	1	1
VDAC1	TRUE	1	1	1
VDAC2	TRUE	1	1	1
VDR	TRUE	1	1	1
WEE1	TRUE	0.434	0.841	1
WRN	TRUE	1	1	1
XPO1	TRUE	1	1	1
YES1	TRUE	1	0.225	1

Table 6.1.5

Univariate Cox proportional hazards regression

Results of the univariate Cox proportional hazards regression. Regression coefficients ("coef"), hazard ratios ("exp(coef)"), standard errors of coefficients ("se(coef)"), Wald statistics ("z") and its p-values ("Pr(>|z|)") are listed. Red colour highlights significant p-values. See also section 2.17.

Overall survival						
Category	Values	coef	exp(coef)	se(coef)	z	Pr(>|z|)
Gender (Ref: Female)	Male	-0.0409	0.96	0.146	-0.279	0.78
Age		0.0184	1.02	0.00823	2.24	0.025
BMI		-0.0279	0.972	0.0167	-1.67	0.0941
Treatment (Ref: CAP)	CAPBEV	0.337	1.4	0.145	2.32	0.0206
Location (Ref: Right colon)	Left colon	-0.351	0.704	0.152	-2.31	0.0207
T (Ref: T2/T3)	T3	-0.512	0.599	0.236	-2.17	0.0299
	T4	0.441	1.55	0.156	2.83	0.00463
N (Ref: N0)	N1	0.338	1.4	0.19	1.78	0.0749
	N2	1.64	5.13	0.19	8.62	0
Stage (Ref: II)	III	0.633	1.88	0.161	3.92	8.85E-05
Dukes (Ref: B)	C	0.633	1.88	0.161	3.92	8.85E-05
MSI (Ref: MSS)	MSI	-0.0869	0.917	0.219	-0.397	0.691
CIN (Ref: Negative)	Positive	0.0763	1.08	0.159	0.479	0.632
MERTK (Ref: <= 5% (n=809))	> 5% (n=146)	0.393	1.48	0.192	2.05	0.0407

Disease-free survival						
Category	**Values**	**coef**	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z|)**
Gender (Ref: Female)	Male	0.0165	1.02	0.121	0.136	0.892
Age		0.00926	1.01	0.00672	1.38	0.168
BMI		-0.0291	0.971	0.0138	-2.11	0.0347
Treatment (Ref: CAP)	CAPBEV	0.298	1.35	0.12	2.49	0.0128
Location (Ref: Right colon)	Left colon	-0.062	0.94	0.127	-0.489	0.625
T (Ref: T2/T3)	T3	-0.634	0.531	0.201	-3.16	0.00159
	T4	0.431	1.54	0.127	3.39	0.000691
N (Ref: N0)	N1	0.417	1.52	0.152	2.75	0.00594
	N2	1.43	4.17	0.161	8.85	0
Stage (Ref: II)	III	0.57	1.77	0.131	4.36	1.29E-05
Dukes (Ref: B)	C	0.57	1.77	0.131	4.36	1.29E-05
MSI (Ref: MSS)	MSI	-0.339	0.712	0.2	-1.7	0.089
CIN (Ref: Negative)	Positive	0.189	1.21	0.133	1.42	0.155
MERTK (Ref: <= 5% (n=809))	> 5% (n=146)	0.463	1.59	0.158	2.93	0.00335

Recurrence-free survival						
Category	**Values**	**coef**	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z|)**
Gender (Ref: Female)	Male	0.0416	1.04	0.13	0.321	0.748
Age		0.00362	1	0.00708	0.511	0.609
BMI		-0.0269	0.974	0.0146	-1.84	0.0665
Treatment (Ref: CAP)	CAPBEV	0.338	1.4	0.128	2.64	0.00834
Location (Ref: Right colon)	Left colon	0.0394	1.04	0.137	0.288	0.773
T (Ref: T2/T3)	T3	-0.73	0.482	0.22	-3.31	0.000921
	T4	0.402	1.5	0.135	2.99	0.00282
N (Ref: N0)	N1	0.491	1.63	0.165	2.97	0.00297
	N2	1.54	4.64	0.174	8.81	0
Stage (Ref: II)	III	0.645	1.91	0.142	4.55	5.39E-06
Dukes (Ref: B)	C	0.645	1.91	0.142	4.55	5.39E-06
MSI (Ref: MSS)	MSI	-0.512	0.599	0.229	-2.24	0.025
CIN (Ref: Negative)	Positive	0.194	1.21	0.141	1.37	0.17
MERTK (Ref: <= 5% (n=809))	> 5% (n=146)	0.491	1.63	0.167	2.94	0.00325

Table 6.1.6

Multivariate Cox proportional hazards regression

Results of the multivariate Cox proportional hazards regression. Regression coefficients ("coef"), hazard ratios ("exp(coef)"), standard errors of coefficients ("se(coef)"), Wald statistics ("z") and its p-values ("Pr(>|z|)") are listed. Red colour highlights significant p-values. See also section 2.17.

Overall survival						
Category	**Values**	**coef**	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z|)**
T (Ref: T2/T3)	T3	0.281	1.32	0.313	0.896	0.37
	T4	0.753	2.12	0.185	4.07	0.000047
N (Ref: N0)	N1	0.395	1.48	0.235	1.68	0.0927
	N2	1.72	5.56	0.231	7.42	1.21E-13
MERTK (Ref: <= 5% (n=809))	> 5% (n=146)	0.47	1.6	0.196	2.4	0.0163

Disease-free survival						
Category	**Values**	**coef**	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z|)**
T (Ref: T2/T3)	T3	0.0547	1.06	0.256	0.214	0.831
	T4	0.736	2.09	0.151	4.87	0.00000109
N (Ref: N0)	N1	0.535	1.71	0.186	2.87	0.00412
	N2	1.5	4.5	0.19	7.92	2.33E-15
MERTK (Ref: <= 5% (n=809))	> 5% (n=146)	0.531	1.7	0.161	3.3	0.00095

Recurrence-free survival						
Category	**Values**	**coef**	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z|)**
T (Ref: T2/T3)	T3	-0.00255	0.997	0.286	-0.00893	0.993
	T4	0.733	2.08	0.161	4.54	0.00000563
N (Ref: N0)	N1	0.585	1.8	0.204	2.87	0.00415
	N2	1.51	4.52	0.209	7.23	4.88E-13
MSI (Ref: MSS)	MSI	-0.573	0.564	0.271	-2.11	0.0348
MERTK (Ref: <= 5% (n=809))	> 5% (n=146)	0.571	1.77	0.172	3.33	0.00087

Table 6.1.7

Associations of high membranous MERTK expression with other clinical and molecular data

Results of two-sided Fisher's exact tests for the association of high membranous MERTK expression (more than 5% MERTK-positive tumour cells) with other clinical and molecular features. Depicted are the p-values for the various tests, which were FDR-adjusted according to the method by Benjamini and Hochberg (1995). Red/blue colour highlights cases with significantly (FDR≤0.05) more/fewer occurrences of the corresponding value of a given category in tumours with high MERTK expression compared to tumours with low MERTK expression.

		BH-adjusted association p-value
Category	**Values**	**> 5% MERTK-positive tumour cells (n=146)**
Gender	Female	0.122
Treatment	CAP	0.59
Location	Right colon	0.156
T	T2/T3	0.011
	T3	0.821
	T4	0.00377
N	N0	0.224
	N1	0.313
	N2	0.799
Stage	II	0.229
Dukes	B	0.229
MSI	MSS	0.588
CIN	Negative	0.619
KRAS	wt	0.137
BRAF_V600E	wt	0.000597

Table 6.1.8

Annotation of drug sensitivity datasets

Annotation information with respect to drugs from four drug sensitivity datasets, compiled from several publications (see section 2.18). These annotations were used in Fisher's exact tests for target space enrichment.

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
palbociclib_GDSC	PD-0332991	FPB	CDK4/6	CDK4;CDK6
NVP-BEZ235_GDSC	NVP-BEZ235	FPB	MTOR;PIK3CA;PIK3CB;PIK3CD;PIK3CG	MTOR;PIK3CA;PIK3CB;PIK3CD;PIK3CG
temsirolimus_GDSC	Temsirolimus	FPB	MTOR	MTOR
Elesclomol_GDSC	Elesclomol	FPC;FPB	HSP70	HSP70
CIL41_CTRP	CIL41	FPB	screening hit	screening hit
AZD8055_GDSC	AZD8055	FPC;FPB	MTOR	MTOR
A-804598_CTRP	A-804598	FPB	P2RX7	P2RX7
hyperforin_CTRP	hyperforin	FPB	TRPC6	TRPC6
SID 26681509_CTRP	SID 26681509	FPB	CTSL1	CTSL1
ML083_CTRP	ML083	FPB	PKM	PKM
cyanoquinoline 11_CTRP	cyanoquinoline 11	FPA	MAP3K8	MAP3K8
PF-543_CTRP	PF-543	FPA	SPHK1	SPHK1
vandetanib_CCLE	ZD-6474	FPA	EGFR;KDR	EGFR;KDR
AICAR_GDSC	AICAR	FPB	AMPK agonist	PRKAA1;PRKAA2;PRKAB1
CIL70_CTRP	CIL70	FPB	screening hit	screening hit
curcumin_CTRP	curcumin	FPA	natural product; modulator of ROS; modulator of NF-kappa-B signaling	natural product;modulator of ROS;modulator of NF-kappa-B signaling
Vinblastine_GDSC	Vinblastine	FPC;FPB	Microtubules	Microtubules
LBH-589_CTRP	LBH-589	FPA	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8
myriocin_CTRP	myriocin	FPB	SPTLC1;SPTLC2;SPTLC3	SPTLC1;SPTLC2;SPTLC3
Compound 23 citrate_CTRP	Compound 23 citrate	FPA	analog of natural product cortistatin	analog of natural product cortistatin
L-685458_CTRP	L-685458	FPA	APH1A;NCSTN;PSEN1;PSENEN	APH1A;NCSTN;PSEN1;PSENEN
belinostat_CTRP	belinostat	FPA	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8
crizotinib_in_PLX-4032 (2_in_1)_CTRP	crizotinib_in_PLX-4032 (2_in_1)	FPA	ALK;MET;BRAF	ALK;MET;BRAF
compound 1B_CTRP	compound 1B	FPB	screening hit	screening hit
BMS-195614_CTRP	BMS-195614	FPB	RARA;RARB;RARG	RARA;RARB;RARG
niclosamide_CTRP	niclosamide	FPB	STAT3	STAT3
CMK_GDSC	CMK	FPB	RSK	RSK
IPA-3_GDSC	IPA-3	FPB	PAK	PAK
selumetinib_in_JQ-1 (4_in_1)_CTRP	selumetinib_in_JQ-1 (4_in_1)	FPA	MAP2K1;MAP2K2;BRDT	MAP2K1;MAP2K2;BRDT
ML334 diastereomer_CTRP	ML334 diastereomer	FPB	KEAP1;NFE2L2	KEAP1;NFE2L2
serdemetan_GDSC	JNJ-26854165	FPB	MDM2	MDM2
MLN2480_CTRP	MLN2480	FPB	ARAF;BRAF;RAF1	ARAF;BRAF;RAF1
bosutinib_GDSC	Bosutinib	FPC;FPB	ABL1;SRC	ABL1;SRC
austocystin D_CTRP	austocystin D	FPA	natural product; inducer of DNA damage	natural product;DNA damage
tacedinaline_CTRP	tacedinaline	FPA	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8
BX-795_GDSC	BX-795	FPC;FPB	TBK1, PDK1, IKK, AURKB/C	TBK1;PDK1;IKK;AURKB;AURKC

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
dasatinib_GDSC	Dasatinib	FPB	EPHA2;KIT;LCK;SRC;YES1	EPHA2;KIT;LCK;SRC;YES1
entinostat_CTRP	entinostat	FPA	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8
neratinib_CTRP	neratinib	FPA	EGFR;ERBB2	EGFR;ERBB2
bortezomib_GDSC	Bortezomib	FPB	PSMB1;PSMB2;PSMB5;PSMD1;PSMD2	PSMB1;PSMB2;PSMB5;PSMD1;PSMD2
WH-4-023_GDSC	WH-4-023	FPB	SRC family, ABL	BLK;FGR;FYN;HCK;LCK;LYN;SRC;YES1;ABL1
Ko-143_CTRP	Ko-143	FPA	ABCG2	ABCG2
navitoclax_in_pluripotin (1_in_1)_CTRP	navitoclax_in_pluripotin (1_in_1)	FPA	BCL2;BCL2L1;BCL2L2;MAPK1;RASAL1	BCL2;BCL2L1;BCL2L2;MAPK1;RASAL1
saracatinib_GDSC	AZD-0530	FPB	ABL1;SRC	ABL1;SRC
TW-37_GDSC	TW 37	FPC;FPB	BCL2;BCL2L1	BCL2;BCL2L1
NSC95397_CTRP	NSC95397	FPB	CDC25A;CDC25B;CDC25C	CDC25A;CDC25B;CDC25C
saracatinib_CCLE	AZD0530	FPA	ABL1;SRC	ABL1;SRC
manumycin A_CTRP	manumycin A	FPB	FNTA;FNTB	FNTA;FNTB
GDC-0941_CTRP	GDC-0941	FPA	PIK3CA;PIK3CB;PIK3CD;PIK3CG	PIK3CA;PIK3CB;PIK3CD;PIK3CG
thalidomide_CTRP	thalidomide	FPB	CRBN	CRBN
XL765_CTRP	XL765	FPB	MTOR;PIK3CA;PIK3CB;PIK3CD;PIK3CG;PRKDC	MTOR;PIK3CA;PIK3CB;PIK3CD;PIK3CG;PRKDC
abiraterone_CTRP	abiraterone	FPA	CYP17A1	CYP17A1
BRD1835_CTRP	BRD1835	FPB	EZH2	EZH2
MG-132_CTRP	MG-132	FPC;FPB	PSMB1;PSMB2;PSMB5;PSMD1;PSMD2	PSMB1;PSMB2;PSMB5;PSMD1;PSMD2
JQ-1_CTRP	JQ-1	FPA	BRDT	BRDT
sunitinib_GDSC	Sunitinib	FPB	FLT1;FLT3;KDR;KIT;PDGFRA;PDGFRB	FLT1;FLT3;KDR;KIT;PDGFRA;PDGFRB
AT-406_CTRP	AT-406	FPB	XIAP	IAP
selumetinib_in_UNC0638 (4_in_1)_CTRP	selumetinib_in_UNC0638 (4_in_1)	FPA	MAP2K1;MAP2K2;EHMT1;EHMT2	MAP2K1;MAP2K2;EHMT1;EHMT2
BIX-01294_CTRP	BIX-01294	FPA	EHMT2	EHMT2
navitoclax_in_sorafenib (1_in_1)_CTRP	navitoclax_in_sorafenib (1_in_1)	FPB	BCL2;BCL2L1;BCL2L2;BRAF;FLT3;KDR;RAF1	BCL2;BCL2L1;BCL2L2;BRAF;FLT3;KDR;RAF1
PL-DI_CTRP	PL-DI	FPA	dimer of piperlongumine; inducer of ROS	dimer of piperlongumine;inducer of ROS
GSK-J4_CTRP	GSK-J4	FPA	KDM6A;KDM6B	KDM6A;KDM6B
BRD-K41334119_CTRP	BRD-K41334119	FPB	product of diversity oriented synthesis	product of diversity oriented synthesis
AZD7545_CTRP	AZD7545	FPB	PDK2	PDK2
A-770041_GDSC	A-770041	FPB	SRC family	BLK;FGR;FYN;HCK;LCK;LYN;SRC;YES1
trifluoperazine_CTRP	trifluoperazine	FPA	DRD2	DRD2
CEP-701_GDSC	CEP-701	FPC;FPB	FLT3, JAK2, NTRK1, RET	FLT3;JAK2;NTRK1;RET
avrainvillamide_CTRP	avrainvillamide	FPB	NPM1	NPM1
MK-1775_CTRP	MK-1775	FPA	WEE1	WEE1
YM-155_CTRP	YM-155	FPB	BIRC5	BIRC5
neuronal differentiation inducer III_CTRP	neuronal differentiation inducer III	FPA	induction of rat HCN cell differentiation as judged by induction of neuronal reporter neuroD activity	induction of rat HCN cell differentiation as judged by induction of neuronal reporter neuroD activity
fingolimod_CTRP	fingolimod	FPA	S1PR1	S1PR1
LY-2157299_CTRP	LY-2157299	FPB	TGFBR1	TGFBR1
ZM-447439_GDSC	ZM-447439	FPB	AURKB	AURKB
azacitidine_CTRP	azacitidine	FPB	DNMT1	DNMT1
TG-100-115_CTRP	TG-100-115	FPB	PIK3CD;PIK3CG	PIK3CD;PIK3CG
indisulam_CTRP	indisulam	FPA	CA9	CA9
MG-132_GDSC	MG-132	FPB	PSMB1;PSMB2;PSMB5;PSMD1;PSMD2	PSMB1;PSMB2;PSMB5;PSMD1;PSMD2
ML312_CTRP	ML312	FPB	SCARB1	SCARB1

