

Predictive modelling of cancer chromosomal instability

Xiaoxiao Zhang

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Julien Gagneur

Prüfer*innen der Dissertation:

1. Prof. Dr. Björn Menze
2. Prof. Dr. Holger Bastians

Die Dissertation wurde am 29.06.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 22.02.2023 angenommen.



Abstract

Chromosomal instability (CIN), a phenotype describing elevated rates of gaining or losing whole chromosomes (W-CIN) or of accumulating structurally abnormal chromosomes (S-CIN), often leads to profound tumour heterogeneity. As a result, CIN enables fast adaptation to selective pressure imposed by the tumour microenvironment during evolution. Previous studies have suggested that CIN is associated with cancer progression, treatment resistance and clinical outcomes. Although W-CIN and S-CIN can co-occur, they have distinct causes. W-CIN results from chromosome missegregations, S-CIN arises due to unfixed DNA double-strand breaks errors. A better understanding of the CIN causes and consequences distinguishing W-CIN versus S-CIN will promote translating the widely noted CIN clinical and evolutionary impacts to therapy strategies.

The development of CIN is complex involving multi-level regulations. In particular, the CIN gene alterations give rise to pathway activity and tumour microenvironment changes, directly or indirectly leading to CIN. In turn, the CIN phenotype together with selective pressure continuously drives the ongoing changes at genetic, pathway and cellular levels. To decipher such complexity, large scale multi-omic data have been accumulated. Integrative predictive models are urgently needed to address the substantial complexity.

If we could gain a better understanding of the CIN process, strategies targeting CIN could be discovered and validated in preclinical models in an ideal case. However, these strategies may still not work effectively in the clinics due to the discrepancy of molecular bases and tumour microenvironment between preclinical models and patients. Predictive models matching the two may help improve the translational capacity.

This thesis first answers the important research question: What are the commonalities and differences between W-CIN and S-CIN? This is done by integrating multi-omic data to characterise the association landscape of W-CIN and S-CIN across 33 cancer types. Extensive correlation analyses have been performed between W-CIN/S-CIN degrees and various characteristics including prognosis, drug sensitivities, pathway activities, somatic point mutations and copy number variations. Our model predicts that the gene *GINS1* has a W-CIN promoting role, which has later been experimentally validated. This suggests the predictive model proposed in this thesis is valuable to understand CIN mechanisms.

Given that the existing computational CIN studies do not distinguish W-CIN from S-CIN, this work thus serves as a complement to these studies to advance our understanding of CIN.

Cancer cell lines are widely used to understand CIN process and to develop new anti-cancer treatments. Selecting the most appropriate cell lines for a given tumour is essential to translate promising results from the *in vitro* experiments to clinical applications. This thesis thus proposes a semi-supervised generative model, MFmap (model fidelity map), matching cell lines to tumours and cancer subtypes, intending to maximise the translational ability of oncological *in vitro* models. MFmap compresses high-dimensional multi-omic data into a small set of cancer subtype specific features and predicts the subtype labels of cell lines, combining very good classification and generative performance. The MFmap embedded features can be used to calculate the pairwise cell-line-tumour similarity, with which one can select the best cell lines for a group of tumours or even individual tumours. By classifying cancer cell lines into subtypes, MFmap allows to predict the subtype specific effect of therapeutic compounds. Additionally, MFmap can be used to study tumour evolutionary processes during the disease course. Taken together, MFmap is useful for a broad range of semi-supervised tasks in the biomedical field where data labelling is difficult to obtain.



Zusammenfassung

Chromosomale Instabilität (CIN), ein Phänotyp, der eine erhöhte Rate des Gewinns oder Verlusts ganzer Chromosomen (W-CIN) oder der Anhäufung strukturell abnormaler Chromosomen (S-CIN) beschreibt, führt häufig zu einer ausgeprägten Tumorerogenität. Infolgedessen ermöglicht CIN eine schnelle Anpassung an den Selektionsdruck, den die Mikroumgebung des Tumors im Laufe der Evolution ausübt. Frühere Studien haben gezeigt, dass CIN mit dem Fortschreiten des Krebses, der Therapieresistenz und dem klinischen Erfolg in Verbindung steht. Obwohl W-CIN und S-CIN gemeinsam auftreten können, haben sie unterschiedliche Ursachen. W-CIN leitet sich aus fehlerhafter Aufteilung der Chromosomen ab. S-CIN entsteht durch nicht fixierte DNA-Doppelstrangbrüche. Ein besseres Verständnis der CIN-Ursachen und -Folgen, bei dem zwischen W-CIN und S-CIN unterschieden wird, wird dazu beitragen, die häufig beobachteten klinischen und evolutionären Auswirkungen von CIN in Therapiestrategien zu berücksichtigen.

Die Entwicklung von CIN ist ein komplexer Vorgang, der auf mehreren Ebenen reguliert wird. Insbesondere die CIN-Genveränderungen führen zu einer Aktivität von Signalwegen und Veränderungen der Tumormikroumgebung, die direkt oder indirekt zu CIN führen. Der CIN-Phänotyp treibt zusammen mit dem Selektionsdruck kontinuierlich die laufenden Veränderungen auf genetischer, pathologischer und zellulärer Ebene voran. Um diese Komplexität zu entschlüsseln, wurden in großem Umfang Multi-Omics-Daten zusammengetragen. Integrative Vorhersagemodelle sind dringend erforderlich, um die erhebliche Komplexität zu bewältigen.

Wenn wir ein besseres Verständnis des CIN-Prozesses erlangen könnten, ließen sich im Idealfall Strategien zur Bekämpfung von CIN in präklinischen Modellen entdecken und validieren. Aufgrund der Diskrepanz zwischen den molekularen Grundlagen und der Mikroumgebung des Tumors in präklinischen Modellen und bei Patienten könnten diese Strategien in der Klinik dennoch nicht wirksam sein. Prädiktive Modelle, die beides aufeinander abstimmen, können helfen, das Translationspotential zu verbessern.

In dieser Arbeit wird zunächst die folgende wichtige Forschungsfrage beantwortet: Was sind die Gemeinsamkeiten und Unterschiede zwischen W-CIN und S-CIN? Dazu werden Multi-Omics-Daten integriert, um die Assoziationslandschaft von W-CIN und S-CIN in 33

Krebsarten zu charakterisieren. Es wurden umfangreiche Korrelationsanalysen zwischen W-CIN/S-CIN und Merkmalen wie Prognose, Empfindlichkeit gegenüber Medikamenten, Aktivität von Signalwegen, somatischen Punktmutationen und Kopienzahlvariationen durchgeführt. Unser Modell sagt voraus, dass das Gen *GINS1* eine W-CIN fördernde Rolle spielt, was später experimentell validiert wurde. Dies deutet darauf hin, dass das in dieser Arbeit vorgeschlagene Prognosemodell wertvoll ist, um CIN-Mechanismen zu verstehen. Angesichts der Tatsache, dass die bestehenden computerbasierten CIN-Studien nicht zwischen W-CIN und S-CIN unterscheiden, dient diese Arbeit als Ergänzung, um unser Verständnis von CIN zu verbessern.

Krebszelllinien werden häufig verwendet, um den CIN-Prozess zu verstehen und neue Krebsbehandlungen zu entwickeln. Die Auswahl der am besten geeigneten Zelllinien für einen bestimmten Tumor ist von entscheidender Bedeutung, um vielversprechende Ergebnisse aus den *in vitro* Experimenten in klinische Anwendungen umzusetzen. In dieser Arbeit wird daher ein halbüberwachtes generatives Modell, MFmap (model fidelity map), vorgeschlagen, das Zelllinien mit Tumoren und Krebssubtypen abgleicht, um die Translationsfähigkeit onkologischer *in vitro* Modelle zu maximieren. MFmap komprimiert hochdimensionale Multi-Omic-Daten in einen kleinen Satz von krebssubtypspezifischen Merkmalen und sagt den Subtyp von Zelllinien voraus, wobei sehr gute Klassifizierungs- und generative Genauigkeit kombiniert werden. Die in MFmap eingebetteten Merkmale können zur Berechnung der paarweisen Zelllinien-Tumor-Ähnlichkeit verwendet werden, mit der man die besten Zelllinien für eine Gruppe von Tumoren oder sogar einzelne Tumoren auswählen kann. Durch die Klassifizierung von Krebszelllinien in Subtypen ermöglicht MFmap die Vorhersage der subtypspezifischen Wirkung von therapeutischen Substanzen. Darüber hinaus kann MFmap zur Untersuchung der Tumorevolution während des Krankheitsverlaufs verwendet werden. Insgesamt ist MFmap für ein breites Spektrum von halbüberwachten Aufgaben im biomedizinischen Bereich nützlich, bei denen die Zuweisung von Klasseneinteilungen schwierig ist.



Acknowledgements

Thanks to my supervisor Prof. Björn Menze for welcoming me to the IBBM group, introducing me to the deep learning field and offering extensive mentorship and support. Thanks to my co-supervisor Prof. Maik Kschischo for his gorgeous guidance, support and encouragement in science and in communication. Just highlight a few: in science he guides me to study Bayesian network, he always gives the freedom to explore whatever I am excited about and gives me timely feedbacks; in communication he offers various great opportunities for collaborative projects.

Thanks to my mentor Prof. Holger Bastians for his valuable advices and support. Thanks to my previous mentor Prof. Andreas Weber for guiding me in cancer evolutionary dynamics modelling, which is critical for cultivating my scientific skills.

Thanks to my collaborators from Prof. Zuzana Storchová's lab and Prof. Holger Bastians's lab for their insightful input and feedbacks. Thanks to my internship mentor Dr. Ulf Reimer at 3B Pharmaceuticals GmbH for his guidance and support. Thanks to Dr. Dominik Kahl and Philipp Wendland for helping handle administrative tasks. Thanks to the Department of Mathematics and Technology at University of Applied Sciences Koblenz where I spend three-year great time. Thanks to DFG FOR2800 project for the financial support.



Contents

Abstract	i
Zusammenfassung	iii
Acknowledgements	v
Contents	vii
List of figures	ix
Acronyms	xi
1 Introduction	1
1.1 Overview	1
1.2 The complexity of cancer	2
1.2.1 Cancer genetic heterogeneity	2
1.2.2 Cancer pathway heterogeneity	3
1.2.3 Tumour microenvironment heterogeneity	4
1.2.4 Cancer dynamics	6
1.3 Chromosomal instability	6
1.3.1 Chromosomal instability classification	7
1.3.2 Chromosomal instability mechanisms	7
1.3.3 Chromosomal instability heterogeneity	9
1.3.4 Chromosomal instability has impacts on cancer complexity	10
1.3.5 Quantification of chromosomal instability	11
1.3.6 Computational models of chromosomal instability	12
1.3.7 Discrepancies between preclinical models and clinical applications	12
1.4 Summary of contributions	13
1.5 Organization	14

2	Background	17
2.1	Supervised learning	17
2.1.1	Single-view learning	18
2.1.2	Multi-view learning	18
2.1.3	Loss function for classification	19
2.1.4	Linear regression	20
2.2	Deep neural networks	20
2.3	Variational Bayesian learning	21
2.3.1	Variational Bayes	21
2.3.2	Stochastic gradient estimation for variational Bayes	23
2.3.3	Variational auto-encoder	24
2.4	Deep generative models for semi-supervised learning	26
3	Distinct and common features of numerical and structural chromosomal instability across different cancer types	29
4	MFmap: A semi-supervised generative model matching cell lines to tumours and cancer subtypes	61
5	Concluding remarks	85
5.1	Conclusion	85
5.2	Outlook	86
5.2.1	Interpretable deep learning to understand CIN mechanisms	86
5.2.2	Self-supervised and semi-supervised learning on tabular data	88
5.2.3	Extending MFmap to regression	88
5.2.4	Extending MFmap to multi-task learning	89
	Appendices	93
A	List of publications	93
	Peer-reviewed journal articles	93
	Bibliography	95



List of figures

1.1	<i>MYC</i> signalling and TME heterogeneity.	5
1.2	Chromosomal instability classification.	8
1.3	Normal chromosome segregations versus chromosome missegregations.	10
2.1	Graphical models for joint distributions.	27
5.1	The design of CINgo architecture.	87



Acronyms

AEVB	auto-encoding variational Bayes
apCAF	antigen presenting CAF
AUC	area under the dose-response curve
BFB	breakage-fusion-bridge
BN	Bayesian network
CAF	carcinoma associated fibroblast
CCL	Cancer Cell Line Encyclopedia
CIN	chromosomal instability
COSMIC	Catalogue of Somatic Mutations in Cancer
DAG	directed acyclic graph
DSB	DNA double-stranded break
ECM	extracellular matrix
ELBO	evidence lower bound
GO	gene ontology
H & E	hematoxylin and eosin
HRD	homologous recombination deficiency
i.i.d.	independent and identically distributed
iCAF	inflammatory CAF

ACRONYMS

KL divergence	Kullback-Leibler divergence
MFmap	model fidelity map
MLE	maximum likelihood estimation
MLP	multilayer perceptron
myCAF	myofibroblastic CAF
NCS	numerical complexity score
NGS	next generation sequencing
OOD	out-of-distribution
RSS	residual sum of squared error
RTK	receptor tyrosine kinase
S-CIN	structural chromosomal instability
SAC	spindle assembly checkpoint
SBS	single base substitution
SCNA	somatic copy number alteration
SCS	structural complexity score
SGD	stochastic gradient descent
SGVB	stochastic gradient variational Bayes
TAM	tumor-associated macrophage
TCGA	The Cancer Genome Atlas
TIL	tumor-infiltrating lymphocyte
TME	tumour microenvironment
VAE	variational auto-encoder
VB	variational Bayesian

- W-CIN whole chromosome instability
- WGD whole genome doubling
- WGII weighted genome instability index

Introduction

1.1 Overview

Cancer is a highly dynamic and complex system, involving diverse elements with different functions and behaviours, as well as multi-layered regulations. These complex properties have been summarised as cancer hallmarks [1, 2, 3]. Among these, [chromosomal instability \(CIN\)](#), a cellular phenotype defined as increased rates of acquiring chromosomal changes [4, 5, 6], is one of the most common cancer features. CIN can be subdivided into two categories: (i) [Whole chromosome instability \(W-CIN\)](#) refers to elevated rates of gaining or losing whole or large parts of chromosomes. (ii) [Structural chromosomal instability \(S-CIN\)](#) describes rapid accumulation of partial chromosomal changes including amplifications, deletions or translocations [4, 5, 6].

Increasing studies have linked CIN to drug resistance [7, 8, 9, 5, 10], poor prognosis [11, 12], elevated metastasis potential [13, 14], rapid adaptive evolution [15, 16, 17] and tumour heterogeneity [15, 18, 19, 20, 21]. Thus targeting CIN may be a promising avenue to kill cancer cells and a deeper understanding of the causes and consequences of CIN is critical for this strategy. With the assumption that integrating multiple data sources allows to construct a complete picture of CIN, a wide range of preclinical CIN models have been developed [22, 23] and abundant cancer omics data accompanied with clinical annotations [24, 25, 26, 27, 28, 29, 30, 31, 32] have been accumulated. Now computational models are urgently needed to complement biological experiments for better understanding of CIN. Specifically, the computational and experimental studies could be performed in an iterative manner: (i) Predictive models integrating multiple layers of information yield quantitative insights on CIN, guiding new experiments in a predictable way; (ii) New data and questions generated by the new experiments further offer opportunities for predictive model development and improvement. Hopefully, with several rounds of iterations, we get not only experientially validated computational models but also better understanding of the CIN mechanisms and consequences.

Previous studies (e.g. [33, 34, 35, 36]) show that computational models are helpful for CIN biological and clinical studies. However existing computational CIN studies focus on W-CIN, leaving S-CIN understudied, though it is known that W-CIN and S-CIN differ in their mechanisms and consequences. Given that we have gained mechanistic insights from computational models distinguishing W-CIN from S-CIN and further validated on preclinical models in an ideal case, the next important question is: How to effectively translate these preclinical findings to clinical targeting. Selecting the most optimal preclinical models for patients is the key to solve this problem.

This thesis aims to address the following CIN related questions using predictive modelling techniques: (i) What are the commonalities and differences between W-CIN and S-CIN? (ii) How to find the best preclinical models that closely mimic the CIN process in a patient? In practice, we have to address question (ii) for general purpose: How to find the best preclinical models (e.g. cell lines) matching a specific tumour?

1.2 The complexity of cancer

The complex cancer system is hierarchically organised: Genetic changes activate or inactivate pathways, leading to cellular process and individual phenotype changes. This hierarchical structure guides current data collection and analysis strategies to decode cancer complexity. For example, DNA alterations, gene expressions and protein expressions are profiled to investigate the cancer complexity at molecular level; pathway activity scores are computed to reveal the pathway-level cancer complexity; microscopic images are analysed to study the cellular level complexity; clinical outcomes of individual tumours are recorded to investigate the phenotype complexity. In literatures cancer complexity and heterogeneity are interchangeably used. In fact, cancer complexity reflects not only heterogeneity, but also its non-linear dynamic properties. I herein describe the cancer complexity from the two aspects.

1.2.1 Cancer genetic heterogeneity

Each cancer cell is theoretically unique because cancers result from combined alterations of a set of oncogenes or tumour suppressor genes and these combinations grossly differ among cancer cells. Multiple integrative analysis have revealed considerable genetic heterogeneity [37, 38, 39, 40, 41, 42, 43, 44]: The alteration distribution is highly skewed and long-fat-tailed. Although recurrent alterations are considered as candidate cancer drivers, they are often undrugable. For example, *KRAS* mutation rate is 30% – 40% in colorectal cancer [45], various approaches targeting *KRAS* do not work effectively (reviewed in [46]). Consensus cancer stratification based on ensemble models using

gene expression data has defined subtypes with clinical implications and clear biological interpretations [47]. Here the biological interpretation is based on the associations between subtypes and features from somatic mutation, copy number variation and methylation.

For those most frequently altered genes, their alteration rates vary across cancer types. For example, a pan-cancer study [48] covering more than 40 cancer types reported that *PIK3CA* mutation tends to hit *HER2*-positive samples and its mutation rate ranges from 0% (mesothelioma) to 37% (endometrial cancer). The same gene can also have different variation types. For instance, *TP53* mutations distribute in all coding exon regions, 30% of which locate in multiple hotspot regions [49]. Even though a mutation is sufficiently frequent so that statistical models have the power to detect it as cancer driver. Its mutant rates vary across different studies [50], suggesting that other latent factors like the cohort might be confounding variables for the distribution of the driver mutations. This striking heterogeneity presents a challenge for cancer driver identification. The [Catalogue of Somatic Mutations in Cancer \(COSMIC\)](#) decomposes 96 sequence context dependent [single base substitutions \(SBSs\)](#) into components which are linked to aetiologies by hypothesizing that different mutagenesis processes induce specific mutation patterns. Although useful, it has at least two limitations: (i) Mutational signatures with limited mutation burdens might be under-represented; (ii) and aetiology space is not fully investigated [51].

1.2.2 Cancer pathway heterogeneity

Molecules including DNA, RNA and proteins are important elements of biological pathways. The molecular heterogeneity naturally provides material for the pathway heterogeneity. Additionally, mutual exclusivity and co-occurrence are two common mutational patterns in the cancer genome. While the former indicates a negative epistatic or synthetic lethal interaction of two pathways, the latter activates two collaborating pathways [52]. Pathways involve a handful of regulations. As a result, the same mutation could have distinct pathway signalling outcomes, or vice versa, different mutations may lead to the same pathway readout.

In general, many causal oncogenic proteins are the network hubs connecting multiple upstream regulation signalling and downstream effectors. These effector proteins selectively bind to binding site of the hub, activating the respective effector subpathway at a given time. Numerous factors like post-translational modifications, effector concentrations, plasma membrane organisations determine which effectors interact with the hub [53]. These effector pathways may function in a competing manner. The *RAS* pathway provides a good example of how a complex signalling regulation can substantially contribute to the cancer pathway heterogeneity. First, there are many *RAS* activation mechanisms

including growth factors, chemokines, Ca^{2+} or [receptor tyrosine kinase \(RTK\)](#) and a wide range of downstream effector pathways like *RAF–MEK–ERK* and *PI3K–AKT–MTOR* signalling [54]. Second, the *RAS* has different isoforms of which activation depends on the upstream signalling strength and the activation of effector pathways shows isoform specificities [55]. Finally, the *RAS* has several complex downstream subpathways. The *PI3K–AKT–MTOR* pathway is one example. *PI3K* can be directly activated by RTK and involve many modes of regulations, ranging from multiple negative and positive feedback loops to crosstalk with other signalling pathways [56].

1.2.3 Tumour microenvironment heterogeneity

The [tumour microenvironment \(TME\)](#) contains not only tumour cells but also stromal cells and immune cells. The different cellular components consist of different sub-populations. The malignant component can be remarkably heterogeneous due to the underlying diverse genetic make-up and pathway activities as mentioned in Subsections 1.2.1 and 1.2.2. Stromal and immune components are constituted by various cell types such as [carcinoma associated fibroblasts \(CAFs\)](#), [tumor-associated macrophages \(TAMs\)](#) and T cells. Three major types of CAF can be distinguished: tumor-suppressive [myofibroblastic CAFs \(myCAFs\)](#) [57], tumor-promoting [inflammatory CAFs \(iCAFs\)](#) [57] and MHC class II⁺ [antigen presenting CAFs \(apCAFs\)](#) [58]. TAMs include pro-inflammatory M1-TAM and pro-tumoral M2-TAM [59]. T cell populations are also heterogeneous: CD8⁺ T cells, CD4⁺ T cells and $\delta\gamma$ T cells belong to tumor-suppressive subsets; Tregs, TH2 and TH17 cells are pro-tumoural subsets (reviewed in [60]).

Beyond the cell type abundance complexity, spatial heterogeneity represents another characteristics of TME. [Tumor-infiltrating lymphocyte \(TIL\)](#) structural patterns are associated with molecular and clinical readouts [61]. Immune cell distribution together with functional activities defines three immunophenotypes: immune-inflamed (high intra-tumour T cell infiltration), immune-excluded (T cell infiltration located in the invasive margin) and immune-desert (absent T cell infiltration) [62]. The [extracellular matrix \(ECM\)](#) that provides structural and mechanical integrity [63] is associated with the TME cellular compositions and their spatial distributions. Excessive dense ECM acts as a physical barrier to TILs, thereby being linked to immune-excluded phenotype [64, 65]. CAFs, TAMs and tumour cells secrete ECM-modifying related enzymes, proteases, cytokines, chemokines and growth factors, directly or indirectly contribute to ECM remodelling [66].

The cancer cell intrinsic properties (genetic and pathway profiles) contribute to TME heterogeneity. The interaction between *MYC* signalling and immune cells provides such a good example (see Fig 1.1). *MYC* signalling is regulated by multiple oncogenic signalling in transcription, translation and post-translation stages [67]. These regulations involve

complex signalling networks. *MYC* activated cells can induce cytokines and chemokines to change the immune cell population [68], resulting in a TME that maximises the cellular fitness.

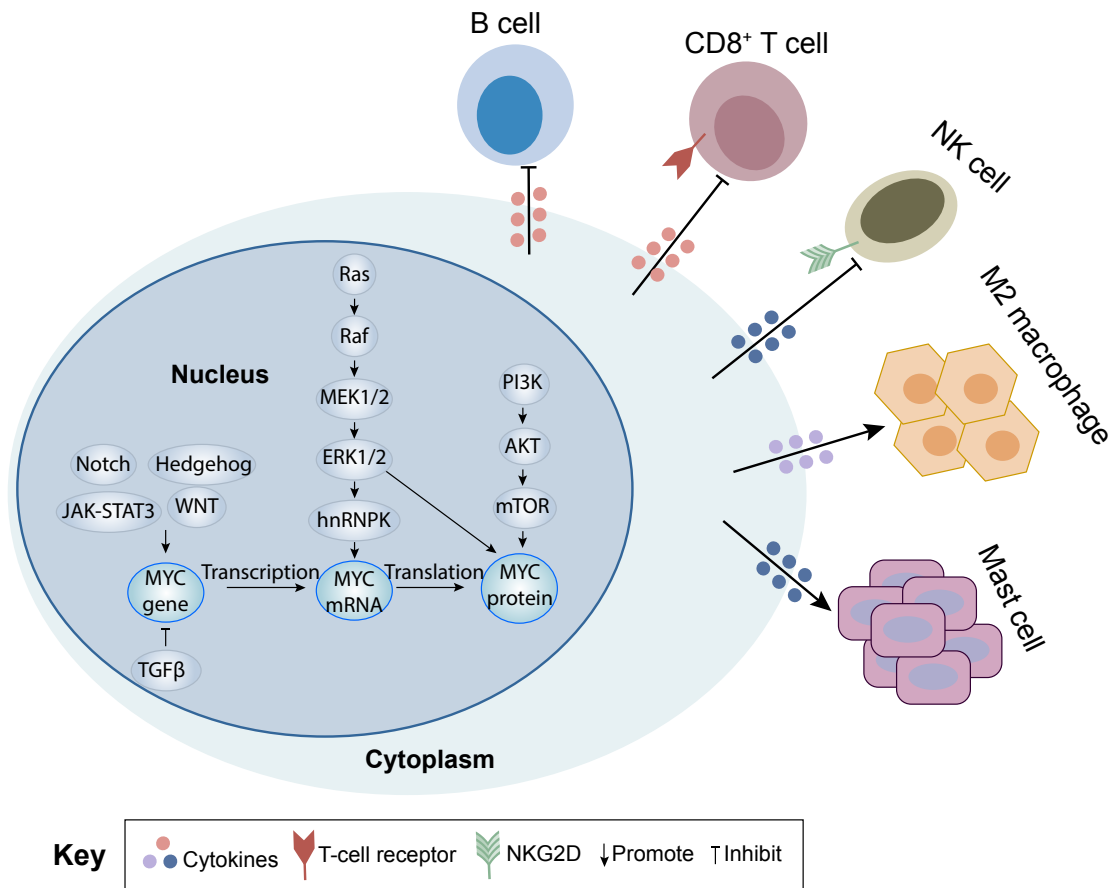


Figure 1.1: MYC signalling and TME heterogeneity. At genetic and pathway level: *MYC* transcription is regulated by *Hedgehog*, *WNT*, *Notch*, *JAK-STAT3* and *TGFβ* signalling; *MYC* translation is mediated by *MAPK-HNRNPK* signalling; *MYC* stabilisation is affected by the *Ras-Raf* and *PI3K-AKT* signalling. At the cellular level: activated *MYC* signalling promotes pro-tumoural immune cell sub-population but inhibiting the tumour-suppressive immune cells by cytokine secretion. The signalling relationship in the nucleus is annotated from [67] and *MYC*-immune relationship is annotated from [68].

1.2.4 Cancer dynamics

Advanced sequencing technologies and software allow to obtain a complete picture of the molecular, pathway and cellular heterogeneity of cancer, thereby improving the disease interventions in an ideal case. But the picture is temporary and the curative results may not last long since the multi-factor cancer system is evolving over time. This evolution proceeds through mutation, selection and adaptation. As a result, cancer cells can adapt to the TME, maximising fitness. The newly acquired mutations may act together with constitutive ones or alone to yield fitness advantage, propagated into the divided daughter cells. Accordingly, other TME parameters including ECM stiffness, within-tumour nutrient concentration, vasculature morphology, TILs, stromal component and spatial distribution may change. These changes serve as the consequences and causes of cancer evolution.

Cancer progression is a good example of the above described dynamics where heterogeneity and evolution are mutually causative. Heterogeneity fosters evolution [69] since tumour cells require heterogeneity to respond to selection pressure imposed by therapies [70, 71, 72] and immune system [73]. The ongoing clonal evolution contributes to clinical and genetic heterogeneity, for example by providing multiple mechanisms to resist interventions [74, 75, 76]. Additionally, the plasticity of tumour cells and immune cells also facilitates such heterogeneity and dynamics [77, 78, 79].

In summary, multiple malignant and non-malignant cellular components coexist in the same complex and dynamic system. They interact with each other and response to the secreted signals, further shaping the system. In turn, tumour cells acquire multi-level changes conferring fitness advantages and thus rapidly adapt to selective pressure exerted by the new environment. In the mutation-selection-adaptation cycle, cancer cells together with their TME continuously evolve. The evolutionary dynamics may yield several problematic scenarios: cancer cells derived from the same patient vary in different cell culture conditions; TME diverges significantly before and after treatment; primary cancer cells differ from recurrent ones in the same location of the same patient; potent drugs in primary tumours do not work for metastatic tumours of the same patient; treatable tumours may become irremediable due to delayed treatment. A large part of this complexity stems from CIN. In the following section, I review the potential mechanisms and impacts of CIN. I discuss methods and data used to measure CIN. I point out that existing computational CIN studies mainly focus on W-CIN, leaving S-CIN understudied.

1.3 Chromosomal instability

CIN substantially contributes to cancer complexity. On one hand, CIN affects a plethora of genes from one cell division, providing genetic materials for selection and adaptation

[4, 80, 81]. On the other hand, the dynamic nature of CIN makes it the major mediator of cancer heterogeneity [5]. The association between CIN and cancer heterogeneity might explain why CIN is a prognostic marker. It is also important to note that there might exist an optimal CIN level for cellular fitness: while intermediate CIN promotes cancer progression, severe CIN is lethal [82, 83]. This concept is consistent with the paradoxical relationship between CIN and clinical outcomes [84]. Although elevated CIN is harmful for cellular fitness, it is often associated with poor prognosis in clinics. This paradox raises two major problems in clinical utility of CIN: (i) How to quantify CIN? (ii) How to define a CIN threshold that is predictive for prognosis? Here, I provide an overview of CIN mechanisms, CIN heterogeneity and its contribution to cancer complexity, discussing the limitations of the current CIN quantification approach.

1.3.1 Chromosomal instability classification

Chromosomal instability (CIN), defined as an increased rate of acquiring numerical and structural chromosome changes, is frequently observed in many cancer types [4, 85, 86, 87]. CIN has two major forms (Fig 1.2): **whole chromosome instability (W-CIN)** and **structural chromosomal instability (S-CIN)**. W-CIN refers to elevated rates of gaining or losing whole or large parts of chromosomes [4, 5, 6] and often leads to aneuploidy [86, 82, 88]. Aneuploidy that refers to the number of imbalanced chromosomes [88, 82], can reversely induce CIN [89] or occur independently [90]. S-CIN is defined as an elevated rate of accumulating structural chromosomal changes, resulting in segmental amplifications and deletions, balanced and unbalanced translocations, inversions [4, 5, 6]. W-CIN and S-CIN have different origins. Erroneous **DNA double-stranded breaks (DSBs)** are major causes of S-CIN [6]. W-CIN mainly arises through chromosome missegregations [91]. They can co-occur and are reciprocally connected [92, 93, 94] (see the Subsection 1.3.2 for detailed explanations). It is also interesting to note that the frequency of arm level **somatic copy number alterations (SCNAs)** is higher than those of focal level [95] in most cancer types, suggesting that W-CIN and S-CIN are differentially selected.

1.3.2 Chromosomal instability mechanisms

In a normal mitotic process, the full chromosome sets of one cell are replicated and equally divided, generating two identical daughter cells (Fig 1.3(A)). Errors in mitotic chromosome segregation cause W-CIN due to mitotic checkpoint defects (Fig 1.3(B)), cohesion defects (Fig 1.3(C)), centrosome amplification (Fig 1.3(D)) and merotelic attachment (Fig 1.3(E)) (all the mechanisms are reviewed in [5, 96]). Merotelic attachments could arise through several paths including aberrant spindle morphology, increased kinetochore microtubule

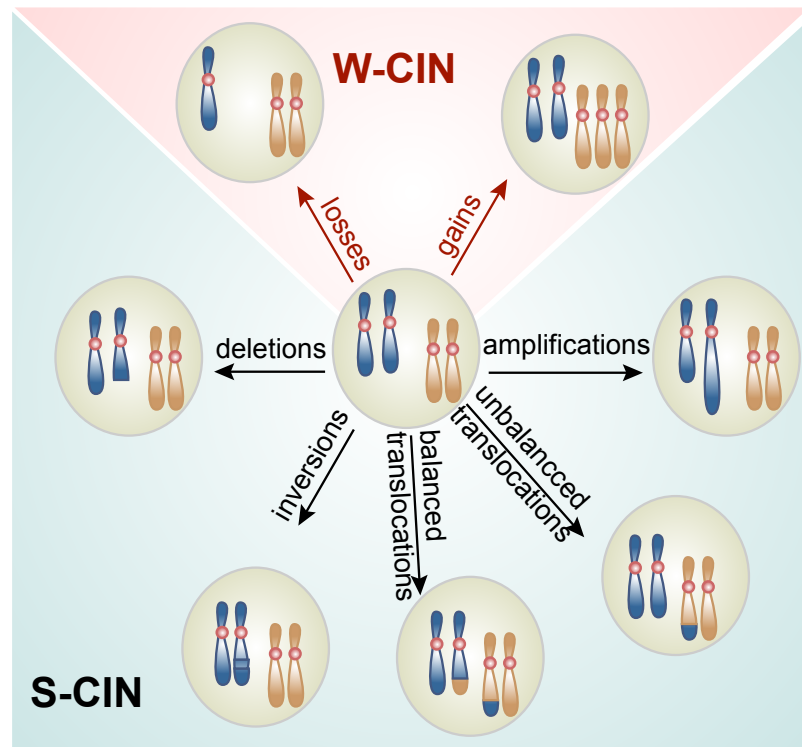
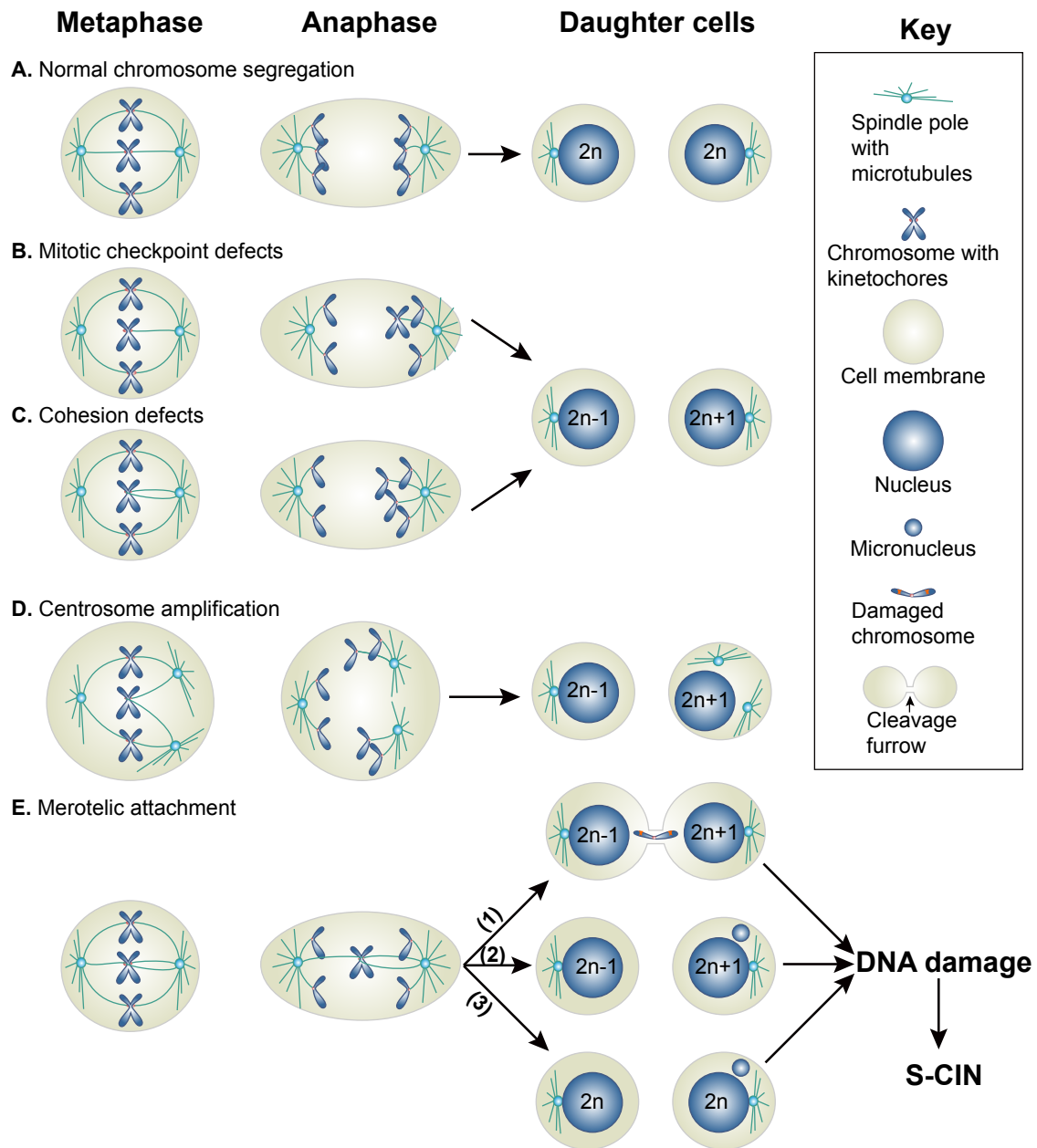


Figure 1.2: Chromosomal instability (CIN) classification. Whole chromosome instability (W-CIN) is characterised by ongoing losing or gaining whole or large parts of chromosomes, leading to aneuploidy. Structural chromosomal instability (S-CIN) is characterised by a tendency to accumulate changes of smaller parts of chromosomes.

stability and multiple microtubule attachment sites (reviewed in [96]). S-CIN arises through erroneous DSBs [96, 6] and breakage-fusion-bridge (BFB) cycles [96, 5]. Replication stress is defined as slowing or stalling of replication fork progression during DNA synthesis [97]. It is a shared way to induce W-CIN and S-CIN and provides a way linking the two types of CIN. Specifically, replication stress generates dicentric chromosomes and acentric chromosomes, leading to chromosome missegregations [94]. The formed dicentric chromosomes can enter into BFB cycles [98] that trigger S-CIN. Alternatively, replication stress directly causes DNA damage [97, 99] or indirectly causes DNA damage via trapped and damaged chromosomes in the cytokinesis cleavage furrow [92, 100] and fragmentation of micronuclei which might trigger S-CIN (Fig 1.3(E)) [101, 100]. In turn, the resulting S-CIN leads to chromosome missegregations [94, 102].



1.3.3 Chromosomal instability heterogeneity

Although W-CIN and S-CIN occur frequently in most common cancer types, the W-CIN and S-CIN levels vary across cancer types or even cancer subtypes [103]. Most

Figure 1.3 (previous page): Chromosome missegregation mechanisms. (A) Normal chromosome segregation. Two kinetochores are attached to microtubules arising from both spindle poles, chromosomes are equally divided into two euploid daughter cells. (B) Defective [spindle assembly checkpoint \(SAC\)](#). With defective SAC signalling, anaphase starts with unattached kinetochores. Two copies of one chromosome are moved towards the same spindle pole. (C) Cohesion defects. Both kinetochores are attached to the microtubules arising from the same centrosome. Two copies of one chromosome are divided into only one daughter cell in (B) and (C), resulting in aneuploid daughter cells. (D) Centrosome amplification. Centrosome amplification leads to an extra centrosome in one spindle side, resulting in one kinetochore being attached to microtubules arising from two centrosomes in one spindle pole, inducing chromosome missegregations. (E) Merotelic attachment. Merotelic attachment is defined as one kinetochore being attached to microtubules arising from both spindle poles. Merotelic attachment causes lagging chromosomes in anaphase, generating euploid (3) or aneuploid (1, 2) daughter cells. The lagging chromosome can either be trapped and damaged in the cleavage furrow in cytokines (1) or form micronucleus (2, 3). In all three scenarios, DNA is damaged and replication is stalled, leading to structurally abnormal chromosomes.

breast and ovarian cancer patients harbouring defective homologous recombination, an important CIN mechanism, tend to have the CIN phenotype. ER⁻ breast cancers have higher CIN levels compared to ER⁺ samples [104]. Colorectal cancer CMS2 and CMS4 subtypes belong to CIN⁺ samples. However, CMS2 samples have higher CIN levels and worse prognosis compared to CMS4 [47]. This might suggest that the CIN levels in CMS2 subtype cancer cells surpass the optimal threshold, thereby being negatively selected. Furthermore, CIN is strongly associated with many phenotype transforming courses including drug resistance, metastasis, disease progression, tumour recurrence and intra-tumour heterogeneity, indicating that CIN status also evolves over time. A recent multi-sample phasing SCNA analysis directly points out that continuous CIN causes SCNA heterogeneity [105]. Taken together, these data reflect the dynamic and heterogeneous nature of CIN.

1.3.4 Chromosomal instability has impacts on cancer complexity

Thanks to its dynamic and heterogeneous attributes, CIN could affect a large part of a cell's cancer genome, yielding complex karyotypes that are continuously evolving. It is apparent that heterogeneous genotypes generated by CIN provide substrates and diversity for selection and adaptation [4, 80, 81]. However, the proportion of affected genome

and the degree of consequent karyotype complexity do not necessarily determine the cellular fitness. A number of copy number alterations caused by CIN may only cover non-functional non-coding genes. Conversely, one single chromosome arm change can allow substantial cellular function change [106, 107]. Even though multiple causal oncogenes are located in CIN affected genome regions thereby being activated simultaneously, CIN can serve as fitness penalty. This could happen if these concurrently activated signalling induces oncogenic stress or they function in divergent and antagonizing manner [52]. Similarly, if the CIN affected genes act in a collaborating manner, CIN is considered as promoting fitness advantage. This idea provides an explanation for mechanisms enabling CIN propagation including aneuploidy tolerance and CIN attenuation [4]. Since the aneuploidy tolerance and CIN attenuation procedures separately cover a wide range of mechanisms that allow cancer cells to cope with ongoing inner molecular and pathway changes, as well as external TME changes, the phenotype implications of CIN remain elusive. It is important to note that the aneuploidy tolerance and other changes acquired by evolved CIN cells should be distinguished from acute response to aneuploidy and CIN [108]. In conclusion, CIN is a context-dependent process. The context mentioned here includes cancer stage, cell types, genotypes, TME, species and cell culture conditions [106, 80, 109].

1.3.5 Quantification of chromosomal instability

The dynamic and heterogeneous nature of CIN poses a challenge to quantitatively assess CIN status. Detailed techniques applied on CIN measurement and their advantages/limitations are reviewed in [5, 110]. In principle, these techniques belong to two broad categories: (i) direct monitoring and (ii) surrogate measures. To capture the CIN dynamic nature, tracking and calculating mitotic error rates or chromosomal aberration rates within a live cell over time is required. However, the involved experimental approaches are complex, rendering it difficult to execute in routine experimental and clinical settings. With the assumption that the CIN degree reflects the cell-to-cell heterogeneity within a cell population, chromosomal structural and numerical changes are measured at single-cell level and intra-tumour variation of these changes are used to infer CIN degrees. In cell population based approaches, averaged intra-tumour heterogeneity, karyotype complexity and SCNA burdens are used as surrogate of CIN measurement. One caveat for using a bulk tumour genomic approach is that the inferred copy number or structural changes can be diluted by the non-neoplastic cells [111], therefore tumour purity effects must be corrected. Similarly, it is difficult to deconvolute the effects of CIN from those of proliferation [36] or non-tumour cell contamination [112, 113] using CIN scores that are derived from the CIN gene signatures of which expressions are highly correlated with CIN

status or aneuploidy (e.g. [34]). It is also important to note that these surrogates just capture a static picture of the complexity at cellular, karyotypic and transcriptomic level, shaped by ongoing CIN and its interaction with selection pressure [103].

1.3.6 Computational models of chromosomal instability

No matter which approach is used for CIN evaluation, computational models could complement experiments to deepen our understanding of CIN. These computational tools are widely used in the following CIN related studies: image processing based karyotype analysis [114, 115, 116], copy number segment based CIN metrics calculation [94, 117], pan-cancer CIN landscape characterisation [107, 33, 118, 119, 120], CIN evolutionary dynamics modelling [120, 121, 122], predicting CIN using [hematoxylin and eosin \(H & E\)](#) images [123, 124]. Although it is widely appreciated that W-CIN and S-CIN differ in their origins and consequences, the above mentioned studies consider only W-CIN or CIN/aneuploidy in general, leaving S-CIN understudied. [Structural complexity score \(SCS\)](#) which counts the number of structurally aberrant regions in the genome of a sample has been proposed to be a good proxy measure of CIN in previous work [94], thereby providing a feasible approach to complement the current CIN computational studies.

1.3.7 Discrepancies between preclinical models and clinical applications

Considering the clinical implications of CIN, targeting CIN might be a promising strategy to kill cancer cells. However, cancer chromosomal instability complexity not only presents a substantial problem for effective interventions but also remains a challenge of translational medicine. In general, significant treatment response differences exist among different types of preclinical models [125]. Promising biomarkers and treatments discovered and validated in preclinical model systems are hardly translated into clinical applications [126, 127]. On the other hand, cancer researches often lack for preclinical model systems that closely mimic the tumour ecosystem of a patient or a group of patients. Although a number of differences like the TME and DNA damage response are proposed to explain the response discrepancy, these factors may interplay to shape the phenotype of individual tumour samples (as underscored in Subsection 1.3.4). Therefore a holistic approach taking into account these factors together with their interactions is needed to match the preclinical models to patients.

1.4 Summary of contributions

Two major challenges for understanding and targeting CIN are highlighted in Section 1.3: (i) The commonalities and differences between W-CIN and S-CIN remain poorly understood. (ii) The discrepancy between preclinical models and patients limits the translational ability of promising *in vitro* experimental results. This thesis presents two predictive modelling based frameworks in Chapter 3 and Chapter 4 to resolve challenges (i) and (ii) respectively. I herein summarise the rationale and contributions of each work.

Chapter 3: Distinct and common features of numerical and structural chromosomal instability across different cancer types

One can distinguish two types of CIN: W-CIN and S-CIN. While W-CIN describes elevated rates of acquiring whole or large parts of chromosome changes, which are caused by persistent chromosome missegregations. S-CIN samples tend to accumulate changes on focal chromosome segments due to DNA damage repair deficiency. Although W-CIN and S-CIN arise through different molecular characteristics, currently available computational CIN studies only focus on W-CIN or CIN in general but leaving S-CIN understudied. Here we analyse cancer genomic data to complement the existing cancer CIN studies, intending to provide comprehensive characterizations of commonalities and differences between W-CIN and S-CIN. In particular, our analysis reveals an almost universal bi-modal pattern in the distribution of W-CIN that are absent in S-CIN. We then show that whole genome doubling is uniformly strongly correlated with W-CIN, but in S-CIN, homologous recombination deficiency shows the strongest consistent association. We demonstrate that prognostic values of W-CIN and S-CIN are cancer type dependent. We identify compounds that selectively target high CIN and reveal that currently available compounds are biased to targeting low CIN tumours. We propose *CKS1B* as potential candidate S-CIN target, its high activity is significantly correlated with S-CIN at the pathway and gene expression levels. We show that high S-CIN is associated with copy number variations rather than somatic mutations in several important cancer driver genes. Finally we propose a copy number based mechanism to promote *PI3K* signalling in high S-CIN tumours.

Chapter 4: MFmap: A semi-supervised generative model matching cell lines to tumours and cancer subtypes

Cell lines are the most commonly used model system to understand mechanisms underlying cancer chromosomal instability and more. Even if promising strategies for targeting CIN could be found in cancer cell lines in an ideal case, fidelitious *in vitro* models that

closely mimic a specific tumour are still needed to maximise the translational capacity. Here we developed **model fidelity map (MFmap)**: a semi-supervised generative neural network integrating copy number, somatic mutation and gene expression data with cancer subtypes to simultaneously predict the cancer subtype labels of a cell line and its similarity to individual tumours. MFmap is a new variant of semi-supervised **variational auto-encoder (VAE)** which can achieve very good classification accuracy and good generative performance simultaneously. The high accuracy (F_1 score $> 90\%$) of MFmap subtype prediction can be seen in ten studied cancer types. With MFmap, one can select the best cell lines for a specific cancer subtype or even an individual tumour. The pair-wise cell-line-tumour similarity is evaluated on the MFmap embedded latent representations, which are invariant between the tumour samples and cell lines. This allows to translate subtype specific drug response to individual tumours. We further demonstrate that the MFmap learned representations are clinically and biologically meaningful and can explain cancer subtype specific features. Furthermore, the generative nature of MFmap allows us to model cellular state transformation during disease course.

1.5 Organization

This paper-based thesis is organised as follows:

Chapter 1 gives the motivation to predictive modelling of cancer chromosomal instability by describing their complexity and relationships. The currently available data collection and analysis strategies to understand the cancer complexity reflect how genetic alterations propagate throughout the biological systems: genetic alterations lead to pathway dysfunctions that give rise to tumour microenvironment and clinical phenotype changes. Section 1.2 is organised according to such a hierarchy, introducing each layer of complexity and discussing their relationships. Section 1.3 briefly introduces the CIN classification, CIN mechanisms, CIN heterogeneity and its clinical/molecular implications on cancer complexity, CIN evaluation and translational gaps between preclinical models and patients. Finally I outline the major contributions of the thesis in Section 1.4.

Chapter 2 summarises the predictive models used to study cancer chromosomal instability complexity. The principles underlying linear regression and stochastic gradient variational Bayes are introduced as the theoretical foundations of Chapter 3 and Chapter 4.

Chapter 3 and Chapter 4 are based on two of my first-author peer-reviewed journal papers in their original forms. Licence notice, simple summary and author contribution corresponding to the paper are presented at the beginning of each chapter, followed by the full-text article.

Chapter 5 briefly summarises the findings and contributions of my research and links them to related works in literatures, pointing out next steps to improve and extend my work. Appendix A contains a list of peer-reviewed journal articles published during my time as a Phd student.

Background

Underscoring the previously mentioned inherent complexity of cancer and chromosomal instability, there is now a strong need to collect and analyse a wide variety of datasets covering all facets of the two systems. For this reason, [next generation sequencing \(NGS\)](#) techniques have been broadly used to generate enormous multi-omic data. As noted in Subsection [1.3.7](#), there are enormously multi-scale discrepancies between *in vitro* model systems and cancer patients, leading to most of the failures in translational medicine. Statistical and machine learning (including deep learning) are vital for integrating these data to (i) uncover the underlying biology of cancer chromosomal instability, (ii) and prioritise the best *in vitro* models for experimental validation. The core computational tools used in this thesis include linear regression model and semi-supervised generative model. The semi-supervised generative model utilises a [variational Bayesian \(VB\)](#) method. To integrate multi-omic data, these models are designed for multi-view learning. I herein give an overview of the key concepts and underlying principles of these models. Please note, the mathematical notations here are specific to genomic data and concepts are introduced briefly. For more broad and detailed discussions on these topics, I recommend [\[128, 129\]](#). For tutorials focusing on VB inference I recommend [\[130, 131, 132\]](#)

2.1 Supervised learning

In supervised settings, a model is trained to predict the discrete label (classification) or continuous target variable (regression) of a sample given its associated inputs. The inputs could be taken from single or multiple data sources. If a model allows to integrate different data modalities such as copy number variations, mutations and gene expressions, it is named as multi-view learning, otherwise is referred as single-view learning. Compared to single-view learning which only provides one aspect of the cancer chromosomal instability, multi-view learning enables to capture the association within single-omic data type, as well as the association between different data types, thereby providing a more comprehensive

picture. Here I discuss single-view learning versus multi-view learning in the supervised learning framework, nevertheless one can easily extend this to other types of machine learning (e.g. unsupervised learning). Depending on the data type of the labels, supervised learning solves two classes of problems: classification (categorical labels) and regression (continuous target variables), necessitating different types of loss functions. I show two exemplary loss functions, cross entropy loss and [residual sum of squared error \(RSS\)](#) that are equivalent to the criterion of [maximum likelihood estimation \(MLE\)](#).

2.1.1 Single-view learning

The base case of supervised learning is single-view learning, suppose we have an [independent and identically distributed \(i.i.d.\)](#) training set comprising n labelled samples $\mathcal{D}^{train} = \{(\mathbf{x}^{train,(i)}, y^{train,(i)})\}_{i=1}^n$, $\mathbf{x}^{train,(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$, $y^{train,(i)} \in \{1, \dots, K\}$ (for K -class classification) or $\mathcal{D}^{train} = \{(\mathbf{x}^{train,(i)}, y^{train,(i)})\}_{i=1}^n$, $\mathbf{x}^{train,(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$, $y^{train,(i)} \in \mathcal{Y} \subseteq \mathbb{R}$ (for regression), where $\mathbf{x}^{train,(i)}$ and $y^{train,(i)}$ denote the feature vector and label of sample i respectively. A predictive model is trained to learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ using \mathcal{D}^{train} such that $f(\mathbf{x}; \theta)$ could make accurate predictions on an unseen test dataset $\mathcal{D}^{test} = \{(\mathbf{x}^{test,(i)}, y^{test,(i)})\}_{i=1}^m$, where $\theta \in \Theta$ (parameter space) is the parameter of $f \in \mathcal{F}$ (hypothesis space). The model performance is quantified by evaluation metrics. The predictive model training process attempts to find optimal values for θ , minimising the defined objective:

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^n l(y^{train,(i)}, f(\mathbf{x}^{train,(i)}; \theta)). \quad (2.1)$$

Here $l(y^{train,(i)}, f(\mathbf{x}^{train,(i)}; \theta))$ is a differentiable loss function.

To avoid overfitting, a complexity penalty $\Omega : \mathcal{F} \rightarrow \mathbb{R}_+$ is added to the objective function, then the optimal θ^* can be expressed as:

$$\theta^* = \arg \min_{\theta \in \Theta} \mathcal{L}(\theta) + \lambda \Omega(\theta). \quad (2.2)$$

The regularization coefficient λ is a hyperparameter, which can be tuned using grid search, randomised search [133] or Bayesian optimization [134], yielding the best evaluation metric on the validation dataset. Additionally cross validation can be used.

2.1.2 Multi-view learning

The single-view setting can be easily extended to the multi-view case. To avoid notational clutter, I hide the data dimensions, data ranges and dataset types from now on. Given a training set with n labelled samples having v views. The j -th view is an [i.i.d.](#) dataset,

written as $\mathcal{D}^j = \{(\mathbf{x}^{(i,j)}, y^{(i)})\}_{i=1}^n$, where $\mathbf{x}^{(i,j)} \in \mathcal{X}^j$ and $y^{(i)} \in \mathcal{Y}$ denote the feature vector from the j -th view and label of sample i respectively and $j \in \{1, \dots, v\}$. For a given sample feature $(\mathbf{x}^{(i,1)}, \dots, \mathbf{x}^{(i,j)}, \dots, \mathbf{x}^{(i,v)}) \in \mathcal{X}^1 \times \dots \times \mathcal{X}^j \times \dots \times \mathcal{X}^v$, v functions $\{f^1, \dots, f^j, \dots, f^v\} \in \mathcal{F}^1 \times \dots \times \mathcal{F}^j \times \dots \times \mathcal{F}^v$ are trained to make accurate predictions, where $f^j : \mathcal{X}^j \rightarrow \mathcal{Y}$. The training is achieved by minimizing the overall objective summing up v single-view objectives:

$$\mathcal{L}(\theta^1, \dots, \theta^j, \dots, \theta^v) = \sum_{j=1}^v \mathcal{L}^j(y, f(\mathbf{x}^j; \theta^j)). \quad (2.3)$$

Here $\mathcal{L}^j(y, f(\mathbf{x}^j; \theta^j))$ is the j -th view objective function parametrised by $\theta^j \in \Theta^j$, where $\{\theta^1, \dots, \theta^j, \dots, \theta^v\} \in \Theta^1 \times \dots \times \Theta^j \times \dots \times \Theta^v$. Taking into account the regularisation Ω and co-regularisation Ω_{co} terms [135], we can resolve the optimal parameters by:

$$\theta^{1*}, \dots, \theta^{v*} = \arg \min_{\theta^1, \dots, \theta^v} \sum_{j=1}^v \mathcal{L}^j(\theta^j) + \lambda \sum_{j=1}^v \Omega(\theta^j) + \lambda_{co} \sum_{\substack{i,j \in \{1, \dots, v\} \\ i \neq j}} \Omega_{co}(\theta^i, \theta^j). \quad (2.4)$$

It is apparent that the major difference between single-view learning and multi-view learning is the co-regularisation term, which can be explicitly defined or implicitly modelled by the neural network architecture.

2.1.3 Loss function for classification

Let $p(y|\mathbf{x})$ be the true conditional distribution over the label y given the input \mathbf{x} , $p(y|\mathbf{x})$ is usually unknown. Supervised learning aims to estimate a distribution $q_\theta(y|\mathbf{x})$ that approximates $p(y|\mathbf{x})$, parametrised by θ , such that for any $(\mathbf{x}^{(i)}, y^{(i)})$:

$$p(y^{(i)}|\mathbf{x}^{(i)}) \approx q_\theta(y^{(i)}|\mathbf{x}^{(i)}). \quad (2.5)$$

The **Kullback-Leibler divergence (KL divergence)** $D_{KL}(p||q_\theta)$ measures the dissimilarity between p and q_θ . In classification problems, we minimise the **KL divergence** loss to find optimal model parameters. In fact, minimising **KL divergence** loss is equivalent to minimising cross entropy loss.

Given a training set $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$, where $y^{(i)} \in \{1, \dots, K\}$ is the label, the categorical cross entropy loss is defined as:

$$\mathcal{L}_{CE} = -\frac{1}{n} \sum_{i=1}^n \log q_\theta(y = y^{(i)}|\mathbf{x}^{(i)}). \quad (2.6)$$

2.1.4 Linear regression

In a regression model, the target variable is continuous. Given a training dataset $\mathcal{D}^{train} = \{(\mathbf{x}^{train,(i)}, y^{train,(i)})\}_{i=1}^n$, $\mathbf{x}^{train,(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$, $y^{train,(i)} \in \mathcal{Y} \subseteq \mathbb{R}$, we aim to find a relationship between \mathbf{x} and y . In a Gaussian setting, we assume:

$$q_\theta(y^{(i)}|\mathbf{x}^{(i)}) = \mathcal{N}(y^{(i)}; \mu_\theta(\mathbf{x}^{(i)}), \sigma^2). \quad (2.7)$$

If the expectation μ_θ is a linear function of \mathbf{x} : $\mu_\theta(\mathbf{x}^{(i)}) = \mathbf{w}^T \mathbf{x}^{(i)} + \epsilon^{(i)}$, where $\epsilon^{(i)}$ denotes the residual error between the prediction and the target and $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$. The MLE estimator of $\theta = (\mu_\theta, \sigma^2)$ is:

$$\begin{aligned} \hat{\theta}^{\text{MLE}} &= \arg \max_{\theta \in \Theta} \sum_{i=1}^n \log q_\theta(y^{(i)}|\mathbf{x}^{(i)}) \\ &= \arg \max_{\theta \in \Theta} - \sum_{i=1}^n \frac{1}{2\sigma^2} (y^{(i)} - \mu_\theta(\mathbf{x}^{(i)}))^2 - \log(\sqrt{2\pi\sigma^2}) \\ &= \arg \min_{\theta \in \Theta} \sum_{i=1}^n \frac{1}{2\sigma^2} (y^{(i)} - \mu_\theta(\mathbf{x}^{(i)}))^2 + \log(\sqrt{2\pi\sigma^2}) \quad (2.8) \\ &= \arg \min_{\theta \in \Theta} \sum_{i=1}^n (y^{(i)} - \mu_\theta(\mathbf{x}^{(i)}))^2 \\ &= \arg \min_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^n (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2. \end{aligned}$$

Eq (2.8) is known as the [residual sum of squared error \(RSS\)](#). Minimising the RRS with respect to \mathbf{w} is equivalent to [MLE](#). Let X be the $n \times d$ data matrix:

$$X = \begin{pmatrix} x^{(1,1)} & x^{(1,2)} & \dots & x^{(1,d)} \\ x^{(2,1)} & x^{(2,2)} & \dots & x^{(2,d)} \\ \vdots & \vdots & \ddots & \vdots \\ x^{(n,1)} & x^{(n,2)} & \dots & x^{(n,d)} \end{pmatrix},$$

and the gradient of RRS equate to 0. Given $X^T X$ is invertible, we can get $\hat{\mathbf{w}}^{\text{MLE}} = (X^T X)^{-1} X^T y$. This estimation method is called ordinary least square.

2.2 Deep neural networks

As noted for the linear regression problem presented in Subsection 2.1.4, the expectation of the conditional distribution $p(y|\mathbf{x})$ is modelled as a linear function of \mathbf{x} . A neural

network allows to model the non-linear relationship between the inputs and outputs. Let's consider the base case: The [multilayer perceptron \(MLP\)](#) is a type of feedforward neural networks. The basic components of MLP are: input layer, output layer and hidden layers that are denoted by $l \in \{1, \dots, L\}$. The number of neurons in layer l is n_l , and n_l can be specified according to the purpose, the number of neurons in the input or the output layer equates to the dimensions of input or output data respectively. Each neuron in layer l is connected to all neurons in layer $l - 1$, i.e. fully connected. The layer l computes a function:

$$\mathbf{a}^{l+1} = \sigma_l(\mathbf{W}_l \mathbf{a}^l + \mathbf{b}_l). \quad (2.9)$$

Here σ_l is the activation function of layer l . The weight matrix $\mathbf{W}_l \in \mathbb{R}^{n_l \times n_{l-1}}$ and the bias vector $\mathbf{b}_l \in \mathbb{R}^{n_l \times 1}$ are the trainable parameters of the network. The MLP can thus be expressed as:

$$\text{MLP}(\mathbf{x}) = \sigma_L(\mathbf{W}_{L-1}(\dots \sigma_2(\mathbf{W}_2 \sigma_1(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2) \dots) + \mathbf{b}_{L-1}) \quad (2.10)$$

Below is a list of commonly used activation function:

- Sigmoid:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \quad (2.11)$$

- Tanh:

$$\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \quad (2.12)$$

- Rectified Linear Unit (ReLU):

$$\sigma(x) = \max(0, x). \quad (2.13)$$

- Softmax:

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^K e^{x_j}}, \quad \text{for } i = 1, 2, \dots, K. \quad (2.14)$$

2.3 Variational Bayesian learning

2.3.1 Variational Bayes

A [Bayesian network \(BN\)](#) represents a probabilistic model as a [directed acyclic graph \(DAG\)](#), of which nodes are random variables, edges linking nodes represent the direct influence between them. If there is no edge between two nodes, we say the two corresponding

variables are conditionally independent. BN corresponds to the factorisation of the joint probability distribution $p(\mathbf{x}_1, \dots, \mathbf{x}_m)$:

$$p(\mathbf{x}_1, \dots, \mathbf{x}_m) = \prod_{i=1}^m p(\mathbf{x}_i | \boldsymbol{\pi}_i), \quad (2.15)$$

where $\boldsymbol{\pi}_i$ is the set of variables corresponding to the parent nodes of \mathbf{x}_i in the graph.

Let us consider a DAG where \mathbf{x} and \mathbf{z} represent a set of observed variables and latent variables respectively, with joint distribution $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$. We are interested in the posterior distribution:

$$p_\theta(\mathbf{z}|\mathbf{x}) = \frac{p_\theta(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{x})}. \quad (2.16)$$

The marginal distribution $p_\theta(\mathbf{x})$ in the denominator of Eq (2.16) requires the integral over \mathbf{z} :

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z}; \quad (2.17)$$

The **variational Bayesian (VB)** learning aims to approximate the posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ (with parameter θ) through optimisation. That is, given a family of tractable candidate distributions $q_\phi(\mathbf{z}|\mathbf{x})$ (ϕ is known as variational parameter), the optimisation finds the variational parameters such that $q_\phi(\mathbf{z}|\mathbf{x})$ is closest in the KL divergence to $p_\theta(\mathbf{z}|\mathbf{x})$ [131]. In fact, the general VB approaches do not minimise the KL divergence, but rather maximise an **evidence lower bound (ELBO)** to find variational parameters that provide a bound on the marginal likelihood as tight as possible [130, 131, 132]. To see the rationale, we need to look how to derive the ELBO and the KL divergence. Taking the logarithm of Eq (2.17), we have:

$$\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x}, \mathbf{z}) \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z} \quad (2.18a)$$

$$\geq \int q_\phi(\mathbf{z}|\mathbf{x}) [\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})] d\mathbf{z} \quad (\text{Jensen's inequality}) \quad (2.18b)$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})] \triangleq \mathcal{L}_{\theta, \phi}(\mathbf{x}), \quad (2.18c)$$

here $\mathcal{L}_{\theta,\phi}(\mathbf{x})$ denotes the ELBO. Now we derive the KL divergence:

$$\begin{aligned}
 D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x})||p_{\theta}(\mathbf{z}|\mathbf{x})) &= \int q_{\phi}(\mathbf{z}|\mathbf{x}) [\log q_{\phi}(\mathbf{z}|\mathbf{x}) - \log p_{\theta}(\mathbf{z}|\mathbf{x})] d\mathbf{z} \\
 &= \int q_{\phi}(\mathbf{z}|\mathbf{x}) \left[\log q_{\phi}(\mathbf{z}|\mathbf{x}) - \log \frac{p_{\theta}(\mathbf{x}, \mathbf{z})}{p_{\theta}(\mathbf{x})} \right] d\mathbf{z} \\
 &= - \int q_{\phi}(\mathbf{z}|\mathbf{x}) \log \frac{p_{\theta}(\mathbf{x}, \mathbf{z})}{q_{\phi}(\mathbf{z}|\mathbf{x})} d\mathbf{z} + \int q_{\phi}(\mathbf{z}|\mathbf{x}) \log p_{\theta}(\mathbf{x}) d\mathbf{z} \quad (2.19) \\
 &= - \underbrace{\int q_{\phi}(\mathbf{z}|\mathbf{x}) \log \frac{p_{\theta}(\mathbf{x}, \mathbf{z})}{q_{\phi}(\mathbf{z}|\mathbf{x})} d\mathbf{z}}_{\triangleq \mathcal{L}_{\theta,\phi}(\mathbf{x})} + \log p_{\theta}(\mathbf{x}) \underbrace{\int q_{\phi}(\mathbf{z}|\mathbf{x}) d\mathbf{z}}_{=1} \\
 &= - \mathcal{L}_{\theta,\phi}(\mathbf{x}) + \log p_{\theta}(\mathbf{x})
 \end{aligned}$$

From Eq (2.19) it's apparent that maximising the ELBO is the same as minimising the KL divergence of $q_{\phi}(\mathbf{z}|\mathbf{x})$ from $p_{\theta}(\mathbf{z}|\mathbf{x})$. It is also interesting to note that $D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x})||p_{\theta}(\mathbf{z}|\mathbf{x}))$ measures the gap between the ELBO and the log marginal likelihood, which is known as the tightness of the bound [132].

2.3.2 Stochastic gradient estimation for variational Bayes

We can optimise the ELBO via [stochastic gradient descent \(SGD\)](#), where differentiating the ELBO with respect to both ϕ and θ is needed. To resolve the problem that differentiating the ELBO with respect to ϕ is problematic and the Monte Carlo estimation is highly variable [130], the following reparameterisation is performed:

$$\epsilon \sim p(\epsilon); \quad \tilde{\mathbf{z}} = g_{\phi}(\epsilon, \mathbf{x}). \quad (2.20)$$

Here $g_{\phi}(\epsilon, \mathbf{x})$ is a differentiable transformation of an axillary noise variable ϵ and $\tilde{\mathbf{z}} \sim q_{\phi}(\mathbf{z}|\mathbf{x})$. Replacing the expectation with respect to $q_{\phi}(\mathbf{z}|\mathbf{x})$ with that with respect to $p(\epsilon)$ in Eq (2.18c), the ELBO is:

$$\mathcal{L}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{p(\epsilon)} [\log p_{\theta}(\mathbf{x}, \mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})] \quad (2.21)$$

The Monte Carlo estimate of the ELBO for the i -th data point is thus:

$$\tilde{\mathcal{L}}_{\theta,\phi}(\mathbf{x}^{(i)}) = \frac{1}{L} \underbrace{\sum_{l=1}^L \log p_{\theta}(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)}) - D_{\text{KL}}(q_{\phi}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})||p_{\theta}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}))}_{\text{negative reconstruction error}}, \quad (2.22)$$

where $\mathbf{z}^{(i,l)} = g_{\phi}(\epsilon^{(l)}, \mathbf{x}^{(i)})$ and $\epsilon^{(l)} \sim p(\epsilon)$.

As a result, the gradient of the ELBO, $\nabla_{\theta, \phi} \tilde{\mathcal{L}}_{\theta, \phi}(\mathbf{x}^{\mathcal{M}}, \epsilon)$ on a mini-batch $\mathbf{X}^{\mathcal{M}} = \{\mathbf{x}^{(i)}\}_{i=1}^{\mathcal{M}}$, can be computed for the SGD based [auto-encoding variational Bayes \(AEVB\)](#) algorithm. The resulting ELBO estimate is named as the [stochastic gradient variational Bayes \(SGVB\)](#) estimator [130, 132].

2.3.3 Variational auto-encoder

The [variational auto-encoder \(VAE\)](#) allows for efficient approximate inference, by combing a deep neural network with the above described SGVB estimation approach. Given a feature vector \mathbf{x} with corresponding latent variables \mathbf{z} , we introduce the variational distribution $q_{\phi}(\mathbf{z}|\mathbf{x})$ that approximates the posterior distribution $p_{\theta}(\mathbf{z}|\mathbf{x})$. The distributions of the generative and inference models for a VAE are:

$$\begin{aligned} p_{\theta}(\mathbf{x}, \mathbf{z}) &= p_{\theta}(\mathbf{x}|\mathbf{z})p_{\theta}(\mathbf{z}) \\ p_{\theta}(\mathbf{x}|\mathbf{z}) &= f_{\theta}(\mathbf{x}; \mathbf{z}) \\ p_{\theta}(\mathbf{z}) &= \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}) \\ q_{\phi}(\mathbf{z}|\mathbf{x}) &= \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mathbf{I}) \end{aligned} \tag{2.23}$$

Here $f_{\theta}(\mathbf{x}; \mathbf{z})$ is a suitable distribution (Normal distribution for continuous data; Bernoulli distribution for binary data) with learnable parameter θ . The decoder network $p_{\theta}(\mathbf{x}|\mathbf{z})$ is a sequence of fully connected neural network layers modelling a non-linear function $d\text{NN}_{\theta} : \mathbf{z} \rightarrow \mathbf{x}$. And $q_{\phi}(\mathbf{z}|\mathbf{x})$ is the inference network (encoder) with learnable parameter ϕ , $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ form the outputs of the encoding network. The non-linear function $e\text{NN}_{\phi} : \mathbf{x} \rightarrow \mathbf{z}$ is modelled by a neural network. The structure of the encoder network mirrors that of the decoding network. The reparameterisation trick (Eq (2.20)) is applied in the sampling step:

$$\begin{aligned} \mathbf{z}^{(i,l)} &= g_{\phi}(\epsilon^{(l)}, \mathbf{x}^{(i)}), \\ \text{where } g_{\phi}(\mathbf{x}^{(i)}, \epsilon^{(l)}) &= \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{2(i)} \odot \epsilon^{(l)} \text{ and } \epsilon^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \end{aligned} \tag{2.24}$$

In Eq (2.24), \odot denotes the elementwise product, i and l represent data point index and Monte Carlo sampling index respectively.

Let us denote the observation data and corresponding latent variables pair as $\{(\mathbf{x}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^n$, $\mathbf{x}^{(i)} \in \mathbf{X} \subseteq \mathbb{R}^d$, $\mathbf{z}^{(i)} \in \mathbb{R}^J$. Plugging the detailed formulas for the distributions into Eq (2.22), the gradient of the ELBO $\nabla_{\theta, \phi} \tilde{\mathcal{L}}_{\theta, \phi}(\mathbf{x}^{\mathcal{M}}, \epsilon)$ on a mini-batch $\mathbf{X}^{\mathcal{M}} = \{\mathbf{x}^{(i)}\}_{i=1}^{\mathcal{M}}$, can be computed for the [AEVB](#) algorithm as:

1. When $f_\theta(\mathbf{x}^{(i)}; \mathbf{z}^{(i)})$ is a Normal distribution:

$$\begin{aligned}
 \nabla_\theta \mathcal{L}_{\theta, \phi}(\mathbf{x}^M, \epsilon) &\simeq \nabla_\theta \frac{1}{L} \sum_{i=1}^M \sum_{l=1}^L \log p_\theta(\mathbf{x}^{(i)} | \mathbf{z}^{(i,l)}) - D_{\text{KL}}(q_\phi(\mathbf{z}^{(i)} | \mathbf{x}^{(i)}) || \mathcal{N}(\mathbf{0}, \mathbf{I})) \\
 &= \nabla_\theta \frac{1}{L} \sum_{i=1}^M \sum_{l=1}^L \log e^{-\frac{1}{2} \|\mathbf{x}^{(i)} - \text{dNN}_\theta(\mathbf{z}^{(i,l)})\|^2} \\
 &= - \nabla_\theta \underbrace{\frac{1}{2L} \sum_{i=1}^M \sum_{l=1}^L \|\mathbf{x}^{(i)} - \text{dNN}_\theta(\mathbf{z}^{(i,l)})\|^2}_{\triangleq \mathcal{L}_{\text{reconstruction}}}.
 \end{aligned} \tag{2.25}$$

Here \simeq denotes estimation.

$$\begin{aligned}
 \nabla_\phi \mathcal{L}_{\theta, \phi}(\mathbf{x}^M, \epsilon) &\simeq \nabla_\phi \sum_{i=1}^M \sum_{l=1}^L \mathbb{E}_{\epsilon^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\log p_\theta(\mathbf{x}^{(i)} | \mathbf{z}^{(i,l)})] \Big|_{\mathbf{z}^{(i,l)} = \boldsymbol{\mu}_\phi^{(i)} + \boldsymbol{\sigma}_\phi^{2(i)} \odot \epsilon^{(l)}} \\
 &\quad - \nabla_\phi \sum_{i=1}^M D_{\text{KL}}(q_\phi(\mathbf{z}^{(i)} | \mathbf{x}^{(i)}) || \mathcal{N}(\mathbf{0}, \mathbf{I})) \\
 &= \nabla_\phi \sum_{i=1}^M \sum_{l=1}^L \mathbb{E}_{\epsilon^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\log p_\theta(\mathbf{x}^{(i)} | \boldsymbol{\mu}_\phi^{(i)} + \boldsymbol{\sigma}_\phi^{2(i)} \odot \epsilon^{(l)})] \\
 &\quad - \nabla_\phi \underbrace{\frac{1}{2} \sum_{i=1}^M \sum_{j=1}^J (1 + \log \sigma_j^{2(i)} - \mu_j^{2(i)} - \sigma_j^{2(i)})}_{\triangleq \mathcal{L}_{\text{KL}}} \\
 &= \nabla_\phi - \frac{1}{2L} \sum_{i=1}^M \sum_{l=1}^L \|\mathbf{x}^{(i)} - \text{dNN}_\theta(\mathbf{z}^{(i,l)})\|^2 \Big|_{\mathbf{z}^{(i,l)} = \boldsymbol{\mu}_\phi^{(i)} + \boldsymbol{\sigma}_\phi^{2(i)} \odot \epsilon^{(l)}} \\
 &\quad - \nabla_\phi \mathcal{L}_{\text{KL}} \\
 &= - \nabla_\phi (\mathcal{L}_{\text{reconstruction}} + \mathcal{L}_{\text{KL}}).
 \end{aligned} \tag{2.26}$$

2. When $f_\theta(\mathbf{x}^{(i)}; \mathbf{z}^{(i)})$ is a Bernoulli distribution:

$$\begin{aligned}
 \nabla_\theta \mathcal{L}_{\theta, \phi}(\mathbf{x}^M, \epsilon) &\simeq \nabla_\theta \frac{1}{L} \sum_{i=1}^M \sum_{l=1}^L \log p_\theta(\mathbf{x}^{(i)} | \mathbf{z}^{(i,l)}) - D_{\text{KL}}(q_\phi(\mathbf{z}^{(i)} | \mathbf{x}^{(i)}) || \mathcal{N}(\mathbf{0}, \mathbf{I})) \\
 &= \nabla_\theta \underbrace{\frac{1}{L} \sum_{i=1}^M \sum_{v=1}^d \sum_{l=1}^L x^{(i,v)} \log \hat{x}^{(i,v,l)} + (1 - x^{(i,v)}) \log(1 - \hat{x}^{(i,v,l)})}_{\triangleq -\mathcal{L}_{\text{reconstruction}}} \Big|_{\hat{\mathbf{x}}^{(i,l)} = \text{dNN}_\theta(\mathbf{z}^{(i,l)})};
 \end{aligned} \tag{2.27}$$

$$\begin{aligned}
\nabla_{\phi} \mathcal{L}_{\theta, \phi}(\mathbf{x}^M, \epsilon) &\simeq \nabla_{\phi} \sum_{i=1}^M \sum_{l=1}^L \mathbb{E}_{\epsilon^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\log p_{\theta}(\mathbf{x}^{(i)} | \mathbf{z}^{(i,l)})] \Big|_{\mathbf{z}^{(i,l)} = \boldsymbol{\mu}_{\phi}^{(i)} + \boldsymbol{\sigma}_{\phi}^{2(i)} \odot \epsilon^{(l)}} \\
&\quad - \nabla_{\phi} \sum_{i=1}^M D_{\text{KL}}(q_{\phi}(\mathbf{z}^{(i)} | \mathbf{x}^{(i)}) || \mathcal{N}(\mathbf{0}, \mathbf{I})) \\
&= \nabla_{\phi} \sum_{i=1}^M \sum_{l=1}^L \mathbb{E}_{\epsilon^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\log p_{\theta}(\mathbf{x}^{(i)} | \boldsymbol{\mu}_{\phi}^{(i)} + \boldsymbol{\sigma}_{\phi}^{2(i)} \odot \epsilon^{(l)})] \\
&\quad - \nabla_{\phi} \underbrace{\frac{1}{2} \sum_{i=1}^M \sum_{j=1}^J (1 + \log \sigma_j^{2(i)} - \mu_j^{2(i)} - \sigma_j^{2(i)})}_{\mathcal{L}_{\text{KL}}} \\
&= \nabla_{\phi} \frac{1}{L} \sum_{i=1}^M \sum_{v=1}^d \sum_{l=1}^L x^{(i,v)} \log \hat{x}^{(i,v,l)} + (1 - x^{(i,v)}) \log(1 - \hat{x}^{(i,v,l)}) \Big|_{\hat{\mathbf{x}}^{(i,l)} = \text{dNN}_{\theta}(\boldsymbol{\mu}_{\phi}^{(i)} + \boldsymbol{\sigma}_{\phi}^{2(i)} \odot \epsilon^{(l)})} - \nabla_{\phi} \mathcal{L}_{\text{KL}} \\
&= - \nabla_{\phi} (\mathcal{L}_{\text{reconstruction}} + \mathcal{L}_{\text{KL}}).
\end{aligned} \tag{2.28}$$

2.4 Deep generative models for semi-supervised learning

Semi-supervised learning is strongly needed in cancer research where only a small proportion of datasets have label annotations and data labelling for large scale data is impractical due to high costs or possible ethical issues. The generative nature of VAE makes it suitable for semi-supervised learning where the learning tasks involve labelled data $\mathcal{D}^l = \{(\mathbf{x}^{l,(i)}, y^{l,(i)})\}_{i=1}^n$, $\mathbf{x}^{l,(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$, $y^{l,(i)} \in \{1, \dots, K\}$ with joint probability distribution $p^l(\mathbf{x}, y)$ and unlabelled data $\mathcal{D}^u = \{\mathbf{x}^{u,(i)}\}_{i=1}^m$, $\mathbf{x}^{u,(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$ that are drawn from probability distribution $p^u(\mathbf{x})$. Depending on the assumed graphical probabilistic model (i.e. the DAG) that represents the joint distribution, one can derive a range of generative models. This enables jointly optimising the semi-supervised classifier and variational approximation parameters.

The M1 and M2 models are among the earliest work extending the VAE to semi-supervised learning [136]. Inspired by this, other new variants of semi-supervised VAE are proposed [137, 138]. Chapter 4 which contains our recent work [139] presents a new variant VAE based on the DAG shown in Fig 2.1(C). Here I shortly introduce the idea of extending the VAE to semi-supervised generative model using the M1 and M2 models as examples. The M1 model is based on a standard VAE (the corresponding DAG is shown in Fig 2.1(A)) that embeds high-dimensional features of both labelled and unlabelled

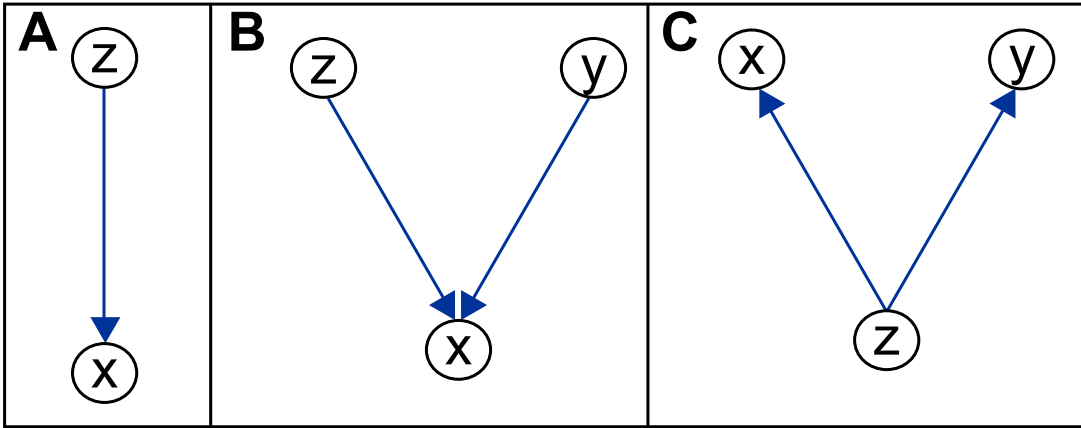


Figure 2.1: Graphical models for joint distributions. (A) Graphical model for standard VAE (M1 model). (B) Graphical model for M2 model. (C) Graphical model for MFmap [139].

samples into low-dimensional latent space. The embedded features of labelled data are then used to train a classifier (e.g. SVM), which also predicts the labels of unlabelled samples.

The M2 model is based on the DAG shown in Fig 2.1(B), where the features x are generated from latent variables y and z for unlabelled data. For the labelled data, y is observed, we only have latent variables z . Following the principle of VAE, the joint distribution and marginal distribution are:

$$p_{\theta}(\mathbf{x}, y) = \int p_{\theta}(\mathbf{x}|y, \mathbf{z})p_{\theta}(y)p_{\theta}(\mathbf{z})d\mathbf{z} \quad (2.29a)$$

$$p_{\theta}(\mathbf{x}) = \sum_y p_{\theta}(\mathbf{x}, y), \quad (2.29b)$$

The approximate variational inference networks $q_{\phi}(\mathbf{z}|y, \mathbf{x})$ and $q_{\phi}(\mathbf{z}, y|\mathbf{x})$ are introduced for labelled and unlabelled data respectively. The authors assumed that the conditional independence $\mathbf{z} \perp y|\mathbf{x}$ in Eq (2.29b) holds in the inference network $q_{\phi}(\mathbf{z}, y|\mathbf{x})$, that is, $q_{\phi}(\mathbf{z}, y|\mathbf{x}) = q_{\phi}(\mathbf{z}|\mathbf{x})q_{\phi}(y|\mathbf{x})$ [136].

The probabilistic model of the generative and inference processes for a semi-supervised

VAE are specified as:

$$p_\theta(\mathbf{x}, y, \mathbf{z}) = p_\theta(\mathbf{x}|y, \mathbf{z})p_\theta(y)p_\theta(\mathbf{z}) \quad (2.30a)$$

$$p_\theta(\mathbf{x}|y, \mathbf{z}) = f_\theta(\mathbf{x}; y, \mathbf{z}) \quad (2.30b)$$

$$p_\theta(y) = \text{Cat}(y|\pi_y) \quad (2.30c)$$

$$p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}) \quad (2.30d)$$

$$q_\phi(\mathbf{z}, y|\mathbf{x}) = q_\phi(\mathbf{z}|\mathbf{x})q_\phi(y|\mathbf{x}) \quad (2.30e)$$

$$= q_\phi(\mathbf{z}|y, \mathbf{x})q_\phi(y|\mathbf{x}) \quad (2.30f)$$

$$q_\phi(\mathbf{z}|y, \mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x})) \quad (2.30g)$$

$$q_\phi(y|\mathbf{x}) = \text{Cat}(y|\pi_\phi(\mathbf{x})) \quad (2.30h)$$

Here $f_\theta(\mathbf{x}; y, \mathbf{z})$ is a suitable distribution represented as the deep decoder network $y, \mathbf{z} \mapsto f_\theta(\cdot; y, \mathbf{z})$; π_y are the probabilities of one-hot encoded class labels; $\pi_\phi(\mathbf{x})$ is the categorical parameter of the classifier network $q_\phi(y|\mathbf{x})$; $\boldsymbol{\mu}_\phi(\mathbf{x})$ and $\boldsymbol{\sigma}^2_\phi(\mathbf{x})$ are the outputs of the encoder network. The ELBO of labelled data is expressed as:

$$\mathcal{L}_{\theta, \phi}^l(\mathbf{x}, y) = \mathbb{E}_{q_\phi(\mathbf{z}|y, \mathbf{x})} [\log p_\theta(\mathbf{x}|y, \mathbf{z})] - D_{\text{KL}}(q_\phi(\mathbf{z}|y, \mathbf{x}) || p_\theta(\mathbf{z})), \quad (2.31)$$

and the ELBO of unlabelled data is

$$\mathcal{L}_{\theta, \phi}^u(\mathbf{x}) = \sum_y q_\phi(y|\mathbf{x}) \mathcal{L}_{\theta, \phi}^l(\mathbf{x}, y) + \underbrace{\mathcal{H}(q_\phi(y|\mathbf{x}))}_{\text{entropy}}. \quad (2.32)$$

To make label prediction $q_\phi(y|\mathbf{x})$ to be trained on both labelled and unlabelled data, the cross entropy loss \mathcal{L}_{CE} is added to the negative ELBO of the labelled data to form the final loss function of the M2 model:

$$\mathcal{L}_{\theta, \phi}(\mathbf{x}, y) = - \sum_{(\mathbf{x}, y) \sim p^l(\mathbf{x}, y)} \mathcal{L}_{\theta, \phi}^l(\mathbf{x}, y) - \sum_{(\mathbf{x}) \sim p^u(\mathbf{x})} \mathcal{L}_{\theta, \phi}^u(\mathbf{x}) - \alpha \cdot \underbrace{\mathbb{E}_{p^l(\mathbf{x}, y)} [\log q_\phi(y|\mathbf{x})]}_{\triangleq \mathcal{L}_{\text{CE}}}. \quad (2.33)$$

Here α is a tunable hyperparameter controlling the relative weight between the generative model and the classifier. With the defined objective of semi-supervised VAE, the stochastic backpropagation strategies and reparametrisation trick discussed in Section 2.3 can be applied to optimise the parameters.