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
vorinostat_CTRP	vorinostat	FPA	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8
tipifarnib-P1_GDSC	Tipifarnib	FPC;FPB	FNTA	FNTA
tretinoin_in_navitoclax (4_in_1)_CTRP	tretinoin_in_navitoclax (4_in_1)	FPB	RARA;RARB;RARG;BCL2;BCL2L1;BCL2L2	RARA;RARB;RARG;BCL2;BCL2L1;BCL2L2
RO-3306_GDSC	RO-3306	FPB	CDK1	CDK1
obatoclax_GDSC	Obatoclax Mesylate	FPC;FPB	BCL2;BCL2L1;MCL1	BCL2;BCL2L1;MCL1
decitabine_CTRP	decitabine	FPA	DNMT1	DNMT1
nutlin-3_GDSC	Nutlin-3a	FPB	MDM2	MDM2
ML031_CTRP	ML031	FPB	S1PR2	S1PR2
olaparib_GDSC	AZD-2281	FPC;FPB	PARP1;PARP2	PARP1;PARP2
GW-405833_CTRP	GW-405833	FPA	CNR2	CNR2
PHA-665752_CCLE	PHA-665752	FPA	MET	MET
BCL-LZH-4_CTRP	BCL-LZH-4	FPA	inhibitor of BCL2, BCL-xL, and MCL1	BCL2;BCL2L1;MCL1
BRD-K63431240_CTRP	BRD-K63431240	FPC;FPA	product of diversity oriented synthesis; screening hit	product of diversity oriented synthesis;screening hit
KU 0060648_CTRP	KU 0060648	FPA	PRKDC	PRKDC
BRD-K30019337_CTRP	BRD-K30019337	FPB	product of diversity oriented synthesis	product of diversity oriented synthesis
NSC23766_CTRP	NSC23766	FPB	RAC1;TIAM1;TRIO	RAC1;TIAM1;TRIO
968_CTRP	968	FPB	GLS	GLS
BRD-K02492147_CTRP	BRD-K02492147	FPB	product of diversity oriented synthesis	product of diversity oriented synthesis
daporinad_CTRP	daporinad	FPA	NAMPT	NAMPT
Bicalutamide_GDSC	Bicalutamide	FPA	Androgen receptor (ANDR)	AR
Embelin_GDSC	Embelin	FPC;FPB	XIAP	IAP
AZD6482_GDSC	AZD6482	FPB	PIK3CB;PIK3CD	PIK3CB;PIK3CD
SN-38_GDSC	Camptothecin	FPC;FPB	TOP1	TOP1
CCT018159_GDSC	CCT018159	FPB	NA	HSP90AA1
JW-74_CTRP	JW-74	FPB	inhibitor of WNT signaling	inhibitor of WNT signaling
AZD8055_CTRP	AZD8055	FPA	MTOR	MTOR
GSK-3 inhibitor IX_CTRP	GSK-3 inhibitor IX	FPA	inhibitor of JAK/STAT signaling	inhibitor of JAK/STAT signaling
LRRK2-IN-1_CTRP	LRRK2-IN-1	FPA	DCLK1;LRRK2	DCLK1;LRRK2
pifithrin-alpha_CTRP	pifithrin-alpha	FPB	inhibitor of p53-dependent signaling	inhibitor of p53-dependent signaling
PLX-4720_CTRP	PLX-4720	FPB	BRAF	BRAF
BRD-K27224038_CTRP	BRD-K27224038	FPB	product of diversity oriented synthesis	product of diversity oriented synthesis
RITA_CTRP	RITA	FPA	MDM2;TP53	MDM2;TP53
ML320_CTRP	ML320	FPA	GSK3B	GSK3B
PRIMA-1-Met_CTRP	PRIMA-1-Met	FPB	TP53	TP53
NPC-26_CTRP	NPC-26	FPB	induces cell death through a non-apoptotic, mitochondrial-dependent mechanism	induces cell death through a non-apoptotic and mitochondrial-dependent mechanism
BEC_CTRP	BEC	FPA	ARG1;ARG2	ARG1;ARG2
BRD-K48477130_CTRP	BRD-K48477130	FPB	screening hit	screening hit
JQ-1_in_vorinostat (2_in_1)_CTRP	JQ-1_in_vorinostat (2_in_1)	FPA	BRDT;HDAC1;HDAC2;HDAC3;HDAC6;HDAC8	BRDT;HDAC1;HDAC2;HDAC3;HDAC6;HDAC8
bardoxolone methyl_CTRP	bardoxolone methyl	FPA	electrophilic inducer of the NFE2L2-KEAP1 pathway	electrophilic inducer of the NFE2L2-KEAP1 pathway
WAY-362450_CTRP	WAY-362450	FPA	NR1H4	NR1H4
cetuximab_MEDICO	cetuximab_M	FPA	NA	EGFR
vismodegib_GDSC	GDC-0449	FPB	SMO	SMO
bleomycin A2_CTRP	bleomycin A2	FPA	inducer of DNA damage	DNA damage

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
Merck60_CTRP	Merck60	FPA	HDAC1;HDAC2	HDAC1;HDAC2
SB-743921_CTRP	SB-743921	FPA	KIF11	KIF11
RG-108_CTRP	RG-108	FPB	DNMT1	DNMT1
BRD9876_in_MK-1775 (4_in_1)_CTRP	BRD9876_in_MK-1775 (4_in_1)	FPA	WEE1	WEE1
DBeQ_CTRP	DBeQ	FPA	VCP	VCP
BRD-K34485477_CTRP	BRD-K34485477	FPB	product of diversity oriented synthesis	product of diversity oriented synthesis
sorafenib_CCLE	Sorafenib	FPB	BRAF;FLT3;KDR;RAF1	BRAF;FLT3;KDR;RAF1
PX-12_CTRP	PX-12	FPA	TXN	TXN
necrostatin-7_CTRP	necrostatin-7	FPB	inhibitor of necroptosis	inhibitor of necroptosis
JW-480_CTRP	JW-480	FPB	NCEH1	NCEH1
sorafenib_CTRP	sorafenib	FPB	BRAF;FLT3;KDR;RAF1	BRAF;FLT3;KDR;RAF1
KU-0063794_CTRP	KU-0063794	FPA	MTOR	MTOR
sirolimus_GDSC	Rapamycin	FPB	MTOR	MTOR
tanespimycin_CTRP	tanespimycin	FPA	HSP90AA1	HSP90AA1
B02_CTRP	B02	FPB	RAD51	RAD51
cyclophosphamide_CTRP	cyclophosphamide	FPB	DNA alkylator	DNA alkylator
JQ-1_in_carboplatin (1_in_1)_CTRP	JQ-1_in_carboplatin (1_in_1)	FPA	BRDT	BRDT
doxorubicin_GDSC	Doxorubicin	FPB	TOP2A	TOP2A
NSC48300_CTRP	NSC48300	FPB	TASP1	TASP1
PDMP_CTRP	PDMP	FPA	UGCG	UGCG
silmitasertib_CTRP	silmitasertib	FPB	CSNK2A1;CSNK2A2	CSNK2A1;CSNK2A2
skepinone-L_CTRP	skepinone-L	FPC;FPB	MAPK14	MAPK14
Salubrinal_GDSC	Salubrinal	FPC;FPB	GADD34-PP1C phosphatase	PPP1R15A
TGX-221_GDSC	TGX221	FPC;FPB	PIK3CB	PIK3CB
docetaxel_GDSC	Docetaxel	FPB	Microtubules	Microtubules
GW-843682X_CTRP	GW-843682X	FPA	PLK1;PLK3	PLK1;PLK3
BRD-K51490254_CTRP	BRD-K51490254	FPB	HDAC6;HDAC8	HDAC6;HDAC8
NSC 74859_CTRP	NSC 74859	FPB	STAT3	STAT3
GSK461364_CTRP	GSK461364	FPA	PLK1	PLK1
NVP-TAE684_CCLE	TAE684	FPA	ALK	ALK
BRD-K27986637_CTRP	BRD-K27986637	FPB	product of diversity oriented synthesis; screening hit	product of diversity oriented synthesis;screening hit
PD 153035_CTRP	PD 153035	FPA	EGFR	EGFR
navitoclax_in_MST-312 (1_in_1)_CTRP	navitoclax_in_MST-312 (1_in_1)	FPB	BCL2;BCL2L1;BCL2L2;TERT	BCL2;BCL2L1;BCL2L2;TERT
lapatinib_CCLE	Lapatinib	FPA	EGFR;ERBB2	EGFR;ERBB2
alvocidib_CTRP	alvocidib	FPA	CDK1;CDK2;CDK4;CDK6	CDK1;CDK2;CDK4;CDK6
MK-0752_CTRP	MK-0752	FPB	APH1A;NCSTN;PSEN1;PSENEN	APH1A;NCSTN;PSEN1;PSENEN
nilotinib_GDSC	Nilotinib	FPC;FPB	ABL1;BCR;KIT	ABL1;BCR;KIT
Cyclopamine_GDSC	Cyclopamine	FPA	SMO	SMO
cediranib_CTRP	cediranib	FPB	FLT1;FLT4;KDR	FLT1;FLT4;KDR
AUY922_GDSC	AUY922	FPC;FPB	HSP90	HSP90
apicidin_CTRP	apicidin	FPA	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8
BRD-K66532283_CTRP	BRD-K66532283	FPA	HDAC1;HDAC2	HDAC1;HDAC2
CAY10618_CTRP	CAY10618	FPA	NAMPT	NAMPT

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
BYL-719_CTRP	BYL-719	FPA	PIK3CA	PIK3CA
JQ-1_in_UNC0638 (2_in_1)_CTRP	JQ-1_in_UNC0638 (2_in_1)	FPA	BRDT;EHMT1;EHMT2	BRDT;EHMT1;EHMT2
VU0155056_CTRP	VU0155056	FPB	PLD1;PLD2	PLD1;PLD2
WP1130_CTRP	WP1130	FPA	UCHL5;USP14;USP5;USP9X	UCHL5;USP14;USP5;USP9X
decitabine_in_carboplatin (1_in_1)_CTRP	decitabine_in_carbopl atin (1_in_1)	FPA	DNMT1	DNMT1
BRD-K84807411_CTRP	BRD-K84807411	FPB	product of diversity oriented synthesis	product of diversity oriented synthesis
PRL-3 inhibitor I_CTRP	PRL-3 inhibitor I	FPB	PTP4A3	PTP4A3
tanespimycin_GDSC	17-AAG	FPB	HSP90AA1	HSP90AA1
serdemetan_in_SCH-529074 (1_in_1)_CTRP	serdemetan_in_SCH-529074 (1_in_1)	FPA	MDM2;TP53	MDM2;TP53
staurosporine_CTRP	staurosporine	FPB	inhibitor of multiple kinases	inhibitor of multiple kinases
phloretin_CTRP	phloretin	FPB	SLC5A1	SLC5A1
GSK525762A_CTRP	GSK525762A	FPA	BRD2;BRD3;BRD4	BRD2;BRD3;BRD4
JQ-1_in_MK-0752 (1_in_1)_CTRP	JQ-1_in_MK-0752 (1_in_1)	FPA	BRDT;APH1A;NCSTN;PSEN1;PSENE N	BRDT;APH1A;NCSTN;PSEN1;PSEN EN
BRD-K78574327_CTRP	BRD-K78574327	FPB	product of diversity oriented synthesis	product of diversity oriented synthesis
CAY10576_CTRP	CAY10576	FPB	IKBKE	IKBKE
FTI-277_GDSC	FTI-277	FPC	Farnesyl transferase (FNTA)	FNTA
PAC-1_CTRP	PAC-1	FPA	CASP3	CASP3
SGX-523_CTRP	SGX-523	FPB	MET	MET
BRD1812_CTRP	BRD1812	FPA	screening hit	screening hit
CHM-1_CTRP	CHM-1	FPA	inhibitor of tubulin polymerization	Microtubules
navitoclax_CTRP	navitoclax	FPB	BCL2;BCL2L1;BCL2L2	BCL2;BCL2L1;BCL2L2
GSK-1904529A_GDSC	GSK-1904529A	FPB	NA	IGF1R
KPT185_CTRP	KPT185	FPB	XPO1	XPO1
selumetinib_in_piperlongumine (8_in_1)_CTRP	selumetinib_in_piperl ongumine (8_in_1)	FPA	MAP2K1;MAP2K2	MAP2K1;MAP2K2
SR-II-138A_CTRP	SR-II-138A	FPA	EIF4A2;EIF4E;EIF4G1	EIF4A2;EIF4E;EIF4G1
BRD-K33514849_CTRP	BRD-K33514849	FPA	screening hit	screening hit
I-BET151_CTRP	I-BET151	FPA	BRD2;BRD3;BRD4	BRD2;BRD3;BRD4
dovitinib_CCLE	TKI258	FPB	EGFR, FGFR1, PDGFRbeta, VEGFR-1, KDR	EGFR;FGFR1;PDGFRB;FLT1;KDR
BRD-K80183349_CTRP	BRD-K80183349	FPA	HDAC1;HDAC2	HDAC1;HDAC2
HLI 373_CTRP	HLI 373	FPA	MDM2	MDM2
BRD-A94377914_CTRP	BRD-A94377914	FPA	HDAC1;HDAC2;HDAC3;HDAC6;HDA C8	HDAC1;HDAC2;HDAC3;HDAC6;HD AC8
BRD-K29313308_CTRP	BRD-K29313308	FPA	HDAC3	HDAC3
BRD-K92856060_CTRP	BRD-K92856060	FPB	screening hit	screening hit
cytarabine hydrochloride_CTRP	cytarabine hydrochloride	FPA	DNA synthesis	DNA synthesis
NVP-231_CTRP	NVP-231	FPA	CERK	CERK
entinostat_GDSC	MS-275	FPC;FPB	HDAC1;HDAC2;HDAC3;HDAC6;HDA C8	HDAC1;HDAC2;HDAC3;HDAC6;HD AC8
SNX-2112_in_bortezomib (250_in_1)_CTRP	SNX-2112_in_bortezomib (250_in_1)	FPB	HSP90AA1;HSP90B1;PSMB1;PSMB 2;PSMB5;PSMD1;PSMD2	HSP90AA1;HSP90B1;PSMB1;PSMB 2;PSMB5;PSMD1;PSMD2
MK-2206_GDSC	MK-2206	FPC;FPB	AKT1	AKT1
CIL55_CTRP	CIL55	FPA	screening hit	screening hit
salermide_in_PLX-4032 (12_in_1)_CTRP	salermide_in_PLX-4032 (12_in_1)	FPB	SIRT1;SIRT2;BRAF	SIRT1;SIRT2;BRAF
ABT-199_CTRP	ABT-199	FPB	BCL2	BCL2
AM-580_CTRP	AM-580	FPB	RARA	RARA
BRD-K90370028_CTRP	BRD-K90370028	FPB	screening hit	screening hit