Distinct and common features of numerical and structural chromosomal instability across different cancer types

This chapter has been published as **peer-reviewed journal paper**:

© 2022 by the authors.

Licensed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0>).

X. Zhang and M. Kschischo. “Distinct and Common Features of Numerical and Structural Chromosomal Instability across Different Cancer Types.” In: *Cancers* 14.6 (2022). DOI: [10.3390/cancers14061424](https://doi.org/10.3390/cancers14061424)

Synopsis: Most cancer cells display chromosomal instability (CIN) phenotype, defined as elevated rates of accumulating whole chromosome changes (W-CIN) or structurally abnormal chromosomes (S-CIN). Both W-CIN and S-CIN have broad clinical implications. While W-CIN could coexist with S-CIN in one cell, they arise through different mechanisms. For better understanding of the commonalities and difference between W-CIN and S-CIN across 33 cancer types, we performed extensive association analyses between W-CIN/S-CIN and various features including prognosis, drug sensitivity, pathway activity, somatic mutation and copy number variation. We found that whole genome doubling is universally strongly associated with high W-CIN, homologous recombination deficiency is strongly associated with high S-CIN in almost all studied cancer types. We show prognostic values of W-CIN and S-CIN are cancer type specific. We report currently available drugs hardly kill high CIN cancer cells. We propose *CKS1B* as a potential candidate S-CIN target. We suggest a copy number dependent mechanism to promote *PI3K* signalling in high S-CIN

cancer cells.

Contributions of thesis author: methodology, formal analysis, investigation, software, data curation, writing, review and editing, visualization.

Article

Distinct and Common Features of Numerical and Structural Chromosomal Instability across Different Cancer Types

Xiaoxiao Zhang ^{1,2}  and Maik Kschischo ^{1,*} 

¹ Department of Mathematics and Technology, University of Applied Sciences Koblenz, 53424 Remagen, Germany; zhang@hs-koblenz.de

² Department of Informatics, Technical University of Munich, 81675 Munich, Germany

* Correspondence: kschischo@rheinahr-campus.de; Tel.: +49-2642932330

Simple Summary: Many cancer cells are chromosomally unstable, a phenotype describing a tendency for accumulating chromosomal aberrations. Entire chromosomes tend to be gained or lost, which is called whole chromosome instability (W-CIN). Structural chromosomal instability (S-CIN) describes an increased rate of gaining, losing or translocating smaller parts of chromosomes. Here, we analyse data from 33 cancer types to find differences and commonalities between W-CIN and S-CIN. We find that W-CIN is strongly linked to whole genome doubling (WGD), whereas S-CIN is associated with a specific DNA damage repair pathway. Both W-CIN and S-CIN are difficult to target using currently available compounds and have distinct prognostic values. The activity of the drug resistance gene *CKS1B* is associated with S-CIN, which merits further investigation. In addition, we identify a potential copy number-based mechanism promoting signalling of the important *PI3K* cancer pathway in high-S-CIN tumours.

Abstract: A large proportion of tumours is characterised by numerical or structural chromosomal instability (CIN), defined as an increased rate of gaining or losing whole chromosomes (W-CIN) or of accumulating structural aberrations (S-CIN). Both W-CIN and S-CIN are associated with tumorigenesis, cancer progression, treatment resistance and clinical outcome. Although W-CIN and S-CIN can co-occur, they are initiated by different molecular events. By analysing tumour genomic data from 33 cancer types, we show that the majority of tumours with high levels of W-CIN underwent whole genome doubling, whereas S-CIN levels are strongly associated with homologous recombination deficiency. Both CIN phenotypes are prognostic in several cancer types. Most drugs are less efficient in high-CIN cell lines, but we also report compounds and drugs which should be investigated as targets for W-CIN or S-CIN. By analysing associations between CIN and bio-molecular entities with pathway and gene expression levels, we complement gene signatures of CIN and report that the drug resistance gene *CKS1B* is strongly associated with S-CIN. Finally, we propose a potential copy number-dependent mechanism to activate the *PI3K* pathway in high-S-CIN tumours.

Keywords: whole chromosomal instability; structural chromosomal instability; whole genome doubling; integrative analysis; *PI3K* oncogenic activation



Citation: Zhang, X.; Kschischo, M. Distinct and Common Features of Numerical and Structural Chromosomal Instability across Different Cancer Types. *Cancers* **2022**, *14*, 1424. <https://doi.org/10.3390/cancers14061424>

Academic Editors: Kozo Tanaka and Henry Heng

Received: 11 January 2022

Accepted: 27 February 2022

Published: 10 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

A large proportion of human tumours exhibits abnormal karyotypes with gains and losses of whole chromosomes or structural aberrations of parts of chromosomes [1–3]. In many cases, these karyotypic changes are the result of ongoing chromosomal instability (CIN), which is defined as an increased rate of chromosomal changes. Accordingly, two major forms of CIN can be distinguished: Whole chromosome instability (W-CIN), which is also called numerical CIN, refers to the ongoing acquisition of gains and losses of whole chromosomes. Structural CIN (S-CIN) is characterised by an increased rate of acquiring structural changes in chromosomes including, amongst other things, amplifications and

deletions, inversions, duplications and balanced or unbalanced translocations [1–4]. CIN is to be distinguished from polyploidy, where the whole set of chromosomes is increased. In cross-sectional tumour samples, W-CIN manifests itself by an abnormal and unequal number of chromosomes, whereas the S-CIN phenotype is characterised by segmental aneuploidy, i.e., gains and losses of chromosome segments.

Although W-CIN can induce S-CIN and vice versa, both types of CIN arise through distinct molecular characteristics. Whilst W-CIN is caused by chromosome missegregation during mitosis, S-CIN is commonly attributed to errors in the repair of DNA double-strand breaks [5,6]. Both types of CIN are intimately related to DNA replication stress [7,8], which not only induces CIN [9,10], but also occurs as an immediate short-term response to aneuploidy and CIN [11].

Aneuploidy and CIN have typically detrimental effects on cell fitness and proliferation [5,11,12]. Therefore, it was unclear why CIN is often associated with poor patient survival and more aggressive disease progression [1,13,14]. Stratification of breast cancer patient samples into low, intermediate and high CIN groups revealed that patients with intermediate levels of CIN had the worst survival, whereas the low and high CIN groups had a better prognosis [15,16]. These results hinted at mechanisms for tolerating CIN in order to survive the stresses provoked by chromosomal aberrations. The CIN tolerance mechanisms are currently not completely understood [17], but one important recurring event is a loss of *TP53* function, which otherwise prevents the propagation of CIN cells [18].

The CIN 70 signature is a set of genes whose expression is correlated with functional segmental aneuploidy [1]. It was one of the first CIN signatures and it is enriched by genes involved in cell cycle regulation and mitosis. CIN 70 was later criticised for rather being a marker for cell proliferation than for CIN, because it reflects evolved aneuploid cancer cell populations which have adapted their genome instead of a primary response to CIN [19]. These studies highlighted that we have to distinguish between acute responses to aneuploidy and CIN [11], mechanisms for tolerating CIN [17] and the cellular programme [20,21] and genetic alterations [22] acquired by evolved CIN cells. These cellular programmes might differ between cancer cell lines and tumours, partially as a result of treatment effects or as a result of interactions with the tumour microenvironment. Recently, it was discovered that chromosome segregation errors as well as replication stress activate the anti-viral immune *cGAS-STING* pathway, which responds to genomic double-stranded DNA in the cytosol [2,23]. This interesting research links cancer cell intrinsic processes with cell to cell communication and immune response in the tumour microenvironment.

The phenotypic plasticity in combination with tumour heterogeneity enables CIN tumours to rapidly adapt to diverse stress conditions. It has been shown that CIN permits and accelerates the acquisition of resistance against anti-cancer therapies by acquiring recurrent copy number changes [24,25]. This acquired drug resistance could potentially exacerbate the intrinsic drug resistance [26] of many CIN cells, which highlights the need to better understand genomic changes of CIN tumours in the context of anti-cancer treatment.

Computational studies of cancer genomic data have provided valuable insights into CIN [1,19–22] and aneuploidy [27] and guided experimental and clinical testing. However, most of these studies did not differentiate between W-CIN and S-CIN. Here, we analyse cancer genomic data to better understand commonalities and differences between both types of CIN. In particular, we analyse, across multiple cancer types, the genomic landscape of S-CIN and W-CIN, their relationship to prognosis and drug sensitivity, the relationship between CIN, somatic point mutations and specific copy number variations and propose a new link between S-CIN and the *PI3K* oncogenic pathway.

2. Materials and Methods

2.1. TCGA Pan-Cancer Clinical and Molecular Data

We analysed chromosome instability of 33 primary tumour types from The Cancer Genome Atlas (TCGA): Adrenocortical carcinoma (ACC, $n = 89$); bladder urothelial carcinoma (BLCA, $n = 399$); breast invasive carcinoma (BRCA, $n = 1039$); cervical and

endocervical cancers (CESC, $n = 294$); cholangiocarcinoma (CHOL, $n = 36$); colon adenocarcinoma (COAD, $n = 420$); lymphoid neoplasm diffuse large B-cell lymphoma (DLBC, $n = 47$); esophageal carcinoma (ESCA, $n = 162$); glioblastoma multiforme (GBM, $n = 556$); head and neck squamous cell carcinoma (HNSC, $n = 510$); kidney chromophobe (KICH, $n = 65$); kidney renal clear cell carcinoma (KIRC, $n = 480$); kidney renal papillary cell carcinoma (KIRP, $n = 280$); acute myeloid leukaemia (LAML, $n = 124$); brain lower grade glioma (LGG, $n = 506$); liver hepatocellular carcinoma (LIHC, $n = 361$); lung adenocarcinoma (LUAD, $n = 490$); lung squamous cell carcinoma (LUSC, $n = 482$); mesothelioma (MESO, $n = 81$); ovarian serous cystadenocarcinoma (OV, $n = 550$); pancreatic adenocarcinoma (PAAD, $n = 165$); pheochromocytoma and paraganglioma (PCPG, $n = 160$); prostate adenocarcinoma (PRAD, $n = 471$); rectum adenocarcinoma (READ, $n = 154$); sarcoma (SARC, $n = 244$); skin cutaneous melanoma (SKCM, $n = 104$); stomach adenocarcinoma (STAD, $n = 427$); testicular germ cell tumours (TGCT, $n = 133$); thyroid carcinoma (THCA, $n = 463$); thymoma (THYM, $n = 106$); uterine corpus endometrial carcinoma (UCEC, $n = 512$); uterine carcinosarcoma (UCS, $n = 56$); uveal melanoma (UVM, $n = 80$).

We also calculated karyotypic complexity scores as surrogate measures for CIN (see Section 2.4) for 391 metastatic tumour tissues, 8719 blood-derived normal tissues and 2207 solid normal tissues.

The TCGA pan-cancer molecular and clinical data were downloaded from the Pan-Cancer Atlas [28]. The file names for different data modalities are: Copy number segment data from `broad.mit.edu_PANCAN_Genome_Wide_SNP_6_whitelisted.seg`; ABSOLUTE [29] inferred ploidy data from `TCGA_mastercalls.abs_tables_JSedit.fixed.txt`; normalised and batch effect-corrected gene expression profile from `EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv`; clinical data from `TCGA-CDR-SupplementalTableS1.xlsx`; PARADIGM [30] inferred pathway activity data from `merge_merged_reals.tar.gz`.

2.2. CCLE Molecular and Sample Annotation Data

Cell line multiomics data were downloaded from the Broad-Novartis Cancer Cell Line Encyclopedia (CCLE) [31]. In particular, the copy number segment data are located in `CCLE_copynumber_2013-12-03.seg.txt`. Gene expression profiles and sample annotations are located in `CCLE_RNAseq_genes_rpkm_20180929.gct.gz` and `Cell_lines_annotations_20181226.txt`. The binary alteration matrix is located in `CCLE_MUT_CNA_AMP_DEL_binary_Revealer.gct`. Sample ploidy data estimated using the ABSOLUTE algorithm [29] are located in `CCLE_ABSOLUTE_combined_20181227.xlsx`.

2.3. CTRP Drug Screening Data

We collected cell line pharmacological profiling data from the Cancer Therapeutics Response Portal (CTRP [32], `CTRPv2.0_2015_ctd2_ExpandedDataset.zip`). The drug resistance quantified by the area under the dose–response curve (AUC) was min–max normalised, i.e., the minimum value was subtracted and the resulting values were rescaled by the original range of the AUC. These min–max normalised AUC values have a range between zero and one. From this, we computed the drug sensitivity index as $1 - \text{normalised AUC}$ with values in the range between 0 (highest resistance) and 1 (most sensitive).

2.4. Karyotypic Complexity Scores (CIN Scores)

We implemented three different karyotypic complexity scores [7] as surrogate measures for CIN in both TCGA bulk tumours and CCLE cell lines: The numerical complexity score (NCS), the structural complexity score (SCS) and the weighted genome instability index (WGII). For brevity, we will refer to these karyotypic complexity scores as CIN scores. Here, we detail the procedures for computing each score.

The NCS is calculated by the following steps:

Step 1: Inferring sample ploidy using the ABSOLUTE algorithm [29].

- Step 2: Rounding the ploidy and segment-wise copy numbers of each sample to the nearest integer.
- Step 3: Identifying whole chromosomal changes in each chromosome. For each chromosome in a sample, this chromosome is counted as a whole chromosomal change if at least 75% of the chromosome has integer copy numbers greater or less than the sample integer ploidy.
- Step 4: Summing up the whole chromosome changes across all 22 autosomes yields the sample NCS.

The SCS is calculated by the following steps:

- Step 1: Rounding the segment-wise copy numbers of each sample to the nearest integer.
- Step 2: Computing the modal copy number for each chromosome in each sample.
- Step 3: Identifying intra-chromosomal changes for each chromosome. Given a chromosome segment of a sample, this segment (with length ≥ 1 Mb) is counted as changed if its integer copy number is greater or less than the modal copy number of this chromosome.
- Step 4: Summing up all intra-chromosomal changes across all 22 autosomes yields the sample SCS.

The WGII is calculated by the following steps:

- Step 1: Inferring sample ploidy using the ABSOLUTE algorithm [29].
- Step 2: Rounding the ploidy and segment-wise copy numbers of each sample to the nearest integer.
- Step 3: Identifying chromosome changes for each chromosome. Given a chromosome segment of a sample, this segment is counted as changed if the integer copy number of this segment is greater or less than the sample integer ploidy.
- Step 4: Calculating the percentage of the chromosome change for each chromosome.
- Step 5: Calculating the mean percentage of the chromosome change of all 22 autosomes, resulting in sample WGII.

2.5. Association Analysis between CIN and Genome Instability

Aneuploidy scores (ASs) of samples are taken from [27], Supplementary Table S2, tumour characteristics including homologous recombination deficiency (HRD), silent mutation rate (SMR), non-silent mutation rate (NSMR), proliferation and intra-tumour heterogeneity (ITH) were collected from [33], Supplementary Table S1. Microsatellite instability (MIN) scores are collected from [34], Supplementary Table S5. The correlations of these genome instability scores and NCS or SCS were quantified by Spearman correlation coefficients.

2.6. Survival Analysis

We performed survival analysis using the survival R package [35]. Patients were stratified according to their median CIN score of all patients from the same cohort. A univariate Cox proportional hazards model was fitted to evaluate the association between patient survival and CIN and the log rank test was applied to calculate the *p*-value for the survival difference between high-CIN and low-CIN groups. Survival curves were visualised using ggsurvplot implemented in the survminer R package [36].

2.7. Treatment Response Analysis

We labelled patients with complete/partial response to chemotherapy or radiation therapy as responders and the other patients as non-responders. A Wilcoxon rank sum test was used to evaluate the differences of the NCS and SCS in the responder and non-responder groups.

2.8. Identification of Candidate Compounds Selectively Targeting CIN

Spearman correlation coefficients between drug sensitivity (defined in Section 2.3) and CIN were computed for 545 CTRP compounds. Compounds with multiple testing adjusted $p \leq 0.05$ and median drug sensitivity >0.5 were considered as candidate compounds selectively targeting low-CIN cancer cells (compounds with negative correlation coefficients) or high-CIN cancer cells (compounds with positive correlation coefficients).

2.9. Association Analysis between CIN and PARADIGM Pathway Activities

We collected the sample-wise PARADIGM pathway activity matrix from the Pan-Cancer Atlas [28] with the file name [merge_merged_reals.tar.gz](#). For each cancer type we computed the Spearman correlation coefficient between CIN score (NCS or SCS) and PARADIGM pathway activity and selected the top pathways corresponding to significant protein coding genes. We filtered genes/proteins whose PARADIGM pathway activities are strongly positively correlated with NCS or SCS (correlation coefficient ≥ 0.3) in more than seven cancer types.

2.10. Association Analysis between Somatic Alterations and CIN

We used the limma R package [37] for multiple linear regression analysis on CIN scores, using alteration status (mutation, copy number amplification or copy number deletion versus wild type) and cohort as predictor variables. To achieve sufficient statistical power, only alterations which occurred in more than 20 samples were included as predictors.

3. Results

3.1. Karyotypic Complexity Scores as Surrogate Measures for CIN

CIN is a dynamic feature of abnormal chromosomes, rendering its assessment in routine experimental settings difficult [38,39]. Assessing the degree of ongoing W-CIN or S-CIN requires time-resolved data to monitor the rate of mitotic errors or the rate of segmental gains or losses, respectively. An alternative is to use single cell analysis to quantify cell to cell karyotype heterogeneity within a population of cells. The latter approach is based on the assumption that the degree of CIN is reflected by the degree of karyotype heterogeneity.

Although these and other approaches have made considerable progress in recent years (see, e.g., [40] for a recent review), the number of patient-derived tumour samples across different cancer types providing such information is not sufficient for a statistically meaningful comparison across different cancer types. Instead, we use established karyotypic complexity scores which have been evaluated as good markers for the CIN phenotype [7,26]. Please note, however, that these scores derived from cross-sectional tumour data quantify the degree of aneuploidy or segmental aneuploidy, which is the result of both CIN and the selective pressures shaping the karyotype. As such, the karyotypic complexity scores cannot quantify ongoing CIN, but only reflect the chromosomal changes resulting from CIN and evolutionary adaptation and selection. Nevertheless, based on previous evidence [26] we assume here that these karyotypic complexity scores reflect features of the evolved CIN phenotype and refer to them as CIN scores.

As a surrogate score for the degree of W-CIN of a given tumour sample, we used the numerical complexity score (NCS) [7], which counts the number of whole chromosome gains/losses (defined as chromosomes with more than 75% of integer copy numbers higher or lower than the sample integer ploidy). The exact computation is given in Section 2.4. The degree of S-CIN was assessed by the structural complexity score (SCS), which is the number of structurally aberrant regions in the genome of a sample. A region in a chromosome is defined as structurally aberrant if it is longer than 1 Mb and its copy number deviates from the modal copy number of the chromosome (Section 2.4).

The weighted genome instability index (WGII) was previously used as a measure integrating both numerical and structural complexity (e.g., [7,12]). The WGII is the average percentage of changed genome relative to the sample ploidy [7], see again Section 2.4. We found that the WGII is highly correlated to the NCS (Pearson correlation coefficient: 0.99)

and we also provide the pan-cancer analysis results using the WGII for comparison in PDF S1.

Please note one important difference between our work and previous analysis (e.g., [7,12]) of karyotypic complexity scores: We used the ABSOLUTE algorithm for estimating the ploidy of the sample, whereas most previous work used the median copy number weighted by segment length across all segments [7]. The ABSOLUTE inferred ploidy has been validated using fluorescence-activated cell sorting, spectral karyotyping and DNA-mixing experiments [29].

3.2. Landscape of W-CIN and S-CIN across Human Cancers

In total, we calculated NCS and SCS for 21,633 samples including 10,308 primary tumours, 391 metastatic tumours and 10,934 normal tissues derived from 33 cancer types. The distribution of NCS varies drastically across cancer types (Figure 1A), but shows a characteristic bimodal pattern, see also the pan-cancer histogram on the right hand side. The colour coding of the whole genome doubling (WGD) status indicates that tumour samples with high levels of NCS are often characterised by a WGD event. Please note that this is not an artefact of the NCS, which is measured relative to the sample ploidy. This suggests that WGD is an important mechanism inducing W-CIN in many cancer types. However, the exception is kidney chromophobe (KICH), where WGD events seem to be rare, but high levels of the NCS can still be observed. In this cancer type, there is also no clear bimodal pattern, suggesting that mechanisms other than WGD drive W-CIN in KICH. Even in cancers where the bimodal pattern suggests a clear separation between numerically unstable and numerically stable tumours, it is difficult to define a universal NCS threshold distinguishing numerically stable from W-CIN tumours across cancer types. For example, in ovarian serous cystadenocarcinoma (OV), one can distinguish low- and high-NCS groups with WGD, but the overall level of the NCS is much higher than that in other cancer types. Similarly, for adrenocortical carcinoma (ACC), there are many patients with high levels of NCS even in the group of samples which did not undergo WGD. This suggests that processes other than WGD can drive a certain degree of W-CIN in these tumours.

In contrast to the NCS distribution, the pan-cancer distribution of SCS peaks at low values and is right skewed (Figure 1B). This indicates that most tumours are structurally chromosomally stable, but some can exhibit extreme levels of S-CIN. Overall, there is no functional relationship between NCS and SCS (Figure A1).

The distribution of SCS indicates a high degree of tumour heterogeneity within the same cancer type and across cancer types. Ovarian serous cystadenocarcinoma (OV), uterine carcinosarcoma (UCS) and sarcoma (SARC) show the highest SCS (Figure 1B) and many samples within these tumours also exhibit high NCS (compare Figure 1A). Both types of CIN occur in many OV, esophageal carcinoma (ESCA) and BRCA samples, whereas thyroid carcinoma (THCA), thymoma (THYM) and acute myeloid leukaemia (LAML) samples are typically both structurally and numerically stable. Cancer types previously recognised as those dominated by the CIN phenotype [41], including stomach adenocarcinoma (STAD), colon adenocarcinoma (COAD), uterine corpus endometrial carcinoma (UCEC), OV, UCS and prostate adenocarcinoma (PRAD) have extremely heterogeneous SCS.

We also checked for associations of CIN with other types of genetic instability by correlating the NCS and SCS with different features: Aneuploidy score (AS), homologous recombination deficiency (HRD), silent mutation rate (SMR), non-silent mutation rate (NSMR) and intra-tumour heterogeneity (ITH). The NCS is positively associated with the aneuploidy score (Figure 1C) across cancer types [42]. HRD is consistently positively associated with the SCS (Figure 1D), suggesting that impaired repair of double-strand DNA breaks might be a key driver of S-CIN.

CIN and microsatellite instability (MIN) are usually considered mutually exclusive [38]. Indeed, most MIN tumours have low NCS and SCS, but some MIN samples which underwent WGD can also exhibit signs of W-CIN and S-CIN (Figure A2A).

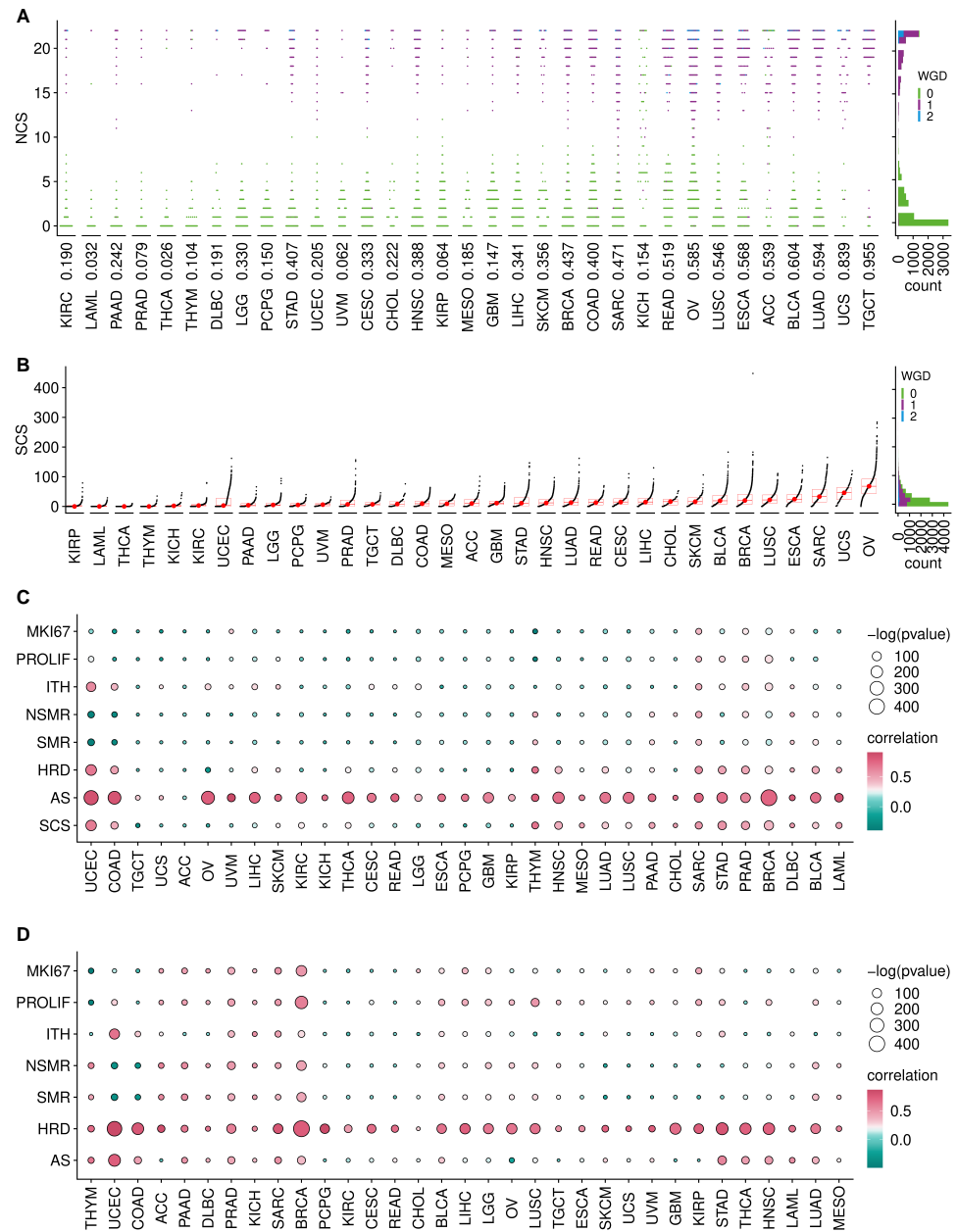


Figure 1. Distribution of CIN scores and their association with genetic instability. **(A)** Left: NCS for TCGA tumour samples (dots) from different cancer types, sorted according to median NCS. The colour coding indicates the WGD status and the number below each beeswarm plot is the proportion of samples which underwent WGD. Right: Pan-cancer histogram of NCS. **(B)** SCS for tumour samples from different cancer types, ordered by their median SCS. Right: Pan-cancer histogram of SCS with colours indicating WGD status. **(C,D)** Correlation between NCS **(C)** or SCS **(D)** with different indices for genetic instability, intra-tumour heterogeneity and proliferation: *MKI67* expression, proliferation rates (PROLIF), intra-tumour heterogeneity (ITH), non-silent mutation rate (NSMR), silent mutation rate (SMR), homologous recombination deficiency (HRD) and aneuploidy score (AS). Data for these indices were collected from [27,33].

To check for a potential link between CIN and proliferation, we used a proliferation index [33] and the expression of the *MKI67* marker for proliferation. In many cancers, including

BRCA, SARC, STAD and PRAD, increasing levels of NCS go along with increasing levels of these proliferation markers (Figure 1C). Proliferation markers are also associated with SCS in some cancers, including BRCA and LUSC. However, this is not the case for many other cancers, reflecting again the complex relationship between CIN and proliferation [43–45]. The balance between the proliferation-promoting effect of CIN as a template for Darwinian selection and the cellular burden of chromosomal aberrations accompanied by CIN might be highly cancer type dependent.

Both NCS and SCS tend to be higher in primary tumours than in normal samples (Figure A2B). Previous findings linked CIN and metastasis [23]. We find that metastatic tumours tend to have higher levels of the SCS. For the NCS, this relationship is unclear. The average NCS is higher in metastatic tumours, but there are many primary tumours with high levels of NCS. The small sample size for metastatic tumours prevents a cancer type-specific analysis of the relationship between CIN and metastatic disease.

These results highlight that W-CIN and S-CIN are two related but distinct phenotypes with different distributions across cancer types. Whole genome doubling is often accompanied by W-CIN, but this does not completely explain the elevated levels of NCS in some cancer types or individual tumours. The bimodal distribution of the NCS in most cancer types separates high-W-CIN from low-W-CIN samples, but does not provide a universal threshold valid across cancer types. However, in some cancers such as OV, even the non-WGD samples can exhibit substantial levels of W-CIN. In contrast, S-CIN is a continuous trait which is strongly associated with HRD, but not with WGD. Please note that these patterns are also observed in cell lines (Figure A2C,D).

3.3. Clinical Significance of CIN in Different Cancer Types

To analyse the relationship between W-CIN and prognosis, we divided the tumour samples in each individual cancer type into disjoint NCS^{high} and NCS^{low} groups using the median as a threshold. For seven of the 33 cancer types, we found that NCS^{high} patients had a significantly shorter overall survival than patients in the NCS^{low} group (Figure 2A, Table A1, log rank test, $p < 0.05$). This includes BRCA, LGG, LIHC, OV, STAD, UCEC and UVM. Disease-free survival is lower in the NCS^{high} group for LGG, OV, PRAD and UCEC patients (Figure A3A, Table A3, log rank test, $p < 0.05$) and progression-free survival is negatively associated with high NCS in KIRC, LGG, OV, PRAD, UCEC and UVM (Figure A4A, Table A5, log rank test, $p < 0.05$).

Using an analogous separation of the tumour samples into SCS^{low} and SCS^{high} groups using the median SCS in each tumour type, we found that the overall survival of patients in 11 out of 33 cancers is negatively associated with S-CIN (Figure 2B, Table A2, log rank test, $p < 0.05$). High SCS is linked to impaired disease-free survival in adrenocortical carcinoma (ACC), KIRC, kidney renal papillary cell carcinoma (KIRP), lung squamous cell carcinoma (LUSC), PRAD, THCA and UCEC (Figure A3B, Table A4, log rank test, $p < 0.05$). For OV, patients with high SCS tend to have slightly better overall survival (Figure 2B, Table A2). However, the effect is very small and at the edge of statistical significance. In addition, the analysis of disease-free survival (Figure A3, Tables A3 and A4) and progression-free survival (Figure A4, Tables A5 and A6) does not provide any evidence for an effect of S-CIN on the prognosis of OV patients.

To further explore the clinical relevance of both types of CIN in therapy, we studied the association between CIN and response to radiotherapy or chemotherapy. Radiotherapy responders tend to have lower NCSs than radiotherapy non-responders (Wilcoxon rank test, $p = 0.0007$), whereas SCS is not significantly associated with radiotherapy response (Figure 2C). On a pan-cancer level, we did not find a significant difference between NCSs in the group of chemotherapy responders versus non-responders (Figure 2D). The median SCS of chemotherapy responders is slightly higher. One possible explanation is that high S-CIN samples tend to have defective homologous recombination repair (see Figure 1B), which renders them slightly more sensitive to chemotherapy [46,47].

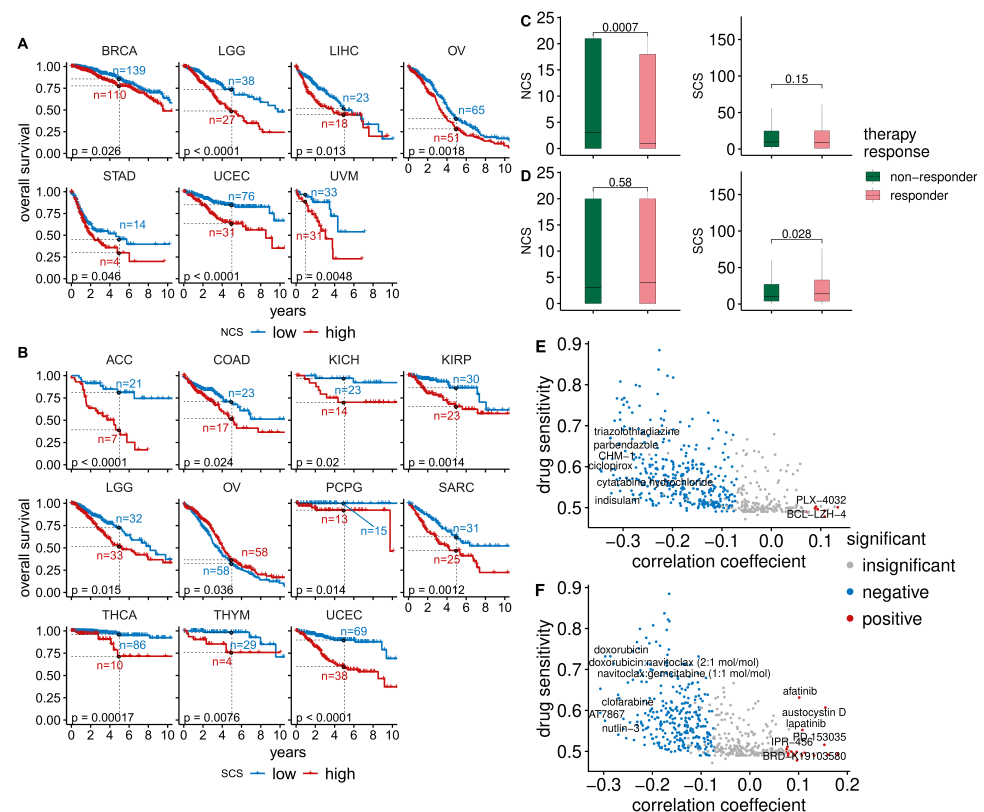


Figure 2. Clinical significance of CIN in different cancer types. (A) For seven cancer types there are significant differences in overall survival between patient samples with low NCS (blue) and high NCS (red). Dashed lines indicate the five-year overall survival probability of the two groups. (B) The SCS is associated with overall survival in 11 cancer types (low-SCS group in blue and high-SCS group in red). (C) Comparison of the NCS and SCS between radiotherapy responders and non-responders using a Wilcoxon rank sum test. (D) Comparison of the NCS and SCS between chemotherapy responders and non-responders using a Wilcoxon rank sum test. (E) The median drug sensitivity of a compound plotted against the correlation coefficient between drug sensitivity and NCS. Drugs with significant positive and negative correlations between their sensitivity and NCS are highlighted in red and blue, respectively. (F) The median drug sensitivity of a compound plotted against the correlation coefficient between drug sensitivity and SCS. Compounds whose sensitivity is significantly negatively or positively correlated with SCS are highlighted in blue and red, respectively.

Next, we asked whether there are drugs suitable for targeting CIN [48]. To this end, we combined data from the Cancer Therapeutics Response Portal (CTRP) and the Cancer Cell Line Encyclopedia (CCLE). We normalised the area under the dose–response curve (AUC) values of 545 compounds and small molecules in all cell lines to values between zero and one and defined drug sensitivity as one minus the normalised AUC. Values of zero indicate the highest resistance level, whereas values of one indicate the highest possible sensitivity. We then computed Spearman rank correlation coefficients between the drug sensitivity of each compound with the NCS or SCS. To analyse the typical drug sensitivity as a function of CIN, we plotted the median drug sensitivity of each compound or small molecule across cell lines against their correlation coefficients with NCS (Figure 2E) or SCS (Figure 2F).

For the majority of compounds, we found negative correlations between their sensitivity and both types of CIN (Figure 2E,F), highlighting that for many compounds CIN confers an intrinsic drug resistance [26]. Only a few compounds are more potent in high-CIN cell

lines than in low-CIN cell lines. However, their overall levels of sensitivity are typically low in comparison to drugs more efficient in low-CIN cell lines.

The strongest positive correlations between drug sensitivity and NCS (Figure 2E) were found for the compounds PLX-4032 and BCL-LZH-4 (median drug sensitivity >0.5 and FDR-adjusted $p < 5\%$). PLX-4032 targets *BRAF* and has been approved by the FDA for clinical use. The *BCL2/BCL-xL/MCL1* inhibitor BCL-LZH-4 is a probe.

Drugs showing increasing sensitivity with the SCS (Figure 2F) include afatinib and lapatinib (median drug sensitivity >0.5 and FDR-adjusted $p < 5\%$). Lapatinib targets *HER2/neu* and is used in combination treatment of *HER2*-positive breast cancer. Afatinib is used to treat non-small lung cancers with *EGFR* mutations [49]. Austocystin D is a natural cytotoxic agent and also more efficient in high-S-CIN tumours. Further details about the correlations between CIN scores and drug sensitivity can be found in the Supplementary Tables (NCS: Table S1; SCS: Table S2; WGII: Table S3).

Overall, the analysis shows that the prognostic value of CIN scores depends on cancer types and that S-CIN and W-CIN provide distinct prognostic information. The prognosis for many cancer types worsens with increased levels of CIN scores. Only for OV did we find a slightly better overall survival for patients with high SCS. It is possible that a stratification of patients according to cancer subtypes might reveal more fine-grained insights regarding the prognostic value of CIN [15,16]. Our drug sensitivity analysis reveals that most compounds are less efficient in high-CIN tumours than in low-CIN tumours. There are a few drugs to which high-CIN cells are more sensitive than low-CIN cells. In particular, we suggest that afatinib, lapatinib and austocystin D merit further investigation for targeting S-CIN tumours. However, current drug sensitivity screens do not include many highly potent drugs specifically targeting CIN.

3.4. PARADIGM Pathway Activity and CIN

To identify pathways with altered activity in W-CIN or S-CIN tumours, we used the PARADIGM framework [30]. PARADIGM is a computational model which represents interactions between biological entities as a factor graph. PARADIGM integrates copy number and gene expression data and computes activities for each PARADIGM pathway feature in an individual tumour sample. These features refer to protein-coding genes, protein complexes, abstract processes and gene families. We focused on the PARADIGM features for protein-coding genes, because these are easier to interpret and can be used to generate experimentally testable predictions. We correlated the PARADIGM pathway features with the NCS or SCS and filtered features with a significant (FDR-adjusted $p < 5\%$) Spearman correlation coefficient ≥ 0.3 in at least seven of the 32 cancer types (NCS: Figure 3A, SCS: Figure 3C).

PARADIGM pathway features corresponding to the mitotic genes *TPX2*, *RAE1*, *UBE2C*, *AURKA* (see Figure 3A) show increased activity in tumours with high NCS, consistent with the known role of chromosome segregation errors in W-CIN [1,20]. Additionally, the PARADIGM features corresponding to the genes *CDC25B* and *DSN1* have higher activity in tumours with high NCS across many cancers. *CDC25B* regulates cell cycle progression and unregulated *CDC25B* induces replication stress, leading to CIN [50]. *DSN1* is required for kinetochore assembly.

The *STX1* (SYNTAXIN 1A) pathway shows increased activity in W-CIN tumours. This finding is surprising, because the *STX1* gene is normally expressed in brain cells and is a key molecule in synaptic exocytosis and ion channel regulation. The reason why *STX1* is upregulated in W-CIN tumours needs further investigation.

It is interesting to note the positive association of the PARADIGM feature for *GINS1* with NCS [10]. The *GINS1* protein is essential for the formation of the *Cdc45-MCM-GINS* (CMG) complex which functions to unwind DNA ahead of the replication fork [51]. As detailed in [10], overexpression of *GINS* in vitro increases replication origin firing and triggers whole chromosome missegregation and W-CIN. Indeed, when we complement our PARADIGM pathway analysis with simple gene-wise correlation of the NCS and gene

expression, we find many genes involved in DNA replication and replication origin firing (see Figure 3A,B).