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
BRD-K48334597_CTRP	BRD-K48334597	FPB	product of diversity oriented synthesis	product of diversity oriented synthesis
etoposide_GDSC	Etoposide	FPC;FPB	TOP2A	TOP2A
cucurbitacin I_CTRP	cucurbitacin I	FPA	natural product; modulator of NFKB1 and STAT3 signaling	natural product;modulator of NFKB1 and STAT3 signaling
epigallocatechin-3-monogallate_CTRP	epigallocatechin-3-monogallate	FPB	natural product	natural product
ISOX_in_bortezomib (250_in_1)_CTRP	ISOX_in_bortezomib (250_in_1)	FPA	HDAC6;PSMB1;PSMB2;PSMB5;PSMD1;PSMD2	HDAC6;PSMB1;PSMB2;PSMB5;PSMD1;PSMD2
BMS-345541_CTRP	BMS-345541	FPA	IKBKB	IKBKB
BI-2536_GDSC	BI-2536	FPC;FPB	PLK1	PLK1
selumetinib_GDSC	AZD6244	FPB	MAP2K1;MAP2K2	MAP2K1;MAP2K2
navitoclax_in_piperlongumine (1_in_1)_CTRP	navitoclax_in_piperlongumine (1_in_1)	FPA	BCL2;BCL2L1;BCL2L2	BCL2;BCL2L1;BCL2L2
FH535_GDSC	FH535	FPB	unknown	unknown
ABT-737_CTRP	ABT-737	FPB	BCL2;BCL2L1;BCL2L2	BCL2;BCL2L1;BCL2L2
CGP-60474_GDSC	CGP-60474	FPB	CDK1/2/5/7/9	CDK1;CDK2;CDK5;CDK7;CDK9
BMS-754807_CTRP	BMS-754807	FPA	IGF1R	IGF1R
necrosulfonamide_CTRP	necrosulfonamide	FPA	inhibitor of downstream signaling of RIP3 associated with MLKL	inhibitor of downstream signaling of RIP3 associated with MLKL
QS-11_GDSC	QS11	FPB	ARFGAP1	ARFGAP1
birinapant_CTRP	birinapant	FPB	DIABLO;XIAP	DIABLO;IAP
methotrexate_CTRP	methotrexate	FPC;FPB	DHFR	DHFR
OSU-03012_GDSC	OSU-03012	FPB	NA	PDK1;PDPK1
oxaliplatin_CTRP	oxaliplatin	FPA	DNA alkylator; organoplatinum reagent	DNA alkylator;organoplatinum reagent
fumonisin B1_CTRP	fumonisin B1	FPB	CERS1;CERS2;CERS3;CERS4;CERS5;CERS6	CERS1;CERS2;CERS3;CERS4;CERS5;CERS6
selumetinib_in_tretinoin (2_in_1)_CTRP	selumetinib_in_tretinoin (2_in_1)	FPA	MAP2K1;MAP2K2;RARA;RARB;RARG	MAP2K1;MAP2K2;RARA;RARB;RARG
selumetinib_in_vorinostat (8_in_1)_CTRP	selumetinib_in_vorinostat (8_in_1)	FPA	MAP2K1;MAP2K2;HDAC1;HDAC2;HDAC3;HDAC6;HDAC8	MAP2K1;MAP2K2;HDAC1;HDAC2;HDAC3;HDAC6;HDAC8
linsitinib_CTRP	linsitinib	FPA	IGF1R;INSR	IGF1R;INSR
FSC231_CTRP	FSC231	FPB	PICK1	PICK1
NVP-BSK805_CTRP	NVP-BSK805	FPA	JAK2	JAK2
ibrutinib_CTRP	ibrutinib	FPB	BTK	BTK
CIL56_CTRP	CIL56	FPB	screening hit	screening hit
RAF265_CCLE	RAF265	FPC;FPB	BRAF;KDR	BRAF;KDR
AT7867_CTRP	AT7867	FPA	AKT1;AKT2;AKT3;RPS6KB2	AKT1;AKT2;AKT3;RPS6KB2
GMX-1778_CTRP	GMX-1778	FPA	NAMPT	NAMPT
vorinostat_in_navitoclax (4_in_1)_CTRP	vorinostat_in_navitoclax (4_in_1)	FPB	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8;BCL2;BCL2L1;BCL2L2	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8;BCL2;BCL2L1;BCL2L2
sitagliptin_CTRP	sitagliptin	FPB	DPP4	DPP4
MK-2206_CTRP	MK-2206	FPA	AKT1	AKT1
BRD-K09344309_CTRP	BRD-K09344309	FPB	screening hit	screening hit
BRD4132_CTRP	BRD4132	FPB	screening hit	screening hit
afatinib_CTRP	afatinib	FPA	EGFR;ERBB2	EGFR;ERBB2
NVP-ADW742_CTRP	NVP-ADW742	FPA	IGF1R	IGF1R
CCT036477_CTRP	CCT036477	FPB	inhibitor of WNT signaling by blocking beta-catenin transcription	inhibitor of WNT signaling
O-6-benzylguanine_CTRP	O-6-benzylguanine	FPB	MGMT	MGMT
fulvestrant_CTRP	fulvestrant	FPB	ESR1;GPER1	ESR1;GPER1
paclitaxel_GDSC	Paclitaxel	FPB	Microtubules	Microtubules
lapatinib_CTRP	lapatinib	FPA	EGFR;ERBB2	EGFR;ERBB2
pazopanib_GDSC	Pazopanib	FPB	FLT1;FLT3;KDR;KIT;PDGFRB	FLT1;FLT3;KDR;KIT;PDGFRB

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
oligomycin A_CTRP	oligomycin A	FPB	ATP5L2	ATP5L2
ETP-46464_CTRP	ETP-46464	FPB	ATM;ATR	ATM;ATR
parthenolide_GDSC	Parthenolide	FPB	NFKB1	NFKB1
KU-55933_GDSC	KU-55933	FPC	ATM	ATM
trametinib_CTRP	trametinib	FPA	MAP2K1;MAP2K2	MAP2K1;MAP2K2
TG-101348_CTRP	TG-101348	FPA	JAK2	JAK2
CR-1-31B_CTRP	CR-1-31B	FPA	EIF4A2;EIF4E;EIF4G1	EIF4A2;EIF4E;EIF4G1
ML162_CTRP	ML162	FPA	selectively kills engineered cells expressing mutant HRAS	selectively kills engineered cells expressing mutant HRAS
parbendazole_CTRP	parbendazole	FPA	inhibitor of microtubule assembly	Microtubules
axitinib_GDSC	Axitinib	FPB	FLT1;FLT3;KDR;KIT;PDGFRA;PDGFRB	FLT1;FLT3;KDR;KIT;PDGFRA;PDGFRB
AKT inhibitor VIII_GDSC	AKT inhibitor VIII	FPC	AKT1/2	AKT1;AKT2
Bleomycin_GDSC	Bleomycin	FPC;FPB	DNA damage	DNA damage
etomoxir_CTRP	etomoxir	FPB	CPT1A	CPT1A
ciclopirox_CTRP	ciclopirox	FPA	RRM1	RRM1
Lenalidomide_GDSC	Lenalidomide	FPC;FPB	TNF alpha	TNF
salermide_CTRP	salermide	FPB	SIRT1;SIRT2	SIRT1;SIRT2
BRD-K55116708_CTRP	BRD-K55116708	FPC;FPB	product of diversity oriented synthesis; inhibitor of leukemic stem cells	product of diversity oriented synthesis;inhibitor of leukemic stem cells
spautin-1_CTRP	spautin-1	FPB	USP10;USP13	USP10;USP13
brefeldin A_CTRP	brefeldin A	FPA	ARF1	ARF1
linsitinib_GDSC	OSI-906	FPA	IGF1R;INSR	IGF1R;INSR
sunitinib_CTRP	sunitinib	FPA	FLT1;FLT3;KDR;KIT;PDGFRA;PDGFRB	FLT1;FLT3;KDR;KIT;PDGFRA;PDGFRB
16-beta-bromoandrosterone_CTRP	16-beta-bromoandrosterone	FPB	dehydroepiandrosterone (DHEA) analog	DHEA analog
vorinostat_in_carboplatin (1_in_1)_CTRP	vorinostat_in_carboplatin (1_in_1)	FPC	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8
3-Cl-AHPC_CTRP	3-Cl-AHPC	FPA	NR0B2	NR0B2
BRD-K88742110_CTRP	BRD-K88742110	FPA	HDAC8	HDAC8
AZD7762_GDSC	AZD7762	FPC	CHEK1;CHEK2	CHEK1;CHEK2
CD-437_CTRP	CD-437	FPA	RARG	RARG
dacarbazine_CTRP	dacarbazine	FPA	DNA alkylator	DNA alkylator
BRD-K51831558_CTRP	BRD-K51831558	FPB	EZH2	EZH2
BRD-K42260513_CTRP	BRD-K42260513	FPA	EZH2	EZH2
BRD-K64610608_CTRP	BRD-K64610608	FPB	product of diversity oriented synthesis; screening hit	product of diversity oriented synthesis;screening hit
cimetidine_CTRP	cimetidine	FPB	HRH2	HRH2
PAC-1_GDSC	PAC-1	FPC	CASP3	CASP3
selumetinib_in_PLX-4032 (8_in_1)_CTRP	selumetinib_in_PLX-4032 (8_in_1)	FPA	MAP2K1;MAP2K2;BRAF	MAP2K1;MAP2K2;BRAF
IPR-456_CTRP	IPR-456	FPB	PLAUR	PLAUR
SU11274_CTRP	SU11274	FPB	MET	MET
SN-38_CCLE	Irinotecan	FPB	TOP1	TOP1
gemcitabine_GDSC	Gemcitabine	FPC	CMPK1;RRM1;TYMS	CMPK1;RRM1;TYMS
XMD8-85_GDSC	XMD8-85	FPB	ERK5 (MK07)	MAPK7
bafilomycin A1_CTRP	bafilomycin A1	FPA	ATP6V0A1	ATP6V0A1
rucaparib_GDSC	AG-014699	FPB	NA	PARP1;PARP2
KHS101_CTRP	KHS101	FPA	TACC3	TACC3
selumetinib_in_decitabine (4_in_1)_CTRP	selumetinib_in_decitabine (4_in_1)	FPA	MAP2K1;MAP2K2;DNMT1	MAP2K1;MAP2K2;DNMT1

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
vandetanib_CTRP	vandetanib	FPA	EGFR;KDR	EGFR;KDR
BRD-A02303741_in_navitoclax (2_in_1)_CTRP	BRD-A02303741_in_navitoclax (2_in_1)	FPB	DOT1L;BCL2;BCL2L1;BCL2L2	DOT1L;BCL2;BCL2L1;BCL2L2
paclitaxel_CTRP	paclitaxel	FPA	Microtubules	Microtubules
selumetinib_in_MK-2206 (8_in_1)_CTRP	selumetinib_in_MK-2206 (8_in_1)	FPA	MAP2K1;MAP2K2;AKT1	MAP2K1;MAP2K2;AKT1
STF-31_CTRP	STF-31	FPA	NAMPT	NAMPT
tipifarnib-P2_CTRP	tipifarnib-P2	FPA	FNTA	FNTA
selumetinib_in_BRD-A02303741 (4_in_1)_CTRP	selumetinib_in_BRD-A02303741 (4_in_1)	FPA	MAP2K1;MAP2K2;DOT1L	MAP2K1;MAP2K2;DOT1L
vincristine_CTRP	vincristine	FPA	inhibitor of mictrotubule assembly	Microtubules
BRD-K96970199_CTRP	BRD-K96970199	FPB	product of diversity oriented synthesis	product of diversity oriented synthesis
bexarotene_GDSC	Bexarotene	FPC	RXRA;RXRB;RXRG	RXRA;RXRB;RXRG
erismodegib_CTRP	erismodegib	FPB	screening hit	screening hit
BRD9876_CTRP	BRD9876	FPB	screening hit	screening hit
parthenolide_CTRP	parthenolide	FPA	NFKB1	NFKB1
dasatinib_CTRP	dasatinib	FPA	EPHA2;KIT;LCK;SRC;YES1	EPHA2;KIT;LCK;SRC;YES1
bendamustine_CTRP	bendamustine	FPB	DNA alkylator	DNA alkylator
SCH-529074_CTRP	SCH-529074	FPA	TP53	TP53
neopeltolide_CTRP	neopeltolide	FPA	inhibitor of cellular respiration	inhibitor of cellular respiration
vorapaxar_CTRP	vorapaxar	FPA	F2R	F2R
GSK1059615_CTRP	GSK1059615	FPA	MTOR;PIK3CA;PIK3CB;PIK3CD;PIK3CG	MTOR;PIK3CA;PIK3CB;PIK3CD;PIK3CG
UNC0638_in_navitoclax (1_in_1)_CTRP	UNC0638_in_navitoclax (1_in_1)	FPB	EHMT1;EHMT2;BCL2;BCL2L1;BCL2L2	EHMT1;EHMT2;BCL2;BCL2L1;BCL2L2
cytochalasin B_CTRP	cytochalasin B	FPA	inhibitor of actin polymerization	inhibitor of actin polymerization
AS601245_GDSC	AS601245	FPC	JNK	JNK
PD-0325901_GDSC	PD-0325901	FPB	MEK1/2	MAP2K1;MAP2K2
RAF265_CTRP	RAF265	FPA	BRAF;KDR	BRAF;KDR
nutlin-3_CCLE	Nutlin-3	FPB	MDM2	MDM2
olaparib_CTRP	olaparib	FPB	PARP1;PARP2	PARP1;PARP2
selumetinib_in_GDC-0941 (4_in_1)_CTRP	selumetinib_in_GDC-0941 (4_in_1)	FPA	MAP2K1;MAP2K2;PIK3CA;PIK3CB;PIK3CD;PIK3CG	MAP2K1;MAP2K2;PIK3CA;PIK3CB;PIK3CD;PIK3CG
CAY10594_CTRP	CAY10594	FPB	PLD2	PLD2
blebbistatin_CTRP	blebbistatin	FPA	MYH1;MYH2	MYH1;MYH2
VER-155008_CTRP	VER-155008	FPA	HSPA1A;HSPA1B;HSPA1L	HSPA1A;HSPA1B;HSPA1L
BRD-K04800985_CTRP	BRD-K04800985	FPC	product of diversity oriented synthesis	product of diversity oriented synthesis
NSC30930_CTRP	NSC30930	FPB	SRD5A2	SRD5A2
KW-2449_CTRP	KW-2449	FPC	AURKA;FLT3	AURKA;FLT3
BRD-K71935468_CTRP	BRD-K71935468	FPA	inducer of ROS	inducer of ROS
dabrafenib_CTRP	dabrafenib	FPB	BRAF	BRAF
PHA-665752_GDSC	PHA-665752	FPB	MET	MET
SR1001_CTRP	SR1001	FPB	RORA;RORC	RORA;RORC
Vinorelbine_GDSC	Vinorelbine	FPC	Microtubules	Microtubules
PLX-4032_CTRP	PLX-4032	FPB	BRAF	BRAF
UNC0321_CTRP	UNC0321	FPB	EHMT2	EHMT2
BRD-K70511574_CTRP	BRD-K70511574	FPA	PLK1	PLK1
nelarabine_CTRP	nelarabine	FPA	POLA1	POLA1
masitinib_CTRP	masitinib	FPA	KIT;PDGFRA;PDGFRB	KIT;PDGFRA;PDGFRB