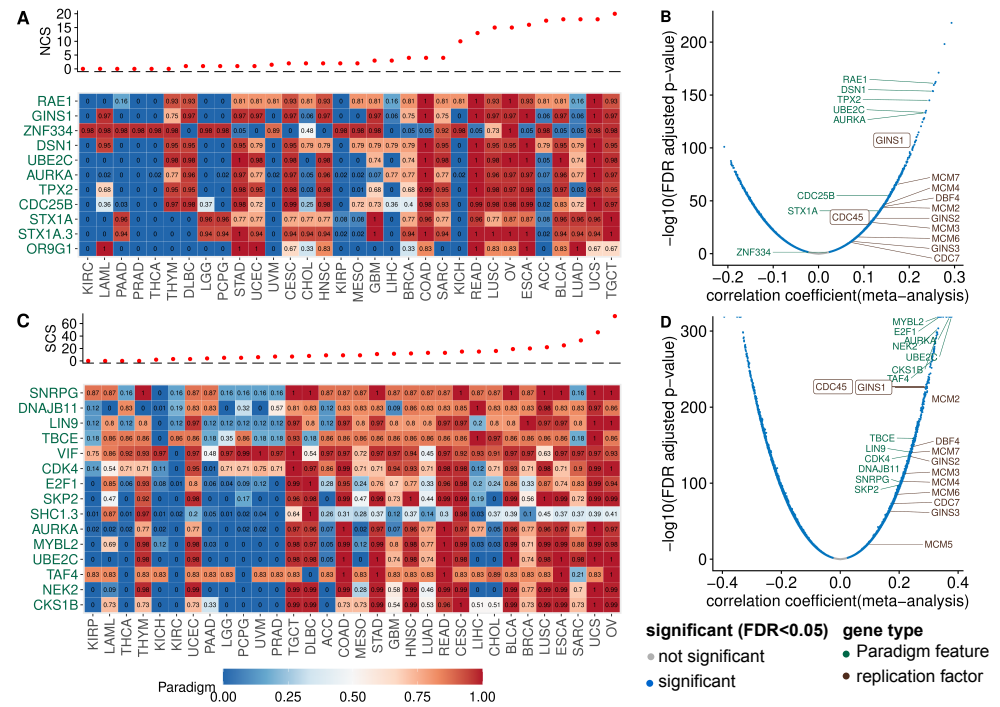


Figure 3. PARADIGM pathway activity and gene expression associated with CIN. (A) The PARADIGM pathway-level activities corresponding to protein-coding genes (rows) were correlated with the NCS. Only pathways with a significant correlation (FDR-adjusted $p < 5\%$) larger than 0.3 in at least seven cancer types were included. The heatmap shows the normalised PARADIGM pathway activity (0–1 from low to high). Cancer types were ordered according to their median NCS, see top panel. (B) Volcano plot for the correlation between gene expression and NCS. (C) Analogous to (A), but for the SCS instead of the NCS. (D) Correlation of SCS and gene expression, analogous to (B).

The analysis of the SCS-associated PARADIGM features (Figure 3C) again revealed proteins involved in kinetochore function, mitotic progression and spindle assembly and chromosome segregation (*AURKA*, *UBE2C*, *NEK2*, *TBCE*) or cell cycle progression (*CDK4*, *E2F1*). The activity of the cyclin-dependent kinase regulatory subunit 1B (*CKS1B*) pathway is positively associated with the SCS. *CKS1B* has recently been linked to cancer drug resistance and was discussed as a new therapeutic target [52]. Our results suggest that the *CKS1B* activity is closely linked to S-CIN, which needs to be considered when studying *CKS1B* as a new target gene or as a marker of drug resistance.

To check for the robustness of these findings, we also performed a gene-wise correlation of the SCS and gene expression (Figures 3C and A5C). We also highlighted genes involved in DNA replication. Gene set enrichment analysis indicates that the top high-SCS-CIN-associated genes are enriched with replication origin factors (Figure A5A,B).

Please note that the analysis of genes and PARADIGM pathways negatively associated with CIN did not reveal a similarly consistent pattern across cancer types (see Figure A6).

Taken together, our analysis of PARADIGM pathway activity and gene expression in the context of CIN not only recovered known CIN genes involved in mitotic processes and spindle assembly, but highlighted, amongst others, the replication factor *GINS1* to be associated with W-CIN [10] and the *CDK* regulator and drug resistance protein *CKS1B* as strongly associated with S-CIN. In addition, we observed that the over-expression of genes involved in DNA replication is positively associated with high CIN.

3.5. Somatic Point Mutation Frequencies in High-CIN Tumours

To investigate the relationship between somatic point mutations and CIN, we identified genes that are more frequently or less frequently mutated in high-CIN tumours. From the 19,171 gene mutations, we included only those occurring in more than 19 samples in the wild type or mutant group across different cancer types. We fitted a linear regression model using NCS or SCS as response and somatic point mutation status (present or absent) and cancer type as predictors. The estimated regression coefficient for mutation status was used to measure its association with CIN, adjusted for tumour type.

As expected, at the pan-cancer level, *TP53* mutation shows the strongest association with CIN. Tumours harbouring a *TP53* mutation have on average more than four more whole chromosome gains or losses (ANOVA p -value $< 2.2 \times 10^{-16}$) than tumours with wild type *TP53* (Figure 4A). The mean difference in the SCS in a tumour sample with a *TP53* mutation compared to wild type samples is approximately 11 structural aberrations (Figure 4B). In line with this, *TP53* mutation is positively associated with high CIN in many individual cancer types (Figure A7A). In fact, even after removing MIN samples, this correlation still holds (Figure A7B), corroborating the well-known role of *TP53* as a gatekeeper of genome stability (see e.g., [53]).

Contrary to the enrichment of *TP53* mutation in both types of CIN, we find that the presence of mutations in 5807 different genes is negatively associated with both NCS and SCS (Figure 4A). A similar negative correlation between the frequencies of recurrent copy number alterations and somatic mutations has previously been reported [54]. Later, it was realised that this negative relationship can be reversed, when the confounding effect of MIN [21,27] is removed. When we exclude these hypermutated samples, we observe a more even distribution between genes more or less frequently mutated in high-CIN compared to low-CIN tumours (Figure 4B). This is also consistent with Figure 1C,D, where we found that neither the silent mutation rate nor the non-silent mutation rate is associated with NCS and SCS.

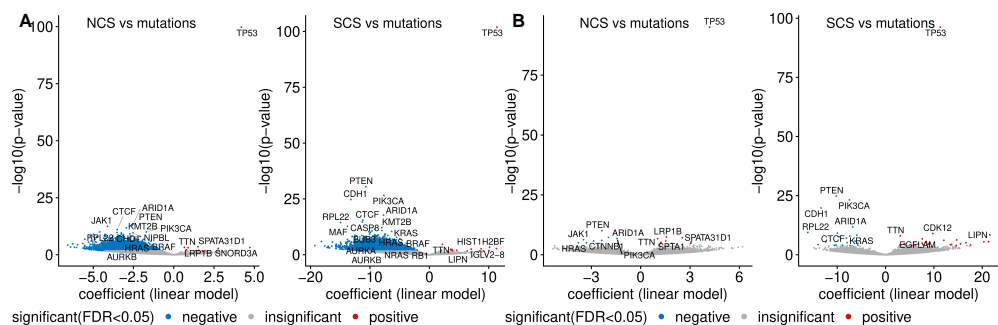


Figure 4. Pan-cancer somatic mutations and CIN. (A) The volcano plots show the association between somatic mutations and the NCS (left) or the SCS (right). The linear model coefficient indicates the mean difference of the respective CIN score when the mutation is present in a tumour sample relative to the wild type. Genes with lowest p -values, well-known CIN genes and cancer driver genes are highlighted. The analysis was performed on genes for which samples sizes for both wild type group and mutated group are larger than 19. Mutations significantly associated (FDR $< 5\%$) with higher or lower CIN score are highlighted in blue and red, respectively. (B) The same as (A), but hypermutated MIN samples are excluded.

Intriguingly, even after excluding hypermutated samples, we find somatic point mutations of important cancer genes including *PI3KCA*, *PTEN* and *ARID1A* to be under-represented in high-CIN bulk tumours (Figure 4B) and high-CIN cancer cell lines (Figure A7D). *HRAS* and *JAK1* mutations are less frequent in tumours with high NCS and *KRAS* mutations are under-represented in samples with high SCS. More remarkably, when only considering validated cancer driver somatic mutations [55], the above observed relationship between *PI3KCA* mutation, *PTEN* mutation and CIN still holds (Figure A7C). The under-representation of

somatic mutations in these key cancer genes in high-CIN tumours cannot be explained by differences in the overall mutation rates of these samples.

3.6. Copy Number Gains and Losses Associated with CIN

Given that somatic mutations of many genes are under-represented in high-CIN tumours, we next investigated copy number alterations which are specifically linked to CIN (Figure 5A). One of the strongest associations between a copy number gain and SCS was found for the *MYC* proto-oncogene. The candidate oncogene *PVT1* is also specifically gained in tumours with high SCS. *PVT1* is involved in the regulation of *MYC* [56] and carries a *TP53*-binding site. In addition, we found high NCS is associated with copy number gains for genes encoding members of the *WFDC-EPPIN* family, which have been linked to proliferation, metastasis, apoptosis and invasion in ovarian cancer (reviewed in [57]).

Genes specifically lost in tumour samples with high NCS include *KIAA1644*, *TAMM41*, *GRM7*, *TTC39B* and *FREM1* (Figure 5B). The top genes whose copy number loss is strongly associated with SCS are *PDE4D*, *PTEN*, *RB1* and *KLIN1* (Figure 5C). The tumour suppressor *RB1* is a key regulator of the G1/S transition of the cell cycle and is required for the stabilisation of heterochromatin.

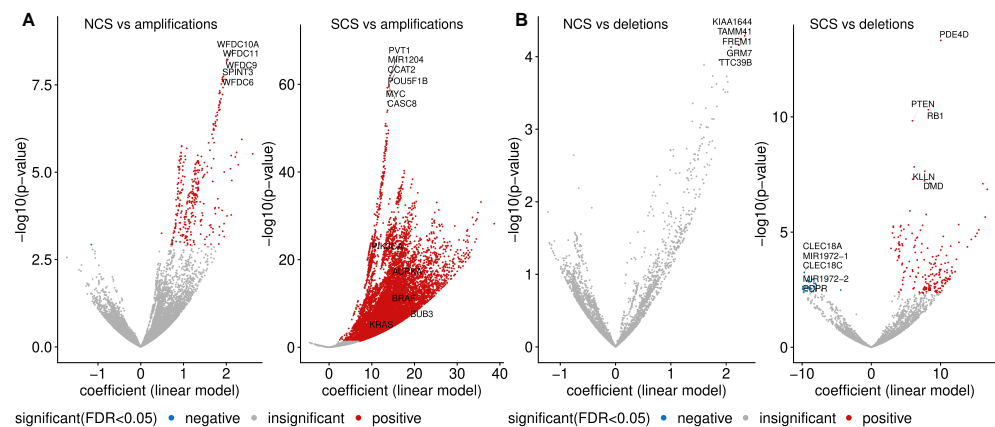


Figure 5. Copy number amplifications and deletions enriched in high-CIN samples. (A) The volcano plots show the gene-wise associations between copy number amplification status and NCS (left) and SCS (right), obtained from a regression model adjusted by cancer type. The linear model coefficient indicates the mean difference in the respective CIN score when the alteration is present in a tumour sample relative to the wild type. Genes with the lowest p-values and well-known CIN genes are highlighted. Blue and red colours encode genes with a significantly higher alteration frequency (FDR < 5%) in samples with low and high CIN scores, respectively. The analysis was performed on 16,922 genes with sample sizes greater than 19 for both wild type and amplified groups. (B) Pan-cancer copy number deletions associated with SCS are displayed in an analogous way to (A).

3.7. *PI3KCA* Copy Number Gains in High-S-CIN Tumours Suggest a Gene Dosage-Dependent Mechanism for *PI3K* Pathway Activation

In Section 3.5, we observed that somatic point mutations of *PTEN* and *PIK3CA* were scarce in high-CIN tumours. In addition, copy number amplification of *PIK3CA* and copy number loss of *PTEN* are very frequent in tumour samples with high SCS. This led us to ask whether there is a link between S-CIN and specific gene copy number alterations in these two genes to activate the *PI3K* oncogenic pathway. The *PIK3CA* gene encodes the catalytic subunit of phosphatidylinositol 3-kinase and the *PI3K* oncogenic pathway is frequently deregulated in many cancers. *PTEN* is a tumour suppressor gene and negatively regulates the growth-promoting *PI3K/AKT/mTOR* signal transduction pathway.

The oncoprint in Figure 6A displays tumour samples from all 33 TCGA cancer types in our investigation, which harbour at least one of the following genetic alterations: Somatic mutation of *PIK3CA* or *PTEN*, copy number amplification of *PIK3CA*, deletion of *PTEN*. It

is apparent that there is only a small number of cancers with an amplification of *PIK3CA* or a deletion of *PTEN*, which simultaneously harbour somatic mutations in any of these genes. The copy number of both genes is also strongly associated with their gene expression. In particular, amplification and simultaneous over-expression of *PIK3CA* are associated with higher levels of SCS.

To check whether this effect is preserved in pure cancer cells, we used cell line data from CCLE and found a very similar pattern. Copy number gains of *PIK3CA* are linked to high levels of its gene expression, and rarely co-occur with somatic mutations, but are associated with high SCS.

Taken together, we suggest a gene dosage effect on *PI3K* pathway activity, which is facilitated in high-S-CIN tumours. This effect is cancer cell intrinsic, because it can also be observed in cancer cell lines.

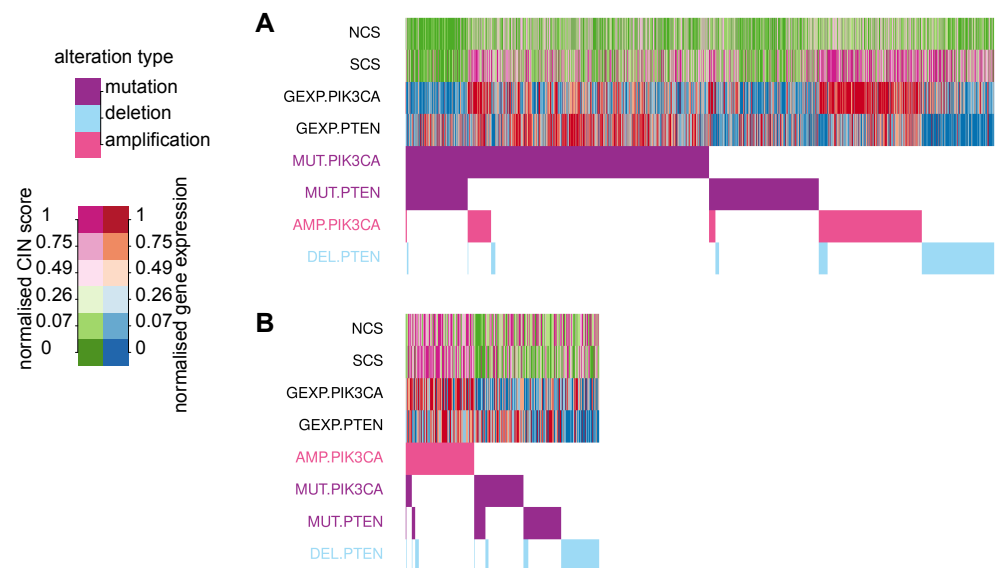


Figure 6. Oncoprint for *PIK3CA* and *PTEN* in relation to CIN. (A) The bottom panel depicts the presence or absence of somatic mutations, copy number amplifications of *PIK3CA* and deletions of *PTEN* in TCGA tumour samples (columns). Alterations are sorted by their frequency. The upper panel shows the NCS, SCS, *PI3KCA* and *PTEN* gene expression. Different levels of CIN scores and gene expression are encoded by colours. (B) The corresponding oncoprint for cell line data from CCLE.

4. Discussion

W-CIN and S-CIN are two distinct but related phenotypes triggered by different biological mechanisms and leading to diverse consequences. A large majority of pan-cancer association studies has focused on CIN in general or exclusively on W-CIN. Here, we present an integrative statistical analysis for 33 cancer types distinguishing between W-CIN and S-CIN. We used the NCS as a proxy measure for W-CIN and the SCS to quantify the degree of S-CIN and associated these karyotypic complexity scores with various molecular and clinical features.

Our analysis reveals that the majority of tumours with high levels of NCS underwent whole genome doubling. Whole genome doubling is an early event in tumourigenesis and has been discussed as a way to rapidly accumulate numerical and structural chromosomal abnormalities and to buffer against negative effects of mutations and aneuploidy [12,58,59]. The results of our analysis suggest that whole genome doubling is typically accompanied by W-CIN, but not S-CIN. Instead, we find that high SCS is linked to homologous recombination deficiency, highlighting the different processes involved in these two different CIN phenotypes [6].

Although whole genome doubling is observed in many tumour samples with high levels of W-CIN, it is not sufficient to explain the elevated NCS in many tumour samples which did not undergo whole genome doubling, as most prominently observed in KICH, ACC and OV. We speculate that replication stress is an alternative mechanism for these elevated levels of W-CIN. This is based on ample evidence that replication stress can induce CIN [7,60] and our observation that replication factors are over-expressed in tumours with high levels of W-CIN and that over-expression of the replication genes *GINS1* and *CDC45* can induce W-CIN [10].

We find that NCS and SCS are associated with poor prognosis in different cancer types. Only in the case of ovarian cancer did we find that high-S-CIN patients have a slightly longer overall survival, but the difference is very small and at the edge of statistical significance. In addition, we observe slightly higher NCS in patients resistant to radiotherapy. However, the relationship between CIN and prognosis is multifaceted and depends on details of the cellular physiology [3]. For instance, extreme levels of CIN in breast cancer subtypes [15,16] were associated with better prognosis. This indicates that a subtype-specific analysis of W-CIN and S-CIN and prognosis might potentially be an interesting future project. This might also apply for the response to radiotherapy, as improved sensitivity against radiotherapy in transplanted human glioblastoma tumours has been reported [61].

From the association of NCS and SCS with in vitro drug sensitivity, it is apparent that both types of CIN are linked to intrinsic drug resistance, corroborating earlier results in colon cancer [3,26]. However, as a new contribution we filtered small molecules and compounds for which drug sensitivity is positively associated with S-CIN or W-CIN. The drug sensitivity of a *BRAF* inhibitor, PLX-4032, is higher in cells with higher NCS. For S-CIN, this includes the approved drugs afatinib and lapatinib and the natural cytotoxic agent austocystin D. It remains to be tested whether these drugs or compounds are indeed efficient against high-CIN tumours in vivo.

In addition to well-known CIN genes including *TPX2*, *UBE2C* and *AURKA*, we identified a number of new candidate CIN genes and corresponding PARADIGM pathway features [30]. One interesting new finding is the chemotherapeutic drug resistance-inducing gene *CKS1B* [52], which is strongly associated with S-CIN. *CKS1B* is a cell cycle progression gene, which is discussed as a new drug target. Here, we show that *CKS1B* is over-expressed in S-CIN tumours, which might be important for the stratification of patients. We also note that the activity of the replication origin firing factor *GINS1* is linked to W-CIN, which was mechanistically verified in a recent collaboration [10]. In this context, we also found many genes involved in DNA replication to be over-expressed in tumours with high levels of W-CIN and S-CIN.

Both W-CIN and S-CIN are strongly correlated with somatic point mutation of *TP53*. We find that many copy number gains of important oncogenes and loss of tumour suppressor genes [62] are strongly associated with W-CIN and S-CIN. Most strikingly, copy number gains of the oncogene *PIK3CA* and deletion of the tumour suppressor gene *PTEN* rarely occur in combination with somatic mutations in these genes. In addition, copy number gain of *PIK3CA* is linked to increased gene expression and strongly associated with S-CIN. Intriguingly, it has recently been reported that mutations in *PIK3CA* increased in vitro cellular tolerance to spontaneous genome doubling [63]. Our results, however, suggest a gene dosage effect for the activation of the *PI3K* pathway in the context of high S-CIN. This copy number-dependent activation of *PI3K* signalling was observed in both bulk tumours and cancer cell lines, indicating that it is an intrinsic property of S-CIN cells. We suggest that copy number gains of *PIK3CA* should be further investigated for both their mechanistic role in S-CIN and for their clinical implications regarding treatment strategies and patient stratification.

As a final remark, we emphasise again that our analysis is based on the karyotypic complexity scores NCS and SCS, which are averaged measures over a population of cancer cells and reflect features of the evolved W-CIN or S-CIN phenotype. As such, our analysis can

stimulate new experimental work, but it cannot cover the spatio-temporal dynamics [62,64] of tumour heterogeneity. In particular, individual chromosome changes in single cells, which still might be important drivers of cancer progression, cannot be detected by bulk data analysis [65]. We believe that the accumulation of single cell-based data from different cancer types will be essential to better understand the effect of ongoing CIN on cancer progression in the future. This will also include the testing of concepts such as karyotype coding [66], the relationship between different karyotypic states within a cellular population and the evolutionary forces shaping cancer evolution at the level of chromosome organisation.

5. Conclusions

In summary, our pan-cancer analysis provides insights into the distinct and common molecular, prognostic and therapeutic characteristics of W-CIN and S-CIN. Our results suggest that whole genome doubling and homologous recombination deficiency might be the most important drivers for W-CIN and S-CIN, respectively. The predictive value of W-CIN and S-CIN depends on the cancer type. We report that most of the existing compounds preferably kill low-CIN cells, but we also suggest a few compounds with increased efficiency in high-CIN cells. High activity of *CKS1B* might be a promising S-CIN target, because its expression is linked high S-CIN. We propose a new copy number-dependent mechanism for an increased activity of the oncogenic *PI3K* pathway in high-S-CIN cancer cells, which merits experimental investigation.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/cancers14061424/s1>, Table S1: The relationship between drug sensitivity and NCS, Table S2: The relationship between drug sensitivity and SCS, Table S3: The relationship between drug sensitivity and WGII, PDF S1: Pan-cancer WGII association analysis results.

Author Contributions: X.Z.: Methodology, formal analysis, investigation, software, data curation, writing, review and editing, visualisation; M.K.: Conceptualisation, methodology, formal analysis, investigation, writing, review and editing, supervision, project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the FOR2800 funded by the Deutsche Forschungsgemeinschaft (sub-project 3).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code and data used to reproduce this work are available at https://github.com/mcmzxx/pancan_cin.

Acknowledgments: This study used data generated by TCGA Research Network: <https://www.cancer.gov/tcga> (accessed on 13 May 2020) and Broad-Novartis CCLE: <https://sites.broadinstitute.org/ccle> (accessed on 13 May 2020).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CIN	Chromosomal instability
W-CIN	Whole chromosome instability
S-CIN	Structural chromosomal instability
WGD	Whole genome doubling
WGII	Weighted genome instability index
NCS	Numerical complexity score
SCS	Structural complexity score
AS	Aneuploidy score
HRD	Homologous recombination deficiency
SMR	Silent mutation rate

NSMR Non-silent mutation rate
 ITH Intra-tumour heterogeneity
 MIN Microsatellite instability
 AUC Area under the dose–response curve
 FDA Food and Drug Administration

Appendix A

Table A1. Association between W-CIN and overall survival across cancer types.

Cohort	Sample_Number	p-Value	Low_surv5 ^a	High_surv5 ^b	Low_surv5_n ^c	High_surv5_n ^d
UCEC	518	0.00	0.85	0.63	76.00	31.00
LGG	509	0.00	0.74	0.49	38.00	27.00
OV	558	0.00	0.40	0.28	65.00	51.00
UVM *	80	0.00	0.97	0.89	33.00	31.00
LIHC	366	0.01	0.52	0.45	23.00	18.00
BRCA	1066	0.03	0.86	0.78	139.00	110.00
STAD	433	0.05	0.45	0.30	14.00	4.00
THYM	122	0.08	0.96	0.81	28.00	5.00
SARC	252	0.08	0.59	0.49	33.00	23.00
HNSC	516	0.10	0.55	0.41	24.00	28.00
LAML *	179	0.13	0.57	0.44	68.00	20.00
GBM	571	0.14	0.07	0.05	12.00	7.00
DLBC	48	0.16	0.73	0.94	6.00	3.00
CESC	294	0.17	0.73	0.59	26.00	15.00
TGCT	133	0.17	0.99	0.95	32.00	19.00
ACC	89	0.30	0.67	0.57	14.00	14.00
KIRP	282	0.31	0.82	0.69	34.00	19.00
ESCA *	182	0.32	0.82	0.70	66.00	49.00
KIRC	506	0.40	0.62	0.63	69.00	78.00
PCPG	161	0.41	0.97	0.96	16.00	12.00
PAAD	183	0.46	0.18	0.40	3.00	5.00
PRAD	489	0.62	0.99	0.96	54.00	30.00
CHOL *	36	0.64	0.75	0.86	15.00	12.00
READ *	154	0.65	0.94	0.96	52.00	63.00
MESO *	86	0.72	0.66	0.70	29.00	27.00
KICH	65	0.77	0.86	0.85	17.00	20.00
BLCA	405	0.78	0.44	0.40	26.00	21.00
LUAD	491	0.81	0.41	0.41	28.00	25.00
THCA	497	0.84	0.93	0.94	80.00	16.00
UCS *	56	0.88	0.76	0.84	22.00	20.00
COAD	425	0.89	0.64	0.57	25.00	15.00
LUSC	481	0.90	0.51	0.44	39.00	41.00
SKCM *	104	0.93	0.84	0.91	33.00	37.00

^a Five-year overall survival probability in low-W-CIN group. ^b Five-year overall survival probability in high-W-CIN group. ^c Number of samples at risk in low-W-CIN group by 5th year. ^d Number of samples at risk in high-W-CIN group by 5th year. * One-year overall survival statistics were reported in these cancer types due to short survival.

Table A2. Association between S-CIN and overall survival across cancer types.

Cohort	Sample_Number	p-Value	Low_surv5 ^a	High_surv5 ^b	Low_surv5_n ^c	High_surv5_n ^d
UCEC	518	0.00	0.90	0.60	69.00	38.00
ACC	89	0.00	0.81	0.39	21.00	7.00
THCA	497	0.00	0.96	0.72	86.00	10.00
SARC	252	0.00	0.62	0.47	31.00	25.00
KIRP	282	0.00	0.86	0.65	30.00	23.00

Table A2. Cont.

Cohort	Sample_Number	p-Value	Low_surv5 ^a	High_surv5 ^b	Low_surv5_n ^c	High_surv5_n ^d
THYM	122	0.01	0.98	0.76	29.00	4.00
PCPG	161	0.01	1.00	0.92	15.00	13.00
LGG	509	0.01	0.73	0.52	32.00	33.00
KICH	65	0.02	0.97	0.70	23.00	14.00
COAD	425	0.02	0.71	0.52	23.00	17.00
OV	558	0.04	0.32	0.37	58.00	58.00
ESCA *	182	0.06	0.81	0.72	60.00	55.00
LUAD	491	0.08	0.48	0.35	28.00	25.00
TGCT	133	0.09	0.95	1.00	25.00	26.00
UCS *	56	0.09	0.75	0.85	20.00	22.00
READ *	154	0.10	0.96	0.93	64.00	51.00
LAML *	179	0.10	0.58	0.45	66.00	22.00
PAAD	183	0.11	0.25	0.26	5.00	3.00
KIRC	506	0.17	0.65	0.59	83.00	64.00
BRCA	1066	0.22	0.83	0.81	129.00	120.00
GBM	571	0.23	0.05	0.08	7.00	12.00
CHOL *	36	0.39	0.76	0.83	13.00	14.00
PRAD	489	0.45	0.99	0.97	40.00	44.00
BLCA	405	0.46	0.40	0.44	21.00	26.00
UVM *	80	0.50	0.93	0.93	38.00	26.00
HNSC	516	0.56	0.51	0.44	27.00	25.00
LUSC	481	0.57	0.46	0.49	36.00	44.00
SKCM *	104	0.63	0.84	0.93	36.00	34.00
MESO *	86	0.64	0.62	0.75	28.00	28.00
LIHC	366	0.73	0.48	0.50	23.00	18.00
CESC	294	0.77	0.68	0.66	20.00	21.00
STAD	433	0.90	0.35	0.43	11.00	7.00
DLBC *	48	0.93	0.96	0.89	23.00	15.00

^a Five-year overall survival probability in low-S-CIN group. ^b Five-year overall survival probability in high-S-CIN group. ^c Number of samples at risk in low-S-CIN group by 5th year. ^d Number of samples at risk in high-S-CIN group by 5th year. * One-year overall survival statistics were reported in these cancer types due to short survival.

Table A3. Association between W-CIN and disease-free survival across cancer types.

Cohort	Sample_Number	p-Value	Low_surv5 ^a	High_surv5 ^b	Low_surv5_n ^c	High_surv5_n ^d
OV	279	0.00	0.24	0.11	17.00	8.00
UCEC	406	0.00	0.87	0.72	60.00	23.00
LGG *	130	0.03	0.97	0.98	65.00	37.00
PRAD	332	0.03	0.85	0.74	36.00	15.00
THCA	352	0.06	0.92	0.84	56.00	13.00
COAD	175	0.07	0.82	0.63	11.00	3.00
LUSC	295	0.07	0.74	0.62	25.00	23.00
CHOL *	24	0.10	0.73	0.43	11.00	3.00
KICH	29	0.16	0.91	1.00	5.00	12.00
CESC	170	0.17	0.83	0.76	18.00	8.00
SARC	148	0.18	0.58	0.41	17.00	10.00
KIRP	180	0.19	0.73	0.91	17.00	14.00
DLBC	28	0.23	1.00	0.90	5.00	3.00
UCS *	26	0.28	1.00	0.77	11.00	9.00
LIHC	315	0.29	0.36	0.28	12.00	6.00
PAAD *	68	0.36	0.85	0.81	28.00	14.00
BLCA	187	0.37	0.69	0.73	13.00	13.00
PCPG	144	0.40	0.95	0.97	12.00	10.00
TGCT	104	0.47	0.70	0.82	10.00	14.00

Table A3. Cont.

Cohort	Sample_Number	p-Value	Low_surv5 ^a	High_surv5 ^b	Low_surv5_n ^c	High_surv5_n ^d
MESO *	15	0.48	0.67	1.00	6.00	2.00
GBM *	3	0.48	1.00	1.00	1.00	2.00
READ *	42	0.53	0.90	1.00	15.00	20.00
KIRC	107	0.60	0.90	0.76	19.00	19.00
ESCA *	87	0.80	0.75	0.82	28.00	22.00
ACC	52	0.83	0.68	0.72	11.00	10.00
LUAD	291	0.85	0.62	0.56	19.00	18.00
HNSC	130	0.85	0.69	0.55	7.00	4.00
BRCA	927	0.87	0.85	0.84	107.00	83.00
STAD	255	0.96	0.62	0.69	9.00	4.00

^a Five-year disease-free survival probability in low-W-CIN group. ^b Five-year disease-free survival probability in high-W-CIN group. ^c Number of samples at risk in low-W-CIN group by 5th year. ^d Number of samples at risk in high-W-CIN group by 5th year. * One-year disease-free survival statistics were reported in these cancer types due to short survival.

Table A4. Association between S-CIN and disease-free survival across cancer types.

Cohort	Sample_Number	p-Value	Low_surv5 ^a	High_surv5 ^b	Low_surv5_n ^c	High_surv5_n ^d
UCEC	406	0.00	0.92	0.65	61.00	22.00
ACC	52	0.00	0.81	0.48	16.00	5.00
PRAD	332	0.00	0.88	0.75	27.00	24.00
KIRP	180	0.02	0.88	0.72	18.00	13.00
LUSC	295	0.03	0.71	0.64	25.00	23.00
THCA	352	0.03	0.92	0.81	64.00	5.00
KIRC	107	0.04	0.94	0.70	24.00	14.00
READ *	42	0.09	1.00	0.89	20.00	15.00
BRCA	927	0.10	0.88	0.80	100.00	90.00
PAAD *	68	0.14	0.91	0.74	23.00	19.00
LIHC	315	0.15	0.33	0.31	12.00	6.00
GBM *	3	0.16	1.00	1.00	1.00	2.00
KICH	29	0.18	0.94	1.00	11.00	6.00
COAD	175	0.27	0.79	0.71	9.00	5.00
CESC	170	0.30	0.86	0.75	15.00	11.00
HNSC	130	0.31	0.66	0.59	8.00	3.00
LUAD	291	0.37	0.63	0.54	19.00	18.00
DLBC *	28	0.45	1.00	1.00	14.00	11.00
MESO *	15	0.46	0.69	1.00	6.00	2.00
ESCA *	87	0.49	0.80	0.76	26.00	24.00
PCPG	144	0.49	0.93	1.00	12.00	10.00
LGG	130	0.61	0.76	0.65	3.00	3.00
OV	279	0.64	0.22	0.14	17.00	8.00
CHOL *	24	0.65	0.67	0.57	10.00	4.00
BLCA	187	0.69	0.73	0.69	14.00	12.00
TGCT	104	0.70	0.75	0.75	14.00	10.00
STAD	255	0.76	0.67	0.64	8.00	5.00
UCS *	26	0.76	0.92	0.83	10.00	10.00
SARC	148	0.79	0.53	0.48	14.00	13.00

^a Five-year disease-free survival probability in low-S-CIN group. ^b Five-year disease-free survival probability in high-S-CIN group. ^c Number of samples at risk in low-S-CIN group by 5th year. ^d Number of samples at risk in high-S-CIN group by 5th year. * One-year disease-free survival statistics were reported in these cancer types due to short survival.

Table A5. Association between W-CIN and progression-free survival across cancer types.

Cohort	Sample_Number	p-Value	Low_surv5 ^a	High_surv5 ^b	Low_surv5_n ^c	High_surv5_n ^d
UCEC	518	0.00	0.80	0.56	67.00	27.00
LGG	509	0.00	0.53	0.29	26.00	10.00
PRAD	489	0.00	0.75	0.60	43.00	20.00
UVM *	79	0.00	0.94	0.69	31.00	24.00
OV	558	0.00	0.17	0.09	24.00	11.00
KIRC	504	0.04	0.69	0.58	60.00	58.00
ESCA *	182	0.07	0.66	0.57	48.00	36.00
THYM	122	0.08	0.83	0.63	24.00	5.00
CESC	294	0.11	0.69	0.63	23.00	13.00
ACC	89	0.13	0.49	0.40	11.00	10.00
SKCM *	104	0.15	0.74	0.68	27.00	25.00
DLBC	48	0.18	0.67	0.93	6.00	3.00
SARC	252	0.20	0.45	0.32	22.00	13.00
LIHC	366	0.23	0.29	0.25	11.00	7.00
GBM *	571	0.27	0.30	0.30	85.00	58.00
LUSC	482	0.31	0.56	0.53	32.00	33.00
HNSC	516	0.32	0.51	0.47	20.00	24.00
CHOL *	36	0.33	0.53	0.41	10.00	5.00
THCA	497	0.33	0.84	0.83	68.00	14.00
READ *	154	0.34	0.90	0.88	48.00	56.00
PCPG	161	0.40	0.80	0.88	11.00	11.00
MESO *	84	0.41	0.57	0.54	22.00	19.00
BLCA	406	0.42	0.39	0.42	21.00	15.00
COAD	425	0.45	0.62	0.56	21.00	9.00
TGCT	133	0.53	0.71	0.80	20.00	17.00
STAD	435	0.68	0.43	0.47	14.00	4.00
KICH	65	0.75	0.87	0.87	17.00	20.00
BRCA	1066	0.76	0.79	0.78	122.00	99.00
UCS *	56	0.79	0.48	0.60	14.00	14.00
KIRP	281	0.96	0.71	0.80	28.00	17.00
LUAD	491	0.99	0.38	0.40	20.00	18.00
PAAD *	183	1.00	0.64	0.61	60.00	32.00

^a Five-year progression-free survival probability in low-W-CIN group. ^b Five-year progression-free survival probability in high-W-CIN group. ^c Number of samples at risk in low-W-CIN group by 5th year. ^d Number of samples at risk in high-W-CIN group by 5th year. * One-year progression-free survival statistics were reported in these cancer types due to short survival.

Table A6. Association between S-CIN and progression-free survival across cancer type.

Cohort	Sample_Number	p-Value	Low_surv5 ^a	High_surv5 ^b	Low_surv5_n ^c	High_surv5_n ^d
UCEC	518	0.00	0.86	0.50	65.00	29.00
ACC	89	0.00	0.66	0.22	16.00	5.00
KIRP	281	0.00	0.86	0.61	27.00	18.00
THCA	497	0.00	0.87	0.64	75.00	7.00
PRAD	489	0.00	0.76	0.65	32.00	31.00
LGG	509	0.00	0.50	0.33	18.00	18.00
CHOL *	36	0.02	0.69	0.27	11.00	4.00
THYM	122	0.03	0.85	0.60	26.00	3.00
GBM *	571	0.06	0.29	0.32	73.00	70.00
COAD	425	0.07	0.65	0.53	19.00	11.00
SARC	252	0.08	0.44	0.34	19.00	16.00
PAAD *	183	0.09	0.67	0.59	46.00	46.00
KICH	65	0.09	0.97	0.72	23.00	14.00
KIRC	504	0.11	0.66	0.60	67.00	51.00
ESCA *	182	0.12	0.61	0.63	41.00	43.00

Table A6. Cont.

Cohort	Sample_Number	p-Value	Low_surv5 ^a	High_surv5 ^b	Low_surv5_n ^c	High_surv5_n ^d
LIHC	366	0.12	0.26	0.27	11.00	7.00
BRCA	1066	0.12	0.81	0.75	118.00	103.00
READ *	154	0.12	0.93	0.84	59.00	45.00
UVM *	79	0.21	0.82	0.80	32.00	23.00
OV	558	0.25	0.14	0.13	22.00	13.00
BLCA	406	0.31	0.44	0.39	16.00	20.00
SKCM *	104	0.31	0.74	0.67	28.00	24.00
LUAD	491	0.35	0.40	0.38	19.00	19.00
UCS *	56	0.41	0.54	0.54	14.00	14.00
CESC	294	0.41	0.70	0.63	19.00	17.00
STAD	435	0.41	0.45	0.44	11.00	7.00
PCPG	161	0.45	0.84	0.84	13.00	9.00
DLBC *	48	0.51	0.83	0.84	19.00	15.00
MESO *	84	0.73	0.63	0.47	24.00	17.00
TGCT	133	0.77	0.72	0.77	19.00	18.00
LUSC	482	0.79	0.51	0.57	29.00	36.00
HNSC	516	0.93	0.49	0.49	23.00	21.00

^a Five-year progression-free survival probability in low-S-CIN group. ^b Five-year progression-free survival probability in high-S-CIN group. ^c Number of samples at risk in low-S-CIN group by 5th year. ^d Number of samples at risk in high-S-CIN group by 5th year. * One-year progression-free survival statistics were reported in these cancer types due to short survival.

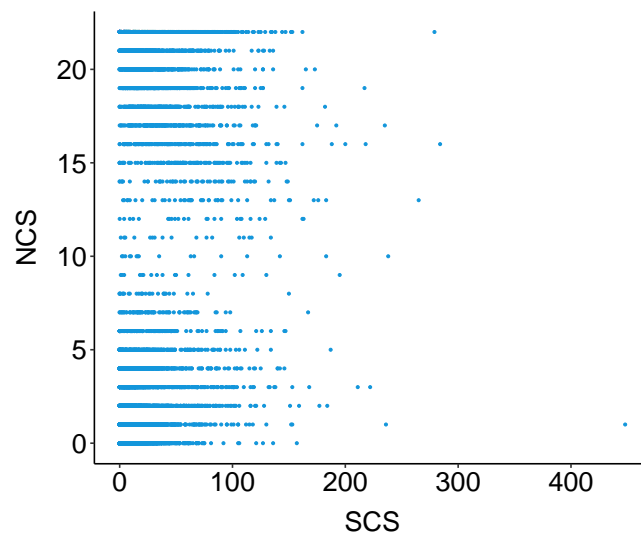


Figure A1. The NCS versus the SCS in TCGA tumours from 33 different cancer types.

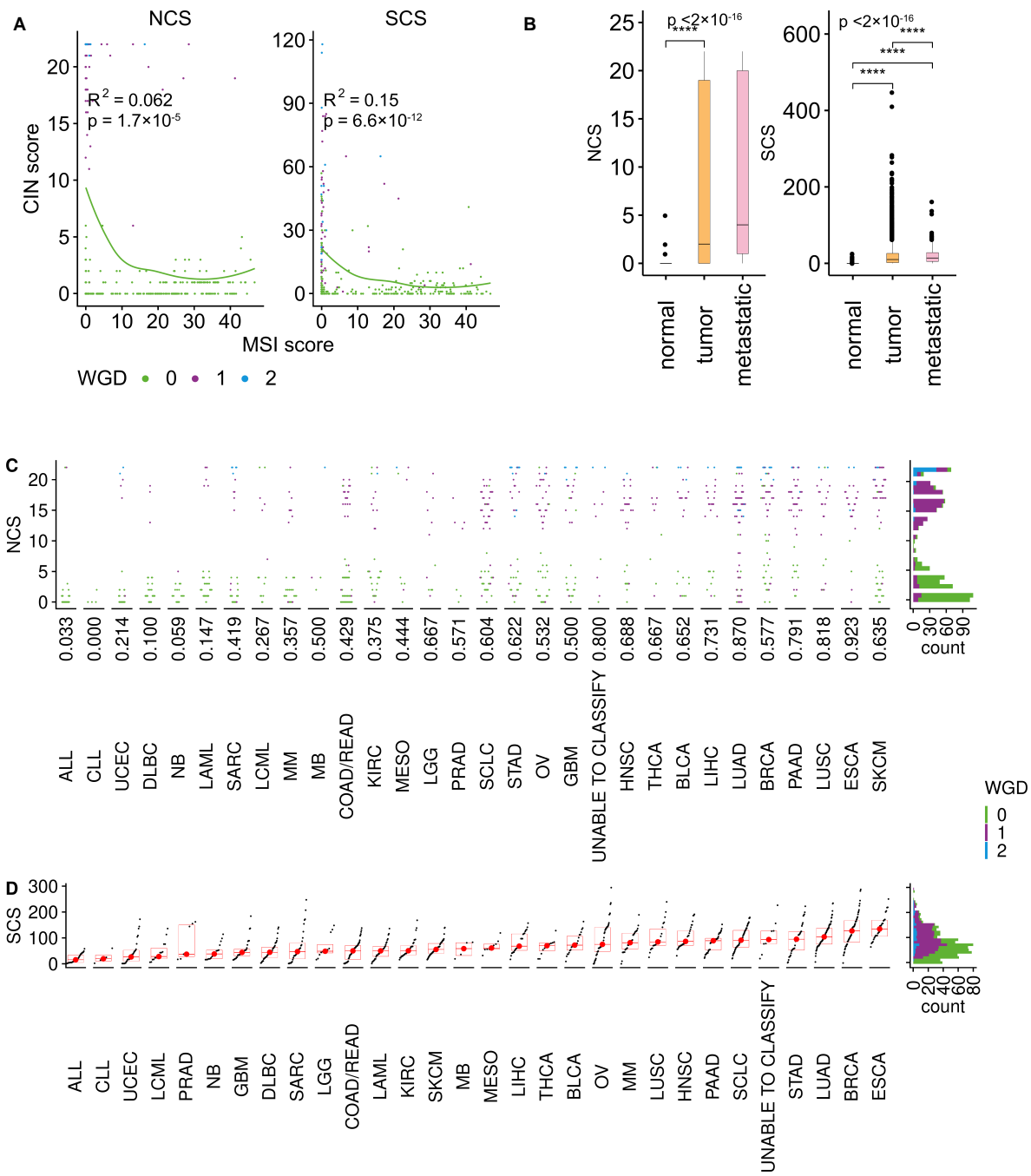


Figure A2. (A) The relationship between microsatellite instability (MIN) scores and the NCS and the SCS. (B) Comparison of the NCS and the SCS between normal samples with primary and metastatic tumour samples. (C) Cancer type-wise NCS distribution in CCLE cell lines, cancer types are ordered by the median NCS; whole genome doubling (WGD) status is encoded by colours. The number reported on x axis is the proportion of samples that underwent WGD. (D) SCS distribution in CCLE cell lines, cancer types are ordered by their median SCS.

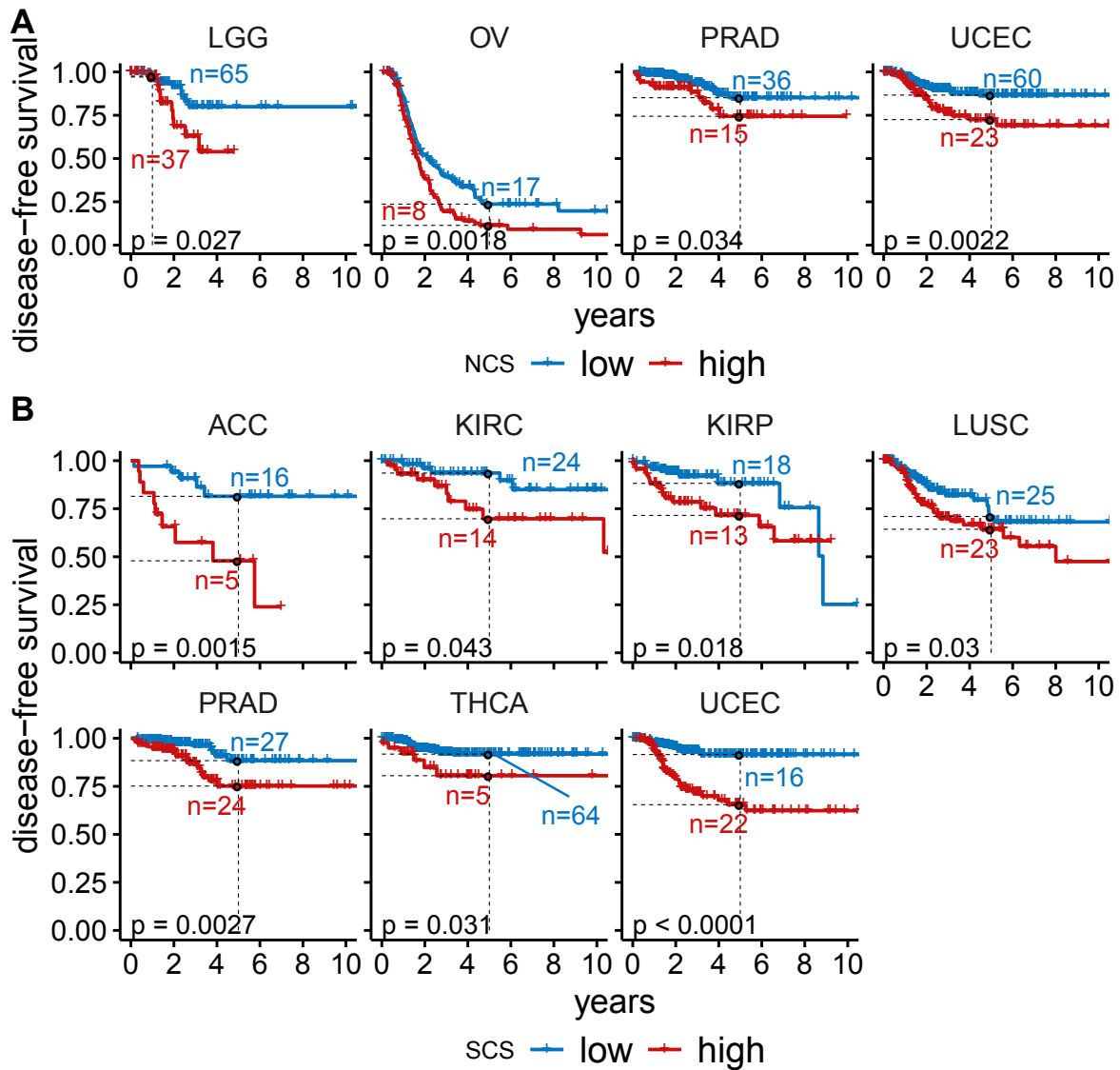


Figure A3. (A) Disease-free survival in four cancer types where significant differences between high- and low-NCS groups were observed. (B) Disease-free survival in seven cancer types where significant differences between high- and low-SCS group were observed.

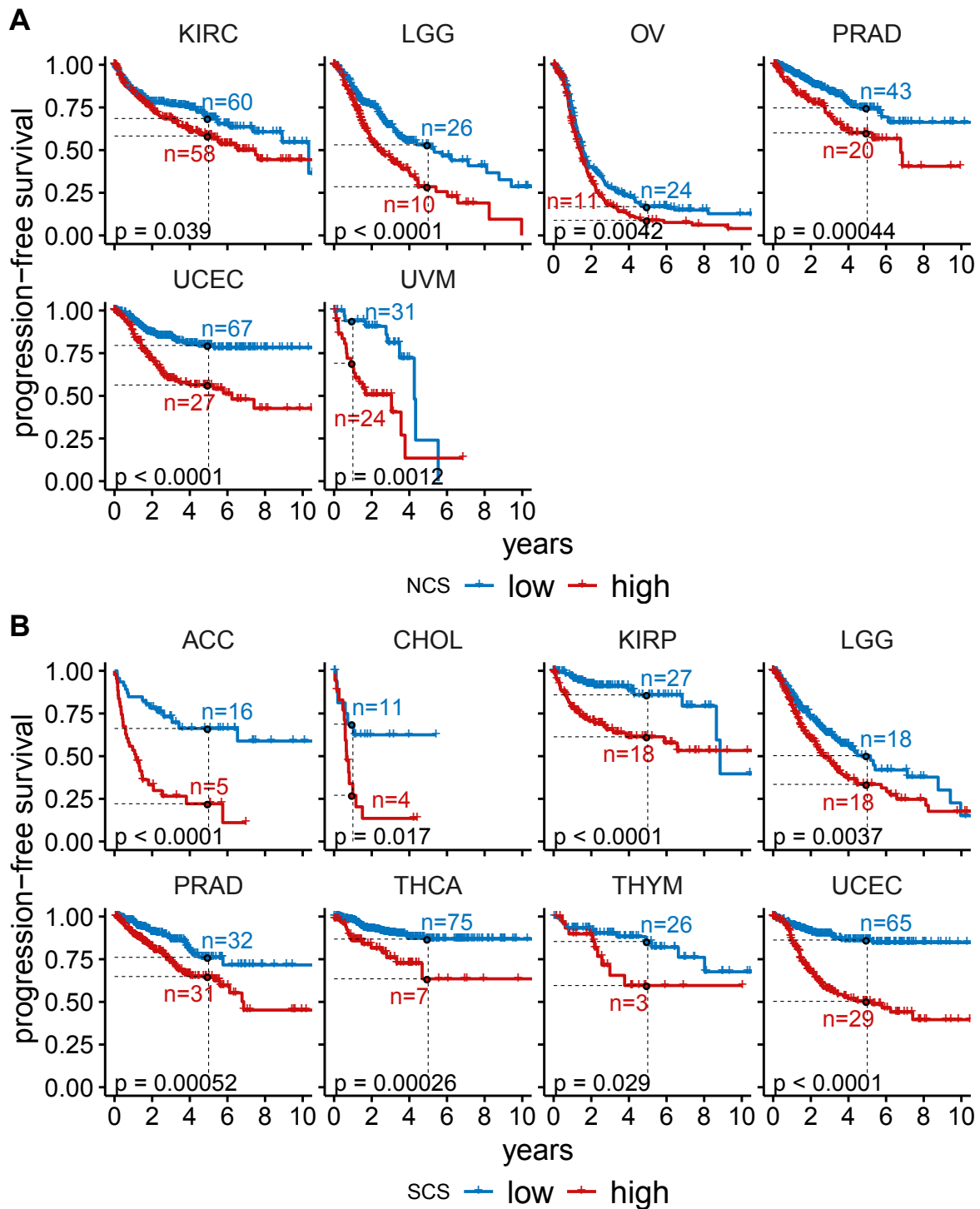


Figure A4. (A) Progression-free survival in six cancer types where significant differences between high- and low-NCS groups were observed. (B) Progression-free survival in eight cancer types where significant differences between high- and low-SCS groups were observed.

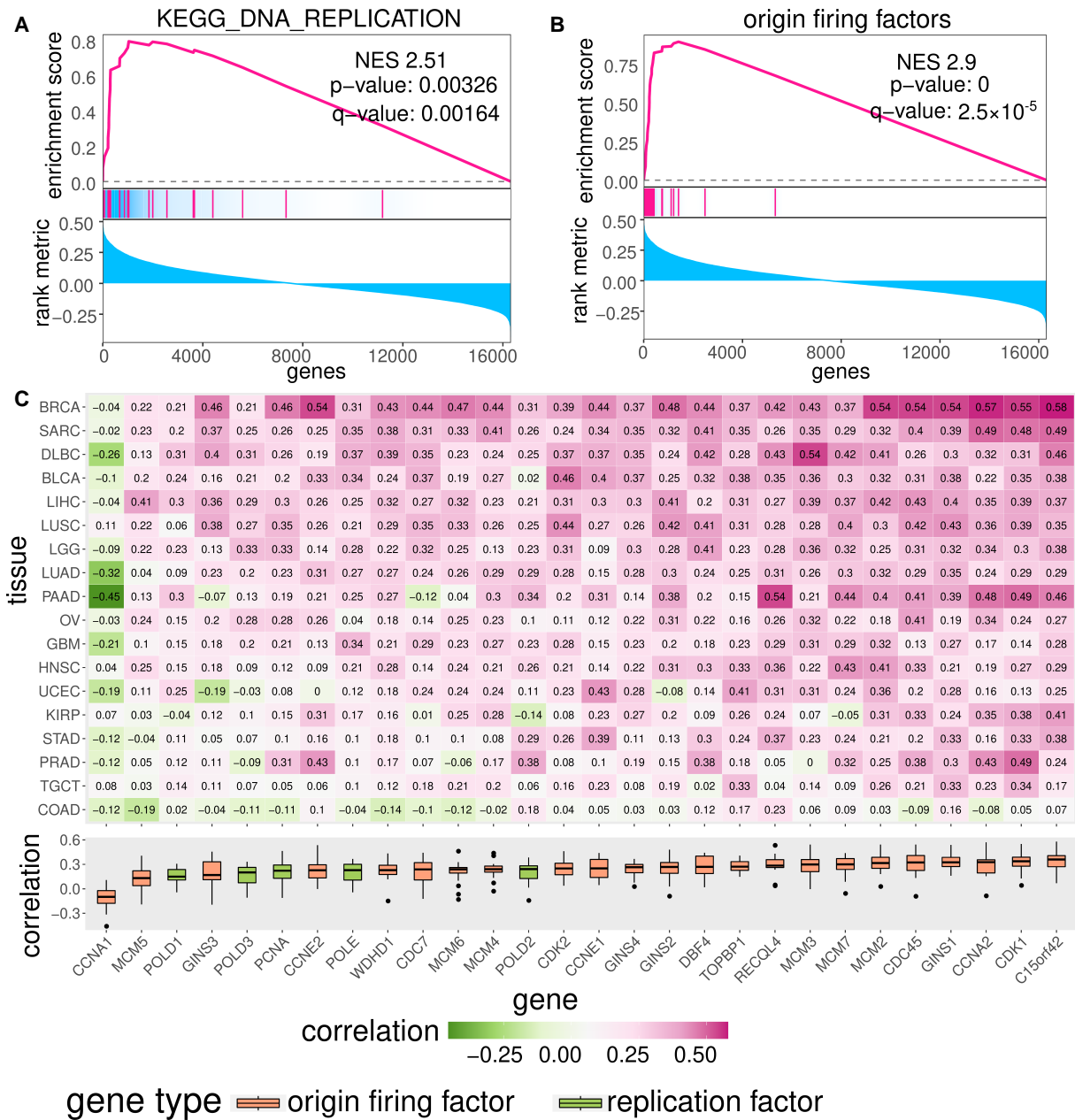


Figure A5. (A) Gene set enrichment analysis (GSEA) [67] for KEGG [68] DNA replication gene set. All genes are ordered according to the correlation of their expression with the SCS and enrichment significance is evaluated using permutation test. (B) GSEA analysis for manually curated origin firing factor gene set (curated in [10]). (C) Gene expression of many origin firing factors is positively correlated with SCS in many cancer types. Rows and columns of the heatmap represent cancer types and origin firing factor genes, respectively. Cancer types are clustered based on their correlation coefficients with origin firing factors. Genes are ordered based on the median correlation coefficient. Colour and values encoded in the heatmap represent the Spearman correlation coefficient.

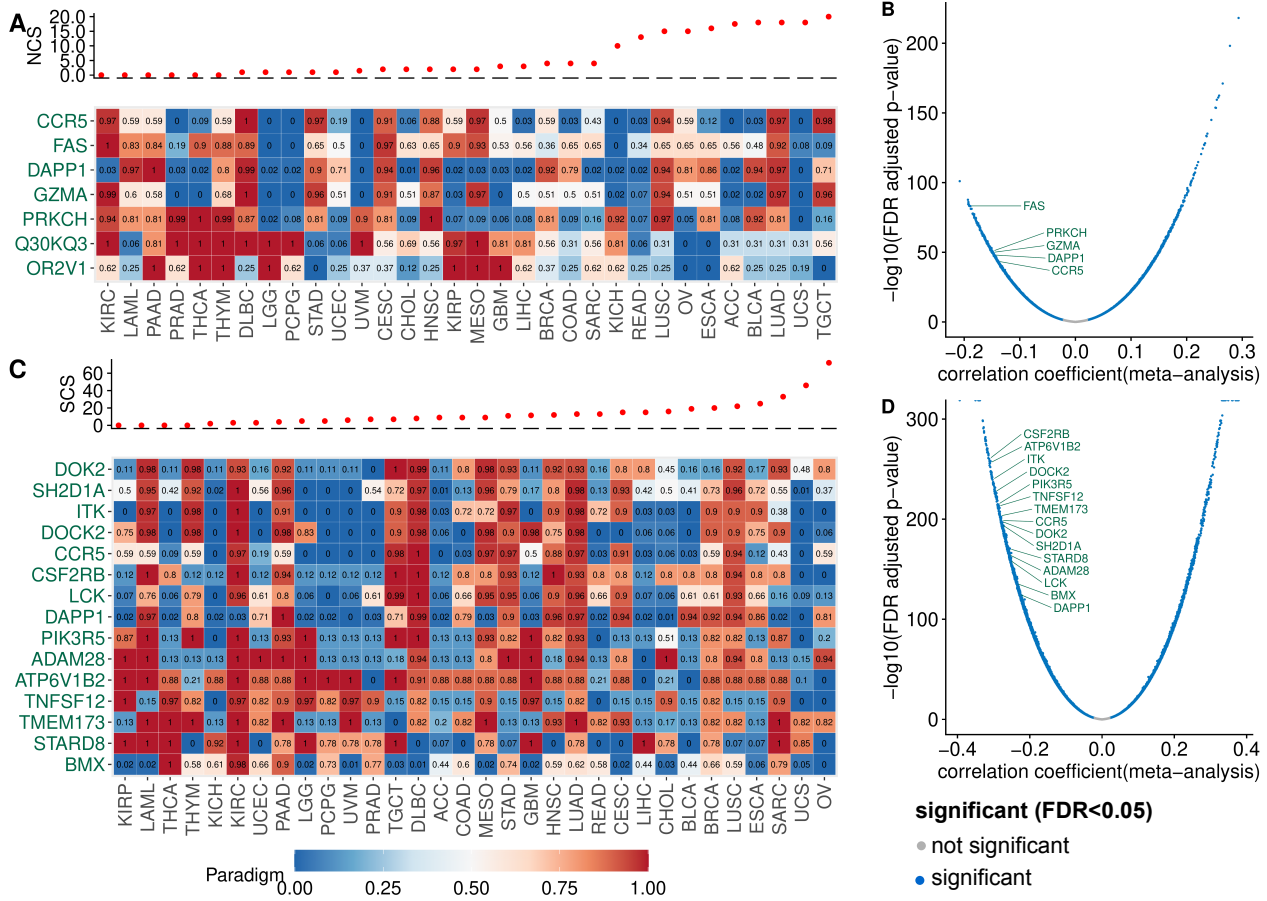


Figure A6. Paradigm pathway activity and gene expression negatively associated with CIN. (A) The PARADIGM pathway-level activities corresponding to protein-coding genes (rows) were correlated with the NCS. Only pathways with a significant negative correlation (FAD-adjusted $p < 5\%$) less than -0.3 in at least seven cancer types were included. The heatmap shows the normalised PARADIGM pathway activity (0–1 from low to high). Cancer types are ordered according to their median NCS, see top panel. (B) Volcano plot for the correlation between gene expression and NCS, highlighting gene names corresponding to PARADIGM features that are significantly negatively associated with NCS. (C) Analogous to (A), but for the SCS instead of the NCS. (D) Correlation of SCS and gene expression, analogous to (B).

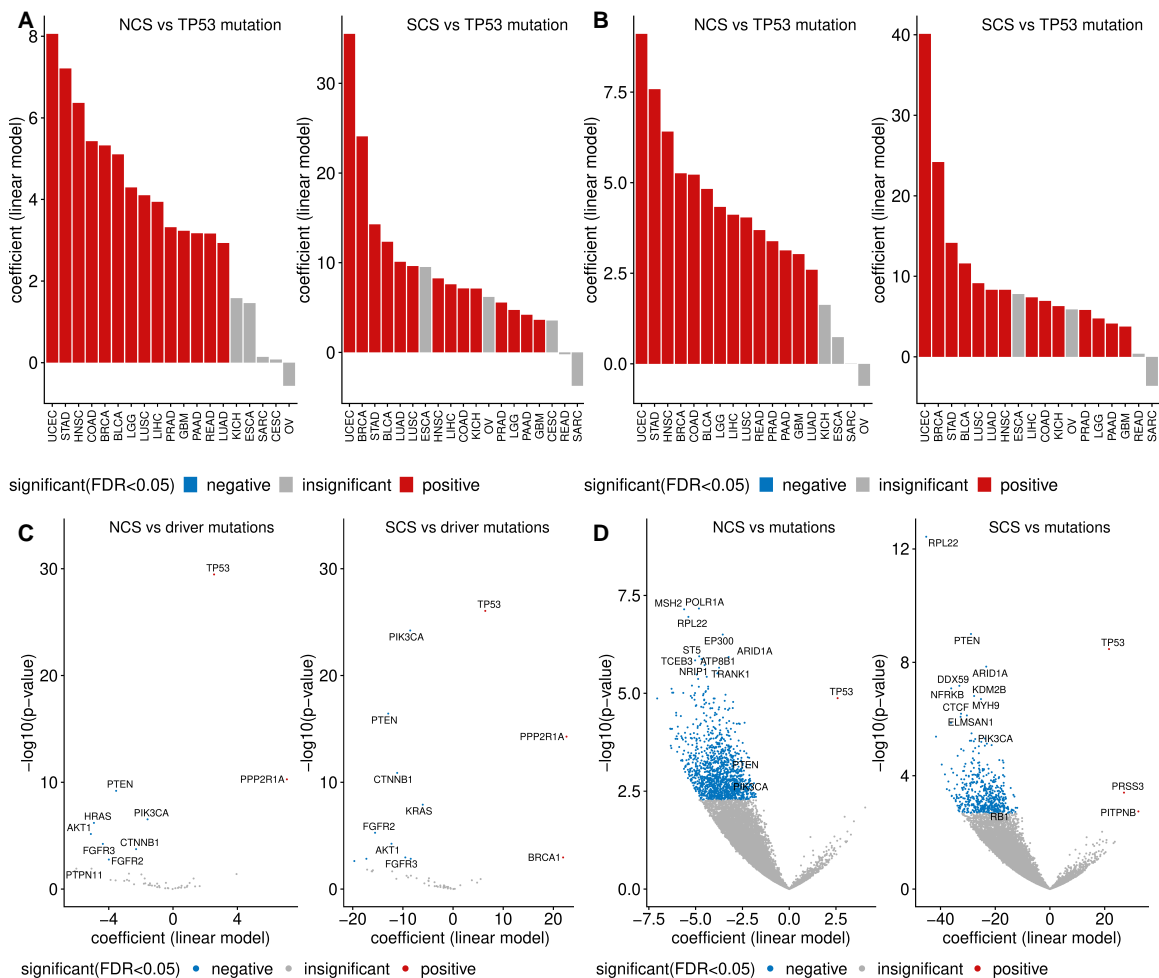


Figure A7. (A) *TP53* mutation is positively associated with high W-CIN in multiple cancer types. The bar shows the linear regression model coefficient using NCS as dependent variable and *TP53* mutation as explanatory variable. Only cancer types with both *TP53* mutant and *TP53* wild type in ≥ 20 samples are considered. (B) *TP53* mutation is positively associated with SCS, the association analysis is performed as in (A), except excluding MIN samples. (C) The volcano plot shows the association between CIN and validated driver mutations, association analysis is performed using non-MIN samples only. (D) The volcano plot shows the correlation between CIN score and somatic mutations in CCLE cell line samples. For all associations in (A–D), red, grey and blue encode positive, insignificant and negative associations. FAD-adjusted $p \leq 0.05$ is considered as significant. Gene names of known important oncogenes and CIN driver genes are annotated in the volcano plot, if significantly associated with CIN.

References

1. Carter, S.L.; Eklund, A.C.; Kohane, I.S.; Harris, L.N.; Szallasi, Z. A Signature of Chromosomal Instability Inferred From Gene Expression Profiles Predicts Clinical Outcome In Multiple Human Cancers. *Nat. Genet.* **2006**, *38*, 1043–1048. [[CrossRef](#)] [[PubMed](#)]
2. Bakhoun, S.F.; Cantley, L.C. The Multifaceted Role of Chromosomal Instability in Cancer and Its Microenvironment. *Cell* **2018**, *174*, 1347–1360. [[CrossRef](#)] [[PubMed](#)]
3. Sansregret, L.; Vanhaesebroeck, B.; Swanton, C. Determinants and clinical implications of chromosomal instability in cancer. *Nat. Rev. Clin. Oncol.* **2018**, *15*, 139–150. [[CrossRef](#)]
4. Roschke, A.; Rozenblum, E. Multi-Layered Cancer Chromosomal Instability Phenotype. *Front. Oncol.* **2013**, *3*, 302. [[CrossRef](#)]
5. Thompson, S.L.; Bakhoun, S.F.; Compton, D.A. Mechanisms of Chromosomal Instability. *Curr. Biol.* **2010**, *20*, R285–R295. [[CrossRef](#)] [[PubMed](#)]

6. Siri, S.O.; Martino, J.; Gottifredi, V. Structural Chromosome Instability: Types, Origins, Consequences, and Therapeutic Opportunities. *Cancers* **2021**, *13*, 3056. [[CrossRef](#)] [[PubMed](#)]
7. Burrell, R.A.; McClelland, S.E.; Endesfelder, D.; Groth, P.; Weller, M.C.; Shaikh, N.; Domingo, E.; Kanu, N.; Dewhurst, S.M.; Gronroos, E.; et al. Replication Stress Links Structural and Numerical Cancer Chromosomal Instability. *Nature* **2013**, *494*, 492–496. [[CrossRef](#)]
8. Wilhelm, T.; Said, M.; Naim, V. DNA Replication Stress and Chromosomal Instability: Dangerous Liaisons. *Genes* **2020**, *11*, 642. [[CrossRef](#)]
9. Böhly, N.; Kistner, M.; Bastians, H. Mild replication stress causes aneuploidy by deregulating microtubule dynamics in mitosis. *Cell Cycle* **2019**, *18*, 2770–2783. [[CrossRef](#)]
10. Schmidt, A.K.; Boehly, N.; Zhang, X.; Slusarenko, B.O.; Hennecke, M.; Kschischo, M.; Bastians, H. Dormant replication origin firing links replication stress to whole chromosomal instability in human cancer. *bioRxiv* **2021**, 463929. [[CrossRef](#)]
11. Passerini, V.; Ozeri-Galai, E.; de Pagter, M.S.; Donnelly, N.; Schmalbrock, S.; Kloosterman, W.P.; Kerem, B.; Storchová, Z. The Presence of Extra Chromosomes Leads to Genomic Instability. *Nat. Commun.* **2016**, *7*, 10754. [[CrossRef](#)] [[PubMed](#)]
12. Dewhurst, S.M.; McGranahan, N.; Burrell, R.A.; Rowan, A.J.; Grönroos, E.; Endesfelder, D.; Joshi, T.; Mouradov, D.; Gibbs, P.; Ward, R.L.; et al. Tolerance of Whole-Genome Doubling Propagates Chromosomal Instability and Accelerates Cancer Genome Evolution. *Cancer Discov.* **2014**, *4*, 175–185. [[CrossRef](#)] [[PubMed](#)]
13. Bakhoun, S.F.; Danilova, O.V.; Kaur, P.; Levy, N.B.; Compton, D.A. Chromosomal Instability Substantiates Poor Prognosis in Patients with Diffuse Large B-cell Lymphoma. *Clin. Cancer Res.* **2011**, *17*, 7704–7711. [[CrossRef](#)] [[PubMed](#)]
14. Tjihuis, A.E.; Johnson, S.C.; McClelland, S.E. The emerging links between chromosomal instability (CIN), metastasis, inflammation and tumour immunity. *Mol. Cytogenet.* **2019**, *12*, 17. [[CrossRef](#)]
15. Roylance, R.; Endesfelder, D.; Gorman, P.; Burrell, R.A.; Sander, J.; Tomlinson, I.; Hanby, A.M.; Speirs, V.; Richardson, A.L.; Birkbak, N.J.; et al. Relationship of extreme chromosomal instability with long-term survival in a retrospective analysis of primary breast cancer. *Cancer Epidemiol. Biomarkers Prev.* **2011**, *20*, 2183–2194. [[CrossRef](#)]
16. Birkbak, N.J.; Eklund, A.C.; Li, Q.; McClelland, S.E.; Endesfelder, D.; Tan, P.; Tan, I.B.; Richardson, A.L.; Szallasi, Z.; Swanton, C. Paradoxical Relationship between Chromosomal Instability and Survival Outcome in Cancer. *Cancer Res.* **2011**, *71*, 3447–3452. [[CrossRef](#)]
17. Gronroos, E.; López-García, C. Tolerance of Chromosomal Instability in Cancer: Mechanisms and Therapeutic Opportunities. *Cancer Res.* **2018**, *78*, 6529–6535. [[CrossRef](#)]
18. Thompson, S.L.; Compton, D.A. Proliferation of aneuploid human cells is limited by a p53-dependent mechanism. *J. Cell Biol.* **2010**, *188*, 369–381. [[CrossRef](#)]
19. Sheltzer, J.M. A Transcriptional and Metabolic Signature of Primary Aneuploidy Is Present in Chromosomally Unstable Cancer Cells and Informs Clinical Prognosis. *Cancer Res.* **2013**, *73*, 6401–6412. [[CrossRef](#)]
20. Endesfelder, D.; Burrell, R.A.; Kanu, N.; McGranahan, N.; Howell, M.; Parker, P.J.; Downward, J.; Swanton, C.; Kschischo, M. Chromosomal Instability Selects Gene Copy-Number Variants Encoding Core Regulators of Proliferation in ER+ Breast Cancer. *Cancer Res.* **2014**, *74*, 4853–4863. [[CrossRef](#)]
21. Buccitelli, C.; Salgueiro, L.; Rowald, K.; Sotillo, R.; Mardin, B.R.; Korbel, J.O. Pan-cancer analysis distinguishes transcriptional changes of aneuploidy from proliferation. *Genome Res.* **2017**, *27*, 501–511. [[CrossRef](#)] [[PubMed](#)]
22. Davoli, T.; Uno, H.; Wooten, E.C.; Elledge, S.J. Tumor Aneuploidy Correlates with Markers of Immune Evasion and with Reduced Response to Immunotherapy. *Science* **2017**, *355*, eaaf8399. [[CrossRef](#)] [[PubMed](#)]
23. Bakhoun, S.F.; Ngo, B.; Laughney, A.M.; Cavallo, J.A.; Murphy, C.J.; Ly, P.; Shah, P.; Sriram, R.K.; Watkins, T.B.K.; Taunk, N.K.; et al. Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature* **2018**, *553*, 467–472. [[CrossRef](#)]
24. Salgueiro, L.; Buccitelli, C.; Rowald, K.; Somogyi, K.; Kandala, S.; Korbel, J.O.; Sotillo, R. Acquisition of chromosome instability is a mechanism to evade oncogene addiction. *EMBO Mol. Med.* **2020**, *12*, e10941. [[CrossRef](#)] [[PubMed](#)]
25. Lukow, D.A.; Sausville, E.L.; Suri, P.; Chunduri, N.K.; Wieland, A.; Leu, J.; Smith, J.C.; Girish, V.; Kumar, A.A.; Kendall, J.; et al. Chromosomal instability accelerates the evolution of resistance to anti-cancer therapies. *Dev. Cell* **2021**, *56*, 2427–2439.e4. [[CrossRef](#)]
26. Lee, A.J.; Endesfelder, D.; Rowan, A.J.; Walther, A.; Birkbak, N.J.; Futreal, P.A.; Downward, J.; Szallasi, Z.; Tomlinson, I.P.; Howell, M.; et al. Chromosomal Instability Confers Intrinsic Multidrug Resistance. *Cancer Res.* **2011**, *71*, 1858–1870. [[CrossRef](#)] [[PubMed](#)]
27. Taylor, A.M.; Shih, J.; Ha, G.; Gao, G.F.; Zhang, X.; Berger, A.C.; Schumacher, S.E.; Wang, C.; Hu, H.; Liu, J.; et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **2018**, *33*, 676–689.e3. [[CrossRef](#)]
28. Chang, K.; Creighton, C.J.; Davis, C.; Donehower, L.; Drummond, J.; Wheeler, D.; Ally, A.; Balasundaram, M.; Birol, I.; Butterfield, Y.S.N.; et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [[CrossRef](#)]
29. Carter, S.L.; Cibulskis, K.; Helman, E.; McKenna, A.; Shen, H.; Zack, T.; Laird, P.W.; Onofrio, R.C.; Winckler, W.; Weir, B.A.; et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **2012**, *30*, 413–421. [[CrossRef](#)]
30. Vaske, C.J.; Benz, S.C.; Sanborn, J.Z.; Earl, D.; Szeto, C.; Zhu, J.; Haussler, D.; Stuart, J.M. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **2010**, *26*, i237–i245. [[CrossRef](#)]

31. Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A.A.; Kim, S.; Wilson, C.J.; Lehár, J.; Kryukov, G.V.; Sonkin, D.; et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **2012**, *483*, 603–607. [[CrossRef](#)] [[PubMed](#)]
32. 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. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov.* **2015**, *5*, 1210–1223. [[CrossRef](#)] [[PubMed](#)]
33. Thorsson, V.; Gibbs, D.L.; Brown, S.D.; Wolf, D.; Bortone, D.S.; Ou Yang, T.H.; Porta-Pardo, E.; Gao, G.F.; Plaisier, C.L.; Eddy, J.A.; et al. The Immune Landscape of Cancer. *Immunity* **2018**, *48*, 812–830.e14. [[CrossRef](#)]
34. Bailey, M.H.; Tokheim, C.; Porta-Pardo, E.; Sengupta, S.; Bertrand, D.; Weerasinghe, A.; Colaprico, A.; Wendl, M.C.; Kim, J.; Reardon, B.; et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **2018**, *173*, 371–385.e18. [[CrossRef](#)] [[PubMed](#)]
35. Therneau, T.M. *A Package for Survival Analysis in R*; R Package Version 3.2-10; 2021. Available online: <https://CRAN.R-project.org/package=survival> (accessed on 4 October 2021).
36. Kassambara, A.; Kosinski, M.; Biecek, P. *Survminer: Drawing Survival Curves Using 'ggplot2'*; R Package Version 0.4.9; 2021. Available online: <https://CRAN.R-project.org/package=survminer> (accessed on 4 October 2021).
37. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. [[CrossRef](#)] [[PubMed](#)]
38. Lengauer, C.; Kinzler, K.W.; Vogelstein, B. Genetic instabilities in human cancers. *Nature* **1998**, *396*, 643–649. [[CrossRef](#)] [[PubMed](#)]
39. McGranahan, N.; Burrell, R.A.; Endesfelder, D.; Novelli, M.R.; Swanton, C. Cancer Chromosomal Instability: Therapeutic and Diagnostic Challenges. *EMBO Rep.* **2012**, *13*, 528–538. [[CrossRef](#)]
40. Lepage, C.C.; Morden, C.R.; Palmer, M.C.L.; Nachtigal, M.W.; McManus, K.J. Detecting Chromosome Instability in Cancer: Approaches to Resolve Cell-to-Cell Heterogeneity. *Cancers* **2019**, *11*, 226. [[CrossRef](#)]
41. Delaney, J.R.; Patel, C.B.; Willis, K.M.; Haghighiabyaneh, M.; Axelrod, J.; Tancioni, I.; Lu, D.; Bapat, J.; Young, S.; Cadassou, O.; et al. Haploinsufficiency networks identify targetable patterns of allelic deficiency in low mutation ovarian cancer. *Nat. Commun.* **2017**, *8*, 14423. [[CrossRef](#)]
42. van Jaarsveld, R.H.; Kops, G.J. Difference Makers: Chromosomal Instability versus Aneuploidy in Cancer. *Trends Cancer* **2016**, *2*, 561–571. [[CrossRef](#)]
43. Sheltzer, J.M.; Amon, A. The aneuploidy paradox: costs and benefits of an incorrect karyotype. *Trends Genet.* **2011**, *27*, 446–453. [[CrossRef](#)] [[PubMed](#)]
44. Salmina, K.; Huna, A.; Kalejs, M.; Pjanova, D.; Scherthan, H.; Cragg, M.S.; Erenpreisa, J. The Cancer Aneuploidy Paradox: In the Light of Evolution. *Genes* **2019**, *10*, 83. [[CrossRef](#)] [[PubMed](#)]
45. Chunduri, N.K.; Storchová, Z. The diverse consequences of aneuploidy. *Nat. Cell Biol.* **2019**, *21*, 54–62. [[CrossRef](#)] [[PubMed](#)]
46. Lord, C.J.; Ashworth, A. BRCAness revisited. *Nat. Rev. Cancer* **2016**, *16*, 110–120. [[CrossRef](#)]
47. Turner, N.; Tutt, A.; Ashworth, A. Hallmarks of 'BRCAness' in sporadic cancers. *Nat. Rev. Cancer* **2004**, *4*, 814–819. [[CrossRef](#)]
48. Thompson, L.; Jeusset, L.; Lepage, C.; McManus, K. Evolving Therapeutic Strategies to Exploit Chromosome Instability in Cancer. *Cancers* **2017**, *9*, 151. [[CrossRef](#)]
49. Wu, Y.L.; Zhou, C.; Hu, C.P.; Feng, J.; Lu, S.; Huang, Y.; Li, W.; Hou, M.; Shi, J.H.; Lee, K.Y.; et al. Afatinib versus cisplatin plus gemcitabine for first-line treatment of Asian patients with advanced non-small-cell lung cancer harbouring EGFR mutations (LUX-Lung 6): An open-label, randomised phase 3 trial. *Lancet Oncol.* **2014**, *15*, 213–222. [[CrossRef](#)]
50. Bugler, B.; Schmitt, E.; Aressy, B.; Ducommun, B. Unscheduled expression of CDC25B in S-phase leads to replicative stress and DNA damage. *Mol. Cancer* **2010**, *9*, 29. [[CrossRef](#)]
51. Kamada, K. The GINS Complex: Structure and Function. In *The Eukaryotic Replisome: A Guide to Protein Structure and Function*; MacNeill, S., Ed.; Springer: Dordrecht, The Netherlands, 2012; Volume 62, pp. 135–156. [[CrossRef](#)]
52. Shi, W.; Huang, Q.; Xie, J.; Wang, H.; Yu, X.; Zhou, Y. CKS1B as Drug Resistance-Inducing Gene—A Potential Target to Improve Cancer Therapy. *Front. Oncol.* **2020**, *10*, 582451. [[CrossRef](#)]
53. Eischen, C.M. Genome Stability Requires p53. *Cold Spring Harb. Perspect. Med.* **2016**, *6*, a026096. [[CrossRef](#)]
54. Ciriello, G.; Miller, M.L.; Aksoy, B.A.; Senbabaoglu, Y.; Schultz, N.; Sander, C. Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **2013**, *45*, 1127–1133. [[CrossRef](#)] [[PubMed](#)]
55. Tamborero, D.; Rubio-Perez, C.; Deu-Pons, J.; Schroeder, M.P.; Vivancos, A.; Rovira, A.; Tusquets, I.; Albanell, J.; Rodon, J.; Tabernero, J.; et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **2018**, *10*, 25. [[CrossRef](#)] [[PubMed](#)]
56. Carramusa, L.; Contino, F.; Ferro, A.; Minafra, L.; Perconti, G.; Giallongo, A.; Feo, S. The PVT-1 oncogene is a Myc protein target that is overexpressed in transformed cells. *J. Cell. Physiol.* **2007**, *213*, 511–518. [[CrossRef](#)] [[PubMed](#)]
57. Zhang, C.; Hu, H.; Wang, X.; Zhu, Y.; Jiang, M. WFDC Protein: A Promising Diagnosis Biomarker of Ovarian Cancer. *J. Cancer* **2021**, *12*, 5404–5412. [[CrossRef](#)]
58. López, S.; Lim, E.L.; Horswell, S.; Haase, K.; Huebner, A.; Dietzen, M.; Mourikis, T.P.; Watkins, T.B.K.; Rowan, A.; Dewhurst, S.M.; et al. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nat. Genet.* **2020**, *52*, 283–293. [[CrossRef](#)]

59. Quinton, R.J.; DiDomizio, A.; Vittoria, M.A.; Kotýnková, K.; Ticas, C.J.; Patel, S.; Koga, Y.; Vakhshoorzadeh, J.; Hermance, N.; Kuroda, T.S.; et al. Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature* **2021**, *590*, 492–497. [[CrossRef](#)]
60. Bakhoun, S.F.; Kabeche, L.; Murnane, J.P.; Zaki, B.I.; Compton, D.A. DNA-Damage Response during Mitosis Induces Whole-Chromosome Missegregation. *Cancer Discov.* **2014**, *4*, 1281–1289. [[CrossRef](#)]
61. Bakhoun, S.F.; Kabeche, L.; Wood, M.D.; Laucius, C.D.; Qu, D.; Laughney, A.M.; Reynolds, G.E.; Louie, R.J.; Phillips, J.; Chan, D.A.; et al. Numerical chromosomal instability mediates susceptibility to radiation treatment. *Nat. Commun.* **2015**, *6*, 5990. [[CrossRef](#)]
62. Davoli, T.; Xu, A.W.; Mengwasser, K.E.; Sack, L.M.; Yoon, J.C.; Park, P.J.; Elledge, S.J. Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome. *Cell* **2013**, *155*, 948–962. [[CrossRef](#)]
63. Berenjano, I.M.; Piñeiro, R.; Castillo, S.D.; Pearce, W.; McGranahan, N.; Dewhurst, S.M.; Meniel, V.; Birkbak, N.J.; Lau, E.; Sansregret, L.; et al. Oncogenic PIK3CA induces centrosome amplification and tolerance to genome doubling. *Nat. Commun.* **2017**, *8*, 1773. [[CrossRef](#)]
64. Laughney, A.; Elizalde, S.; Genovese, G.; Bakhoun, S. Dynamics of Tumor Heterogeneity Derived from Clonal Karyotypic Evolution. *Cell Rep.* **2015**, *12*, 809–820. [[CrossRef](#)] [[PubMed](#)]
65. Heng, H.; Liu, G.; Stevens, J.; Abdallah, B.; Horne, S.; Ye, K.; Bremer, S.; Chowdhury, S.; Ye, C. Karyotype Heterogeneity and Unclassified Chromosomal Abnormalities. *Cytogenet. Genome Res.* **2013**, *139*, 144–157. [[CrossRef](#)] [[PubMed](#)]
66. Ye, C.J.; Stilgenbauer, L.; Moy, A.; Liu, G.; Heng, H.H. What Is Karyotype Coding and Why Is Genomic Topology Important for Cancer and Evolution? *Front. Genet.* **2019**, *10*, 1082. [[CrossRef](#)] [[PubMed](#)]
67. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [[CrossRef](#)] [[PubMed](#)]
68. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [[CrossRef](#)] [[PubMed](#)]

MFmap: A semi-supervised generative model matching cell lines to tumours and cancer subtypes

This chapter has been published as **peer-reviewed journal paper**:

© 2021 by the authors.

Licensed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0>).

X. Zhang and M. Kschischo. “MFmap: A semi-supervised generative model matching cell lines to tumours and cancer subtypes.” In: *PLOS ONE* 16.12 (2021), pp. 1–21. DOI: [10.1371/journal.pone.0261183](https://doi.org/10.1371/journal.pone.0261183)



Synopsis: Cell lines are the most commonly used model systems for better understanding of cancer biology (e.g. chromosomal instability process) and for developing new anti-cancer treatments. Selecting the best cell lines that closely mimic a given tumour or a cancer subtype is critical to translate the promising *in vitro* experiments to clinical treatment. We develop MFmap (model fidelity map), a semi-supervised generative model that combines very good generative and classification performance to integrate multi-omic data of cell lines and bulk tumours. MFmap embeds high dimensional features from somatic mutations, copy number variations and gene expressions into cancer subtype specific latent representations and predicts cancer subtype labels for cell lines. Pairwise cell-line-tumour similarity can be calculated based on the MFmap embedded latent representations, with which one can select the best cell lines for a cancer subtype or even individual tumours. We show that the MFmap embedded latent representations capture the known and novel features of cancer subtypes. We demonstrate the usefulness of MFmap by two cases:

(i) Translating the *in vitro* drug screening results to individual tumours. (ii) *In silico* modelling the cell state transformation during cancer progression.

Contributions of thesis author: conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing.

RESEARCH ARTICLE

MFmap: A semi-supervised generative model matching cell lines to tumours and cancer subtypes

Xiaoxiao Zhang ^{1,2}, Maik Kschischo ^{1*}

1 Department of Mathematics and Technology, RheinAhrCampus, University of Applied Sciences Koblenz, Remagen, Germany, **2** Department of Informatics, Technical University of Munich, Munich, Germany

* kschischo@rheinahrcampus.de OPEN ACCESS

Citation: Zhang X, Kschischo M (2021) MFmap: A semi-supervised generative model matching cell lines to tumours and cancer subtypes. PLoS ONE 16(12): e0261183. <https://doi.org/10.1371/journal.pone.0261183>

Editor: Tao Huang, Chinese Academy of Sciences, CHINA

Received: July 21, 2021

Accepted: November 24, 2021

Published: December 16, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pone.0261183>

Copyright: © 2021 Zhang, Kschischo. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data used in this study are publicly available online. Detailed references to access the data can be found in the main text and Supplemental Information. We have also added a cloud folder at where all preprocessed

Abstract

Translating *in vitro* results from experiments with cancer cell lines to clinical applications requires the selection of appropriate cell line models. Here we present MFmap (model fidelity map), a machine learning model to simultaneously predict the cancer subtype of a cell line and its similarity to an individual tumour sample. The MFmap is a semi-supervised generative model, which compresses high dimensional gene expression, copy number variation and mutation data into cancer subtype informed low dimensional latent representations. The accuracy (test set F_1 score >90%) of the MFmap subtype prediction is validated in ten different cancer datasets. We use breast cancer and glioblastoma cohorts as examples to show how subtype specific drug sensitivity can be translated to individual tumour samples. The low dimensional latent representations extracted by MFmap explain known and novel subtype specific features and enable the analysis of cell-state transformations between different subtypes. From a methodological perspective, we report that MFmap is a semi-supervised method which simultaneously achieves good generative and predictive performance and thus opens opportunities in other areas of computational biology.