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
erlotinib_CCLE	Erlotinib	FPA	EGFR;ERBB2	EGFR;ERBB2
Thapsigargin_GDSC	Thapsigargin	FPB	ATPase, Ca++ transporting, cardiac muscle, slow twitch 2	ATPase;Ca2plus transporting;cardiac muscle;slow twitch 2
cytarabine hydrochloride_GDSC	Cytarabine	FPC	DNA synthesis	DNA synthesis
Ki8751_CTRP	Ki8751	FPB	KDR;KIT;PDGFRA	KDR;KIT;PDGFRA
doxorubicin_CTRP	doxorubicin	FPA	TOP2A	TOP2A
VAF-347_CTRP	VAF-347	FPB	AHR	AHR
tretinoin_CTRP	tretinoin	FPB	RARA;RARB;RARG	RARA;RARB;RARG
PD318088_CTRP	PD318088	FPA	MAP2K1;MAP2K2	MAP2K1;MAP2K2
pluripotin_CTRP	pluripotin	FPA	MAPK1;RASAL1	MAPK1;RASAL1
lenvatinib_CTRP	lenvatinib	FPB	FLT1;FLT3;KDR;KIT;PDGFRA;PDGFRB	FLT1;FLT3;KDR;KIT;PDGFRA;PDGFRB
ML239_CTRP	ML239	FPC	ML239; inhibitor of breast cancer stem cell proliferation	inhibitor of breast cancer stem cell proliferation
VX-680_CTRP	VX-680	FPA	AURKA;AURKB;AURKC	AURKA;AURKB;AURKC
veliparib_CTRP	veliparib	FPC	PARP1;PARP2	PARP1;PARP2
crizotinib_GDSC	PF-02341066	FPC	ALK;MET	ALK;MET
bosutinib_CTRP	bosutinib	FPA	ABL1;SRC	ABL1;SRC
LE-135_CTRP	LE-135	FPB	RARB	RARB
myricetin_CTRP	myricetin	FPB	flavonoid antioxidant	flavonoid antioxidant
KH-CB19_CTRP	KH-CB19	FPB	CLK1;CLK4	CLK1;CLK4
purmorphamine_CTRP	purmorphamine	FPB	SMO	SMO
BRD-K14844214_CTRP	BRD-K14844214	FPA	product of diversity oriented synthesis	product of diversity oriented synthesis
OSI-027_CTRP	OSI-027	FPA	MTOR	MTOR
BMS-536924_CTRP	BMS-536924	FPA	IGF1R;INSR	IGF1R;INSR
importazole_CTRP	importazole	FPB	KPNB1	KPNB1
topotecan_CTRP	topotecan	FPA	TOP1	TOP1
YK 4-279_CTRP	YK 4-279	FPA	DHX9;ERG;ETV1	DHX9;ERG;ETV1
BRD-K13185470_CTRP	BRD-K13185470	FPA	MCL1	MCL1
LBW242_CCLE	LBW242	FPB	IAP	IAP
FGIN-1-27_CTRP	FGIN-1-27	FPA	TSPO	TSPO
isonicotinohydroxamic acid_CTRP	isonicotinohydroxamic acid	FPB	HDAC6	HDAC6
ceranib-2_CTRP	ceranib-2	FPA	ACER1;ACER2;ACER3;ASAH1;ASAH2;ASAH2B	ACER1;ACER2;ACER3;ASAH1;ASAH2;ASAH2B
BAY 61-3606_GDSC	BAY 61-3606	FPC	SYK	SYK
GSK269962A_GDSC	GSK269962A	FPB	ROCK	ROCK
ZSTK474_CTRP	ZSTK474	FPA	PIK3CB;PIK3CD;PIK3CG	PIK3CB;PIK3CD;PIK3CG
foretinib_CTRP	foretinib	FPA	KDR;MET	KDR;MET
nilotinib_CCLE	Nilotinib	FPB	ABL1;BCR;KIT	ABL1;BCR;KIT
Compound 7d-cis_CTRP	Compound 7d-cis	FPA	XPO1	XPO1
avicin D_CTRP	avicin D	FPA	natural product	natural product
IC-87114_CTRP	IC-87114	FPB	PIK3CD	PIK3CD
bortezomib_CTRP	bortezomib	FPA	PSMB1;PSMB2;PSMB5;PSMD1;PSMD2	PSMB1;PSMB2;PSMB5;PSMD1;PSMD2
PLX-4720_CCLE	PLX4720	FPB	BRAF	BRAF
BRD8958_CTRP	BRD8958	FPB	EP300	EP300
Pyrimethamine_GDSC	Pyrimethamine	FPC	Dihydrofolate reductase (DHFR)	DHFR
PYR-41_CTRP	PYR-41	FPB	UBA1	UBA1

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
betulinic acid_CTRP	betulinic acid	FPB	natural product; inhibitor of specificity protein 1 transcription factor in cells	natural product;inhibitor of specificity protein 1 transcription factor in cells
cabozantinib_CTRP	cabozantinib	FPA	FLT3;KDR;MET;RET	FLT3;KDR;MET;RET
navitoclax_in_birinapant (1_in_1)_CTRP	navitoclax_in_birinapant (1_in_1)	FPB	BCL2;BCL2L1;BCL2L2;DIABLO;XIAP	BCL2;BCL2L1;BCL2L2;DIABLO;IAP
ML210_CTRP	ML210	FPA	selectively kills engineered cells expressing mutant HRAS	selectively kills engineered cells expressing mutant HRAS
BRD-K86535717_CTRP	BRD-K86535717	FPA	product of diversity oriented synthesis	product of diversity oriented synthesis
PD-173074_GDSC	PD-173074	FPB	FGFR1/3	FGFR1;FGFR3
BRD6340_CTRP	BRD6340	FPC	screening hit	screening hit
BRD-K44224150_CTRP	BRD-K44224150	FPB	product of diversity oriented synthesis	product of diversity oriented synthesis
tanespimycin_CCLE	17-AAG	FPC	HSP90AA1	HSP90AA1
BRD-K35604418_CTRP	BRD-K35604418	FPB	MCL1	MCL1
cisplatin_CTRP	Platin	FPB	DNA alkylator; organoplatinum reagent	DNA alkylator;organoplatinum reagent
motesanib_GDSC	AMG-706	FPB	VEGFR, RET, c-KIT, PDGFR	KDR;FLT1;FLT4;RET;KIT;PDGFRA;PDGFRB
BI-2536_CTRP	BI-2536	FPA	PLK1	PLK1
navitoclax_in_PLX-4032 (1_in_1)_CTRP	navitoclax_in_PLX-4032 (1_in_1)	FPB	BCL2;BCL2L1;BCL2L2;BRAF	BCL2;BCL2L1;BCL2L2;BRAF
isoliquiritigenin_CTRP	isoliquiritigenin	FPA	natural product	natural product
1S,3R-RSL-3_CTRP	1S,3R-RSL-3	BG	GPX4	GPX4
681640_GDSC	681640	BG	WEE1, CHK1	WEE1;CHK1
A-443654_GDSC	A-443654	BG	AKT1/2/3	AKT1;AKT2;AKT3
AA-COCF3_CTRP	AA-COCF3	BG	FAAH;PLA2G4A;PLA2G4B;PLA2G4C;PLA2G4D	FAAH;PLA2G4A;PLA2G4B;PLA2G4C;PLA2G4D
AC55649_CTRP	AC55649	BG	RARB	RARB
AEW541_CCLE	AEW541	BG	IGF-1R	IGF1R
AGK-2_CTRP	AGK-2	BG	SIRT2	SIRT2
AT13387_CTRP	AT13387	BG	HSP90AA1	HSP90AA1
AZ-3146_CTRP	AZ-3146	BG	TTK	TTK
AZ628_GDSC	AZ628	BG	BRAF	BRAF
AZD1480_CTRP	AZD1480	BG	JAK1;JAK2	JAK1;JAK2
AZD4547_CTRP	AZD4547	BG	FGFR1;FGFR2;FGFR3	FGFR1;FGFR2;FGFR3
AZD6482_CTRP	AZD6482	BG	PIK3CB;PIK3CD	PIK3CB;PIK3CD
AZD7762_CTRP	AZD7762	BG	CHEK1;CHEK2	CHEK1;CHEK2
BIBR-1532_CTRP	BIBR-1532	BG	TERT	TERT
BIRB-796_CTRP	BIRB-796	BG	MAPK11;MAPK11;MAPK14	MAPK11;MAPK11;MAPK14
BIRB-796_GDSC	BIRB 0796	BG	MAPK11;MAPK11;MAPK14	MAPK11;MAPK11;MAPK14
BMS-270394_CTRP	BMS-270394	BG	RARG	RARG
BMS-509744_GDSC	BMS-509744	BG	ITK	ITK
BMS-536924_GDSC	BMS-536924	BG	IGF1R;INSR	IGF1R;INSR
BMS-708163_GDSC	BMS-708163	BG	NA	g-secretase
BMS-754807_GDSC	BMS-754807	BG	IGF1R	IGF1R
BRD-A02303741_CTRP	BRD-A02303741	BG	DOT1L	DOT1L
BRD-A02303741_in_carboplatin (1_in_1)_CTRP	BRD-A02303741_in_carboplatin (1_in_1)	BG	DOT1L	DOT1L
BRD-A05715709_CTRP	BRD-A05715709	BG	IDH1	IDH1
BRD-A71883111_CTRP	BRD-A71883111	BG	IDH1	IDH1
BRD-A86708339_CTRP	BRD-A86708339	BG	screening hit	screening hit
BRD-K01737880_CTRP	BRD-K01737880	BG	screening hit	screening hit

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
BRD-K02251932_CTRP	BRD-K02251932	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K03536150_CTRP	BRD-K03536150	BG	BAX	BAX
BRD-K03911514_CTRP	BRD-K03911514	BG	stapled helical peptide	stapled helical peptide
BRD-K07442505_CTRP	BRD-K07442505	BG	BAX	BAX
BRD-K09587429_CTRP	BRD-K09587429	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K11533227_CTRP	BRD-K11533227	BG	HDAC1;HDAC2	HDAC1;HDAC2
BRD-K13999467_CTRP	BRD-K13999467	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K16130065_CTRP	BRD-K16130065	BG	stapled helical peptide	stapled helical peptide
BRD-K16147474_CTRP	BRD-K16147474	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K17060750_CTRP	BRD-K17060750	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K19103580_CTRP	BRD-K19103580	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K20514654_CTRP	BRD-K20514654	BG	stapled helical peptide	stapled helical peptide
BRD-K24690302_CTRP	BRD-K24690302	BG	HDAC1	HDAC1
BRD-K26531177_CTRP	BRD-K26531177	BG	analog of the natural product piperlongumine	analog of the natural product piperlongumine
BRD-K27188169_CTRP	BRD-K27188169	BG	stapled helical peptide	stapled helical peptide
BRD-K27188169_in_navitoclax (2_in_1)_CTRP	BRD-K27188169_in_navitoclax (2_in_1)	BG	BCL2;BCL2L1;BCL2L2	BCL2;BCL2L1;BCL2L2
BRD-K28456706_CTRP	BRD-K28456706	BG	HNF4A	HNF4A
BRD-K29086754_CTRP	BRD-K29086754	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K30748066_CTRP	BRD-K30748066	BG	CDK9	CDK9
BRD-K33199242_CTRP	BRD-K33199242	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K34099515_CTRP	BRD-K34099515	BG	screening hit	screening hit
BRD-K34222889_CTRP	BRD-K34222889	BG	analog of the natural product piperlongumine	analog of the natural product piperlongumine
BRD-K37390332_CTRP	BRD-K37390332	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K41597374_CTRP	BRD-K41597374	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K45681478_CTRP	BRD-K45681478	BG	product of diversity oriented synthesis; screening hit	product of diversity oriented synthesis;screening hit
BRD-K49290616_CTRP	BRD-K49290616	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K50799972_CTRP	BRD-K50799972	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K52037352_CTRP	BRD-K52037352	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K55473186_CTRP	BRD-K55473186	BG	NOTCH1	NOTCH1
BRD-K58730230_CTRP	BRD-K58730230	BG	stapled helical peptide	stapled helical peptide
BRD-K61166597_CTRP	BRD-K61166597	BG	inhibitor of HDAC1 and HDAC2	HDAC1;HDAC2
BRD-K66453893_CTRP	BRD-K66453893	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K71781559_CTRP	BRD-K71781559	BG	screening hit	screening hit
BRD-K75293299_CTRP	BRD-K75293299	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K79669418_CTRP	BRD-K79669418	BG	MDM2;MDM4	MDM2;MDM4
BRD-K85133207_CTRP	BRD-K85133207	BG	HDAC1	HDAC1
BRD-K94991378_CTRP	BRD-K94991378	BG	inducer of ROS	inducer of ROS
BRD-K96431673_CTRP	BRD-K96431673	BG	product of diversity oriented synthesis	product of diversity oriented synthesis
BRD-K97651142_CTRP	BRD-K97651142	BG	aphrocallistin derivative	aphrocallistin derivative
BRD-K99006945_CTRP	BRD-K99006945	BG	screening hit	screening hit
BRD-K99584050_CTRP	BRD-K99584050	BG	stapled helical peptide	stapled helical peptide
BRD-M00053801_CTRP	BRD-M00053801	BG	BCL2	BCL2
BRD8899_CTRP	BRD8899	BG	STK33	STK33