Introduction

Tumour-derived cell lines are important model systems for developing new anti-cancer treatments and for understanding cancer biology [1–3]. They are comparably cost efficient, easy to handle under laboratory conditions and do not inflict ethical issues arising in research involving human or animal subjects. Yet, promising cell line experiments are rarely translated to clinical applications. In some cases, there are remarkable differences between cell lines and the primary tumours they were derived from [2–4]. This is also the reason why the assignment of clinically informative tumour subtypes to cell line models [3–5] is not a straightforward task.

To narrow the gap between preclinical findings and tumour treatment, it is necessary to select appropriate cell line models for a given tumour sample or a given cancer subtype. Several attempts to evaluate similarities and differences between cell lines and bulk tumours have

data are available: <https://cloud.hs-koblenz.de/s/yfFKkzcK78AekL4>.

Funding: This work was supported by the FOR2800 research unit funded by the Deutsche Forschungsgemeinschaft (DFG project number 395736209). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

focused on associations between corresponding data modalities including mutation, copy number, gene expression and methylation [6–12]. An important data resource comes from collaborative projects like NCI-60 [13] and the Cancer Cell Line Encyclopaedia (CCLE) [5, 14], who have generated large-scale pharmacogenomics data from patient-derived cell lines across organs. Other efforts like Sanger Genomics of Drug Sensitivity in Cancer (GDSC) [15], Connectivity Map (CMAP) [16], the Cancer Therapeutics Response Portal (CTRP v1 and CTRP v2) [17, 18] further expanded the datasets. On the other hand, The Cancer Genome Atlas (TCGA) [19] and the International Cancer Genome Consortium (ICGC) [20] systematically characterised molecular profiles of thousands of tumours. These complementary data resources are valuable for understanding the complexity of cancer biology and connecting *in vitro* pharmacogenomic profiles to patient molecular characteristics, potentially informing anti-cancer treatment strategies.

Integrative analyses considering multiple data types of both cell lines and bulk tumours are still challenging and new analysis concepts tailored towards specific questions are an ongoing research topic. For instance, Collector [21] preselects the most frequent genomic alterations and defines cancer subtypes based on a sequence of these alterations. Although such a preselection of genomic alterations integrates prior knowledge about cancer mutational patterns, it neglects complementary information contained in other data types. Furthermore, Collector relies on a binary matrix of genomic alterations. This matrix is very sparse, since samples harbouring the same alterations are very rare. Therefore, the statistical power to detect appropriate cell lines for tumours might be limited.

A recent study [22] highlighted that independent classifiers based on different data types to predict cell line identity often yield inconsistent results. For example, predictions based on the mutation spectrum and oncogenic mutations can be contradictory, although both features are derived from mutation data. Complementary information from different data sources is integrated by the MAGNETIC-framework [23] into gene modules. Gene set enrichment analysis (GSEA) is then used to interpret these modules as pathways. MAGNETIC is indeed a powerful technique for integrating multiple molecular datasets and prior knowledge, but it does not conclude to what extent a cell line is suitable as a tumour model. The maui framework assigns cancer subtype labels to cell lines by extracting relevant features from multiple data types using a variational autoencoder (VAE) [24]. However, most of the maui embedded features are weakly associated with subtype labels and are therefore difficult to interpret.

Here, we propose MFmap, a new semi-supervised VAE architecture and objective function which combines good classification accuracy with good generative performance. We exploit these properties to derive subtype informed low dimensional representations for both cell lines and bulk tumours from high dimensional multi-omics data including gene expression, mutation and copy number variation. The latent representations can then be used to assess the similarity between a cell line and a tumour. We provide cell line by tumour dissimilarity matrices for CCLE and TCGA for the ten different cancer types listed in Table 1. In addition, MFmap predicts cancer subtype labels for cell lines. We demonstrate, how these predicted cancer subtypes can be used to transfer information from cell-line-based drug sensitivity screens to patient cohorts. We also show, that the latent representations learnt by MFmap are biologically interpretable. Finally, we illustrate how the generative nature of the MFmap model can be exploited for studying subtype transformations during cancer progression. At http://h2926513.stratoserver.net:3838/MFmap_shiny/ we provide a resource enabling researchers to select the most relevant cell line for a cancer patient.

Table 1. The sample size of TCGA and CCLE data used for training and testing MFmap.

TCGA code	study name	number of subtypes	TCGA sample size	CCLE sample size
BRCA	Breast invasive carcinoma	4	484	51
COADREAD	Colon adenocarcinoma	4	414	54
ESCA	Esophageal carcinoma	2	169	27
HNSC	Head and neck squamous cell carcinoma	4	278	29
LUAD	Lung adenocarcinoma	3	227	70
LUSC	Lung squamous cell carcinoma	4	178	22
PAAD	Pancreatic adenocarcinoma	2	149	40
SKCM	Skin cutaneous melanoma	3	260	49
UCEC	Uterine corpus endometrial carcinoma	3	234	28
GBMLGG	Glioblastoma multiforme and lower grade glioma	7	621	55

<https://doi.org/10.1371/journal.pone.0261183.t001>

Materials and methods

Matching cell lines and tumours as a semi-supervised learning problem

MFmap is a semi-supervised deep neural network which integrates gene expression, copy number variation (CNV) and somatic mutation data with subtype classification. Each tumour sample t consists of a pair of (\mathbf{x}_t, y_t) , where $\mathbf{x}_t \in \mathbb{R}^D$ denotes the high dimensional molecular features and $y_t \in \{1, \dots, h\}$ is the cancer subtype label. For a cell line c , the cancer subtype is unknown and only the molecular features \mathbf{x}_c are available. The index c or t will be suppressed, whenever we refer to a single observation. The MFmap neural network is trained in a semi-supervised manner using both cell line data $\mathcal{D}_{cl}^{train} = \{\mathbf{x}_c\}_{c=1}^{C_{train}}$ and tumour data $\mathcal{D}_{tu}^{train} = \{(\mathbf{x}_t, y_t)\}_{t=1}^{T_{train}}$. Here, we used cell line data from CCLE and tumour data from TCGA.

One aim of MFmap is to use semi-supervised classification to infer the cancer subtype y_c of a cell line c . A second aim is to assess the similarity between a cell line and a tumour. Instead of comparing the high dimensional molecular features \mathbf{x}_t and \mathbf{x}_c directly, we first encode them into low dimensional latent representations \mathbf{z} (see next section for details). Then, the similarity of a tumour sample t and a cell line c is measured as the cosine coefficient between the corresponding latent representation vectors \mathbf{z}_t and \mathbf{z}_c . We will also show that these latent representations \mathbf{z} carry interpretable biological information.

The molecular data $\mathbf{x} = (\mathbf{x}_{DNA}, \mathbf{x}_{RNA})$ consist of gene expression profiles \mathbf{x}_{RNA} and network smoothed mutation and CNV profiles \mathbf{x}_{DNA} . We will refer to these two parts as RNA and DNA view, respectively. The DNA view is obtained from the original binary mutation and CNV matrices (Fig 1(A)), which indicate the occurrence of a mutation or CNV event targeting a gene in a given tumour sample or cell line. These very sparse matrices are first projected onto an annotated cancer network [25]. By using a network diffusion algorithm [26], a mutation or CNV signal hitting a single gene is propagated to neighbouring nodes in the network, thereby enriching the mutation or CNV data by cancer network information. All molecular features were translated and scaled to the interval between zero and one.

Specification of MFmap as a semi-supervised generative model

The MFmap neural network (Fig 1(B)) is a new variant of a semi-supervised VAE [27]. The observable data are considered to be drawn from the probability distributions $p(\mathbf{x}, y)$ for tumour samples and $p(\mathbf{x})$ for cell lines. These distributions are modelled as marginals over the

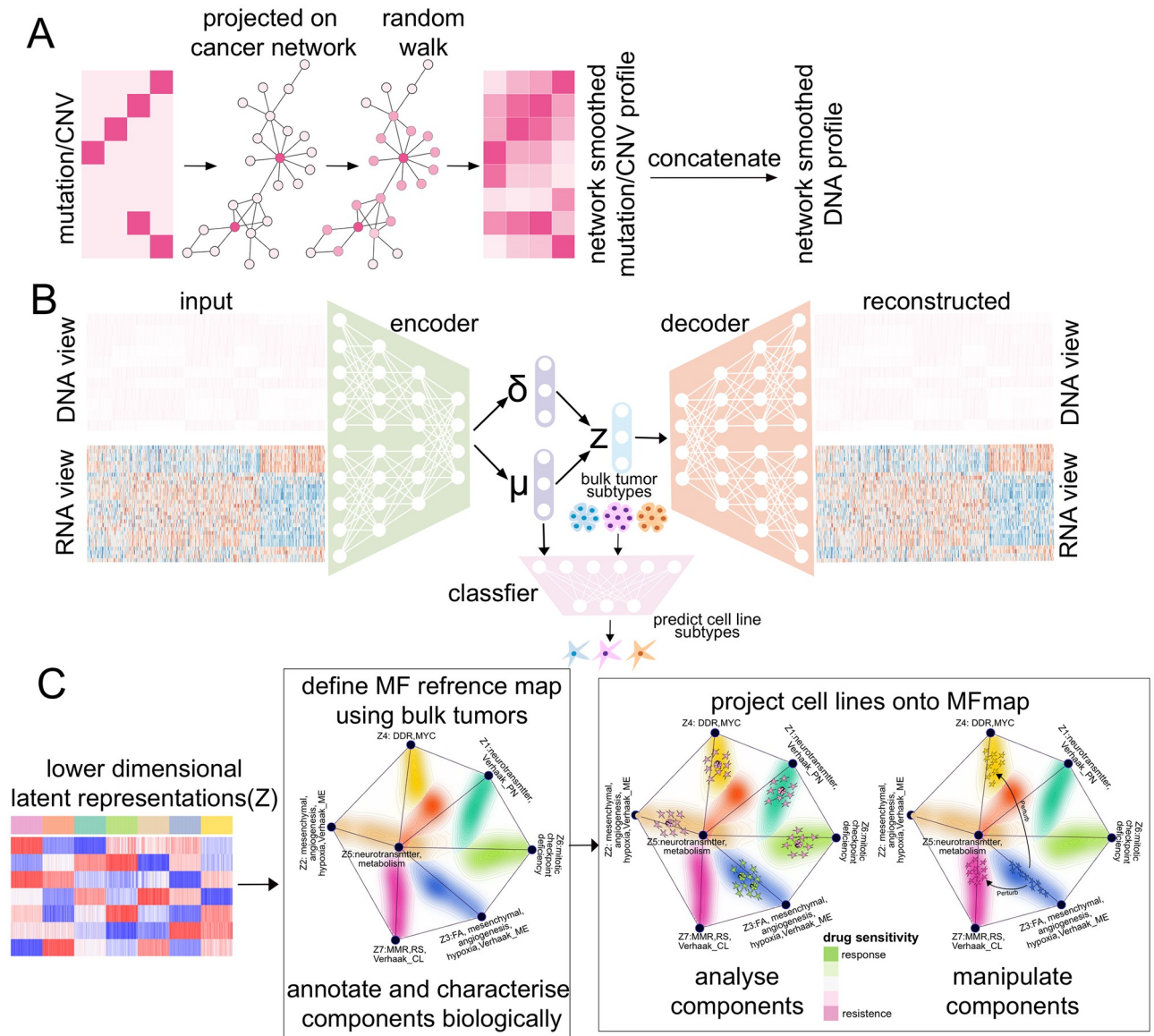


Fig 1. Overview of MFmap. (A) In a preprocessing step, mutation and CNV profiles are transformed to network smoothed DNA profiles. The original mutation and CNV data are represented as a binary matrix indicating the presence/absence of a DNA alteration in a given tumour sample or cell line. This sparse matrix is projected onto a cancer reference network (CRN) [25] and a network diffusion algorithm propagates this information to network neighbours, resulting in a dense DNA mutation or CNV matrix (DNA features). (B) The smoothed DNA features (DNA view) combined with gene expression data (RNA view) form the input of MFmap. The neural network architecture of MFmap has three components: encoder, decoder and classifier, encoded by different colours. The encoder maps sample features to a distribution $q(z|x)$ for the latent representation z with mean value $\mu(x)$ and covariance $\sigma^2(x)$. The classifier outputs a molecular subtype probability $p(y|z)$ and the decoder models a density $p(x|z)$ for the reconstruction of the DNA and RNA views. During semi-supervised training, the molecular subtypes of tumour samples are used. (C) For visualisation, the latent representations of bulk tumour samples are used to generate a reference map. Cell lines are then projected to the reference map. The colour coding of individual samples or cell lines (dots) indicates the tumour subtype or the predicted subtype, respectively. The density of the tumour samples is indicated by background contour lines coloured according to the subtypes.

<https://doi.org/10.1371/journal.pone.0261183.g001>

latent variable $\mathbf{z} = (z_1, \dots, z_d)^T \in \mathbb{R}^h$, such that

$$p(\mathbf{x}, y) = \int p(\mathbf{x}, y, \mathbf{z}) d\mathbf{z}, \quad p(\mathbf{x}) = \sum_{y=1}^h p(\mathbf{x}, y). \quad (1)$$

To facilitate biological interpretation of the latent representations, we set the dimension d of the latent space equal to the number of cancer subtypes h . In other applications of the MFmap model, one could also consider d as a tuneable hyper-parameter.

For the generative model, we assume \mathbf{x} and y to be conditionally independent given the latent variable \mathbf{z} . Accordingly, the joint distribution can be factorised as

$$p(\mathbf{x}, y, \mathbf{z}) = p(\mathbf{x}|\mathbf{z}) p(y|\mathbf{z}) p(\mathbf{z}). \quad (2)$$

These distributions are specified as

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|0, \mathbf{I}) \quad (3a)$$

$$p(y|\mathbf{z}) = \text{Cat}(y|\boldsymbol{\pi}_\theta(\mathbf{z})) \quad (3b)$$

$$p(\mathbf{x}|\mathbf{z}) = \mathbf{f}_\theta(\mathbf{x}|\mathbf{z}). \quad (3c)$$

Here, $p(\mathbf{z})$ is the prior distribution for the latent representation vector. We denote the Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ by $\mathcal{N}(\cdot|\boldsymbol{\mu}, \Sigma)$. The parameter $\boldsymbol{\pi}_\theta(\mathbf{z})$ of the categorical distribution $p(y|\mathbf{z})$ depends on the latent representation \mathbf{z} . For the decoder $p(\mathbf{x}|\mathbf{z})$ one can choose a suitable distribution \mathbf{f}_θ with parameters depending on the latent representations \mathbf{z} [27]. The functions $\mathbf{z} \mapsto \boldsymbol{\pi}_\theta(\mathbf{z})$ and $\mathbf{z} \mapsto \mathbf{f}_\theta(\cdot|\mathbf{z})$ are represented as neural networks. The parameters of these decoder networks are jointly denoted as θ .

For the mfMAP model we initially used a Gaussian distribution $\mathbf{f}_\theta(\mathbf{x}|\mathbf{z})$ to model the outputs. However, we found that rescaling the molecular features \mathbf{x} to the interval $[0, 1]$ and using a Bernoulli distribution for \mathbf{f}_θ improved the semi-supervised classification accuracy (see Results section). Then, each single output of the decoder neural network $\mathbf{z} \mapsto \mathbf{f}_\theta(\cdot|\mathbf{z})$ can be interpreted as the probability, that the corresponding molecular feature is active or not. For instance, for the i -th component $(\mathbf{x}_{RNA})_i$ of the RNA-view, the corresponding output can be regarded as the probability that the i -th gene is expressed.

Posterior inference, i.e. the evaluation of $p(y, \mathbf{z}|\mathbf{x})$ using Bayes theorem, is often intractable, because the marginal likelihood $p(\mathbf{x})$ in Eq (1) requires integrating over \mathbf{z} . Therefore, a variational distribution $q(y, \mathbf{z}|\mathbf{x})$ is introduced to approximate the true posterior [24, 27]. We assume that the variational distribution reflects the conditional independence $\mathbf{x} \perp y|\mathbf{z}$ of the generative model in Eq (2). This implies

$$q(\mathbf{x}, y|\mathbf{z}) = q(\mathbf{x}|\mathbf{z}) q(y|\mathbf{z}). \quad (4)$$

For consistency we assume that $q(y|\mathbf{z})$ in Eq (4) is identical to $p(y|\mathbf{z})$ in Eq (3b) and is represented by the same neural network mapping \mathbf{z} to the categorical parameter $\boldsymbol{\pi}_\theta(\mathbf{z})$. For the variational distribution $q(\mathbf{z}|\mathbf{x})$ we choose a Gaussian

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}(\mathbf{x}), \text{diag}(\boldsymbol{\sigma}(\mathbf{x}))) \quad \text{with} \quad (\boldsymbol{\mu}(\mathbf{x}), \log \boldsymbol{\sigma}(\mathbf{x})) = \mathbf{g}_\phi(\mathbf{x}) \quad (5)$$

with parameters $\boldsymbol{\mu}(\mathbf{x})$ and $\boldsymbol{\sigma}(\mathbf{x})$. The parameters are represented by the encoder neural network \mathbf{g}_ϕ , which is itself parametrised by ϕ . The overall architecture of MFmap (Fig 1(B)) is thus formed by three neural networks, the encoder Eq (5), the classifier Eq (3b) and the decoder Eq (3c).

Training of MFmap using a semi-supervised loss function

Variational inference involves maximising an evidence lower bound (ELBO) to the log-likelihood of the observational data [24, 27]. For a single cell line sample $\mathbf{x}_c \in \mathcal{D}_{cl}$ one can derive a lower bound to the log-likelihood

$$\log p(\mathbf{x}_c) = \log \left(\sum_y \int p(\mathbf{x}_c, y, \mathbf{z}) d\mathbf{z} \right) \geq \mathcal{L}(\mathbf{x}_c), \tag{6}$$

which is identical to the ELBO of the basic VAE [24] for unsupervised learning

$$\mathcal{L}(\mathbf{x}) = E_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \tag{7}$$

consisting of a reconstruction loss term and a Kullback-Leibler (KL) divergence term. For a single labelled tumour sample $(\mathbf{x}_t, y_t) \in \mathcal{D}_{tu}$ we have for the log-likelihood

$$\log p(\mathbf{x}_t, y_t) = \log \left(\int p(\mathbf{x}_t, y_t, \mathbf{z}) d\mathbf{z} \right) \geq \mathcal{L}_{tu}(\mathbf{x}_t, y_t), \tag{8}$$

where the ELBO for labelled examples reads

$$\mathcal{L}_{tu}(\mathbf{x}, y) = \mathcal{L}(\mathbf{x}) + E_{q(\mathbf{z}|\mathbf{x})}[\log p(y|\mathbf{z})]. \tag{9}$$

To derive this ELBO (see S1 File), we exploited the conditional independence assumption $\mathbf{x} \perp y|\mathbf{z}$ for both the generative model (Eq (2)) and the inference model (Eq (4)). The additional term in Eq (9) in comparison to Eq (7) can be interpreted as a classification loss. Given a tumour sample (\mathbf{x}_t, y_t) , the probability for the cancer subtype label $p(y_t|\mathbf{z})$ is a function of \mathbf{z} , which is inferred from $q(\mathbf{z}|\mathbf{x}_t)$. This distribution is in turn determined by the molecular feature vector \mathbf{x}_t .

We found empirically that the semi-supervised classification accuracy during training was relatively poor when using these exact negative ELBOs as loss functions. This is in line with previous findings that achieving both good semi-supervised classification accuracy and good generative performance is often difficult in VAEs [28] or other generative models [29]. Motivated by the work from [30], we added the negative entropy $\mathcal{H}[p(y|\mathbf{z})]$ of the distribution $p(y|\mathbf{z})$ to the unsupervised ELBO \mathcal{L} in Eq (7) and to the supervised ELBO \mathcal{L}_{tu} in Eq (9). In summary, the MFmap loss functions for the unlabelled cell line and the labelled tumour data are respectively given by

$$\begin{aligned} \mathcal{U}(\mathbf{x}) &= -\mathcal{L}(\mathbf{x}) + \mathcal{H}[p(y|\mathbf{z})] \\ &= -E_{q(\mathbf{z}|\mathbf{x})}[p(\mathbf{x}|\mathbf{z})] + D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + \mathcal{H}[p(y|\mathbf{z})] \end{aligned} \tag{10a}$$

$$\begin{aligned} \mathcal{S}(\mathbf{x}, y) &= -\mathcal{L}_{tu}(\mathbf{x}) + \mathcal{H}[p(y|\mathbf{z})] \\ &= -E_{q(\mathbf{z}|\mathbf{x})}[p(\mathbf{x}|\mathbf{z})] + D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + \mathcal{H}[p(y|\mathbf{z})] - E_{q(\mathbf{z}|\mathbf{x})}[\log p(y|\mathbf{z})]. \end{aligned} \tag{10b}$$

This entropy regularisation encourages the classification boundaries to be located in low sample density regions [30] in the latent space, which improves the generalisation performance of the model. As shown below (see Results section), the semi-supervised classification accuracy was very convincing, when using this entropy regularisation.

During training, mini-batches $b = 1, \dots, B$ from the cell line $\mathcal{D}_{cl}^{(b)} \subset \mathcal{D}_{cl}^{Train}$ and tumour data $\mathcal{D}_{tu}^{(b)} \subset \mathcal{D}_{tu}^{Train}$ are used to minimise

$$\sum_{\mathbf{x}_c \in \mathcal{D}_{cl}^{(b)}} \mathcal{U}(\mathbf{x}_c) + \sum_{(\mathbf{x}_t, y_t) \in \mathcal{D}_{tu}^{(b)}} \mathcal{S}(\mathbf{x}_t, y_t) \tag{11}$$

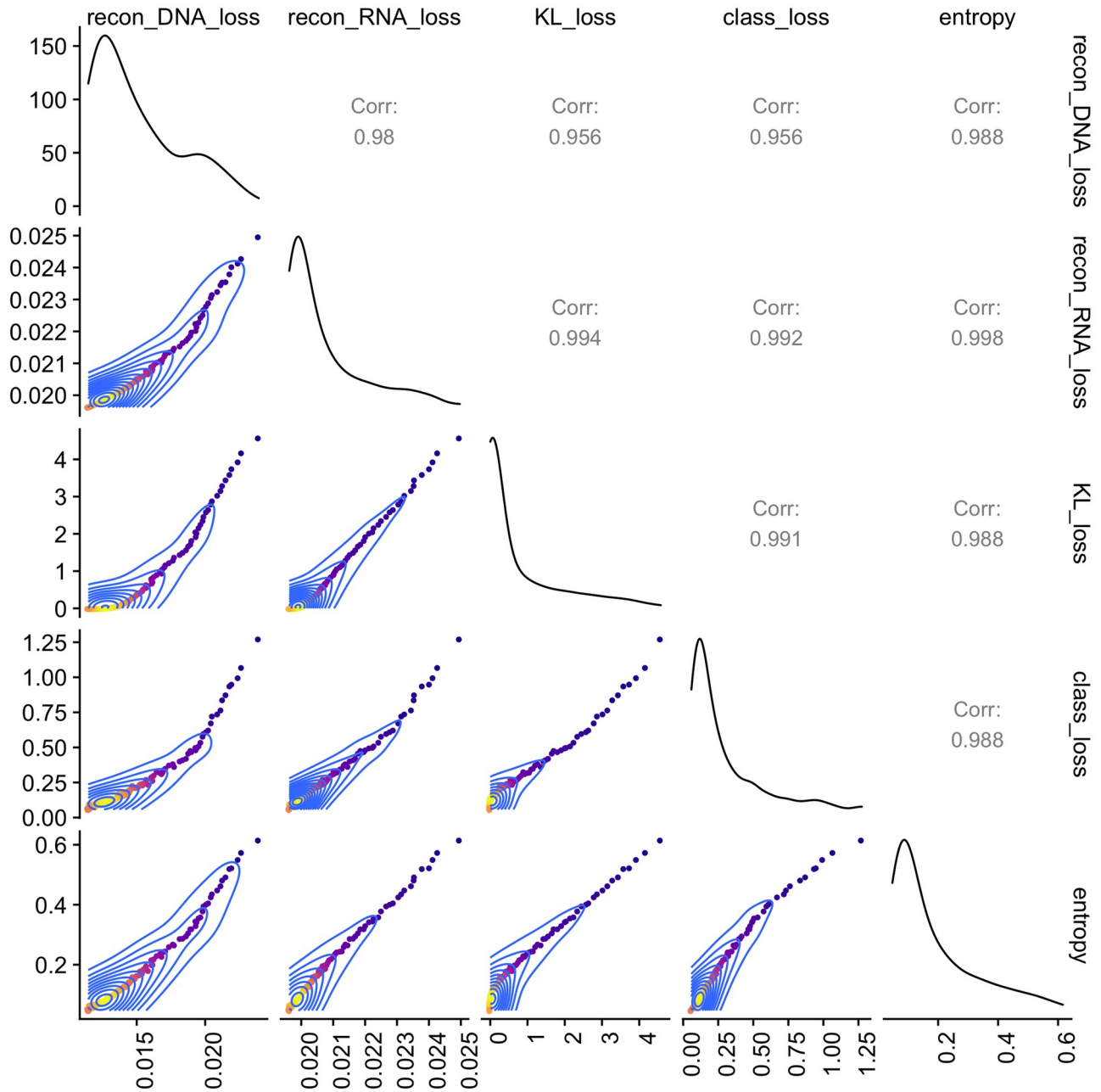


Fig 2. Joint optimisation of the reconstruction loss, the KL divergence, entropy and the classification loss with the MFmap loss function. The plot shows the pairwise correlation of different terms in the MFmap loss function Eq (10) during different training epochs.

<https://doi.org/10.1371/journal.pone.0261183.g002>

over different epochs. To check whether all terms in the MFmap loss function in Eq (10) can be jointly optimised, we recorded the values of each term in each training epoch and calculated their pair-wise correlations. The reconstruction loss $-E_{q(z|x)}[p(x|z)]$, the KL-divergence $D_{KL}(q(y|x)||p(z))$, the entropy $\mathcal{H}[p(y|z)]$ and the classification loss $-E_{q(z|x)}[\log p(x|z)]$ are highly correlated (Fig 2), what suggests that they are optimised simultaneously.

Visualisation of individual samples

The MFmap latent representation \mathbf{z} can be used to visualise and organise the associations of individual tumour samples and cell lines (Fig 1(C)). Inspired by the visualisation concept of Onco-GPS (OncoGenic Positioning System) [31], we used the tumour samples with known subtypes to generate a reference map for the cancer subtypes. In this reference map, the components z_1, \dots, z_h of the latent representation are presented as a graph with h corner points in a plane. The location of these corner points is determined by multidimensional scaling and is chosen so as to reflect the distances in the h -dimensional latent space as good as possible (see S1 File for details). An individual tumour sample can now be visualised as a point located in the area between the corner points. The location of such a point is given by a superposition of the corner positions weighted by the latent representation magnitudes of individual samples. In addition, the subtypes of the tumour samples are colour coded. The contour lines and the background colour shading represent the sample density in the region.

Once the reference map is established, individual cell lines can be projected to this map, where the colour of each dot encodes the subtype *predicted* by the MFmap classifier. This projection is based on the latent representation values of the cell line samples. Since our aim is to analyse the fidelity of a cell line as an oncological model for a given tumour or a cancer subtype, we name our framework the model fidelity map (MFmap).

Results

Evaluating the MFmap classification and generative performance

A direct evaluation of the MFmap subtype prediction for cell lines is impossible because there are no ground truth labels available. However, the classification accuracy on an unseen test dataset of bulk tumours provides an indirect evaluation of the subtype prediction performance. In Table 2 we used 20% of the tumour samples as independent test set and evaluated the classification performance using four multi-class classification metrics: overall accuracy, weighted precision, weighted recall, and weighted F_1 score. Similar results can be obtained, when 10% of the tumour samples are used for testing (see Table 1 in the S2 File). We also tested the effect of increasing the latent space dimension d and found that the classification accuracy was typically not higher, indicating that our choice of setting d equal to the number of cancer subtypes did not impair the classification accuracy (see Table 2 in the S2 File).

The good classification results for GBMLGG are intriguing, because the G-CIMP-High, G-CIMP-Low and LGM6-GBM subtypes were derived from methylation data [32], which

Table 2. MFmap subtype classification performance estimated for unseen tumour samples. Here, 20% of the bulk tumour data were randomly selected as an independent test set.

accuracy	precision	recall	F_1 score	organ
0.97	0.97	0.97	0.97	BRCA
0.96	0.96	0.96	0.96	COADREAD
1.00	1.00	1.00	1.00	ESCA
0.99	0.99	0.99	0.99	GBMLGG
0.91	0.92	0.91	0.91	HNSC
0.96	0.96	0.96	0.96	LUAD
0.94	0.95	0.94	0.94	LUSC
0.97	0.97	0.97	0.97	PAAD
1.00	1.00	1.00	1.00	SKCM
0.96	0.96	0.96	0.96	UCEC

<https://doi.org/10.1371/journal.pone.0261183.t002>

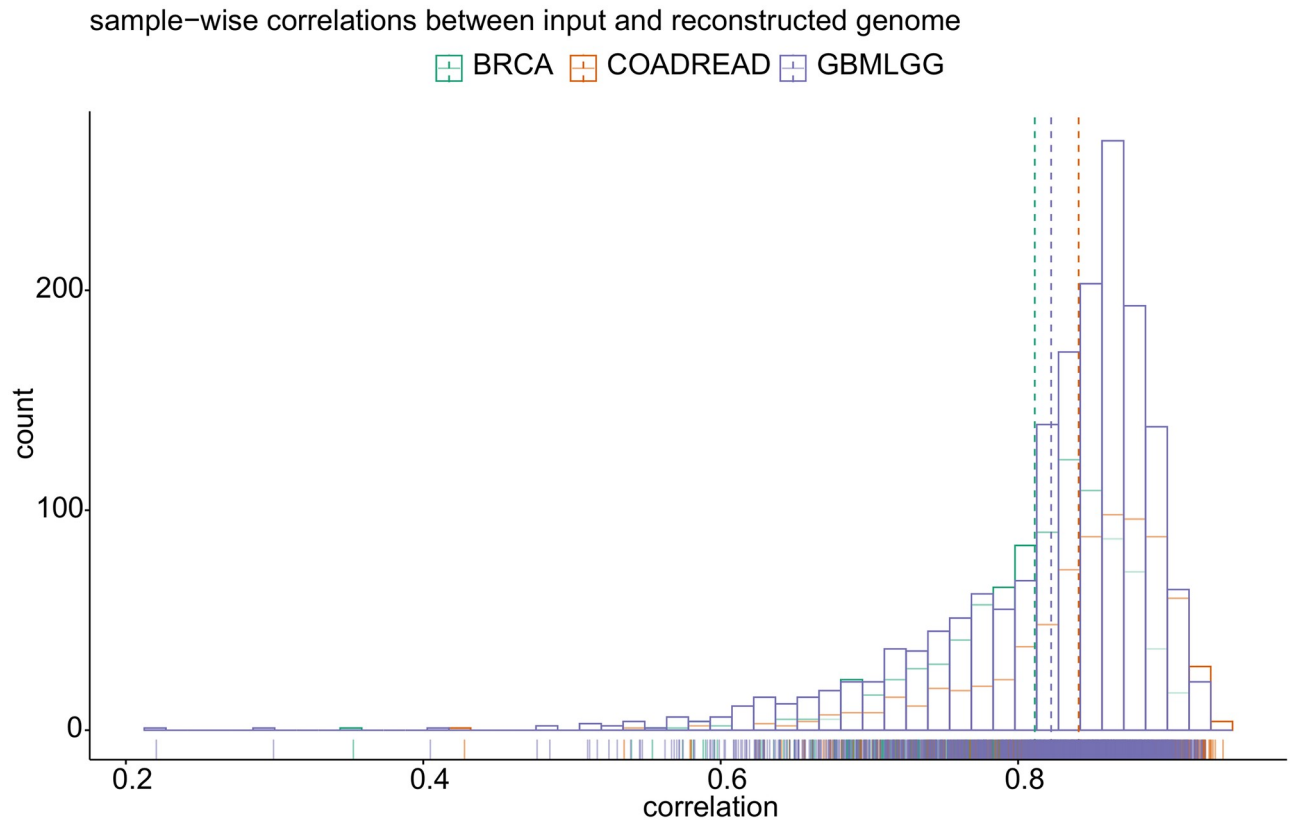


Fig 3. The generative performance of MFmap. The histogram shows sample-wise correlation coefficients between input features (DNA and RNA views) and reconstructed features output by the MFmap decoder.

<https://doi.org/10.1371/journal.pone.0261183.g003>

were not used to train MFmap. This indicates that MFmap is able to extract DNA and RNA patterns reflecting features originally derived from different methylation status.

In addition, we tested how well the MFmap autoencoder part reconstructs the molecular features x . To this end, we first sampled a latent representations from the encoder $q(z|x)$ for a given input x from the real data. Then, we correlated these original molecular features with the output sampled from the decoder distribution $p(x|z)$. The histogram of Pearson correlation coefficients in Fig 3 shows a high input-output correlation for most molecular features for three exemplary cancer types: breast invasive carcinoma (BRCA), colorectal adenocarcinoma (COADREAD) and glioblastoma multiforme and lower grade glioma (GBMLGG). Taken together, MFmap can combine very good classification accuracy with good generative performance.

Future applications of MFmap will include the analysis of query samples input to a reference model trained on a large data set. To check how well MFmap can perform in such a setting, we checked various measures for the quality of integrating these data from different sources [33–35]. Since this is not the focus of this paper, we have relegated the very promising results to the Supporting Information (see S2 File).

Selecting the optimal cell line for a given tumour

The heatmaps in Fig 4 represent pairwise cell line by tumour dissimilarity matrices for three exemplary cancer types BRCA, COADREAD and GBMLGG. In addition, the subtypes of bulk tumours annotated from [32, 36, 37] and the subtypes of cell lines predicted by the MFmap

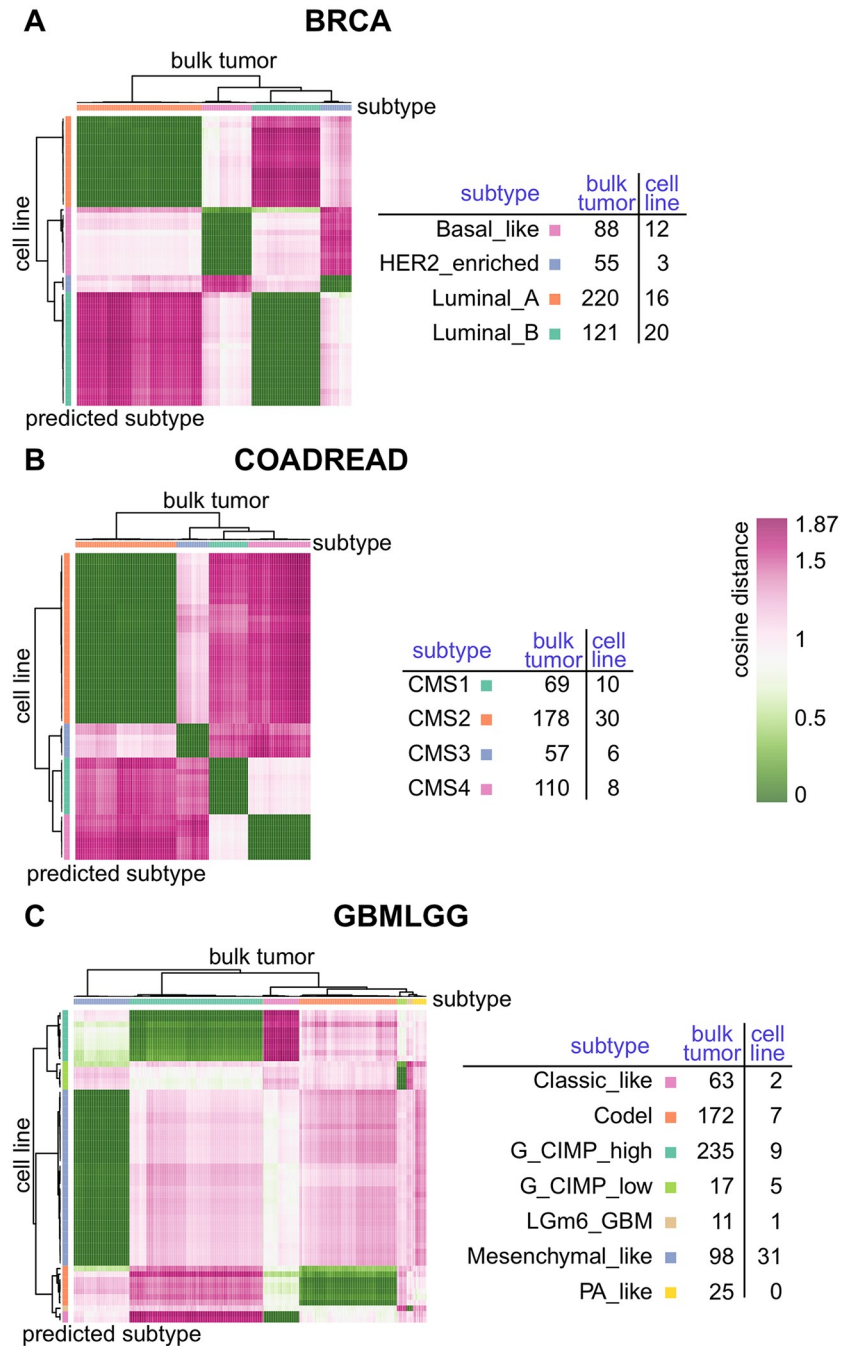


Fig 4. Pairwise dissimilarity between CCLE cell lines and TCGA bulk tumours. The colour coding in the heatmaps indicates the pairwise dissimilarity which was obtained from the latent representations of cell lines and tumours for the three exemplary cancer types (A) breast invasive carcinoma (BRCA), (B) colorectal adenocarcinoma (COADREAD) and (C) glioblastoma multiforme and lower grade glioma (GBMLGG). Tumours (columns) and cell lines (rows) were clustered according to the dissimilarity score, which ranges from 0 (very similar) to 2 (very dissimilar). The subtype classification of each cell line was predicted from the classification layer of the MFmap neural network. The tables display the sample size for the different subtypes or predicted subtypes.

<https://doi.org/10.1371/journal.pone.0261183.g004>

classifier are displayed. For a better visualisation, cell lines and tumours are clustered based on their pairwise cosine dissimilarity scores. The similarity of a cell line c to a tumour t is defined as the cosine of the angle between their latent representations z_c and z_t . Accordingly, the dissimilarity between c and t is defined as $d(c, t) = 1 - \frac{z_c \cdot z_t}{\|z_c\| \|z_t\|}$. A dissimilarity of $d(c, t) = 0$ indicates perfect alignment between the latent representations of the cell line and the tumour, whereas a dissimilarity $d(c, t) = 1$ indicates orthogonal latent representations. The highest dissimilarity of $d(c, t) = 2$ would be achieved for antipodal latent vectors. Based on this dissimilarity matrix, researchers can select the best cell lines for a given tumour or a given tumour subtype. And, vice versa, the relevance of promising experimental results observed *in vitro* can be checked by selecting a subset of tumours most likely resembling the cell line characteristics. The pairwise dissimilarity matrices between TCGA bulk tumours and CCLE cell lines and cell line subtype predictions for all tumour types listed in Table 1 are provided on our website (http://h2926513.stratoserver.net:3838/MFmap_shiny/).

These results also indicate, for which subtypes suitable cell line models exist and for which subtypes cell lines should be prioritised for future *in vitro* model development [21]. Each BRCA subtype is represented by at least three cell lines (Fig 4(A)) and the heatmap shows that these cell lines are very similar to the corresponding tumours of the same subtype. However, only three cell lines represent the HER2-enriched subtype. The four subtypes of COADREAD tumours are also well represented by at least six highly similar cell lines in CCLE (Fig 4(B)).

For GBMLGG, the Mesenchymal-like tumour subtype is represented by 31 cell lines with high similarity scores. Many TCGA tumour samples have the molecular subtype Codel and G-CIMP-high, but they are only represented by seven and nine cell lines, respectively. Only two cell lines were classified as Classic-like and a single cell line has the predicted subtype LGm6-GBM. The PA-like tumour subtype is not represented by any cell line.

Predicting drug sensitivity in cancer patient sub-cohorts using MFmap and *in vitro* drug screens

Predicting patient therapeutic response is one important goal of subtype stratification. To explore the translational potential of the subtypes predicted by MFmap we estimated the association between predicted subtypes and drug sensitivity of all compounds available in the CTRP dataset [18]. For each cancer type listed in Table 1 and each compound, we compared the drug sensitivity among different cell line subtypes predicted by the MFmap classifier. Drug sensitivity is quantified in CTRP by the area under the dose response curve (AUC). We used an ANOVA to test for differences in the mean AUC among the predicted subtypes. At a false discovery rate (FDR) cutoff of 25%, we found 18, six and 16 compounds in BRCA, GBMLGG and UCEC to show significant subtype specificity, respectively. For the other seven cancer types in Table 1, there are no significant AUC differences across the different subtypes. Note that the sample size per subtype is very small, which might explain why statistically significant results can only be obtained for three cancer types.

For BRCA, the compound with the strongest association between subtype and drug sensitivity is Lapatinib (ANOVA p-value = 2.95e-05). Lapatinib is a tyrosine kinase inhibitor used in combination therapy for HER2-positive breast cancer [38]. Our results suggest that cell lines of molecular subtype HER2-enriched are more sensitive to Lapatinib treatment (Fig 5(A)) in comparison to other three subtypes. Although there are only three cell lines representing the HER2-enriched subtype, this finding is in line with the known inhibitive mechanism of Lapatinib on the HER2/neu and epidermal growth factor receptor (EGFR) pathways. This result highlights the potential of MFmap as a tool for translating *in vitro* drug screening results to patient sub-cohorts. Our analysis also suggests that larger sample sizes and a better coverage

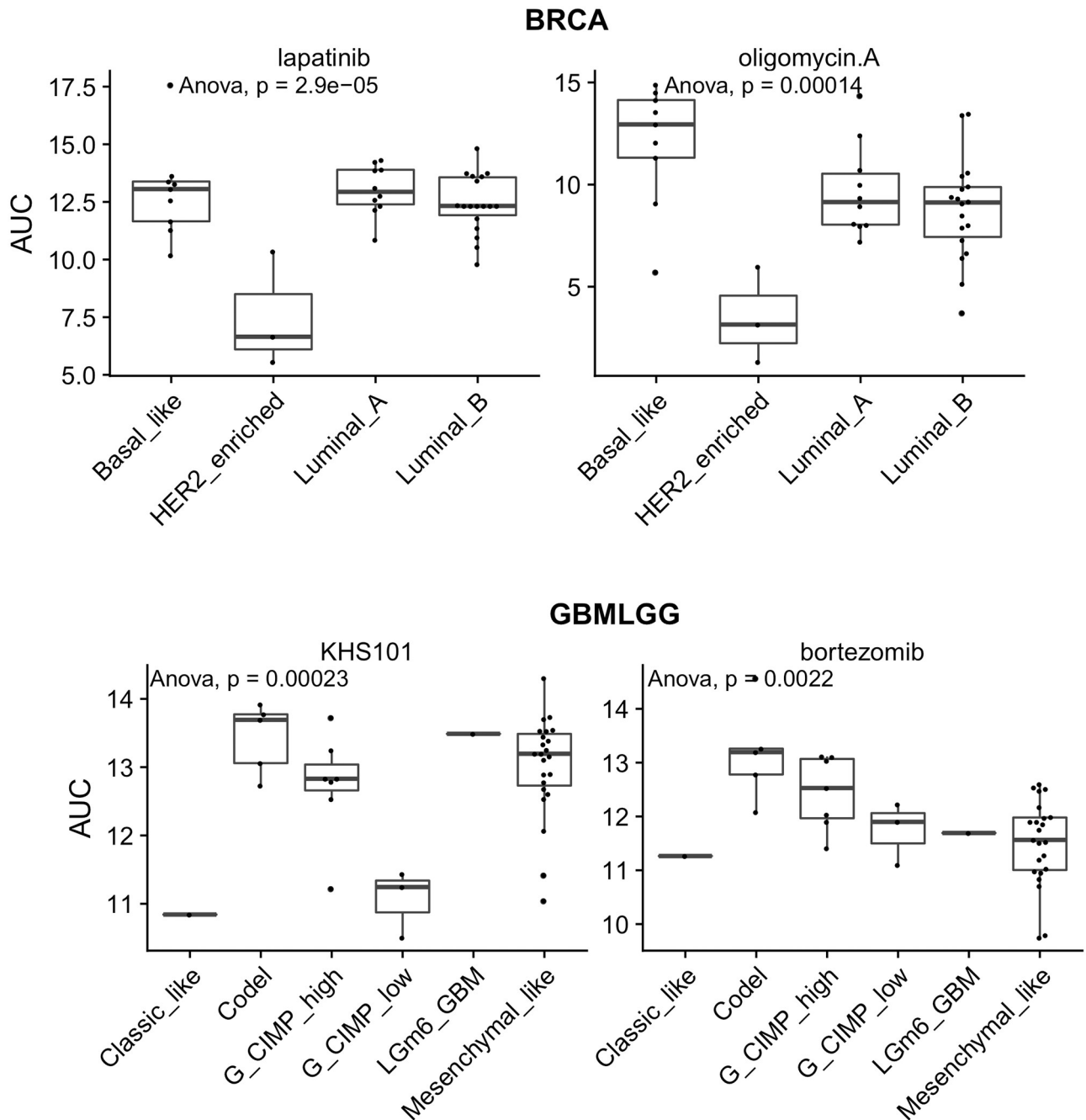


Fig 5. Cancer subtype specific drug sensitivity of CCLL cell lines. The subtypes of breast invasive carcinoma (BRCA) cell lines respond differentially to the compounds Lapatinib and Oligomycin A. Treatment response to the compounds KHS101 and Bortezomib in glioblastoma multiforme and lower grade glioma (GBMLGG) cell lines is subtype specific. The drug sensitivity is summarised by the area under the dose response curve (AUC) and p-values refer to an ANOVA of the AUC differences among different subtypes.

<https://doi.org/10.1371/journal.pone.0261183.g005>

of underrepresented subtypes are essential to increase the statistical power for detecting subtype specificity from cell line drug screens.

Another drug with significant variations of the AUC values across the different BRCA subtypes is Oligomycin A (ANOVA p-value = $1.39e-4$), a compound targeting oxidative

phosphorylation via an inhibition of the ATP synthase. The potential of Oligomycin A as a therapeutic compound to prevent metastatic spread in breast cancer has recently been highlighted [39]. The results in Fig 5(B) suggest that treatment with Oligomycin A might be most efficient for the HER2-enriched and Luminal A or Luminal B subtypes.

The drug sensitivities of KHS101 and Bortezomib are significantly associated with GBMLGG subtypes (KHS101: ANOVA p-value = 2.3e-04; Bortezomib: ANOVA p-value = 2.3e-04). The synthetic small molecule KHS101 was shown to promote tumour cell death in diverse glioblastoma multiforme cell line models [40]. Our analysis suggests that the G-CIMP-low subtype is more sensitive to KSH101 treatment (Fig 5(C)) compared to the other six GBMLGG subtypes. G-CIMP-low is an IDH mutant glioma subtype with poor clinical outcome in recurrent glioma [32].

Bortezomib targets the ubiquitin-proteasome pathway and is used for the treatment of multiple myeloma, but has also been discussed as treatment for glioma [41]. Our results in Fig 5(D) show that the Codel and G-CIMP-high subtypes have larger AUCs. The results for LGm6-GBM and Classic-like are not conclusive because there are not enough cell lines representing these subtypes.

Biological characterisation of latent representations learnt by MFmap

The pattern of MFmap learnt latent representations z can be used as a signature for cancer subtypes. For example, in BRCA, the basal-like subtype is characterised by a pattern of low values of components z_1 and z_4 and high values of z_2 and z_3 (Fig 6(A)). HER2-enriched tumours are characterised by high values of z_1 and z_3 and z_4 . Luminal A and B subtypes can be distinguished by z_4 . Similarly, cancer subtypes in COADREAD and GBMLGG are highly associated with their latent representations learnt by MFmap (Fig 6(B) and 6(C)).

To further investigate the biological meaning of the latent representations we analysed the association between z and pathway activities in TCGA reference datasets. We used single sample gene set enrichment analysis (ssGSEA) [42] to assess sample-wise pathway activities. The pathway signatures were compiled from several sources including 10 curated oncogenic signaling pathways [43], 19 curated specific DNA damage repair (DDR) pathways [44], 14 expert-curated specific DDR processes and DDR associated processes [45]. This collection was combined with MsigDB (v7.0) [46] chemical and genetic perturbations (CGP) and canonical pathways (CP) collections (MsigDB C2 collection) and MsigDB (v7.0) hallmark gene sets (MsigDB H collection). The degree of associations was quantified by the information coefficient and the Pearson correlation coefficient and the statistical significance was assessed by permutation tests. To tackle class imbalance in the different subtypes, we applied SMOTE upsampling [47].

We used COADREAD as a proof of concept, because it has four well characterised molecular subtypes CMS1-CMS4 [37]. The CMS1 subtype is characterised by micro-satellite instability (MSI), whereas CMS4 tumours are micro-satellite stable. The CMS4 subtype is also distinguished from CMS1 by epithelial mesenchymal transformation (EMT) characteristics, accompanied by prominent stromal invasion and angiogenesis. These mutually exclusive characteristics are clearly reflected in the magnitudes of the latent representation components. The top gene sets associated with component z_2 are “WATANABE COLON CANCER MSI VS MSS UP” and “KOINUMA COLON CANCER MSI UP”, whereas z_4 is associated with the activity of gene sets annotated as “HALLMARK ANGIOGENESIS” and “HALLMARK EPI-THELIAL MESENCHYMAL TRANSITION”. Clearly, high values of z_2 are a characteristics of the CMS1 subtype, whereas high values of z_4 are a distinctive feature of CMS4 tumours. This example illustrates that a meaningful way to guide biological interpretation of the latent representations is to associate them to single sample pathway activity.

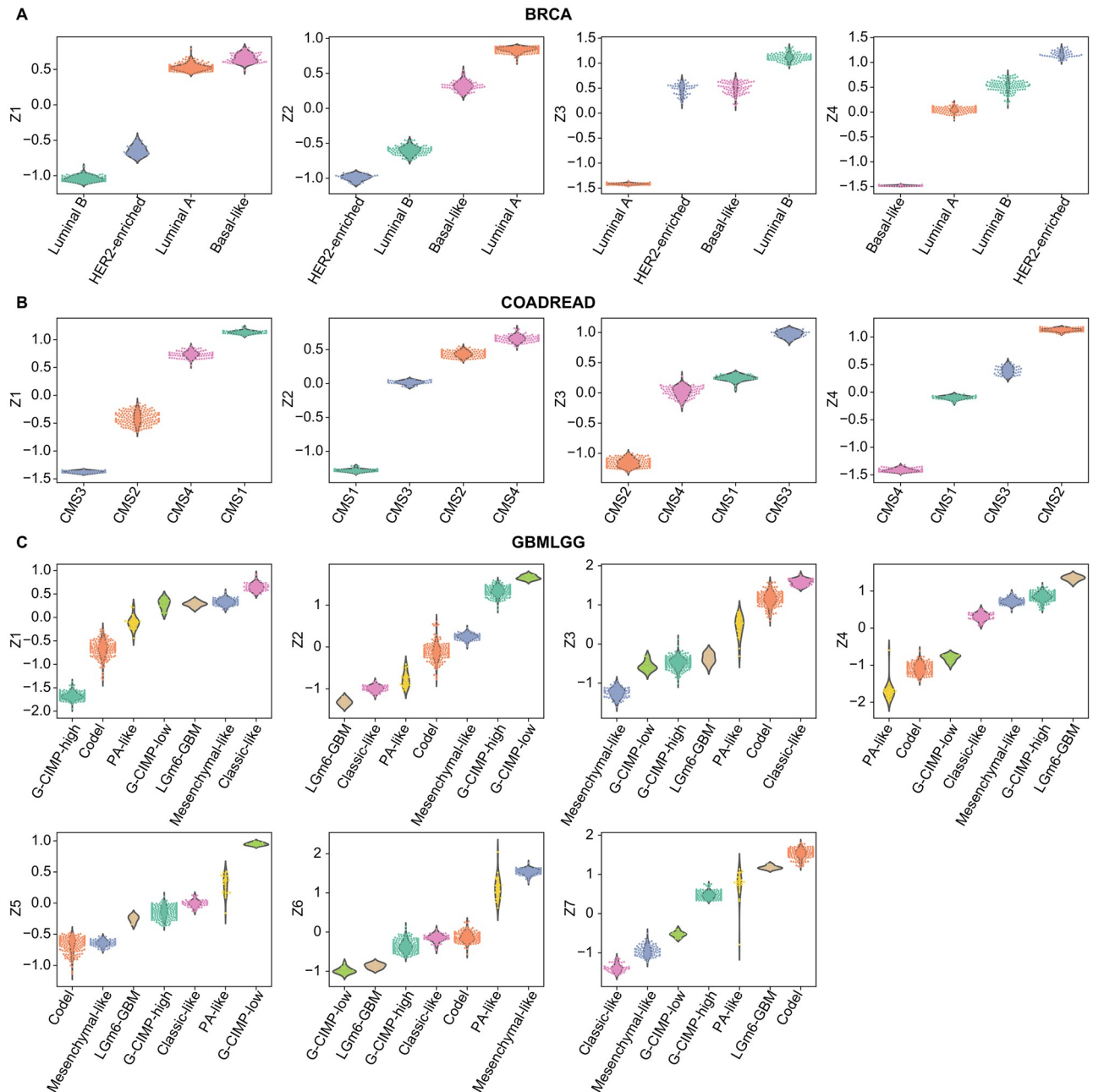


Fig 6. Association of MFmap latent representations and cancer subtypes. The dimension of the latent representation h is set to the number of cancer subtypes. The boxplots display latent representations of different subtypes of TCGA samples in the three exemplary cancer types (A) breast invasive carcinoma (BRCA), (B) colorectal adenocarcinoma (COADREAD) and (C) glioblastoma multiforme and lower grade glioma (GBMLGG). Cancer subtypes are colour encoded and sorted by their median latent representations.

<https://doi.org/10.1371/journal.pone.0261183.g006>

The same method was applied to annotate latent representations of GBMLGG (Fig 7(A)), which has seven subtypes [32]. The Mesenchymal-like and PA-like are stratified by gene expression profiles and the G-CIMP-high, G-CIMP-low and LGM6-GBM are methylation based. The Codel subtype describes IDH-mutant samples harbouring a co-deletion of chromosome arm 1p and 19q. Many pathways associated with latent representation z_1 are related to the neurotransmitter release cycle, which is also a characteristics of the Verhaak proneuronal

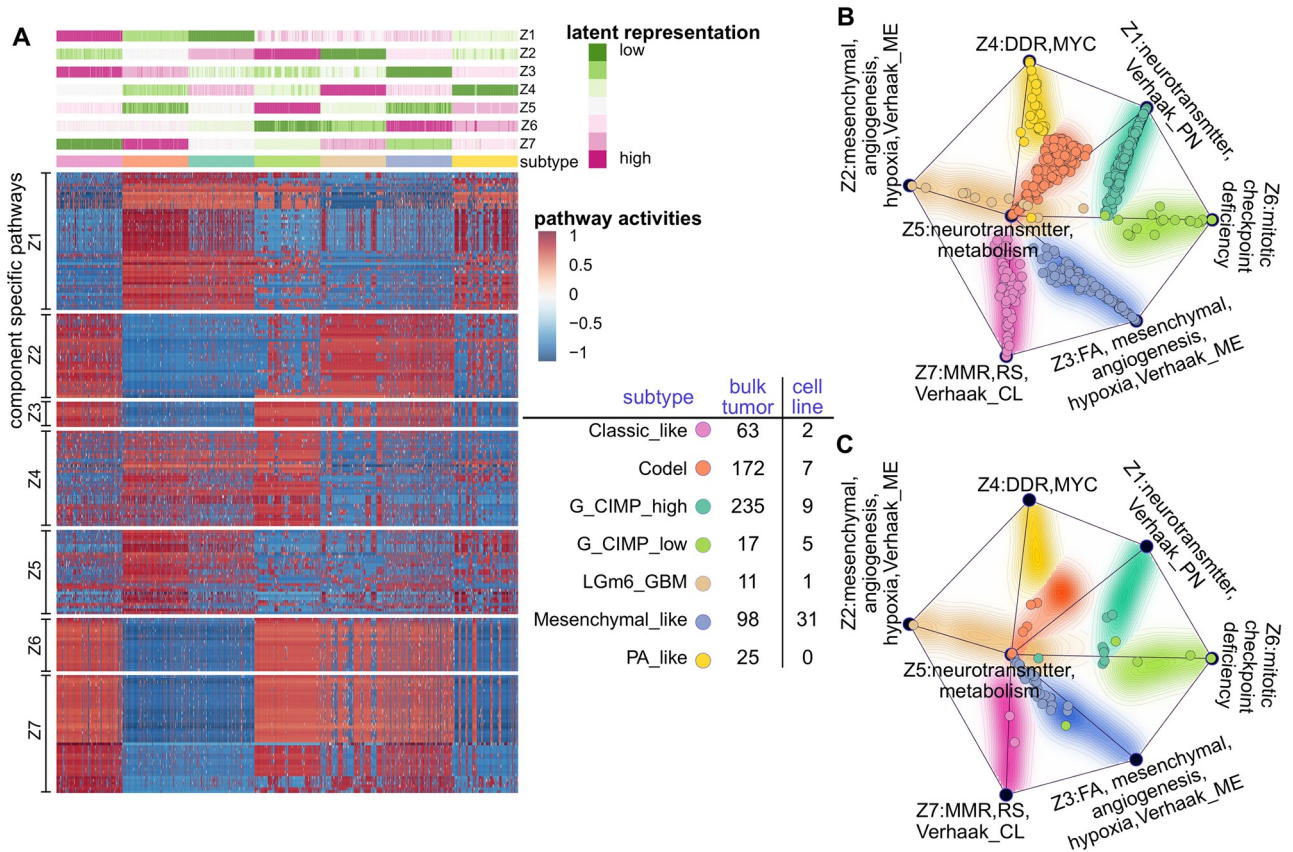


Fig 7. Characterising the MFmap learnt latent representations in glioblastoma multiforme and lower grade glioma (GBMLGG). (A) The top heatmap shows the latent representation z of TCGA tumour samples (columns). The tumour samples are ordered based on a hierarchical clustering of z and their subtypes are colour encoded. The heatmap at the bottom displays sample-wise pathway activities that are significantly associated with the latent representations z_1, \dots, z_7 . Pathway activities were computed using the ssGSEA algorithm [42]. For better visualisation, we upsampled the input data of MFmap and ssGSEA to get a balanced sample size in each subtype. (B) The MFmap reference map is formed by projecting the latent representations z of bulk tumours into two dimensions using multidimensional scaling. It consists of seven dominant components represented by black nodes. The length of their connections is given by the Euclidean distance of the dominant components in the latent space. The annotation of the seven dominant nodes is based on the correlation between z and pathway activity scores (see A). The background colour encodes sample subtypes, and the background contour encodes sample density. Individual bulk tumours are displayed as dots on the MFmap reference map. (C) Cell line samples are projected to the MFmap reference map. In both (B) and (C), the subtype of bulk tumours and predicted subtype of cell lines are colour coded. Subtype specific sample size for bulk tumours and cell lines is reported in the legend table.

<https://doi.org/10.1371/journal.pone.0261183.g007>

subtype [48]. Pathways correlated to latent representation z_2 are related to the mesenchymal cell type, hypoxia and angiogenesis, which characterises the Verhaak mesenchymal subtype. The activity of the Fanconi Anemia (FA) DNA repair pathway is highly correlated with latent representation z_3 . DNA damage response deficiency and amplified oncogenic MYC signalling characterises tumours with large values of latent representation z_4 . Latent representation z_5 is related to the neurotransmitter release cycle and dysfunctional metabolism; latent representation z_6 to mitotic checkpoint deficiency. Many pathways associated with latent representation z_7 are involved in mismatch repair deficiency, replication stress and cell cycle dysregulation and also related to the classical subtype in the earlier classification of Verhaak [48].

Individual samples and their relationships can be displayed in the MFmap reference map (Fig 7(B)), a visualisation tool adapted from OncoGPS [31]. Here, the seven corners of the map correspond to the respective latent representations z_1, \dots, z_7 in GBMLGG. The corner locations are determined by multidimensional scaling on the latent representations of bulk

tumours. Individual bulk tumour samples are displayed as dots in the regions between the corner points with locations determined by a weighted vector sum of the seven corner locations (see [S1 File](#) for details). The subtype of each tumour sample is indicated by colours. The density of the tumour samples of a given subtype is depicted by the contour lines and the corresponding colour shading. [Fig 7\(B\)](#) shows that samples of the same subtype clustered together and the inter-cluster distance is large. Projecting cell lines to the MFmap reference map ([Fig 7\(C\)](#)) helps to visualise the relationship between their predicted subtypes and their latent representations.

Modelling cellular state transformations using latent space arithmetics

Cancerous neoplasms undergo various biochemical changes during cancer evolution and in response to selective pressure. One example is the transition from a proneural to a mesenchymal phenotype in glioblastoma, which is characterised by acquired therapeutic resistance and more aggressive potential [49]. In the DNA methylation based subtype classification of [32], the G-CIMP-high methylation phenotype tends to have the proneural molecular subtype [48] (see [Fig 7\(B\)](#)). Given that the latent representations learnt by MFmap clearly distinguish these different subtypes, we asked, whether the generative nature of the semi-supervised VAE can also be exploited to study such cancer subtype transformations.

To this end, we used the latent representations of the G-CIMP-high tumours and the Mesenchymal-like tumours (see [Fig 7\(B\)](#)) and computed the centroid vectors $\bar{z}_{\text{G-CIMP-high}}$ and $\bar{z}_{\text{Mesenchymal-like}}$ for the corresponding tumour samples. The difference $\delta = \bar{z}_{\text{Mesenchymal-like}} - \bar{z}_{\text{G-CIMP-high}}$ was used as a latent perturbation vector. By adding δ to the latent representation of each G-CIMP-high tumour ([Fig 8\(A\)](#)) we obtained the latent representation of *in silico* samples ([Fig 8\(B\)](#)), which are located in the “Mesenchymal-like region” of the reference map. We used these latent representation vectors of the *in silico* samples as input to the decoder of the MFmap network. We then checked, whether key molecular features of real Mesenchymal-like samples are reflected by these generated samples. Based on the available biological knowledge, we focussed on the most prominent onco-markers of the G-CIMP-high subtype: mutation status of the alpha thalassemia/mental retardation syndrome X-linked (ATRX), isocitrate dehydrogenase (IDH) and TP53 genes. The original G-CIMP-high tumours show a high propensity towards mutations in these genes, indicated by relatively higher network smoothed mutation scores ([Fig 8\(C\)](#)), although not all samples are necessarily harbouring these mutations. In contrast, the predicted mutation scores for the perturbed *in silico* samples in [Fig 8\(B\)](#) are much lower, indicating a lower propensity to IDH1, ATRX or TP53 mutations. This is in agreement with the observed tendency of Mesenchymal-like tumours for these mutations [49]. This example not only highlights the good generative performance of MFmap but also hints at potential applications on integrative analysis of cancer evolution dynamics.

Discussion

Limited success in translating *in vitro* therapeutic markers to clinical applications highlights that not all cell lines are good models for a given cancer subtype. Selecting the most appropriate cell line for a given tumour or a set of tumours is crucial for understanding cancer biology and developing new anti-cancer treatments. Here, we provide a computational framework and a resource for cancer researchers to select the best cell lines for a TCGA tumour or a cancer subtype from ten different cancer types (http://h2926513.stratoserver.net:3838/MFmap_shiny/). The quantitative similarity score enables researchers to judge, whether a given tumour or a subtype of tumours is well represented by a cell line.

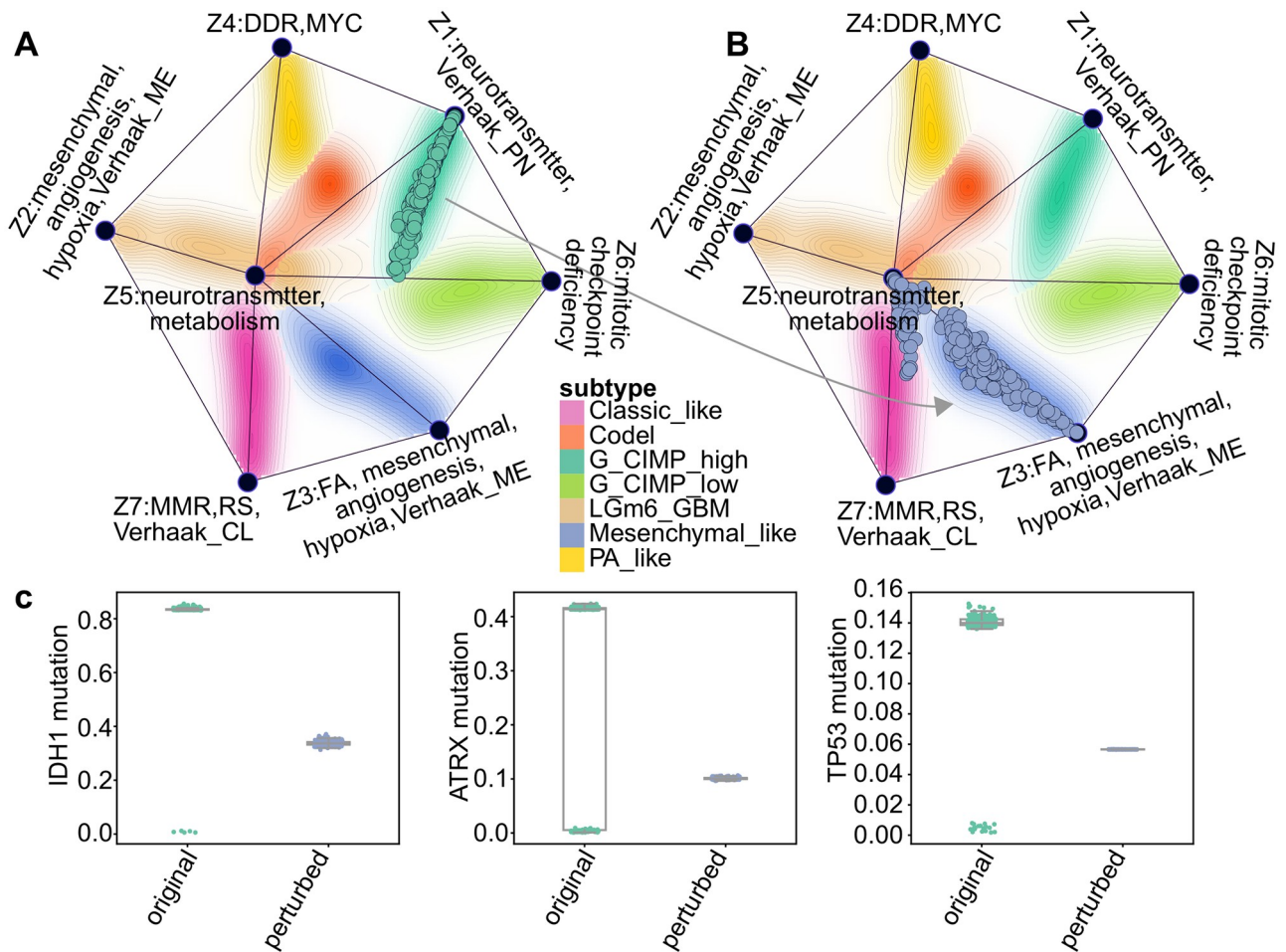


Fig 8. In-silico perturbation analysis of cellular state changes during disease transformation from the G-CIMP-high to the Mesenchymal-like subtype in glioblastoma multiforme and lower grade glioma (GBMLGG). (A) The G-CIMP-high tumours from TCGA are projected to the MFmap reference map. (B) By perturbing the latent representation vectors of these G-CIMP-high tumours we generate artificial tumour samples located in the Mesenchymal-like region of the MFmap reference map (compare Fig 7(B)). (C) Boxplots of the sample mutation status (network smoothed mutation scores) of marker genes IDH1, ATRX1 and TP53 before and after perturbation.

<https://doi.org/10.1371/journal.pone.0261183.g008>

The assignment of cancer subtype labels to cell lines enables cell biologists to optimise experimental planning and to focus their research on clinically relevant model systems. We found that our semi-supervised MFmap model can classify tumours with a very high accuracy. Further analysis of drug sensitivity profiles supports that the subtype prediction for cell lines is biologically meaningful. Our analysis shows that HER2-enriched cell lines are most sensitive to Lapatinib, in agreement with prior knowledge about drug efficiency of this compound. As an example for the translation of *in vitro* pharmacogenomic data, we predict that the G-CIMP-low subtype is more sensitive to the new synthetic compound KHS101 compared to other GBMLGG subtypes.

Our finding that only BRCA, GBMLGG and UCEC show significant subtype specific drug sensitivity variation merits further investigation. One important reason is the small number of cell lines representing some cancer subtypes, which prevents us from finding statistically significant variations of drug sensitivity across the different subtypes. This highlights the need to prioritise cell line development for underrepresented disease variants [21]. However, it can not be ruled out that for some cancers the known subtype classifications are not predictive of drug

sensitivity. This suggests that clinically relevant subtype stratification should take into account drug sensitivity.

By embedding the original gene expression space, somatic mutation space and copy number space of bulk tumours and cell lines into a lower dimensional latent space, MFmap extracts latent features that are strongly associated with cancer subtypes. For COADREAD and GLMBGG, we have illustrated that the abstract latent representations can be annotated biologically using their associations with pathway activities. This makes the latent representations interpretable and allows to study the molecular and clinical heterogeneity of this disease. In principle, MFmap can be complemented by other modalities such as methylation or proteomics data. However, for our purpose we found that gene expression and DNA features in combination with the prior knowledge about tumour subtypes contains sufficient information.

Our proof of principle analysis of the transformation between two different tumour subtypes presents a new approach for studying tumour evolutionary processes in a more integrative way [50]. The small sample size of some multi-region sequencing or single-cell sequencing studies limits the ability to infer robust evolutionary patterns. By projecting these data to the MFmap reference map obtained from training on large sets of bulk tumour data one could deduce useful phenotypic information for individual patients. We believe that this can leverage information gathered in large cancer genomic studies like TCGA to guide personalised clinical decision making.

The MFmap is based on a new semi-supervised neural network architecture combining a basic VAE with an additional classifier. Such semi-supervised learning tasks are very common in the biomedical research field, because it is often easier to acquire a large number of measurements than to obtain the corresponding labels. Based on the good predictive and generative performance of MFmap together with the evidence provided here, that MFmap can learn biologically and clinically meaningful information, we are convinced that the MFmap model can be adapted to other semi-supervised tasks in oncology and beyond.

Supporting information

S1 File. Extended method details.

(PDF)

S2 File. Further evaluation of the MFmap performance.

(PDF)

S1 Data.

(TXT)

Author Contributions

Conceptualization: Xiaoxiao Zhang, Maik Kschischo.

Data curation: Xiaoxiao Zhang.

Formal analysis: Xiaoxiao Zhang, Maik Kschischo.

Funding acquisition: Maik Kschischo.

Investigation: Xiaoxiao Zhang, Maik Kschischo.

Methodology: Xiaoxiao Zhang, Maik Kschischo.

Project administration: Maik Kschischo.

Resources: Maik Kschischo.

Software: Xiaoxiao Zhang, Maik Kschischo.

Supervision: Maik Kschischo.

Visualization: Xiaoxiao Zhang.

Writing – original draft: Xiaoxiao Zhang, Maik Kschischo.

Writing – review & editing: Maik Kschischo.

References

1. Sharma SV, Haber DA, Settleman J. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nature Reviews Cancer*. 2010; 10(4):241–253. <https://doi.org/10.1038/nrc2820> PMID: 20300105
2. Kim N, He N, Yoon S. Cell line modeling for systems medicine in cancers (Review). *International Journal of Oncology*. 2014; 44(2):371–376. <https://doi.org/10.3892/ijo.2013.2202> PMID: 24297677
3. Goodspeed A, Heiser LM, Gray JW, Costello JC. Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. *Molecular Cancer Research*. 2016; 14(1):3–13. <https://doi.org/10.1158/1541-7786.MCR-15-0189> PMID: 26248648
4. Kaur G, Dufour JM. Cell lines: Valuable tools or useless artifacts. *Spermatogenesis*. 2012; 2(1):1–5. <https://doi.org/10.4161/spmg.19885> PMID: 22553484
5. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012; 483(7391):603–607. <https://doi.org/10.1038/nature11003>
6. Mouradov D, Sloggett C, Jorissen RN, Love CG, Li S, Burgess AW, et al. Colorectal Cancer Cell Lines Are Representative Models of the Main Molecular Subtypes of Primary Cancer. *Cancer Research*. 2014; 74(12):3238–3247. <https://doi.org/10.1158/0008-5472.CAN-14-0013> PMID: 24755471
7. Ince TA, Sousa AD, Jones MA, Harrell JC, Agoston ES, Krohn M, et al. Characterization of twenty-five ovarian tumour cell lines that phenocopy primary tumours. *Nature Communications*. 2015; 6(1):7419. <https://doi.org/10.1038/ncomms8419> PMID: 26080861
8. Cheng H, Yang X, Si H, Saleh AD, Xiao W, Coupar J, et al. Genomic and Transcriptomic Characterization Links Cell Lines with Aggressive Head and Neck Cancers. *Cell Reports*. 2018; 25(5):1332–1345. e5. <https://doi.org/10.1016/j.celrep.2018.10.007> PMID: 30380422
9. Domcke S, Sinha R, Levine DA, Sander C, Schultz N. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature Communications*. 2013; 4(1):2126. <https://doi.org/10.1038/ncomms3126> PMID: 23839242
10. Liu K, Newbury PA, Glicksberg BS, Zeng WZD, Paithankar S, Andrechek ER, et al. Evaluating cell lines as models for metastatic breast cancer through integrative analysis of genomic data. *Nature Communications*. 2019; 10(1):2138. <https://doi.org/10.1038/s41467-019-10148-6> PMID: 31092827
11. Sinha R, Winer AG, Chevinsky M, Jakubowski C, Chen YB, Dong Y, et al. Analysis of renal cancer cell lines from two major resources enables genomics-guided cell line selection. *Nature Communications*. 2017; 8(1):15165. <https://doi.org/10.1038/ncomms15165> PMID: 28489074
12. Yu K, Chen B, Aran D, Charalel J, Yau C, Wolf DM, et al. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nature Communications*. 2019; 10(1):3574. <https://doi.org/10.1038/s41467-019-11415-2> PMID: 31395879
13. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*. 2006; 6(10):813–823. <https://doi.org/10.1038/nrc1951> PMID: 16990858
14. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*. 2019; 569(7757):503–508. <https://doi.org/10.1038/s41586-019-1186-3> PMID: 31068700
15. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*. 2016; 166(3):740–754. <https://doi.org/10.1016/j.cell.2016.06.017> PMID: 27397505
16. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*. 2006; 313(5795):1929. <https://doi.org/10.1126/science.1132939> PMID: 17008526

17. Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules. *Cell*. 2013; 154(5):1151–1161. <https://doi.org/10.1016/j.cell.2013.08.003> PMID: 23993102
18. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discovery*. 2015; 5(11):1210. <https://doi.org/10.1158/2159-8290.CD-15-0235> PMID: 26482930
19. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*. 2013; 45(10):1113–1120. <https://doi.org/10.1038/ng.2764>
20. Hudson (Chairperson) TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. *Nature*. 2010; 464(7291):993–998. <https://doi.org/10.1038/nature08987>
21. Najgebauer H, Yang M, Francies HE, Pacini C, Stronach EA, Garnett MJ, et al. CELLector: Genomics-Guided Selection of Cancer In Vitro Models. *Cell Systems*. 2020; 10(5):424–432.e6. <https://doi.org/10.1016/j.cels.2020.04.007> PMID: 32437684
22. Salvadores M, Fuster-Tormo F, Supek F. Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. *Science Advances*. 2020; 6(27):eaba1862. <https://doi.org/10.1126/sciadv.aba1862> PMID: 32937430
23. Webber JT, Kaushik S, Bandyopadhyay S. Integration of Tumor Genomic Data with Cell Lines Using Multi-dimensional Network Modules Improves Cancer Pharmacogenomics. *Cell Systems*. 2018; 7(5):526–536.e6. <https://doi.org/10.1016/j.cels.2018.10.001> PMID: 30414925
24. Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv:13126114 [Preprint]. 2013; Available from: <https://arxiv.org/pdf/1312.6114.pdf>.
25. Huang JK, Jia T, Carlin DE, Ideker T. pyNBS: a Python implementation for network-based stratification of tumor mutations. *Bioinformatics*. 2018; 34(16):2859–2861. <https://doi.org/10.1093/bioinformatics/bty186> PMID: 29608663
26. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nature Methods*. 2013; 10(11):1108–1115. <https://doi.org/10.1038/nmeth.2651> PMID: 24037242
27. Kingma DP, Rezende DJ, Mohamed S, Welling M. Semi-Supervised Learning with Deep Generative Models. arXiv:14065298v2[Preprint]. 2014; Available from: <https://arxiv.org/pdf/1406.5298.pdf>.
28. Feng H, Kong K, Chen M, Zhang T, Zhu M, Chen W. SHOT-VAE: Semi-supervised Deep Generative Models With Label-aware ELBO Approximations. arXiv:201110684[Preprint]. 2020; abs/2011.10684. Available from: <https://arxiv.org/pdf/2011.10684.pdf>.
29. Dai Z, Yang Z, Yang F, Cohen WW, Salakhutdinov R. Good Semi-supervised Learning that Requires a Bad GAN. arXiv:170509783[Preprint]. 2017; Available from: <https://arxiv.org/pdf/1705.09783.pdf>.
30. Grandvalet Y, Bengio Y. Semi-Supervised Learning by Entropy Minimization. In: Proceedings of the 17th International Conference on Neural Information Processing Systems. NIPS'04. Cambridge, MA, USA: MIT Press; 2004. p. 529–536.
31. Kim JW, Abudayyeh OO, Yeerna H, Yeang CH, Stewart M, Jenkins RW, et al. Decomposing Oncogenic Transcriptional Signatures to Generate Maps of Divergent Cellular States. *Cell Systems*. 2017; 5(2):105–118.e9. <https://doi.org/10.1016/j.cels.2017.08.002> PMID: 28837809
32. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*. 2016; 164(3):550–563. <https://doi.org/10.1016/j.cell.2015.12.028> PMID: 26824661
33. Lotfollahi M, Naghipourfar M, Luecken MD, Khajavi M, Büttner M, Wagenstetter M, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*. 2021. <https://doi.org/10.1038/s41587-021-01001-7> PMID: 34462589
34. Stein-O'Brien GL, Clark BS, Sherman T, Zibetti C, Hu Q, Sealfon R, et al. Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. *Cell Systems*. 2019; 8(5):395–411.e8. <https://doi.org/10.1016/j.cels.2019.04.004> PMID: 31121116
35. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck I William M, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019; 177(7):1888–1902.e21. <https://doi.org/10.1016/j.cell.2019.05.031> PMID: 31178118
36. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012; 486(7403):346–352. <https://doi.org/10.1038/nature10983> PMID: 22522925
37. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*. 2015; 21(11):1350–1356. <https://doi.org/10.1038/nm.3967> PMID: 26457759

38. Higa GM, Abraham J. Lapatinib in the treatment of breast cancer. *Expert Review of Anticancer Therapy*. 2007; 7(9):1183–1192. <https://doi.org/10.1586/14737140.7.9.1183> PMID: 17892419
39. Davis RT, Blake K, Ma D, Gabra MBI, Hernandez GA, Phung AT, et al. Transcriptional diversity and bioenergetic shift in human breast cancer metastasis revealed by single-cell RNA sequencing. *Nat Cell Biol*. 2020; 22(3):310–320. <https://doi.org/10.1038/s41556-020-0477-0> PMID: 32144411
40. Polson ES, Kuchler VB, Abbosh C, Ross EM, Mathew RK, Beard HA, et al. KHS101 disrupts energy metabolism in human glioblastoma cells and reduces tumor growth in mice. *Science Translational Medicine*. 2018; 10(454):eaar2718. <https://doi.org/10.1126/scitranslmed.aar2718> PMID: 30111643
41. Tang JH, Yang L, Chen JX, Li QR, Zhu LR, Xu QF, et al. Bortezomib inhibits growth and sensitizes glioma to temozolomide (TMZ) via down-regulating the FOXM1–Survivin axis. *Cancer Commun*. 2019; 39(1):81. <https://doi.org/10.1186/s40880-019-0424-2> PMID: 31796105
42. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*. 2013; 14(1):7. <https://doi.org/10.1186/1471-2105-14-7> PMID: 23323831
43. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*. 2018; 173(2):321–337.e10. <https://doi.org/10.1016/j.cell.2018.03.035> PMID: 29625050
44. Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, et al. Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Reports*. 2018; 23(1):239–254.e6. <https://doi.org/10.1016/j.celrep.2018.03.076> PMID: 29617664
45. Pearl LH, Schierz AC, Ward SE, Al-Lazikani B, Pearl FMG. Therapeutic opportunities within the DNA damage response. *Nature Reviews Cancer*. 2015; 15(3):166–180. <https://doi.org/10.1038/nrc3891> PMID: 25709118
46. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011; 27(12):1739–1740. <https://doi.org/10.1093/bioinformatics/btr260> PMID: 21546393
47. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002; 16:321–357. <https://doi.org/10.1613/jair.953>
48. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. 2010; 17(1):98–110. <https://doi.org/10.1016/j.ccr.2009.12.020> PMID: 20129251
49. Behnan J, Finocchiaro G, Hanna G. The landscape of the mesenchymal signature in brain tumours. *Brain*. 2019; 142(4):847–866. <https://doi.org/10.1093/brain/awz044> PMID: 30946477
50. Williams MJ, Sottoriva A, Graham TA. Measuring Clonal Evolution in Cancer with Genomics. *Annu Rev Genom Hum Genet*. 2019; 20(1):309–329. <https://doi.org/10.1146/annurev-genom-083117-021712> PMID: 31059289

Concluding remarks

Predictive modelling is helpful for better understanding of cancer [chromosomal instability \(CIN\)](#) complexity, by yielding quantitative insights on CIN mechanisms and consequences, and by suggesting new experiments. Conversely, such multi-scale complexity provides opportunities for integrative predictive model development. This thesis first reviews how complex the cancer CIN system is, thereby providing the rationale to address it using predictive modelling techniques. This thesis then answers one important question: What are the commonalities and differences between W-CIN and S-CIN? It thus serves as a complementary to the current CIN computational studies that only focus on W-CIN. This thesis then proposes a novel semi-supervised generative model matching cell lines to tumours, which has broader applications beyond cancer CIN. Here I shortly summarise the significance of the work presented in Chapters 3 and 4 and outlook future directions.

5.1 Conclusion

In Chapter 3, we implemented three types of karyotypic complexity scores as a proxy measurement of CIN: The [numerical complexity score \(NCS\)](#) measuring W-CIN, the [structural complexity score \(SCS\)](#) measuring S-CIN and the [weighted genome instability index \(WGII\)](#) measuring the overall CIN levels. We then performed extensive association analysis between these karyotypic complexity scores and molecular and clinical features for 33 cancer types from [Cancer Cell Line Encyclopedia \(CCLE\)](#) and [The Cancer Genome Atlas \(TCGA\)](#) datasets, intending to provide a better understanding of the commonalities and differences between W-CIN and S-CIN. The analysed features include genomic landscape, clinical outcomes, drug sensitivity, somatic mutations and somatic copy numbers. We found that [whole genome doubling \(WGD\)](#) is uniformly associated with high W-CIN but [homologous recombination deficiency \(HRD\)](#) is associated with high S-CIN in most cancer types. We report W-CIN and S-CIN have cancer type specific prognostic values and are hard to target using currently available drugs. We propose *CKS1B* as a potential S-CIN

target. We suggest a copy number based mechanism to promote *PI3K* signalling in high S-CIN cancer cells. Additionally, our model predicted *GINS1* as a W-CIN promoting gene, which has subsequently been experimentally validated [140]. In summary, our predictive model is valuable for understanding the mechanisms of CIN.