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
BRD9647_CTRP	BRD9647	BG	screening hit	screening hit
Bax channel blocker_CTRP	Bax channel blocker	BG	BAX	BAX
Bryostatin 1_GDSC	Bryostatin 1	BG	PRKC	PRKC
C6-ceramide_CTRP	C6-ceramide	BG	MAPK1;PPP2CA;UGCG	MAPK1;PPP2CA;UGCG
CAL-101_CTRP	CAL-101	BG	PIK3CD	PIK3CD
CBB-1007_CTRP	CBB-1007	BG	KDM1A	KDM1A
CCT007093_GDSC	CCT007093	BG	NA	PPM1D
CD-1530_CTRP	CD-1530	BG	RARG	RARG
CGP-082996_GDSC	CGP-082996	BG	CDK4	CDK4
CHIR-99021_CTRP	CHIR-99021	BG	GSK3B	GSK3B
CHIR-99021_GDSC	CHIR-99021	BG	GSK3B	GSK3B
CI-1040_GDSC	CI-1040	BG	MEK1/2	MAP2K1;MAP2K2
CI-976_CTRP	CI-976	BG	ACAT1	ACAT1
CID-5951923_CTRP	CID-5951923	BG	KLF5	KLF5
CIL55A_CTRP	CIL55A	BG	screening hit	screening hit
COL-3_CTRP	COL-3	BG	analog of tetracycline	analog of tetracycline
Ch-55_CTRP	Ch-55	BG	RARA;RARB;RARG	RARA;RARB;RARG
Cisplatin_GDSC	Cisplatin	BG	DNA crosslinker	DNA crosslinker
Compound 1541A_CTRP	Compound 1541A	BG	CASP3;CASP6;CASP7	CASP3;CASP6;CASP7
DMOG_GDSC	DMOG	BG	Prolyl-4-Hydroxylase	Prolyl-4-Hydroxylase
EHT 1864_GDSC	EHT 1864	BG	NA	Rac GTPases
ELCPK_CTRP	ELCPK	BG	screening hit	screening hit
EX-527_CTRP	EX-527	BG	SIRT1	SIRT1
Epothilone B_GDSC	Epothilone B	BG	Microtubules	Microtubules
FQI-1_CTRP	FQI-1	BG	inhibitor of LSF1-mediated transcription	inhibitor of LSF1-mediated transcription
FQI-2_CTRP	FQI-2	BG	inhibitor of LSF1-mediated transcription	inhibitor of LSF1-mediated transcription
GANT-61_CTRP	GANT-61	BG	inhibitor of hedgehog signaling pathway	inhibitor of hedgehog signaling pathway
GDC-0879_CTRP	GDC-0879	BG	BRAF	BRAF
GDC-0941_GDSC	GDC0941	BG	PIK3CA;PIK3CB;PIK3CD;PIK3CG	PIK3CA;PIK3CB;PIK3CD;PIK3CG
GNF-2_GDSC	GNF-2	BG	BCR-ABL	BCR-ABL1
GSK-650394_GDSC	GSK-650394	BG	SGK3	SGK3
GSK2636771_CTRP	GSK2636771	BG	PIK3CB	PIK3CB
GSK4112_CTRP	GSK4112	BG	NR1D1	NR1D1
GW 441756_GDSC	GW 441756	BG	NTRK1	NTRK1
GW-843682X_GDSC	GW843682X	BG	PLK1;PLK3	PLK1;PLK3
HBX-41108_CTRP	HBX-41108	BG	USP7	USP7
HC-067047_CTRP	HC-067047	BG	TRPV4	TRPV4
ISOX_CTRP	ISOX	BG	HDAC6	HDAC6
IU1_CTRP	IU1	BG	USP14	USP14
JNK Inhibitor VIII_GDSC	JNK Inhibitor VIII	BG	JNK	JNK
JNK-9L_GDSC	JNK-9L	BG	JNK	JNK
JQ-1_in_navitoclax (2_in_1)_CTRP	JQ-1_in_navitoclax (2_in_1)	BG	BRDT;BCL2;BCL2L1;BCL2L2	BRDT;BCL2;BCL2L1;BCL2L2
JW-55_CTRP	JW-55	BG	TNKS	TNKS

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
JW-7-52-1_GDSC	JW-7-52-1	BG	MTOR	MTOR
KIN001-135_GDSC	KIN001-135	BG	IKKE	IKKE
KU-55933_CTRP	KU-55933	BG	ATM	ATM
KU-60019_CTRP	KU-60019	BG	ATM	ATM
KX2-391_CTRP	KX2-391	BG	SRC	SRC
L-685458_CCLE	L-685458	BG	APH1A;NCSTN;PSEN1;PSENEN	APH1A;NCSTN;PSEN1;PSENEN
LFM-A13_GDSC	LFM-A13	BG	BTK	BTK
LY-2183240_CTRP	LY-2183240	BG	FAAH	FAAH
MGCD-265_CTRP	MGCD-265	BG	FLT1;FLT3;KDR;MET	FLT1;FLT3;KDR;MET
MI-1_CTRP	MI-1	BG	MEN1	MEN1
MI-2_CTRP	MI-2	BG	MEN1	MEN1
ML006_CTRP	ML006	BG	S1PR3	S1PR3
ML029_CTRP	ML029	BG	inhibitor of antigen receptor-mediated NFkappaB activity	inhibitor of antigen receptor-mediated NFkappaB activity
ML050_CTRP	ML050	BG	GPER1	GPER1
ML203_CTRP	ML203	BG	PKM	PKM
ML258_CTRP	ML258	BG	BCL2L10	BCL2L10
ML311_CTRP	ML311	BG	MCL1	MCL1
MLN2238_CTRP	MLN2238	BG	PSMB5	PSMB5
MST-312_CTRP	MST-312	BG	TERT	TERT
Mdivi-1_CTRP	Mdivi-1	BG	DNM1	DNM1
Midostaurin_GDSC	Midostaurin	BG	KIT	KIT
N9-isopropylolomoucine_CTRP	N9-isopropylolomoucine	BG	CCNB1;CDK1;CDK5;CDK5R1	CCNB1;CDK1;CDK5;CDK5R1
NSC-87877_GDSC	NSC-87877	BG	SHP1/2 (PTN6/11)	PTPN6;PTPN11
NSC19630_CTRP	NSC19630	BG	WRN	WRN
NSC632839_CTRP	NSC632839	BG	USP13;USP5	USP13;USP5
NU-7441_GDSC	NU-7441	BG	DNAPK	DNAPK
NVP-BEZ235_CTRP	NVP-BEZ235	BG	MTOR;PIK3CA;PIK3CB;PIK3CD;PIK3CG	MTOR;PIK3CA;PIK3CB;PIK3CD;PIK3CG
NVP-TAE684_CTRP	NVP-TAE684	BG	ALK	ALK
NVP-TAE684_GDSC	NVP-TAE684	BG	ALK	ALK
OSI-930_CTRP	OSI-930	BG	KDR;KIT	KDR;KIT
PD-0325901_CCLE	PD-0325901	BG	MEK1/2	MAP2K1;MAP2K2
PF-184_CTRP	PF-184	BG	IKBKB	IKBKB
PF-3758309_CTRP	PF-3758309	BG	PAK4	PAK4
PF-4708671_GDSC	PF-4708671	BG	NA	RPS6KB1
PF-4800567 hydrochloride_CTRP	PF-4800567 hydrochloride	BG	CSNK1D;CSNK1E	CSNK1D;CSNK1E
PF-562271_GDSC	PF-562271	BG	FAK	FAK
PF-573228_CTRP	PF-573228	BG	PTK2	PTK2
PF-750_CTRP	PF-750	BG	FAAH	FAAH
PHA-793887_CTRP	PHA-793887	BG	CDK1;CDK2;CDK4;CDK5;CDK7;CDK9	CDK1;CDK2;CDK4;CDK5;CDK7;CDK9
PI-103_CTRP	PI-103	BG	MTOR;PIK3CA;PIK3CB;PIK3CD;PIK3CG;PRKDC	MTOR;PIK3CA;PIK3CB;PIK3CD;PIK3CG;PRKDC
PIK-93_CTRP	PIK-93	BG	PIK3CG	PIK3CG
PLX-4720_GDSC	PLX4720	BG	BRAF	BRAF
PRIMA-1_CTRP	PRIMA-1	BG	TP53	TP53

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
Panobinostat_CCLE	Panobinostat	BG	HDAC	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8
QS-11_CTRP	QS-11	BG	ARFGAP1	ARFGAP1
QW-BI-011_CTRP	QW-BI-011	BG	EHMT2	EHMT2
R428_CTRP	R428	BG	AXL	AXL
RDEA119_GDSC	RDEA119	BG	MEK1/2	MAP2K1;MAP2K2
RO4929097_CTRP	RO4929097	BG	APH1A;NCSTN;PSEN1;PSENEN	APH1A;NCSTN;PSEN1;PSENEN
Repligen 136_CTRP	Repligen 136	BG	HDAC3	HDAC3
S-Trityl-L-cysteine_GDSC	S-Trityl-L-cysteine	BG	KIF11	KIF11
SB 216763_GDSC	SB 216763	BG	GSK3A/B	GSK3A;GSK3B
SB-225002_CTRP	SB-225002	BG	CXCR2	CXCR2
SB-431542_CTRP	SB-431542	BG	TGFBR1	TGFBR1
SB-525334_CTRP	SB-525334	BG	TGFBR1	TGFBR1
SB590885_GDSC	SB590885	BG	BRAF	BRAF
SCH-79797_CTRP	SCH-79797	BG	F2R	F2R
SJ-172550_CTRP	SJ-172550	BG	MDM2;TP53	MDM2;TP53
SKI-II_CTRP	SKI-II	BG	SPHK1	SPHK1
SL 0101-1_GDSC	SL 0101-1	BG	RSK, AURKB, PIM3	RSK;AURKB;PIM3
SMER-3_CTRP	SMER-3	BG	CUL1;SKP1	CUL1;SKP1
SN-38_CTRP	SN-38	BG	TOP1	TOP1
SNS-032_CTRP	SNS-032	BG	CDK16;CDK17;CDK2;CDK7;CDK9;CDKL5	CDK16;CDK17;CDK2;CDK7;CDK9;CDKL5
SNX-2112_CTRP	SNX-2112	BG	HSP90AA1;HSP90B1	HSP90AA1;HSP90B1
SR8278_CTRP	SR8278	BG	NR1D1	NR1D1
SRT-1720_CTRP	SRT-1720	BG	SIRT1	SIRT1
SZ4TA2_CTRP	SZ4TA2	BG	BCL2L1	BCL2L1
Shikonin_GDSC	Shikonin	BG	unknown	unknown
StemRegenin 1_CTRP	StemRegenin 1	BG	AHR	AHR
TGX-221_CTRP	TGX-221	BG	PIK3CB	PIK3CB
TPCA-1_CTRP	TPCA-1	BG	IKBKB	IKBKB
TW-37_CTRP	TW-37	BG	BCL2;BCL2L1	BCL2;BCL2L1
UNC0638_CTRP	UNC0638	BG	EHMT1;EHMT2	EHMT1;EHMT2
VX-680_GDSC	VX-680	BG	AURKA;AURKB;AURKC	AURKA;AURKB;AURKC
VX-702_GDSC	VX-702	BG	p38	MAPK11;MAPK12;MAPK13;MAPK14
WZ-1-84_GDSC	WZ-1-84	BG	BMX, Tec-family	BMX;TEC;BTK;ITK;BMX;TXK
WZ4002_CTRP	WZ4002	BG	EGFR	EGFR
WZ8040_CTRP	WZ8040	BG	EGFR	EGFR
YL54_CTRP	YL54	BG	stapled helical peptide	stapled helical peptide
Z-LLNle-CHO_GDSC	Z-LLNle-CHO	BG	g-secretase	g-secretase
afatinib_GDSC	BIBW2992	BG	EGFR;ERBB2	EGFR;ERBB2
alisertib_CTRP	alisertib	BG	AURKA;AURKB	AURKA;AURKB
alisertib_in_navitoclax (2_in_1)_CTRP	alisertib_in_navitoclax (2_in_1)	BG	AURKA;AURKB;BCL2;BCL2L1;BCL2L2	AURKA;AURKB;BCL2;BCL2L1;BCL2L2
axitinib_CTRP	axitinib	BG	FLT1;FLT3;KDR;KIT;PDGFRA;PDGFRB	FLT1;FLT3;KDR;KIT;PDGFRA;PDGFRB
barasertib_CTRP	barasertib	BG	AURKB	AURKB
bexarotene_CTRP	bexarotene	BG	RXRA;RXRB;RXRG	RXRA;RXRB;RXRG

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
brivanib_CTRP	brivanib	BG	FLT1;KDR	FLT1;KDR
canertinib_CTRP	canertinib	BG	EGFR;ERBB2	EGFR;ERBB2
carboplatin_CTRP	carboplatin	BG	inducer of DNA damage	DNA damage
carboplatin_in_UNC0638 (2_in_1)_CTRP	carboplatin_in_UNC0638 (2_in_1)	BG	EHMT1;EHMT2	EHMT1;EHMT2
carboplatin_in_etoposide (40_in_17)_CTRP	carboplatin_in_etoposide (40_in_17)	BG	TOP2A	TOP2A
cerulenin_CTRP	cerulenin	BG	FASN;HMGCS1	FASN;HMGCS1
chlorambucil_CTRP	chlorambucil	BG	DNA alkylator	DNA alkylator
ciclosporin_CTRP	ciclosporin	BG	PPID	PPID
clofarabine_CTRP	clofarabine	BG	inducer of DNA damage	DNA damage
crizotinib_CCLE	PF2341066	BG	ALK;MET	ALK;MET
crizotinib_CTRP	crizotinib	BG	ALK;MET	ALK;MET
dacinostat_GDSC	LAQ824	BG	NA	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8
darinaparsin_CTRP	darinaparsin	BG	inducer of ROS; inhibitor of microtubule assembly	inducer of ROS;Microtubules
decitabine_in_navitoclax (2_in_1)_CTRP	decitabine_in_navitoclax (2_in_1)	BG	DNMT1;BCL2;BCL2L1;BCL2L2	DNMT1;BCL2;BCL2L1;BCL2L2
dexamethasone_CTRP	dexamethasone	BG	NR3C1	NR3C1
dinaciclib_CTRP	dinaciclib	BG	CDK1;CDK2;CDK5;CDK9	CDK1;CDK2;CDK5;CDK9
docetaxel_CTRP	docetaxel	BG	Microtubules	Microtubules
docetaxel_in_tanespimycin (2_in_1)_CTRP	docetaxel_in_tanespimycin (2_in_1)	BG	HSP90AA1	HSP90AA1
doxorubicin_in_navitoclax (2_in_1)_CTRP	doxorubicin_in_navitoclax (2_in_1)	BG	TOP2A;BCL2;BCL2L1;BCL2L2	TOP2A;BCL2;BCL2L1;BCL2L2
elocalcitol_CTRP	elocalcitol	BG	VDR	VDR
erastin_CTRP	erastin	BG	SLC7A11;VDAC1;VDAC2	SLC7A11;VDAC1;VDAC2
erlotinib_CTRP	erlotinib	BG	EGFR;ERBB2	EGFR;ERBB2
erlotinib_GDSC	Erlotinib	BG	EGFR;ERBB2	EGFR;ERBB2
erlotinib_in_PLX-4032 (2_in_1)_CTRP	erlotinib_in_PLX-4032 (2_in_1)	BG	EGFR;ERBB2;BRAF	EGFR;ERBB2;BRAF
etoposide_CTRP	etoposide	BG	TOP2A	TOP2A
fluorouracil_CTRP	fluorouracil	BG	TYMS	TYMS
fluvastatin_CTRP	fluvastatin	BG	HMGCR	HMGCR
gefitinib_CTRP	gefitinib	BG	AKT1;EGFR	AKT1;EGFR
gefitinib_GDSC	Gefitinib	BG	AKT1;EGFR	AKT1;EGFR
gemcitabine_CTRP	gemcitabine	BG	CMPK1;RRM1;TYMS	CMPK1;RRM1;TYMS
gossypol_CTRP	gossypol	BG	BCL2;BCL2L1;LDHA;LDHB;LDHC	BCL2;BCL2L1;LDHA;LDHB;LDHC
ifosfamide_CTRP	ifosfamide	BG	DNA alkylator	DNA alkylator
imatinib_CTRP	imatinib	BG	ABL1;BCR;KIT	ABL1;BCR;KIT
imatinib_GDSC	Imatinib	BG	ABL1;BCR;KIT	ABL1;BCR;KIT
isoevodiamine_CTRP	isoevodiamine	BG	natural product	natural product
istradefylline_CTRP	istradefylline	BG	ADORA2A	ADORA2A
itraconazole_CTRP	itraconazole	BG	anti-fungal agent; inhibitor of hedgehog signaling pathway	anti-fungal agent;inhibitor of hedgehog signaling pathway
lapatinib_GDSC	Lapatinib	BG	EGFR;ERBB2	EGFR;ERBB2
leptomycin B_CTRP	leptomycin B	BG	XPO1	XPO1
linifanib_CTRP	linifanib	BG	FLT1;FLT3;KDR	FLT1;FLT3;KDR
lomeguatrib_CTRP	lomeguatrib	BG	MGMT	MGMT
lovastatin_CTRP	lovastatin	BG	HMGCR	HMGCR
marinopyrrole A_CTRP	marinopyrrole A	BG	MCL1	MCL1

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
methotrexate_GDSC	Methotrexate	BG	DHFR	DHFR
methylstat_CTRP	methylstat	BG	KDM3A;KDM4A;KDM4B;KDM4C;KDM4D	KDM3A;KDM4A;KDM4B;KDM4C;KDM4D
mitomycin_CTRP	mitomycin	BG	DNA crosslinker	DNA crosslinker
mitomycin_GDSC	Mitomycin C	BG	DNA crosslinker	DNA crosslinker
momelotinib_CTRP	momelotinib	BG	JAK1;JAK2	JAK1;JAK2
nakiterpiosin_CTRP	nakiterpiosin	BG	natural product; inhibitor of microtubule assembly	natural product;Microtubules
narciclasine_CTRP	narciclasine	BG	RHOA	RHOA
navitoclax_GDSC	ABT-263	BG	BCL2;BCL2L1;BCL2L2	BCL2;BCL2L1;BCL2L2
navitoclax_in_gemcitabine (1_in_1)_CTRP	navitoclax_in_gemcitabine (1_in_1)	BG	BCL2;BCL2L1;BCL2L2;CMPK1;RRM1;TYMS	BCL2;BCL2L1;BCL2L2;CMPK1;RRM1;TYMS
necrostatin-1_CTRP	necrostatin-1	BG	RIPK1	RIPK1
nilotinib_CTRP	nilotinib	BG	ABL1;BCR;KIT	ABL1;BCR;KIT
nintedanib_CTRP	nintedanib	BG	FGFR1;FGFR2;FGFR3;FLT1;FLT3;KDR;PDGFRA;PDGFRB	FGFR1;FGFR2;FGFR3;FLT1;FLT3;KDR;PDGFRA;PDGFRB
nutlin-3_CTRP	nutlin-3	BG	MDM2	MDM2
obatoclax_CTRP	obatoclax	BG	BCL2;BCL2L1;MCL1	BCL2;BCL2L1;MCL1
omacetaxine mepesuccinate_CTRP	omacetaxine mepesuccinate	BG	inhibitor of protein translation by preventing protein elongation	inhibitor of protein translation by preventing protein elongation
ouabain_CTRP	ouabain	BG	ATP1A1;ATP1A2;ATP1A3;ATP1A4;ATP1B1;ATP1B2;ATP1B3;ATP1B4	ATP1A1;ATP1A2;ATP1A3;ATP1A4;ATP1B1;ATP1B2;ATP1B3;ATP1B4
paclitaxel_CCLE	Paclitaxel	BG	Microtubules	Microtubules
palbociclib_CCLE	PD-0332991	BG	CDK4/6	CDK4;CDK6
palmostatin B_CTRP	palmostatin B	BG	LYPLA1	LYPLA1
pandacostat_CTRP	pandacostat	BG	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8
pazopanib_CTRP	pazopanib	BG	FLT1;FLT3;KDR;KIT;PDGFRB	FLT1;FLT3;KDR;KIT;PDGFRB
pevonedistat_CTRP	pevonedistat	BG	NAE1	NAE1
pifithrin-mu_CTRP	pifithrin-mu	BG	HSPA1A;HSPA1B;HSPA1L;TP53	HSPA1A;HSPA1B;HSPA1L;TP53
piperlongumine_CTRP	piperlongumine	BG	natural product; modulator of ROS levels	natural product;modulator of ROS
piperlongumine_in_MST-312 (1_in_1)_CTRP	piperlongumine_in_MST-312 (1_in_1)	BG	TERT	TERT
pitstop2_CTRP	pitstop2	BG	CLTA;CLTB;CLTC;CLTCL1	CLTA;CLTB;CLTC;CLTCL1
ponatinib_GDSC	AP-24534	BG	ABL	ABL1
procarbazine_CTRP	procarbazine	BG	inducer of DNA damage	DNA damage
prochlorperazine_CTRP	prochlorperazine	BG	DRD2	DRD2
pyrazolanthrone_CTRP	pyrazolanthrone	BG	MAPK10;MAPK8;MAPK9	MAPK10;MAPK8;MAPK9
quizartinib_CTRP	quizartinib	BG	FLT3	FLT3
regorafenib_CTRP	regorafenib	BG	BRAF;KDR;KIT;RET	BRAF;KDR;KIT;RET
rigosertib_CTRP	rigosertib	BG	PIK3CA;PIK3CB;PLK1	PIK3CA;PIK3CB;PLK1
ruxolitinib_CTRP	ruxolitinib	BG	JAK1;JAK2	JAK1;JAK2
saracatinib_CTRP	saracatinib	BG	ABL1;SRC	ABL1;SRC
seliciclib_GDSC	Roscovitine	BG	CDKs	CDKs
selumetinib_CCLE	AZD6244	BG	MAP2K1;MAP2K2	MAP2K1;MAP2K2
selumetinib_CTRP	selumetinib	BG	MAP2K1;MAP2K2	MAP2K1;MAP2K2
selumetinib_in_navitoclax (8_in_1)_CTRP	selumetinib_in_navitoclax (8_in_1)	BG	MAP2K1;MAP2K2;BCL2;BCL2L1;BCL2L2	MAP2K1;MAP2K2;BCL2;BCL2L1;BCL2L2
semagacestat_CTRP	semagacestat	BG	APH1A;NCSTN;PSEN1;PSENEN	APH1A;NCSTN;PSEN1;PSENEN
serdemetan_CTRP	serdemetan	BG	MDM2	MDM2
sildenafil_CTRP	sildenafil	BG	PDE5A	PDE5A
simvastatin_CTRP	simvastatin	BG	HMGCR	HMGCR

Drug or Combination	Alternative name	FPS	Target or activity	Targets for Fisher's exact test
sirolimus_CTRP	sirolimus	BG	MTOR	MTOR
sirolimus_in_bortezomib (250_in_1)_CTRP	sirolimus_in_bortezomib (250_in_1)	BG	MTOR;PSMB1;PSMB2;PSMB5;PSMD1;PSMD2	MTOR;PSMB1;PSMB2;PSMB5;PSMD1;PSMD2
sorafenib_GDSC	Sorafenib	BG	BRAF;FLT3;KDR;RAF1	BRAF;FLT3;KDR;RAF1
sotrastaurin_CTRP	sotrastaurin	BG	PRKCB	PRKCB
tacrolimus_CTRP	tacrolimus	BG	PPP3CA;PPP3CB;PPP3CC;PPP3R1;PPP3R2	PPP3CA;PPP3CB;PPP3CC;PPP3R1;PPP3R2
tamatinib_CTRP	tamatinib	BG	SYK	SYK
tamoxifen_CTRP	tamoxifen	BG	ESR1;ESR2	ESR1;ESR2
tandutinib_CTRP	tandutinib	BG	FLT3;KIT	FLT3;KIT
tanespimycin_in_bortezomib (250_in_1)_CTRP	tanespimycin_in_bortezomib (250_in_1)	BG	HSP90AA1;PSMB1;PSMB2;PSMB5;PSMD1;PSMD2	HSP90AA1;PSMB1;PSMB2;PSMB5;PSMD1;PSMD2
tanespimycin_in_gemcitabine (1_in_1)_CTRP	tanespimycin_in_gemcitabine (1_in_1)	BG	HSP90AA1;CMPK1;RRM1;TYMS	HSP90AA1;CMPK1;RRM1;TYMS
temozolomide_CTRP	temozolomide	BG	DNA alkylator	DNA alkylator
temsirolimus_CTRP	temsirolimus	BG	MTOR	MTOR
teniposide_CTRP	teniposide	BG	TOP2A;TOP2B	TOP2A;TOP2B
tigecycline_CTRP	tigecycline	BG	analog of tetracycline	analog of tetracycline
tipifarnib-P1_CTRP	tipifarnib-P1	BG	FNTA	FNTA
tivantinib_CTRP	tivantinib	BG	MET	MET
tivozanib_CTRP	tivozanib	BG	FLT1;FLT3;KDR	FLT1;FLT3;KDR
topotecan_CCLE	Topotecan	BG	TOP1	TOP1
tosedostat_CTRP	tosedostat	BG	ANPEP;LAP3;NPEPPS	ANPEP;LAP3;NPEPPS
tretinoin_GDSC	ATRA	BG	RARA;RARB;RARG	RARA;RARB;RARG
tretinoin_in_carboplatin (2_in_1)_CTRP	tretinoin_in_carboplatin (2_in_1)	BG	RARA;RARB;RARG	RARA;RARB;RARG
triazolothiadiazine_CTRP	triazolothiadiazine	BG	PDE4A;PDE4B;PDE4D	PDE4A;PDE4B;PDE4D
triptolide_CTRP	triptolide	BG	natural product; inhibitor of RNA polymerase II	natural product;inhibitor of RNA polymerase II
tubastatin A_CTRP	tubastatin A	BG	HDAC6	HDAC6
valdecoxib_CTRP	valdecoxib	BG	PTGS2	PTGS2
veliparib_GDSC	ABT-888	BG	PARP1;PARP2	PARP1;PARP2
vorinostat_GDSC	Vorinostat	BG	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8	HDAC1;HDAC2;HDAC3;HDAC6;HDAC8
zebularine_CTRP	zebularine	BG	DNMT1	DNMT1

Table 6.1.9

Annotation of the CRC65 and CPTAC datasets

Annotation information with respect to cell lines from the CRC65 and tumours from the CPTAC dataset, compiled from several publications (see section 2.18 for a complete list of sources). These annotations were used in Fisher's exact tests for association with FPSs/KSs.

Sample name	Alternative Cell line name	TCGA Patient ID	TCGA mRNA Source ID	CPTAC sample ID	Dataset	CPTAC subtype	Transcriptomic subtype	Methylation subtype	CMS	CMS mRNA	CMS Protein	Marisa subtype	Budinska subtype	Roepman subtype	Sadanandam subtype	De Sousa E Melo subtype	Tumor stage
0A	NA	TCGA-AA-A02O	TCGA-AA-A02O-01A-21R-A16W-07	TCGA-AA-A02O-01A-23	Tumours	D	NA	CIMP-high	CMS3	CMS3	CMS3	NA	NA	NA	NA	CCS1	IIA
0B	NA	TCGA-AA-3534	TCGA-AA-3534-01A-01R-0821-07	TCGA-AA-3534-01A-22	Tumours	C	CIN	Cluster 3	CMS2	CMS2	NA	NA	NA	NA	TA	CCS1	IIA
0C	NA	TCGA-AA-3818	TCGA-AA-3818-01A-01R-0905-07	TCGA-AA-3818-01A-22	Tumours	C	Invasive	CIMP-low	CMS3	CMS3	CMS4	NA	NA	NA	Enterocyte/Goblet-like	CCS1	IIA
0D	NA	TCGA-AG-A02X	TCGA-AG-A02X-01A-01R-A00A-07	TCGA-AG-A02X-01A-32	Tumours	D	Invasive	CIMP-high	CMS3	CMS3	CMS3	NA	NA	NA	Enterocyte/Goblet-like	CCS2	I
0E	NA	TCGA-AG-A020	TCGA-AG-A020-01A-21R-A083-07	TCGA-AG-A020-01A-23	Tumours	D	NA	NA	CMS3	CMS3	CMS3	NA	NA	NA	NA	CCS1	IIIB
1A	NA	TCGA-AF-2691	TCGA-AF-2691-01A-01R-0821-07	TCGA-AF-2691-01A-41	Tumours	E	Invasive	CIMP-low	CMS4	CMS4	NA	NA	NA	NA	Enterocyte/Goblet-like	CCS3	I
1B	NA	TCGA-AA-A01F	TCGA-AA-A01F-01A-01R-A002-07	TCGA-AA-A01F-01A-23	Tumours	E	CIN	Cluster 3	CMS2	CMS2	CMS2	NA	NA	NA	Enterocyte/Goblet-like	CCS1	IIIB
1C	NA	TCGA-AG-A00C	TCGA-AG-A00C-01A-01R-A002-07	TCGA-AG-A00C-01A-23	Tumours	E	Invasive	Cluster 3	CMS2	CMS2	CMS2	NA	NA	NA	TA	CCS1	IIIB
1D	NA	TCGA-AG-A016	TCGA-AG-A016-01A-01R-A002-07	TCGA-AG-A016-01A-23	Tumours	NA	CIN	Cluster 3	CMS2	CMS2	CMS2	NA	NA	NA	TA	CCS1	IV
1E	NA	TCGA-AG-4007	TCGA-AG-4007-01A-01R-1119-07	TCGA-AG-4007-01A-22	Tumours	NA	CIN	CIMP-low	NA	CMS4	CMS4	NA	NA	NA	TA	CCS1	IV
1F	NA	TCGA-AA-A00J	TCGA-AA-A00J-01A-02R-A002-07	TCGA-AA-A00J-01A-12	Tumours	NA	MSI/CIMP	CIMP-high	CMS1	CMS1	CMS1	NA	NA	NA	Enterocyte/Goblet-like	CCS2	IIIB
1G	NA	TCGA-AG-A036	TCGA-AG-A036-01A-12R-A083-07	TCGA-AG-A036-01A-22	Tumours	NA	NA	NA	CMS2	CMS2	NA	NA	NA	NA	NA	CCS1	IIIB
1H	NA	TCGA-AA-A01X	TCGA-AA-A01X-01A-21R-A083-07	TCGA-AA-A01X-01A-23	Tumours	E	NA	Cluster 3	CMS2	CMS4	NA	NA	NA	NA	NA	CCS1	IIIA
1I	NA	TCGA-AA-3561	TCGA-AA-3561-01A-01R-0821-07	TCGA-AA-3561-01A-22	Tumours	E	Invasive	Cluster 3	CMS2	NA	CMS2	NA	NA	NA	Stem-like	CCS1	IIA
1J	NA	TCGA-AA-A03J	TCGA-AA-A03J-01A-21R-A16W-07	TCGA-AA-A03J-01A-23	Tumours	E	NA	CIMP-low	CMS2	NA	CMS2	NA	NA	NA	NA	CCS1	I