In Chapter 4, we developed a semi-supervised generative model named as **model fidelity map (MFmap)** matching cell lines to tumours. MFmap embeds high-dimensional gene expression, copy number and somatic mutation features into cancer subtype specific representations and predicts the subtypes of cell lines. The MFmap embedded representations are invariant between tumours and cell lines and can be used to quantify the pair-wise cell-line-tumour similarity. MFmap can achieve good generative and classification performance simultaneously, making it useful in several scenarios: selecting the best cell lines for a specific cancer subtype or an individual tumour; revealing novel cancer subtype specific features; translating *in vitro* drug screening to patients; modelling cancer disease transformation course.

5.2 Outlook

5.2.1 Interpretable deep learning to understand CIN mechanisms

Deep learning has been criticized as a class of “black box” approach, which hinders its wide application on cancer medicine where interpretability and trust are most essential. Appreciating that the cancer complex system is hierarchically organised (see Section 1.2), one possible interpretable solution is embedding the biological hierarchical structure to the feedforward neural network. **Gene ontology (GO)** represents such hierarchical structure as a **DAG** where nodes are GO terms and directed edges are parent-child relationships. A child term connects to one or more parent terms and represents more specific meanings than its parent terms in three forms of ontologies (biological process, molecular function, and cellular component). For CIN mechanisms analysis, a neural network mirroring GO substructures that are relevant to cell division (to avoid large neural network) receives genetic measurements as input and outputs the CIN status. I name it as CINgo (gene ontology guided neural network for CIN analysis). A conceptual design of CINgo is shown in Fig 5.1, each neuron of CINgo encodes a GO term and only receives the outputs from its children nodes. This is analogous to the CIN causal genetic alteration flow as previously described in Chapter 1. The CINgo will allow to nominate the potential CIN driver pathways based on the node relative predictability metrics such as relative local improvement in predictive power (RLIPP) score [141, 142], which can be further tested in the lab. We previously proposed a potential copy number dependent mechanism to promote S-CIN in *PI3K* pathway in Chapter 3, CINgo could improve our understanding

on this aspect by checking the high weight genes corresponding to the leaf nodes of the *PI3K* pathway ontology.

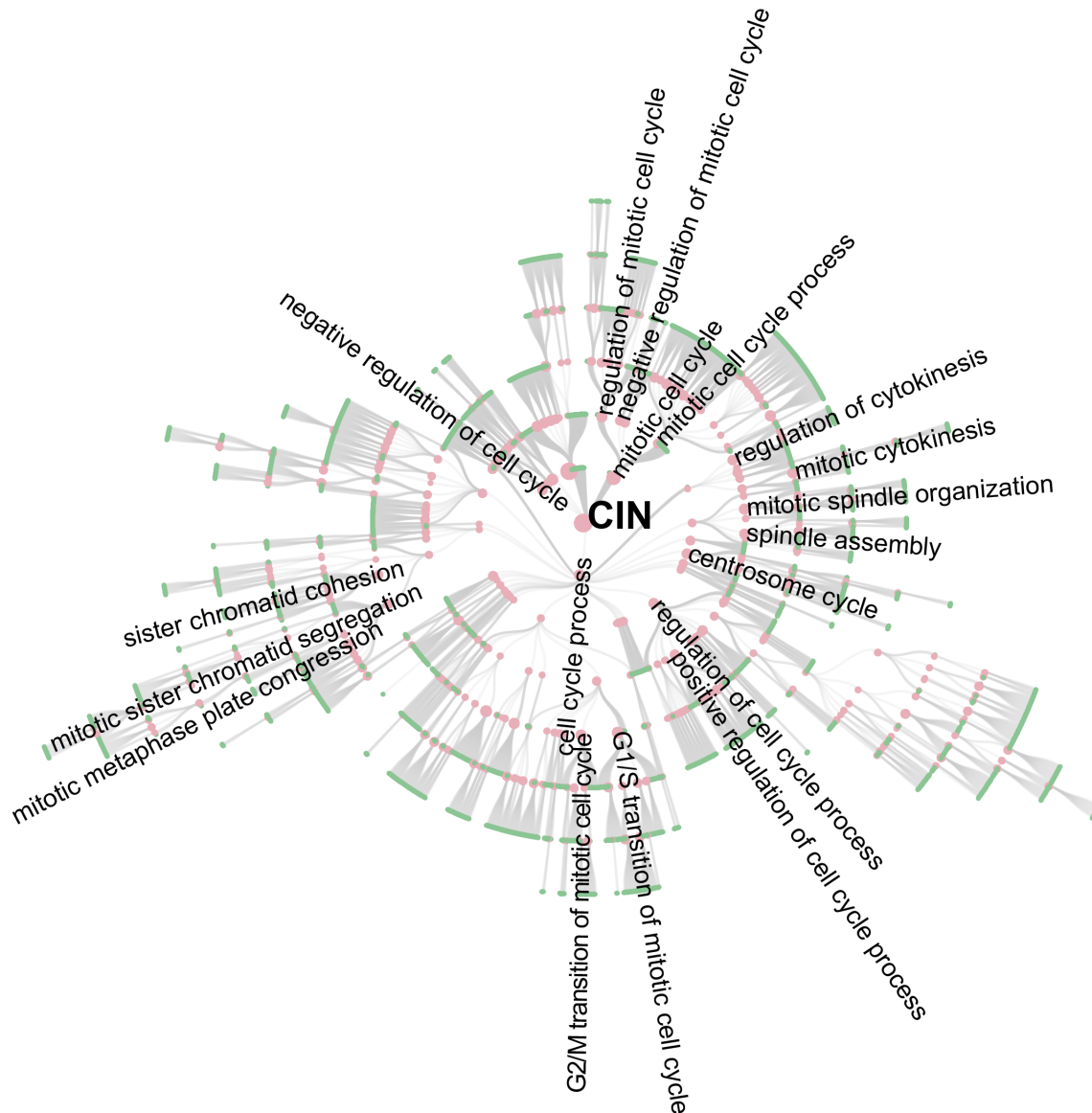


Figure 5.1: The design of CINgo architecture. CINgo is a feedforward neural network of which architecture mirrors the cell division related GO substructure. Pink dots represent pathways, green dots represent genes. Dot size encodes out degree, only representative pathways with large out-degrees are annotated.

5.2.2 Self-supervised and semi-supervised learning on tabular data

In spite of the lack of interpretability, deep learning has the strength in integrating multi-level heterogeneous data. Additionally, a variety of self-supervised learning and semi-supervised learning frameworks have achieved great advances in computer vision and nature language processing, some are even superior to supervised learning [143, 144]. Given the large amount of unlabelled datasets generated by big cancer genome projects, self-supervised and semi-supervised learning are urgently needed. The MFmap presented in Chapter 4 is a semi-supervised deep generative model for multi-modal cancer genome where the majority of the data are tabular data. For this type of data, most successful self-supervised and semi-supervised learning algorithms that make use of the semantic or spatial structures presented in languages or images do not work effectively [145, 146, 147]. A future direction would be how to extend the general self-supervised or semi-supervised model to tabular data? Since this question is understudied, we may face a lot of challenges on experiment setup, evaluation and visualisation. For resolving these challenges, we need to answer several important questions: How to preprocess the sparse and unbalanced categorical data (e.g. binary somatic mutation matrix)? How to design sufficiently difficult pretext tasks for tabular data? How to design appropriate evaluation metrics? How to display and interpret the learned representations? How to judge whether a learned representation is clinically and biologically meaningful?

5.2.3 Extending MFmap to regression

Recall that the MFmap presented in Chapter 4 is a new variant of VAE extending to semi-supervised learning. We have demonstrated that MFmap is able to learn cancer subtype specific features using *in silico* perturbation analysis and association analysis between latent representations and pathway activities as well as drug sensitivities. All of these good results could benefit from (i) large inter-class distance and small intra-class distance in the latent representation space; (ii) MFmap learned latent representations that are invariant between the labelled tumour samples and unlabelled cell line samples. These two aspects could also explain why MFmap achieves high data-integration performance (see <https://doi.org/10.1371/journal.pone.0261183.s002>). However, a large number of target variables in cancer genome research are continuous numbers (e.g. drug sensitivities or gene dependencies), the next interesting question would be how to extend the MFmap for regression problems? For this, the evaluation designation is challenging because we lack for validated biology experimental data. Using drug sensitivity prediction as example, we could design an *in silico* experiment by perturbing the latent representations

of a sample with low [area under the dose–response curve \(AUC\)](#) values to ones with other AUC values and reconstruct the gene profile of these generated artificial samples. We can select the genes with greater fold changes using differential gene profile analysis and compare it to experimental validation results which are often not available and are expensive to obtain. Altogether, extending MFmap to regression is useful, but the latent representation interpretation and evaluation require new biological experimental data.

5.2.4 Extending MFmap to multi-task learning

Considering the substantial complexity of cancer CIN, a practical way to target CIN might be to stratify patients into subtypes and cure each subtype based on their clinical and genetic characteristics. This is known as precision oncology, accurate and trustworthy treatment response prediction is essential to maximise the benefit of precision medicine. For this purpose, a number of pre-clinical pharmacogenomics datasets have been generated [26, 27, 28, 29, 30, 31, 32]. And various computational models have been developed to predict treatment responses using these pharmacogenomics data (reviewed in [148]). Even though such concept and execution of precision medicine sound rationale, many clinical trials fail to develop effective drugs. Although many factors contribute to the failure, a drug response predictive model could significantly improve the translational capability between preclinical models and patient treatments by taking into account the following input and output distribution differences between them. The input distribution difference can be addressed by [out-of-distribution \(OOD\)](#) generalization methodologies such as representation learning [149] and invariant risk minimization [150]. The output discrepancies: the AUC values for cell lines versus binary response readouts for patients can be addressed by multi-task learning. MFmap as a representation learning tool is able to extract clinically and biologically meaningful latent representations that are invariant between patients and cell lines. One more step is to extend MFmap for multi-task learning by incorporating a regression subnetwork predicting treatment response of cell lines and a classification subnetwork predicting patient treatment response. Again we also need to solve the the evaluation problem discussed in Subsection 5.2.3.

Appendices

List of publications

The following publications were written *during this thesis*.

Peer-reviewed journal articles

* indicates equal contributions

- **X. Zhang** and M. Kschischo. “Distinct and Common Features of Numerical and Structural Chromosomal Instability across Different Cancer Types.” In: *Cancers* 14.6 (2022). DOI: [10.3390/cancers14061424](https://doi.org/10.3390/cancers14061424).
- N. K. Chunduri, P. Menges, **X. Zhang**, A. Wieland, V. L. Gotsmann, B. R. Mardin, C. Buccitelli, J. O. Korb, F. Willmund, M. Kschischo, M. Raeschle, and Z. Storchova. “Systems approaches identify the consequences of monosomy in somatic human cells.” In: *Nature Communications* 12.1 (2021), p. 5576. DOI: [10.1038/s41467-021-25288-x](https://doi.org/10.1038/s41467-021-25288-x).
- **X. Zhang** and M. Kschischo. “MFmap: A semi-supervised generative model matching cell lines to tumours and cancer subtypes.” In: *PLOS ONE* 16.12 (2021), pp. 1–21. DOI: [10.1371/journal.pone.0261183](https://doi.org/10.1371/journal.pone.0261183).
- **X. Zhang**, H. Fröhlich, D. Grigoriev, S. Vakulenko, J. Zimmermann, and A. G. Weber. “A Simple 3-Parameter Model for Cancer Incidences.” In: *Scientific Reports* 8.1 (2018), p. 3388. DOI: [10.1038/s41598-018-21734-x](https://doi.org/10.1038/s41598-018-21734-x).
- T. Wang*, G. Wang*, **X. Zhang***, D. Wu, L. Yang, G. Wang, and D. Hao. “The expression of miRNAs is associated with tumour genome instability and predicts the outcome of ovarian cancer patients treated with platinum agents.” In: *Scientific Reports* 7.1 (2017), p. 14736. DOI: [10.1038/s41598-017-12259-w](https://doi.org/10.1038/s41598-017-12259-w).



Bibliography

- [1] D. Hanahan and R. A. Weinberg. “The Hallmarks of Cancer.” In: *Cell* 100.1 (2000), pp. 57–70.
- [2] D. Hanahan and R. A. Weinberg. “Hallmarks of Cancer: The Next Generation.” In: *Cell* 144.5 (2011), pp. 646–674.
- [3] D. Hanahan. “Hallmarks of Cancer: New Dimensions.” In: *Cancer Discovery* 12.1 (2022), pp. 31–46.
- [4] L. Sansregret, B. Vanhaesebroeck, and C. Swanton. “Determinants and clinical implications of chromosomal instability in cancer.” In: *Nature Reviews Clinical Oncology* 15.3 (2018), pp. 139–150.
- [5] N. McGranahan, R. A. Burrell, D. Endesfelder, M. R. Novelli, and C. Swanton. “Cancer chromosomal instability: therapeutic and diagnostic challenges.” In: *EMBO reports* 13.6 (2012), pp. 528–538.
- [6] S. O. Siri, J. Martino, and V. Gottifredi. “Structural Chromosome Instability: Types, Origins, Consequences, and Therapeutic Opportunities.” In: *Cancers* 13.12 (2021), p. 3056.
- [7] A. J. Lee, D. Endesfelder, A. J. Rowan, et al. “Chromosomal Instability Confers Intrinsic Multidrug Resistance.” In: *Cancer Research* 71.5 (2011), pp. 1858–1870.
- [8] D. A. Lukow and J. M. Sheltzer. “Chromosomal instability and aneuploidy as causes of cancer drug resistance.” In: *Trends in Cancer* 8.1 (2022), pp. 43–53.
- [9] D. A. Lukow, E. L. Sausville, P. Suri, et al. “Chromosomal instability accelerates the evolution of resistance to anti-cancer therapies.” In: *Developmental Cell* 56.17 (2021), 2427–2439.e4.
- [10] C. Swanton, B. Nicke, M. Schuett, et al. “Chromosomal instability determines taxane response.” In: *Proceedings of the National Academy of Sciences* 106.21 (2009), pp. 8671–8676.

- [11] C.-M. Choi, K. W. Seo, S. J. Jang, et al. "Chromosomal instability is a risk factor for poor prognosis of adenocarcinoma of the lung: Fluorescence in situ hybridization analysis of paraffin-embedded tissue from Korean patients." In: *Lung Cancer* 64.1 (2009), pp. 66–70.
- [12] S. F. Bakhoun, O. V. Danilova, P. Kaur, N. B. Levy, and D. A. Compton. "Chromosomal Instability Substantiates Poor Prognosis in Patients with Diffuse Large B-cell Lymphoma." In: *Clinical Cancer Research* 17.24 (2011), pp. 7704–7711.
- [13] S. F. Bakhoun, B. Ngo, A. M. Laughney, et al. "Chromosomal instability drives metastasis through a cytosolic DNA response." In: *Nature* 553.7689 (2018), pp. 467–472.
- [14] C. Gao, Y. Su, J. Koeman, et al. "Chromosome instability drives phenotypic switching to metastasis." In: *Proceedings of the National Academy of Sciences* 113.51 (2016), pp. 14793–14798.
- [15] S. F. Bakhoun and D. A. Landau. "Chromosomal Instability as a Driver of Tumor Heterogeneity and Evolution." In: *Cold Spring Harbor Perspectives in Medicine* 7.6 (2017).
- [16] S.-L. Chang, H.-Y. Lai, S.-Y. Tung, and J.-Y. Leu. "Dynamic Large-Scale Chromosomal Rearrangements Fuel Rapid Adaptation in Yeast Populations." In: *PLOS Genetics* 9.1 (2013), pp. 1–15.
- [17] G. Rancati, N. Pavelka, B. Fleharty, et al. "Aneuploidy Underlies Rapid Adaptive Evolution of Yeast Cells Deprived of a Conserved Cytokinesis Motor." In: *Cell* 135.5 (2008), pp. 879–893.
- [18] R. R. Beach, C. Ricci-Tam, C. M. Brennan, et al. "Aneuploidy Causes Non-genetic Individuality." In: *Cell* 169.2 (2017), 229–242.e21.
- [19] M. Gerlinger, A. J. Rowan, S. Horswell, et al. "Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing." In: *New England Journal of Medicine* 366.10 (2012), pp. 883–892.
- [20] A. Bashashati, G. Ha, A. Tone, et al. "Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling." In: *The Journal of Pathology* 231.1 (2013), pp. 21–34.
- [21] M. Castellarin, K. Milne, T. Zeng, et al. "Clonal evolution of high-grade serous ovarian carcinoma from primary to recurrent disease." In: *The Journal of Pathology* 229.4 (2013), pp. 515–524.

- [22] J.-M. Schvartzman, R. Sotillo, and R. Benezra. "Mitotic chromosomal instability and cancer: mouse modelling of the human disease." In: *Nature Reviews Cancer* 10.2 (2010), pp. 102–115.
- [23] S. J. Pfau and A. Amon. "Chromosomal instability and aneuploidy in cancer: from yeast to man." In: *EMBO reports* 13.6 (2012), pp. 515–527.
- [24] T. J. Hudson (Chairperson), W. Anderson, A. Aretz, et al. "International network of cancer genome projects." In: *Nature* 464.7291 (2010), pp. 993–998.
- [25] K. Chang, C. J. Creighton, C. Davis, et al. "The Cancer Genome Atlas Pan-Cancer analysis project." In: *Nature Genetics* 45.10 (2013), pp. 1113–1120.
- [26] R. H. Shoemaker. "The NCI60 human tumour cell line anticancer drug screen." In: *Nature Reviews Cancer* 6.10 (2006), pp. 813–823.
- [27] J. Lamb, E. D. Crawford, D. Peck, et al. "The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease." In: *Science* 313.5795 (2006), pp. 1929–1935.
- [28] J. Barretina, G. Caponigro, N. Stransky, et al. "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity." In: *Nature* 483.7391 (2012), pp. 603–607.
- [29] M. Ghandi, F. W. Huang, J. Jané-Valbuena, et al. "Next-generation characterization of the Cancer Cell Line Encyclopedia." In: *Nature* 569.7757 (2019), pp. 503–508.
- [30] B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, et al. "Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset." In: *Cancer Discovery* 5.11 (2015), pp. 1210–1223.
- [31] A. Basu, N. E. Bodycombe, J. H. Cheah, et al. "An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules." In: *Cell* 154.5 (2013), pp. 1151–1161.
- [32] F. Iorio, T. A. Knijnenburg, D. J. Vis, et al. "A Landscape of Pharmacogenomic Interactions in Cancer." In: *Cell* 166.3 (2016), pp. 740–754.
- [33] C. Buccitelli, L. Salgueiro, K. Rowald, R. Sotillo, B. R. Mardin, and J. O. Korbel. "Pan-cancer analysis distinguishes transcriptional changes of aneuploidy from proliferation." In: *Genome Research* 27.4 (2017), pp. 501–511.
- [34] S. L. Carter, A. C. Eklund, I. S. Kohane, L. N. Harris, and Z. Szallasi. "A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers." In: *Nature Genetics* 38.9 (2006), pp. 1043–1048.

- [35] D. Endesfelder, R. A. Burrell, N. Kanu, et al. "Chromosomal Instability Selects Gene Copy-Number Variants Encoding Core Regulators of Proliferation in ER+ Breast Cancer." In: *Cancer Research* 74.17 (2014), pp. 4853–4863.
- [36] J. M. Sheltzer. "A Transcriptional and Metabolic Signature of Primary Aneuploidy Is Present in Chromosomally Unstable Cancer Cells and Informs Clinical Prognosis." In: *Cancer Research* 73.21 (2013), pp. 6401–6412.
- [37] M. Gerstung, C. Jolly, I. Leshchiner, et al. "The evolutionary history of 2,658 cancers." In: *Nature* 578.7793 (2020), pp. 122–128.
- [38] S. C. Dentro, I. Leshchiner, K. Haase, et al. "Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes." In: *Cell* 184.8 (2021), 2239–2254.e39.
- [39] A. R. J. Lawson, F. Abascal, T. H. H. Coorens, et al. "Extensive heterogeneity in somatic mutation and selection in the human bladder." In: *Science* 370.6512 (2020), pp. 75–82.
- [40] M. Liu, H. An, Y. Zhang, et al. "Molecular analysis of Chinese oesophageal squamous cell carcinoma identifies novel subtypes associated with distinct clinical outcomes." In: *eBioMedicine* 57 (2020), p. 102831.
- [41] M. S. Lawrence, P. Stojanov, P. Polak, et al. "Mutational heterogeneity in cancer and the search for new cancer-associated genes." In: *Nature* 499.7457 (2013), pp. 214–218.
- [42] C. Kandoth, M. D. McLellan, F. Vandin, et al. "Mutational landscape and significance across 12 major cancer types." In: *Nature* 502.7471 (2013), pp. 333–339.
- [43] N. Andor, T. A. Graham, M. Jansen, et al. "Pan-cancer analysis of the extent and consequences of intratumor heterogeneity." In: *Nature Medicine* 22.1 (2016), pp. 105–113.
- [44] P. J. Campbell, G. Getz, J. O. Korb, et al. "Pan-cancer analysis of whole genomes." In: *Nature* 578.7793 (2020), pp. 82–93.
- [45] A. K. Arrington, E. L. Heinrich, W. Lee, et al. "Prognostic and Predictive Roles of KRAS Mutation in Colorectal Cancer." In: *International Journal of Molecular Sciences* 13.10 (2012), pp. 12153–12168.
- [46] M. Porru, L. Pompili, C. Caruso, A. Biroccio, and C. Leonetti. "Targeting KRAS in metastatic colorectal cancer: current strategies and emerging opportunities." In: *Journal of Experimental & Clinical Cancer Research* 37.1 (2018), p. 57.

- [47] J. Guinney, R. Dienstmann, X. Wang, et al. "The consensus molecular subtypes of colorectal cancer." In: *Nature Medicine* 21.11 (2015), pp. 1350–1356.
- [48] S. Z. Millis, S. Ikeda, S. Reddy, Z. Gatalica, and R. Kurzrock. "Landscape of Phosphatidylinositol-3-Kinase Pathway Alterations Across 19784 Diverse Solid Tumors." In: *JAMA Oncology* 2.12 (2016), pp. 1565–1573.
- [49] N. Rivlin, R. Brosh, M. Oren, and V. Rotter. "Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis." In: *Genes & Cancer* 2.4 (2011), pp. 466–474.
- [50] M. Meng, K. Zhong, T. Jiang, Z. Liu, H. Y. Kwan, and T. Su. "The current understanding on the impact of KRAS on colorectal cancer." In: *Biomedicine & Pharmacotherapy* 140 (2021), p. 111717.
- [51] L. B. Alexandrov, J. Kim, N. J. Haradhvala, et al. "The repertoire of mutational signatures in human cancer." In: *Nature* 578.7793 (2020), pp. 94–101.
- [52] G. El Tekle, T. Bernasocchi, A. M. Unni, et al. "Co-occurrence and mutual exclusivity: what cross-cancer mutation patterns can tell us." In: *Trends in Cancer* 7.9 (2021), pp. 823–836.
- [53] R. Nussinov, H. Jang, C.-J. Tsai, and F. Cheng. "Review: Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers." In: *PLOS Computational Biology* 15.3 (2019), e1006658.
- [54] L. Huang, Z. Guo, F. Wang, and L. Fu. "KRAS mutation: from undruggable to druggable in cancer." In: *Signal Transduction and Targeted Therapy* 6.1 (2021), p. 386.
- [55] A. Nair, S. Chakraborty, L. A. Banerji, A. Srivastava, C. Navare, and B. Saha. "Ras isoforms: signaling specificities in CD40 pathway." In: *Cell Communication and Signaling* 18.1 (2020), p. 3.
- [56] M. Ghomlaghi, A. Hart, N. Hoang, S. Shin, and L. K. Nguyen. "Feedback, Crosstalk and Competition: Ingredients for Emergent Non-Linear Behaviour in the PI3K/mTOR Signalling Network." In: *International Journal of Molecular Sciences* 22.13 (2021), p. 6944.
- [57] D. Öhlund, A. Handly-Santana, G. Biffi, et al. "Distinct populations of inflammatory fibroblasts and myofibroblasts in pancreatic cancer." In: *Journal of Experimental Medicine* 214.3 (2017), pp. 579–596.

- [58] E. Elyada, M. Bolisetty, P. Laise, et al. “Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts.” In: *Cancer Discovery* 9.8 (2019), pp. 1102–1123.
- [59] K. Wu, K. Lin, X. Li, et al. “Redefining Tumor-Associated Macrophage Subpopulations and Functions in the Tumor Microenvironment.” In: *Frontiers in Immunology* 11 (2020), p. 1731.
- [60] D. Chraa, A. Naim, D. Olive, and A. Badou. “T lymphocyte subsets in cancer immunity: Friends or foes.” In: *Journal of Leukocyte Biology* 105.2 (2019), pp. 243–255.
- [61] J. Saltz, R. Gupta, L. Hou, et al. “Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images.” In: *Cell Reports* 23.1 (2018), 181–193.e7.
- [62] D. S. Chen and I. Mellman. “Elements of cancer immunity and the cancer-immune set point.” In: *Nature* 541.7637 (2017), pp. 321–330.
- [63] A. Malandrino, M. Mak, R. D. Kamm, and E. Moeendarbary. “Complex mechanics of the heterogeneous extracellular matrix in cancer.” In: *Extreme Mechanics Letters* 21 (2018), pp. 25–34.
- [64] D. Hammerl, J. W. M. Martens, M. Timmermans, et al. “Spatial immunophenotypes predict response to anti-PD1 treatment and capture distinct paths of T cell evasion in triple negative breast cancer.” In: *Nature Communications* 12.1 (2021), p. 5668.
- [65] V. Cremasco, J. L. Astarita, A. L. Grauel, et al. “FAP Delineates Heterogeneous and Functionally Divergent Stromal Cells in Immune-Excluded Breast Tumors.” In: *Cancer Immunology Research* 6.12 (2018), pp. 1472–1485.
- [66] J. Winkler, A. Abisoye-Ogunniyan, K. J. Metcalf, and Z. Werb. “Concepts of extracellular matrix remodelling in tumour progression and metastasis.” In: *Nature Communications* 11.1 (2020), p. 5120.
- [67] T. R. Kress, A. Sabò, and B. Amati. “MYC: connecting selective transcriptional control to global RNA production.” In: *Nature Reviews Cancer* 15.10 (2015), pp. 593–607.
- [68] R. Dhanasekaran, A. Deutzmann, W. D. Mahauad-Fernandez, A. S. Hansen, A. M. Gouw, and D. W. Felsher. “The MYC oncogene — the grand orchestrator of cancer growth and immune evasion.” In: *Nature Reviews Clinical Oncology* 19.1 (2022), pp. 23–36.

- [69] N. McGranahan and C. Swanton. "Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future." In: *Cell* 168.4 (2017), pp. 613–628.
- [70] C. D. DiNardo, I. S. Tiong, A. Quaglieri, et al. "Molecular patterns of response and treatment failure after frontline venetoclax combinations in older patients with AML." In: *Blood* 135.11 (2020), pp. 791–803.
- [71] J. Malcikova, S. Pavlova, B. Kunt Vonkova, et al. "Low-burden TP53 mutations in CLL: clinical impact and clonal evolution within the context of different treatment options." In: *Blood* 138.25 (2021), pp. 2670–2685.
- [72] M. J. J. Rose-Zerilli, J. Gibson, J. Wang, et al. "Longitudinal copy number, whole exome and targeted deep sequencing of 'good risk' IGHV-mutated CLL patients with progressive disease." In: *Leukemia* 30.6 (2016), pp. 1301–1310.
- [73] M. Angelova, B. Mlecnik, A. Vasaturo, et al. "Evolution of Metastases in Space and Time under Immune Selection." In: *Cell* 175.3 (2018). Publisher: Elsevier, 751–765.e16.
- [74] A. Marusyk, M. Janiszewska, and K. Polyak. "Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance." In: *Cancer Cell* 37.4 (2020), pp. 471–484.
- [75] Z.-F. Lim and P. C. Ma. "Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy." In: *Journal of Hematology & Oncology* 12.1 (2019), p. 134.
- [76] R. Somasundaram, J. Villanueva, and M. Herlyn. "Intratumoral heterogeneity as a therapy resistance mechanism: role of melanoma subpopulations." In: *Advances in Pharmacology* 65 (2012), pp. 335–359.
- [77] M. Hölzel, A. Bovier, and T. Tüting. "Plasticity of tumour and immune cells: a source of heterogeneity and a cause for therapy resistance?" In: *Nature Reviews Cancer* 13.5 (2013), pp. 365–376.
- [78] M. DuPage and J. A. Bluestone. "Harnessing the plasticity of CD4+ T cells to treat immune-mediated disease." In: *Nature Reviews Immunology* 16.3 (2016), pp. 149–163.
- [79] S. Xiong, Y. Feng, and L. Cheng. "Cellular Reprogramming as a Therapeutic Target in Cancer." In: *Trends in Cell Biology* 29.8 (2019), pp. 623–634.
- [80] S. Turajlic, A. Sottoriva, T. Graham, and C. Swanton. "Resolving genetic heterogeneity in cancer." In: *Nature Reviews Genetics* 20.7 (2019), pp. 404–416.
- [81] R. A. Burrell, N. McGranahan, J. Bartek, and C. Swanton. "The causes and consequences of genetic heterogeneity in cancer evolution." In: *Nature* 501.7467 (2013), pp. 338–345.

- [82] R. H. van Jaarsveld and G. J. P. L. Kops. "Difference Makers: Chromosomal Instability versus Aneuploidy in Cancer." In: *Trends in Cancer* 2.10 (2016), pp. 561–571.
- [83] A. D. Silk, L. M. Zasadil, A. J. Holland, B. Vitre, D. W. Cleveland, and B. A. Weaver. "Chromosome missegregation rate predicts whether aneuploidy will promote or suppress tumors." In: *Proceedings of the National Academy of Sciences* 110.44 (2013), E4134–E4141.
- [84] N. J. Birkbak, A. C. Eklund, Q. Li, et al. "Paradoxical Relationship between Chromosomal Instability and Survival Outcome in Cancer." In: *Cancer Research* 71.10 (2011), pp. 3447–3452.
- [85] S. Negrini, V. G. Gorgoulis, and T. D. Halazonetis. "Genomic instability —an evolving hallmark of cancer." In: *Nature Reviews Molecular Cell Biology* 11.3 (2010), pp. 220–228.
- [86] J. B. Geigl, A. C. Obenauf, T. Schwarzbraun, and M. R. Speicher. "Defining 'chromosomal instability'." In: *Trends in Genetics* 24.2 (2008), pp. 64–69.
- [87] F. Mitelman, B. Johansson, N. Mandahl, and F. Mertens. "Clinical significance of cytogenetic findings in solid tumors." In: *Cancer Genetics and Cytogenetics* 95.1 (1997), pp. 1–8.
- [88] A. J. Holland and D. W. Cleveland. "Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis." In: *Nature Reviews Molecular Cell Biology* 10.7 (2009), pp. 478–487.
- [89] J. M. Nicholson, J. C. Macedo, A. J. Mattingly, et al. "Chromosome mis-segregation and cytokinesis failure in trisomic human cells." In: *eLife* 4 (2015), e05068.
- [90] A. Valind, Y. Jin, B. Baldetorp, and D. Gisselsson. "Whole chromosome gain does not in itself confer cancer-like chromosomal instability." In: *Proceedings of the National Academy of Sciences* 110.52 (2013), pp. 21119–21123.
- [91] S. L. Thompson, S. F. Bakhom, and D. A. Compton. "Mechanisms of Chromosomal Instability." In: *Current Biology* 20.6 (2010), R285–R295.
- [92] A. Janssen, M. van der Burg, K. Szuhai, G. J. P. L. Kops, and R. H. Medema. "Chromosome Segregation Errors as a Cause of DNA Damage and Structural Chromosome Aberrations." In: *Science* 333.6051 (2011), pp. 1895–1898.
- [93] S. F. Bakhom, L. Kabeche, J. P. Murnane, B. I. Zaki, and D. A. Compton. "DNA-Damage Response during Mitosis Induces Whole-Chromosome Missegregation." In: *Cancer Discovery* 4.11 (2014), pp. 1281–1289.

- [94] R. A. Burrell, S. E. McClelland, D. Endesfelder, et al. "Replication stress links structural and numerical cancer chromosomal instability." In: *Nature* 494.7438 (2013), pp. 492–496.
- [95] R. Beroukhim, C. H. Mermel, D. Porter, et al. "The landscape of somatic copy-number alteration across human cancers." In: *Nature* 463.7283 (2010), pp. 899–905.
- [96] A. Janssen and R. H. Medema. "Genetic instability: tipping the balance." In: *Oncogene* 32.38 (2013), pp. 4459–4470.
- [97] M. K. Zeman and K. A. Cimprich. "Causes and consequences of replication stress." In: *Nature Cell Biology* 16.1 (2014), pp. 2–9.
- [98] M. Marotta, X. Chen, T. Watanabe, et al. "Homology-mediated end-capping as a primary step of sister chromatid fusion in the breakage-fusion-bridge cycles." In: *Nucleic Acids Research* 41.21 (2013), pp. 9732–9740.
- [99] A. Mazouzi, G. Velimezi, and J. I. Loizou. "DNA replication stress: Causes, resolution and disease." In: *Experimental Cell Research* 329.1 (2014), pp. 85–93.
- [100] S. Santaguida and A. Amon. "Short- and long-term effects of chromosome mis-segregation and aneuploidy." In: *Nature Reviews Molecular Cell Biology* 16.8 (2015), pp. 473–485.
- [101] B. Xu, Z. Sun, Z. Liu, et al. "Replication Stress Induces Micronuclei Comprising of Aggregated DNA Double-Strand Breaks." In: *PLOS ONE* 6.4 (2011), pp. 1–11.
- [102] T. Kawabata, S. W. Luebben, S. Yamaguchi, et al. "Stalled Fork Rescue via Dormant Replication Origins in Unchallenged S Phase Promotes Proper Chromosome Segregation and Tumor Suppression." In: *Molecular Cell* 41.5 (2011), pp. 543–553.
- [103] X. Zhang and M. Kschischo. "Distinct and Common Features of Numerical and Structural Chromosomal Instability across Different Cancer Types." In: *Cancers* 14.6 (2022).
- [104] R. Roylance, D. Endesfelder, P. Gorman, et al. "Relationship of Extreme Chromosomal Instability with Long-term Survival in a Retrospective Analysis of Primary Breast Cancer." In: *Cancer Epidemiology, Biomarkers & Prevention* 20.10 (2011), pp. 2183–2194.
- [105] T. B. K. Watkins, E. L. Lim, M. Petkovic, et al. "Pervasive chromosomal instability and karyotype order in tumour evolution." In: *Nature* 587.7832 (2020), pp. 126–132.
- [106] U. Ben-David and A. Amon. "Context is everything: aneuploidy in cancer." In: *Nature Reviews Genetics* 21.1 (2020), pp. 44–62.

- [107] A. M. Taylor, J. Shih, G. Ha, et al. “Genomic and Functional Approaches to Understanding Cancer Aneuploidy.” In: *Cancer Cell* 33.4 (2018), 676–689.e3.
- [108] V. Passerini, E. Ozeri-Galai, M. S. de Pagter, et al. “The presence of extra chromosomes leads to genomic instability.” In: *Nature Communications* 7.1 (2016), p. 10754.
- [109] R. M. Ricke, J. H. van Ree, and J. M. van Deursen. “Whole chromosome instability and cancer: a complex relationship.” In: *Trends in Genetics* 24.9 (2008), pp. 457–466.
- [110] C. C. Lepage, C. R. Morden, M. C. L. Palmer, M. W. Nachtigal, and K. J. McManus. “Detecting Chromosome Instability in Cancer: Approaches to Resolve Cell-to-Cell Heterogeneity.” In: *Cancers* 11.2 (2019), p. 226.
- [111] F. Zare, M. Dow, N. Monteleone, A. Hosny, and S. Nabavi. “An evaluation of copy number variation detection tools for cancer using whole exome sequencing data.” In: *BMC Bioinformatics* 18.1 (2017), p. 286.
- [112] D. de Ridder, C. E. van der Linden, T. Schonewille, et al. “Purity for clarity: the need for purification of tumor cells in DNA microarray studies.” In: *Leukemia* 19.4 (2005), pp. 618–627.
- [113] M. Meyerson, S. Gabriel, and G. Getz. “Advances in understanding cancer genomes through second-generation sequencing.” In: *Nature Reviews Genetics* 11.10 (2010), pp. 685–696.
- [114] M. Sharma, O. Saha, A. Sriraman, R. Hebbalaguppe, L. Vig, and S. S. Karande. “Crowdsourcing for Chromosome Segmentation and Deep Classification.” In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017), pp. 786–793.
- [115] M. S. Al-Kharraz, L. A. Elrefaei, and M. A. Fadel. “Automated System for Chromosome Karyotyping to Recognize the Most Common Numerical Abnormalities Using Deep Learning.” In: *IEEE Access* 8 (2020), pp. 157727–157747.
- [116] N. Xie, X. Li, K. Li, Y. Yang, and H. T. Shen. “Statistical Karyotype Analysis Using CNN and Geometric Optimization.” In: *IEEE Access* 7 (2019), pp. 179445–179453.
- [117] S. F. Chin, A. E. Teschendorff, J. C. Marioni, et al. “High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer.” In: *Genome Biology* 8.10 (2007), R215.
- [118] A. Shukla, T. H. M. Nguyen, S. B. Moka, et al. “Chromosome arm aneuploidies shape tumour evolution and drug response.” In: *Nature Communications* 11.1 (2020), p. 449.

- [119] T. Davoli, H. Uno, E. C. Wooten, and S. J. Elledge. “Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy.” In: *Science* 355.6322 (2017), eaaf8399.
- [120] H. Hieronymus, R. Murali, A. Tin, et al. “Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death.” In: *eLife* 7 (2018), e37294.
- [121] S. Elizalde, A. M. Laughney, and S. F. Bakhoun. “A Markov chain for numerical chromosomal instability in clonally expanding populations.” In: *PLOS Computational Biology* 14.9 (2018), e1006447.
- [122] M. A. Nowak, N. L. Komarova, A. Sengupta, et al. “The role of chromosomal instability in tumor initiation.” In: *Proceedings of the National Academy of Sciences* 99.25 (2002), pp. 16226–16231.
- [123] Z. Xu, A. Verma, U. Naveed, S. F. Bakhoun, P. Khosravi, and O. Elemento. “Deep learning predicts chromosomal instability from histopathology images.” In: *iScience* 24.5 (2021), p. 102394.
- [124] M. Bilal, S. E. A. Raza, A. Azam, et al. “Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study.” In: *The Lancet Digital Health* 3.12 (2021), e763–e772.
- [125] I. Tinhofer, D. Braunholz, and K. Klinghammer. “Preclinical models of head and neck squamous cell carcinoma for a basic understanding of cancer biology and its translation into efficient therapies.” In: *Cancers of the Head & Neck* 5.1 (2020), p. 9.
- [126] A. A. Seyhan. “Lost in translation: the valley of death across preclinical and clinical divide –identification of problems and overcoming obstacles.” In: *Translational Medicine Communications* 4.1 (2019), p. 18.
- [127] S. K. Golombek, J.-N. May, B. Theek, et al. “Tumor targeting via EPR: Strategies to enhance patient responses.” In: *Advanced Drug Delivery Reviews* 130 (2018), pp. 17–38.
- [128] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.
- [129] K. P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.
- [130] D. P. Kingma and M. Welling. “Auto-Encoding Variational Bayes.” In: *arXiv:1312.6114 [Preprint]* (2013).

- [131] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. “Variational Inference: A Review for Statisticians.” In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.
- [132] D. P. Kingma and M. Welling. “An Introduction to Variational Autoencoders.” In: *Foundations and Trends in Machine Learning* 12.4 (2019), pp. 307–392.
- [133] J. Bergstra and Y. Bengio. “Random Search for Hyper-Parameter Optimization.” In: *Journal of Machine Learning Research* 13.10 (2012), pp. 281–305.
- [134] J. Snoek, H. Larochelle, and R. P. Adams. “Practical Bayesian Optimization of Machine Learning Algorithms.” In: *arXiv:1206.2944 [Preprint]* (2012).
- [135] N. D. Nguyen and D. Wang. “Multiview learning for understanding functional multiomics.” In: *PLOS Computational Biology* 16.4 (2020), pp. 1–26.
- [136] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. “Semi-Supervised Learning with Deep Generative Models.” In: *arXiv:1406.5298 [Preprint]* (2014).
- [137] H.-Z. Feng, K. Kong, M. Chen, T. Zhang, M. Zhu, and W. Chen. “SHOT-VAE: Semi-supervised Deep Generative Models With Label-aware ELBO Approximations.” In: *arXiv:2011.10684 [Preprint]* (2020).
- [138] J. Gordon and J. M. Hernández-Lobato. “Bayesian Semisupervised Learning with Deep Generative Models.” In: *arXiv:1706.09751 [Preprint]* (2017).
- [139] X. Zhang and M. Kschischo. “MFmap: A semi-supervised generative model matching cell lines to tumours and cancer subtypes.” In: *PLOS ONE* 16.12 (2021), pp. 1–21.
- [140] A.-K. Schmidt, N. Böhly, X. Zhang, et al. “Dormant replication origin firing links replication stress to whole chromosomal instability in human cancer.” In: *bioRxiv:463929 [Preprint]* (2021).
- [141] J. Ma, M. K. Yu, S. Fong, et al. “Using deep learning to model the hierarchical structure and function of a cell.” In: *Nature Methods* 15.4 (2018), pp. 290–298.
- [142] B. M. Kuenzi, J. Park, S. H. Fong, et al. “Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells.” In: *Cancer Cell* 38.5 (2020), 672–684.e6.
- [143] T. Wang, Z. Yue, J. Huang, Q. Sun, and H. Zhang. “Self-Supervised Learning Disentangled Group Representation as Feature.” In: *arXiv:2110.15255 [Preprint]* (2021).
- [144] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. “Masked Autoencoders Are Scalable Vision Learners.” In: *arXiv:2111.06377 [Preprint]* (2021).

- [145] R. Shwartz-Ziv and A. Armon. “Tabular data: Deep learning is not all you need.” In: *Information Fusion* 81 (2022), pp. 84–90.
- [146] J. Yoon, Y. Zhang, J. Jordon, and M. van der Schaar. “VIME: Extending the Success of Self- and Semi-supervised Learning to Tabular Domain.” In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 11033–11043.
- [147] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci. “Deep Neural Networks and Tabular Data: A Survey.” In: *arXiv:2110.01889 [Preprint]* (2021).
- [148] G. Adam, L. Rampášek, Z. Safikhani, P. Smirnov, B. Haibe-Kains, and A. Goldenberg. “Machine learning approaches to drug response prediction: challenges and recent progress.” In: *npj Precision Oncology* 4.1 (2020), p. 19.
- [149] K. Muandet, D. Balduzzi, and B. Schölkopf. “Domain Generalization via Invariant Feature Representation.” In: *arXiv:1301.2115 [Preprint]* (2013).
- [150] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. “Invariant Risk Minimization.” In: *arXiv:1907.02893 [Preprint]* (2019).