Sample name	Alternative Cell line name	TCGA Patient ID	TCGA mRNA Source ID	CPTAC sample ID	Dataset	CPTAC subtype	Transcriptomic subtype	Methylation subtype	CMS	CMS mRNA	CMS Protein	Marisa subtype	Budinska subtype	Roepman subtype	Sadanandam subtype	De Sousa E Melo subtype	Tumor stage
2A	NA	TCGA-AA-A00O	TCGA-AA-A00O-01A-02R-A089-07	TCGA-AA-A00O-01A-13	Tumours	NA	CIN	Cluster 4	CMS4	CMS4	CMS4	NA	NA	NA	Stem-like	CCS3	IIIC
2B	NA	TCGA-AA-A017	TCGA-AA-A017-01A-01R-A00A-07	TCGA-AA-A017-01A-22	Tumours	NA	CIN	Cluster 4	CMS2	CMS2	CMS2	NA	NA	NA	Stem-like	CCS1	IIA
2C	NA	TCGA-AG-A02N	TCGA-AG-A02N-01A-11R-A083-07	TCGA-AG-A02N-01A-31	Tumours	E	NA	NA	CMS3	CMS3	CMS3	NA	NA	NA	NA	CCS2	IIA
2D	NA	TCGA-AF-3400	TCGA-AF-3400-01A-01R-0821-07	TCGA-AF-3400-01A-41	Tumours	C	MSI/CIMP	Cluster 4	CMS4	CMS4	CMS4	NA	NA	NA	Stem-like	CCS3	IIA
2E	NA	TCGA-AA-3715	TCGA-AA-3715-01A-01R-0905-07	TCGA-AA-3715-01A-22	Tumours	C	MSI/CIMP	CIMP-high	CMS1	CMS1	CMS4	NA	NA	NA	Inflammatory	CCS3	IIA
2F	NA	TCGA-AG-A014	TCGA-AG-A014-01A-02R-A002-07	TCGA-AG-A014-01A-12	Tumours	E	CIN	Cluster 4	CMS2	CMS2	NA	NA	NA	NA	TA	CCS1	I
2G	NA	TCGA-AG-A01Y	TCGA-AG-A01Y-01A-41R-A083-07	TCGA-AG-A01Y-01A-43	Tumours	C	NA	NA	CMS4	NA	NA	NA	NA	NA	NA	CCS3	IIA
2H	NA	TCGA-AA-3710	TCGA-AA-3710-01A-01R-1022-07	TCGA-AA-3710-01A-22	Tumours	B	MSI/CIMP	CIMP-high	CMS1	CMS1	CMS1	NA	NA	NA	Inflammatory	CCS2	IIA
2I	NA	TCGA-AG-A011	TCGA-AG-A011-01A-01R-A002-07	TCGA-AG-A011-01A-32	Tumours	E	Invasive	CIMP-low	CMS2	NA	NA	NA	NA	NA	Enterocyte/Goblet-like	CCS1	IIA
2J	NA	TCGA-AA-A01R	TCGA-AA-A01R-01A-21R-A083-07	TCGA-AA-A01R-01A-23	Tumours	B	NA	CIMP-low	CMS1	CMS1	CMS1	NA	NA	NA	NA	CCS2	IIIB
3A	NA	TCGA-AA-A01Z	TCGA-AA-A01Z-01A-11R-A083-07	TCGA-AA-A01Z-01A-13	Tumours	A	NA	CIMP-low	NA	CMS2	NA	NA	NA	NA	NA	CCS1	IIA
3B	NA	TCGA-AA-A02E	TCGA-AA-A02E-01A-01R-A00A-07	TCGA-AA-A02E-01A-23	Tumours	A	CIN	Cluster 3	CMS2	NA	CMS4	NA	NA	NA	Stem-like	CCS1	IV
3C	NA	TCGA-AA-A02Y	NA	TCGA-AA-A02Y-01A-31	Tumours	A	NA	CIMP-low	NA	NA	CMS3	NA	NA	NA	NA	NA	I
3D	NA	TCGA-AF-2692	TCGA-AF-2692-01A-01R-0821-07	TCGA-AF-2692-01A-41	Tumours	C	CIN	Cluster 4	CMS2	CMS2	CMS4	NA	NA	NA	TA	CCS1	IIA
3F	NA	TCGA-AA-3518	TCGA-AA-3518-01A-02R-0826-07	TCGA-AA-3518-01A-11	Tumours	C	MSI/CIMP	CIMP-high	CMS1	CMS1	CMS1	NA	NA	NA	Enterocyte/Goblet-like	CCS2	IIA
3G	NA	TCGA-AG-3593	TCGA-AG-3593-01A-01R-0821-07	TCGA-AG-3593-01A-22	Tumours	E	CIN	Cluster 3	CMS2	CMS2	CMS2	NA	NA	NA	TA	CCS1	IIA
3H	NA	TCGA-AG-A015	TCGA-AG-A015-01A-01R-A002-07	TCGA-AG-A015-01A-51	Tumours	A	Invasive	Cluster 3	CMS3	CMS3	CMS3	NA	NA	NA	Enterocyte/Goblet-like	CCS1	I
3I	NA	TCGA-AA-A03F	TCGA-AA-A03F-01A-11R-A16W-07	TCGA-AA-A03F-01A-41	Tumours	A	NA	CIMP-high	CMS3	CMS3	CMS3	NA	NA	NA	NA	CCS1	IIIB
3J	NA	TCGA-AA-A010	TCGA-AA-A010-01A-01R-A089-07	TCGA-AA-A010-01A-31	Tumours	C	MSI/CIMP	Cluster 4	NA	CMS1	CMS3	NA	NA	NA	Enterocyte/Goblet-like	CCS1	IIB
4A	NA	TCGA-AG-3584	TCGA-AG-3584-01A-01R-0821-07	TCGA-AG-3584-01A-22	Tumours	A	CIN	CIMP-low	CMS2	CMS2	NA	NA	NA	NA	TA	CCS1	IV

Sample name	Alternative Cell line name	TCGA Patient ID	TCGA mRNA Source ID	CPTAC sample ID	Dataset	CPTAC subtype	Transcriptomic subtype	Methylation subtype	CMS	CMS mRNA	CMS Protein	Marisa subtype	Budinska subtype	Roepman subtype	Sadanandam subtype	De Sousa E Melo subtype	Tumor stage
4C	NA	TCGA-AA-A029	TCGA-AA-A029-01A-01R-A00A-07	TCGA-AA-A029-01A-41	Tumours	A	Invasive	CIMP-high	CMS3	CMS3	CMS3	NA	NA	NA	Enterocyte/Goblet-like	CCS1	IIA
4D	NA	TCGA-AG-3594	TCGA-AG-3594-01A-02R-0821-07	TCGA-AG-3594-01A-12	Tumours	NA	MSI/CIMP	Cluster 3	CMS3	NA	CMS1	NA	NA	NA	Inflammatory	CCS2	IIA
4E	NA	TCGA-AA-A00E	TCGA-AA-A00E-01A-01R-A002-07	TCGA-AA-A00E-01A-31	Tumours	NA	MSI/CIMP	CIMP-high	CMS1	CMS1	CMS1	NA	NA	NA	Inflammatory	CCS2	IIA
4F	NA	TCGA-AA-A01C	TCGA-AA-A01C-01A-01R-A00A-07	TCGA-AA-A01C-01A-22	Tumours	A	CIN	Cluster 4	CMS4	CMS2	NA	NA	NA	NA	Stem-like	CCS1	IIIA
4G	NA	TCGA-AG-A00Y	TCGA-AG-A00Y-01A-02R-A002-07	TCGA-AG-A00Y-01A-12	Tumours	A	Invasive	Cluster 3	CMS2	CMS2	CMS2	NA	NA	NA	Inflammatory	CCS1	IIA
4H	NA	TCGA-AA-3684	TCGA-AA-3684-01A-02R-0905-07	TCGA-AA-3684-01A-31	Tumours	C	MSI/CIMP	CIMP-low	CMS4	CMS4	CMS4	NA	NA	NA	Stem-like	CCS3	IV
4I	NA	TCGA-A6-3807	TCGA-A6-3807-01A-01R-1022-07	TCGA-A6-3807-01A-22	Tumours	C	MSI/CIMP	Cluster 4	CMS4	CMS4	NA	NA	NA	NA	Stem-like	CCS3	IIIC
4J	NA	TCGA-AA-A004	TCGA-AA-A004-01A-01R-A00A-07	TCGA-AA-A004-01A-22	Tumours	C	MSI/CIMP	CIMP-low	NA	CMS4	CMS3	NA	NA	NA	Inflammatory	CCS1	IIA
5A	NA	TCGA-AG-A026	TCGA-AG-A026-01A-01R-A00A-07	TCGA-AG-A026-01A-71	Tumours	E	NA	Cluster 3	CMS2	NA	CMS2	NA	NA	NA	NA	CCS1	NA
5B	NA	TCGA-AA-A02H	TCGA-AA-A02H-01A-01R-A089-07	TCGA-AA-A02H-01A-32	Tumours	E	CIN	Cluster 3	CMS2	NA	CMS2	NA	NA	NA	TA	CCS1	IV
5C	NA	TCGA-AF-3913	TCGA-AF-3913-01A-02R-1119-07	TCGA-AF-3913-01A-12	Tumours	C	CIN	Cluster 4	CMS4	CMS2	CMS4	NA	NA	NA	Stem-like	CCS1	IV
5D	NA	TCGA-AA-A00N	TCGA-AA-A00N-01A-02R-A00A-07	TCGA-AA-A00N-01A-32	Tumours	C	MSI/CIMP	Cluster 4	CMS4	CMS4	CMS4	NA	NA	NA	Stem-like	CCS3	IIB
5E	NA	TCGA-A6-3810	TCGA-A6-3810-01A-01R-1022-07	TCGA-A6-3810-01A-22	Tumours	C	MSI/CIMP	Cluster 4	CMS4	CMS4	CMS4	NA	NA	NA	Stem-like	CCS3	IIA
5F	NA	TCGA-AA-3666	TCGA-AA-3666-01A-02R-0905-07	TCGA-AA-3666-01A-31	Tumours	E	Invasive	Cluster 4	NA	NA	CMS2	NA	NA	NA	Enterocyte/Goblet-like	CCS1	IIIB
5G	NA	TCGA-AA-A00R	TCGA-AA-A00R-01A-01R-A002-07	TCGA-AA-A00R-01A-22	Tumours	NA	MSI/CIMP	CIMP-low	CMS1	CMS1	CMS1	NA	NA	NA	Inflammatory	CCS2	I
5H	NA	TCGA-AA-A02R	TCGA-AA-A02R-01A-01R-A00A-07	TCGA-AA-A02R-01A-23	Tumours	C	MSI/CIMP	CIMP-high	CMS1	CMS1	CMS1	NA	NA	NA	Inflammatory	CCS2	IIA
5I	NA	TCGA-AA-A01D	TCGA-AA-A01D-01A-01R-A00A-07	TCGA-AA-A01D-01A-23	Tumours	C	MSI/CIMP	CIMP-low	CMS4	CMS4	CMS4	NA	NA	NA	Stem-like	CCS3	IIIC
5J	NA	TCGA-AG-A032	TCGA-AG-A032-01A-01R-A00A-07	TCGA-AG-A032-01A-31	Tumours	E	CIN	Cluster 4	CMS2	CMS2	CMS2	NA	NA	NA	TA	CCS1	IIIB
6A	NA	TCGA-AA-3558	NA	TCGA-AA-3558-01A-22	Tumours	E	Invasive	CIMP-low	NA	CMS2	CMS2	NA	NA	NA	TA	CCS1	NA
6B	NA	TCGA-AA-3989	TCGA-AA-3989-01A-01R-1022-07	TCGA-AA-3989-01A-22	Tumours	E	Invasive	CIMP-low	CMS2	CMS2	CMS2	NA	NA	NA	TA	CCS1	IV

Sample name	Alternative Cell line name	TCGA Patient ID	TCGA mRNA Source ID	CPTAC sample ID	Dataset	CPTAC subtype	Transcriptomic subtype	Methylation subtype	CMS	CMS mRNA	CMS Protein	Marisa subtype	Budinska subtype	Roepman subtype	Sadanandam subtype	De Sousa E Melo subtype	Tumor stage
6C	NA	TCGA-AG-A01W	TCGA-AG-A01W-01A-21R-A083-07	TCGA-AG-A01W-01A-23	Tumours	E	NA	NA	CMS2	CMS2	NA	NA	NA	NA	NA	CCS1	IIA
6D	NA	TCGA-AA-A01K	TCGA-AA-A01K-01A-01R-A00A-07	TCGA-AA-A01K-01A-31	Tumours	C	MSI/CIMP	Cluster 3	CMS4	CMS4	CMS4	NA	NA	NA	Stem-like	CCS3	IIIC
6E	NA	TCGA-AA-A01I	TCGA-AA-A01I-01A-02R-A089-07	TCGA-AA-A01I-01A-12	Tumours	A	Invasive	CIMP-low	CMS2	CMS2	CMS2	NA	NA	NA	TA	CCS1	I
6F	NA	TCGA-AA-A01S	TCGA-AA-A01S-01A-21R-A083-07	TCGA-AA-A01S-01A-23	Tumours	A	NA	Cluster 4	CMS2	CMS2	CMS2	NA	NA	NA	NA	CCS1	IIIB
6G	NA	TCGA-AA-3554	TCGA-AA-3554-01A-01R-0826-07	TCGA-AA-3554-01A-22	Tumours	C	MSI/CIMP	Cluster 4	CMS1	CMS1	NA	NA	NA	NA	Inflammatory	CCS3	IIA
6H	NA	TCGA-AA-A01V	TCGA-AA-A01V-01A-23R-A083-07	TCGA-AA-A01V-01A-24	Tumours	C	NA	CIMP-low	CMS3	CMS3	NA	NA	NA	NA	NA	CCS1	I
6I	NA	TCGA-AA-A02J	TCGA-AA-A02J-01A-01R-A00A-07	TCGA-AA-A02J-01A-23	Tumours	NA	CIN	Cluster 4	CMS2	CMS2	CMS2	NA	NA	NA	TA	CCS1	IV
7A	NA	TCGA-AG-A00H	TCGA-AG-A00H-01A-01R-A00A-07	TCGA-AG-A00H-01A-22	Tumours	C	Invasive	Cluster 4	CMS4	CMS4	CMS4	NA	NA	NA	Enterocyte/Goblet-like	CCS3	IIA
7B	NA	TCGA-AA-3529	TCGA-AA-3529-01A-02R-0821-07	TCGA-AA-3529-01A-12	Tumours	C	CIN	Cluster 4	CMS2	CMS2	CMS4	NA	NA	NA	TA	CCS1	IIIC
7C	NA	TCGA-AA-3848	TCGA-AA-3848-01A-01R-0905-07	TCGA-AA-3848-01A-22	Tumours	E	CIN	CIMP-low	CMS2	CMS2	CMS2	NA	NA	NA	TA	CCS1	IIIC
7D	NA	TCGA-A6-3808	TCGA-A6-3808-01A-01R-1022-07	TCGA-A6-3808-01A-22	Tumours	C	MSI/CIMP	Cluster 3	CMS4	CMS4	CMS4	NA	NA	NA	Stem-like	CCS3	IIA
7E	NA	TCGA-AG-3574	TCGA-AG-3574-01A-01R-0821-07	TCGA-AG-3574-01A-22	Tumours	C	CIN	Cluster 3	CMS2	CMS2	NA	NA	NA	NA	TA	CCS1	IIA
7F	NA	TCGA-AA-3525	TCGA-AA-3525-01A-02R-0826-07	TCGA-AA-3525-01A-12	Tumours	B	MSI/CIMP	CIMP-high	CMS1	CMS1	NA	NA	NA	NA	Enterocyte/Goblet-like	CCS2	IIIB
7G	NA	TCGA-AG-A002	TCGA-AG-A002-01A-01R-A002-07	TCGA-AG-A002-01A-23	Tumours	B	MSI/CIMP	Cluster 3	CMS1	CMS1	CMS1	NA	NA	NA	Enterocyte/Goblet-like	CCS2	I
7H	NA	TCGA-AA-A00K	TCGA-AA-A00K-01A-02R-A002-07	TCGA-AA-A00K-01A-12	Tumours	B	Invasive	Cluster 3	CMS3	CMS2	NA	NA	NA	NA	TA	CCS1	IIA
7I	NA	TCGA-AA-A01P	TCGA-AA-A01P-01A-21R-A083-07	TCGA-AA-A01P-01A-23	Tumours	B	NA	CIMP-high	CMS1	CMS1	CMS1	NA	NA	NA	NA	CCS3	IIIB
7J	NA	TCGA-AA-3672	TCGA-AA-3672-01A-01R-0905-07	TCGA-AA-3672-01A-22	Tumours	B	MSI/CIMP	CIMP-high	CMS1	CMS1	CMS1	NA	NA	NA	Inflammatory	CCS2	IIIB
8A	NA	TCGA-AA-A00A	TCGA-AA-A00A-01A-01R-A002-07	TCGA-AA-A00A-01A-22	Tumours	B	MSI/CIMP	CIMP-high	CMS1	NA	NA	NA	NA	NA	Inflammatory	CCS2	IIA
8B	NA	TCGA-AA-A00F	TCGA-AA-A00F-01A-01R-A002-07	TCGA-AA-A00F-01A-31	Tumours	D	CIN	Cluster 3	CMS4	NA	NA	NA	NA	NA	Stem-like	CCS3	IIIC
8C	NA	TCGA-AG-3580	TCGA-AG-3580-01A-01R-0821-07	TCGA-AG-3580-01A-22	Tumours	E	Invasive	Cluster 3	CMS2	CMS2	CMS2	NA	NA	NA	TA	CCS1	I

Sample name	Alternative Cell line name	TCGA Patient ID	TCGA mRNA Source ID	CPTAC sample ID	Dataset	CPTAC subtype	Transcriptomic subtype	Methylation subtype	CMS	CMS mRNA	CMS Protein	Marisa subtype	Budinska subtype	Roepman subtype	Sadanandam subtype	De Sousa E Melo subtype	Tumor stage
8D	NA	TCGA-AA-3864	TCGA-AA-3864-01A-01R-1022-07	TCGA-AA-3864-01A-22	Tumours	NA	MSI/CIMP	Cluster 3	CMS1	CMS1	CMS3	NA	NA	NA	Enterocyte/Goblet-like	CCS3	IIA
8F	NA	TCGA-AG-A01J	TCGA-AG-A01J-01A-01R-A00A-07	TCGA-AG-A01J-01A-22	Tumours	C	CIN	Cluster 4	CMS2	CMS2	CMS4	NA	NA	NA	TA	CCS1	IIA
8G	NA	TCGA-AA-3531	TCGA-AA-3531-01A-01R-0821-07	TCGA-AA-3531-01A-22	Tumours	A	CIN	Cluster 4	CMS2	CMS2	CMS2	NA	NA	NA	TA	CCS1	IIA
8H	NA	TCGA-AA-A01T	TCGA-AA-A01T-01A-21R-A16W-07	TCGA-AA-A01T-01A-23	Tumours	A	NA	Cluster 3	CMS2	CMS2	CMS3	NA	NA	NA	NA	CCS1	IIIB
8I	NA	TCGA-AG-A01N	TCGA-AG-A01N-01A-01R-A00A-07	TCGA-AG-A01N-01A-23	Tumours	A	CIN	Cluster 4	CMS2	CMS2	CMS2	NA	NA	NA	TA	CCS1	IV
8J	NA	TCGA-AA-3526	NA	TCGA-AA-3526-01A-11	Tumours	A	MSI/CIMP	Cluster 3	NA	CMS3	NA	NA	NA	NA	Enterocyte/Goblet-like	NA	I
9A	NA	TCGA-AA-A024	TCGA-AA-A024-01A-02R-A00A-07	TCGA-AA-A024-01A-32	Tumours	D	Invasive	Cluster 3	CMS3	CMS3	CMS3	NA	NA	NA	Enterocyte/Goblet-like	CCS1	IIA
9B	NA	TCGA-AA-3986	TCGA-AA-3986-01A-02R-1022-07	TCGA-AA-3986-01A-12	Tumours	C	Invasive	Cluster 3	CMS2	CMS2	CMS4	NA	NA	NA	Inflammatory	CCS1	I
9C	NA	TCGA-AA-3664	TCGA-AA-3664-01A-01R-0905-07	TCGA-AA-3664-01A-22	Tumours	D	MSI/CIMP	CIMP-high	CMS1	CMS1	NA	NA	NA	NA	Enterocyte/Goblet-like	CCS2	IIA
9D	NA	TCGA-AA-A022	TCGA-AA-A022-01A-21R-A16W-07	TCGA-AA-A022-01A-23	Tumours	B	NA	CIMP-high	CMS1	CMS1	CMS1	NA	NA	NA	NA	CCS2	NA
9F	NA	TCGA-AA-3552	TCGA-AA-3552-01A-01R-0821-07	TCGA-AA-3552-01A-22	Tumours	D	Invasive	Cluster 3	CMS2	CMS4	NA	NA	NA	NA	TA	CCS1	IIIC
9G	NA	TCGA-AG-A008	TCGA-AG-A008-01A-01R-A002-07	TCGA-AG-A008-01A-23	Tumours	D	Invasive	Cluster 4	CMS3	CMS3	CMS3	NA	NA	NA	Enterocyte/Goblet-like	CCS1	I
9H	NA	TCGA-AG-A01L	TCGA-AG-A01L-01A-01R-A002-07	TCGA-AG-A01L-01A-22	Tumours	D	Invasive	Cluster 3	CMS2	CMS2	NA	NA	NA	NA	Enterocyte/Goblet-like	CCS1	IIIB
9I	NA	TCGA-AA-3695	TCGA-AA-3695-01A-01R-0905-07	TCGA-AA-3695-01A-22	Tumours	D	CIN	CIMP-low	NA	NA	CMS4	NA	NA	NA	TA	CCS2	IV
9J	NA	TCGA-AA-A00U	TCGA-AA-A00U-01A-01R-A002-07	TCGA-AA-A00U-01A-41	Tumours	D	Invasive	CIMP-low	CMS3	CMS2	NA	NA	NA	NA	Enterocyte/Goblet-like	CCS1	IIIB
C10	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	CMS4	C4	Subtype D	C-Type	Stem-like	CCS3	NA
C106	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	NA	C1	Subtype E	B-Type	TA	CCS1	NA
C125-PM	C125PM	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	CMS3	C5	Subtype B	B-Type	Enterocyte/Goblet-like	CCS1	NA
C32	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS4	CMS1	C6	Subtype E	B-Type	Enterocyte/Goblet-like	CCS1	NA
C70	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	CMS2	C1	Subtype B	B-Type	TA	CCS1	NA
C75	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	NA	C1	Subtype B	B-Type	TA	CCS1	NA
C80	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	CMS2	C3	Subtype B	B-Type	Enterocyte/Goblet-like	CCS1	NA
C84	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	CMS3	C1	Subtype E	B-Type	Enterocyte/Goblet-like	CCS1	NA
C99	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	CMS3	C3	Subtype A	B-Type	Enterocyte/Goblet-like	CCS1	NA
CaCo-2	C2BBe1; Caco-2; CACO2	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	CMS4	C4	Subtype D	B-Type	Stem-like	CCS1	NA
CaR-1	CAR1	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	CMS4	C4	Subtype E	C-Type	Stem-like	CCS3	NA
CC07	CCO7	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CC20	CC2O	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS4	NA	NA	NA	NA	NA	NA	NA

Sample name	Alternative Cell line name	TCGA Patient ID	TCGA mRNA Source ID	CPTAC sample ID	Dataset	CPTAC subtype	Transcriptomic subtype	Methylation subtype	CMS	CMS mRNA	CMS Protein	Marisa subtype	Budinska subtype	Roepman subtype	Sadanandam subtype	De Sousa E Melo subtype	Tumor stage
CCK-81	CCK81	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	NA	C1	Subtype B	B-Type	TA	CCS2	NA
CoCM-1	COCM1	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	NA	C6	Subtype B	B-Type	TA	CCS1	NA
Colo 320DM	Colo 320; COLO320; COLO-320; COLO320DM; COLO 320DM; COLO320-DM; COLO320HSR; COLO-320-HSR	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	C4	Subtype D	C-Type	Stem-like	CCS3	NA
Colo-678	COLO 678; COLO678; COLO-678	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS4	NA	C4	Subtype D	B-Type	Inflammatory	CCS3	NA
Colo 741	COLO 741; COLO741; COLO-741	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS4	CMS4	NA	NA	NA	NA	NA	NA
DLD-1	DLD1; DLD-1 DKO1; DLD-1 DKO3; DLD-1 DKO4; C170; HCT-8; HCT8; HCT-15; HCT15; C146; MIP101	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	NA	C4	Subtype D	C-Type	Stem-like	CCS3	NA
GP2d	GP2D; GP5d; Gp5D	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	C2	Subtype B	A-Type	TA	CCS1	NA
HCA-46	HCA46	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	CMS2	C1	Subtype A	B-Type	TA	CCS1	NA
HCA-7	HCA7	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	NA	C2	Subtype D	C-Type	Stem-like	CCS2	NA
HCT 116	ATRFLOX; HCT116; HCT 116 HKH2; HCT 116 TSP531026; HCT-116	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	NA	C4	Subtype D	C-Type	Stem-like	CCS3	NA
HDC-111	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	CMS3	NA	NA	NA	NA	NA	NA
HDC-114	HDC114	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	NA	C2	Subtype C	A-Type	Inflammatory	CCS2	NA
HDC-135	HDC135	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	CMS1	C2	Subtype C	A-Type	Inflammatory	CCS2	NA
HDC-142	HDC142	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	NA	C5	Subtype B	B-Type	TA	CCS1	NA
HDC-143	HDC143	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS3	CMS3	C6	Subtype B	B-Type	Enterocyte/Goblet-like	CCS3	NA
HDC-54	HDC54; HDC-57	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	CMS2	C6	Subtype B	B-Type	Enterocyte/Goblet-like	CCS1	NA
HDC-57	HDC-54; HDC54	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
HDC-73	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
HDC-8	HDC8	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	NA	C4	Subtype C	A-Type	Stem-like	CCS3	NA
HDC-82	HDC82	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	NA	C5	Subtype B	B-Type	TA	CCS1	NA
HDC-9	HDC9	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS3	NA	C3	Subtype A	C-Type	Enterocyte/Goblet-like	CCS2	NA
HRA-19	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	NA	NA	NA	NA	NA	NA	NA
HT-29	HT-29 GLUCC1; CX-1; CX1; WiDr; WIDR; KM20	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	CMS1	C1	Subtype C	A-Type	Enterocyte/Goblet-like	CCS3	NA
HT55	HT-55	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	CMS2	C1	Subtype B	B-Type	TA	CCS1	NA
LIM1863	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	CMS2	NA	NA	NA	NA	NA	NA
LoVo	LOVO	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	CMS1	C2	Subtype E	A-Type	Inflammatory	CCS2	NA
LS1034	LS-1034	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	CMS2	C1	Subtype B	B-Type	TA	CCS1	NA
LS123	LS-123	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	CMS4	C4	Subtype C	C-Type	Inflammatory	CCS3	NA
LS 174T	LS174T; LS 180; LS180; HM7	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	C3	Subtype A	A-Type	Enterocyte/Goblet-like	CCS2	NA
LS 180	LS180; LS 174T; LS174T; HM7	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS3	NA	C2	Subtype A	C-Type	Enterocyte/Goblet-like	CCS2	NA

Sample name	Alternative Cell line name	TCGA Patient ID	TCGA mRNA Source ID	CPTAC sample ID	Dataset	CPTAC subtype	Transcriptomic subtype	Methylation subtype	CMS	CMS mRNA	CMS Protein	Marisa subtype	Budinska subtype	Roepman subtype	Sadanandam subtype	De Sousa E Melo subtype	Tumor stage
LS411	LS411N; LS-411N	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	NA	C2	Subtype B	A-Type	Inflammatory	CCS2	NA
LS513	LS-513	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS3	CMS3	C3	Subtype A	B-Type	Enterocyte/Goblet-like	CCS1	NA
NCI-H548	COLO 201; COLO201; COLO 205; COLO205; COLO-205; COLO 206F	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	C5	Subtype C	C-Type	TA	CCS2	NA
NCI-H716	NCIH716	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	C4	Subtype D	C-Type	Stem-like	CCS3	NA
NCI-H747	NCIH747	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	C4	Subtype D	C-Type	Inflammatory	CCS3	NA
OXCO-1	OXCO1	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	CMS4	C4	Subtype D	C-Type	Stem-like	CCS3	NA
OXCO-3	OXCO3	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	NA	C5	Subtype C	A-Type	Inflammatory	CCS3	NA
PC/JW	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
RCM-1	RCM1	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	CMS2	C1	Subtype E	B-Type	TA	CCS1	NA
RKO	RKO-E6; RKO AS451; RKO E6; RKO AS45	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	CMS1	C2	Subtype E	C-Type	Inflammatory	CCS2	NA
SK-CO-1	SKCO1	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS3	NA	C3	Subtype A	C-Type	Enterocyte/Goblet-like	CCS1	NA
SNU-C2B	SNUC2B	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	NA	C2	Subtype C	A-Type	Enterocyte/Goblet-like	CCS2	NA
SW1116	SW 1116	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	C5	Subtype B	B-Type	TA	CCS1	NA
SW1417	SW 1417	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	C5	Subtype C	B-Type	TA	CCS1	NA
SW403	SW 403	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	CMS2	C5	Subtype B	B-Type	TA	CCS1	NA
SW48	SW 48	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	CMS1	C2	Subtype C	A-Type	Inflammatory	CCS2	NA
SW480	SW 480	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	C4	Subtype D	C-Type	Stem-like	CCS3	NA
SW620	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	CMS4	C5	Subtype D	C-Type	TA	CCS3	NA
SW837	SW 837	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS1	NA	C5	Subtype C	C-Type	Inflammatory	CCS3	NA
SW948	SW 948	NA	NA	NA	Cell lines	NA	NA	NA	NA	CMS2	CMS3	C5	Subtype B	B-Type	TA	CCS1	NA
T84	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	CMS3	C2	Subtype C	B-Type	Inflammatory	CCS1	NA
VACO 4A	NA	NA	NA	NA	Cell lines	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA